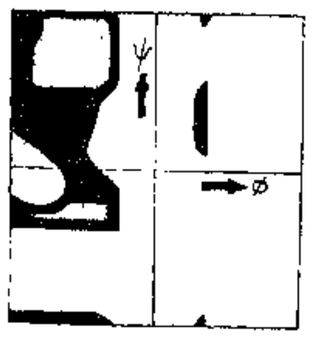


Special issue:
The Ramachandran map and
protein and peptide structures

Glycyl residues in proteins and peptides: An analysis

C. Ramakrishnan* and N. Srinivasan
Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560 012, India

Glycyl residue, simplest of all the residues, is well known for its conformational freedom. An analysis of the conformational and structural aspects of this residue occurring in proteins and peptides has been made making use of the Ramachandran (ϕ, ψ) angles and their distribution. The chief observations are: (i) By and large, there is no bias for an amino-acid residue to precede or succeed Gly. (ii) The conformational points show clustering in the 'bridge' region. (iii) While in general, glycyl residue plays a passive role when it occurs in a helix and helps in helix propagation, it also acts as a helix breaker in some instances. (iv) It is a poor former of extended strands. (v) The conformational freedom is effectively used by Gly to prefer those positions in turns that are less favourable for non-glycyl residues. (vi) Analysis of doublet data reinforces the propagative tendency of glycyl residues. X-Gly doublet is a better turn former than Gly-Y doublet, (vii) Only one third of the glycyl residues are situated on the surface of the proteins. The results can be useful in modelling studies on proteins.

ANALYSIS and documentation of protein backbone conformation are being widely carried out using the (ϕ,ψ) angles. The 'time-tested' Ramachandran map[1] provides an excellent tool to study the general stereochemical feasibility of the peptide backbone. The secondary structural features such as α-helix, β-sheet and β-turn can be represented by one or two points in the (ϕ,ψ) space[2] and this aspect enables one to quickly recognize such structural motifs in proteins. Thus the regularities in three-dimensional complexity of globular proteins are reduced on to a comprehendable two-dimensional Ramachandran (ϕ, ψ) plane. In the literature, the Ramachandran plot is commonly used to assess the stereochemical quality of any modelled protein. Glycyl residues (Gly), in most cases are not considered separately in spite of its high degree of conformational freedom.

Over the years, extensive attempts have been made to understand the significance of the role of Gly in various synthetic polypeptides and fibrous proteins like collagen and silk[3]. The poly(Gly) I adopts a β-sheet structure[4] and poly(Gly) II a three-fold left-handed triple-helical structure[5]. Silk (*Bombyx mori*) contains nearly 50% of Gly and the presence of glycine in almost every alternate position in the sequence facilitates close packing of β-sheets[6]. Gly occupies every third location of the collagen chains. The chains of the triple-helical collagen get closely packed due to the glycyl residues[7].

Glycyl has some nearly direct roles in the function of some peptides and proteins. For example, the sequence Arg–Gly–Asp–Ser is a part of the cell attachment domain of fibronectin[8]. It is also involved in the debatable mechanism of binding of carboxypeptidase with its ligand[9–11] and in the activation of oncogenic p21 protein where Gly-12 is crucial for activity[12].

In view of these interesting sequence and structural characteristics of Gly in various systems, the present study is motivated towards extracting its conformational peculiarities and positional preferences in crystal structures of proteins and peptides. The Ramachandran (ϕ,ψ) angles are extensively used as a potential tool to investigate the conformational aspects.

Glycyl residues in the primary structure

The occurrence of certain sequence patterns involving glycyl residues is well known for fibrous proteins such as collagen and silk. In the former, not only does glycyl form about one third of the total number of residues but occur as every third residue[7]. So also in β-proteins glycyl occurs more as . . . Gly-X–Gly-X–Gly-X. . . .

In order to find out whether there are any amino acids which have extreme bias of occurrence adjacent to glycyl residues in globular proteins, a data set comprising of 119 non-homologous proteins chosen from the protein data available in the Brookhaven Protein Data Bank[13] has been used. This set contains 2135 Gly occurring in these proteins. Considering a

*For correspondence.

triplet of the type X–Gly–Y, the ratio of the number of examples of any particular amino acid X when it occurs as X–Gly, to the total number of amino acids X in the set is worked out. This is also done for the Gly-Y doublets. These doublet data are given in Figure 1 as a bar diagram. The figure reveals that there is no outstanding bias for any amino acid. It is interesting to note that the percentages at position X and at Y are nearly the same for most of the residues. For a few, such as Cys, Pro and Trp, the X-Gly has a higher preference over Gly-Y and *vice versa* for a few others such as Arg, Ile and Phe. These conclusions are to be treated as more of an indicative rather than conclusive nature and need better substantiation using a much larger data set such as the one available in protein sequence bank.

The (ϕ,ψ) distribution of glycyl residues in peptides and proteins

The usefulness of the Ramachandran map in the study of protein conformation is now well established. The oft-quoted map is applicable to L-amino-acid residues, in which a considerable region is disallowed due to the presence of β-carbon atom. The map, appropriate for the glycyl residue, has also been worked[14] out as early as 1965. However, this map is not being referred to as often as the other map and in most cases the glycyl conformations are plotted in the alanyl map. Compared with the L-Ala map, the glycyl map is centrosymmetric with respect to the origin. In addition, the percentage of the (ϕ,ψ) region allowed by extreme limit, for Gly (57.4) is more than double that for Ala (20.3), spanning both the right and left halves of the map[14].

Information regarding the conformation of Gly occurring in small oligopeptides has been extracted from the literature and this is shown in Figure 2,*a* against the background of the Ramachandran glycyl map. The data have been collected from both linear and cyclic peptides. In the case of linear peptides only those examples where glycyl occurs as a 'middle' residue has been collected. The distribution brings out the following features:

(i) The conformations are clustered around specific regions. In particular, the points occupy the two 'bridge' regions on either side of the line $\psi = 0$ and another concentration around the extended structural region. The points largely conform to the low energy contour levels obtained by using a two-fold ψ (torsional) potential[15].

(ii) There is a conspicuous absence of conformation around α-helical region as well as around $\psi = -90$. A few points occurring around $\psi = 90°$ belong mostly to cyclic peptides. The percentage distribution of the points is shown in Figure 2,*c* as a grid map and from this it can be seen that 75% of the examples favour extended conformation.

(iii) As far as the parameter ψ is concerned, the values can be taken to be concentrated around $\psi = 0$ and 180°, and paucity of conformation around $\psi = \pm 90$. A histogram showing the distribution of ψ was given by Manjula *et al.*[16] based upon the data then available. This also shows that there is a peak around 0° and 180° of ψ.

(iv) The major portion of disallowed region occurring around $\phi = \psi = 0°$ is fully devoid of conformational points (except an isolated example). The region around ($\pm 180°$, 0°) which is also disallowed, is again free of conformational points.

A set of 65 proteins carefully chosen from the protein data bank[13] was used for the present analysis. The set is so chosen that only non-homologous proteins and the structures determined to a reasonably good accuracy (resolution $\leqslant$ 2 Å) are included. This data set contains various functional proteins such as immuno-globulins, globins, proteases, etc.

The distribution of the 994 glycyl conformations occurring is proteins in the present data set is shown in Figure 2,*b* along with the Ramachandran map for glycyl residue. The distribution plot of the glycyl conformations in the (ϕ,ψ) plane has been done by other workers also[17–24]. The percentage distribution of the (ϕ,ψ) plot is shown in Figure 2,*d*. It can be seen from the figure that the points are in general scattered over the entire allowed region of the map. Most of the points in fact lie within the allowed region of the Ramachandran glycyl map, with just about 20 points (2%) occurring in the core of disallowed region.

The (ϕ,ψ) map is obviously centrosymmetric about the origin ($\phi = \psi = 0°$) due to the inherent achiral nature of the glycyl residue. The distribution of points

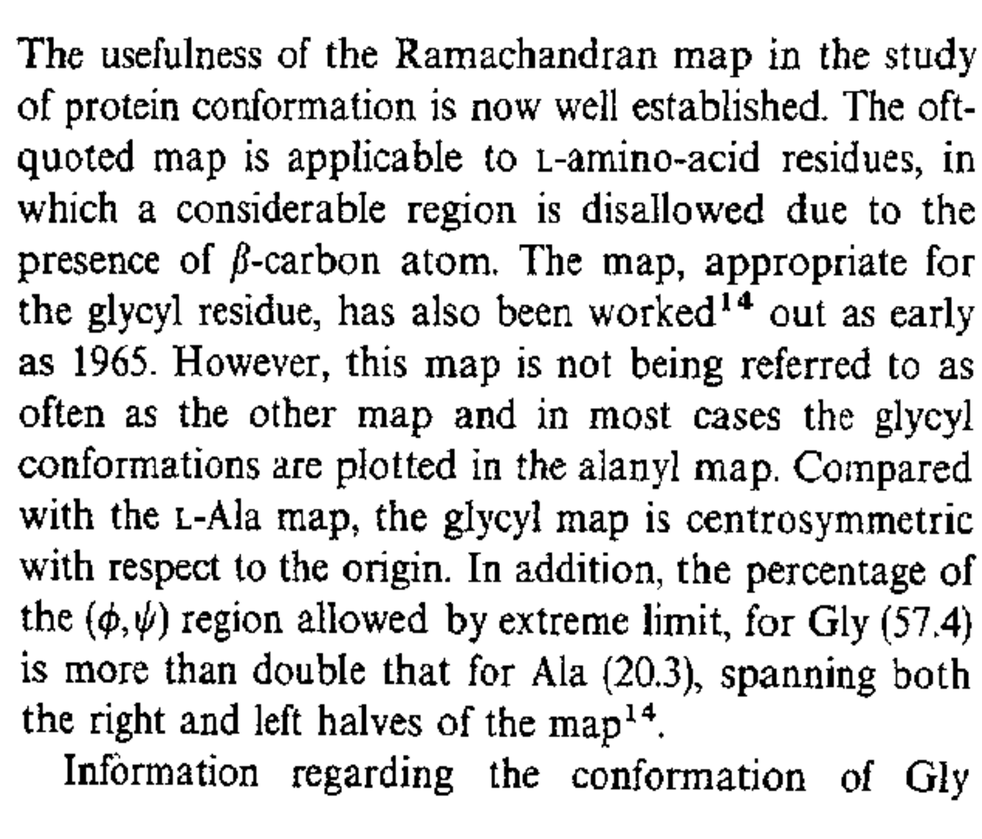
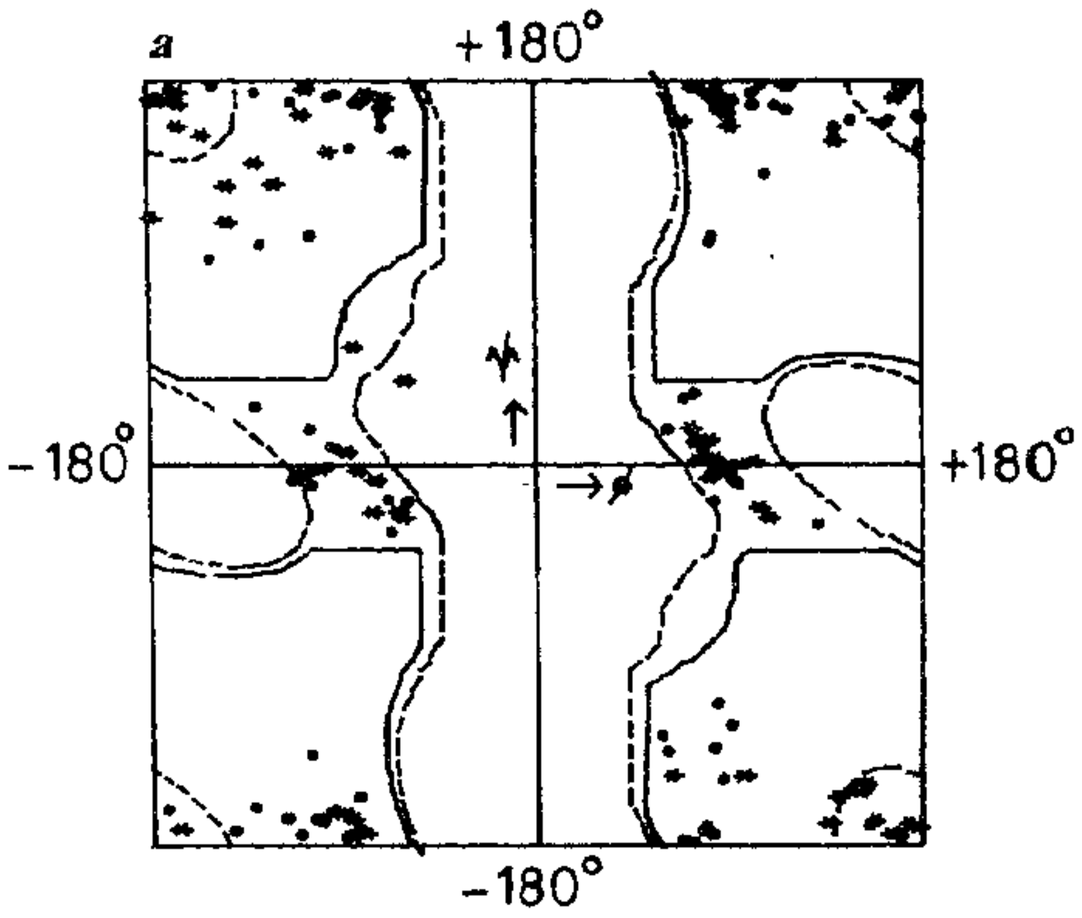
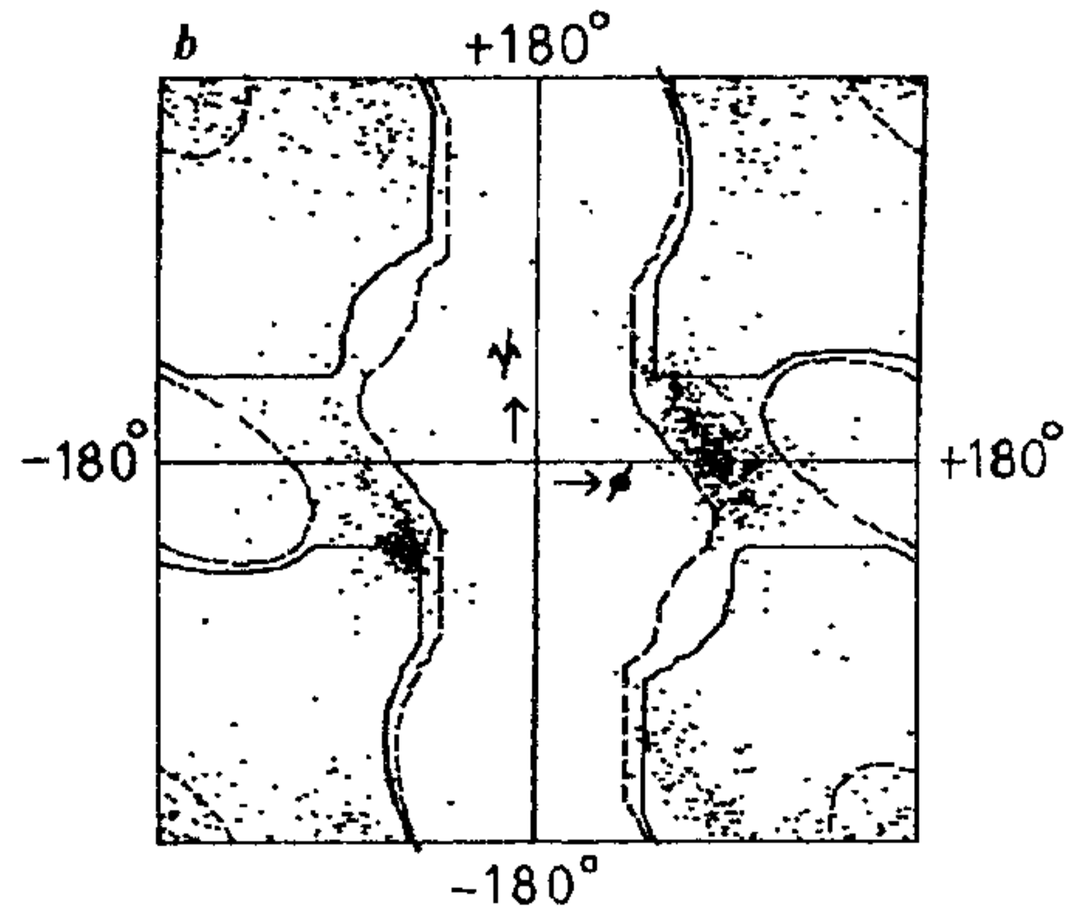


Figure 1. Normalized distribution (percentage) of number of occurrences of X-Gly and Gly-Y doublets in a chosen set of 119 unrelated proteins. The letters shown along the X-axis correspond to the single letter codes of the amino acids X or Y in X-Gly and Gly-Y doublets respectively.





c

+180°

6.4	0.6	4.5	5.1					5.1	5.8	2.6	5.1
	1.9	0.6	1.3						0.6		
1.3	1.3	0.6						1.3			
			1.3					1.3			
	0.6	1.9	0.6					5.1	1.3		
		3.2	4.5					2.6	3.8	0.6	
			0.6								
								0.6			
		0.6					0.6	2.6	1.3		
1.3	1.3	5.1	5.1					3.2		1.3	5.1

−180° +180° −180°

d

+180°

3.4	1.3	1.3	3.5					1.2	1.9	1.3	1.5
0.4	0.4	0.6	1.5	0.4	0.1	0.1		0.3	1.5	0.1	0.1
	0.1	0.1	0.2		0.1			0.1	0.2	0.1	0.1
	0.1	0.1	0.1	0.1		0.1	0.1	0.2	0.2		0.1
	0.1	0.3	0.3				1.5	3.2	0.1		
0.1	0.3	1.3	0.5	0.1		0.1	0.3	12.2	3.2	0.3	0.1
	0.3	0.6	3.5	0.4			0.1	4.6	6.4	0.5	
	0.1	0.2	6.5	3.9				0.2	0.6		0.1
0.1		0.1	0.2	0.5					0.4	0.1	
		0.1					0.5	0.3		0.2	
0.4	0.2	0.4	0.1				1.3	1.9	0.9	0.1	0.1
2.5	0.7	1.7	1.4	0.2			0.1	3.5	1.5	1.5	2.1

−180° +180° −180°

Figure 2. ***a***, The (ϕ, ψ) distribution of glycyl residues in peptides superimposed on the Ramachandran glycyl map (●–in cyclic peptides and *–in linear peptides). ***b***, The (ϕ, ψ) distribution of glycyl residues in proteins. ***c***, The *percentage* distribution of conformational points in grids of 30° of (ϕ, ψ) corresponding to ***a*** for peptides. Absence of any number in a square means that there are no examples occurring in that square. ***d***, Same as in ***c***, for the distribution given in ***b*** for proteins.

however is only roughly symmetric and there are certain obvious clusters that are definitely of an asymmetric nature. The most conspicuous cluster occurs in the bridge region, in the right half of the map. The extended strand regions (in the neighbourhood of $\phi = \psi = \pm 180°$) also contain some clusters. Quantitatively, about 30% of the conformations occur in the right bridge region, about 39% in the extended or near-extended regions and 17% in the left bridge region. The last category includes Gly in the right handed α-helical conformation (α_R) and the percentage in this category is 14.3.

Our earlier analysis on Gly in proteins was focused on extracting conformational preferences of Gly dependent on its adjoining residues. The pattern noticed is that, if Gly is sandwiched between two specific amino-acid residues, then the glycyl conformation prefers to lie either in the right or in the left half of the (ϕ, ψ) plane[22]. The present study attempts to extract conformational preferences or otherwise of Gly in the local regions of the Ramachandran glycyl map.

When the distribution of glycyl conformations in proteins given in Figure 2,*b* is compared with the corresponding distribution of peptide examples (Figure 2,*a*), it is possible to extract some common as well as differing features. The concentration of points in the bridge as well as in the extended region is common to both. However, the percentage of examples occurring in the extended region in peptides is considerably large and this can be taken to mean that the tendency of the glycyl residue to occur in extended strands is highly reduced when it occurs in proteins.

In both the maps, the clustering of points in the bridge region around (90°, 0°) is indicative of the inherent conformational preference of Gly both in peptides and in proteins. As will be seen later these points form a significant component of turns.

The almost total absence of points in the α_R region for peptides and the reasonable percentage (14) in proteins show that a few of the glycyl residues in the latter, are α-helical accommodative. This aspect is considered in more detail in a later section.

Though β-turn is the better known feature for producing chain reversals in proteins, γ-turn can also bring about the chain reversal in a sharper way. For the γ-turn to occur, the conformation at the hinge must be in the neighbourhood of (−80°, 80°) or (80°, −80°). The present data set contains as many as 115 such turns*, with 110 of them occurring in the left top quadrant of the map which is allowed for any residue, glycyl or non-glycyl. Quite interestingly, only four out of these 115 are glycyl residues. Thus, in proteins, the glycyl is a non-former of γ-turns. Of these four examples, two are in the inverse γ-turn region (around (−80°, 80°)), and the remaining two occur in the γ-region. The same inference can be drawn from the peptide distribution also, in which only one example can be taken to be in the γ-turn conformation.

In the (ϕ,ψ) plane, the poly(proline)II (refs. 23, 24) (or single helix of collagen) conformation lies around (−70°, 145°) and poly(glycine) II (ref. 25) is known to take up this conformation. A search in the present data for conformations around this region yields 1174 residues occurring in this region (the limits, −90° to −50° is used for ϕ and 125° to 165° for ψ). However, the number of glycyl residues is only 33 and this will mean that the tendency of the poly(glycine) II to take up this conformation must be due to the large number of contiguous glycyl residues.

Glycyl residues in sub-structures of proteins

One of the aspects that can be looked into, is the occurrence of Gly in regular secondary structural regions. The three well-known secondary structural regions are (i) helix (primarily α-helix), (ii) extended strand (which in many cases form β-sheets), and (iii) different types of turns. Identification of the secondary structural regions can be done using some specified criteria and different workers have followed differing criteria for this purpose[26-36].

In the present case we have used the knowledge of (ϕ,ψ) values at the various residues to determine the secondary structural characteristics of Gly. The criterion used in the present case is based upon certain regularities in the (ϕ,ψ) values occurring at consecutive residues. The details are as follows:

The helical, extended and β-turn regions in the proteins were recognized based on suitably defined (ϕ,ψ) limits.

(i) A segment i to j is identified as a helical segment if the (ϕ,ψ) values at all the residues i to j are within the range −120° to −10° for ϕ and −120° to 20° for ψ. A minimum of four consecutive residues should satisfy these limits for the segment to be identified as helical.

(ii) Identification of an extended strand is similar to that of the helical segment except that the limits used are −180° to −60° for ϕ and 60° to 180° for ψ. Here also four consecutive residues should have their (ϕ,ψ) values within these limits.

(iii) The different types of β-turn are identified using the criterion that four torsion angles (two pairs of consecutive (ϕ,ψ) values) under examination should not vary by more than 30° from the ideal values proposed by Venkatachalam[37]. A slight relaxation is permitted, that is, only one of the four angles can vary up to 45°.

(iv) If a residue does not fall under in any of (i) to (iii) above, it is designated as uncharacterized.

Using the above criteria any residue can be classified into one of the four structural classes and file* containing the information about secondary structural assignments at each of the amino-acid residues can be prepared for each one of the 65 proteins in the data set. A flag H or B or T or U is assigned to each residue depending on whether the residue is situated in a helix (H) or extended strand (B) or β-turn (T) or uncharacterized (U) segment respectively. This sub-data base obtained forms a knowledge bank of sequence and structural information and is extremely useful for quick retrieval of information. Analysis on glycyl residues in secondary structures is performed using these (ϕ,ψ)-based secondary structure assignments.

The glycyl residues occurring in 65 proteins are sorted out based on their secondary structural assignments. The distribution in terms of H or B or T or U is shown in Figure 3. It is clear from the figure that a majority (52.2%) of Gly are in uncharacterized structural segments. Occurrence of Gly in helices and extended strands are 13.2% and 8.8% respectively. Its occurrence in 'turns' is also substantial (22.9%).

Glycyl residues in helices

Several workers have derived the propensities of each of the amino-acid residues to occur in different secondary structures. According to Levitt[38] the propensity of Gly to occur in helix is 0.56. This will mean that Gly cannot be considered as helix promoter. However, an idea of

*For this purpose a margin of ±20° was allowed for each of ϕ and ψ and conformations lying within this box were taken to represent γ-turn.

*This file will be referred to as map file further in the paper.

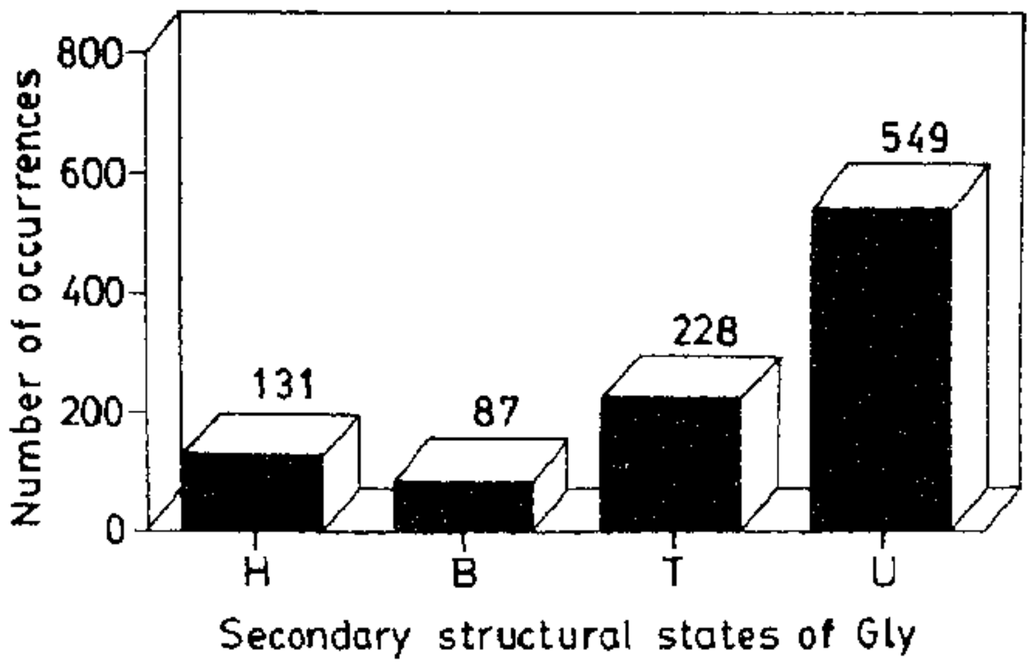


Figure 3. Bar diagram showing the distribution of glycyl residues in a set of 65 proteins, in the four secondary structural states, H (helix), B (extended strand), T (turn) and U (uncharacterized segments).

the occurrence of Gly in α-helical conformation can be obtained from Figure 2,*b*. It can be seen that a reasonable number of conformations do lie in the α-helical region. While it is true that Gly occurring in a helix must have its conformation in the α-helical region of the Ramachandran map, the converse need not be true, since the definition of the helical segment is not dependent on a single (ϕ,ψ) value but upon a collection of consecutive (ϕ,ψ) values. The analysis shows that, (a) of the 994 glycyl residues in the complete data set, 131 (13.2%) occur in helical segments; (b) there are 322 helical segments in all in the data set, of which 102 (31.7%) do have one or more Gly in them; (c) considering the length of helices, quite interestingly, 78 of these 102 helices, are 10 or more residues long. The number of Gly in these 78 long helices is 105 out of the total of 131 (81%). This shows that Gly is not 'averse' to occurring in helix and has a better preference to occur in longer segments than shorter ones.

Another feature that can be analysed is the relative position of Gly in the helix. According to our definition, the minimum number of residues required for a segment to be classified as helical is four. The location of Gly in a helical segment is designated as N, C or M depending upon whether it occurs at N-end or C-end or in the middle of the helix. The criterion for designation of N or C in a helical segment depends on the number of residues in the helix and it is as follows:

No. of residues in the helical segment	No. of residues from the beginning, designated as N	No. of residues from the end, (backwards) designated as C
>7	3	3
6 or 7	2	2
4 or 5	1	1

The distribution of 131 glycyl residues (Figure 4) among N, M and C categories is 36, 71 and 24

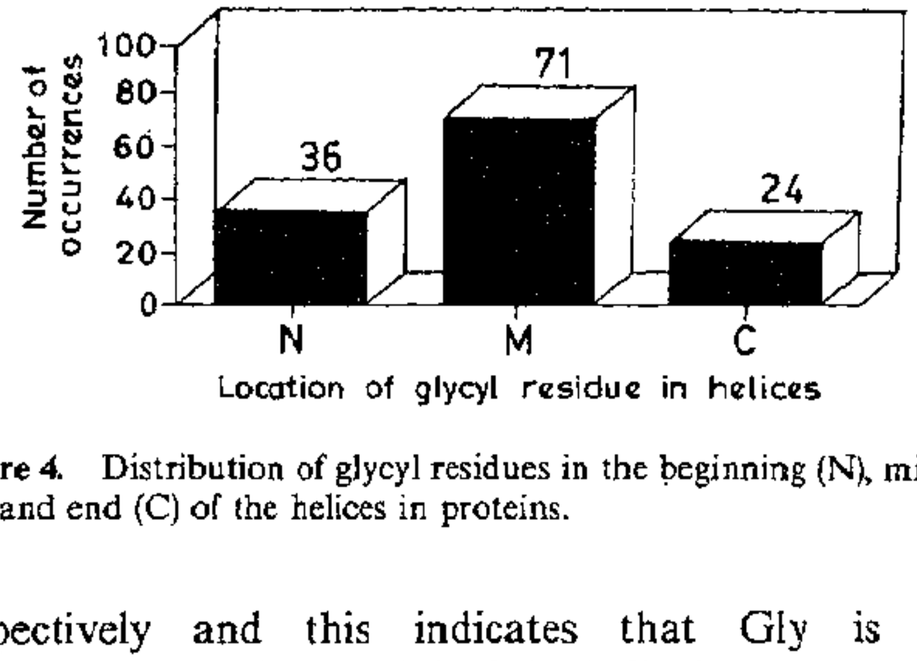


Figure 4. Distribution of glycyl residues in the beginning (N), middle (M) and end (C) of the helices in proteins.

respectively and this indicates that Gly is not concentrated at the termini of the helical segments. The larger number of residues occurring in the middle of the helix is more indicative of the fact that Gly favours continuance of the helix that has already been formed. In fact, a similar distribution is exhibited by Gly occurring in long helices (N : M : C = 27 : 61 : 16).

Examining the percentage of Gly in the helical segments in the individual proteins, the values vary widely (0–80%). The low values in some cases can be attributed to the poor helix content in those proteins. However, an interesting observation emerges in the case of proteins that are helix-rich. The percentage of Gly occurring in the helical segments in these proteins is given in Table 1 in decreasing order. In the present set, there are only eight proteins for which the percentage of Gly occurring in helix is $\geq$ 30. It can be seen that for all these proteins the helix content is $\geq$ 70%. In fact in the present set, these are the only proteins which can be considered as helix-rich ($\geq$ 70%). Of these eight proteins, six belong to globin family. From these data one can reasonably be sure that for helix-rich proteins, the expected percentage of Gly occurring in the helical region can be higher than the average value normally obtained. No other definite conclusion can be arrived at from the individual percentage values and hence they are not given here.

In order to find out how many Gly have just missed the helical segments, the number of Gly occurring at the position just before the beginning as well as that immediately following the helix is computed and the values are 50 (5%) and 90 (9.1%) respectively. Occurrence of Gly at the carboxyl end of the helices in

Table 1. Abundance of glycine in helical segments in some helix-rich proteins.

Protein code	% of Gly residues in helices	Helix content (%)
1ECA	82	84
4HHB	75	79
2MHR	67	70
1MBD	55	82
2LHB	50	76
2WRP	40	80
1MLT	33	85
1HMQ	33	70

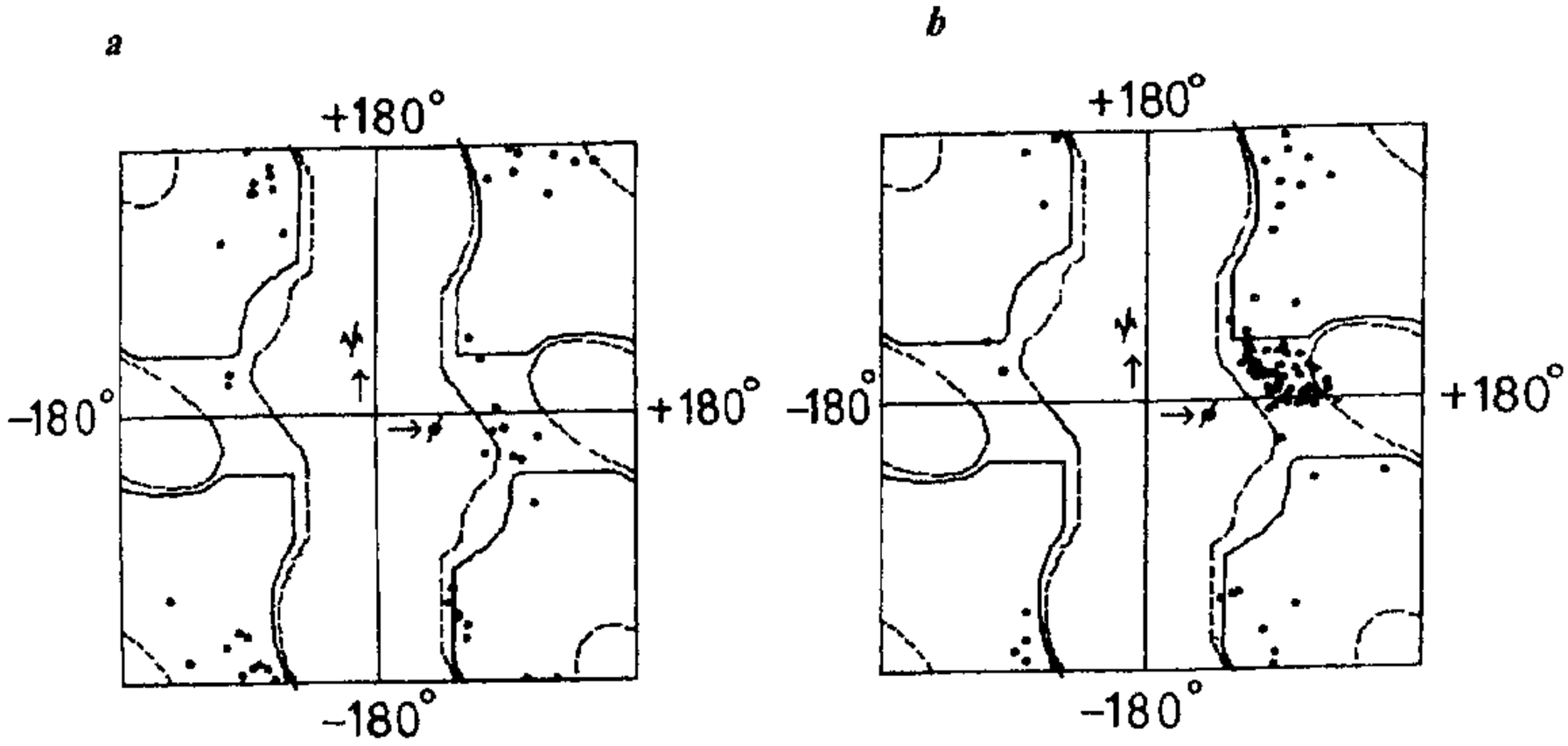


Figure 5. The (ϕ, ψ) plot of glycyl residues situated (*a*) immediately before the beginning of the helices, and (*b*) immediately after the end of the helices.

proteins has also been noticed by Schellman[39] and Richardson and Richardson[40].

The (ϕ, ψ) plot of these examples is shown in Figure 5, *a* and *b*. It is logical that these plots must be devoid of conformation in the vicinity of the α-helical region as otherwise they would have been part of the helix. However, there is a total absence of conformations in most of the quadrant of the map containing the α-helical conformation. This can be interpreted to mean that if the conformation of Gly is in the vicinity of that of a helix and its location in the sequence is immediate to a helix, it will most probably become a part of the helix itself.

There is a very good resemblance between Figure 5, *a* and Figure 2, *a*. The points are concentrated at the selective regions of the map with a notable absence around $\psi = \pm 90°$. Figure 5, *b* shows a high concentration of points with positive ϕ, and $\psi = 0°$. Comparing this figure with Figure 2, *b*, which gives the general glycyl distribution, the percentage of Gly occurring in this clustered region that are situated immediately after the end of the helices, works out to 18.9%.

The examples at the helix termini are further analysed for their occurrence in the other three categories namely, B, T and U. The results are shown in Figure 6, *a* and *b*, the former giving the details of the 50 examples occurring just before the beginning of the helix and the latter of the 90 examples occurring immediately after the end of the helix. It may be noticed from (ϕ, ψ) plot of the 90 examples (Figure 5, *b*) that there are some points in the extended conformation region of the Ramachandran map. Comparing with the Figure 6, *b*, it can be realized that none of these conformational points correspond to an extended strand. Occurrence of 9% of Gly immediately after the helices shows that Gly does also act as a helix breaker in as much as it does not prevent propagation of the helix. A more indepth investigation is warranted to find out the pre-requisites of Gly to take up either of the two roles.

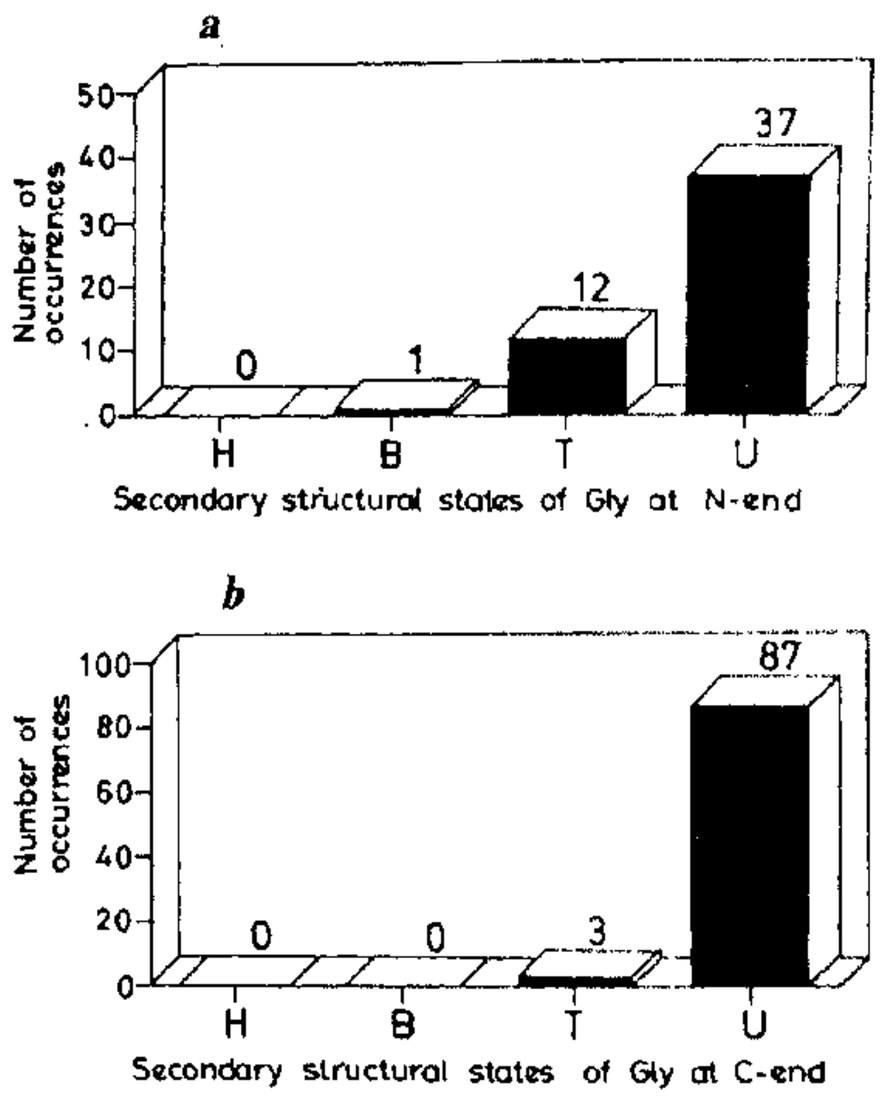


Figure 6. Secondary structure distribution of glycyl residues situated (*a*) immediately before the beginning of the helices, and (*b*) immediately after the end of the helices.

Glycyl residues in extended strands in proteins

Based on the overwhelming abundance of Gly in β-proteins such as silk which takes up a pleated β-sheet structure[6] and poly(glycine) I (ref. 4) which also takes up extended strand arrangement, it is natural to expect

reasonable preference for Gly to occur in the extended strands of globular proteins also. In addition, the conformation (of a single helix) of collagen, in which glycyl occupies every third position is also in the vicinity of the extended region.

However, the results, from the present study as far as globular proteins are concerned, turn out to be contrary to expectations. In the present data set, out of 994 Gly only 87 (8.8%) are situated in extended strands. The total number of residues occurring in the extended strands (category 'B' in map files), is 2825 and Gly constitute only 3.1%. Of the total of 468 extended strands available in the present data set, only 75 (16%) contain one or more Gly in them. In fact, 63 of these strands (84%) have only one glycyl residue in them and the rest (12) have just two glycyl residues per strand. Another observation, *vis-a-vis* the length of the strands, is that two thirds of the glycyl residues occur in short strands containing less than eight residues.

The propensity for Gly to occur in β-structures (P_β) as computed from the present data set is 0.35 (the P_β value for glycine derived by Levitt[38] is 0.92) which again points to the fact that Gly are poor candidates for inclusion in the extended strands. In order to rule out any element of artefact, the P_β values, for some of the other β-sheet-promoting residues worked out from the present data set are checked with those derived by Levitt[38]. The values for the amino acids Val, Ile, Phe, Tyr and Thr are 1.69, 1.46, 1.17, 1.33 and 1.27 respectively and the corresponding values derived by Levitt are 1.49, 1.45, 1.32, 1.25 and 1.21. This good agreement can be taken to mean that the present conclusion regarding the disfavour of Gly in extended strands is a realistic one.

The two parametric criteria used to identify the extended strands involve a specific region of the (ϕ, ψ) plane and a minimum number of four consecutive residues to occur in this region. In order to check whether the conclusions are critically dependent on these, relaxations of these were attempted. First, the minimum requirement of consecutive residues was reduced from 4 to 2. Though the number of extended strands was almost doubled (468 to 920), there were only 25 glycyl residues which occur in the additional extended strands. Secondly, the ψ range was extended to include the region $-180°$ to $-150°$ and this relaxation resulted in 40 more Gly in the extended strands. This increased the P_β value from 0.35 to 0.46. Even this value is considerably low compared to the value obtained by Levitt. *Thus these results clearly suggest that glycyl residues can only be considered as poor formers/promoters of extended structures in globular proteins.*

Glycyl residues in β-turns

It is well known that glycyl is one of the potential turn-forming residues in proteins (see for eg. ref. 41) the other conspicuous residue being proline. Interestingly, these two residues having opposite conformational characteristics, namely flexibility for the former and rigidity for the latter seem to play a big role in chain reversals of proteins. In the present data set, 228 out of 994 (22.9%) glycyl residues occur in β-turns. The propensity of glycyl residue to occur in β-turns works out to 2.49, indicating reasonably high preference of glycyl residue to occur in turns. In fact, of the four conformational states defined (H, B, T and U), glycine has maximum propensity for turns.

The conformational points of the standard turn types proposed by Venkatachalam[37] can occur only in certain regions of the (ϕ, ψ) plane. These are shown as four boxes (α_R, α_L, e and e*) in Figure 7 where the conformations of Gly occurring in turns are also plotted. The boxes are drawn in accordance with the criteria used for identifying turns. The types and the positions of the turns corresponding to each box are also shown in the figure. The number of conformations occurring in a box is also marked within the box.

From the figure it can be seen that a majority of conformations (69.3%) occur in α_L box, around (90°, 0°). This, as well as the α_R box occupy the bridge regions of the map. About 80% of the points occur in the right half of the map, which are generally disallowed for non-glycyl residues in L-configuration. This suggests that there is a strong preference for Gly to make use of its inherent conformational freedom and produce the

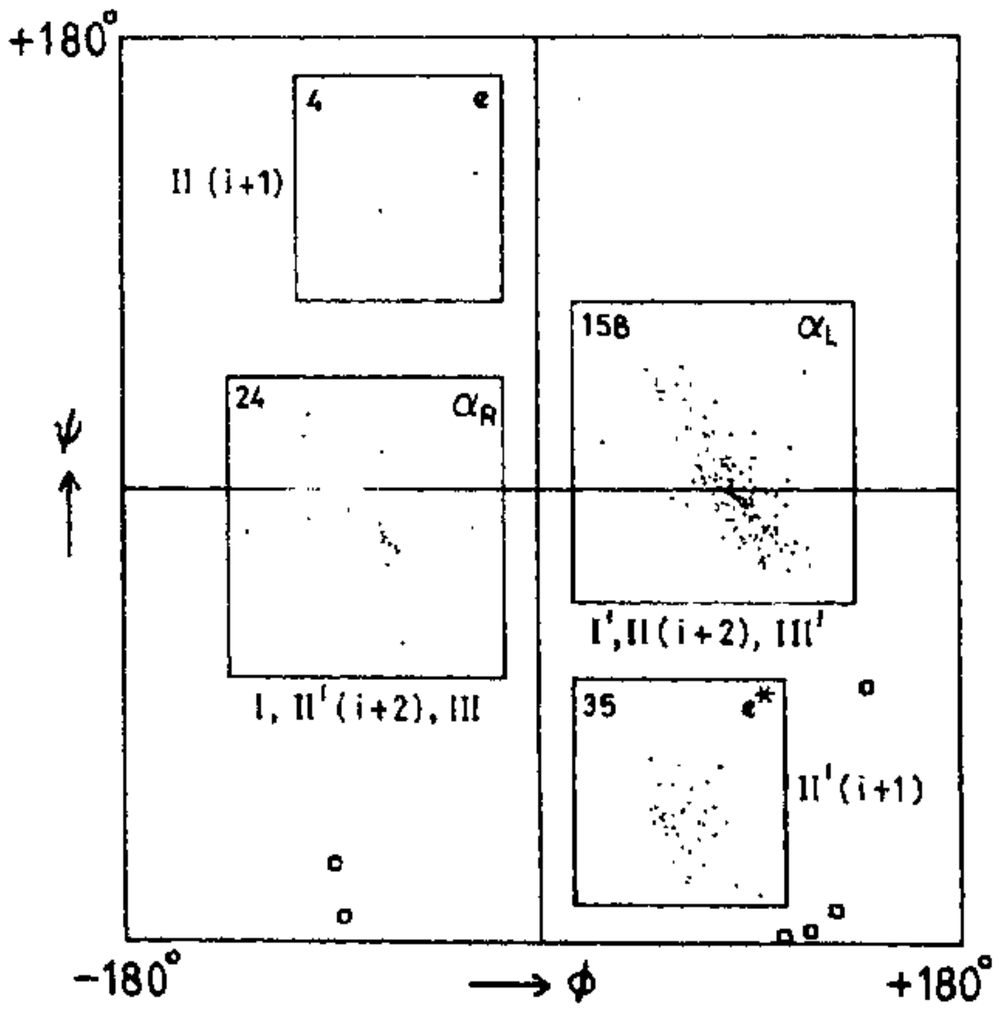


Figure 7. The (ϕ, ψ) plot of glycyl residues occurring in β-turns in proteins. The (ϕ, ψ) regions where turn conformations can occur are shown as boxes (α_R, e, α_L and e*). The number of points inside each box is shown in the left top corner of the box. Turns and the positions possible for a given box are also indicated. (□ – turn conformations involving a *cis*-peptide unit.)

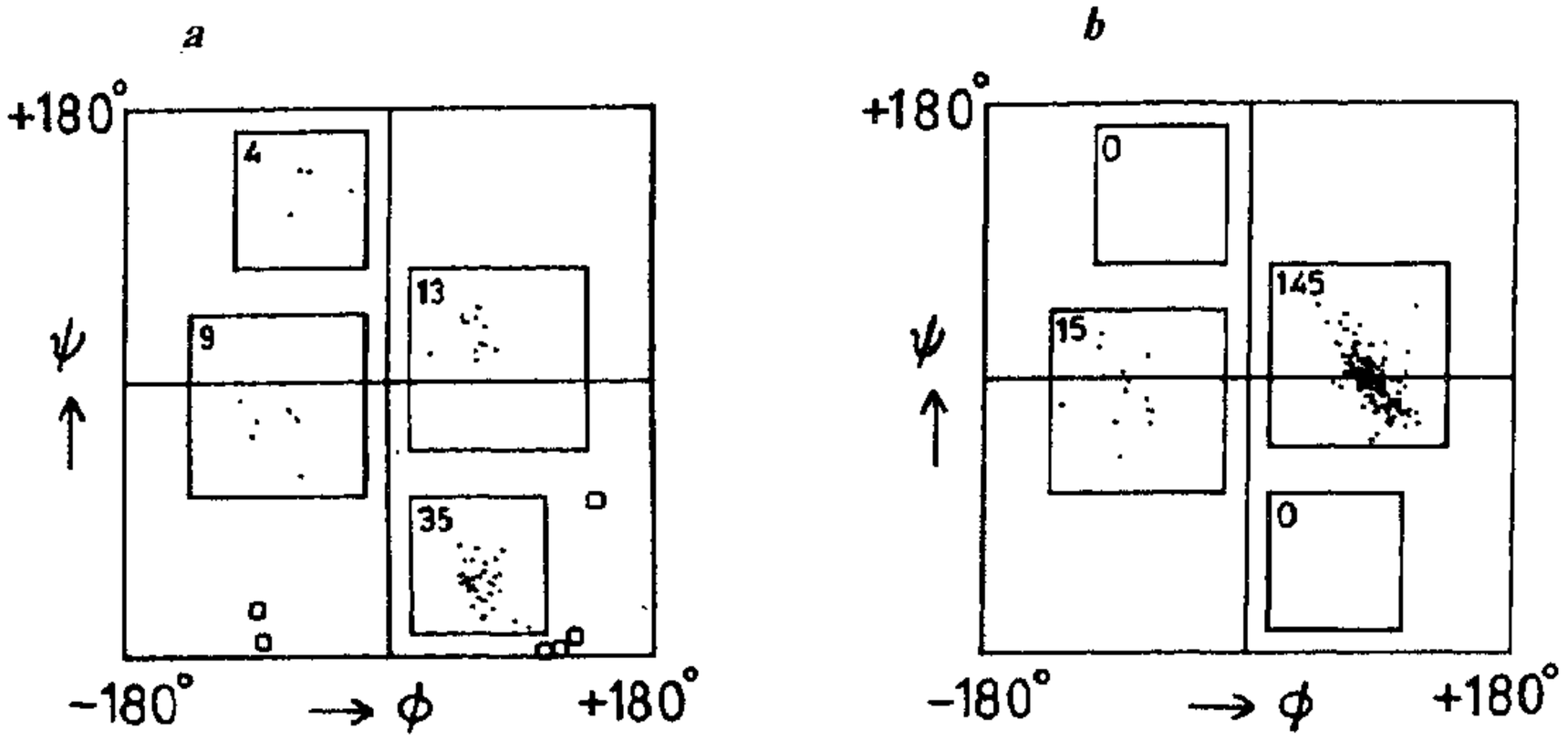


Figure 8. The (ϕ, ψ) distribution of glycyl residues occurring at (*a*) position $i+1$ of the turn; (*b*) position $i+2$ of the turn.

necessary β-turns of types, less favourable with other residues.

Designating any turn to be from residue i to $i+3$, the glycyl residues responsible for the turn conformation must occupy positions $i+1$ or $i+2$. The number of examples in position $i+1$ is 66 (28.9%) and that in position $i+2$ is 162 (71.1%). The (ϕ, ψ) plot of Gly occurring in turns is shown in Figure 7, those occurring in position $i+1$ and $i+2$ of the β-turn are shown in Figure 8,*a* and *b* respectively. Figure 8,*a* shows that about 50% of Gly occurring in position $i+1$ are of type II and the remaining are distributed in other types. Figure 8,*b* shows the clustering of points in the two bridge regions and comparing this with Figure 7, it can be seen that these are the only two regions that can be taken up by conformations in positions $i+2$ of the β-turn. What is more interesting, is the observation that 90.6% of the points occur in the right half of the map indicating a strong tendency for glycyl residues in position $i+2$ to have conformations in this region.

As already pointed out, Gly is abundant in the uncharacterized structural elements which often connect secondary structures. An analysis carried out in this laboratory on $\alpha\alpha$, $\beta\beta$, $\alpha\beta$ and $\beta\alpha$ connectors in proteins, clearly shows that the glycyl is in fact the most abundant of the residues in short loops (to be published).

Secondary structure preferences of glycine containing doublets

Sequences of the segments having Gly sandwiched between any two residues X and Y, were analysed for secondary structural preferences. This is done by extracting secondary structural states of different doublets of the types X-Gly and Gly-Y. The number of examples of each doublet in different structural states at the constituent residues is conveniently represented as a (4×4) matrix*. The consolidated matrices for X-Gly and Gly-Y doublets are shown in Figures 9,*a* and *b* respectively. In Figure 9,*a* the information in each row corresponds to the distribution of Gly (in different states) for a given state of X and in Figure 9,*b* the information in each row corresponds to the distribution of Y (in different states) for a given state of Gly. It is possible to compute propensity values of the various doublets for occurrences in specific secondary structures following the method used by Chou and Fasman[42].

The consolidated doublet distributions for X-Gly and Gly-Y show that the values at the diagonal elements are in general high. This feature of the X-Gly distribution can be interpreted to mean that, *the glycyl residue prefers to have the same conformational state, as that of the previous residue.* The same aspect of the Gly-Y distribution indicates that *the conformational state which glycyl adopts is propagated to the next residue Y in the sequence.* Some of the values in this distribution (H–B, H–T, B–H, U–H, U–B, T–H and T–B in X-Gly and H–B, H–U, H–T, B–H and B–T in Gly-Y) are very small (<10) and hence this combinatorial states can generally be treated as unfavourable for the doublets involving glycine.

The doublet propensity values for the three homogeneous states H–H, B–B and T–T of X-Gly and Gly-Y were compared for various residues X and Y. Those with extreme values of the propensities are picked out and given in Table 2. This can be an useful input for those algorithms dealing with prediction and folding aspects of proteins. A few other observations are:

(i) By examining the individual matrices it is found that for many doublets, of both X-Gly and Gly-Y types,

*Corresponding to 20 different amino acids, one can have 20 matrices for X-Gly and 20 more for Gly-Y doublets. These are not given here, but can be had from the authors on request.

X – Gly

X \ Gly	H	B	U	T
H	120	0	87	3
B	2	71	65	29
U	4	7	323	44
T	2	6	67	167

(*a*)

Gly – Y

Gly \ Y	H	B	U	T
H	121	0	8	0
B	1	62	16	6
U	37	86	382	36
T	12	55	100	73

(*b*)

Figure 9. *a*, Number of occurrences of all X-Gly doublets in the various combinations of four secondary structural states at X and Gly. *b*, Same as *a* for Gly-Y doublets.

Table 2. Potential glycine containing doublets favouring or disfavouring helix, extended strand and turn.

Doublet	No. of examples	Sec. str. at the residues	Favour or disfavour (F or D)
His–Gly	26	HH	F
Leu–Gly	60	HH	F
Cys–Gly	45	HH	D
Pro–Gly	58	HH	D
Gly–Glu	43	HH	F
Gly–Met	20	HH	F
Gly–Asp	58	HH	D
Gly–Gly	100	HH	D
Val–Gly	64	BB	F
Phe–Gly	27	BB	F
His–Gly	26	BB	D
Asn–Gly	36	BB	D
Gly–Pro	34	BB	F
Gly–Val	70	BB	F
Gly–Gln	28	BB	D
Gly–Asn	28	BB	D
Pro–Gly	58	TT	F
Lys–Gly	51	TT	F
Ile–Gly	42	TT	D
Trp–Gly	20	TT	D
Gly–Pro	34	TT	F
Gly–Cys	34	TT	F
Gly–Ile	58	TT	D
Gly–Gln	28	TT	D

the number of examples of U–U state is the largest. This is in tune with the total distribution of X-Gly and Gly-Y doublets (Figures 9,*a* and *b*). This means that when Gly is sandwiched between two residues occurring in state U, it tends to take up the same state, irrespective of X and Y.

(ii) Examination of individual matrices allows one to pick out some combinations involving the three states H, B or T, for which the number of examples is relatively large. These are Ala–Gly (H–H and T–T), Gly–Gly (T–T), Leu–Gly (H–H), Lys–Gly (T–T), Pro–Gly (T–T), Gly–Glu (H–H), and Gly–Leu (H–H). It is interesting to note that in the four examples of the states H–H given above, the non-glycyl residues are Ala, Leu and Glu which by themselves are good helix promoters.

(iii) The best preferred state of Pro–Gly is T–T while that of Gly–Pro is U–U. This is in agreement with the idea that Pro–Gly is a better turn-former than Gly–Pro.

Accessibility of glycyl residues

In the previous section, the location and preference of Gly were looked into from a secondary structural viewpoint. Another aspect that can be analysed is the exposure of Gly to the surroundings. The widely used method for such purpose is that of Lee and Richards[43] which gives the accessible surface area of each of the atoms in a protein. However, in the present study we have used a simplified procedure based upon geometri-

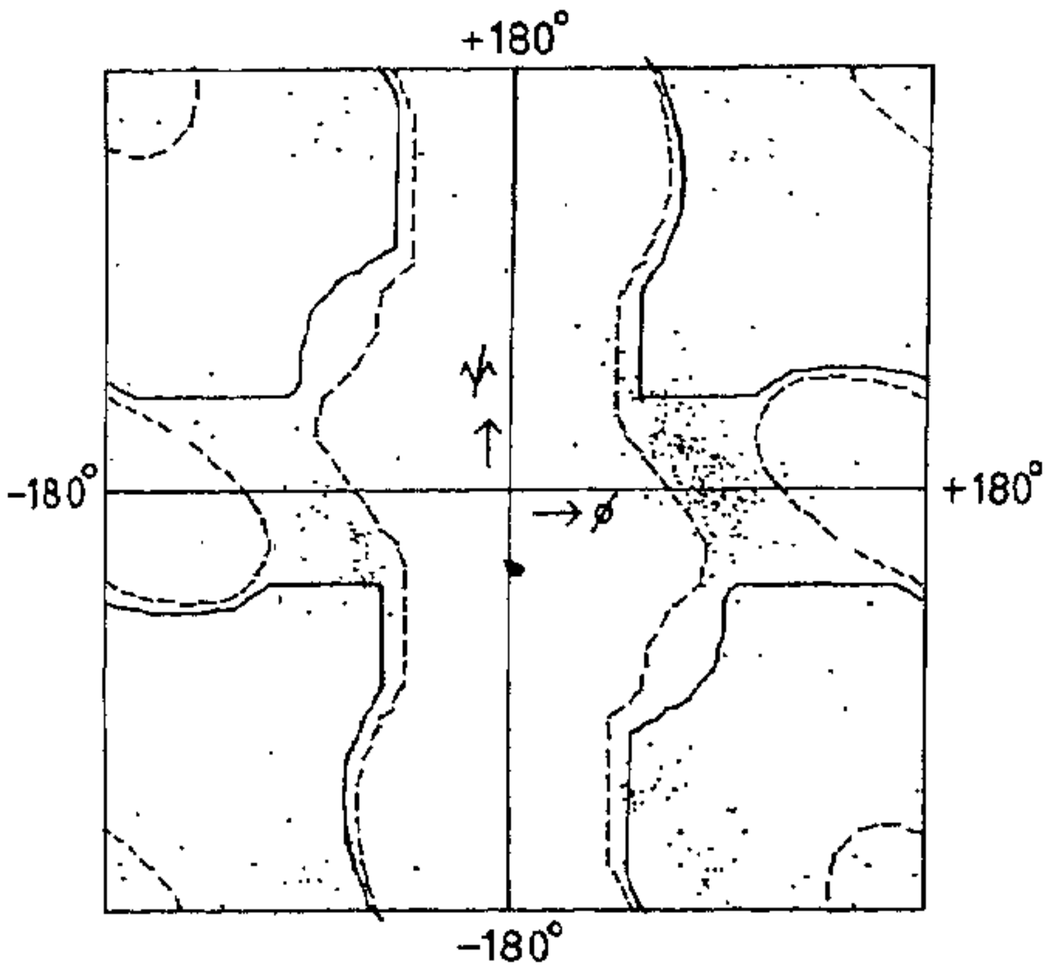


Figure 10. The (ϕ, ψ) plot of glycyl residues situated at the surface of proteins.

cal considerations to identify those residues in a protein that are on the surface and hence exposed to the solvent (C. Ramakrishnan and N. Srinivasan, unpublished). This method, which is computationally very fast and makes use of positions of α-carbons alone, has been tested with the results of Lee and Richards program and the agreement factor is good.

Of the 994 glycyl residues in the present data set, 333 are found to be located on the surface (33.5%). This shows that a considerable proportion of glycyl residues is buried within the protein.

The (ϕ, ψ) distribution of the 333 glycyl residues occurring on the surface is shown in Figure 10. The conspicuous feature of this figure is the preference of these residues to the positive ϕ region of the map (75.7%). Here again the concentration is in the bridge region of the map occurring around (90°, 0°).

Distribution of these glycyl residues among the four secondary structural states H, B, U and T is 24, 5, 181 and 123 respectively. Thus about 91% of the residues are concentrated in U and T, leaving a meagre 9% for the regular secondary structures. This is in agreement with the well-understood feature that loops and turns are predominantly found in surfaces[27, 44] and Gly occurring in these are naturally to be found on the surface.

As shown in Figure 3, the distribution of Gly into the four states, H, B, U and T is 131, 87, 549 and 228 respectively. Comparing this distribution with that of surface glycyl residues, one can get an idea of the proportion of 'exposed' and 'buried' residues in the four states. The ratio of surface to buried for the four secondary structural states is H, 18:82; B, 6:94; U, 33:67 and T, 54:46. This again shows that those glycyl residues that occur in helices and strands are more buried than exposed, whereas in 'turns' it has more or less equal tendency to occur either on surface or inside.

Conclusions

The present analysis on glycyl residues in proteins, as examined both from conformational angles ϕ and ψ as well as its locations in the different organized secondary structural regions, brings out a number of observable features. Concentration of conformational points in the bridge region, taking part in turns and located on the surface of the proteins can all be used in a very effective way when one develops prediction/folding algorithms for proteins. The results of the analysis can also be used in choosing locations for site-directed mutagenesis studies involving replacement of a non-glycyl by a glycyl residue. The present study reveals Gly to be a passive member in helices, a poor promoter of extended strands (as contrasted to the situation in fibrous proteins) and an active member in turns. Occurrence of conformationally flexible glycyl residues in large number in uncharacterized regions is fully justified in view of the topological requirement of such regions, to act as connectors between regular secondary structures. Thus Gly plays a crucial role in the ultimate folded state of proteins. More detailed and protein-specific analysis of the role of glycyl is the next logical step that can be taken.

1. Ramachandran, G. N., Ramakrishnan, C. and Sasisekharan, V., *J. Mol. Biol.*, 1963, **7**, 95.
2. Ramachandran, G. N. and Sasisekharan, V., *Adv. Protein Chem.*, 1968, **23**, 283.
3. Fraser, R. D. B. and Mc Rae, T. P., *Conformation of Fibrous Proteins and Related Synthetic Polypeptides*, Academic Press, New York, 1973.
4. Bradbury, E. M. and Elliot, A., *Polymer*, 1963, **4**, 47.
5. Ramachandran, G. N., Sasisekharan, V. and Ramakrishnan, C., *Biochim. Biophys. Acta*, 1966, **112**, 168.
6. Marsh, R. E., Corey, R. B. and Pauling, L., *Biochim. Biophys. Acta*, 1955, **16**, 1.
7. Ramachandran, G. N. and Reddi, A. H., *Biochemistry of Collagen*, Plenum Press, New York, 1976.
8. Pierschbacher, M. D. and Ruoslahti, E., *Nature*, 1984, **309**, 30.
9. Lipscomb, W. N., *Proc. Robert A. Welch Found. Conf. Chem. Res.*, 1971, **15**, 140.
10. Gardell, S. J., Craik, C. S., Hilvert, D., Urdea, M. S. and Rutter, W. J., *Nature*, 1985, **317**, 551.
11. Hilvert, D., Gardell, S. J., Rutter, W. J. and Kaiser, E. T., *J. Am. Chem. Soc.*, 1986, **108**, 5298.
12. Tong, L., de Vas, A. M., Milburn, M. V., Jancarik, J., Naguchi, S., Nishimura, S., Miura, K., Ohtsuka, E. and Kim, S.-H., *Nature*, 1989, **337**, 90.
13. Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Jr., Brice, M. D., Rodgers, J. R., Kannard, O., Shimanouchi, T. and Tasumi, M., *J. Mol. Biol.*, 1977, **112**, 535.
14. Ramakrishnan, C. and Ramachandran, G. N., *Biophys. J.*, 1965, **5**, 909.

15. Kolaskar, A. S., Sarathy, K. P. and Sasisekharan, V., *Curr. Sci.*, 1975, **44**, 35.
16. Manjula, G., Ramakrishnan, C. and Sarathy, K. P., *Proc. Indian Acad. Sci.*, 1977, **A86**, 443.
17. Richardson, J. S., *Adv. Protein Chem.*, 1981, **34**, 167.
18. Ravichandran, V. and Subramanian, E., *Int. J. Pept. Protein Res.*, 1981, **18**, 121.
19. Lambert, M. H. and Scheraga, H. A., *J. Comp. Chem.*, 1989, **10**, 817.
20. Nicholson, H., Soderlind, E., Tronrud, D. E. and Matthews, B. W., *J. Mol. Biol.*, 1989, **210**, 181.
21. Richardson, J. S. and Richardson D. C., *Prediction of Protein Structure and Principles of Protein Conformation* (ed. Fasman, G. D.), Plenum, New York, 1989.
22. Ramakrishnan, C., Srinivasan, N. and Prashanth, D., *Int. J. Pept. Protein Res.*, 1987, **29**, 629.
23. Cowan, P. M. and Mc Gavin, S., *Nature*, 1955, **176**, 501.
24. Sasisekharan, V., *Acta Crystallogr.*, 1959, **12**, 897.
25. Crick, F. H. C. and Rich, A., *Nature*, 1955, **176**, 780.
26. Lewis, P. N., Momany, F. A. and Scheraga, H. A., *Proc. Natl. Acad. Sci. (USA)*, 1971, **68**, 2293.
27. Kuntz, I. D., *J. Am. Chem. Soc.*, 1972, **94**, 4009.
28. Crawford, J. L., Lipscomb, W. N. and Schellman, C. G., *Proc. Natl. Acad. Sci. (USA)*, 1973, **70**, 538.
29. Levitt, M. and Greer, G., *J. Mol. Biol.*, 1977, **114**, 181.
30. Rose, G. D. and Seltzer, J. P., *J. Mol. Biol.*, 1977, **113**, 153.
31. Chou, P. Y. and Fasman, G. D., *J. Mol. Biol.*, 1977, **115**, 135.
32. Kolaskar, A. S., Ramabrahmam, V. and Soman, K. V., *Int. J. Pept. Protein Res.*, 1980, **16**, 1.
33. Ramakrishnan, C. and Soman, K. V., *Int. J. Pept. Protein Res.*, 1982, **20**, 218.
34. Kabsch, W. and Sander, C., *Biopolymers*, 1983, **22**, 2577.
35. Hohne, E. and Kretschmer, R. G., *Stud. Biophys.*, 1985, **108**, 165.
36. Richards, F. M. and Kundrot, C. E., *Proteins: Str. Fn. Gen.*, 1988, **3**, 71.
37. Venkatachalam, C. M., *Biopolymers*, 1968, **6**, 1425.
38. Levitt, M., *Biochemistry*, 1978, **17**, 4277.
39. Schellman, C., *Protein Folding* (ed. Janicke, R.), Elsevier/North Holland, Amsterdam, New York, 1980.
40. Richardson, J. S. and Richardson, D. C., *Science*, 1988, **240**, 1648.
41. Wilmot, C. M. and Thornton, J. M., *J. Mol. Biol.*, 1988, **203**, 221.
42. Chou, P. Y. and Fasman, G. D., *Biochemistry*, 1974, **13**, 211.
43. Lee, B. and Richards, F. M., *J. Mol. Biol.*, 1971, **55**, 379.
44. Rose, G. D., Young, W. B. and Gierasch, L. M., *Nature*, 1983, **304**, 654.

ACKNOWLEDGEMENTS. We thank Prof. P. Balaram for his comments and suggestions. Thanks are due to Mr H. A. Nagarajaram and Mr D. V. Nataraj for compiling the data on peptides and to Ms K. P. Viji for reproducing the glycyl steric map.