*J. Mol. Biol.* (1998) **275**, 917–932

JMB

# Stereochemical Punctuation Marks in Protein Structures: Glycine and Proline Containing Helix Stop Signals

**K. Gunasekaran, H. A. Nagarajaram, C. Ramakrishnan and P. Balaram***

*Molecular Biophysics Unit*
*Indian Institute of Science*
*Bangalore, India*

An analysis on the nature of α-helix stop signals has been carried out, using a dataset of 1057 helices identified from 250 high resolution (≤2.0 Å), non-homologous, protein crystal structures. The backbone dihedral angles (φ, ψ) of the terminating residue (T) were found to cluster either in the left-handed helical region ($\alpha_L$: $\phi = 20^\circ$ to $125^\circ$ and $\psi = -45^\circ$ to $90^\circ$; 469 helices (44%)) or in the extended region (E: $\phi = -180^\circ$ to $-30^\circ$ and $\psi = 60^\circ$ to $180^\circ$ and $-180^\circ$ to $-150^\circ$; 459 helices (43%)) of the Ramachandran map. These two broad categories of helix stop signals, $\alpha_L$ and E-terminated helices, were further examined for sequence preferences. Gly residues were found to have an overwhelming preference to occur as the "$\alpha_L$-terminator (T)" resulting in the classical Schellman motif, with a strong preference for hydrophobic residues at position T − 4 and T + 1. In the case of E-terminated helices His, Asn, Leu and Phe were found to occur with high propensity at position T. Quite remarkably Pro residues, with single exception, were absent at position T, but had the highest propensity at position T + 1. Examination of the frequencies of hydrophobic (h) and polar (p) residues at positions flanking Gly/Pro permitted delineation of exclusive patterns and predictive rules for Gly-terminated helices and Pro-terminated helices. The analysis reveals that Pro residues flanked by polar amino acids have a very strong tendency to terminate helices. Examination of a segment ranging from T − 4 to T + 3 appeared to be necessary to determine whether helix termination or continuation occur at Gly residues. The two types of helix termination ($\alpha_L$, E) signals also differed dramatically in their solvent accessibility. Gly and Pro residues at helix termini appeared to be strongly conserved in homologous sequences.



*Keywords:* α-helix; helix design; stop signals; structure prediction; protein engineering

*Corresponding author

## Introduction

The hypothesis that amino acid sequences are translated into well defined three dimensional-conformations by means of a stereochemical code (Anfinsen, 1973), has stimulated many attempts to derive specific rules for secondary structure formation by short stretches of polypeptides (Chou & Fasman, 1974, 1978; Gibrat *et al.*, 1987; Rooman & Wodak, 1988; Kametkar *et al.*, 1993; Sippl, 1995; Aurora *et al.*, 1997; Rost & Sander, 1994). The α-helix, one of the two most widely occurring secondary structures in proteins (Pauling *et al.*, 1951; Richardson, 1981), has been the focus of several investigations, which have aimed at analysing sequence effects on stability (Lyu *et al.*, 1990, 1992; Horovitz *et al.*, 1992; Serrano *et al.*, 1992; Chakrabartty & Baldwin, 1995; Scholtz & Baldwin, 1992; Bruch *et al.*, 1991; Serrano & Fersht, 1989; Blaber *et al.*, 1993; Zhou *et al.*, 1994; O'Neil & DeGrado, 1990; Padmanabhan & Baldwin, 1994) and N and C terminus signals which act to define start and the end points of the helix (Richardson & Richardson, 1988; Presta & Rose, 1988; Harper & Rose, 1993; Aurora *et al.*, 1994; Preissner & Bork, 1991; Dasgupta & Bell, 1993; Jimenez *et al.*, 1994; Serrano & Fersht, 1989; Kim & Baldwin, 1984; Seale *et al.*, 1994; Doig *et al.*, 1997; Doig & Baldwin,

Present address: H. A. Nagarajaram, Department of Biochemistry, University of Cambridge, Tennis Court Road, Cambridge CB2 1WQ, UK.

0022–2836/98/050917–16 $25.00/0/mb971505

1995). It would be desirable to inspect an amino acid sequence, recognize segments with a high helical propensity (Chou & Fasman, 1974, 1978) and precisely identify the initiation and termination sites for the helix. Several illuminating analyses of helices in protein structures have provided valuable information on both helix start and stop signals (Richardson & Richardson, 1988; Presta & Rose, 1988; Harper & Rose, 1993; Aurora *et al.*, 1994; Preissner & Bork, 1991; Dasgupta & Bell, 1993; Doig *et al.*, 1997). An α-helix can effectively be terminated by a C terminus non-glycyl residue straying into either extended (E, β) region or left-handed helical ($\alpha_L$) regions of conformational space, since only three broad regions (E, $\alpha_L$, $\alpha_R$) are sterically allowed for the L-amino acids (Ramachandran *et al.*, 1963; Ramakrishnan & Ramachandran, 1965; Ramachandran & Sasisekharan, 1968). In order to obtain detailed insights into the nature of helix stop signals which may be useful for both peptide design (DeGrado, 1988; Betz *et al.*, 1993; Richardson *et al.*, 1992; Kametkar *et al.*, 1993; Baldwin, 1995) and protein structure prediction (Chou & Fasman, 1974, 1978; Gibrat *et al.*, 1987; Rooman & Wodak, 1988; Sippl, 1995), we have undertaken an analysis of 1057 helices identified from 250 high resolution protein structures. Two broad categories of helix stop signals, with the terminating residue lying in $\alpha_L$ or E conformations, have been identified. The former leads to the classical Schellman motif (Schellman, 1980; Milner-White, 1988) with Gly being overwhelmingly preferred at the terminating position (Aurora *et al.*, 1994; Viguera & Serrano, 1995). Specific sequence effects have been identified which determine whether the Gly residue will continue or terminate a helix. For the E-terminated structures there is a remarkable preference for Pro residues to follow the helix-terminating residue, a feature rationalized by the local interactions involving the pyrrolidine ring (Hurley *et al.*, 1992; MacArthur & Thornton, 1991). The local sequence patterns that determine whether a helix will terminate near a Pro residue or continue with incorporation of Pro into the body of the helix have been analysed. The results described here focus on the C-terminal end of helices and establish that an α-helix can effectively be terminated either by left-handed helical ($\alpha_L$) or by extended (E) conformations, with the two modes of termination exhibiting unique local structural features.

## Results

### Definitions

Helices were identified using the criterion that the backbone dihedral angles (ϕ, ψ) of the successive residues, in a segment of length $\geqslant 7$, should lie in the helical region ($\alpha_R$: $\phi = -140^\circ$ to $-30^\circ$ and $\psi = -90^\circ$ to $45^\circ$). This procedure resulted in the identification of 1057 helices. The length of the helical segment varied from 7 residues to 42 residues with mean of 14 residues. The C-terminal tail (C-tail) comprises of three residues which follow the terminating residues i.e. T + 1, T + 2, T + 3. N-cap refers to residues 1 to 3 at the N terminus of a helix while C-cap refers to residues $n$, $n-1$ and $n-2$ in a helical segment of length $n$. (It may be noted that the N-cap, C-cap definition used by Richardson & Richardson (1988) differ from that used here). Residues in the helix which form the segment 4 to $n-3$ are termed as ''middle'' residues. The terminating or the terminus residue T is the residue at the C terminus of the helix which shows non-helical (ϕ, ψ) values signifying the end of the helical segment.

### Amino acid positional preferences in helices

A total of 1057 helices of length $\geqslant 7$ residues were identified from a data set of 250 protein crystal structures. Figure 1(a) shows the Ramachandran plot of all the helical residues and Figure 1(b) shows the plot of the terminating residue (T). The propensity of the amino acids to occur in specific segments of helices was computed for the data set of 1057 helices. Propensities of all the 20 amino acids to occur at a given position ($P_{ij}$) were computed using the formula:

$$P_{ij} = \frac{F_{ij}}{\sum_{i=1}^{20} F_{ij}} \Big/ \frac{D_i}{\sum_{i=1}^{20} D_i}$$

where, $F_{ij}$ is the number of times residue $i$ occurs in the position $j$ and $D_i$ is the number of times residue $i$ occurs in the data set. A value of $P_{ij} > 1$ indicates preference, and a value less than 1 indicates disfavour. The results are summarized in Figure 2. An earlier analysis of helices in proteins specifically focused on the N-cap and C-cap segments leading to several important conclusions (Richardson & Richardson, 1988). In particular, the presence of Pro to a significant extent at the N terminus of helices pointed to its importance as a helix initiator. The presence of glycine at the C-terminal end of helices has also been noted by Richardson & Richardson (1988). The present analysis (Figure2) confirms both these observations on a larger data set. Interestingly, Asn, which has earlier been reported to have a dramatic N-cap preference, appears to have greater propensity to occur at the C-terminal end of helices (in the present analysis, Asn has propensity of 0.78 to occur at the N-cap positions and 1.83 for the C terminus). A striking feature of the data in Figure 2 is the remarkable preference of Gly to act as a helix terminator. It is also noteworthy that Pro, a residue which was traditionally considered to be a helix breaker, is never found at the terminating position T, but occurs with a very high propensity in the C-tail positions, a feature which is considered in greater detail later. The occurrence of Pro immediately after the C termini of α-helices has been noted before in statistical analyses (Chou & Fasman, 1978; Richardson & Richardson, 1988; MacArthur & Thornton, 1991; Dasgupta & Bell, 1993).

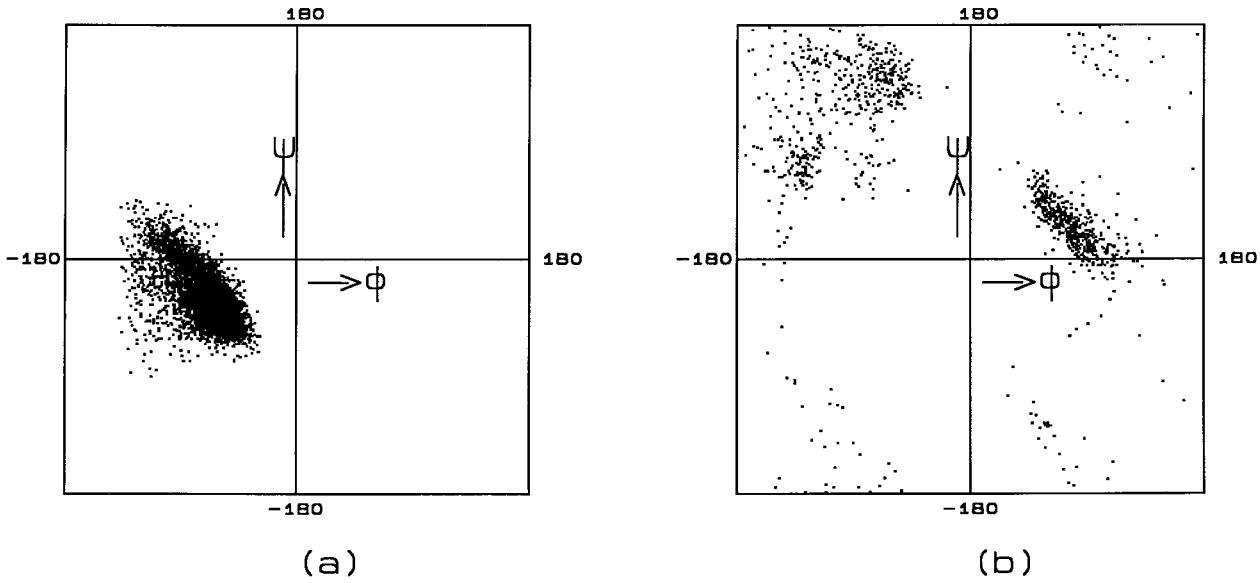


**Figure 1.** (a) Distribution of backbone dihedral angles $\phi$, $\psi$ for all the helical residues in the 1057 helices extracted from a data set of 250 protein crystal structures. (b) Plot showing the $\phi$, $\psi$ distribution of the terminator (T) residues that are found at the C-terminal end of the 1057 helices (see the text for definition).

## Helix-termination signals

Inspection of Figure 1(b) reveals that there are only two substantially populated regions of ($\phi$, $\psi$) space in which the conformation of the terminating residue T are clustered. These are the left-handed helical region ($\alpha_L$: $\phi = 20°$ to $125°$ and $\psi = -45°$ to $90°$), which requires positive $\phi$ values and the extended region (E: $\phi = -180°$ to $-30°$ and $\psi = 60°$ to $180°$ and $-180°$ to $-150°$) of the Ramachandran map (Ramachandran *et al.*, 1963). Of the 1057 helices, 44% ($N = 469$) are terminated by $\alpha_L$ conformations (Nagarajaram *et al.*, 1993), 43% ($N = 459$) are terminated by E conformations, while the remainder primarily correspond to conformations with positive $\phi$ values in the non-helical regions, largely populated by Gly residues. Thus we may consider two broad general categories of helices namely $\alpha_L$-terminated helices and E-terminated helices.

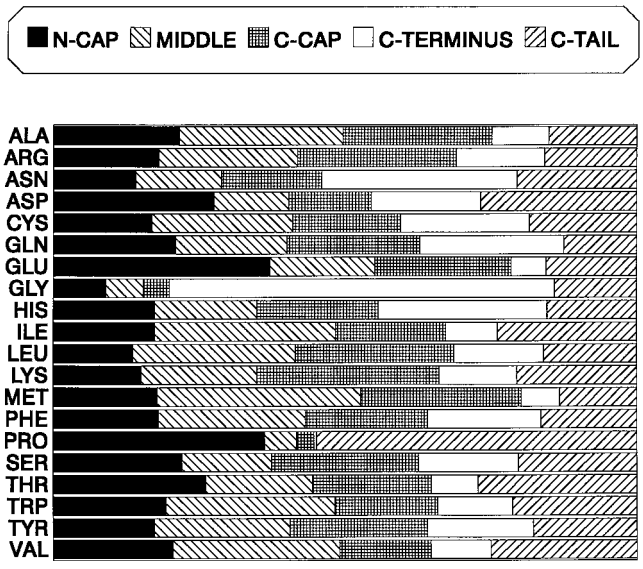


**Figure 2.** Bar diagram showing fractional propensities of the amino acids to occur at the N-cap, middle, C-cap, C terminus and C-tail positions (see the text for definition). For individual amino acids, the length of the respective bar represents its contribution. Bars of equal length for each position correspond to a situation where no specific positional preference is observed.

The propensities of amino acids to occur at the terminating position T and position T + 1 in both types of helices were computed (Table 1). Gly residues have an overwhelming preference to occur as the $\alpha_L$-terminator consistent with the strong intrinsic preference of this achiral residue for $\alpha_L$-conformations (Ramakrishnan & Srinivasan, 1990). The only other residue with a significant propensity to occur at position T in $\alpha_L$-terminated helices is Asn, a residue with a relatively high propensity for $\alpha_L$ conformation (Srinivasan *et al.*, 1994). Quite remarkably, in the case of E-terminated helices, only a single example of a Pro residue was identified at position T. The four residues with the greater propensity to occur at position T are His, Asn, Leu and Phe. The much lower propensity of the β-branched residues Val, Ile and Thr suggests that these residues have a significantly diminished tendency to drift away from the helical region when placed at helix C termini. Although β-branching has been considered to enhance propensities for extended conformations (Smith & Regan, 1995; Miner & Kim, 1994), it is likely that the ($\phi$, $\psi$) preferences at individual residues (Munoz & Serrano, 1994; Swindells *et al.*, 1995) may be dominated by sequence context effects.

Despite its absence at position T, in the E-terminated helices, Pro is a very strong helix breaker as

**Table 1.** Amino acid propensity at helix termini in T and T + 1 positions

| Amino acids | Residue at T | | | | Residue at T + 1 | | | |
|---|---|---|---|---|---|---|---|---|
| | $\alpha_L$[a] | | E[b] | | $\alpha_L$ | | E | |
| | No.[c] | P[d] | No. | P | No. | P | No. | P |
| ALA | 7 | 0.17 | 41 | 1.00 | 43 | 1.03 | 23 | 0.56 |
| ARG | 9 | 0.47 | 22 | 1.18 | 21 | 1.10 | 17 | 0.91 |
| ASN | 48 | **2.15** | 40 | **1.84** | 10 | 0.45 | 22 | 1.01 |
| ASP | 21 | 0.73 | 36 | 1.29 | 10 | 0.35 | 44 | **1.57** |
| CYS | 1 | 0.12 | 10 | 1.23 | 4 | 0.48 | 2 | 0.24 |
| GLN | 10 | 0.59 | 12 | 0.73 | 12 | 0.71 | 17 | 1.03 |
| GLU | 10 | 0.38 | 7 | 0.27 | 20 | 0.76 | 15 | 0.58 |
| GLY | 318 | **8.13** | 20 | 0.52 | 33 | 0.84 | 32 | 0.83 |
| HIS | 12 | 1.23 | 22 | **2.31** | 10 | 1.02 | 6 | 0.63 |
| ILE | 0 | 0.00 | 18 | 0.75 | 47 | **1.91** | 4 | 0.17 |
| LEU | 1 | 0.03 | 62 | **1.69** | 62 | **1.66** | 10 | 0.27 |
| LYS | 25 | 0.88 | 21 | 0.76 | 35 | 1.23 | 38 | 1.37 |
| MET | 0 | 0.00 | 5 | 0.59 | 10 | 1.15 | 3 | 0.35 |
| PHE | 2 | 0.11 | 31 | **1.69** | 27 | **1.45** | 7 | 0.38 |
| PRO | 0 | 0.00 | 1 | 0.05 | 3 | 0.14 | 153 | **7.37** |
| SER | 3 | 0.09 | 37 | 1.21 | 13 | 0.41 | 30 | 0.98 |
| THR | 0 | 0.00 | 16 | 0.56 | 31 | 1.01 | 25 | 0.88 |
| TRP | 0 | 0.00 | 7 | 1.03 | 10 | **1.45** | 1 | 0.14 |
| TYR | 2 | 0.11 | 23 | 1.35 | 21 | 1.21 | 3 | 0.17 |
| VAL | 0 | 0.00 | 27 | 0.83 | 46 | 1.34 | 7 | 0.22 |

[a] Terminating residue adopting left-handed helical ($\alpha_L$) conformation.
[b] terminating residue adopting extended (E) conformation.
[c] Number of occurrences.
[d] Propensity of the residue, values $\geqslant 1.45$ are in bold.

seen by its very high propensity at position T + 1 (Table 1). One third of the examples (153) of the E-terminated helices have Pro at position T + 1. The reasons for this are not hard to see. The presence of the bulky $C^{\delta}$ $H_2$ substituent on the nitrogen atom results in the disruption of a potential intra-helical hydrogen bond. Steric adjustments to accommodate the Pro residues can result in changes of ($\phi$, $\psi$) values of the preceding residue (Hurley *et al.*, 1992; MacArthur & Thornton, 1991). As a consequence, Pro appears almost exclusively at position T + 1, with position T being occupied by some other residue. Local stereochemical consequences of Pro in helices have indeed been analysed extensively in the literature (Barlow & Thornton, 1988; MacArthur & Thornton, 1991; von Heijne, 1991; Hurley *et al.*, 1992; Strehlow *et al.*, 1991). Although Gly has a very high propensity to occur at the position T in the $\alpha_L$-terminated helices, and Pro a correspondingly high propensity to occur at the position T + 1 in the E-terminated helices, we found only seven examples of Gly-Pro segments occupying the helix terminus T and T + 1 positions. Significantly, in none of these examples did Gly adopt an $\alpha_L$ conformation. This is presumably a consequence of the absence of the $6 \rightarrow 1$ hydrogen bond interaction in the Schellman motif (*vide infra*), formed in the $\alpha_L$-terminated helices, since the amide hydrogen is missing in Pro.

## Schellman motifs and $\alpha_L$-terminated helices

Charlotte Schellman recognized many years ago that Gly residues in $\alpha_L$ conformations occur frequently at the C-terminal end of helices in proteins (Schellman, 1980). A consequence of this stereochemical feature is the occurrence of a $6 \rightarrow 1$ hydrogen bond between the N–H group of residue T + 1 and the C=O group of residue T-4. In many cases, the second $4 \rightarrow 1$ (sometimes denoted as $5 \rightarrow 2$) hydrogen bond is observed between the N–H group of residue T and the C=O group of residue T-3. While the presence of both $6 \rightarrow 1$ and $4 \rightarrow 1$ hydrogen bonds have been considered as characteristics of the Schellman motif (Aurora *et al.*, 1994; Milner-White, 1988; Karle *et al.*, 1993; Nagarajaram *et al.*, 1993), we would prefer to describe this stereochemical feature by the backbone conformational designation, namely $\alpha_R$-$\alpha_R$-$\alpha_R$-$\alpha_L$, where a succession of three residues in right-handed $\alpha$-helical conformations is followed by a screw sense reversal at the C-terminal end of the helix. As much as 44% (469) of the helices in the present data set of 1057 terminate in the Schellman motif. An overwhelming preponderance of Gly is noted at the terminating position. Although Gly is frequently the helix terminator, glycine residues are also found to occur with reasonable frequency within the body of the helix. In the present data set, 27% ($N = 289$) of the examples had Gly residues within the helix and 30% ($N = 318$) had Gly acting as $\alpha_L$-terminators.

The propensities of the amino acid residues to occur at the T + 1 position in $\alpha_L$-terminated helices were examined (Table 1). The hydrophobic residues Ile, Leu, Phe, Trp and Val have significant propensities for the T + 1 positions. Indeed, earlier analyses of the Schellman motif have, in fact, pointed to the occurrence of hydrophobic residues at positions T − 4 and T + 1 (Preissner & Bork, 1991; Aurora *et al.*, 1994). In order to examine whether the motif hxxxGh (h, hydrophobic resi-

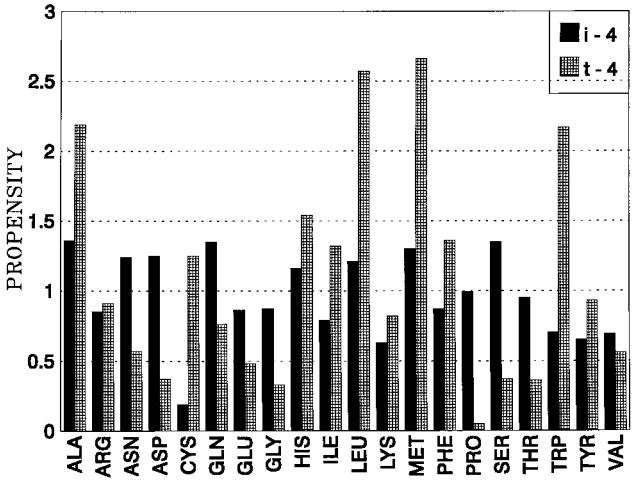


**Figure 3.** Histogram showing the propensity of the amino acids to occur at position T − 4 when Gly occurs at position T and terminates the helix, compared with the propensity to occur at the position $i - 4$ when Gly, placed at the position $i$, continues the helix.

dues: A, L, I, M, F, P, Y, W, V; x, any residue) is a strong determinant of the terminating Schellman motif, we searched for such patterns in cases where Gly residues occur within helices. Of the 318 $\alpha_L$-terminated helices, this motif was identified in 46% ($N = 148$) of the examples. In the case of 289 examples of Gly within helices, 25% ($N = 72$) of the examples had this motif. It is therefore difficult to classify such a motif as a strong Schellman terminator. Figure 3 summarizes the propensity of the amino acids to occur four residues (T − 4) ahead of Gly in both the Schellman motif and continuing helices. A preponderance of hydrophobic residues at position T − 4 was observed. In order to provide a stronger predictor for helix terminating Schellman motifs, we examined sequence segments ranging from position T − 4 and T + 3 with Gly at position T (xxxxGxxx). Classifying residues as hydrophobic (h: A, I, L, M, P, F, Y, W, V) and polar (p: R, N, D, C, Q, E, G, H, K, S, T) as many as ($2^7$) 128 motifs may be generated. Of these, only 16 were exclusively observed in Gly terminated Schellman motifs, with 57 out of 318 examples being represented (Table 2). 25 belonged uniquely to the cases where Gly was accommodated in the helix (Table 3). When the characterizing motif is limited to seven residues (xxxxGxx) only two motifs were unique in helix terminating position. Further truncation of motifs to six residues resulted in no unique examples. These results suggest that local sequence (T − 4 to T + 3) effects are a strong determinant of Gly conformation in helices and that unique helix stop signals are generated by local interaction between the amino acids flanking the achiral residue, which has an intrinsically high $\alpha_L$ propensity.

## Hydrogen bonding patterns in the Schellman motif

In the conventional definition of Schellman motifs, the presence of both 6 → 1 and 5 → 2 intramolecular hydrogen bonds within the six residue segment is taken as clear characteristic of this specific stereochemical feature (Schellman, 1980; Milner-White, 1989; Karle *et al*., 1993; Datta *et al*., 1997). Since Schellman motifs are frequently found at helix termini, solvent invasion and consequent hydrogen bond distortions must be considered. It has frequently been observed in peptide structures

**Table 2.** Sequence patterns exclusively observed in $\alpha_L$ (Gly) terminated helices

| Patterns | | | | | | | | No. of |
|---|---|---|---|---|---|---|---|---|
| T − 4 | T − 3 | T − 2 | T − 1 | T | T + 1 | T + 2 | T + 3 | occurrences |
| Segment characterized: T − 4 to T + 3 | | | | | | | | |
| h | h | h | h | G | h | p | p | 4 |
| h | h | h | p | G | h | h | h | 1 |
| h | h | p | h | G | h | p | p | 11 |
| h | h | p | p | G | p | h | h | 2 |
| h | h | p | p | G | p | h | p | 7 |
| h | p | h | p | G | p | h | p | 2 |
| h | p | h | p | G | p | p | h | 4 |
| h | p | p | h | G | h | h | p | 4 |
| h | p | p | p | G | h | h | h | 5 |
| h | p | p | p | G | h | h | p | 2 |
| h | p | p | p | G | p | h | p | 4 |
| p | h | p | p | G | p | h | p | 4 |
| p | p | h | h | G | h | p | p | 2 |
| p | p | h | h | G | p | p | p | 1 |
| p | p | p | h | G | p | h | p | 3 |
| p | p | p | p | G | p | h | h | 1 |
| Segment characterized: T − 4 to T + 2 | | | | | | | | |
| h | h | p | p | G | p | h | x | 9 |
| h | p | p | p | G | h | h | x | 7 |
| Segment characterized: T − 4 to T + 1 (no examples found) | | | | | | | | |

Symbols used: h, hydrophobic residues: A, L, I, M, F, P, Y, W, V; p, polar residues: R, N, D, C, Q, E, G, H, K, S, T; x, any residue. There are 318 Gly ($\alpha_L$) terminated helices present in the data set.

**Table 3.** Sequence patterns exclusively observed in cases where Gly (position i), accommodated in the helical segment

| Patterns | | | | | | | | No. of |
|---|---|---|---|---|---|---|---|---|
| $i-4$ | $i-3$ | $i-2$ | $i-1$ | $i$ | $i+1$ | $i+2$ | $i+3$ | occurrences |
| Segment characterized: $i-4$ to $i+3$ | | | | | | | | |
| h | h | h | h | G | h | h | h | 3 |
| h | h | h | h | G | p | p | h | 3 |
| h | h | h | h | G | p | p | p | 1 |
| h | h | h | p | G | h | h | p | 3 |
| h | h | p | h | G | p | p | h | 2 |
| h | p | h | p | G | h | h | p | 2 |
| p | h | h | h | G | h | h | p | 4 |
| p | h | h | h | G | h | p | p | 7 |
| p | h | h | h | G | p | p | h | 5 |
| p | h | h | p | G | h | h | h | 5 |
| p | h | h | p | G | h | h | p | 3 |
| p | h | h | p | G | h | p | p | 1 |
| p | h | p | h | G | h | h | h | 1 |
| p | h | p | h | G | p | h | h | 3 |
| p | h | p | h | G | p | p | h | 1 |
| p | h | p | p | G | p | p | h | 1 |
| p | p | h | h | G | p | h | h | 6 |
| p | p | h | p | G | h | h | h | 3 |
| p | p | h | p | G | p | h | p | 2 |
| p | p | h | p | G | p | p | h | 3 |
| p | p | p | h | G | p | h | h | 4 |
| p | p | p | h | G | p | p | h | 1 |
| p | p | p | h | G | p | p | p | 1 |
| p | p | p | p | G | h | h | p | 4 |
| p | p | p | p | G | h | p | p | 2 |
| Segment characterized: $i-4$ to $i+2$ | | | | | | | | |
| h | h | h | h | G | p | p | x | 4 |
| p | h | h | p | G | h | h | x | 8 |
| p | p | p | h | G | p | p | x | 2 |
| Segment characterized: $i-4$ to $i+1$ (no examples found) | | | | | | | | |

Symbols used: h, hydrophobic residues: A, L, I, M, F, P, Y, W, V; p, polar residues: R, N, D, C, Q, E, G, H, K, S, T; x, any residue. 289 examples of Gly accommodated helical segments were observed in the data set.

that solvent insertion into intra-helical hydrogen bonds results in dramatic lengthening of N—O distance even though backbone dihedral angles still lie within the broadly helical region (Karle *et al.*, 1996; Thanki *et al.*, 1991; Baker & Hubbard, 1984; Blundell *et al.*, 1983). In the present study, backbone ($\phi$, $\psi$) angles have been considered to be a better descriptor of the Schellman motif, which is defined as an $\alpha_R$-$\alpha_R$-$\alpha_R$-$\alpha_L$ motif. Of the 469 $\alpha_L$-terminated helices, 72% ($N = 340$) of the examples possessed $6 \rightarrow 1$ hydrogen bonds, while 62% ($N = 290$) possessed both $6 \rightarrow 1$ and $5 \rightarrow 2$ hydrogen bonds as defined by N---O distances $\leqslant$3.5 Å (Figure 4(a)). All the 129 cases with N---O distance >3.5 Å were examined in detail. It was found that in a very large majority of these examples water insertion into the hydrogen bond (Figure 4(b)) or side-chain insertion (Ser, Thr hydroxyl group; Figure 4(c)) was observed. We re-examined the amino acid propensities at position $T+1$ in $\alpha_L$-terminated helices which lack the internal $6 \rightarrow 1$ hydrogen bond within the Schellman motif. The highest propensities were observed for Val(1.9), Thr(1.7), Trp(1.6), Ile(1.6) and Leu (1.5), suggesting that hydrophobicity was still a major determining factor.

## E-terminated helices

Table 1 reveals that unlike in the case of $\alpha_L$-terminated helices there is no overwhelmingly strong preference for any single residue at the terminating position T in the case of E-terminated helices. Inspection of position $T+1$ propensities in Table 1 reveals a very high propensity for Pro, with about a third of the examples in the data set falling into this group. We have therefore chosen a sub-class of E-terminated helices with Pro at $T+1$ position for further analysis. Figure 5 summarizes the occurrence of residues which follow the Pro residue at helix termini. A clear preponderance of polar and charged residues like Asp, Glu and Gln is evident. Since Pro residues are also found within helices it was of interest to examine the propensities of amino acids following the Pro in these cases. There are 78 examples of Pro residues in distinct helices after removal of Pro in N-cap and C-cap segments. Figure 5 shows that the propensity of amino acids to follow Pro in helices is dramatically different from that observed at helix termini. Helix continuation at Pro appears to be facilitated by the presence of hydrophobic residues like Ile, Leu and Met in the position succeeding Pro. His also has a very high propensity at this position.

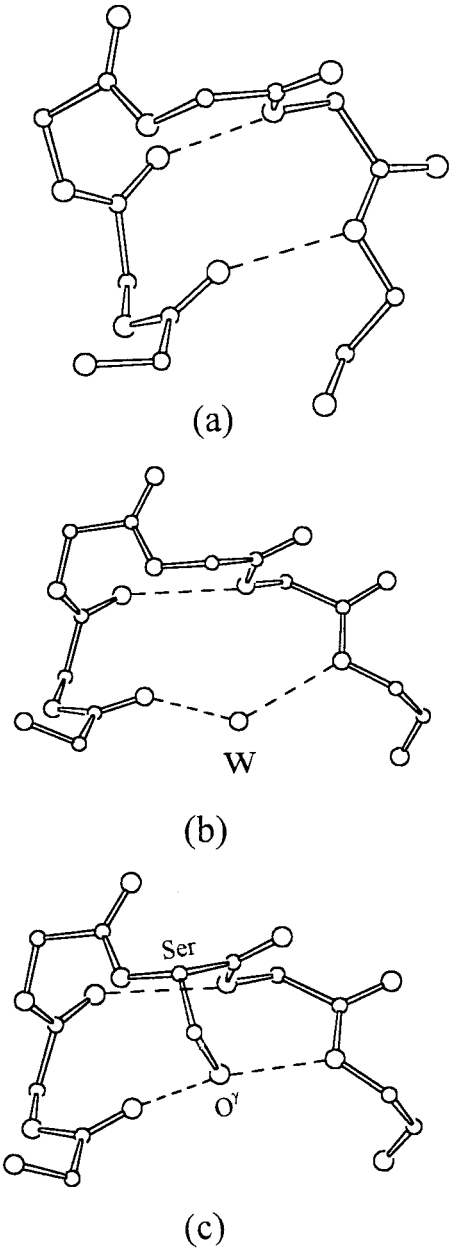


**Figure 4.** Structural representatives of Schellman motifs. (a) An ideal motif with $6 \rightarrow 1$ and $5 \rightarrow 2$ hydrogen bond observed in serum amyloid P component (1SAC 171 to 176). (b) Schellman motif with water insertion into the $6 \rightarrow 1$ hydrogen bond observed in histidine-containing phosphocarrier protein. (c) Ser162 OG insertion into the $6 \rightarrow 1$ hydrogen bond observed in serine carboxypeptidase II (1WHT 159 to 164). For clarity, only backbone atoms are shown.

In order to establish contextual effects on the conformations of Pro residues in helical segments, we searched for specific patterning of hydrophobic (h) and polar (p) residues, using windows of variable length. Using a seven-residue motif, out of the 64 possible patterns 49 were found to occur in helices. Of these, 20 were uniquely observed only in cases where the helix is terminated at the residue preceding Pro (Table 4), while 10 were unique to the cases where the helix continued after the Pro residue (Table 5). A clear feature of the amino acid sequence patterns that emerges is that helix continuation at Pro is favoured when the flanking residues are largely hydrophobic, while termination is favoured when the segment containing Pro is polar.

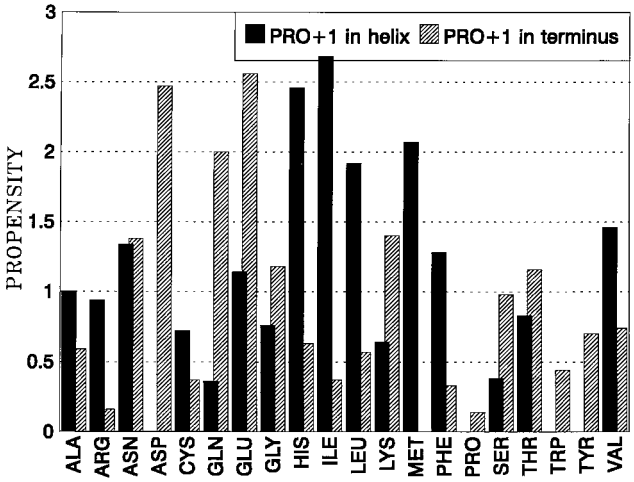


**Figure 5.** Histogram showing the propensity of the amino acids to occur following Pro when Pro occurs at the helix terminus compared with the case when Pro occurs in the middle of a helix.

## Helix terminating sequence patterns

In order to derive rules which will permit identification of Gly and Pro signals for helix continuation or termination we have examined the frequencies of these two stereochemical possibilities using sequence windows of eight residues in both cases. Gly was placed at position 5 (T) in the eight residue window since $(T-4)/(T+1)$ interactions are important in the Schellman motif, when Gly occupies position T with an $\alpha_L$-conformation. Pro ($T+1$ residue) was placed at position 4 so that $T+5/T+1$ interactions were accommodated at the C termini (note that the placement of the terminating residue T within the chosen window differs for the Gly and Pro cases). The frequencies of hydrophobic and polar residues at each of the seven variable positions of the motif was computed for the helix data set (Table 6). It is clear that relatively strong preferences for one class of residues are observed at specific positions, while other positions appear to be indifferent to amino acid type. We therefore generated 2187 ($3^7$) motifs of the type (rrrrGrrr, r = h, p or h/p), where hydrophobic (h: A, I, L, M, P, F, Y, W, V), polar (p: R, N, D, C, Q, E, G, H, K, S, T) or both (x) residues could be accommodated at each of the seven positions. These motifs were then examined for their occurrence as helix terminator or continuing helical seg-

**Table 4.** Sequence patterns exclusively observed in cases where Pro (T + 1) terminates helix

| Patterns | | | | | | | | No. of |
|---|---|---|---|---|---|---|---|---|
| T − 2 | T − 1 | T | T + 1 | T + 2 | T + 3 | T + 4 | T + 5 | occurrences |
| | | | Segment characterized: T − 2 to T + 5 (several examples found) | | | | | |
| | | | Segment characterized: T − 2 to T + 4 | | | | | |
| h | h | h | P | h | p | p | x | 1 |
| h | h | h | P | p | p | h | x | 2 |
| h | h | h | P | p | p | p | x | 3 |
| h | h | p | P | p | p | p | x | 3 |
| h | p | h | P | p | h | p | x | 2 |
| h | p | h | P | p | p | h | x | 1 |
| h | p | h | P | p | p | p | x | 2 |
| h | p | p | P | p | h | h | x | 10 |
| h | p | p | P | p | h | p | x | 8 |
| h | p | p | P | p | p | h | x | 6 |
| h | p | p | P | p | p | p | x | 2 |
| p | h | p | P | h | p | h | x | 1 |
| p | h | p | P | p | p | h | x | 1 |
| p | p | h | P | h | h | h | x | 1 |
| p | p | h | P | h | p | p | x | 3 |
| p | p | h | P | p | p | p | x | 3 |
| p | p | p | P | h | h | h | x | 2 |
| p | p | p | P | h | p | h | x | 3 |
| p | p | p | P | p | h | p | x | 7 |
| p | p | p | P | p | p | h | x | 4 |
| | | | Segment characterized: T − 2 to T + 3 | | | | | |
| h | h | h | P | p | p | x | x | 5 |
| h | p | h | P | p | p | x | x | 3 |
| h | p | p | P | p | h | x | x | 18 |
| h | p | p | P | p | p | x | x | 8 |
| p | h | p | P | p | p | x | x | 1 |
| p | p | h | P | h | p | x | x | 3 |
| p | p | p | P | h | p | x | x | 3 |
| | | | Segment characterized: T − 2 to T + 2 | | | | | |
| h | p | p | P | p | x | x | x | 26 |
| | | | Segment characterized: T − 2 to T + 1 (no examples found) | | | | | |

Symbols used: h, hydrophobic residues: A, L, I, M, F, P, Y, W, V; p, polar residues: R, N, D, C, Q, E, G, H, K, S, T; x, any residue. Out of 173 Pro (T + 1) terminated helices, 153 have extended conformation for the residue preceding Pro.

**Table 5.** List of patterns exclusively observed in cases where Pro (position *i*), accommodated in the helical segment

| Patterns | | | | | | | | No. of |
|---|---|---|---|---|---|---|---|---|
| *i* − 3 | *i* − 2 | *i* − 1 | *i* | *i* + 1 | *i* + 2 | *i* + 3 | *i* + 4 | occurrences |
| | | | Segment characterized: *i* − 3 to *i* + 4 (several examples found) | | | | | |
| | | | Segment characterized: *i* − 3 to *i* + 3 | | | | | |
| h | h | h | P | h | h | h | x | 2 |
| h | h | h | P | h | h | p | x | 3 |
| h | h | h | P | p | h | h | x | 3 |
| h | h | p | P | h | p | h | x | 3 |
| h | p | h | P | h | h | h | x | 1 |
| h | p | h | P | h | h | p | x | 5 |
| h | p | h | P | h | p | p | x | 2 |
| p | h | h | P | h | h | h | x | 3 |
| p | h | h | P | p | p | p | x | 1 |
| p | h | p | P | h | p | p | x | 1 |
| | | | Segment characterized: *i* − 3 to *i* + 2 | | | | | |
| h | h | h | P | h | h | x | x | 5 |
| h | p | h | P | h | h | x | x | 6 |
| | | | Segment characterized: *i* − 3 to *i* + 1 (no examples found) | | | | | |

Symbols used: h, hydrophobic residues: A, L, I, M, F, P, Y, W, V; p, polar residues: R, N, D, C, Q, E, G, H, K, S, T; x, any residue. 78 examples of Pro accommodated helical segments were observed in the data set.

**Table 6.** The occurrence of hydrophobic (h) and polar (p) residues in proximal positions

A. *When Gly (position T) terminates the helix and when Gly (position 'i') occurs in the helical segment*

| Type | Position | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Gly-Terminus | | | | | | | |
| | T − 4 | T − 3 | T − 2 | T − 1 | T | T + 1 | T + 2 | T + 3 |
| h | 228 | 136 | 117 | 146 | 0 | 189 | 129 | 161 |
| p | 90 | 182 | 201 | 172 | 318 | 129 | 189 | 157 |
| | Gly-Middle | | | | | | | |
| | *i* − 4 | *i* − 3 | *i* − 2 | *i* − 1 | *i* | *i* + 1 | *i* + 2 | *i* + 3 |
| h | 129 | 143 | 165 | 167 | 0 | 169 | 150 | 165 |
| p | 160 | 146 | 124 | 122 | 289 | 120 | 139 | 124 |

B. *When Pro (position T + 1) terminates the helix and when Pro (position i) occurs in the helical segment*

| Type | Position | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Pro-Terminus | | | | | | | |
| | T − 2 | T − 1 | T | T + 1 | T + 2 | T + 3 | T + 4 | T + 5 |
| h | 69 | 69 | 78 | 173 | 37 | 99 | 79 | 88 |
| p | 104 | 104 | 95 | 0 | 136 | 74 | 94 | 83 |
| | Pro-Middle | | | | | | | |
| | *i* − 3 | *i* − 2 | *i* − 1 | *i* | *i* + 1 | *i* + 2 | *i* + 3 | *i* + 4 |
| h | 37 | 50 | 52 | 78 | 45 | 57 | 39 | 38 |
| p | 41 | 28 | 26 | 0 | 33 | 21 | 39 | 40 |

ments. The highest scoring motifs, that is, motifs which were found most frequently at helix terminating positions as compared to continuing helical segments are listed in Table 7. For Gly terminating motifs, five patterns were found to occur at least three times as often as in continuing helical segments. An earlier analysis of Gly terminated helices (Aurora *et al.*, 1994) reported a motif (hxpxGh) as having high propensity for termination. This motif is in fact listed in Table 7. It should be noted that the remaining four motifs show similar propensities to occur at helix termini. In Table 7 the high preference of motifs for helix termini is statistically significant. However, ranking of motifs on statistical criteria is not desirable at present. Examination of helix length for Gly terminated helices and for Gly continuing helices reveals that there are only six helices, having length <ten residues, with Gly in the helical segment as against 90 helices with Gly at the terminating (T) position. This might

**Table 7.** Highest scoring helix terminating motifs

| | | | | | Gly | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| T − 4 | T − 3 | T − 2 | T − 1 | T | T + 1 | T + 2 | T + 3 | RS[a] | RE[b] | Total[c] |
| h | x | p | p | Gly | x | x | x | 87 | 21 | 586 |
| h | x | p | x | Gly | x | x | p | 72 | 18 | 575 |
| h | x | p | x | Gly | h | x | x | 98 | 33 | 494 |
| h | x | x | p | Gly | x | x | h | 68 | 22 | 535 |
| h | p | x | p | Gly | x | x | x | 67 | 19 | 649 |
| | | | | | Pro | | | | | |
| T − 2 | T − 1 | T | T + 1 | T + 2 | T + 3 | T + 4 | T + 5 | RS[a] | RE[b] | Total[c] |
| x | x | x | Pro | p | x | p | x | 67 | 11 | 688 |
| x | p | x | Pro | p | x | x | x | 67 | 10 | 729 |
| x | x | p | Pro | p | x | x | x | 65 | 10 | 728 |
| x | x | x | Pro | p | x | x | h | 63 | 14 | 608 |
| x | p | p | Pro | x | x | x | x | 58 | 8 | 727 |
| h | x | x | Pro | p | x | x | x | 51 | 12 | 587 |
| p | p | x | Pro | x | x | x | x | 50 | 11 | 695 |
| x | x | x | Pro | p | h | p | x | 49 | 9 | 323 |
| x | x | x | Pro | p | p | x | x | 46 | 5 | 719 |
| p | x | x | Pro | p | x | p | x | 43 | 8 | 381 |
| x | p | p | Pro | p | x | x | x | 43 | 2 | 400 |
| x | p | x | Pro | x | p | x | x | 39 | 8 | 688 |
| x | p | x | Pro | p | h | x | x | 38 | 8 | 311 |
| x | x | p | Pro | p | h | x | x | 38 | 8 | 316 |
| h | x | p | Pro | p | x | x | x | 38 | 5 | 334 |
| x | p | x | Pro | p | x | x | h | 38 | 5 | 348 |

The highest scoring motifs are the motifs which occur in at least 20% (25% for Pro case) of the examples and have greater than 3:1 (4:1 for Pro case) preference for the helix terminating region over the continuing segment. Symbols used: h-hydrophobic residues: A, L, I, M, F, P, Y, W, V; p, polar residues: R, N, D, C, Q, E, G, H, K, S, T ; x, any residue.

[a] RS, Number of occurrences of the motif at the helix terminating region.

[b] RE, Number of occurrences of the motif within continuing helical segments.

[c] Total, Number of occurrences of the motif in the data set of 250 proteins, regardless of secondary structure.

indicate that the continuing helices with long intramolecular hydrogen bonding networks may compensate for the intrinsic tendency of Gly to break the helix.

In the case of E-terminated helices with Pro at position T + 1, several motifs which have high helix terminating propensity are identifiable even using a more stringent criterion, that the number of termination examples must be at least four times as large as helix continuation examples. The results in Table 7 establish that Pro residues flanked by polar amino acids have a very strong tendency to terminate helices. Identification of Pro containing helix terminating motifs may be far less ambiguous than recognition of Gly at helix stop signals. Qualitatively, it is easy to answer the question as to when a Pro containing segment will act as a helix terminator and when the segment is likely to form a continuous helix. The presence of Pro within a helical segment will interrupt the regular intramolecular hydrogen bonding scheme. Accommodation of the $C^{\delta}$ methylene group on the nitrogen further results in kinking or bending of the helix (Ramachandran *et al.*, 1963; Barlow & Thornton, 1988; MacArthur & Thornton, 1991; von Heijne, 1991; Sankararamakrishnan & Vishveshwara, 1992). It is therefore not surprising that helix continuation is favoured in a hydrophobic environment where hydration and consequent unwinding is disfavoured. It is noteworthy that Pro residues are found quite extensively in transmembrane helices which are almost completely hydrophobic in sequence (von Heijne, 1991; Sansom, 1992; Brandl & Deber, 1986; Woolfson *et al.*, 1991). When a Pro residue is found in a hydrophobic stretch, helix continuation appears to be the favoured alternative. In the present data set there were only ten examples of Pro residues flanked on either side by polar amino acids found in continuing helical segments (Table 7). Examination of each of these cases revealed that in many instances compensating local interactions like salt bridge formation between positions $i/i+3$ or $i+4$ or side-chain backbone hydrogen bonding may have been the determining factor.

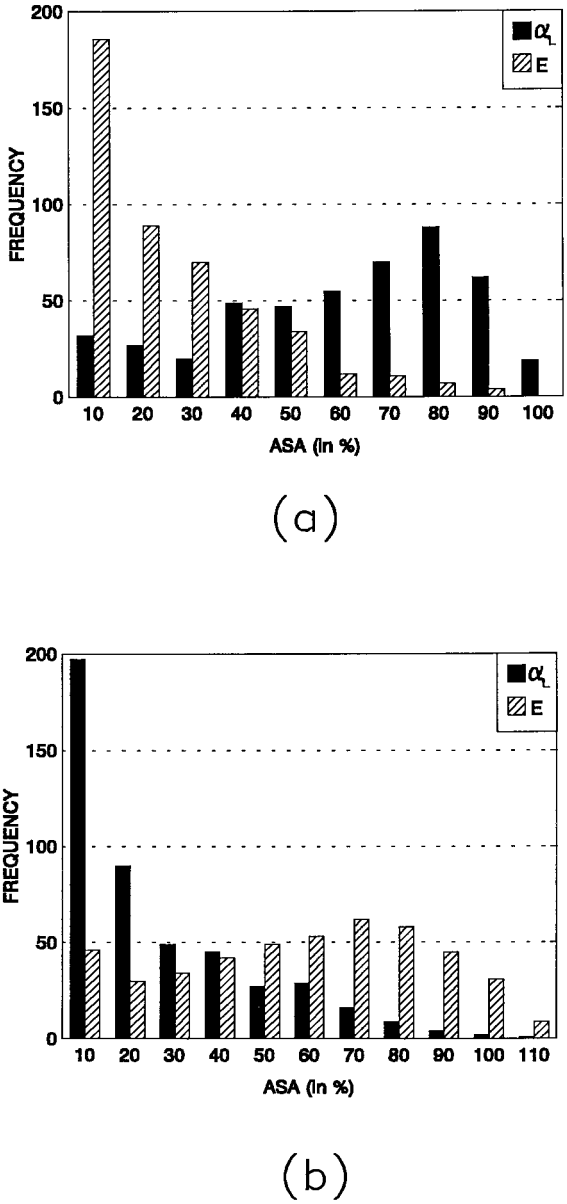


**Figure 6.** Distribution of accessible solvent area (ASA; Lee & Richards, 1971) for the position (a) T and (b) T + 1 in the case of $\alpha_L$ and E-terminated helices.

## Environment of helix termini

The distinctive amino acid propensities at the C-terminal end of $\alpha_L$ and E-terminated structures prompted us to examine the solvent accessibilities of the residues at position T and T + 1. Figure 6 summarizes the results of accessible solvent area (ASA) (Lee & Richards, 1971) calculations for $\alpha_L$ and E-terminated helices. At position T, E-terminated helices showed diminished solvent accessibility as compared to the $\alpha_L$-terminated structures. In sharp contrast, the situation is reversed at position T + 1, which is almost completely inaccessible in $\alpha_L$-terminated helices, but substantially solvent exposed in E-terminated helices. Figure 7 shows typical views of the two kinds of helix termini observed in proteins. In most cases, in $\alpha_L$-terminated helices the T + 1 residue is buried by pushing it back into contact with the helix, a feature leading to T − 4/T + 1 interactions (Aurora *et al.*, 1994, Viguera & Serrano, 1995). In the E-terminated helices, the structure following the terminator residue is more open and consequently, solvated.

The occurrence of Pro and the positively charged residues at the T + 1 position, in the case of E-terminated helices, prompted us to examine the possibility of T + 1 initiating a succeeding helix or strand. Of the 459 E-terminated helices, in 16% ($N = 72$) of the cases the T + 1 position was found to initiate a helix (of length $\geqslant 4$ residues) and in 10% ($N = 45$) T + 1 initiated strands. The strands were identified based on the criterion that the backbone dihedral

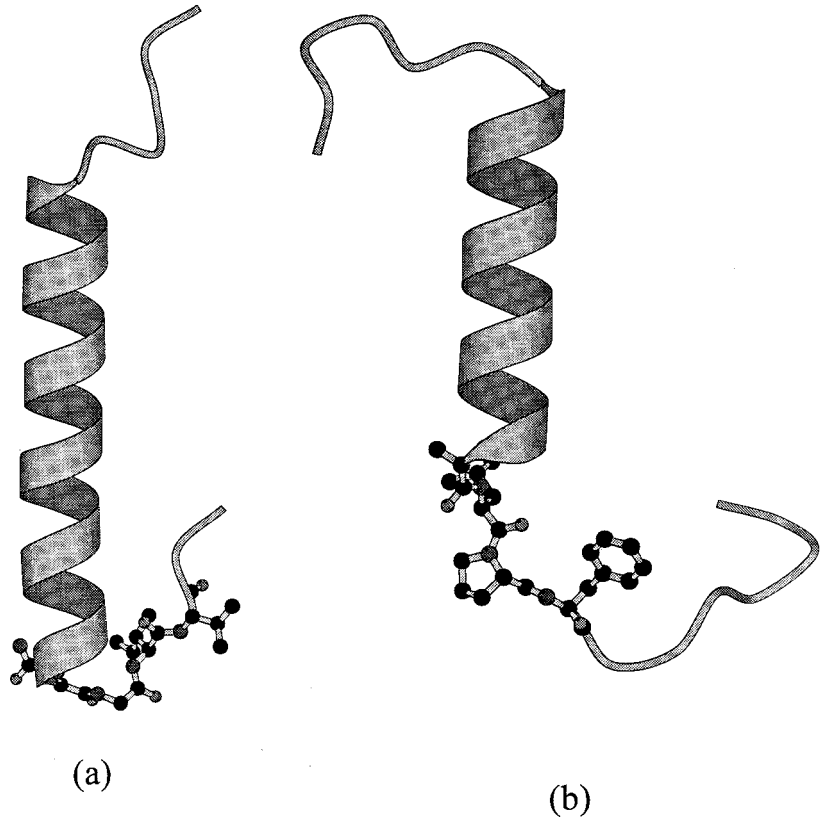


**Figure 7.** Ribbon representation of (a) left-handed helical ($\alpha_L$) and (b) extended (E) conformation terminated helices as observed in proteins aldolase A (1ALD 155 to 185; Asn180, Gly181, Ile182 and Val183 are shown in ball and stick model) and α-momorcharin (1AHC 5 to 34; Ala24, Leu25, Pro26 and Phe27 are shown in ball and stick), respectively. The picture was drawn using the MOLSCRIPT program (Kraulis, 1991).

angles (ϕ,ψ) of the successive residues, in a segment of length $\geqslant 4$, should lie in the extended region (E:$\phi = -180^\circ$ to $-30^\circ$ and $\psi = 60^\circ$ to $180^\circ$ and $-180^\circ$ to $-150^\circ$; Sowdhamini *et al.*, 1992). Of the 72 examples, which initiated helix, in 38 cases T + 1 position was occupied by the residue Pro. Interestingly, in the case of $\alpha_L$-terminated helices, only in 2% ($N = 9$) of the cases was the T + 1 position found to initiate succeeding helix formation as against 22% ($N = 102$) initiating strands.

## Conservation of helix terminating residues

The conservation of residues and local stereochemistry at helix termini was examined by considering homologous protein structures available in the PDB at a resolution $\leqslant 2.0$ Å (Gunasekaran *et al.*, 1996). Of the 469 $\alpha_L$-terminated helices, homologous structures were available for 53% ($N = 249$) cases. In 225 examples, both residue and local conformation ($\alpha_L$) were conserved and in 20 cases residue changes were observed but local conformation at position T was conserved. Interestingly, in these cases the terminating residues in the parent structure were G, N, D, R and K and the replacements were always by polar residues (G, N, S, Q, H, K and R) with only one example each of A and Y. In two cases (parent examples:- P21 protein 5P21, Pai *et al.*, 1990: helix segment Ser127 to Tyr137; adipocyte lipid-binding protein 1LIB, Xu *et al.*, 1993: helix segment Phe27 to Ala36) conformational changes at the helix terminus were observed although the terminating residues were conserved. Examination of the backbone dihedral angels at these positions suggests that peptide bond flips may indeed restore the conformational similarity in the homologous structures. A single example of a residue replacement accommodated by a conformational change was observed (parent: actinidin 2ACT; Baker & Dodson, 1980: helix segment Ile70 to Asp80; homologous protein: papain 1PIP; Yamamoto *et al.*, 1992: Pro68 to Gly77). However, in this case there is a residue deletion in the body of the helix and poor sequence identity in the helical segment.

We also examined sequence variability in an eight residue segment with the terminating residue in a central position using the Homology Derived Secondary Structure of Proteins (HSSP) database (Sander & Schneider, 1991). Only those proteins for which the HSSP data base is available with number of aligned sequences $\geqslant 5$ were considered for the analysis. The results are summarized in Figure 8. For $\alpha_L$-terminated helices with Gly at position T, a very high degree of conservation is observed at position T and T − 4. The importance of a hydrophobic residue at position T − 4 in the Schellman motif had been noted earlier (Aurora *et al.*, 1994; Preissner & Bork, 1991). In the case of E-terminated helices with Pro at T + 1, a slightly greater degree of conservation is observed for residue T + 1 as compared to residue T. For E-terminated helices which have non-Pro residues at

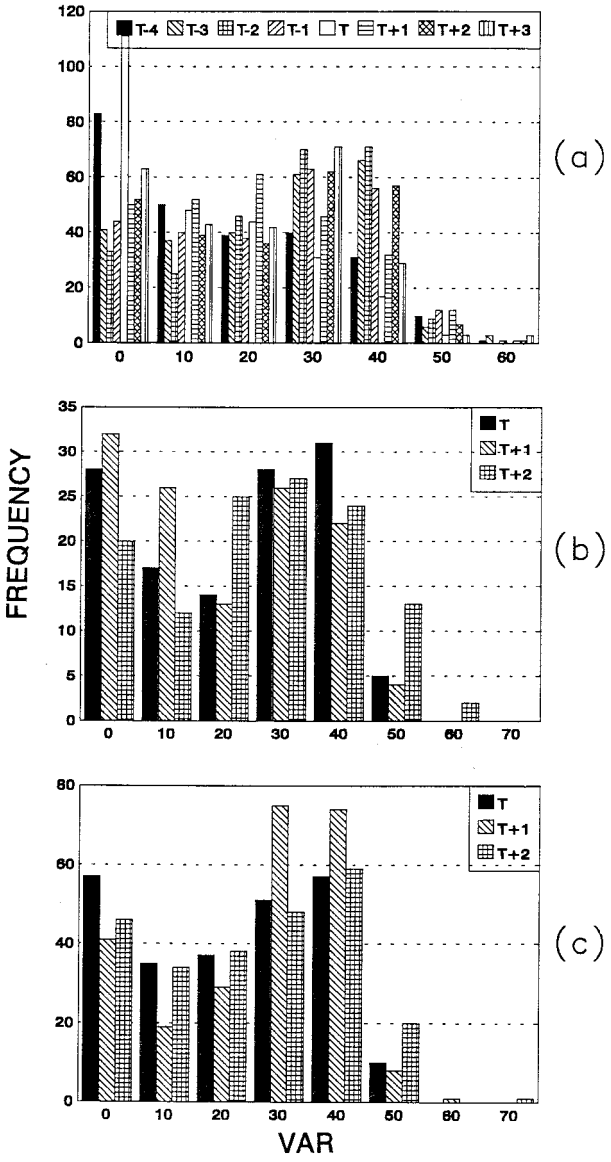


**Figure 8.** Histogram showing the distribution of position dependent sequence variability (VAR) on a scale of 0 to 100 (Sander & Schneider, 1991) (a) for Gly (at T) terminated helices, (b) for Pro (at T + 1) E-terminated helices, and (c) for non-Pro (at T + 1) E-terminated helices. The lower the value of VAR the better is the conservation of the residue in the homologous sequence entries.

position T + 1 no clearly discernible pattern of conservation is evident. In the case of $\alpha_L$-terminated helices, the requirement for a positive $\phi$ value at residue T greatly favours the occurrence of Gly, which is the only achiral residue in the genetically coded set of amino acids. In the case of E-terminated helices with Pro at T + 1, proline can, in principle, be replaced by other residues if other local structural factors contribute to helix termination. We also examined the possibility of conversion of one helix-termination motif into another by mutation. In the data set of homologous protein structures extracted from the PDB no example of a conformational switch in homologous protein was observed. Analysis of aligned sequences of homologous proteins from the HSSP database, in which the parent crystal structure was part of our data set revealed a few interesting examples where mutational changes could indeed result in a conformational transition. These include lactate dehydrogenase (parent 6LDH; Abad-Zapatero *et al.*, 1987: Ser128-Pro129 placed at the end of helix was mutated to Gly-Phe in some of the aligned sequences), actinidin (parent 2ACT; Baker & Dodson, 1980: Gln131-Pro132 mutated to Gly-Leu) and glucose permease (parent: 1GPR; for description on 2.2 Å resolution crystal structure see Liao & Herzberg, 1991: Val 124-Pro125 mutated to Gly-Tyr or Gly-Leu).

## Implications for Design and Engineering

The present analysis of a large data base of helices in proteins has permitted delineation of sequence features which determine whether a potential "stop signal" will indeed terminate a helical stretch. In the case of Pro in E-terminated helices, a fairly strong local sequence effect emerges. While Gly in $\alpha_L$-conformations constitutes the single most abundant helix termination signal, there is greater ambiguity in deciding by inspection of a given amino acid sequence, whether a specific Gly residue will act to terminate a helix. Despite the inherent ambiguities, the present analysis may prove of use in developing approaches to synthetic helices with stereochemically well defined terminating segments. An important, recent study by Viguera & Serrano (1995) attempts to design a synthetic helical sequence (Tyr-Gly-Gly-Ser-Lys-Ala-Glu-Ala-Ala-Arg-Ala-X-Ala-Lys-His-Gly-Y-Gly-Gly-$NH_2$) terminating in a Schellman motif. The 19-residue peptide did indeed adopt a helical conformation although no NMR evidence could be obtained for the interaction between the residues X and Y, which are placed at the positions T − 4 and T + 1. The limited stability of intramolecular hydrogen bonds in an exposed environment will undoubtedly facilitate helix fraying at the ends due to solvation, unless appropriate reinforcing, complementary side-chain interactions are introduced. Our analysis of helix termination motifs presented above suggests that design of synthetic Schellman motifs in water soluble peptides requires careful choice of sequence over the entire segment ranging from T − 4 to T + 3 (Tables 2 and 7).

As pointed out by Aurora *et al.* (1994) there are a few interesting examples of site directed mutagenesis in which sequence changes affect helix termini. The mutagenesis study on staphylococcal nuclease (Shortle *et al.*, 1990) and T4 phage lysozyme (Alber *et al.*, 1987) and an analysis of barnase (Horovitz *et al.*, 1991) have in fact been already noted and discussed in an earlier analysis on helix termini

(Aurora *et al*., 1994). A noteworthy study has been carried out on *Escherichia coli* ribonuclease H (Ishikawa *et al*., 1993) in order to increase the thermostability of *E. coli* ribonuclease H by directed mutagenesis, based on the sequence variation between the *E. coli* and *T. thermophilus* enzymes. Insertion of a Gly residue (between Gln80 and Trp81) at the C-terminal end of the αII-helix (Gln72 to Thr79) resulted in the formation of a Schellman motif (Gly77 to Trp81), characterized crystallographically in the mutant, which enhanced the protein stability by 0.4 kcal/mol. An additional mutation within the αII-helix at position T-4, Gly77 → Ala, enhanced the stability by 0.8 kcal/mol, with retention of the Schellman motif. In contrast, the Gly77 → Ala mutation alone (without the insertion of Gly at the C-terminal end) reduced the stability by 0.9 kcal/mol. These observations indicate the importance of the hydrophobic amino acid residue at position T − 4.

## Conclusions

The present analysis suggests that motifs, largely determined by local interactions, which act to terminate secondary structure elements may be recognizable in protein sequences. The definition of local sequences which act as a helix stop signal may be of value in peptide design, protein structure prediction and engineering. The importance of short range structural features like β-turns (Venkatachalam, 1968) in determining folding and stability has been emphasized in a recent study on an immunoglobulin variable domain, which correlates database analysis with the results of site-directed mutagenesis (Ohage *et al*., 1997). Although polypeptide helices have been known for nearly half-a-century, the precise sequences and structural features which act as helix start and stop signals have still not been completely defined. A detailed understanding of stereochemical punctuation marks encoded in protein sequences will facilitate more definitive structure prediction and permit rational design of synthetic mimics for protein structures.

## Methods

A data set of 250, largely non-homologous, high-resolution (⩽2.0 Å) protein structures from the Brookhaven Data Bank (PDB; Bernstein *et al*., 1977) was examined. The data set consisted of the following PDB entries (polypeptide chain identifiers are indicated wherever homologous multiple chains are present)

| | | | | | | |
|---|---|---|---|---|---|---|
| 1AAN | 1AAZ A | 1ABE | 1ABK | 1ACF | 1ACX | 1AFG A |
| 1AHC | 1AK3 A | 1ALC | 1ALD | 1ALK A | 1AMP | 1ANK A |
| 1AOZ A | 1APM E | 1ARB | 1ARP | 1ARS | 1AST | 1BBH A |
| 1BBP A | 1BGC | 1BGH | 1BMD A | 1BRS D | 1BSA A | 1BYB |
| 1CBN | 1CCR | 1CEW I | 1CGT | 1CHM A | 1CMB A | 1COT |
| 1CPC A | 1CPC B | 1CPN | 1CSE E | 1CSE I | 1CSH | 1CTF |
| 1CUS | 1DDT | 1DFN A | 1DMB | 1DRI | 1DSB A | 1ECA |
| 1ESL | 1EZM | 1FAS | 1FDN | 1FGV H | 1FIA A | 1FKF |
| 1FLP | 1FLV | 1FNA | 1FRR A | 1FUS | 1FX1 | 1FXD |
| 1GD1 O | 1GIA | 1GKY | 1GLQ A | 1GLT | 1GOG | 1GOX |
| 1GP1 A | 1GPR | 1HEL | 1HIP | 1HLE A | 1HLE B | 1HOE |
| 1HPI | 1HSB A | 1HSB B | 1HSL A | 1HUW | 1HVK A | 1HYP |
| 1IAG | 1IFB | 1ISA A | 1ISU A | 1LCF | 1LEC | 1LIB |
| 1LIS | 1LLD A | 1LTS A | 1LTS C | 1LTS D | 1MBA | 1MBD |
| 1MDC | 1MJC | 1MOL A | 1MPP | 1NAR | 1NBA A | 1NLK R |
| 1NPC | 1NSC A | 1OLB A | 1ONC | 1OPA A | 1OVA A | 1PDA |
| 1PGB | 1PHC | 1PHP | 1PII | 1PK4 | 1PMY | 1POC |
| 1POH | 1PPA | 1PPB H | 1PPB L | 1PPF E | 1PPT | 1PRN |
| 1PTF | 1PTS A | 1R69 | 1RBP | 1RDG | 1REC | 1RIS |
| 1RNH | 1ROP A | 1SAC A | 1SBP | 1SGT | 1SHA A | 1SHF A |
| 1SHG | 1SIM | 1SLT A | 1SMR A | 1SRD A | 1STN | 1TCA |
| 1TEN | 1TFG | 1TGN | 1TGS I | 1TGX A | 1THB A | 1TML |
| 1TON | 1TRB | 1TRK A | 1UBQ | 1UTG | 1WHT A | 1WHT B |
| 1XIB | 1YPI A | 256B A | 2ACQ | 2ACT | 2ALP | 2APR |
| 2BBK H | 2BBK L | 2BMH A | 2CAB | 2CCY A | 2CDV | 2CHS A |
| 2CI2 I | 2CMD | 2CPL | 2CTV A | 2CY3 | 2CYP | 2END |
| 2FCR | 2GBP | 2GST A | 2HAD | 2HBG | 2HMQ A | 2LH7 |
| 2LHB | 2LTN A | 2LTN B | 2LZM | 2MCM | 2MLT A | 2MNR |
| 2MSB A | 2OHX A | 2OVO | 2PAB A | 2PIA | 2PLT | 2POR |
| 2PRK | 2RHE | 2RSP A | 2SAR A | 2SCP A | 2SGA | 2SN3 |
| 2SPC A | 2TRX A | 2TSC A | 2WRP R | 2ZTA A | 351C | 3APP |
| 3B5C | 3BCL | 3BLM | 3C2C | 3CHY | 3CLA | 3COX |
| 3DFR | 3DNI | 3DRC A | 3EBX | 3EST | 3GRS | 3IL8 |
| 3MDS A | 3PSG | 3RP2 A | 3RUB L | 3RUB S | 3SDH A | 3TGL |
| 4AZU A | 4BP2 | 4CPV | 4ENL | 4FXN | 4GCR | 4I1B |
| 4ICB | 4INS C | 4INS D | 4MT2 | 4TNC | 5CHA A | 5CPA |
| 5FD1 | 5P21 | 5PTI | 5RUB A | 6LDH | 7ACN | 7RSA |
| 8DFR | 8FAB A | 8FAB B | 9WGA A | | | |

## Acknowledgments

The work was supported by a grant from the Council of Scientific and Industrial Research, Government of India.

## References

Abad-Zapatero, C., Griffith, J. P., Sussman, J. & Rossmann, M. G. (1987). Refined crystal structure of dogfish M4 apo-lactate dehydrogenase. *J. Mol. Biol.* **198**, 445–467.

Alber, T., Dao-pin, S., Wilson, K., Wozniak, J. A., Cook, S. P. & Matthews, B. W. (1987). Contributions of hydrogen bonds of Thr157 to the thermodynamic stability of phage T4 lysozyme. *Nature,* **330**, 41–46.

Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science,* **181**, 223–230.

Aurora, R., Srinivasan, R. & Rose, G. D. (1994). Rules for α-helix termination by glycine. *Science,* **264**, 1126–1130.

Aurora, R., Creamer, T. P., Srinivasan, R. & Rose, G. D. (1997). Local interactions in protein folding; Lessons from the α-helix. *J. Biol. Chem.* **272**, 1413–1416.

Baker, E. N. & Dodson, E. J. (1980). Crystallographic refinement of the structure of actinidin at 1.7 angstroms resolution by fast fourier last-squares methods. *Acta Crystallog. sect. A,* **36**, 559–572.

Baker, E. N. & Hubbard, R. E. (1984). Hydrogen bonding in globular proteins. *Prog. Biophys. Mol. Biol.* **44**, 97–179.

Baldwin, R. L. (1995). α-helix formation by peptides of defined sequence. *Biophys. Chem.* **55**, 127–135.

Barlow, D. J. & Thornton, J. M. (1988). Helix geometry in proteins. *J. Mol. Biol.* **201**, 601–619.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). The Protein Data Bank: a computer based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–542.

Betz, S. F., Raleigh, D. P. & DeGrado, W. F. (1993). *De novo* protein design: from molten globules to native-like states. *Curr. Opin. Struct. Biol.* **3**, 601–610.

Blaber, M. Z., Zhang, X. & Matthews, B. W. (1993). Structural basis of amino acid α-helix propensity. *Science,* **260**, 1637–1640.

Blundell, T. M., Barlow, D. J., Borkakoti, N. & Thornton, J. M. (1983). Solvent induced distortions and the curvature of α-helices. *Nature,* **306**, 281–283.

Brandl, C. J. & Deber, C. M. (1986). Hypothesis about the function of membrane-buried proline residues in transport proteins. *Proc. Natl Acad. Sci. USA,* **83**, 917–921.

Bruch, M. D., Dhingra, M. M. & Gierasch, L. M. (1991). Side chain backbone hydrogen bonding contributes to helix stability in peptides derived from an α-helical region of carboxypeptidase A. *Proteins: Struct. Funct. Genet.* **10**, 130–139.

Chakrabartty, A. & Baldwin, R. L. (1995). Stability of α-helices. *Advan. Protein Chem.* **46**, 141–176.

Chou, P. Y. & Fasman, G. D. (1974). Conformational parameters for amino acids in helical, β-sheet and random coil regions calculated from proteins. *Biochemistry,* **13**, 211–222.

Chou, P. Y. & Fasman, G. D. (1978). Empirical predictions of protein conformation. *Annu. Rev. Biochem.* **47**, 251–276.

Dasgupta, S. & Bell, J. A. (1993). Design of helix ends. Amino acid preferences, hydrogen bonding and electrostatic interactions. *Int. J. Peptide Protein Res.* **41**, 499–511.

Datta, S., Shamala, N., Banerjee, A., Pramanik, A., Bhattacharjya, S. & Balaram, P. (1997). Characterization of helix terminating Schellman motif in peptides. Crystal structure and nuclear Overhauser effect analysis of a synthetic heptapeptide helix. *J. Am. Chem. Soc.* **119**, 9246–9251.

DeGrado, W. F. (1988). Design of peptides and proteins. *Advan. Protein Chem.* **39**, 51–124.

Doig, A. J. & Baldwin, R. L. (1995). N and C capping preferences for all 20 amino acids in helical peptides. *Protein Sci.* **4**, 1325–1336.

Doig, A. J., MacArthur, M. W., Stapley, B. J. & Thornton, J. M. (1997). Structures of N termini of helices in proteins. *Protein Sci.* **6**, 147–155.

Gibrat, J.-F., Garnier, J. & Robson, B. (1987). Further developments of protein secondary structure prediction using information theory. *J. Mol. Biol.* **198**, 425–443.

Gunasekaran, K., Ramakrishnan, C. & Balaram, P. (1996). Ramachandran disallowed conformations of amino acid residues in proteins. *J. Mol. Biol.* **264**, 191–198.

Harper, E. T. & Rose, G. D. (1993). Helix stop signals in proteins and peptides: the capping box. *Biochemistry,* **32**, 7606–7609.

Horovitz, A., Serrano, L. & Fersht, A. R. (1991). COSMIC analysis of the major α-helix of barnase during folding. *J. Mol. Biol.* **219**, 5–9.

Horovitz, A., Matthews, J. M. & Fersht, A. R. (1992). α-Helix stability in proteins. II. Factors that influence stability at an internal position. *J. Mol. Biol.* **227**, 560–568.

Hurley, J. H., Mason, D. A. & Matthews, B. W. (1992). Flexible-geometry conformational energy maps for the amino acid residue preceding a proline. *Biopolymers,* **32**, 1443–1446.

Ishikawa, K., Nakamura, H., Morikawa, K., Kimura, S. & Kanaya, S. (1993). Cooperative stabilization of *Escherichia coli* ribonuclease HI by insertion of Gly80b and Gly-77 → Ala substitution. *Biochemistry,* **32**, 7136–7142.

Jimenez, M. A., Munoz, V., Rico, M. & Serrano, L. (1994). Helix stop and start signals in peptides and proteins. The capping box does not necessarily prevent helix elongation. *J. Mol. Biol.* **242**, 487–496.

Kametkar, S., Schiffer, J. M., Xiong, H., Babik, J. M. & Hecht, M. H. (1993). Protein design by binary patterning of polar and non-polar amino acids. *Science,* **262**, 1680–1685.

Karle, I. L., Flippen-Anderson, J. L., Uma, K. & Balaram, P. (1993). Peptide mimics for structural features in proteins. Crystal structures of three heptapeptide helices with a C-terminal 6 → 1 hydrogen bond. *Int. J. Peptide Protein Res.* **23**, 411–419.

Karle, I. L., Banerjee, A., Bhattacharjya, S. & Balaram, P. (1996). Solid state and solution conformations of a helical peptide with a central Gly-Gly segment. *Biopolymers,* **38**, 515–526.

Kim, P. S. & Baldwin, R. L. (1984). A helix-stop signal in the isolated peptide of ribonuclease A. *Nature,* **307**, 1080–1085.

Kraulis, P. J. (1991). MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallog.* **24**, 946–950.

Lee, B. & Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **55**, 379–400.

Liao, D.-I. & Herzberg, O. (1991). Structure of the IIA domain of the glucose permease of *Bacillus subtilis* at 2.2 Å resolution. *Biochemistry,* **30**, 9583–9594.

Lyu, P. C., Liff, M. I., Marky, L. A. & Kallenbach, N. R. (1990). Sidechain contributions to the stability of α-helical structures in peptides. *Science,* **250**, 669–673.

Lyu, P. C., Zhou, H. X., Jelveh, N., Wemmer, D. E. & Kallenbach, N. R. (1992). Position-dependent stabilizing effects in α-helices: N-terminal capping in synthetic model peptides. *J. Am. Chem. Soc.* **114**, 6560–6562.

MacArthur, M. W. & Thornton, J. M. (1991). Influence of proline residues on protein conformation. *J. Mol. Biol.* **218**, 397–412.

Milner-White, E. J. (1988). Recurring loop motif in proteins that occurs in right-handed and left-handed forms. Its relationship with alpha-helices and beta-bulge loops. *J. Mol. Biol.* **199**, 503–511.

Miner, D. L., Jr & Kim, P. S. (1994). Measurement of the β-sheet propensities of amino acids. *Nature,* **367**, 660–663.

Munoz, V. & Serrano, L. (1994). Intrinsic secondary structure propensities of the amino acids, using statistical ϕ-ψ matrices: comparison with experimental scales. *Proteins: Struct. Funct. Genet.* **20**, 301–311.

Nagarajaram, H. A., Sowdhamini, R., Ramakrishnan, C. & Balaram, P. (1993). Termination of right handed helices in proteins by residues in left handed helical conformations. *FEBS Letters,* **321**, 79–83.

Ohage, E. C., Graml, W., Walter, M. M., Steinbacher, S. & Steipe, B. (1997). β-Turn propensities as paradigms for the analysis of structural motifs to engineer protein stability. *Protein Sci.* **6**, 233–241.

O'Neil, K. T. & DeGrado, W. F. (1990). A thermodynamic scale for the helix forming tendencies of the commonly occuring amino acids. *Science,* **250**, 646–650.

Padmanabhan, S. & Baldwin, R. L. (1994). Tests for helix stabilizing interaction between various peptides. *Protein Sci.* **3**, 1992–1997.

Pai, E. F., Krengel, U., Petsko, G. A., Goody, R. S., Kabsch, W. & Wittinghofer, A. (1990). Refined crystal structure of the triphosphate conformation of h-ras p21 at 1. 35 Å resolution: implications for the mechanism of GTP hydrolysis. *EMBO J.* **9**, 2351–2359.

Pauling, L., Corey, R. B. & Branson, H. R. (1951). The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. *Proc. Natl Acad. Sci. USA,* **37**, 205–211 (1951).

Preissner, R. & Bork, P. (1991). On α-helices terminated by glycine: 1. Identification of common structural features. *Biochem. Biophys. Res. Commun.* **180**, 660–665.

Presta, L. G. & Rose, G. D. (1988). Helix signals in proteins. *Science,* **240**, 1632–1641.

Ramachandran, G. N. & Sasisekharan, V. (1968). Conformation of polypeptides and proteins. *Advan. Protein Chem.* **23**, 283–437.

Ramachandran, G. N., Ramakrishnan, C. & Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.* **7**, 95–99.

Ramakrishnan, C. & Ramachandran, G. N. (1965). Stereochemical criteria for polypeptide and protein chain conformations. II. Allowed conformations for a pair of peptide units. *Biophys. J.* **5**, 909–933.

Ramakrishnan, C. & Srinivasan, N. (1990). Glycyl residues in proteins and peptides: an analysis. *Curr. Sci.* **59**, 851–861.

Richardson, J. S. (1981). The anatomy and taxonomy of protein structure. *Advan. Protein Chem.* **34**, 167–330.

Richardson, J. S. & Richardson, D. C. (1988). Amino acid preferences for specific locations at the ends of α-helices. *Science,* **240**, 1648–1652.

Richardson, J. S., Richardson, D. C., Tweedy, N. B., Gernert, K. M., Quinn, T. P., Hecht, M. H., Erickson, B. W., Yan, Y., McClain, R. D., Dolan, M. E. & Surles, M. E. (1992). Looking at proteins: representations, folding, packing, and design. *Biophys. J.* **63**, 1186–1209.

Rooman, M. & Wodak, S. J. (1988). Identification of predictive sequence motifs limited by protein structure data base size. *Nature,* **335**, 45–49.

Rost, B. & Sander, C. (1994). Structure prediction of proteins: where are we now? *Curr. Opin. Biotechnol.* **5**, 372–380.

Sander, C. & Schneider, R. (1991). Database of homology-derived protein secondary structures and the structural meaning of sequence alignment. *Proteins: Struct. Funct. Genet.* **9**, 56–68.

Sankararamakrishnan, R. & Vishveshwara, S. (1992). Geometry of proline-containing α-helices. *Int. J. Peptide Protein Res.* **39**, 356–363.

Sansom, M. S. P. (1992). Proline residues in transmembrane helices of channel and transport proteins: a molecular modelling study. *Protein Eng.* **5**, 53–60.

Schellman, C. (1980). The $\alpha_L$-conformation at the ends of helices. In *Protein Folding* (Jaenicke, R., ed.), pp. 53–61, Elsevier/North-Holland, New York.

Scholtz, J. M. & Baldwin, R. L. (1992). The mechanism of α-helix formation by peptides. *Annu. Rev. Biophys. Biomol. Struct.* **21**, 95–118.

Seale, J. W., Srinivasan, R. & Rose, G. D. (1994). Sequence determinants of the capping-box, a stabilizing motif at the N-termini of α-helices. *Protein Sci.* **3**, 1741–1745.

Serrano, L. & Fersht, A. R. (1989). Capping and α-helix stability. *Nature,* **342**, 296–299.

Serrano, L., Sancho, J., Hirshberg, M. & Fersht, A. R. (1992). α-Helix stability in proteins. I. Empirical correlations concerning substitution of side-chains at the N and C-caps and the replacement of alanine by glycine or serine at solvent-exposed surfaces. *J. Mol. Biol.* **227**, 544–559.

Shortle, D., Stites, W. E. & Meeker, A. K. (1990). Contributions of the large hydrophobic amino acids to the stability of staphylococcal nuclease. *Biochemistry,* **29**, 8033–8041.

Sippl, M. J. (1995). Knowledge-based potentials for proteins. *Curr. Opin. Struct. Biol.* **5**, 229–235.

Smith, C. K. & Regan, L. (1995). Guidelines for protein design: the energetics of β-sheet side chain interactions. *Science,* **270**, 980–982.

Sowdhamini, R., Srinivasan, N., Ramakrishnan, C. & Balaram, P. (1992). Orthogonal ββ motifs in proteins. *J. Mol. Biol.* **223**, 845–851.

Strehlow, K. G., Robertson, A. D. & Baldwin, R. L. (1991). Proline for alanine substitutions in the C-

peptide helix of ribonuclease A. *Biochemistry,* **30**, 5810–5814.
Srinivasan, N., Anuradha, V. S., Ramakrishnan, C., Sowdhamini, R. & Balaram, P. (1994). Conformational characteristics of asparaginyl residues in proteins. *Int. J. Peptide Protein Res.* **44**, 112–122.
Swindells, M. B., MacArthur, M. W. & Thornton, J. M. (1995). Intrinsic $\phi$, $\psi$ propensities of amino acids, derived from the coil regions of known structures. *Nature Struct. Biol.* **2**, 596–603.
Thanki, N., Umrania, Y., Thornton, J. M. & Goodfellow, J. M. (1991). Analysis of protein main-chain solvation as a function of secondary structure. *J. Mol. Biol.* **221**, 669–691.
Venkatachalam, C. M. (1968). Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units. *Biopolymers,* **6**, 1425–1436.
Viguera, A. R. & Serrano, L. (1995). Experimental analysis of the Schellman motif. *J. Mol. Biol.* **251**, 150–160.
von Heijne, G. (1991). Proline kinks in transmembrane $\alpha$-helices. *J. Mol. Biol.* **218**, 499–503.
Woolfson, D. N., Mortishire-Smith, R. J. & Williams, D. H. (1991). Conserved positioning of proline residues in membrane-spanning helices of ion-channel proteins. *Biochem. Biophys. Res. Commun.* **175**, 733–737.
Xu, Z., Bernlohr, D. A. & Banaszak, L. J. (1993). The adipocyte lipid-binding protein at 1.6 Å resolution: crystal structures of the apoprotein and with bound saturated and unsaturated fatty acids. *J. Biol. Chem.* **268**, 7874–7884.
Yamamoto, A., Tomoo, K., Doi, M., Ohishi, H., Inoue, M., Ishida, T., Yamamoto, D., Tsuboi, S., Okamoto, H. & Okada, Y. (1992). Crystal structure of 2 papain-succinyl-gln-val-val-ala-ala-p-nitroanilide complex at 1.7 Å resolution: noncovalent binding mode of a common sequence of endogenous thiol protease inhibitors. *Biochemistry,* **31**, 11305–11309.
Zhou, H. X., Lyu, P. C., Wemmer, D. E. & Kallenbach, N. R. (1994). $\alpha$-Helix capping in synthetic model peptides by reciprocal side chain-main chain interactions: evidence for an N-terminal ''capping box''. *Proteins: Struct. Funct. Genet.* **18**, 1–7.

***Edited by J. Thornton***

*(Received 12 May 1997; received in revised form 23 October 1997; accepted 24 October 1997)*