*Protein Science* (1999), *8*:587–595. Cambridge University Press. Printed in the USA.


# A survey of left-handed polyproline II helices

BENJAMIN J. STAPLEY AND TREVOR P. CREAMER

Kentucky Center for Structural Biology, Department of Biochemistry, University of Kentucky, 800 Rose Street, Lexington, Kentucky 40536-0298

(RECEIVED May 18, 1998; ACCEPTED November 12, 1998)

**Abstract**

Left-handed polyproline II helices (PPII) are contiguous elements of protein secondary structure in which the $\phi$ and $\psi$ angles of constituent residues are restricted to around $-75°$ and 145°, respectively. They are important in structural proteins, in unfolded states and as ligands for signaling proteins. Here, we present a survey of 274 nonhomologous polypeptide chains from proteins of known structure for regions that form these structures. Such regions are rare, but the majority of proteins contain at least one PPII helix. Most PPII helices are shorter than five residues, although the longest found contained 12 amino acids. Proline predominates in PPII, but Gln and positively charged residues are also favored. The basis of Gln's prevalence is its ability to form an $i$, $i + 1$ side-chain to main-chain hydrogen bond with the backbone carbonyl oxygen of the proceeding residue; this helps to fix the $\psi$ angle of the Gln and the $\phi$ and $\psi$ of the proceeding residue in PPII conformations and explains why Gln is favored at the first position in a PPII helix. PPII helices are highly solvent exposed, which explains why apolar amino acids are disfavored despite preferring this region of $\phi/\psi$ space when in isolation. PPII helices have perfect threefold rotational symmetry and within these structures we find significant correlation between the hydrophobicity of residues at $i$ and $i + 3$; thus, PPII helices in globular proteins can be considered to be amphipathic.

**Keywords:** polyproline-II-helix; proline; proline-rich-region; protein structure

Proline-rich regions (PRRs) occur widely in both eukaryotes and prokaryotes. Such polypeptide segments are believed to adopt left-handed polyproline II helical conformations (PPII) similar to poly-L-proline and collagen (Fraser et al., 1979; Williamson, 1994). These regions have been classified into repetitive short PRRs that are important both structurally and in binding and signaling (Hannavy et al., 1990), tandemly repeated PRRs such as salivary PRRs that have polyphenol binding properties (Hagerman & Butler, 1981; Murray et al., 1994), and nonrepetitive PRRs that are involved in a large variety of cellular processes. Of these last phenomena, perhaps the best characterized is the regulation of src tyrosine kinase by the binding of a short PRR to a 60 residue beta-barrel domain (SH3 domain) (Pawson, 1995). SH3 domains have been identified as a modular component of many eukaroyote proteins (e.g., SOS, PI3K, and $p47^{phox}$), and a PXXP motif has been identified as a consensus binding sequence for these domains (Feng et al., 1994). Other proteins that bind ligands in PPII conformations are profilin (Mahoney et al., 1997), WW domains (Macias et al., 1996), and class II MHC proteins (Murthy & Stern, 1997).

Although SH3 domain ligands are predominantly proline, the SH2/kinase domain linker regions that complex the SH3 domains of Hck and c-src tyrosine kinases in their inactive forms are considerably less proline-rich (Sicheri et al., 1997; Xu et al., 1997). The ligands of class II MHC proteins can contain no proline residues and yet adopt PPII conformations (Jardetzky et al., 1996; Murthy & Stern, 1997). The physical basis of PPII formation in polypeptides is of physiological relevance but at present is unknown.

There is evidence for the presence of PPII conformations in nonproline containing polypeptides (Tiffany & Krimm, 1968; Woody, 1992). The unfolded state of hen egg white lysozyme shows some PPII helix by Raman spectroscopy (Wilson et al., 1996). In Ala-based peptides, which possess a strong tendency to form $\alpha$-helices, it has been suggested that the PPII conformation contributes significantly to the ensemble of unfolded states (Park et al., 1997).

The PPII helix (Fig. 1A) is characterized by having its $\phi$ and $\psi$ angles restricted to the regions $-78°$ and $+146°$, respectively (Cowan & McGavin, 1955); this imparts a perfect threefold rotational symmetry to the structure (Fig. 1B). PPII helices in globular proteins of known structure contain a large proportion of proline residues (Adzhubei & Sternberg, 1993). Proline is unique among naturally occurring amino acids; a proline residue in a polypeptide has its $\phi$ torsion angle restricted to $-63°$ ($\pm 15°$) (MacArthur & Thornton, 1991) and the $\delta$ carbon of the proline ring interacts with the preceding residue hindering that residue from adopting helical $\psi$ angles.

Here we present a survey of PPII helices from a set of 274 nonhomologous, high-resolution polypeptide chains from proteins of known structure (Bernstein et al., 1977; Hobohm & Sander,

Reprint requests to: Trevor P. Creamer, Kentucky Center for Structural Biology, Department of Biochemistry, University of Kentucky, 800 Rose Street, Lexington, Kentucky 40536-0298; e-mail: trevor@euripides.gws.uky.edu.

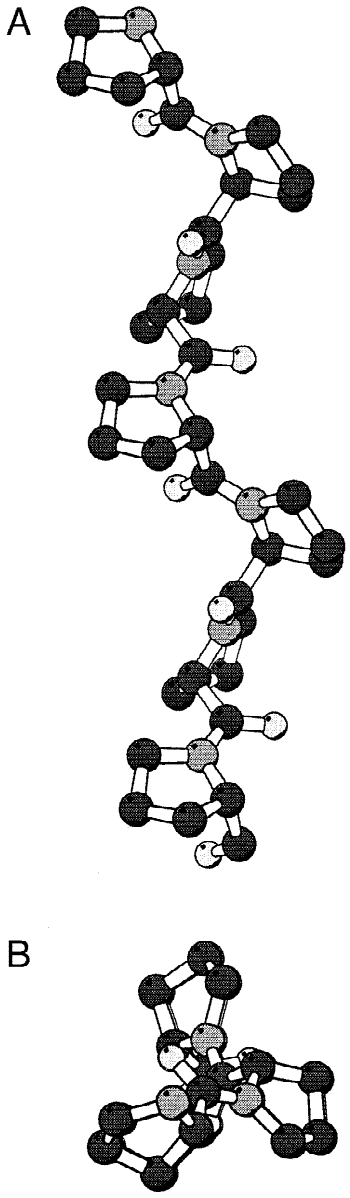


**Fig. 1.** The poly-L-proline II helix; $\phi = -78°$, $\psi = +146°$; generated with MolScript (Kraulis, 1992).

1994). Although PPII helices are not common (2% of residues), more than half the polypeptide chains in the dataset contain at least one region of PPII helix of length greater than three. Proline predominates in PPII helices and polar amino acids are favored over nonpolar (especially over aromatic residues). Within PPII helices, Gln and positively charged residues are strongly favored at the first position, while Leu is favored at central positions. Our amino acid propensities for PPII helices are poorly correlated with those for isolated random coil residues with similar $\phi$ and $\psi$ angles (Swindells et al., 1995). We believe the preference for long side-chain residues that contain hydrogen bond donors is caused by the ability of these residues to interact with the exposed backbone carbonyl oxygens of the proceeding residue of the PPII—contrawise for negatively charged residues. Rotamer preferences for residues within PPII helices are weak; however, aromatic side chains favor the *trans* rotamer while for others *gauche*$^+$ is preferred. As reported previously, our study confirms that PPII helices in proteins of known structure are generally shorter than six residues in length and highly solvent exposed (Adzhubei & Sternberg, 1993). The threefold symmetry of the PPII helix and its high surface exposure results in a correlation between the polarity of residues pairs spaced $i$, $i + 3$.

## Results and discussion

### *PPII distribution and prevalence*

The application of our criteria for assessing the presence of PPII helices to a dataset of 274 nonhomologous, high-resolution polypeptide structures produced 272 occurrences of helices longer than three residues. This represents 1,231 residues of the 62,504 in the dataset, which is 2%, similar to the estimate of Adzhubei and Sternberg (1993). Sreerama and Woody (1994) suggested about 10% of residues are in PPII conformations that give rise to characteristic CD spectra; however, they included single residues in their analysis. Ten percent of prolines in the proteins of our study are contained within PPII structures. Although PPII helices represent a small fraction of total residues, 144 of the 274 protein chains in our survey contain one or more helix. Figure 2 shows a histogram of the number of polypeptide chains that contain certain fractions of secondary structure. We consider four or more consecutive H or E Kabsch and Sander (1983) assignments constitute a helix and sheet, respectively. In no chain does PPII constitute more than 12% of the secondary structure. This contrasts with $\beta$-sheet and $\alpha$-helix for which approximately 1/3 and 1/2 of the chains contain more than 25% of the respective secondary structures. We found no relationship between $\alpha$-helical or $\beta$-sheet content and PPII content (data not shown).

### *PPII helix length and composition*

PPII helices are short—the majority are just four residues in length (Table 1); however, a very few contain more than nine residues. Our distribution of lengths is similar to that obtained by Adzhubei and Sternberg (1993). Not all PPII helices contain proline and few are entirely proline (Table 1). Slightly fewer helices than expected contain no proline (25%), while more than expected contain a single proline (50%) (see Materials and methods).

### *Amino acid propensities*

Table 2 shows the amino acid propensities for residues in PPII helices. Of the 1,231 residues observed, nearly one-quarter are proline. The only other significantly favored amino acid is Gln (as was noted by Adzhubei & Sternberg, 1993). Gly and apolar amino acids are disfavored; Gly probably because of its preference for turns and apolar amino acids presumably due to the high surface exposure of PPII helices (see below). Figure 3A shows the correlation between the propensities obtained by this study and those obtained by Adzhubei and Sternberg (excluding proline). The correlation coefficient is low ($r^2 = 0.3$) and this probably reflects the scarcity of data in the earlier study. A comparison of the propensities obtained in this work with those from a survey of all coil residues from a similar region of the Ramachandran plot (Fig. 3B) (Swindells et al., 1995) reveals that although hydrophobic amino

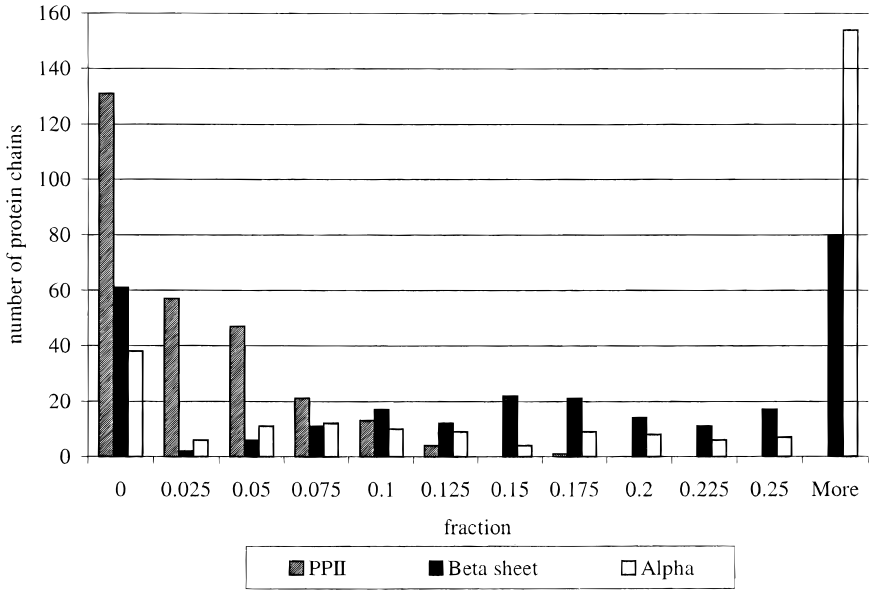


**Fig. 2.** Histogram of fractional contents of secondary structures. Hatched bars, polyproline II helix; solid bars, $\beta$-sheet (defined as four or more consecutive residues having Kabsch and Sander definition "E"); open bars, $\alpha$-helix (defined as four or more consecutive residues having Kabsch and Sander definition "H" (Kabsch & Sander, 1983)).

acids (Ala, Met, Cys, Leu, Trp, and Tyr) prefer these $\phi$ and $\psi$ angles, they are not prevalent in PPII helices (contrawise polar amino acids Gln, Arg, Lys, and Asp). The correlation coefficient (excluding proline) between these data is low ($r^2 = 0.1$) indicating that the PPII helix cannot be considered as simply a sequence of coil residues with similar $\phi/\psi$ angles.

A more detailed analysis of amino acid preferences within PPII helices is shown in Table 3. Here we have considered the residue preferences at each position in the helix and, although the data are sparse, some trends emerge. Proline is favored at all positions within the helix while Gly is highly unfavorable. This is in accord with CD studies of Gly/Pro co-polymers that show that the presence of Gly in polyproline peptides reduces the apparent PPII content (Petrella et al., 1996). The first residue in the structure is the most likely to be an amino acid other than Pro, with Gln and positively charged residues being most common. At central positions Leu is favored, but Ile disfavored. A possible reason for this is that the $\chi_1$ *gauche*$^-$ rotamer for Ile is not observed in PPII helices because the methyl groups on the $C_\beta$ effectively bury the amide backbone when the side chain adopts this conformation. By contrast, this rotamer is more favorable for Leu because both methyl groups are linked to the $C_\gamma$, thus allowing the peptide backbone to hydrogen bond with solvent or protein.

### *Side-chain rotamer preference*

As has been noted many times before, backbone dihedral angles influence the preferred side-chain $\chi_1$ and $\chi_2$ rotamers (McGregor et al., 1987; Ponder & Richards, 1987; Dunbrack & Karplus, 1994; Muñoz & Serrano, 1994; Swindells et al., 1995). These rotamers

**Table 1.** *Length distribution and proline content of PPII helices*

| | Length | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| Number of helices | 183 | 62 | 13 | 8 | 3 | 1 | 1 | 0 | 1 |
| Fraction | 0.67 | 0.23 | 0.05 | 0.03 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 |

| | Number of prolines | | | | |
|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 |
| Number of helices[a] | 68 (81) | 134 (111) | 53 (60) | 13 (17) | 4 (3) |
| Fraction | 0.25 | 0.49 | 0.19 | 0.05 | 0.01 |

[a]Numbers in parentheses are expected values from the fractional content of proline in PPII helices and are calculated by assuming that the probability of Pro occurring at one position in the helix is uncorrelated with the probability at a different position.

**Table 2.** *Amino acid propensities in polyproline II helices*

| Amino acid | Occurrences | | Propensity | $Pr(\chi^2)$ | Sampling probability[a] |
|---|---|---|---|---|---|
| | In PPII helices | In all structures | | | |
| Pro | 295 | 2,963 | 5.06 | 1E-210 | 0.00 |
| Gln | 56 | 2,284 | 1.24 | 0.1 | 0.06 |
| Arg | 60 | 2,806 | 1.09 | 0.5 | 0.28 |
| Lys | 77 | 3,700 | 1.06 | 0.6 | 0.33 |
| Thr | 76 | 3,870 | 1.00 | 1.0 | 0.47 |
| Leu | 93 | 5,197 | 0.91 | 0.4 | 0.15 |
| Asp | 64 | 3,680 | 0.88 | 0.3 | 0.14 |
| Met | 22 | 1,305 | 0.86 | 0.5 | 0.20 |
| Ala | 88 | 5,294 | 0.84 | 0.1 | 0.04 |
| Cys | 16 | 987 | 0.82 | 0.4 | 0.19 |
| Val | 69 | 4,277 | 0.82 | 0.1 | 0.03 |
| Glu | 60 | 3,756 | 0.81 | 0.1 | 0.04 |
| Asn | 42 | 2,995 | 0.71 | 0.03 | 0.01 |
| Phe | 35 | 2,556 | 0.70 | 0.03 | 0.01 |
| Ser | 53 | 3,868 | 0.70 | 0.008 | 0.00 |
| Ile | 46 | 3,374 | 0.69 | 0.01 | 0.00 |
| Trp | 11 | 911 | 0.61 | 0.1 | 0.03 |
| Tyr | 26 | 2,298 | 0.57 | 0.004 | 0.00 |
| His | 15 | 1,396 | 0.55 | 0.02 | 0.00 |
| Gly | 27 | 4,987 | 0.27 | 7E-13 | 0.00 |
| Total | 1,231 | 62,504 | | | |

[a]The probability, given the fraction of amino acids of that type occurring in "all" structures, that a random sample gives rise to the fraction observed in PPII helices. See Materials and methods.

**Table 3.** *Position dependent propensities for polyproline II helices*[a]

| | Preceding | Position within helix | | | | | Proceeding |
|---|---|---|---|---|---|---|---|
| | | First | Second | Middle | Penultimate | Final | |
| Ala | 0.7 (15) | 0.87 (20) | 0.78 (18) | 0.74 (9) | 1.09 (25) | 0.69 (16) | 0.82 (19) |
| Leu | 0.8 (17) | 0.8 (18) | 1.11 (25) | 1.35 (16) | 1.11 (25) | 0.4 (9) | 0.66 (15) |
| Ile | 0.8 (11) | 1.09 (16) | 0.89 (13) | 0.39 (3) | 0.68 (10) | 0.27 (4) | 0.95 (14) |
| Pro | 1.2 (15) | 1.47 (19) | 5.35 (69) | 8.26 (56) | 6.05 (78) | 5.66 (73) | 0.54 (7) |
| Gly | 1.8 (38) | 0.55 (12) | 0.09 (2) | 0.09 (1) | 0.32 (7) | 0.23 (5) | 1.43 (31) |
| Phe | 0.9 (10) | 0.9 (10) | 0.9 (10) | 0.68 (4) | 0.45 (5) | 0.54 (6) | 1.53 (17) |
| Met | 1.6 (9) | 1.23 (7) | 0.7 (4) | 0 (0) | 1.06 (6) | 0.88 (5) | 0.53 (3) |
| Cys | 0.5 (2) | 0.7 (3) | 1.63 (7) | 0 (0) | 1.16 (5) | 0.23 (1) | 0.47 (2) |
| Trp | 1.0 (4) | 0.25 (1) | 0.25 (1) | 1.92 (4) | 0.76 (3) | 0.5 (2) | 0.5 (2) |
| Tyr | 0.8 (8) | 0.4 (4) | 0.5 (5) | 0.76 (4) | 0.6 (6) | 0.7 (7) | 1.1 (11) |
| His | 1.0 (6) | 0.66 (4) | 0.49 (3) | 0 (0) | 0.66 (4) | 0.66 (4) | 1.65 (10) |
| Val | 0.7 (13) | 1.18 (22) | 0.75 (14) | 0.41 (4) | 0.91 (17) | 0.64 (12) | 0.81 (15) |
| Asn | 1.3 (17) | 0.84 (11) | 0.46 (6) | 0.44 (3) | 0.77 (10) | 0.92 (12) | 1.38 (18) |
| Gln | 1.0 (10) | 2.41 (24) | 1.01 (10) | 0.96 (5) | 0.8 (8) | 0.91 (9) | 1.01 (10) |
| Asp | 0.9 (14) | 0.37 (6) | 1.12 (18) | 0.48 (4) | 0.75 (12) | 1.5 (24) | 1.06 (17) |
| Glu | 0.9 (15) | 0.86 (14) | 0.98 (16) | 0.58 (5) | 0.67 (11) | 0.86 (14) | 0.98 (16) |
| Ser | 1.2 (20) | 0.89 (15) | 0.48 (8) | 0.57 (5) | 0.42 (7) | 1.07 (18) | 0.89 (15) |
| Thr | 1.0 (16) | 1.37 (23) | 1.13 (19) | 0.56 (5) | 0.42 (7) | 1.31 (22) | 1.48 (25) |
| Lys | 1.2 (19) | 1.37 (22) | 0.75 (12) | 1.06 (9) | 0.93 (15) | 1.18 (19) | 0.99 (16) |
| Arg | 0.8 (10) | 1.72 (21) | 0.98 (12) | 0.93 (6) | 0.9 (11) | 0.82 (10) | 0.74 (9) |
| Total | 269 | 272 | 272 | 143 | 272 | 272 | 272 |
| Aromatic | 0.89 (22) | 0.60 (15) | 0.64 (16) | 0.91 (12) | 0.56 (14) | 0.60 (15) | 1.20 (30) |
| Aliphatic | 0.76 (67) | 0.97 (86) | 0.91 (81) | 0.68 (32) | 0.99 (88) | 0.53 (47) | 0.76 (68) |
| Charged | 0.97 (58) | 1.04 (63) | 0.96 (58) | 0.75 (24) | 0.81 (49) | 1.10 (67) | 0.96 (58) |
| Polar (neutral) | 1.12 (63) | 1.29 (73) | 0.76 (43) | 0.60 (18) | 0.56 (32) | 1.08 (61) | 1.20 (68) |

[a]Numbers in parentheses are the number of occurrences.

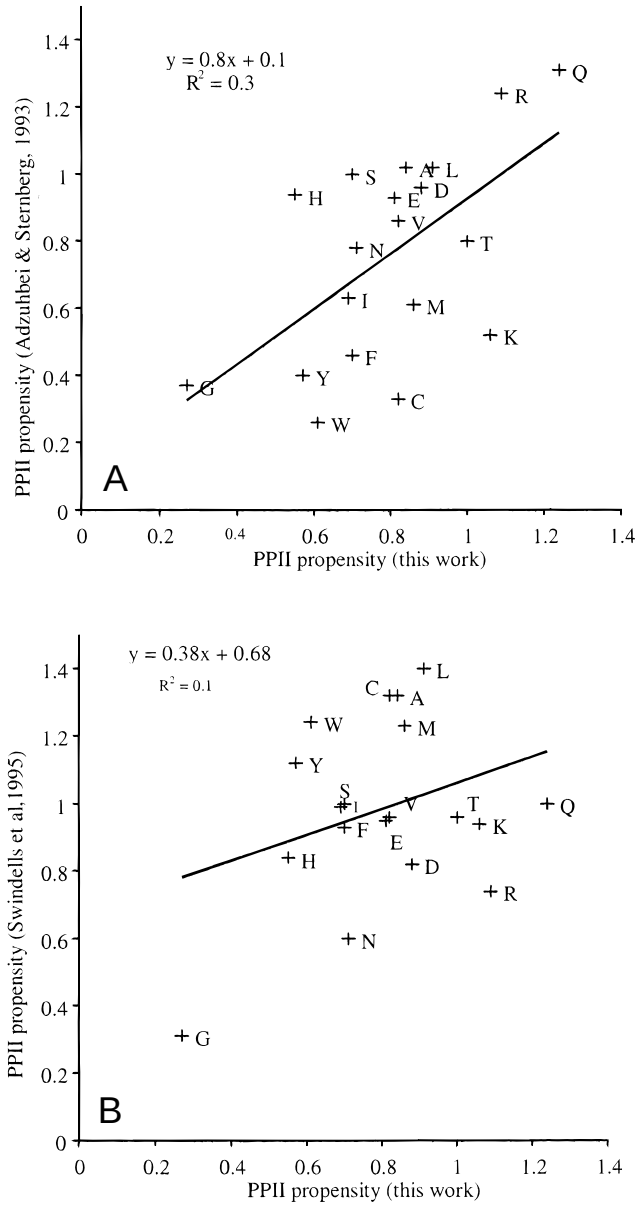


**Fig. 3.** Comparison of PPII amino acid propensities from this study with (**A**) that of Adzhubei and Sternberg (1993) and (**B**) that of Swindells et al. (1995).

are classed *trans* ($-120° > \chi_1 > 120°$), *gauche*$^+$ ($0° > \chi_1 > -120°$) and *gauche*$^-$ ($120° > \chi_1 > 0°$) (IUPAC-IUB, 1970). Table 4 shows the results of our survey of side-chain rotamers in PPII helices. The PPII region of $\phi$ space generally precludes the *gauche*$^-$ rotamer through clashes with the backbone amide hydrogen and perhaps its hydrogen bond acceptor (Dunbrack & Karplus, 1994). The extended nature of the PPII helix means that contacts between side chains are uncommon and that there is very little effect on rotamer preference from neighboring side chains; however, aromatic side chains favor the *trans* rotamer while for others *gauche*$^+$ is preferred. Our side-chain rotamer preferences are similar to those obtained from a survey of all random coil residues (including non-PPII structures) from a equivalent region of $\phi/\psi$ space (Swindells et al., 1995). The qualitative agreement in side-chain rotamer preferences between these two surveys supports the hypothesis that the distribution of side-chain rotamers for residues with these $\phi$ and $\psi$ angles is a result of the steric effect of adjacent peptide bonds rather than any interactions with neighboring amino acid side chains.

**Table 4.** *Side-chain $\chi_1$ rotamer propensities for polyproline II helices*[a]

| | *gauche*$^-$ | *trans* | *gauche*$^+$ |
|---|---|---|---|
| Leu | 0.0 (0) | 1.0 (32) | 1.1 (61) |
| Ile | 1.1 (8) | 0.7 (4) | 1.0 (34) |
| Phe | 0.9 (4) | 0.8 (10) | 1.1 (21) |
| Met | 0.9 (2) | 0.6 (4) | 1.2 (16) |
| Cys | 0.7 (2) | 0.7 (3) | 1.3 (11) |
| Trp | 1.2 (2) | 1.0 (4) | 0.9 (5) |
| Tyr | 0.0 (0) | 1.4 (13) | 1.0 (14) |
| His | 0.6 (1) | 1.4 (6) | 1.0 (7) |
| Val | 0.9 (7) | 0.9 (44) | 1.4 (20) |
| Asn | 1.1 (8) | 1.2 (16) | 0.8 (18) |
| Gln | 0.6 (3) | 1.0 (19) | 1.1 (34) |
| Asp | 0.7 (8) | 0.6 (12) | 1.4 (42) |
| Glu | 1.2 (8) | 0.9 (17) | 1.0 (34) |
| Ser | 1.0 (23) | 1.2 (16) | 0.8 (16) |
| Thr | 1.0 (33) | 1.4 (11) | 0.9 (32) |
| Lys | 0.7 (5) | 0.6 (15) | 1.3 (54) |
| Arg | 1.4 (9) | 0.7 (14) | 1.1 (37) |
| Total | 0.9 (123) | 0.9 (240) | 1.1 (456) |

[a] Numbers in parentheses are the number of occurrences.

### *Hydrogen bonding within polyproline II helices*

To assess the extent to which hydrogen bonding contributes to the stability of PPII helices, we determined the presence/absence of hydrogen bonding between residues within PPII helices. The extended nature of the PPII helix precludes main-chain–main-chain hydrogen bonds, but interaction between side chains and backbone oxygens or nonprolyl amides are sterically possible. Table 5 shows how frequently the side chains of Gln, Asn, Arg, Lys, Ser, Thr, and His donate hydrogen bonds to backbone carbonyl oxygens within PPII helices. Eleven percent of Glns within PPII helices are involved in this kind of interaction and the predominant topology is a hydrogen bond between the Gln side chain NE2 and the carbonyl of the proceeding residue. This *i*, *i* + 1 interaction may help to explain why Gln predominates at the first position in the PPII helix. Figure 4 shows a superimposition of four cases of PPII helices where Gln NE2 hydrogen bonds to the peptide oxygen of the proceeding residue. They are from rhizomucor miehei triacylglycerol lipase (3tgl), residues 176–179; a clostridium molybdenum-iron protein (1mio) chain C, residues 204–207; bacteriophage PHIX174 capsid protein (2bpa) chain 2, residues 3–74; bovine pancreatic carboxypeptidase A (2ctc), residues 211–214. The Gln side chains are all in similar conformations with the $\chi_1$ angles *trans*. The creation of this pattern of hydrogen bonding helps fix the $\psi$ angle of the Gln within the range of the PPII structures. The side chain of Asn is less capable of forming such interactions; however, Arg and Lys can participate in this hydrogen bonding network and are also favored at the first position in a PPII helix.

**Table 5.** *Side-chain donor–main-chain acceptor hydrogen bonds within polyproline II helices*

| | $i-3$ | $i-2$ | $i-1$ | $i$ | $i+1$ | $i+2$ | Total | Fraction of residues |
|---|---|---|---|---|---|---|---|---|
| Gln | 0 | 2 | 0 | 0 | 4 | 0 | 6 | 0.11 |
| Asn | 0 | 0 | 0 | 1 | 2 | 0 | 3 | 0.07 |
| Lys | 0 | 1 | 0 | 0 | 2 | 0 | 3 | 0.04 |
| Arg | 1 | 0 | 0 | 0 | 2 | 0 | 3 | 0.05 |
| Thr | 0 | 0 | 0 | 0 | 2 | 0 | 2 | 0.03 |
| Ser | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.00 |
| His | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0.07 |
| Total | 1 | 4 | 2 | 1 | 14 | 0 | 22 | 0.07 |

It is interesting to contrast the N-termini of PPII helices with those of $\alpha$-helices. It has been noted that the first position in an $\alpha$-helix is often a Ser, Thr, Asn, or Asp (Richardson & Richardson, 1988) and that the reason for is the ability of these residues to cap the helix by hydrogen bonding to the exposed amide protons and the terminus of the helix (Presta & Rose, 1988; Doig et al., 1997). By contrast, Gln and Glu are a poor N-caps and this has been attributed to the fact that the presence of an extra methylene group over Asn/Asp prevents effective hydrogen bonding to the backbone amides at positions N2 or N3 (Doig & Baldwin, 1995). In the case of PPII helices, the situation is reversed; Gln is favored over Asn because it has a sufficiently long side chain so as to interact with the proceeding peptide carbonyl group. Interestingly, many proline rich regions are also rich in Gln (Takagi et al., 1984; Laurent et al., 1990) and it has been suggested that Gln rich regions share similar properties to Pro rich regions in forming linkers between domains and inducing oligomerization.

### *Solvent accessibilities*

Adzhubei and Sternberg (1993) reported that PPII helices were more surface exposed than other structures. We concur with their findings as Table 6 demonstrates. Residues within PPII helices expose 60% more polar surface area and 50% more nonpolar surface area to solvent than do "average" residues. All amino acid residues have greater than average nonpolar solvent exposure and there is a weak but significant negative correlation between average residual nonpolar accessible surface area and the logarithm of propensity (Fig. 5). We use this quantity because if the partitioning of residues between the PPII state and other states is Boltzmann-like then the free energy difference between the "average" state of a residue and that of PPII helix would be inversely proportional to the logarithm of the propensity of that residue. This quantity appears to be positively correlated with the nonpolar ASA, which is generally accepted to be proportional to solvation free energy (Sharp et al., 1991). It appears that nonpolar amino acids are disfavored in PPII helices because such structures are generally highly solvent exposed.

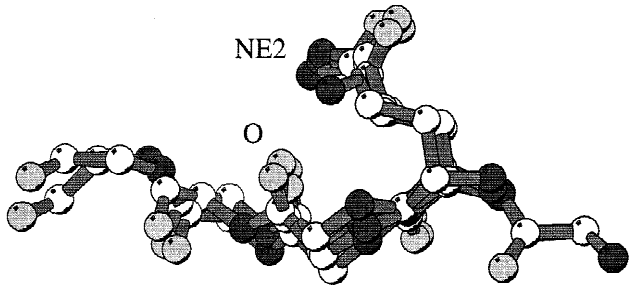


**Fig. 4.** Overlay of four structures from PPII helices in which a Gln side chain forms a hydrogen bond to the backbone carbonyl oxygen of the preceding residue. They are from rhizomucor miehei triacylglycerol lipase (3tgl), residues 176–179; a clostridium molybdenum-iron protein (1mio) chain C, residues 204–207; bacteriophage PHIX174 capsid protein (2bpa) chain 2, residues 3–74; bovine pancreatic carboxypeptidase A (2ctc), residues 211–214. Generated with MolScript (Kraulis, 1992).

**Table 6.** *Solvent accessibilities of amino acid residues in polyproline II helices in Å²*

| | All structures | | PPII helices | | % change | |
|---|---|---|---|---|---|---|
| | Nonpolar | Polar | Nonpolar | Polar | Nonpolar | Polar |
| Ile | 18.2 | 3.4 | 43.4 | 11.9 | 138 | 253 |
| Phe | 22.2 | 3.9 | 48.7 | 11.0 | 119 | 182 |
| Val | 17.8 | 3.7 | 37.8 | 8.2 | 113 | 118 |
| Leu | 19.4 | 4.0 | 40.1 | 10.9 | 106 | 174 |
| Tyr | 23.8 | 14.9 | 46.8 | 27.8 | 97 | 87 |
| Trp | 24.9 | 8.0 | 42.4 | 15.3 | 70 | 91 |
| Met | 24.7 | 4.5 | 41.4 | 8.8 | 68 | 98 |
| Ser | 20.0 | 18.5 | 32.7 | 26.8 | 63 | 45 |
| Thr | 26.6 | 15.0 | 42.3 | 23.9 | 59 | 59 |
| Ala | 18.9 | 6.3 | 29.7 | 12.0 | 57 | 91 |
| His | 29.5 | 19.0 | 45.9 | 30.8 | 56 | 63 |
| Asp | 20.3 | 36.7 | 28.5 | 46.2 | 40 | 26 |
| Asn | 16.1 | 39.5 | 21.2 | 47.3 | 32 | 20 |
| Cys | 8.5 | 4.8 | 10.9 | 11.9 | 28 | 150 |
| Glu | 31.3 | 43.1 | 39.7 | 52.4 | 27 | 22 |
| Lys | 58.4 | 33.9 | 71.4 | 33.3 | 22 | −2 |
| Pro | 42.2 | 6.1 | 50.6 | 6.5 | 20 | 7 |
| Arg | 29.4 | 51.9 | 33.3 | 48.9 | 13 | −6 |
| Gln | 23.7 | 42.9 | 24.9 | 47.5 | 5 | 11 |
| Gly | 14.8 | 9.5 | 15.4 | 12.3 | 4 | 29 |
| Mean | | | | | 49.9 | 59.0 |

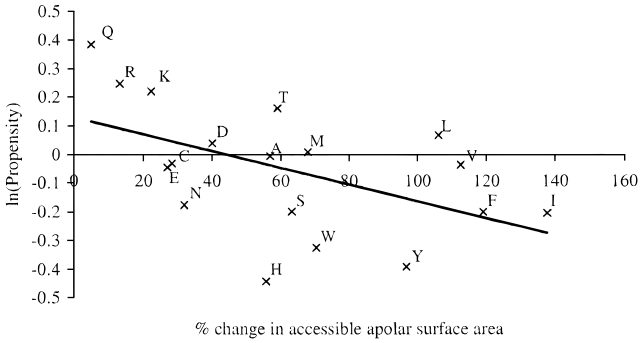


**Fig. 5.** Relationship between amino acid propensity and nonpolar accessible surface area in polyproline II helices. The values on the *x*-axis are the percentage differences between the average nonpolar exposed surface area for a residue in a PPII and that of the residue averaged over the entire dataset. (Glycine and proline excluded.)

### *Amino acid pairwise preferences*

Since PPII helices are generally near the surface of proteins and the structure possesses perfect threefold symmetry, one would expect residues spaced *i*, *i* + 3 to show a similarity in their hydrophobicity. Table 7 lists the observed and expected numbers of side-chain pairs—we exclude proline since its hydropathic properties are hard to categorize. Significantly more hydrophobic pairs of amino acids are spaced *i*, *i* + 3 than would be expected if the probability of finding the individual residue at positions *i* and *i* + 3 were uncorrelated. We can therefore conclude that PPII helices in globular proteins are amphipathic.

### *Long PPII helices*

Although we find PPII helices to be short, there are several cases of proteins with helices longer than seven residues. The longest helix found was 12 residues from toward the C-terminus of soybean $\beta$-amylase (Mikami et al., 1993). The last 40 residues of this protein consists of two PPII helices of length 12 and 8 residues interrupted by a short $\alpha$-helix. A C-terminal truncation of the homologous protein from barley reduced the thermostability of the enzyme (Yoshigi et al., 1995). The enzyme lignin peroxidase contains a cluster of PPII helices at its C-terminus with 50% of the last 54 residues being in this conformation (Poulos et al., 1993). These residues lie across the active site of the enzyme. Their role in regulating the activity of this enzyme, if any, is uncertain at present.

**Table 7.** *Pairwise i, i + 3 amino acid preferences in polyproline II helices*[a]

| *i* | *i* + 3 | Observed | Expected | Chi-squared |
|---|---|---|---|---|
| Phobic | Phobic | 76 | 42 | 26 |
| Phillic | Phillic | 119 | 87 | 11 |
| Phobic | Phillic | 76 | 67 | 1 |
| Phillic | Phobic | 79 | 55 | 10 |

[a]Excluding proline.

### *Summary*

In this work we have observed and rationalized some of the influences that predispose PPII helix formation. A high content of proline, a tendency to contain Gln and positively charged residues, high solvent exposure, and a hydrophobic periodicity of three all appear to be important. Although PPII helices are rare in proteins of known structure, proline rich regions are common in nature; this may indicate that such structures are not readily amenable to traditional methods of structure determination, possibly due to their inherent flexibility. The PPII conformation is significant in the unfolded states of proteins (Woody, 1992; Park et al., 1997) and in molecular recognition (Williamson, 1994). A better understanding of the forces that govern PPII formation might throw light on the nature of the unfolded state and would aid in the identification of possible ligands for proteins that bind polypeptides in such conformations. To this end, accurate quantification of the extent of PPII formation in peptides might be useful.

## Materials and methods

The database is a set of 274 nonhomologous polypeptide chains from high resolution protein structures take from the Brookhaven Protein Data Base (Bernstein et al., 1977; Hobohm & Sander, 1994). The Brookhaven four letter ID codes followed by the chain identifier are listed below: 129l, 1aaj, 1aak, 1aapA, 1aba, 2abk, 1abmA, 1add, 1ads, 1aep, 1alkA, 1aozA, 1apa, 1apmE, 1arb, 1atnA, 1avdA, 1avhA, 2ayh, 2baa, 1babB, 1bbpA, 1bbt1, 1bbt2, 1bgeB, 1bgh, 1bllE, 1bovA, 1bsaA, 1btc, 1caj, 1cauB, 1cbn, 1ccr, 1cde, 1cdtA, 1cgt, 1chrA, 1cid, 1cmbA, 1cobA, 1colA, 1cpcA, 1cpcL, 1cpt, 1crl, 1cseI, 1d66A, 1dfnA, 1dhr, 1dog, 2dri, 1dsbA, 1eaf, 1eco, 1ede, 2end, 1ezm, 1faiL, 1fas, 1fbaA, 1fc1A, 1fc2C, 1fdd, 1fha, 1fiaB, 1fnb, 1fod4, 1fxiA, 1gal, 1gd1O, 1gdhA, 1gky, 1glaF, 1glaG, 2glt, 1gmfA, 1gof, 1gox, 1gpb, 1gsrA, 1hbq, 1hc6, 1hddC, 1hdxA, 1hgeB, 1hivA, 1hleA, 1hleB, 1hmy, 1huw, 1ifc, 1ipd, 1isuA, 1le4, 1lenA, 1lgaA, 1lis, 1ltsD, 1ltsA, 1ltsC, 1mdaA, 1mdc, 1mfbH, 1mgn, 1minB, 1mioC, 2ms2A, 1mup, 1mypC, 1nar, 1ndk, 1nipB, 1nxb, 1ofv, 2omf, 1omp, 1onc, 1osa, 1pda, 1pdgB, 1pfkA, 2pgd, 1phh, 1plc, 1poa, 1poc, 1poxA, 1ppbL, 1ppfE, 1ppn, 1ppt, 1prcC, 1prcM, 1pyp, 1r094, 1r1a2, 1rcb, 1rec, 1rhd, 1ribA, 1rinB, 1rnd, 1rveA, 1s01, 1sbp, 1sgt, 1shaA, 1shfA, 2sim, 1sltA, 1smrA, 1snc, 1spa, 1sryA, 1tabI, 1tbpA, 1ten, 1tgsI, 1tie, 1tlk, 1tml, 1tnfA, 1tplA, 1trb, 1troA, 1ttbA, 1ula, 1utg, 2vaaB, 1vsgA, 1wsyA, 1wsyB, 1zaaC, 2aaiB, 2achB, 2atcB, 2azaA, 2bbkH, 2bopA, 2bpa1, 2bpa2, 2bpa3, 2cas, 2ccyA, 2cdv, 2cmd, 2cp4, 2cpl, 2cro, 2ctc, 2cts, 2cyp, 2dnjA, 2er7E, 2hipA, 2hpdA, 2ihl, 2lh2, 2liv, 2madL, 2mev1, 2mhr, 2mnr, 2msbA, 2mtaC, 2pf1, 2pia , 2plv1, 2plv3, 3pmgA, 2por, 2reb, 2rn2, 2sas, 2scpA, 2sga, 2sn3, 2snv, 2spo, 2stv, 2tbvA, 2tgi, 2tmdA, 2tmvP, 2tscA, 2ztaA, 3adk, 3b5c, 3cd4, 3chy, 3cla, 3cox, 3dfr, 3gapA, 3gbp, 3grs, 3inkC, 3monA, 3pgk, 3pgm, 3rubS, 3sdhA, 3sgbI, 3tgl, 4blmA, 4cpaI, 4enl, 4fgf, 4fxn, 4gcr, 4htcI, 4insB, 4rcrH, 4sbvA, 4sgbI, 4ts1A, 4xis, 5fbpA, 5nn9, 5p21, 5timA, 6taa, 7apiB, 8abp, 8acn, 8atcA, 8catA, 8i1b, 8rxnA, 9ldtA, 9rnt, 9rubB, 9wgaA.

Using the program SSTRUCT (S. Hubbard, pers. comm.), $\phi$, $\psi$, $\omega$, $\chi_1$ angles were calculated for all residues. Secondary structure was assigned using the criteria of Kabsch and Sander (1983).

Our assessment of PPII regions is as follows: first, PPII helices must be at least four residues in length. This requires the fixation of three $\phi$ and three $\psi$ angles (the $\phi$ angles of residues $i + 1$, $i + 2$, $i + 1$ and the $\psi$ of $i$, $i + 1$, $i + 2$, where $i$ is the first residue in the

helix). Second, the mean $\phi$ and $\psi$ of the helix must between $-55°$ and $-95°$ and between 125° and 165°, respectively ($-55° \leq \bar{\phi} \leq -95°$, $125° \leq \bar{\psi} \leq 165°$). None of these angles can lie more than 20° from the mean value for the helix and the standard deviation of either $\phi$ or $\psi$ cannot be greater than 20°($|\phi_i - \bar{\phi}| \leq 20°$ and $\sum_i^N[(\phi_i - \bar{\phi})^2/N] \leq 400°$). Because the PPII helix is close in phase space to a $\beta$-sheet or strand, none of the residues in the helix are allowed Kabsch and Sander secondary structural assignments B, E, or H (Kabsch & Sander, 1983); this is the fourth criterion. Lastly, we do not allow *cis* peptide bonds to occur in PPII helices.

The propensity of a given residue for a particular attribute is defined as the ratio of the actual number of observations to the expected number of observations, where the expected number of observations is given by $n^{\dagger}(n_{\mathrm{Xaa}}/n_{\mathrm{total}})$, where $n_{\mathrm{Xaa}}$ is the total number of residue type Xaa in the dataset, $n_{\mathrm{total}}$ is the total number of all residues in the dataset, and $n^{\dagger}$ is the total number of residues in the dataset with the attribute of interest.

To assess the significance of observed counts from expected we used two methods. First, we applied a standard $\chi^2$ test to the data where $\chi^2 = (E - O)^2/E$, where $E$ and $O$ are the expected and observed values, respectively. We also apply a more probabilistic approach; given the frequency of occurrence of an amino acid residue in the dataset $(n_{\mathrm{Xaa}}/n_{\mathrm{total}})$, we enumerate the probability that, given that the frequency of the amino acid in the structure of interest is the same as that in the dataset as a whole, a random sample of $n^*$ residues would give rise to a number of counts of the amino acid that is equal to or further from the expected value. In this way we are assessing the probability that the frequency of the amino acid in the structure of interest is different from the dataset as a whole. This probability is given by

$$P = \sum_{i=0}^{O} \frac{n^*!}{i!(n^* - i)!} \left(\frac{n_{\mathrm{Xaa}}}{n_{\mathrm{total}}}\right)^i \left(1 - \frac{n_{\mathrm{Xaa}}}{n_{\mathrm{total}}}\right)^{O-i}$$

for $E > O$ and

$$P = 1 - \sum_{i=0}^{O} \frac{n^*!}{i!(n^* - i)!} \left(\frac{n_{\mathrm{Xaa}}}{n_{\mathrm{total}}}\right)^i \left(1 - \frac{n_{\mathrm{Xaa}}}{n_{\mathrm{total}}}\right)^{O-i}$$

for $E < O$.

For the assessment of the proline content of PPII helices (Table 1), expected values of occurrences are calculated from the fractional proline content using the assumption that the probability of proline occurring at one position is independent from its occurrence elsewhere in the helix. Thus the expected number of PPII helices containing $n$ prolines $\mathrm{C}(n)$ is given by a binomial distribution.

$$C(n) = \sum_{i}^{N} \frac{L_i!}{n!(L_i - n)!} (f_{PRO})^n (1 - f_{PRO})^{L_i - n}$$

where $N$ is the total number of PPII helices (272), $L_i$ is the length of helix $i$, and $f_{PRO}$ is the fractional content of proline in PPII helices. The summation is over all helices and the elements of the summation are simply the fraction or probability that each helix; $i$ of length $L_i$ has exactly $n$ prolines.

Side-chain rotamer $\chi_1$ angles were defined accord to the IUPAC-IUB convention (IUPAC-IUB, 1970) with *gauche*$^+$ and *gauche*$^-$ centered at $-60°$ ($\pm 60°$) and $+60°$ ($\pm 60°$), respectively. The *trans* rotamer is assigned for $\chi_1 > 120°$ or $\chi_1 < -120°$.

Residue accessibilities were calculated using NACCESS v2.1 (Hubbard & Thornton, 1993) using a probe radius of 1.4 Å. For comparison, we calculated the average accessible surface area for all amino acids in our dataset.

The program HBplus (McDonald & Thornton, 1994) was used to assess the hydrogen bonding properties of PPII helices. For the side chains of Gln and Asn, for which the side chains are difficult to resolve crystallographically with certainty, we allowed the relative positions of the side-chain nitrogen and oxygen to be exchanged; thus we are including potential hydrogen bonds. The maximum allowed separation between donor and acceptor atoms was 3.9 Å and the maximum donor-H-acceptor angle was 90°.

## Acknowledgments

We thank Brett Johnson for helpful discussions. Funded in part by the University of Kentucky Medical Center Research Fund, an American Cancer Society Institutional Research Grant award and by a Kentucky-NSF EPSCOR award to the Kentucky Center for Structural Biology. BJS is supported in part by a University of Kentucky Center for Computational Sciences Postdoctoral Appointment.

## References

Adzhubei AA, Sternberg MJE. 1993. Left-handed polyproline II helices commonly occur in globular proteins. *J Mol Biol 229*:472–493.

Bernstein FC, Koetzle TF, Williams G, Meyer DJ, Brice MD, Rodgers JR, 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.

Cowan PM, McGavin S. 1955. Structure of poly-L-proline. *Nature 176*:470–478.

Doig AJ, Baldwin RL. 1995. N- and C-capping preferences for all 20 amino acids in alpha-helical peptides. *Protein Sci 4*:1325–1336.

Doig AJ, MacArthur MW, Stapley BJ, Thornton JM. 1997. Structures of N-termini of helices in proteins. *Protein Sci 6*:147–155.

Dunbrack RL Jr, Karplus M. 1994. Conformational analysis of the backbone-dependent rotamer preferences of protein side chains. *Nature Struct Biol 1*:334–340.

Feng S, Chen JK, Yu H, Simon JA, Schreiber SL. 1994. Two binding orientations for peptides to the Src SH3 domain: Development of a general model for SH3-ligand interactions. *Science 266*:1241–1247.

Fraser RDB, MacRae TP, Suzuki E. 1979. Chain conformation in the collagen molecule. *J Mol Biol 129*:463–481.

Hagerman AE, Butler LG. 1981. The specificity of proanthocyanidin-protein interactions. *J Biol Chem 256*:4494–4497.

Hannavy K, Barr GC, Dorman CJ, Adamson J, Mazengera LR, Gallagher MP, Evans JS, Levine BA, Trayer IP, Higgins CF. 1990. TonB protein of *Salmonella typhimurium*. A model for signal transduction between membranes. *J Mol Biol 216*:897–910.

Hobohm U, Sander C. 1994. Enlarged representative set of protein structures. *Protein Sci 3*:522–524.

Hubbard SJ, Thornton JM. 1993. "NACCESS," computer program. London: Department of Biochemistry and Molecular Biology, University College London.

IUPAC-IUB. 1970. IUPAC-IUB Commission on Biochemical Nomenclature. Abbreviations and symbols for the description of the conformation of polypeptide chains. *J Mol Biol 52*:1–17.

Jardetzky TS, Brown JH, Gorga JC, Stern LJ, Urban RG, Strominger JL, Wiley DC. 1996. Crystallographic analysis of endogenous peptides associated with HLA-DR1 suggests a common, polyproline II-like conformation for bound peptides. *Proc Natl Acad Sci USA 93*:734–738.

Kabsch W, Sander C. 1983. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers 22*:2577–2637.

Kraulis PJ. 1992. MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. *J Applied Crystallography 24*:946–950.

Laurent BC, Gram M, Pongs O. 1990. The SNF5 protein of *Saccharomyces cerevisiae* is a glutamine- and proline-rich transcriptional activator that affects expression of a broad spectrum of genes. *Mol Cell Biol 10*:5616–5625.

MacArthur MW, Thornton JM. 1991. Influence of proline residues on protein conformation. *J Mol Biol 218*:397–412.

Macias MJ, Hyvönen M, Baraldi E, Schultz J, Sudol M, Saraste M, Oschkinat

H. 1996. Structure of the WW domain of a kinase-associated protein complexed with a proline-rich peptide. *Nature 382*:646–649.
Mahoney NM, Janmey PA, Almo SC. 1997. Structure of the profilin-poly-L-proline complex involved in morphogenesis and cytoskeletal regulation. *Nature Struct Biol 4*:953–960.
McDonald IK, Thornton JM. 1994. Satisfying hydrogen bonding potential in proteins. *J Mol Biol 238*:777–793.
McGregor MJ, Islam SA, Sternberg MJE. 1987. Analysis of the relationship between side-chain conformation and secondary structure in globular proteins. *J Mol Biol 198:*295–310.
Mikami B, Hehre EJ, Sato M, Katsube Y, Hirose M, Morita Y, Sacchettini JC. 1993. The 2.0-Å resolution structure of soybean beta-amylase complexed with alpha-cyclodextrin. *Biochemistry 32*:6836–6845.
Muñoz V, Serrano L. 1994. Intrinsic secondary structure propensities of the amino acids, using statistical phi-psi matrices: Comparison with experimental scales. *Proteins Struct Funct Genet 20*:301–311.
Murray NJ, Williamson MP, Lilley TH, Haslam E. 1994. Study of the interaction between salivary proline-rich proteins and a polyphenol by 1H-NMR spectroscopy. *Eur J Biochem 219*:923–935.
Murthy VL, Stern LJ. 1997. The class II MHC protein HLA-DR1 in complex with an endogenous peptide: Implications for the structural basis of the specificity of peptide binding. *Current Biol 5*:1385–1396.
Park S-H, Shalongo W, Stellwagen E. 1997. The role of PPII conformations in the calculation of peptide fractional helix content. *Protein Sci 6*:1694–1700.
Pawson T. 1995. Protein modules and signaling networks. *Nature 373*:573–580.
Petrella EC, Machesky LM, Kaiser DA, Pollard TD. 1996. Structural requirements and thermodynamics of the interaction of proline peptides with profilin. *Biochemistry 35:*16535–16543.
Ponder JW, Richards FM. 1987. Tertiary templates for proteins—Use of packing criteria in the enumeration of allowed sequences for different structural classes. *J Mol Biol 193*:775–791.
Poulos TL, Edwards SL, Wariishi H, Gold MH. 1993. Crystallographic refinement of lignin peroxidase at 2 Å. *J Biol Chem 268*:4429–4440.
Presta LG, Rose GD. 1988. Helix signals in proteins. *Science 240*:1632–1641.
Richardson JS, Richardson DC. 1988. Amino acid preferences for specific locations at the ends of $\alpha$-helices. *Science 240*:1648–1652.
Sharp KA, Nicholls A, Friedman R, Honig B. 1991. Extracting hydrophobic free energies from experimental data: Relationship to protein folding and theoretical models. *Biochemistry 30*:9686–9697
Sicheri F, Moarefi I, Kuriyan J. 1997. Crystal structure of the src family tyrosine kinase Hck. *Nature 385*:602–609.
Sreerama N, Woody RW. 1994. Poly(pro)II helices in globular proteins: Identification and circular dichroic analysis. *Biochemistry 33*:10022–10025.
Swindells MB, MacArthur MW, Thornton JM. 1995. Intrinsic phi, psi propensities of amino acids, derived from the coil regions of known structures. *Nature Struct Biol 2*:596–603.
Takagi T, Suzuki M, Baba T, Minegishi K, Sasaki S. 1984. Complete amino acid sequence of amelogenin in developing bovine enamel. *Biochem Biophys Res Commun 121*:592–597.
Tiffany ML, Krimm S. 1968. New chain conformations of poly(glutamic acid) and polylysine. *Biopolymers 6*:1379–1382.
Williamson MP. 1994. The structure and function of proline-rich regions in proteins. *Biochem J 297*:249–260.
Wilson G, Hecht L, Barron LD. 1996. Residual structure of unfolded proteins revealed by Raman optical activity. *Biochemistry 35*:12518–12525.
Woody RW. 1992. Circular dichroism and conformation of unordered polypeptides. *Advan Biophys Chem 2*:37–79.
Xu W, Harrison SC, Eck MJ. 1997. Three-dimensional structure of the tyrosine kinase c-src. *Nature 385*:595–602.
Yoshigi N, Sahara H, Koshino S. 1995. Role of the C-terminal region of beta-amylase from barley. *J Biochem 117*:63–67.