FOR THE RECORD

On the occurrence of linear groups in proteins

Scott A. Hollingsworth, Donald S. Berkholz, and P. Andrew Karplus*

Department of Biochemistry and Biophysics, Oregon State University, Corvallis, Oregon 97331-7305

Received 19 March 2009; Revised 6 April 2009; Accepted 9 April 2009
DOI: 10.1002/pro.133
Published online 16 April 2009 proteinscience.org

Abstract: Linear groups—polypeptide conformations based on a single repeating ϕ,ψ-pair—are a foundational concept in protein structure, yet how they are presented in textbooks is based largely on theoretical studies from the early days of protein structure analysis. Now, ultra-high resolution protein structures provide a resource for an accurate empirical and systematic assessment of the linear groups that truly exist in proteins. Here, a purely conformation-based survey of linear groups shows that only three distinct ϕ,ψ-regions occur: a diverse set of extended conformations mostly present as β-strands, a broad population of polyproline-II-like spirals, and a tight cluster that includes the highly populated α-helix and the conformationally-similar but much less populated 3_{10}-helix. Rare, short left-handed α-/3_{10}-helical turns with repeating ϕ,ψ-angles occur, but none are longer than three residues. Misperceptions dispelled by this study are the existence of 2.2_7- and π-helices as linear groups, the existence of specific ideal ϕ,ψ-angles for each linear group, and the existence of a substantive difference in the ϕ,ψ-preferences for parallel versus antiparallel β-strands. This study provides a concrete basis for updating and enhancing how we think about and teach the basics of protein structure.

Keywords: Ramachandran plot; linear group; α-helix; β-sheet; polyproline; left-handed helix; protein standard conformation; secondary structure

Statement for Broader Audience

The fundamentals of protein conformation are currently presented in textbooks and taught based on out-of-date information that misses the true nature of the conformations that build proteins. This work asks one simple question that provides key information and insights needed for revising our understanding and teaching of these basics.

In 1950, based on crystal structures of five small-molecule peptides, Corey and Donohue[1] published the first set of standard bond lengths and angles for guiding polypeptide modeling. Pauling and coworkers then used this geometry to search for linear groups, defined as a series of residues all having identical conformations[†], that could enter regular hydrogen-bonded interactions: they predicted the γ- and α-helices[2] and the β-pleated sheet.[3] Subsequently, additional linear groups were predicted including the π-helix,[4] the 2.2_7-ribbon and the 3_{10}-helix.[5] Later, structural studies led to the recognition of the poly-L-proline II (P_{II})/

Additional Supporting Information may be found in the online version of this article.

Grant sponsor: NIH; Grant number: GM083136; Grant sponsor: Howard Hughes Medical Institute; Grant number: 52005883.

*Correspondence to: P. Andrew Karplus, Department of Biochemistry and Biophysics, Oregon State University, Corvallis, OR 97331. E-mail: karplusp@science.oregonstate.edu

[†]Technically, linear groups may also be called helices, but as the term helix is now commonly used to refer to any spiral-like structure, we use the less ambiguous term here.

polyglycine II/collagen-like helix as a linear group that does not undergo regular hydrogen bonding.[6–8] Although some predicted linear groups, such as Pauling's γ-helix,[2] were never seen in proteins and are no longer referred to, a subset of linear groups compiled in 1970,[9] with a set of associated ϕ,ψ-angles, are often presented in contemporary biochemistry[10–13] and structural biology texts[14,15] as standard protein conformations (see Fig. 1).

Although much has been learned about structural motifs in proteins,[16,17] we were surprised to find no single systematic survey in the literature directly addressing the occurrence of linear groups in proteins that updates the information summarized by IUPAC in 1970. What is needed is a simple, direct analysis that, independent of hydrogen bonding patterns, assesses which single conformations are seen to repeat in real proteins. Here, we present such an analysis.

The 1.2 Å Resolution Data Set

The primary data set was created by a search of a Protein Geometry Database (DSB, PAK, unpublished) for three-residue segments in protein structures determined at ≤1.2 Å resolution and which according to the PDBSelect March 2006 release[18] had ≤25% sequence identity with any other included structure. Each residue furthermore was required to have an average backbone B-factor ≤25 Å^2 and a trans peptide bond (|ω| > 140°). The search resulted in 30,692 segments from 209 protein chains. Considering just the central residue of each segment, all amino acid types were well-represented (numbers of occurrences ranging from 2,710 for Ala to 472 for Trp) and the ϕ,ψ-angles [Fig. 2(A)] were distributed as expected based on previous analyses using strict selection criteria.[20–22] In terms of regular hydrogen-bonded secondary structures, 10,028 (33%), 8,756 (29%), 1,200 (4%), and 9 (0.03%) residues were classified by DSSP[23] as α-helix (α), β-sheet (β), 3_{10}-helix, and π-helix, respectively.

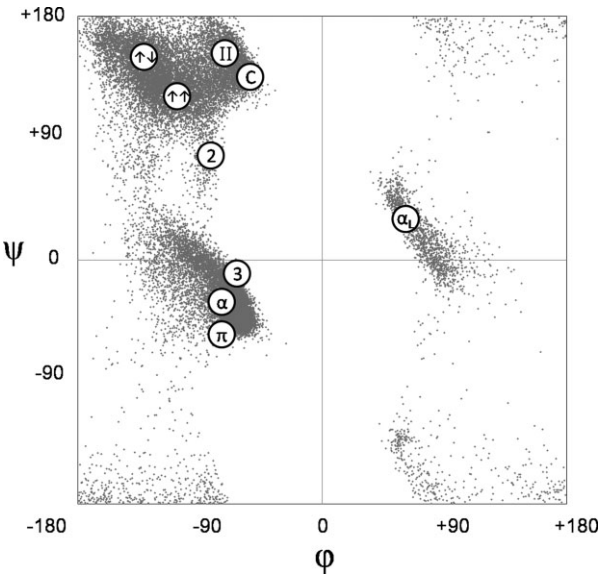


Figure 1. Linear groups as commonly cited in textbooks. Linear groups identified on the Ramachandran plot include the α-helix (α), 3_{10}-helix (3), π-helix (π), left-handed α-helix (α_L), 2.2_7-ribbon (2), polyproline-II (II), collagen (C), parallel β-sheet (↑↑), and anti-parallel β-sheet (↑↓). For reference, the background is a scatter plot of the 30,692 central residues used in this study. The figure is based on corresponding figures in textbooks cited in the text.[10–15]

Identification of Linear Groups

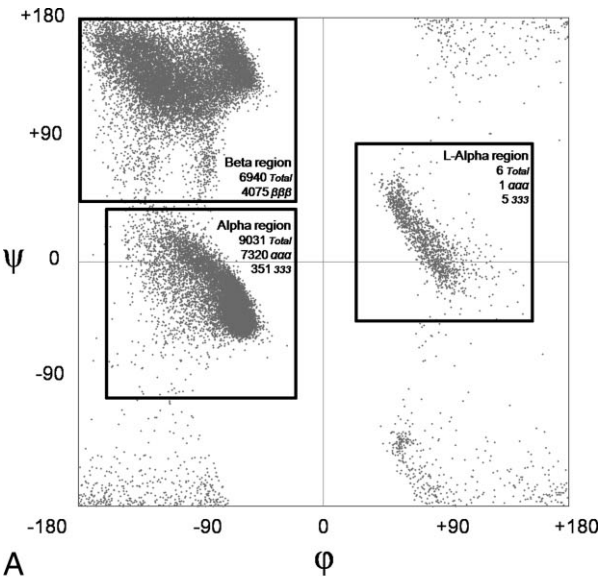
Identifying linear groups requires selecting criteria for the minimal number of residues and the ϕ,ψ-variation allowed: the longer the minimal length and the narrower the allowed ϕ,ψ-variation, the fewer groups will be found but the more truly linear those groups will be. A minimal length of one is meaningless because it would simply define every observed conformation as a linear group. In addition, two ϕ,ψ-pairs are considered to define turn types rather than "regular" secondary structures.[24] While three residue segments also make up many turn structures,[25] we settled on three residues as the shortest (i.e. least stringent) reasonable length requirement, acknowledging that some of the groups satisfying these criteria will not actually represent true linear groups that occur at longer lengths.

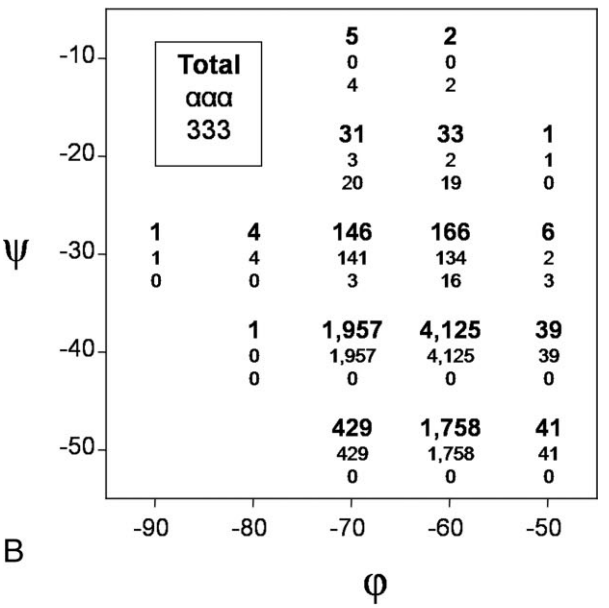
In terms of ϕ,ψ-variation allowed, we first carried out a low-stringency study, considering each residue to simply belong to one of the three main well-populated regions of the Ramachandran plot often referred to as the alpha, beta, and L-alpha regions [Fig. 2(A)]. Segments with all three residues residing in the alpha, beta, or L-alpha regions accounted for approximately 30%, 23%, and 0.02% of all residues respectively, leaving over 45% of the segments adopting more conformationally diverse structures. Whereas no hydrogen-bonding criteria were used in the search, in all cases, the majority of qualifying segments were part of regularly hydrogen-bonded β-strands and right-handed α- or 3_{10}-helices [Fig. 2(A)]. The six segments qualifying in the L-alpha region included four 3-residue segments and one 4-residue segment, leading us to conclude that whereas they passed this minimal filter, they do not occur in segments long enough to be considered true linear groups (Supporting Information Fig. S1).

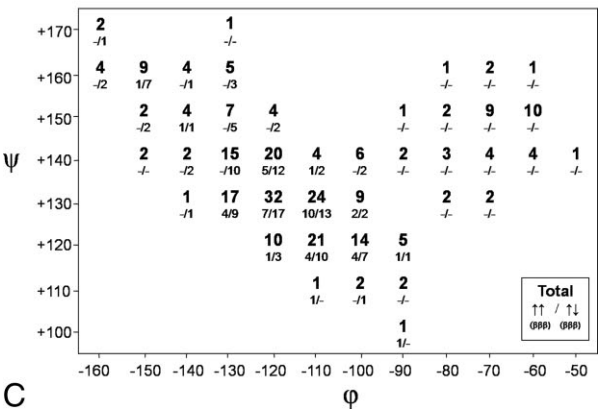
At a more discriminatory and informative level, we assessed linear groups as three consecutive residues all having the same ϕ,ψ-angles within ±10° by mapping ϕ,ψ-space in 20° × 20° boxes at 10° intervals. There were substantial numbers of observations in the alpha and beta regions [Figs. 2(B,C)]. Qualitatively equivalent results were obtained searching 30° × 30° boxes, showing that the results are not highly sensitive to the choice of box size (Supporting Information Fig. S2).

The alpha region shows a rather tight distribution of qualifying linear groups, with a single 20° × 20°

box (centered at φ,ψ = −60, −40) having more than 4,000 observations [Fig. 2(B)]. In terms of hydrogen-bonding patterns, all occurrences at lower ψ-values were α-helical with 3_{10}- and mixed α-/3_{10}-helices occurring at the higher ψ-values.







For the beta region, this more stringent mapping reveals a natural division into two populations: a broad elliptical grouping with nearly all observations having β-sheet hydrogen-bonding patterns and a smaller but also well-dispersed grouping around the classical P_{II} conformation that is nearly devoid of regular β-sheet hydrogen bonding [Fig. 2(C)]. The narrow bridge between the two regions is centered near φ,ψ = −95,+140. Despite the standard depiction of distinct ideal values for anti-parallel and parallel β-strands (see Fig. 1), both types of strands are seen to have their highest density of observations surrounding φ,ψ = −115, +130, and both occur throughout the elliptical β-sheet forming region. In agreement with Nagano,[26] it appears that the parallel strands are somewhat less diverse in conformation [Fig. 2(C), Supporting Information Figure S3].

Discussion

This survey shows that considering conformational properties alone, only three distinct linear groups comprise the protein-building toolkit. The most populated is a right-handed helical conformation dominated by the α-helix but also including the 3_{10}-helix and mixed α/3_{10} forms. The second is a diverse group of extended conformations that are dominated by residues

Figure 2. Linear groups in the 1.2 Å resolution dataset. (A) Ramachandran scatter plot of the 30,692 central residues qualifying for this study. Boxes outline the three major regions searched for occurrences of three contiguous residues all falling within the box. Each box is labeled (alpha, beta, and L-alpha) and the total number of qualifying segments in that box is given along with the subset of the results that fall completely within an α-helix (ααα), 3_{10}-helix (333), or β-strand (βββ) according to DSSP. Note that DSSP does not distinguish between left- and right-handed helices in its nomenclature. (B) Results for the fine search of the α-region. Each 20° × 20° box includes angles ±10° from its central value. As specified in the legend in the figure, within each box the top number reports the total number of segments having three consecutive residues in that box, and two numbers below this are how many of these are fully involved in purely α-helical (ααα) and purely 3_{10}-helical (333) H-bonding. Boxes outside of the displayed regions had zero observations. (C) Same as B but for the beta region. As noted in the legend in the figure given in each box are total occurrences (top), and the numbers of fully β-strand (βββ), residues involved in parallel and antiparallel strands. Mixed and outer strands are not included here, but can be seen in Supporting Information Figure S3. Assignments of β-strand orientation (parallel, anti-parallel, mixed or outer) were done using PDBsum.[19] A parallel stand was defined as a strand that was H-bonded on both sides to parallel stands, an anti-parallel stand as H-bonded on both sides to anti-parallel strands, a mixed strand as H-bonded to both a parallel and an anti-parallel strand, and an outer strand as one that was H-bonded to only one strand. Supporting Information Figure S2 gives plots similar to Figures 2(B,C) but based on 30° × 30° boxes.

occurring in β-strands in both parallel and antiparallel β-sheets. The third is a set of left-handed spiral conformations in the P_{II} area that generally are not a part of β-sheets. Although these results may not surprise many structural biologists, they illustrate some basic features of protein structure that are not generally appreciated or incorporated into current curricula.

A first point worth emphasizing is that beyond these three clusters, *no other true linear building blocks of proteins exist*. Textbooks can now be clear that although isolated residues and rare short segments may exist with the 2.2_7-ribbon, the π-helical, and the left-handed α- or 3_{10}-helical conformations, those conformations do not occur in extended segments having repeating φ,ψ-angles. This conclusion is in apparent contradiction to reports describing π-helices in proteins as uncommon but real.[27,28] The resolution to this apparent contradiction is that the reported π-helices were defined not by conformation, but simply by the presence of two or more consecutive i+5,i hydrogen-bonds. They do not satisfy the definition of true linear groups as they are formed by a series of residues conformations varying by up to 60° in both φ and ψ (see Fig. 2 of Fodje and Al-Karadaghi[28]); also, the large majority are short with only two hydrogen-bonds.

A related inference is that the 3_{10}-helices identified by hydrogen-bonding are mostly not built from a narrowly repeating conformation. Considering just the central residue of each three-residue segment (see above), there are 12% (1200/10,028) as many residues in 3_{10}-helices as in α-helices. In contrast, for the three-residue segments having consistent φ,ψ-angles, the ratio is only 0.8% [67/8,623 derived from the sums of the occurrences in Fig. 2(B)]. This means over 90% of residues identified as 3_{10}-helical in proteins are involved in turn-like conformations rather than true linear groups. This fits with the observation that 3_{10}-helices are mostly short and associated with the beginnings or ends of α-helices.[29]

A second main point is that the β-strand and P_{II} represent fully distinct linear groups [Fig. 2(B)]. The P_{II} conformation was overlooked for many years in proteins because it is not defined by hydrogen bonding. It came to be recognized as an important element of folded proteins in the 1990s[30], occurring in lengths up to 12 residues,[31] and since then has become recognized as a significant conformation in unfolded peptides and proteins.[32–34] It has been noted that P_{II} is a confusing and unfortunate designation, since the conformation is not just associated with Pro but can be adopted by all amino acids. In the study here, about one-third of the residues in the center of P_{II} tripeptides are Pro; the rest include all types of amino acids. Perhaps the common name could be changed to a more general "polypeptide-II" conformation. This would maintain the familiar P_{II} acronym, avoid the misleading association with only Pro, and be consistent with the observation that it is a prominent conformation in unfolded polypeptide chains. Similarly, we suggest that the region of the Ramachandran plot broadly referred to as the β-region [Fig. 2(A)], be renamed the β/P_{II}-region so the nomenclature used lays a foundation for proper recognition of both of the two main contributing conformations.

A third important point is that the β-strand and P_{II} populations are both very spread out and cannot be well-characterized by a single φ,ψ-conformation. In contrast, the α-/3_{10}-helix group is more tightly clustered. The tighter clustering of the φ,ψ-preferences for the α-/3_{10}-helices compared to the b-strand and P_{II} spirals is consistent with different hydrogen-bonding constraints: the α-/3_{10}-helices are constrained by the need to satisfy local hydrogen bonding, whereas both β-strands and P_{II} spirals exhibit variety as they access many relatively isoenergetic conformations to optimize tertiary hydrogen-bonding interactions with other parts of the protein and solvent. Given these spreads, it is misleading to give a single φ,ψ as the "ideal" β or P_{II} value. Also misleading is the assignment of distinct ideal values for parallel and anti-parallel β-strands (see Fig. 1) that are here shown to exhibit no large difference in preferred value [Fig. 2(C)]. Instead, we suggest a simple figure illustrating the full range of ideal values for each type of linear group would better convey the reality that a range of values can all be

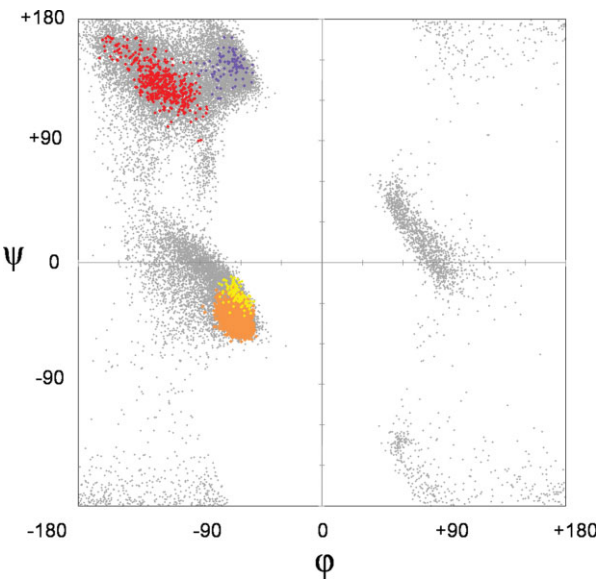


Figure 3. Locations and breadths of common linear groups in proteins. The composite scatter plot shows all triplets of residues found in the fine (20° × 20°) searches. Separate colors indicate the α (orange), 3_{10} (yellow), β (red), and P_{II} (purple) residues. For reference, a background plot (gray) shows all the residues in the 1.2 Å data set. The α and 3_{10} regions overlap. The most densely populated centers of each region are at α = (−63, −43), 3_{10} = (−62, −22), β = (−116,129), P_{II} = (−65, 145). Supporting Information Figure S4 is similar, but based on 90,211 residues from diverse crystal structures at 1.75 Å resolution or better.

considered normal (Fig. 3, and Supporting Information Fig. S4).

Acknowledgments

The authors thank Kevin Ahern for discussions and encouragement. They also thank all the crystallographers who have graciously deposited their coordinate sets in the PDB.

References

1. Corey R, Donohue J (1950) Interatomic distances and bond angles in the polypeptide chain of proteins. Proc Natl Acad Sci USA 72:2899–2900.
2. Pauling L, Corey RB, Branson HR (1951) The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. Proc Natl Acad Sci USA 37: 205–211.
3. Pauling L, Corey RB (1951) The pleated sheet, a new layer configuration of polypeptide chains. Proc Natl Acad Sci USA 37:251–256.
4. Low B, Baybutt R (1952) The pi-helix—a hydrogen bonded configuration of the polypeptide chain. J Am Chem Soc 74: 5806–5807.
5. Donohue J (1953) Hydrogen bonded helical configurations of the polypeptide chain. Proc Natl Acad Sci USA 39:470–478.
6. Arnott S, Dover SD (1968) The structure of poly-L-proline II. Acta Crystallogr B 24:599–601.
7. Sasisekharan V (1959) Structure of poly-L-proline II. Acta Crystallogr 12:897–903.
8. Ramachandran GN, Ramakrishnan C, Sasisekharan V (1963) Stereochemistry of polypeptide chain configurations. J Mol Biol 7:95–99.
9. IUPAC (1970) IUPAC-IUB Commission on biochemical nomenclature. Abbreviations and symbols for the description of the conformation of polypeptide chains. Biochemistry 9:3471–3479.
10. Garrett R, Grisham CM (2005) Biochemistry, 3rd ed. Belmont, CA: Thomson Brooks/Cole.
11. Mathews CK, Van Holde KE, Ahern KG (2000) Biochemistry, 3rd ed. San Francisco, California: Benjamin Cummings.
12. Voet D, Voet JG (2004) Biochemistry, 3rd ed. Hoboken, NJ: Wiley.
13. Lehninger AL, Nelson DL, Cox MM (2008) Lehninger principles of biochemistry, 5th ed. New York: W.H. Freeman.
14. Lesk AM (2001) Introduction to protein architecture: the structural biology of proteins. Oxford: Oxford University Press.
15. Van Holde KE, Johnson WC, Ho PS (2006) Principles of physical biochemistry, 2nd ed. Upper Saddle River, NJ: Pearson/Prentice Hall.
16. Beck DA, Alonso DO, Inoyama D, Daggett V (2008) The intrinsic conformational propensities of the 20 naturally occurring amino acids and reflection of these propensities in proteins. Proc Natl Acad Sci USA 105:12259–12264.
17. Perskie LL, Street TO, Rose GD (2008) Structures, basins, and energies: A deconstruction of the protein coil library. Protein Sci 17:1151–1161.
18. Hobohm U, Sander C (1994) Enlarged representative set of protein structures. Protein Sci 3:522–524.
19. Laskowski RA, Chistyakov VV, Thornton JM (2005) PDBsum more: new summaries and analyses of the known 3D structures of proteins and nucleic acids. Nucleic Acids Res 33:D266–D268.
20. Karplus PA (1996) Experimentally observed conformation-dependent geometry and hidden strain in proteins. Protein Sci 5:1406–1420.
21. Kleywegt GJ, Jones TA (1996) Phi/psi-chology: Ramachandran revisited. Structure 4:1395–1400.
22. Lovell SC, Davis IW, Arendall WB, III, de Bakker PI, Word JM, Prisant MG, Richardson JS, Richardson DC (2003) Structure validation by Cα geometry: phi, psi, and Cβ deviation. Proteins 50:437–450.
23. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637.
24. Venkatachalam CM (1968) Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units. Biopolymers 6:1425–1436.
25. Street TO, Fitzkee NC, Perskie LL, Rose GD (2007) Physical-chemical determinants of turn conformations in globular proteins. Protein Sci 16:1720–1727.
26. Nagano K (1977) Logical analysis of the mechanism of protein folding. IV. Super-secondary structures. J Mol Biol 109:235–250.
27. Weaver TM (2000) The pi-helix translates structure into function. Protein Sci 9:201–206.
28. Fodje MN, Al-Karadaghi S (2002) Occurrence, conformational features and amino acid propensities for the pi-helix. Protein Eng 15:353–358.
29. Richardson JS (1981) The anatomy and taxonomy of protein structure. Adv Protein Chem 34:167–339.
30. Adzhubei AA, Sternberg MJ (1993) Left-handed polyproline II helices commonly occur in globular proteins. J Mol Biol 229:472–493.
31. Stapley BJ, Creamer TP (1999) A survey of left-handed polyproline II helices. Protein Sci 8:587–595.
32. Shi Z, Woody RW, Kallenbach NR (2002) Is polyproline II a major backbone conformation in unfolded proteins?Adv Protein Chem 62:163–240.
33. Creamer TP, Campbell MN (2002) Determinants of the polyproline II helix from modeling studies. Adv Protein Chem 62:263–282.
34. Whittington SJ, Chellgren BW, Hermann VM, Creamer TP (2005) Urea promotes polyproline II helix formation: implications for protein denatured states. Biochemistry 44:6269–6275.