DEFINING α-HELIX GEOMETRY BY C^{α} ATOM TRACE *vs* (φ-ψ) TORSION ANGLES: A COMPARATIVE ANALYSIS

ASHISH SHELAR[*], PRASUN KUMAR[*] AND MANJU BANSAL

Molecular Biophysics Unit, Indian Institute of Science
Bangalore, 560 012, India

A regular secondary structure is described by a well defined set of values for the backbone dihedral angles (φ, ψ and ω) in a polypeptide chain. However in real protein structures small local variations give rise to distortions from the ideal structures, which can lead to considerable variation in higher order organization. Protein structure analysis and accurate assignment of various structural elements, especially their terminii, are important first step in protein structure prediction and design. Various algorithms are available for assigning secondary structure elements in proteins but some lacunae still exist. In this study, results of a recently developed in-house program ASSP have been compared with those from STRIDE, in identification of α-helical regions in both globular and membrane proteins. It is found that, while a combination of hydrogen bond patterns and backbone torsional angles (φ-ψ) are generally used to define secondary structure elements, the geometry of the C^{α} atom trace by itself is sufficient to define the parameters of helical structures in proteins. It is also possible to differentiate the various helical structures by their C^{α} trace and identify the deviations occurring both at mid-positions as well as at the terminii of α-helices, which often lead to occurrence of 3_{10} and π-helical fragments in both globular and membrane proteins.

1. Introduction

The most general description of 'helix' is a smooth 3D curve that lies on a conical or cylindrical surface. Helices which constitute the major secondary structure elements in proteins take up this structure due to repeating values of the backbone torsion angles (φ, ψ and ω) accompanied by regular hydrogen bond patterns between the backbone NH and CO groups of amino acids. Numerous secondary structure assignment algorithms which use atomic coordinate data have been developed and can be broadly classified into three categories: (i) algorithms based on backbone torsion angles and hydrogen bond patterns (ii) algorithms based on 3D geometry (iii) hybrid methods, which use both (i) and (ii). Programs like DSSP[1], STRIDE[2] and PROSS[3] fall into the first category, DEFINE-S[4], P-CURVE[5] come under second category whereas KAKSI[6], PALSSE[7] etc. fall under the third category. All these algorithms

[*] Both authors contributed equally.

identify the main body of the α-helix, however they often differ in their definition of helix terminii. Even in the main body, sometimes differences occur in the assignment because of small local deviations from the uniform helical character, arising due to solvent induced distortions[8], peptide bond distortions[9,10] or presence of Proline[11-13], Serine and Threonine residues[14,15]. Blundell *et.al.*[8] carried out the first survey on the irregularities in helices and concluded that a majority of the helices are curved, which has been confirmed by subsequent studies[16]. α-helices also show distinct preferences for amino acids at their N and C terminal regions which define the helix initiation and termination motifs[17-19] as well as at mid-positions[20].

Assignment of β-strands is comparatively more complex, as the residues involved in the hydrogen bonds, can be far from each other in sequence space. Several experimental[21-23] and statistical[24,25] analyses have not conclusively indicated any specificity for the terminal residues in β-strands.

In this article, we describe the results of an analysis of helix terminii in α-helices only, since they are most numerous and of sufficient length to be well defined geometrically. An α-helix is defined as extending from residues N1 to C1 which make up the main body of the helix while Ncap and Ccap are the immediately preceding and succeeding non-helical residues. We have recently developed a program, designated ASSP (Assignment of Secondary Structure in Proteins) as an extension of our in-house program HELANAL-Plus[26], which identifies helical regions based on the geometry of the C^α trace. The most widely used program STRIDE (**Str**uctural **ide**ntification) uses hydrogen bond patterns along with the backbone torsional angles φ and-ψ to define helical regions. Results of analysis using both the algorithms, for helices in globular as well as membrane proteins, have been compared. Our analysis using ASSP indicates that in some cases, even if the main chain-main chain hydrogen bonds are missing at the terminii, the overall helical geometry is retained by the C^α atoms and hence the corresponding residues should be categorized as being part of a helix, while in other cases an α-helix is found to be interspersed or flanked by a small, but identifiable, fragment of 3_{10} or π-helix.

2. Methods

2.1 Dataset of globular and membrane proteins

A non-redundant dataset of globular proteins was created at the fold level, using the ASTRAL[27] compendium in the SCOP[28] database. The ASTRAL compendium groups proteins with the same major secondary structures, with the same arrangement and topological connections. Total of 1195 representative folds were downloaded from ASTRAL-1.75 release database. The dataset was further refined by the following steps: (i) Domains with SPACI score less than 0.4 (resolution worse than 2.5 Å) were excluded (total 266 folds removed) (ii) Proteins which had a missing ATOM record for any of the residues were excluded (total 1 fold removed) (iii) proteins with 'all beta' fold were excluded (total 174 folds removed) and (iv) membrane and cell surface proteins were excluded (total 58 folds removed). The final dataset consisted of 626 representative folds.

A separate non-homologous X-ray crystal structure dataset of membrane proteins was created by selecting non redundant protein structures from PDBTM[29]. Proteins with a resolution better than 2.5 Å, R factor less than 0.25 and sequence identity less than 25% were selected using the PISCES server[30]. The dataset contains 75 proteins comprising 181 chains and includes 48 out of the 58 folds defined by the SCOP database.

Coordinates for the culled datasets of globular and membrane proteins were then downloaded from PDB[31].

2.2 Helix assignment and identification of H-bonds

In-house program ASSP and commonly used STRIDE program were used for defining helices in both the datasets. Helical parameters and local bending angles for successive steps have been calculated using HELANAL-Plus[26]. Hydrogen bonds were identified using HBPLUS v3.06[32].

3. Results and Discussion

3.1 Length distribution of the helices in both datasets

The globular protein dataset contained a total of 3615 and 4028 α-helices as assigned by STRIDE and ASSP respectively. The length of helices defined by STRIDE for the globular and membrane protein dataset varies from 1 to 80 and 1 to 68 residues respectively. However, for ASSP it varies from 4 to 80 and 4 to 69 residues respectively, since ASSP defines helices based on the geometric trace of C^{α} atoms and requires a minimum of four residues. The mean length of helices assigned by ASSP in both datasets is hence found to be smaller, particularly in case of membrane proteins, where the mean length is 16 according to STRIDE and 13 according to ASSP. The length distribution of the helices with length >4 residues, in both the datasets, is shown in Figure 1. Interestingly a few of the long helices assigned by STRIDE are broken and defined as two or more helices by ASSP, thereby increasing the number of shorter helices (≤15 residues). In this study, we have analyzed in detail only α-helices with length >8 residues, since in these helices at least one residue has both N-H and C=O groups involved in hydrogen bond formation and hence these are expected to be structurally well defined**.** In globular protein dataset, 2680 and 2665 STRIDE and ASSP assigned helices were found to be of length >8 residues. In membrane proteins, out of 1164 and 1404 STRIDE and ASSP assigned helices, 865 and 883 helices respectively were found to be of length >8 residues.

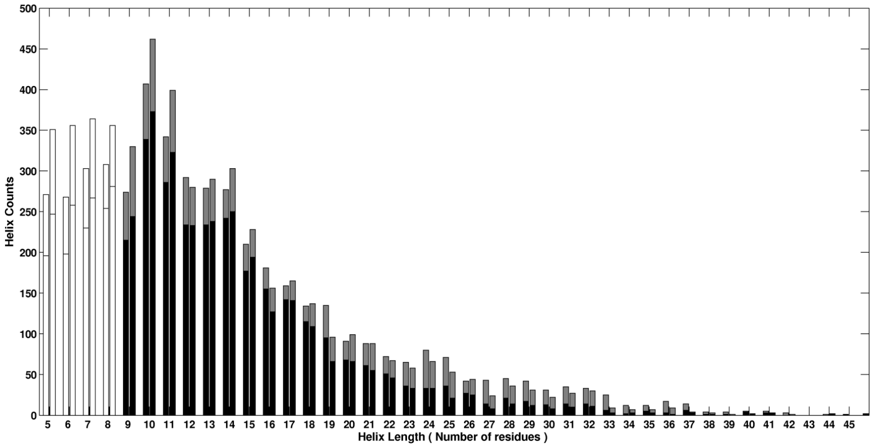


Figure 1: Stacked bar diagram showing length distribution for α-helices assigned by STRIDE and ASSP in globular (black bars) and membrane (grey bars) proteins. α-helices with ≤8 amino acid residues are shown as white bars and not considered in present analysis. In each pair of stacked bars, first bar represents helix assignment by STRIDE and second one corresponds to ASSP. A few α-helices with length >45 residues are also defined by both programs. Such helices are very few in number (<5) and the data are not displayed in the plot.

3.2 Mismatch in the helix terminus assignment

The variation in helix length found by STRIDE and ASSP analysis clearly indicates that there are differences in identifying the helix terminii by the two programs. Table 1 gives the statistics of the STRIDE and ASSP helix assignments with exact match at both terminii, match at either N or C terminii, as well as mismatch at both ends. 1307 and 320 helices in globular and membrane proteins match at both N and C terminii, as assigned by both STRIDE and ASSP. A significantly higher number of helices are observed to have only the N-terminii matching, as compared to those with only their C-termini matching.

Table 1: Statistics of the matching and non-matching helix terminal assignments by STRIDE and ASSP algorithms for α-helices in globular and membrane protein datasets.

Dataset	No. of Helices		No. of Helices (Length > 8)		Both ends matching	Only N-terminii matching	Only C-terminii matching	Both terminii not matching
	STRIDE	ASSP	STRIDE	ASSP				
Globular	3615	4028	2680	2665	1307	918	403	37
Membrane	1164	1404	865	883	320	339	190	16

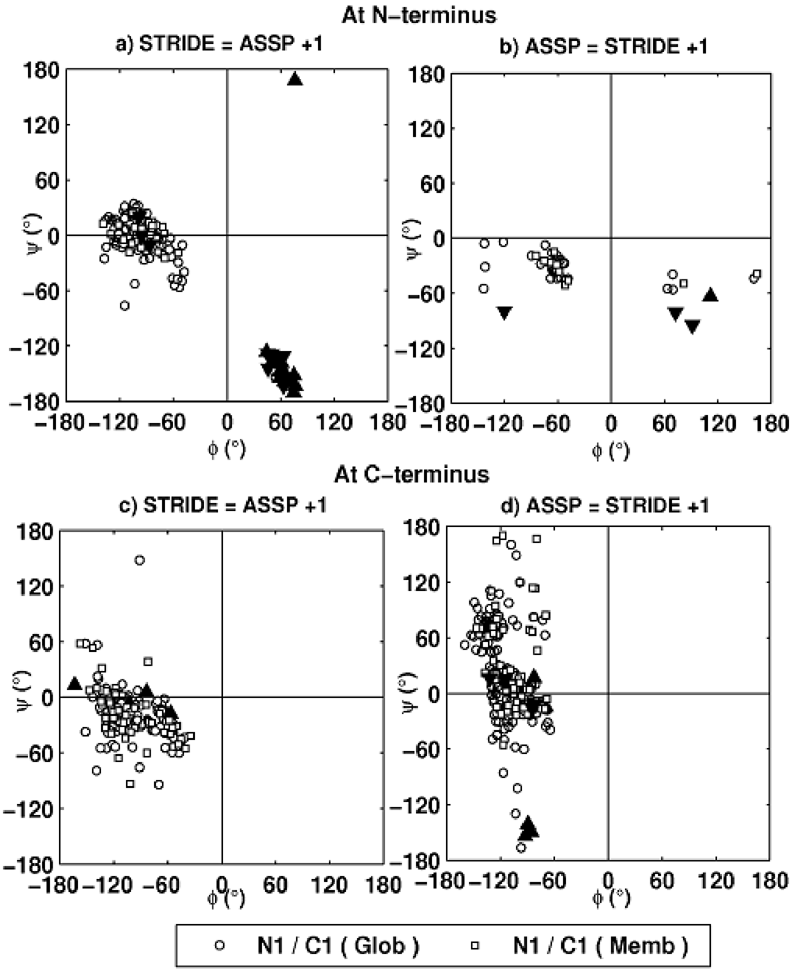
The terminal regions of the helices show a wider spread in their (φ-ψ) values as compared to the MID regions. However, the commonly used structural parameters, viz. unit rise and unit twist, calculated for these α-helices, do not show large variations, as also reported in our earlier analysis of long α-helices[33].

The (φ-ψ) distribution for helices with identical assignments by STRIDE and ASSP for globular (1307) and membrane (320) protein datasets lies well within the allowed regions of the Ramachandran Map (data not shown). The (φ-ψ) distribution for N2 to C2 residues in helices with one residue mismatch at either terminus, according to STRIDE and ASSP assignments, shows a similar spread.

3.3.1 Helices with one residue mismatch at the N terminus

In helices where STRIDE assignment of N terminus starts one residue before ASSP, the STRIDE assigned N1 position becomes the Ncap position for ASSP. There are 154 and 46 helices showing such mismatches in globular and membrane proteins datasets respectively. The N1 residues in this case show two distinct clusters for backbone torsion angles (φ-ψ), as seen on the Ramachandran map (Figure 2a).

The largest cluster of torsion angles is observed with (φ-ψ) values shifted away from α-helical region, towards the 3_{10} helical and bridge regions. STRIDE identifies these residues as N1 of the helix whereas ASSP defines these positions as Ncap and starts the helix assignment from the next residue. The twist value for the Ncap-N1-N2-N3 step in these helices also deviates considerably from α-helical value, in both datasets, leading to the one residue later start assigned by ASSP. A small cluster is also seen in the bottom right quadrant (φ = 35 to 70° and ψ = -120 to -180°) a region disallowed for non-glycine residues. Interestingly all 24 residues in this cluster are Glycines that can take up these torsion angle values, but are not expected to be categorized as being helical. Since only the carbonyl group at N1 position is involved in H-bond formation, it is not affected by these non-helical torsion angles.



Dataset	N–terminus		C–terminus	
	a	b	c	d
Globular	154	27	170	203
Membrane	46	14	67	78

Figure 2: (φ-ψ) distribution for amino acid residues at N1 position (subplots a and b) and C1 position (subplots c and d) in helices with one residue mismatch in assignment by STRIDE and ASSP. Glycine residues in globular and membrane proteins are represented by '▲' and '▼' respectively. The residue position nomenclature in subplots a) and c) is according to STRIDE (ASSP identifies helices that are one residue shorter) whereas in b) and d) it is according to ASSP assignment.

In helices where ASSP assignment of the N terminus starts one residue earlier than STRIDE, the STRIDE assigned Ncap position is the N1 position for ASSP. Only 27 and 14 such cases are seen, with 9 and 5 residues respectively showing non α-helical backbone torsions, in globular and membrane proteins (Figure 2b) The (φ-ψ) of all other residues fall within the allowed α-helical regions of the Ramachandran Map, validating their assignment as N1 residues by ASSP, contrary to STRIDE which designates them as Ncap.

3.3.2 Helices with one residue mismatch at the C terminus

In helices where STRIDE assigns the C terminus one residue later than ASSP, the STRIDE assigned C1 position corresponds to the Ccap for ASSP. 170 and 67 helices with these mismatches are reported in globular and membrane proteins respectively. In both the datasets the (φ-ψ) for residues at C1 position shows a large variation, with several examples lying in the 3_{10} and π as well as the bridge regions (Figure 2c). This leads to the twist and rise values at the C3-C2-C1-Ccap step deviating considerably from the α-helical values in both globular and membrane proteins (as shown in Figure 3a) due to which ASSP truncates the helix one residue earlier. It is interesting to note that even though it is not essential for the backbone amide of Cap to form a main chain-main chain hydrogen-bond, 115 of 4→1 type, 14 of 5→1 and 1 of 6→1 type hydrogen bonds, corresponding to 3_{10}, α and π-helical hydrogen bond patterns, were found in globular proteins, while the corresponding figures for membrane proteins were 45, 4 and 11 respectively (Figure 3a).

In helices where ASSP definition at the C terminus ends one residue later than STRIDE, the ASSP assigned C1 position corresponds to the Ccap according to STRIDE. 203 and 78 helices with mismatches of this type are seen in globular and membrane proteins respectively. In spite of deviations in the (**φ** – ψ) values from α-helical values at C1 position according to ASSP (Figure 2d), the twist values for the C3-C2-C1-Ccap step lie between 78° to 115.6° and 80.3° to 112° for globular and membrane proteins respectively (Figure 3b). The corresponding ranges for rise values are between 0.8 Å to 2.3 Å and 0.9 Å to 2 Å for the two datasets. The Ccap residues in these helices were examined for the presence of a hydrogen bond involving their backbone amide group. 102 of 4→1 type and 37 of 5→1 type hydrogen bonds, corresponding to 3_{10} and α-helical hydrogen bond patterns, were found in globular proteins, while the corresponding figures for membrane proteins were 26 and 29 respectively (Figure 3b). A correlation was observed between the helical parameters

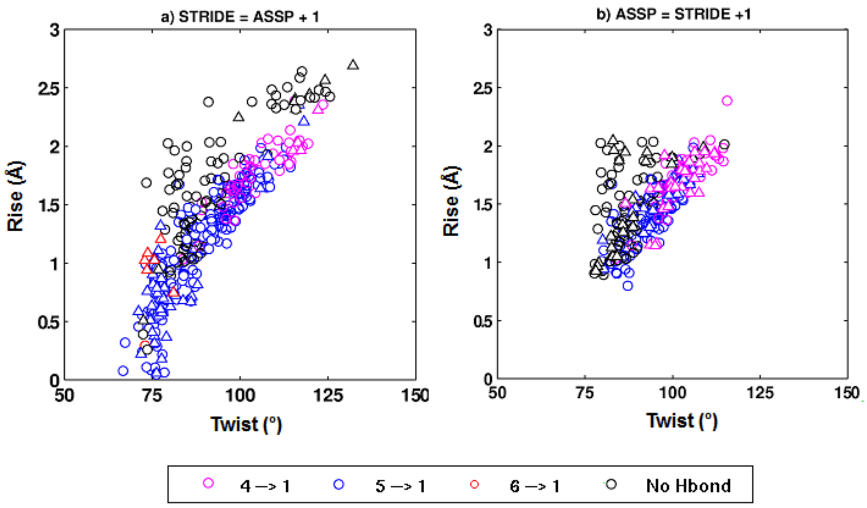
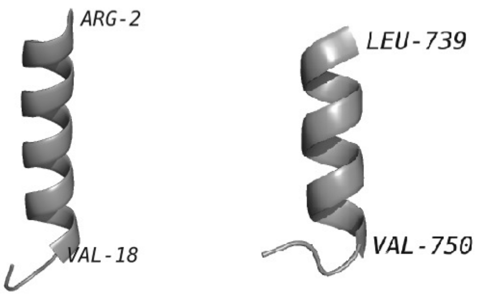


Figure 3: Rise *vs* Twist plots for the step defined by C^{α} atoms of the C3-C2-C1-Ccap residues in the helices with mismatch at C terminii. The color scheme is based on the hydrogen bond pattern observed for the backbone N-H of the Ccap residue. Globular protein data points are represented by 'o' while those of membrane proteins are represented by 'Δ' of the corresponding color.



PDB ID	1L9L; A	1OFC; X
STRIDE	2-17	739-749
ASSP	2-18	739-750
C1 (ϕ, ψ)	-108.8, 160.6	-97.3, -166.4
Twist (°)	99.7^{*}, $81.3^{\#}$	99.2^{*}, $84.6^{\#}$
Rise (Å)	1.8^{*}, $1.9^{\#}$	1.6^{*}, $1.8^{\#}$

Figure 4: Examples of helices with (φ-ψ) of C1 residue (as defined by ASSP) showing deviations from helical values, as seen in Figure 2d. The twist and rise values indicated by '*' and '#' correspond to the C4-C3-C2-C1 and C3-C2-C1-Ccap steps respectively and lie within the accepted helical range.

calculated and the type of hydrogen bond observed. A proline is present at Ccap position in several cases where no hydrogen bond is observed.

Figure 2d shows that, in a few cases the ψ values for the residues at C1 position deviate considerably, taking up either large positive or negative values. An analysis of two representative examples from among these helices (shown in Figure 4) reveals that twist values at the C3-C2-C1-Ccap step lie within the α-helical range and the backbone amide group of C1 is involved in a 5→1 hydrogen bond. The local bending angles calculated by HELANAL-Plus[25] at this step, for the two example helices, are 11° and 1° respectively, indicating that residues assigned as C1 by ASSP are indeed a part of the helix and do not produce a discontinuity or even a kink.

3.3.3 Helices with more than one residue mismatch at the terminii

The length distribution of helices shown in Figure 1 indicated that several STRIDE assigned long α-helices were divided into two or more α-helices by ASSP. In some cases these are interspersed with 3_{10} or π-helical fragments.

For example, in periplasmic hydrogenase protein (PDB id: 1WUI), STRIDE assigns residues 252-284 of 'L' chain as a single α-helix (Figure 5a), while ASSP breaks this 33 residue helix into two α-helical fragments (252-268 and 274-284) with an intervening π-helix (269-273). HELANAL-Plus gives the average twist and rise per residue for the region 269-273 as 77.9° and 0.99 Å, respectively, close to the π-helix values and 6→1 hydrogen bonds are also observed between 273(NH)→268(CO) and 274(NH)→269(CO), confirming the correct assignment of this region as a π-helix by ASSP.

Similarly in Aquaglyceroporin, a membrane protein (PDB id: 3C02), STRIDE assigns residue numbers 210–248 of ‘A’ chain as a single helix (Figure 5b), while ASSP breaks this into three α-helical segments (211-214; 220-239 and 242-248) and surprisingly also identifies a left-handed α-helical fragment (between residues 216-218). HELANAL-Plus gives (Twist, Rise) values of (-96.59°, 1.53Å) and (-99.3°, 1.56Å) for the first and second step respectively. When the (φ-ψ) for the residues in this helix were checked, residue numbers 216-218 were found to have positive values for both φ and ψ, confirming the occurrence of a left-handed region, as assigned by ASSP.

There are also a few cases where ASSP defined helices are much longer than STRIDE. For example, in both chains ‘A’ and ‘B’ of protein yheA (PDB id: 2OEE), ASSP defines a single kinked α-helix between residues 81-111 with large bending angles of ~44° at position M105 and K106 (Figure 5c), while STRIDE breaks it into two α-helical parts (81-85 and 96-111) and a 3_{10} helix

(86-89) for both ‘A’ and ‘B’ chains. Interestingly PDB identifies two α-helices (80-105 and 105-112) in chain ‘A’ while in the ‘B’ chain, a single helix spanning residues 80-112 is identified. The (φ-ψ) values and helical parameters in the 86-89 region do not indicate the presence of a 3_{10} helix.

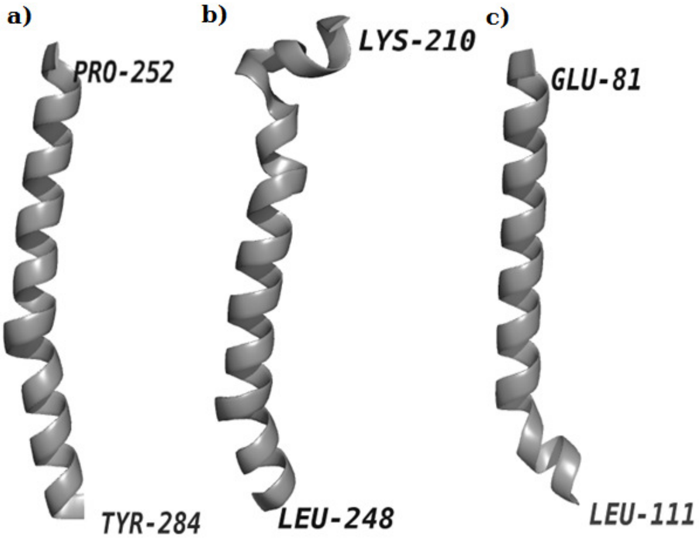


Figure 5: Helices with STRIDE and ASSP helix assignment differing significantly. Representative helices taken from a) Periplasmic hydrogenase (PDB id: 1WUI) b) Aquaglyceroporin (PDB id: 3C02) and c) Protein yheA (PDB id: 2OEE). In examples a) and b) the helix start and end residues indicated are as defined by STRIDE for a single long helix, while in c) the residue assignment is according to ASSP.

4. Conclusion

In this study we have compared α-helix assignments by two algorithms, one based on C^{α} trace (ASSP) and the other using a combination of hydrogen bond pattern along with backbone torsion angles φ and ψ (STRIDE). We find that local variations/deviations in the helical geometries are better detected by ASSP, leading to it breaking longer helices into shorter ones and even identifying some short 3_{10} or π helical regions. A systematic comparison shows higher number of mismatches at the C terminii as compared to the N terminii of α-helices in both globular and membrane protein datasets, indicating that the helix initiation site is better defined. ASSP often identifies a π-helical region at the C-terminii, leading to a large number of mismatches in assignment between the two algorithms.

A geometrically uniform C$^{\alpha}$ trace defines the helical nature of the protein chain, but the exact orientation of the peptide unit about the virtual C$^{\alpha}$ - C$^{\alpha}$ bond determines the backbone torsion angles (φ-ψ) and hydrogen bond pattern between the backbone amides and carbonyls. Specifically in helices where ASSP extends the helix at the C terminus (ASSP = STRIDE+1), the (φ-ψ) of residues at the C1 position assigned by ASSP deviate from α-helical values in some cases, which is accompanied by absence of hydrogen bonds involving the backbone amide of Ccap. Hence while STRIDE truncates the helix one residue earlier, these residues are included as C1 in the helix definition by ASSP, since the C3-C2-C1-Ccap step has acceptable twist and rise values. Thus, while the backbone torsion angles and hydrogen bonds are required to describe the detailed features of a helix and other secondary structures, geometry defined by the C$^{\alpha}$ atom trace seems to be a good alternative to define variations in helical parameters, which can better determine the extent of a helix and differentiate between various helical conformations. It can also identify helices in low resolution structures, where (φ-ψ) values may not be highly reliable. We believe that insights from our study will help in improving the understanding of residue preferences at helix initiation and termination sites, as well as in modeling tertiary structures of large protein complexes.

Acknowledgments

MB would like to express her gratitude to Professor Ramachandran for his unfailing encouragement and inspiring mentorship. This work was supported by DBT India. Prasun Kumar acknowledges DBT-BINC, India for a fellowship. MB is a recipient of J. C. Bose National Fellowship of Department of Science and Technology, Government of India.

References

1. W. Kabsch and C. Sander, *Biopolymers* **22**, 2577 (1983).
2. D. Frishman and P. Argos, *Proteins* **23**, 566 (1995).
3. R. Srinivasan and G.D. Rose, *Proc. Natl. Acad. Sci. USA,* **96**, 14258. (1999).
4. F.M. Richards and C.E. Kundrot, *Proteins*, **3**, 71(1988).
5. H. Sklenar, et al., *Proteins* **6**, 46 (1989).
6. Martin,J. et al., *BMC Struct. Biol.*, **5**, 17 (2005).
7. Majumdar,I. et al. (2005). *BMC Bioinformatics*, **6**, 202.8 15 (2005).
8. T. Blundell, D. Barlow, N. Borkakoti and J. Thornton, *Nature* **306**, 281 (1983).

9. W. E. Love, P.A. Klock, E. E. Lattman, E. A. Padlan, K. B. Ward, Jr. and W. A. Hendrickson, *Cold Spring Harb Symp Quant Biol* **36**, 349 (1972).
10. D. J. Barlow and J. M. Thornton, *J Mol Biol* **201**, 601 (1988).
11. P. Chakrabarti, M. Bernard and D. C. Rees. *Biopolymers* **25**, 1087 (1986).
12. M. W. MacArthur and J. M. Thornton, *J Mol Biol* **264**, 1180 (1996).
13. R. Sankararamakrishnan and S. Vishveshwara, *Biopolymers* **30**, 287 (1990).
14. J. A. Ballesteros, X. Deupi, M. Olivella, E. E. Haaksma and L. Pardo, *Biophys J* **79**, 2754 (2000).
15. X. Deupi, M. Olivella, C. Govaerts, J. A. Ballesteros, M. Campillo and L. Pardo, *Biophys J* **86**, 105 (2004).
16. S. Kumar and M. Bansal, *Biophys J* **75,** 1935 (1998).
17. R. Aurora, G. D. Rose, *Protein Sci.* **7**, 21 (1998).
18. S. Kumar, M. Bansal, *Proteins: Struct Funct Genet* **31**, 460 (1998).
19. K. Gunasekaran, H. A. Nagarajaram, C. Ramakrishnan, P. Balaram, *J Mol Biol* **275**, 917 (1998).
20. D. E. Engel and W.F. DeGrado, *J Mol Biol* **337***,* 1195 (2004).
21. S. M. Zaremba, L. M. Gregoret, *J Mol Biol* **291**, 463 (1999).
22. J. S. Merkel, L. Regan, *Fold Des* **3,** 449 (1998).
23. J. S. Merkel, J. M. Sturtevant, L. Regan, *Fold Des* **7,** 1333 (1999).
24. M. A. Wouters, P. M. Curmi, *Proteins* **22**, 119 (1995).
25. E. G. Hutchinson, R. B. Sessions, J. M. Thornton, D. N. Woolfson, *Protein Sci* **7**, 2287 (1998).
26. P. Kumar, M. Bansal, *J Biomol Struct Dyn* **30**, 773 (2012).
27. S. E. Brenner, P. Koehl and M. Levitt, *Nucleic Acids Res* **28**, 254 (2000).
28. A. Andreeva, D. Howorth, J. M. Chandonia, S. E. Brenner, T. J. Hubbard, C. Chothia and A. G. Murzin, *Nucleic Acids Res* **36**, D419 (2008).
29. G. E. Tusnady, Z. Dosztanyi and I. Simon, *Nucleic Acids Res* **33**, D275 (2005).
30. G. Wang and R. L. Dunbrack, Jr, *Bioinformatics* **19**, 1589 (2003).
31. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne. *Nucleic Acids Res* **28**, 235 (2000).
32. I. K. McDonald and J. M. Thornton. *J Mol Biol* **238**, 777 (1994).
33. S. Kumar, M. Bansal, *Biophys J* **75**, 1935 (1998).