BIOLOGICAL
CRYSTALLOGRAPHY

Volume 71 (2015)

Supporting information for article:

Identification of local variations within secondary structures of proteins

Prasun Kumar and Manju Bansal

S1. Brief description of *ASSP* algorithm

ASSP first calculates the local structure parameters like twist, rise, virtual torsion angle and radius by considering four contiguous C^{α} atoms (repeating unit or step) with a sliding window of one C^{α} atom at a time along the length of the protein structure. Subsequently, the continuity in the protein structure is checked based on these parameters and continuous stretches are further divided into the SSEs.

S1.1. Calculation of the parameters

Let $\mathbf{CA_1}$*(x1, y1, z1)*, $\mathbf{CA_2}$*(x2, y2, z2)*, $\mathbf{CA_3}$*(x3, y3, z3)* and $\mathbf{CA_4}$*(x4, y4, z4)* are the Cartesian coordinates of four contiguous C^{α} atoms (Supplementary Fig. S9). Two successive points are joined to each other by pseudo bonds and can be given as:

- **B1** = $\mathbf{CA_1CA_2}$ = *(x2-x1)***i** + *(y2-y1)***j** + *(z2-z1)***k**
- **B2** = $\mathbf{CA_2CA_3}$ = *(x3-x2)***i** + *(y3-y2)***j** + *(z3-z2)***k**
- **B3** = $\mathbf{CA_3CA_4}$ = *(x4-x3)***i** + *(y4-y3)***j** + *(z4-z3)***k**

Where, **i**, **j**, **k** are unit vectors along X, Y and Z axes respectively. The vectors obtained by subtracting the pseudo bond vectors are:

- **V1** = **B1 - B2**
- **V2** = **B2 - B3**

V1 and **V2** lie in a plane perpendicular to the axis of a helix described by these four atoms. The direction cosines *(l, m, n)* of the helix axis **U** were obtained from the cross product of vectors **V1** and **V2**:

- **U** ***(l, m, n)*** = (**V1** x **V2**) / (|**V1** x **V2**|)

Finally, various geometric parameters were derived using following equations:

- **Twist (t)** = $\cos^{-1}$ ((**V1.V2**) / (|**V1**||**V2**|))
- **Vtor (θ)** = Calculation is similar to that of (φ, ψ). The only difference is that the four atoms involved are C^{α}. Since, the bonds between them are virtual, the torsional angle is called Vtor (Virtual torsional angle)
- **Rise (h)** = ((**B2.U**) / |**U**|
- **Radius (r)** = *(sqrt (|**V1**||**V2**|)* / (2(1-$\cos^{-1}$t)))
- **Bending Angle (BA) =** It is defined as an angle between successive local helical axes corresponding to two sets of four C^{α} atoms $\mathbf{CA_1}$**,** $\mathbf{CA_2}$**,** $\mathbf{CA_3}$**,** $\mathbf{CA_4}$ and $\mathbf{CA_4}$**,** $\mathbf{CA_5}$**,** $\mathbf{CA_6}$**,** $\mathbf{CA_7}$. The bending angle obtained will be at $\mathbf{CA_4}$

S1.2. Identification of continuous stretches

Two contiguous steps will be said to be a part of a continuous stretch, if and only if the absolute value of (Twist difference (ΔTwist), Rise difference (Δh), Vtor difference (ΔVtor)) ≤ (35°, 1.1Å, 50°), where

- **ΔTwist** = |Twist1 - Twist2|
- **Δh** = |h1 - h2|
- **ΔVtor** = |Vtor1 - Vtor2|

Here (Twist1, h1, Vtor1) and (Twist2, h2, Vtor2) are the structural parameters of repeating units 1 and 2 respectively.

The twist *vs* rise and twist difference *vs* rise difference plots for α-helices identified by *DSSP*, *STRIDE* and *ASSP* are shown in Supplementary Fig. S10.

S1.3. Classifying the continuous stretches

Continuous stretches as a whole or part of it are further classified into different type of SSEs by first assigning characters (**A/ a, G/ g, I/ i, P, S** and **U**) to individual step based on the step parameter values.

S1.4. Final arrangement

No two secondary structures should be overlapping and in order to address this issue, final check of the same is done. Here we make use of minimum length criteria of different Secondary structure elements (SSEs) identified by *ASSP*. The minimum possible length for π, α, 310 and PPII helices are 5, 4, 3 and 3 residues respectively. The overlap between two SSEs (assuming 1st SSE is S1 and 2nd is S2) can only be of one residue at the junction. Broadly two cases are possible:

Case 1: When the two SSEs are of different SSE types

Based on the length of two SSEs, it is further divided into four different categories.

i. **Length of S1 is the minimum possible and that of S2 is more than the minimum possible.** In this case, the termini of S1 remain the same, while N2 of S2 becomes N1, with the length of S2 reduced by one.

ii. **Length of S1 is more than the minimum possible and that of S2 is the minimum possible.** In this case, the termini of S2 remain the same, while C2 of S1 becomes the C1 with the length of S1 reduced by one.

iii. **Length of S1 and S2 both are more than the minimum possible.** The treatment to this case will be similar to the case (ii).

iv. **Length of S1 and S2 both are minimum possible.** If the SSEs are of same type, both are merged to give a single SSE of the same type and in this case the bending angle is not

checked. In case, when the SSE types are different, the SSEs will be merged and the new SSE will be of the longer one.

v. **When the length of either of S1 or S2 is less than minimum possible and other SSE is longer than the minimum possible:** The SSE with the length < minimum possible, will be merged to the longer one

Case 2: When the two SSEs are of same types

When two secondary structure elements (SSE) are of same type and the bending angle at the overlapping residue ≤ 60°, the two SSEs are combined with the N1 coming from the 1st SSE and C1 from 2nd. In other case, where the bending angle at the overlapping residue > 60°, we treat them as two separate SSEs and further arrangement is same as of case1.

Table S1 Brief description about different Secondary Structure assignment algorithms. The algorithms are divided according to the categories mentioned in the main text.

Sl. No.	Algorithm	Description	Reference
Category (i)			
1	***DSSP***	Detects the hydrogen-bond patterns using bond energy criterion	(Kabsch & Sander, 1983)
2	***STRIDE***	Uses (φ, ψ) along with hydrogen bond pattern	(Frishman & Argos, 1995)
3	***PROSS****	Uses only on the backbone dihedral angles (φ, ψ)	(Srinivasan & Rose, 1999)
4	***SECSTR***	Uses DSSP like hydrogen bond definition and was developed to identify and analyze π-helices	(Fodje & Al-Karadaghi, 2002)
5	***DSSP-PPII****	Identifies the PPII helices in the region not assigned as a major SSE by DSSP and gives the output in the DSSP format	(Mansiaux *et al.*, 2011)
Category (ii)			
6	**Levitt *et.al.***	Uses distance and virtual torsion angle made by the CA atoms over a sliding window of four residues	(Levitt & Greer, 1977)
7	***DEFINE-S***	Uses only CA coordinates and compares the distance between various CAs with the distances	(Richards & Kundrot, 1988)

		in ideal SSEs	
8	***P-CURVE***	To start with, it chooses the successive repeating unit and does the analysis of mathematical analysis of protein curvature	(Sklenar *et al.*, 1989)
9	***P-SEA***	Solely based on the CA atoms. Uses three distance, one angle and one dihedral angle	(Labesse *et al.*, 1997)
10	***XTLSSTR****	Calculates two angles and three distances for assigning SSEs. The algorithm is driven by the concept of circular dichroism (CD) of a protein in the far ultraviolet range.	(King & Johnson, 1999)
11	***STICK***	Finds a set of best fit axes and later takes the average rise of the residues along each axis	(Taylor, 2001)
12	***SST***	Uses minimum message length inference for SSEs assignment	(Konagurthu *et al.*, 2012)
Category (iii)			
13	***KAKSI***	Uses CA distances and backbone dihedral angles to show the concordance with the assignments found in the Protein Data Bank	(Martin *et al.*, 2005)
14	***PALSSE***	Mainly uses distance and torsion angle constraints to identify core elements and later extends them to longer segments	(Majumdar *et al.*, 2005)
15	***SEGNO****	The CA atoms along with the backbone dihedral angles (φ, ψ) and the angle-distance hydrogen bond	(Cubellis *et al.*, 2005)

***** Algorithms identify PPII-helices also

Table S2 Percentage content in-terms of amino acid residues assigned as α, 3_{10}, π helix and strand content by different algorithms in four mentioned datasets (HRes, MRes, LRes and NMR). KAKSI and PALSSE do not differentiate between α, 3_{10} and π and assign them as helix only. π value for many of the algorithms is 0 because values are rounded to first decimal place.

	HRes				**MRes**				**LRes**				**NMR**			
	α	**3_{10}**	**π**	**β**	**α**	**3_{10}**	**π**	**β**	**α**	**3_{10}**	**π**	**β**	**α**	**3_{10}**	**π**	**β**
ASSP	35.2	3.5	0.9	22	35.7	3.4	0.8	22.6	32.1	3.6	0.8	19.4	32.5	3.7	0.8	16.1
DSSP	35.3	4.8	0	22.5	36.1	4.2	0	22.9	33.7	3.3	0	20.4	33.7	1.6	0	17.3
KAKSI	36.4			22	38			22.5	35.1			19	32.2			15.2
PALSSE	57.6			23.2	57.3			24.1	54.6			22.8	54.6			20
SST	34.8	1.5	0.5	22.3	35.3	1.5	0.4	23	33.1	1.7	0	20.7	30.8	2.2	0.5	18.2
STRIDE	36.5	5	0	22.6	37.3	4.4	0	23.3	34.3	3.5	0.3	21.2	35.1	2	0	18.8

Table S3 Mean (std. dev.) values of twist, rise and radius for the PPII helices assigned by different algorithms

	ASSP	**DSSP-PPII**	**PROSS**	**SEGNO**	**XTLSSTR**
Twist (°)	237.6 (9.2)	224.4 (44)	231.3 (35.9)	234.6 (27.8)	234.4 (33.7)
Rise (Å)	3.0 (0.1)	2.8 (0.5)	2.9 (0.4)	2.9 (0.4)	2.9 (0.5)
Radius(Å)	1.3 (0.1)	1.4 (0.4)	1.4 (0.3)	1.4 (0.3)	1.3 (0.4)

Table S4 Comparison of left handed helices as identified by ASSP and (Novotny & Kleywegt, 2005). The corresponding assignments by other algorithms, which have been discussed in the referenced paper, are not tabulated here.

Sl. No.	**PDB ID: Chain ID**	**Protein Name**	**ASSP**	(Novotny & Kleywegt, 2005)
1	1BD0:A	Alanine racemase	40-44 (α)	40-44 (α)
2	1AUT:L	Activated protein C	101-104 (3_{10})	101-104 (3_{10})
3*	1B9W:A	Merozoite surface protein 1 (P. cynomolgi)	52-55 (3_{10})	52-55 (3_{10})
4	1G2L:B	Coagulation factor X	258-261 (3_{10})	258-261 (3_{10})
5	1KLI:L	Coagulation factor VII	94-97 (3_{10})	94-97 (3_{10})
6	1N1I:A	Merozoite surface protein 1 (P. knowlesi)	57-60 (3_{10})	57-60 (3_{10})

7	1OB1:C	Merozoite surface protein 1 (P. falciparum)	52-55 (3_{10})	52-55 (3_{10})
8	1RFN:_	Coagulation factor IX	91-94 (3_{10})	91-94 (3_{10})
9	1PB5:A	Lnr module from Notch	29-32 (3_{10})	28-32 (3_{10})
10	1KDG:A	Cellobiose dehydrogenase	532-535 (3_{10})	532-535 (3_{10})
11	1HZM:A	Protein phosphatase 6	60-63 (α)	61-64 (α)
12[ζ]	1GTX:A	4-Aminobutyrate aminotransferase	70-73 (α)	70-73 (α)
13	1QJ5:A	7, 8-Diaminopelargonic acid synthase	50-53 (α)	50-53 (3_{10})
14	2GSA:A	Glutamate semialdehyde aminotransferase	65-67 (3_{10})	65-68 (3_{10})
15	2OAT:A	Ornithine aminotransferase	83-86 (α)	83-86 (3_{10})
16	1BNL:A	Endostatin (Homo sapiens)	135-138 (α)	135-138 (α)
17	1DY2:A	Endostatin (M. musculus)	207-210 (α)	207-210 (α)
18	1KOE:S	Endostatin (M. musculus)	266-269 (α)	266-269 (α)
19	1BQB:A	Aureolysin	223-226 (α)	223-226 (α)
20	1NPC:E	Neutral protease	227-230 (α)	227-230 (α)
21	8TLN:E	Thermolysin	266-269 (α)	266-269 (α)
22	1J9Q:A	Nitrate reductase (A. feacalis)	105-108 (α)	105-108 (3_{10})
23	1NIF:A	Nitrate reductase (A. cycloclates)	105-108 (α)	105-108 (3_{10})
24	1OE1:A	Nitrate reductase (A. xylosoxidans)	99-102 (α)	99-102 (3_{10})
25[‡]	1PTM:A	4-Hydroxythreonine-4-phosphate dehydrogenase	-	211-216 (3_{10})
26	1AK0:E	P1 nuclease	131-134 (α)	131-134 (α)
27[‡]	1MZR:_	2, 5-Diketo-d-gluconate reductase	-	191-194 (3_{10})
28	1H21:A	Split-soret cytochrome C	77-81 (π)	77-81 (3_{10})
29	1HXX:A	Ompf porin	143-146 (α)	143-146 (3_{10})
30	1JV1:A	Glcnac1p uridyltransferase	182-185 (3_{10})	182-185 (3_{10})
31	1KWS:A	Beta-1,3-glucuronyltransferase 3	298-301 (α)	298-301 (α)

Figure S1 Plots showing twist (°) vs. rise (Å) for the steps constituting the *ASSP* identified right handed α, 3_{10}, π and left handed PPII-helices. The corresponding total number of helices and steps is given in the parentheses above each plot. The mean values of parameters are listed in Table 1.

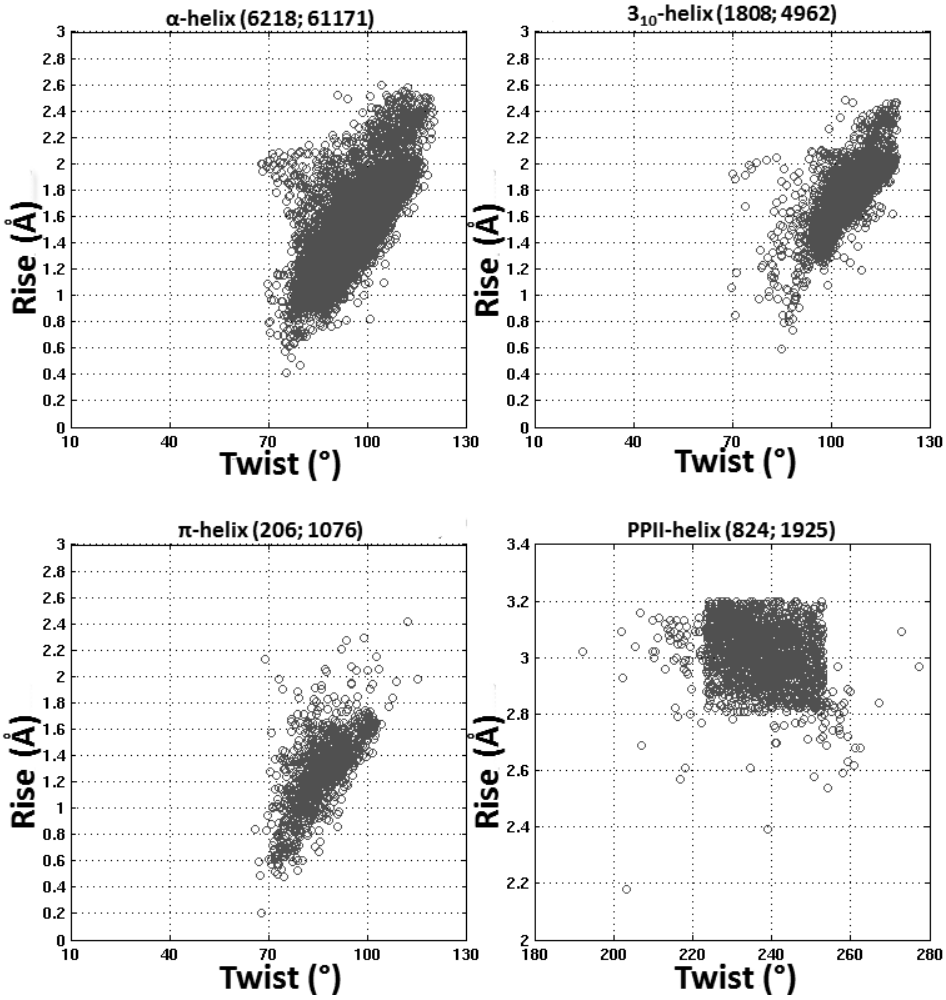


Figure S2 Bar diagram showing length distribution for α, 3_{10} and π-helices assigned by *ASSP*, *DSSP* and *STRIDE*. PPII-helices are not identified by *DSSP* and *STRIDE*.

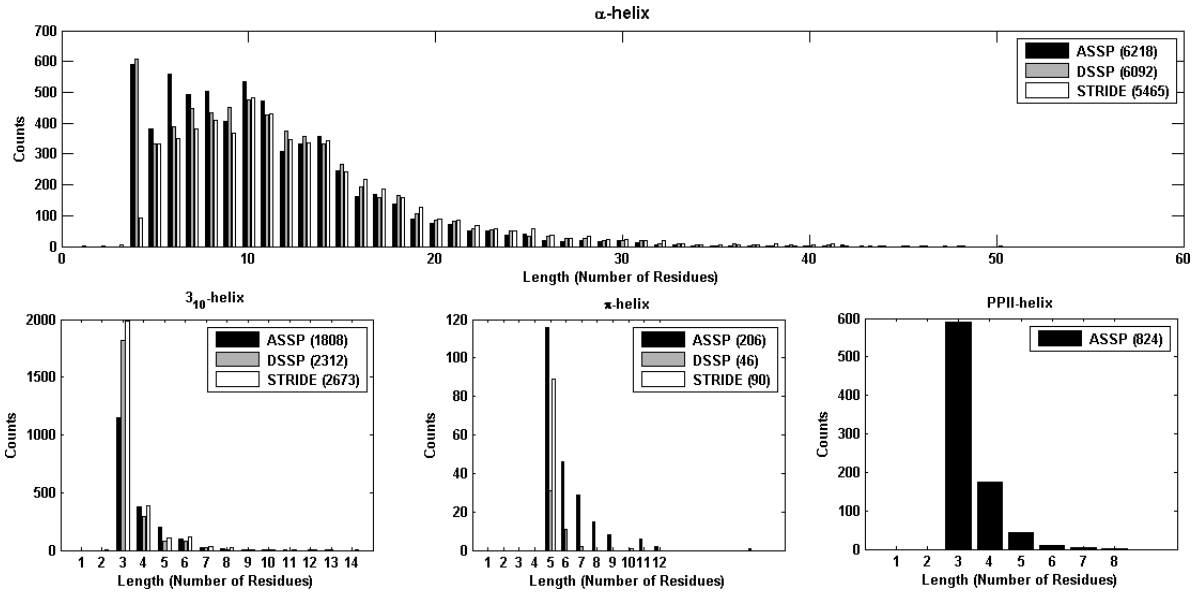


Figure S3 Comparison of SSEs assigned by *ASSP*, *DSSP* and *STRIDE* for full length protein chain (PDB ID: 1H4P:B)

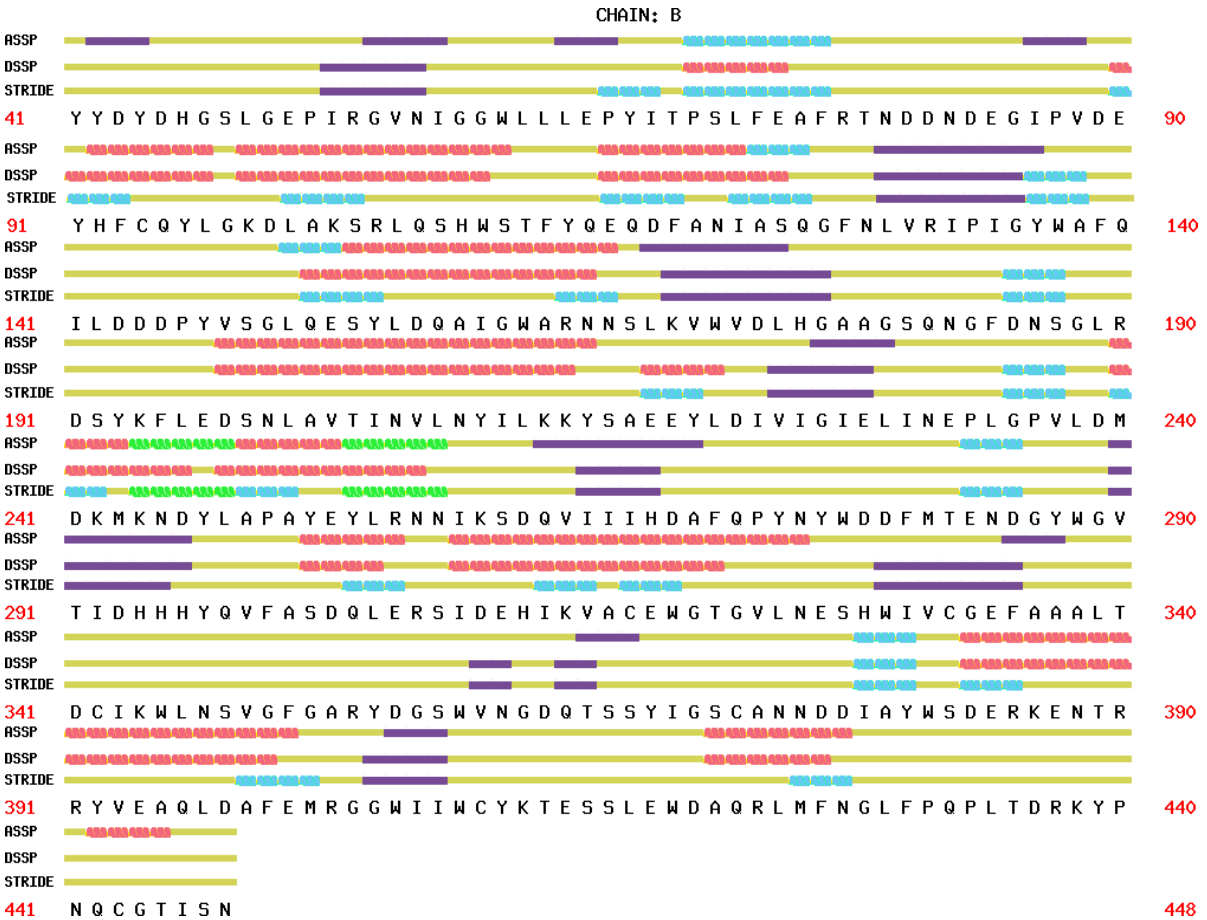


Figure S4 Cartoon diagram showing the SSEs according to *PyMol* for a) 1JSD: A (84S-87N); b) 1EL6: A (25G-29V); c) 1KPF: A (50T-52F; 89G-91N; 116L-118G) and d) 1KIC: A (7L-9H; 174V-176L). Uniform segments identified as a part of cluster (iv) are shown in magenta.

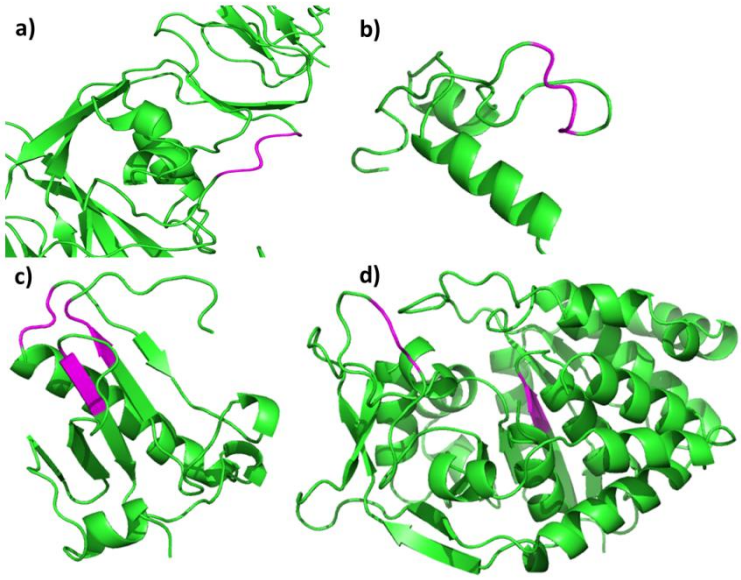


Figure S5 Plots showing backbone torsion angles (φ, ψ) for *ASSP* assigned right handed α, 3_{10}, π and left handed PPII helices. Only the helical residues (N1 to C1) are considered. GLY residues are shown as 'Δ'. The number of helices and residues constituting the helix are given within parentheses above each plot.

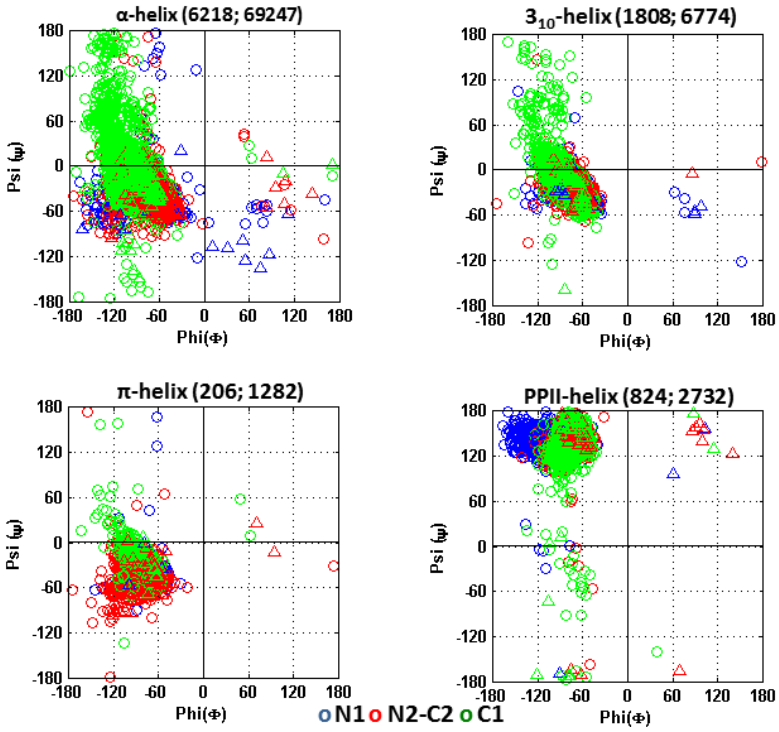
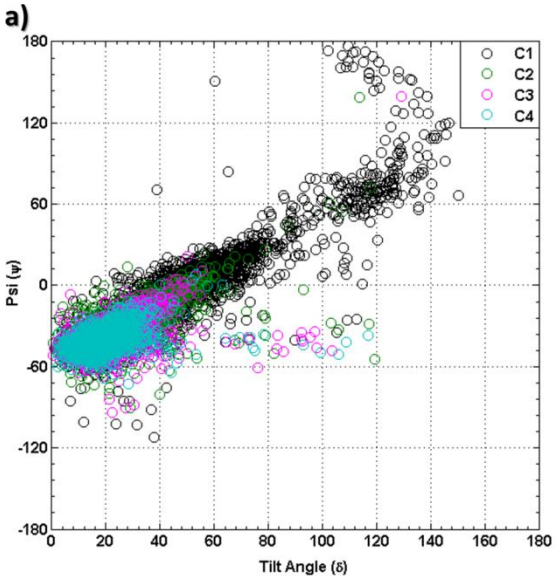
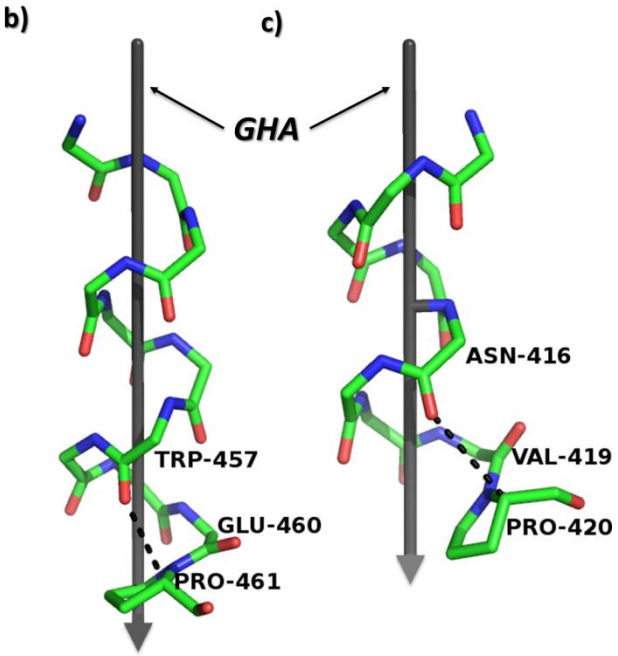


Figure S6 a) Plot showing the correlation between tilt angle (δ) and backbone torsion angle Psi (ψ) for residues at the C-terminal positions C4 to C1 of 4646 α-helices with length >6 residues. b) and c) Representative examples of *ASSP* assigned α-helices with $C1_{\psi}$>40° and C^{α} of PRO at Ccap making a C-H…O hydrogen bond (shown in black dotted line) with backbone carbonyl oxygen of residue at C4 position.





PDB ID: Chain ID	**1FZ7: A**	**2NSZ: A**
Helix (N1-C1)	**451Q-460E**	**412I-419V**
C1 (ϕ, ψ)	**-133.5°, 62.3°**	**-133.8°, 92.0°**
Ccap (ϕ, ψ)	**-53.6°, -38.5°**	**-53.5°, 149.7°**
Tilt Angle (δ)	**110.9°**	**145.3°**
C^{α}_{Ccap}…O_{C4} distance	**3.2Å**	**3.4Å**
C^{α}_{Ccap}-H^{α}_{Ccap}…O_{C4} Angle	**102.6°**	**128.8°**

Figure S7 Distribution of twist (°) *vs* rise (Å) plot for the steps constituting PPII-helices identified by different algorithms. Total number of helices is given in parentheses.

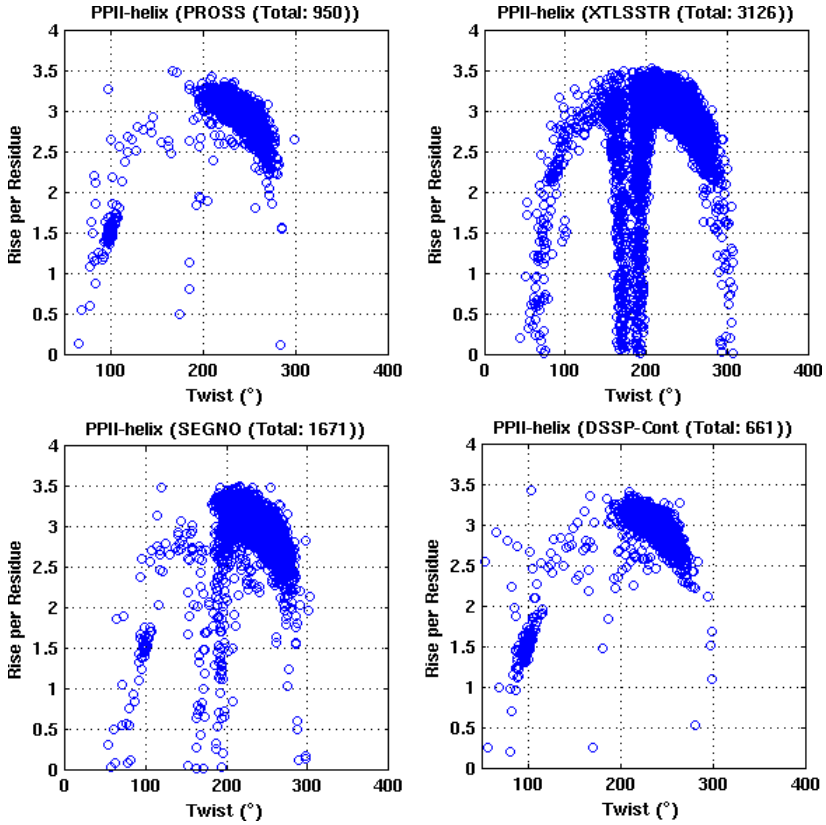


Figure S8 Comparison of SSEs assigned by *ASSP* and *STRIDE* for full length protein chains: a) PDB ID: 3EE4:A b) 1NM8:A c) PDB ID:1B0B:A and d) 3I5O:A.

Figure S9 Down the helix axis view of 4 C^{α} atoms in an ideal α-helix. Various symbols used in the diagram are defined in the text.

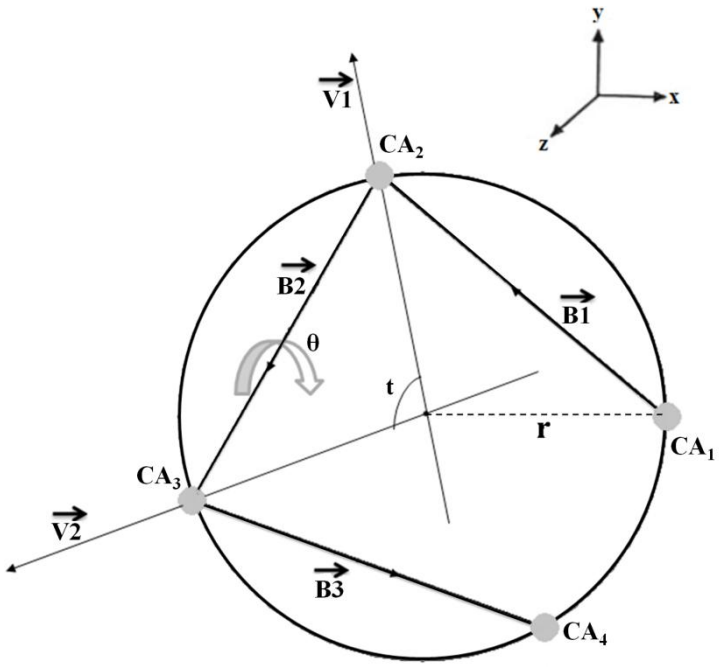
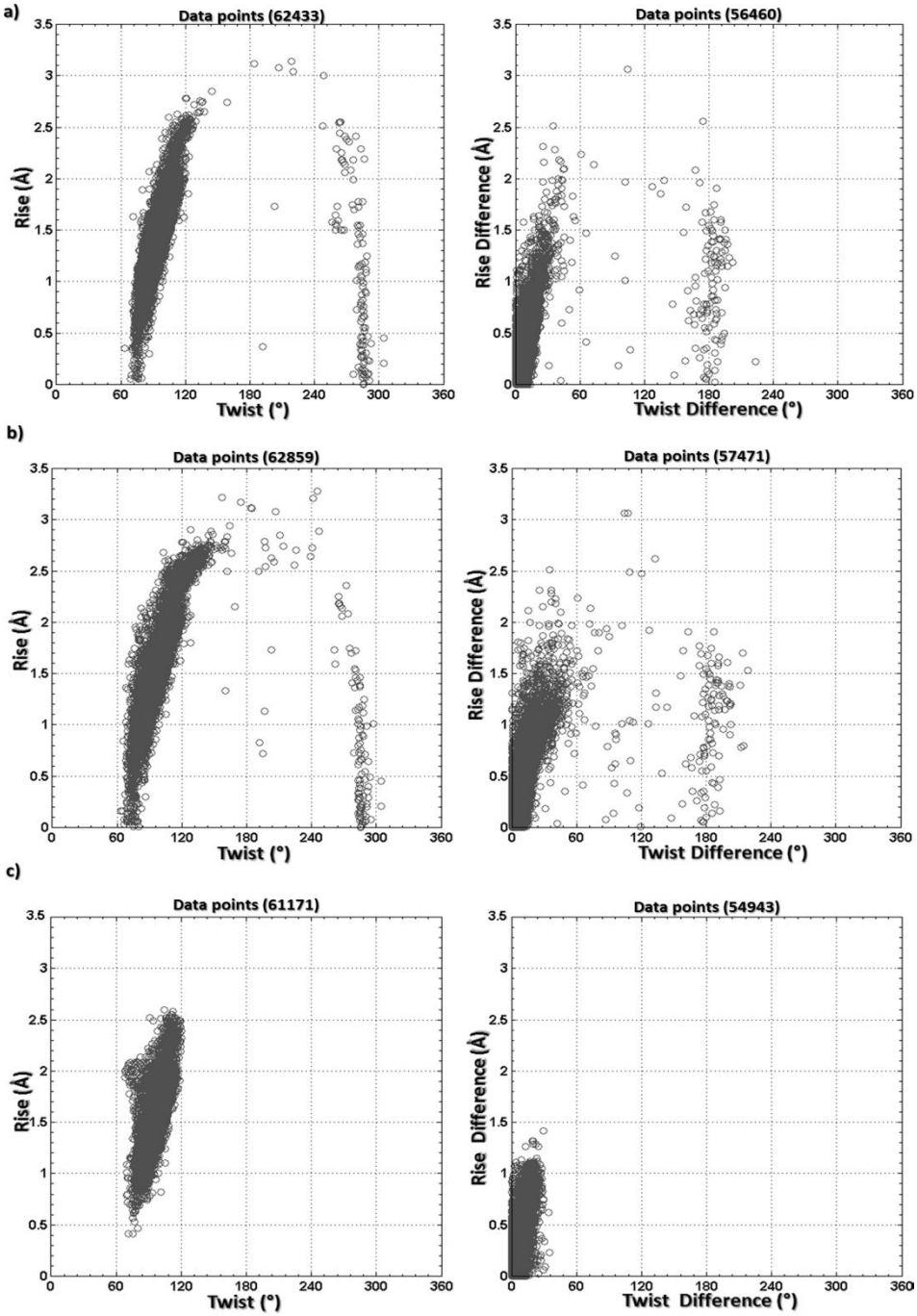


Figure S10 a) Distribution of twist (°) *vs* rise (Å) (column 1) and twist difference (°) *vs* rise difference (Å) (column 2) for the steps constituting α-helices. a) *DSSP* assigned 5465 α-helices; b) *STRIDE* assigned 6092 α-helices and c) *ASSP* identified 6218 α-helices. The number of data points is given above each plot.



References

Cubellis, M. V., Cailliez, F. & Lovell, S. C. (2005). *BMC Bioinformatics* **6 Suppl 4**, S8.
Fodje, M. N. & Al-Karadaghi, S. (2002). *Protein Eng* **15**, 353-358.
Frishman, D. & Argos, P. (1995). *Proteins* **23**, 566-579.

Kabsch, W. & Sander, C. (1983). *Biopolymers* **22**, 2577-2637.
King, S. M. & Johnson, W. C. (1999). *Proteins* **35**, 313-320.
Konagurthu, A. S., Lesk, A. M. & Allison, L. (2012). *Bioinformatics* **28**, i97-105.
Labesse, G., Colloc'h, N., Pothier, J. & Mornon, J. P. (1997). *Comput Appl Biosci* **13**, 291-295.
Levitt, M. & Greer, J. (1977). *J Mol Biol* **114**, 181-239.
Majumdar, I., Krishna, S. S. & Grishin, N. V. (2005). *BMC Bioinformatics* **6**, 202.
Mansiaux, Y., Joseph, A. P., Gelly, J. C. & de Brevern, A. G. (2011). *PLoS One* **6**, e18401.
Martin, J., Letellier, G., Marin, A., Taly, J. F., de Brevern, A. G. & Gibrat, J. F. (2005). *BMC Struct Biol* **5**, 17.
Novotny, M. & Kleywegt, G. J. (2005). *J Mol Biol* **347**, 231-241.
Richards, F. M. & Kundrot, C. E. (1988). *Proteins* **3**, 71-84.
Sklenar, H., Etchebest, C. & Lavery, R. (1989). *Proteins* **6**, 46-60.
Srinivasan, R. & Rose, G. D. (1999). *Proceedings of the National Academy of Sciences of the United States of America* **96**, 14258-14263.
Taylor, W. R. (2001). *J Mol Biol* **310**, 1135-1150.