BIOLOGICAL CRYSTALLOGRAPHY

ISSN 1399-0047

Identification of local variations within secondary structures of proteins

Prasun Kumar and Manju Bansal*

Molecular Biophysics Unit, Indian Institute of Science, Bangalore, Karnataka 560 012, India. *Correspondence e-mail: mb@mbu.iisc.ernet.in

Received 11 December 2014
Accepted 13 February 2015

Edited by Z. Dauter, Argonne National Laboratory, USA

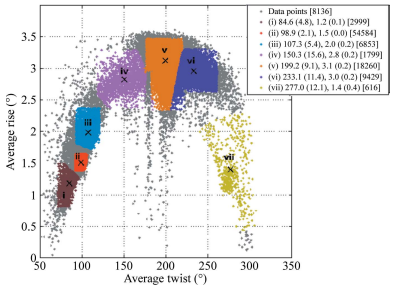
Keywords: *ASSP*; π-helix; polyproline II (PPII) helix; helical wheel; wire diagrams; protein secondary structure.

Supporting information: this article has supporting information at journals.iucr.org/d

Secondary-structure elements (SSEs) play an important role in the folding of proteins. Identification of SSEs in proteins is a common problem in structural biology. A new method, *ASSP* (**A**ssignment of **S**econdary **S**tructure in **P**roteins), using only the path traversed by the C^{α} atoms has been developed. The algorithm is based on the premise that the protein structure can be divided into continuous or uniform stretches, which can be defined in terms of helical parameters, and depending on their values the stretches can be classified into different SSEs, namely α-helices, 3_{10}-helices, π-helices, extended β-strands and polyproline II (PPII) and other left-handed helices. The methodology was validated using an unbiased clustering of these parameters for a protein data set consisting of 1008 protein chains, which suggested that there are seven well defined clusters associated with different SSEs. Apart from α-helices and extended β-strands, 3_{10}-helices and π-helices were also found to occur in substantial numbers. *ASSP* was able to discriminate non-α-helical segments from flanking α-helices, which were often identified as part of α-helices by other algorithms. *ASSP* can also lead to the identification of novel SSEs. It is believed that *ASSP* could provide a better understanding of the finer nuances of protein secondary structure and could make an important contribution to the better understanding of comparatively less frequently occurring structural motifs. At the same time, it can contribute to the identification of novel SSEs. A standalone version of the program for the Linux as well as the Windows operating systems is freely downloadable and a web-server version is also available at http://nucleix.mbu.iisc.ernet.in/assp/index.php.

1. Introduction

The concept of repeating backbone torsion angles (φ, ψ) and patterns of main chain–main chain (MM) N—H$\cdots$O hydrogen bonds defining a regular secondary-structure element (SSE) in proteins is well established. α-Helices and extended β-strands, the major regular SSEs found in proteins, were first predicted by theoretical studies (Pauling *et al.*, 1951; Pauling & Corey, 1951) and were subsequently confirmed by X-ray diffraction analysis (Perutz, 1951; Blake *et al.*, 1965). Other SSEs such as 3_{10}-helices (Donohue, 1953), π-helices (Low & Grenville-Wells, 1953), polyproline II (PPII) helices (Cowan & McGavin, 1955) and left-handed α-helices, 3_{10}-helices and π-helices were also proposed from model-building studies and found to occur occasionally in proteins (Ramachandran & Sasisekharan, 1968). The information about these SSEs is used in a number of structural biology applications, such as structure comparison (Gibrat *et al.*, 1996), visualization (Sayle & Milner-White, 1995; Humphrey *et al.*, 1996) and classification (Murzin *et al.*, 1995; Orengo *et al.*, 1997). The formation of SSEs also plays a major role in protein folding (Murzin *et al.*, 1995; Orengo *et al.*, 1997). Hence, several methods have been proposed over the years to identify the SSEs in protein structures, which can be broadly classified into three



categories: (i) algorithms based on (φ, ψ) and/or hydrogen-bond patterns, (ii) algorithms based on three-dimensional geometry and (iii) hybrid methods which use both (i) and (ii). Programs such as *DSSP* (Kabsch & Sander, 1983), *STRIDE* (Frishman & Argos, 1995) and *PROSS* (Srinivasan & Rose, 1999) fall into the first category, while *DEFINE* (Richards & Kundrot, 1988), *P-CURVE* (Sklenar *et al.*, 1989), *P-SEA* (Labesse *et al.*, 1997) and *SST* (Konagurthu *et al.*, 2012) come under the second category and *KAKSI* (Martin *et al.*, 2005), *PALSSE* (Majumdar *et al.*, 2005) *etc.* fall into the third category. A few programs that specifically identify π-helices (Fodje & Al-Karadaghi, 2002) and PPII helices (King & Johnson, 1999; Srinivasan & Rose, 1999; Cubellis *et al.*, 2005; Mansiaux *et al.*, 2011) have also been developed. A brief description of various algorithms is provided in Supplementary Table S1. Most algorithms correctly identify the main bodies of the SSEs, but the identification of their termini varies considerably. Even in the main body, differences sometimes occur in the assignment because of small deviations from uniform helical character arising owing to solvent-induced distortions (Blundell *et al.*, 1983), peptide-bond distortions (Barlow & Thornton, 1988) or the presence of proline (MacArthur & Thornton, 1996; Chakrabarti *et al.*, 1986; Sankararamakrishnan & Vishveshwara, 1990), serine and threonine residues (Deupi *et al.*, 2004; Ballesteros *et al.*, 2000). The bias of the existing algorithms towards α-helices over non-α-helices, especially when these segments are interspersed between two α-helices, is also a persistent problem.

With the objective of providing accurate and reliable assignment of all regular SSEs, here we describe a method called *ASSP*, an extension of the in-house algorithm *HELANAL-Plus* (Kumar & Bansal, 1996, 1998; Bansal *et al.*, 2000; Kumar & Bansal, 2012), which uses the geometry of the path traversed by the C^{α} atoms in a protein chain and hence places it in category (ii) mentioned above. The protein structure is assumed to consist of uniform stretches interspersed with non-uniform segments. These uniform stretches are defined based on the local geometrical parameters, namely, twist, rise and virtual torsion angle (Vtor), calculated for each block of four consecutive C^{α} atoms and then classified into different SSEs depending on the average values of the local parameters. It is observed that the (φ, ψ) torsion angles for residues in a particular SSE, especially at the termini, are sometimes not in the expected range for that SSE. This leads to the peptide plane involving a terminal residue to tilt away from the helix axis, although the C^{α} atoms follow the helical path and hence provide a better definition for helical termini. The user-friendly web server, which includes visualization of the assignments and comparison with those by *DSSP* and *STRIDE*, should make *ASSP* very useful for structural biologists.

2. Methods

2.1. Preparation of training and testing data sets

A data set representing 929 folds with an ASTRAL-SPACI score of >0.4 was downloaded from the ASTRAL-1.75 database (Brenner *et al.*, 2000). 79 protein chains with resolutions of $\leq$2.5 Å and <30% sequence similarity were added to the data set for enrichment in π-helices. A total of 5465 α-helices, 2340 3_{10}-helices and 46 π-helices assigned by both *STRIDE* and *DSSP* were considered to determine the corresponding cutoff values for use in *ASSP*.

The performance of *ASSP* was tested and compared with other methods for four data sets (Hres, Mres, Lres and NMR) comprising of 689, 624, 332 and 296 protein chains, respectively (Martin *et al.*, 2005). For the analysis of π-helices, a data set of 85 protein chains containing π-helices (Fodje & Al-Karadaghi, 2002) was considered. The performance of *ASSP* was also evaluated using the structures of five randomly selected proteins solved using X-ray crystallography, NMR and electron microscopy (EM).

2.2. Working of *ASSP*

ASSP is an extension of the in-house program *HELANAL-Plus*, which uses the path traversed by the C^{α} atoms to define the geometry of already assigned helical segments in a protein chain by calculating the local geometric parameters following the method of Sugeta & Miyazawa (1967). The algorithm used in *ASSP* can be divided into four levels as follows.

(i) Calculation of local geometric parameters. The full length of the protein chain is scanned and geometric parameters for each block of four C^{α} atoms are calculated.

(ii) Identification of uniform stretches. A uniform stretch is defined if the absolute difference between the geometric

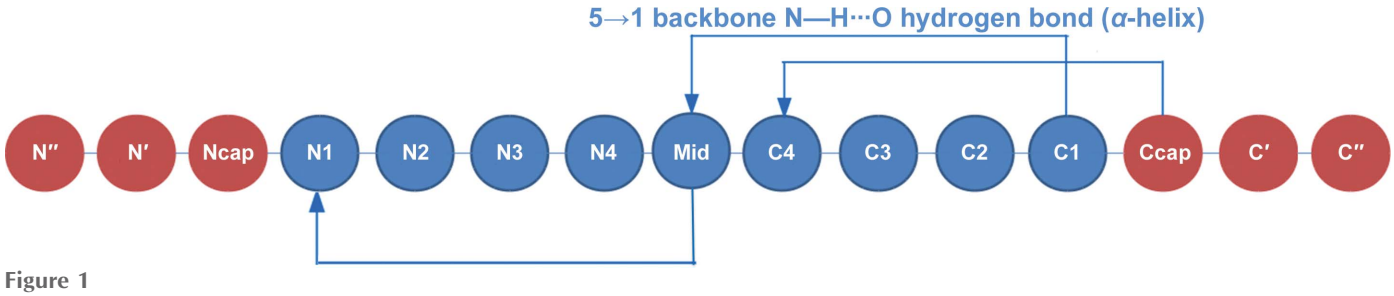


Figure 1
Nomenclature of various positions within and around an α-helix. Nine helical and six nonhelical positions along with characteristic 5→1 backbone N—H···O hydrogen bonds are shown. For an α-helix of more than nine amino-acid residues in length, there will be at least one residue for which both the amide and the carboxyl group are involved in backbone N—H···O hydrogen bonds. The 'Mid' position will have (N − 8) residues, where N is the length of the helix.

Table 1
Mean and standard deviation (std dev.) values of various local step geometrical parameters for *ASSP* assigned as right-handed π-helices, α-helices and 3_{10}-helices, left-handed PPII helices and extended β-strands in the training data set.

The number of SSEs in each category is given in parentheses in the first row.

	π-Helix (206)	α-Helix (6218)	3_{10}-Helix (1808)	PPII helix (824)	Extended β-strand (8717)
Parameter	Mean (std dev.)	Mean (std dev.)	Mean (std dev.)	Mean (std dev.)	Mean (std dev.)
Twist (°)	86.8 (7.4)	98.9 (3.8)	107.1 (4.9)	238.2 (8.6)	199.9 (18.4)
Rise (Å)	1.2 (0.3)	1.5 (0.2)	1.8 (0.2)	3.0 (0.1)	3.0 (0.4)
Vtor (°)	36.3 (10.4)	50.1 (6.4)	65.4 (9.0)	249.5 (10.2)	201.9 (21.3)
Radius (Å)	2.6 (0.2)	2.3 (0.1)	2.1 (0.1)	1.3 (0.1)	1.0 (0.2)
Length (residues)	5.6 (1.8)	12.6 (6.0)	3.6 (1.2)	3.5 (0.7)	4.3 (1.1)
φ (°)	−79.7 (23.2)	−65.3 (20.8)	−68.6 (25.2)	−77.6 (25.21)	−106.1 (34.0)
ψ (°)	−38.0 (25.0)	−43.7 (15.9)	−22.2 (19.9)	138.1 (31.52)	122.4 (51.8)

parameters for two consecutive steps is less than a defined value.

(iii) Classifying the uniform stretches. Uniform stretches are further classified into various types of SSEs. *ASSP* assigns right-handed and left-handed α-helices, 3_{10}-helices and π-helices as well as extended β-strands and left-handed PPII helices.

(iv) Final arrangement. The format of the final output is similar to the HELIX record of the PDB file. A colour-coded visual representation of the SSE assignments is also provided. The six nonhelical (N″, N′, Ncap, Ccap, C′ and C″) and nine helical (N1, N2, N3, N4, Mid, C4, C3, C2 and C1) positions within and around α-helices are shown in Fig. 1. A detailed description of the algorithm is given in the Supporting Information and in the algorithm section of the *ASSP* web server.

2.3. Calculation of the tilt angle

We define the tilt angle (δ_i) as the angle made by the vector along the backbone carboxyl group $C_i{=}O_i$ (CO_i) of the ith residue in a helix with the corresponding global helix axis (GHA). The direction cosines (l, m, n) for the GHA were calculated using *HELANAL-Plus* for α-helices of length >6 residues.

2.4. Programs used

The nonbonded interactions were calculated using *HBPLUS* (McDonald & Thornton, 1994) and *MolBridge* (Kumar *et al.*, 2014). Figures were generated using *PyMOL* v.1.3r1 (Schrödinger) and *MATLAB* v.7.10.0 (The MathWorks). The *k*-means algorithm available in *MATLAB* was used for clustering the twist and rise data.

3. Results and discussion

ASSP identifies right-handed and left-handed α-helices, 3_{10}-helices and π-helices, as well as left-handed PPII helices and extended β-strands, in a protein structure. A stretch that does not fall into any of these categories is assigned as unidentified. The mean and standard deviation values of the local geometrical parameters along with (φ, ψ) and the corresponding mean length of the SSEs identified by *ASSP* are given in Table 1. Plots of twist *versus* rise for various SSEs are shown in Supplementary Fig. S1.

3.1. Distribution of different types of SSEs

In the training data set, 6218, 5465 and 6092 α-helices were identified by *ASSP*, *DSSP* and *STRIDE*, respectively. The length of α-helices defined by *ASSP* spans from four to 80 residues, while for *STRIDE* it varies between one and 80 residues (Supplementary Fig. S2*a*). Interestingly, *STRIDE* assigned seven α-helices with a length of less than four residues. Often, longer α-helices assigned by *STRIDE* were divided into two or more helices by *ASSP*, thereby increasing the number of shorter helices ($\leq$10 residues) and hence leading to the smaller values for the median/mean length.

The number of 3_{10}-helices identified by *ASSP*, *DSSP* and *STRIDE* is 1808, 2312 and 2673, respectively (Supplementary Fig. S2*b*). The larger number of *STRIDE*-assigned 3_{10}-helices can be attributed to two factors: (i) *STRIDE* in some cases defines a protein segment as a 3_{10}-helix or a group of 3_{10}-helices when it is identified as an α-helix by *ASSP* and *DSSP* and (ii) many *STRIDE*-assigned 3_{10}-helices were found to be part of a non-uniform stretch by *ASSP*.

ASSP identified 206 π-helices, with the 18-residue fragment Met65–Leu82 of chain *A* from an oxidoreductase (PDB entry 1v54) being the longest. Both *DSSP* and *STRIDE* were found to be biased towards α-helices over π-helices and identified only 46 and 90 helices, respectively (Supplementary Fig. S2*c*). Our analysis indicates that π-helices are more abundant than generally believed, as suggested by an earlier study (Fodje & Al-Karadaghi, 2002).

The differences in the helix assignments are clearly illustrated (Supplementary Fig. S3) in the case of a hydrolase (PDB entry 1h4p chain *B*). Several α-helices according to *ASSP* and *DSSP* were assigned as 3_{10}-helices by *STRIDE*, despite MM 5→1 N—H···O hydrogen bonds being present and the average twist (98.3°) and average rise (1.5 Å) being very close to ideal α-helix values. The wrong assignment of the SSEs by *STRIDE* can be attributed to the miscalculation of (φ, ψ), with these values being positive. The (φ, ψ) values of

all of the residues of chain *A* were correctly calculated; the problem occurs only for chain *B*.

ASSP identified three, 11, zero and 824 left-handed α-helices, 3_{10}-helices, π-helices and PPII helices, respectively. The length distribution for PPII helices is also given in Supplementary Fig. S2(*d*). Although we found a few left-handed π-helices of length four residues, they do not satisfy our minimum-length criterion and hence are not listed in the final output.

3.2. Validation of the results

In order to show that any newly developed algorithm has wide applicability and provides acceptable results, it needs to be validated using a test data set. In addition to validating the *ASSP*-assigned SSEs by comparison with other algorithms, we adopted a novel method for this algorithm. This is based on the assumption that the local geometrical parameters of the steps constituting each of the regular SSEs must form a distinct cluster. Hence, the average twist and average rise of two contiguous steps (the total number of data points is 102 676) constituting all of the uniform stretches, irrespective of whether they have been assigned to any SSE or not, were selected and clustered. The *k*-means clustering algorithm indicated that these data points belonged to seven well defined groups (Fig. 2*a*). The mean and the standard deviation of the average twist and average rise for each cluster were obtained and compared with those assigned by *ASSP*. We found that the values for five of the clusters corresponded to those obtained for *ASSP*-assigned π-helices, α-helices, 3_{10}-helices, extended β-strands and PPII helices. Representative examples taken from each cluster are shown in Fig. 2(*b*). Interestingly, two additional small clusters are seen in Fig. 2(*a*). Cluster (iv) corresponds to a mean twist of ∼150° ($n = 2.4$) and a mean rise of 2.8 Å. Cluster (vii) encompasses the small number of left-handed π-helices, α-helices and 3_{10}-helices in the data set.

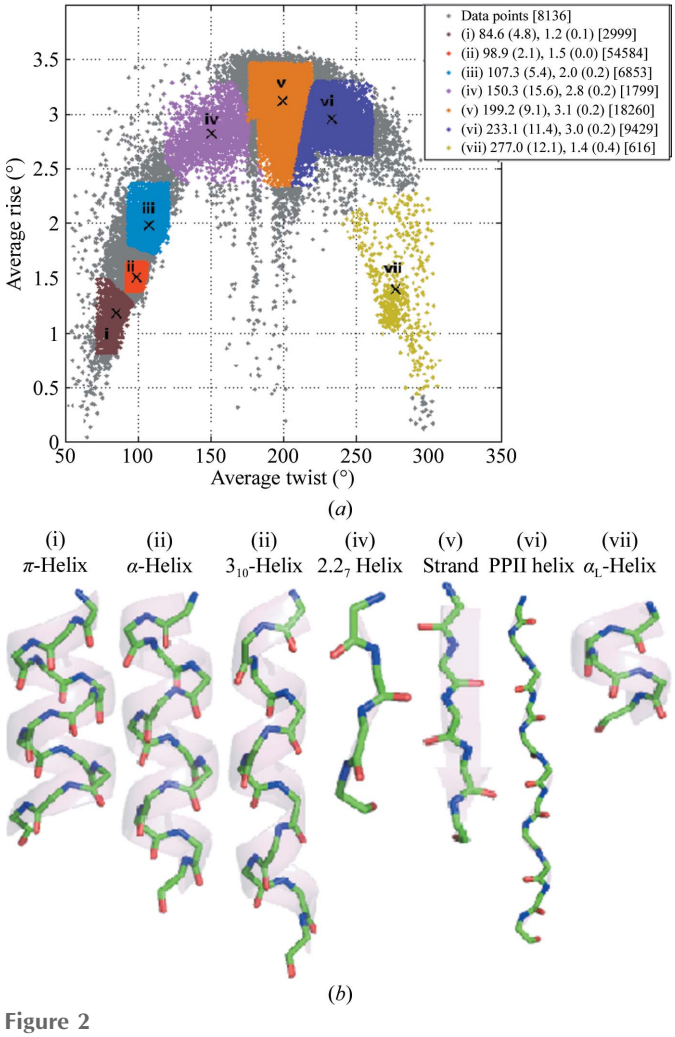


Figure 2
The clustered data of average twist *versus* average rise for two consecutive steps which form part of a uniform stretch in 1008 protein chains. The mean and standard deviation values along with the number of data points in each cluster are given in (*a*). Data points lying within 2σ deviation from the mean values for each cluster are shown in a different colour, while the remaining 8136 data points (that are not part of any cluster) are coloured grey. A representative example from each cluster is displayed in (*b*). PDB code and segment description: (i) 1x38 chain *A* (Gln265–Leu276), (ii) 2v8u chain *A* (Lys57–Thr68), (iii) 1szh chain *A* (Leu84–Gly94), (iv) 1jsd chain *A* (Ser84–Asn87), (v) 1h4p chain *B* (Leu129–Pro133), (vi) 3boi chain *A* (Gly15–Gly25), (vii) 1rcq chain *A* (Ala34–Gly38).

It should be mentioned that assigned π-helices, α-helices and 3_{10}-helices sometimes have overlapping values of twist and rise, while extended β-strands and PPII helices similarly share some space. In many instances, a lower value of twist or rise for one step is compensated by higher values for the next step, so that the mean values of the two steps lie in the corresponding SSE range. The observations mentioned above are exemplified by an example taken from methionyl-tRNA synthase (PDB entry 3h9c chain *A*). *ASSP* assigned the segment Val298–Phe308 as a π-helix, but the average twist (98.7°) and average rise (1.5 Å) for first two contiguous steps lie in the α-helical region. Since the first step has a helical character this cannot be excluded, and at the same time the first three residues cannot be assigned as α-helix because of the minimum-length criterion imposed during the assignment. Hence, these two steps were assigned as being part of a π-helix, since all subsequent steps had clear π-helix geometry. The number of data points in *ASSP*-assigned SSEs was found to be lower than the number of points in all of the clusters obtained by *k*-means clustering. This is owing to the minimum-length cutoff applied in the *ASSP* algorithm, as a consequence of which isolated π-helical and α-helical steps often do not appear in the final output. Only cluster (v) has a larger number of data points in *ASSP*-assigned extended β-strands than in the cluster. This is because a deviation of up to 2σ from the mean values is allowed for extended β-strand assignment, while for other SSEs it is only 1σ. The majority of the (φ, ψ) values for residues corresponding to these data points were found to lie in the characteristic region of the Ramachandran map expected for each SSE.

Cluster (iv) has not been associated by *ASSP* with any well characterized SSE. The mean value of average twist and average rise for this cluster were found to be 150.3° and 2.8 Å, respectively, corresponding to 2.4 residues per turn. These values are very close to the 2.27 residues per turn structure

first proposed by Donohue (1953). Although the spread of (φ, ψ) values was found to be greater compared with those of residues in other clusters, many residues were found to have (φ, ψ) values near the suggested values for a 2.2_7 helix (Porter & Rose, 2011). The structural and functional importance of such helices is not clear owing to their rare occurrence. Out of 1799 uniform stretches or data points, 242 were found to have at least one MM $i + 2 \rightarrow i$ N—H···O hydrogen bond. These uniform stretches were manually searched in the corresponding protein structure and found to be present either in proteins with very little SSE content or in all-β proteins. At the same time, we often found another segment almost parallel to such identified uniform stretches (Supplementary Fig. S4), suggesting that these structures could be intermediates between either extended β-strands or PPII helices and turns. Detailed analysis of cluster (iv) could lead to a more reliable identification of 2.2_7 helices.

The identified SSEs were also confirmed based on the hydrogen-bond patterns as well as the (φ, ψ) values of the constituent residues (Supplementary Fig. S5). Helices were searched for characteristic MM N—H···O hydrogen-bond patterns and the results were found to be in concordance with the assignment. It was found that *DSSP*-assigned and *STRIDE*-assigned α-helices have several residues which are involved in a stronger $6 \rightarrow 1$ MM N—H···O H-bond rather than the expected $5 \rightarrow 1$ N—H···O hydrogen bond, thus confirming their bias towards α-helices over other helices. It was found that *ASSP*-assigned helices almost always have the characteristic types of MM N—H···O hydrogen-bond patterns for π-helices, α-helices and 3_{10}-helices.

3.3. Unusual backbone torsion angle at the C-termini

A majority of the (φ, ψ) values in assigned helices have values characteristic of π-helices, α-helices, 3_{10}-helices or PPII helices. However, for 340 α-helices, of which 244 have a length of greater than six residues, we observed a large deviation of the torsion angle ψ for residues at the C1 position ($C1_\psi > 40^\circ$), although the parameters calculated using C^α atoms lie in the helical range. The deviation in ψ leads to the peptide plane being considerably tilted away from its normal orientation in an α-helix. The tilt angle (δ) for all of the α-helical residues at the C-terminus (C4–C1) was calculated and plotted against the respective ψ value (Supplementary Fig. S6*a*). It was observed that in 244 helices in which the C1 residues have large positive values of ψ, δ ranges between 90 and 150°, while for 4402 α-helices with $C1_\psi$ in the range −120 to 40°, δ is between 0 and 90°. Out of 244 helices, 126 (51.6%) were found to have proline at the Ccap position. The C^α atom of a proline at Ccap forms a $5 \rightarrow 1$ CA—HA···O bond to the C=O at the C4 position in 67 (53.17%) of 126 α-helices. Representative examples illustrating this feature are shown in Supplementary Figs. S6(*b*) and S6(*c*) for α-helices.

3.4. Comparison of the SSEs assigned by various algorithms

α-Helices assigned by *ASSP* and other available algorithms, namely *DSSP*, *STRIDE*, *XTLSSTR*, *SST*, *KAKSI* and *PALSSE*, were compared with each other and the residue-wise percentage agreements are listed in Table 2. With *ASSP* as a reference, the percentage agreement varies from 85.4% (for *XTLSSTR*) to 99.1% (for *PALSSE*). The high percentage of agreement seen between *ASSP* and *KAKSI* or *PALSSE* is because both these programs define all-helical fragments as being α-helices and also extend the helices, leading to much larger numbers of residues being selected. As expected, *DSSP* and *STRIDE* showed high agreement with each other owing to their similar algorithms. *ASSP* assigned 90.4 and 90.2% of the residues of α-helices identified by *DSSP* and *STRIDE*, respectively, while *DSSP* and *STRIDE* identified 93.8 and 92.5% of *ASSP*-assigned α-helical residues, respectively. The difference in the number of residues can be attributed to the bias of these algorithms towards α-helices over other types of helices. The differences between the assignments, especially at the termini, by *ASSP* and *STRIDE* was analysed extensively in a previous study (Shelar *et al.*, 2013). Differences in the assignments according to various algorithms are highlighted in Fig. 3 for residues Thr87–Leu134 of an oxidoreductase (PDB entry 1syy chain A). *STRIDE* assigns a

Table 2
Comparison of the assignment of α-helices at the residue level between *ASSP* and six other algorithms.

Each cell gives the percentage agreement at the residue level between the pair of algorithms given in the first row and the first column. The algorithm in the column header was taken as a reference and the number of residues assigned to α-helices by the algorithm is given in parentheses.

	ASSP (69247)	*DSSP* (69936)	*STRIDE* (69153	*XTLSSTR* (69519)	*SST* (71675)	*KAKSI* (82644)	*PALSSE* (104772)
ASSP	—	90.4	90.2	85.8	83.3	80.2	63.7
DSSP	93.8	—	93.9	88.3	85.2	81.9	66.1
STRIDE	92.5	92.9	—	85.2	83.7	80.8	65.2
XTLSSTR	85.4	87.2	85.1	—	79.8	77.8	64.1
SST	89.6	87.3	86.8	82.9	—	81.5	67.6
KAKSI	96.4	96.8	83.4	93.1	94.1	—	78.6
PALSSE	99.1	99.1	98.9	97.3	98.1	78.1	—

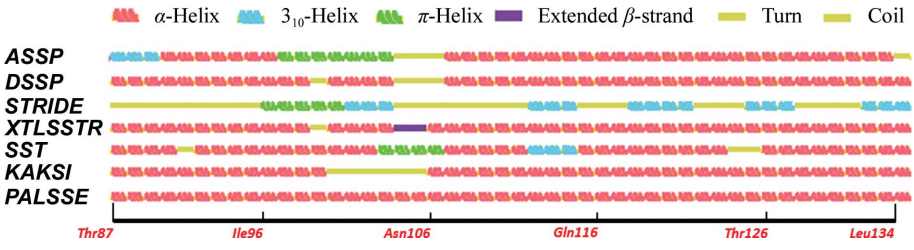


Figure 3
Pictorial representation comparing the secondary structures assigned by different algorithms. Amino-acid residues Thr87–Leu134 of an oxidoreductase (PDB entry 1syy chain *A*) were taken as an example.

Table 3
Agreement between the secondary-structure assignments by *ASSP* for the same protein structure solved by different methods or at different resolutions.

The structures in the Hres column correspond to the highest resolutions. The number of residues common to all data sets is indicated in column 4. Agreement is expressed as percentage of residues.

Stucture	Hres	Resolution (Å)	No. of residues	Lres	Resolution (Å)	Agreement	NMR	Agreement	EM	Agreement
1	1mms chain *A*	2.57	133	1giy chain *L*	5.5	100	1oln chain *A*	100	1eg0 chain *K*	100
2	1ya7 chain *A*	2.3	221	1pma chain *D*	3.4	84.7	2ku1 chain *A*	84.7	3c91 chain *D*	81.9
3	4j9z chain *R*	1.66	141	1a29 chain *A*	2.74	86.5	1cfc chain *D*	94.6	3j41 chain *E*	82.4
4	132l chain *A*	1.8	129	2znw chain *Y*	2.71	96.2	1e8l chain *A*	92.3	4a8a chain *M*	88.5
5	3h47 chain *A*	1.9	175	3p0a chain *A*	5.95	97.5	2lf4 chain *A*	92.6	1vu4 chain 0	92.6

shorter π-helix in the same region as *ASSP*, but the helical segments identified as α-helices by other algorithms are assigned as small 3_{10}-helices. SST also identified a π-helix, but the position was found to be shifted right compared with that assigned by *ASSP* and *STRIDE*. *XTLSSTR* assigned extended β-strand to Glu120-Ala121, whereas the same segment remained unassigned by *ASSP* and *DSSP*, while *STRIDE* assigned a turn. Not surprisingly, *PALSSE* assigned the whole segment (Thr87–Leu134) as α-helix, while *KAKSI* identified two α-helices (Thr87–Ala99 and Asn106–Leu134). The assignments were checked for the presence of the corresponding MM N—H$\cdots$O hydrogen bonds and (φ, ψ) of the constituting residues and were found to be in accordance with the *ASSP* assignment.

The sensitivity of *ASSP* and other algorithms was also judged based on the percentage residue content of various SSEs in the Hres, Mres, Lres and NMR data sets and the results are listed in Supplementary Table S2. *ASSP* identified the highest percentage of residues in π-helix in all data sets. The residue-wise comparison between different algorithms in the Hres data set follows a similar trend to that observed in the training data set (data not shown).

ASSP was also tested using a data set that was considered earlier for analysis of π-helices (Fodje & Al-Karadaghi, 2002). *ASSP* identified 102 π-helices (540 residues), whereas 104 (550 residues) were reported by the authors. However, only 80% of the residues were common to the two sets of π-helices. Some of the originally assigned π-helices were identified as part of a non-uniform stretch by *ASSP*, while 16 new π-helices were picked up. For the same data set, *DSSP*, *SST* and *STRIDE* identified only nine, 25 and six π-helices, respectively.

The PPII helices assigned by *ASSP* ($\mathrm{PPII_{ASSP}}$) were compared with those assigned by *DSSP-PPII*, *PROSS*, *SEGNO* and *XTLSSTR* on the training data set and a maximum of 70% agreement (*XTLSSTR*) was observed. The means and standard deviations of the PPII helices assigned by different methods are listed in Supplementary Table S3, while twist *versus* rise plots for methods other than *ASSP* are given in Supplementary Fig. S7. For algorithms other than *ASSP*, many of the residues from N2 to C2 were found to have (φ, ψ) values in the α-helical region of the Ramachandran map, indicating a wrong assignment. Although the (φ, ψ) values of residues were not used as a criterion for SSE assignment, the segments identified as PPII by *ASSP* were not identified as being part of a non-uniform stretch. It has been shown earlier that PPII helices assigned by one algorithm generally have some overlap with the extended β-strands assigned by either *DSSP* or *STRIDE* (Carter *et al.*, 2003). Keeping this in mind, $\mathrm{PPII_{ASSP}}$ were compared with the extended β-strands identified by *STRIDE* and out of 2787 residues, 300 were found to be part of extended β-strands. The negative value of $\mathrm{C1}_{\psi}$ for 24 out of 824 $\mathrm{PPII_{ASSP}}$ helices can be attributed to the algorithm using only C^{α} atoms.

The consistency of *ASSP* was also shown using the coordinates of the same protein solved by different experimental methods and at different resolutions (Table 3). The difference of >10% in agreement is owing to actual differences in structure, rather than a shortcoming of the algorithm.

A total of 25 and 93 π-helices were found at the N-termini and C-termini of α-helices, respectively, whereas 36 π-helices were sandwiched between two α-helices. Similarly, 354 and 559 3_{10}-helices were found at the N-termini and C-termini of α-helices, respectively, while 92 were interspersed between two α-helices. The significant difference in identifying a non-α-helical segment flanked by α-helices is illustrated by chain *A* of a ribonucleotide-diphosphate reductase protein (PDB entry 3ee4). *STRIDE* assigned α-helix to Pro152–Ile165 and Met169–Val183, while *ASSP* identified an interspersed π-helix (Val160–Ala171) with two flanking α-helices (Pro152–Ser159 and Leu172–Arg185) (Fig. 4*a*). Although a π-helix (Asn163–Glu167) was identified by *DSSP*, its length was found to be different. Similarly, on many occasions an interspersed 3_{10}-helix has been identified as part of an α-helix. In human carnitine acetyltransferase (PDB entry 1nm8 chain *A*), *ASSP* identified the segment Phe128–Val130 as a 3_{10}-helix interspersed between two α-helices (Leu111–Asp127 and Met131–Asn134), while *DSSP* and *STRIDE* assigned the whole segment (Leu111–Asn134) as α-helix (Fig. 4*b*). The MM N—H$\cdots$O hydrogen-bond patterns confirmed the presence of a 3_{10}-helix, reaffirming the *ASSP* assignment as being more accurate, especially when it comes to identifying the local deformations in the helical segments.

Compared with those from *DSSP* and those mentioned in PDB files, a larger number of *STRIDE*-assigned α-helices have a kinked geometry (Kumar & Bansal, 2012). The kinked region often creates a discontinuity, and hence *ASSP* either divides the helix into more than one part or truncates it. For example, a maximum bending angle of 72.1° was reported at residue Trp21 by *HELANAL-Plus* for the *STRIDE*-assigned α-helix Ala4–Ala35 in monomeric haemoglobin I (PDB entry

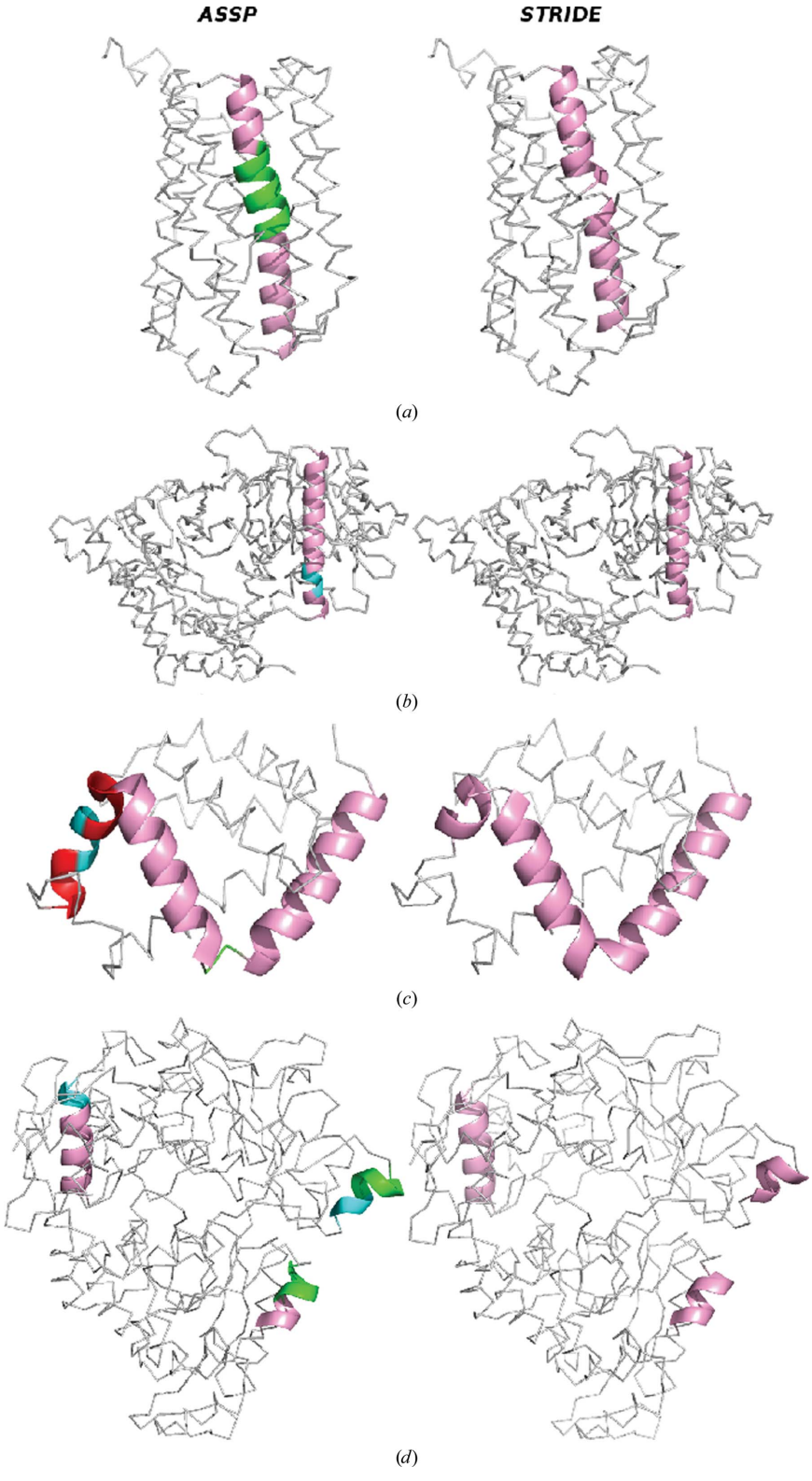


Figure 4
Examples of differences in assignment between *ASSP* and *STRIDE*. The representative segment is shown as a cartoon, while the rest of the structure is shown as a grey ribbon. α-Helices are coloured pink and red, while 3_{10}-helices and π-helices are shown in cyan and green, respectively. Two different colours are used to differentiate between neighbouring α-helices. (*a*) Oxidoreductase (PDB entry 3ee4 chain *A*); (*b*) human carnitine acetyltransferase (PDB entry 1nm8 chain *A*); (*c*) haemoglobin I (PDB entry 1b0b chain *A*) and (*d*) sugar-binding protein (PDB entry 3i5o chain *A*). A comparison of SSE assignment by *ASSP* and *STRIDE* for full-length proteins is shown in Supplementary Fig. S8.

1b0b chain *A*) and hence it was classified as kinked. The same segment was divided by *ASSP* into two α-helices Ala4–Ala20 and Thr23–His36, with *HELANAL-Plus* classifying them as a curved and a linear helix, respectively. It was also observed that *ASSP* assigned Asp37–Phe40, Ala41–Phe43 and Ser44–Phe47 as α-helix, 3_{10}-helix and α-helix, respectively, while *STRIDE* designated residues Asp37–Lys42, Phe43–Leu46 and Ser44–Phe47 as α-helix, turn I and turn IV, respectively (Fig. 4*c*).

A relaxation in the backbone torsion angles (φ, ψ) at C1 of a 3_{10}-helix at the N-terminus of an α-helix produces a kink at the interface (Pal *et al.*, 2003). In integrin (PDB entry 1aox chain *A*), Glu318–Glu323 was identified as a 3_{10}-helix by *ASSP*, which is followed by an α-helix (Lys324–Ser334). *HELANAL-Plus* assigns a maximum bending angle of 32.3° at Ala325 and hence a kinked geometry to the helical segment encompassing both a 3_{10}-helix and α-helix (Glu318–Ser334). *STRIDE* divides the 3_{10}-helix into two 3_{10}-helices (Glu318–Ala320 and Leu322–Lys324), while *DSSP* assigns Glu318–Lys324 as a 3_{10}-helix. Although all three algorithms were able to identify the 3_{10}-helical region in chain *A*, only *ASSP* assigned a 3_{10}-helix to the same fragment of chain *B*, whereas *DSSP* and *STRIDE* assigned an α-helix to the segment Lys324–Ser334.

In a sugar-binding protein (PDB entry 3i5o chain *A*), *ASSP* identified a 3_{10}-helix (Ala85–Asp87) at the N-terminus of an α-helix (Phe88–Leu98) and a π-helix (Leu295–Tyr299) at the C-terminus of an α-helix (Tyr291–Arg294). Moreover, *ASSP* was also able to identify a helical segment comprised of only a 3_{10}-helix (Val238–Glu240) and a π-helix (Leu241–Lys245). Neither *DSSP* nor *STRIDE* identified these local structural changes (Fig. 4*d*).

Compared with their right-handed counterparts, left-handed α-helices, 3_{10}-helices and π-helices are far less abundant and the existing automated algorithms do not identify them. The performance of *ASSP* was checked with the help of previously identified left-handed helices from Novotny & Kleywegt (2005), where the authors discussed the structural and functional importance of 31 such helices. *ASSP* identified one, 17 and 11 left-handed π-helices, α-helices and 3_{10}-helices, respectively, against the reported 11 α-helices and 20

3_{10}-helices (Supplementary Table S4). Two helices which were part of a non-uniform stretch were not identified by *ASSP*. Out of 20 reported 3_{10}-helices, six were identified as α-helices and one as a π-helix. The average twist and average rise for the protein segment Gly77–Asp81 in split-Soret cytochrome *c* (PDB entry 1h21 chain *A*) was found to be −82.0° and 0.97 Å, respectively, with 4.39 residues per turn. The MM hydrogen-bond analysis showed the presence of a 6→1 N—H···O hydrogen bond. Another example taken from a copper-nitrate reductase protein (PDB entry 1nif chain *A*), previously identified as a 3_{10}-helix (Ala105–Gly108), was identified as an α-helix by *ASSP*. The mean twist and mean rise were found to be −101.16° and 1.59 Å, respectively. The 5→1 and 4→1 MM N—H···O hydrogen bonds were of comparable strength in terms of angle and distance, with the 4→1 hydrogen bond being stronger. We observe trifurcated (4→1, 5→1, 6→1) MM N—H···O hydrogen bonds of comparable strength for the residues at the N1 position in most of the cases.

In general, comparison of *ASSP* with other available algorithms shows that the local variations in the protein structure can be better identified with the help of C^{α} atoms alone.

3.5. The *ASSP* web server

ASSP is also available in a web-mounted format and can be accessed freely at http://nucleix.mbu.iisc.ernet.in/assp/index.html. Linux/Unix and Windows compatible standalone executables are also available for download. The user has the option of either providing the four-character PDB code or uploading a `.pdb` file containing the three-dimensional

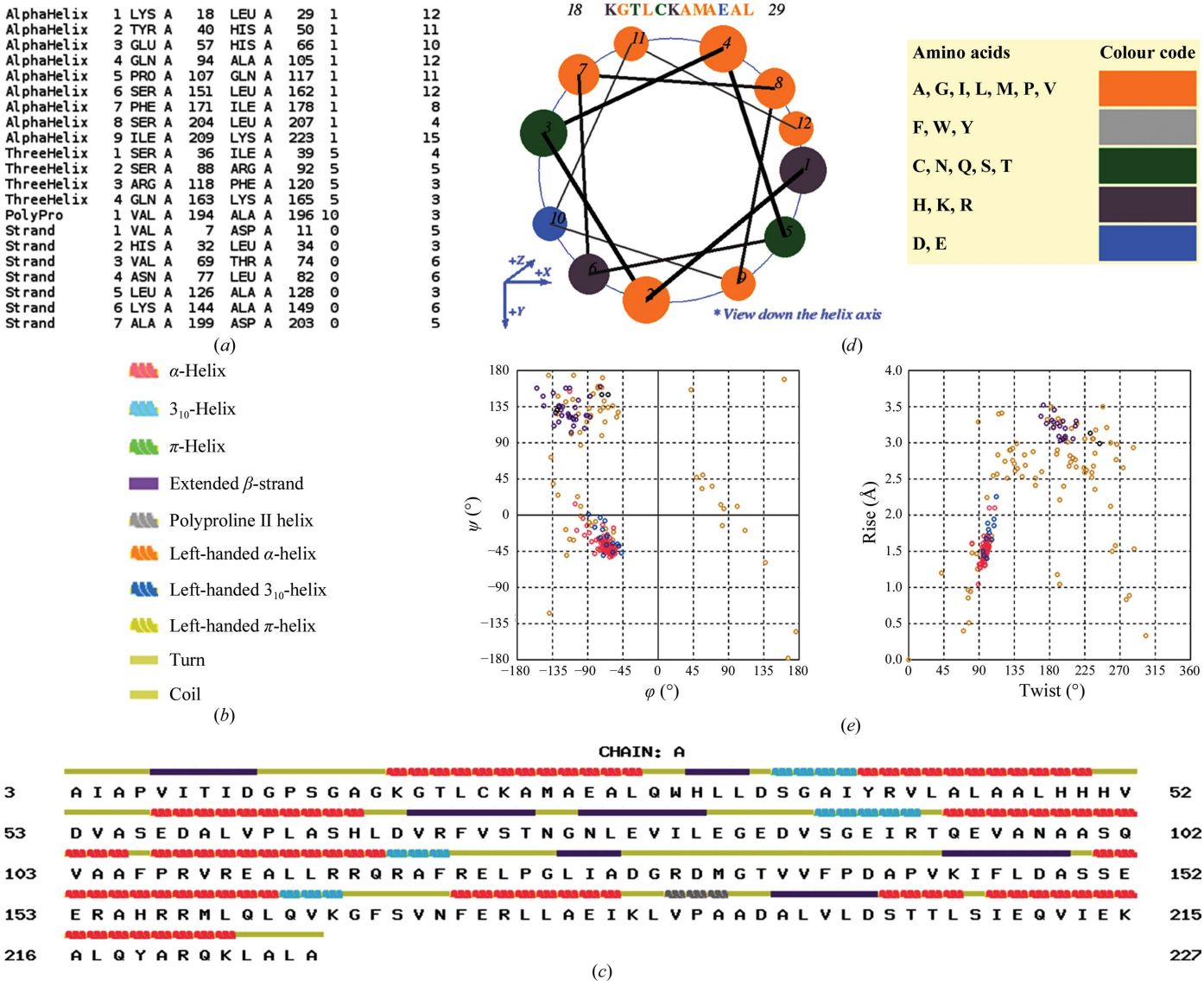


Figure 5
Different outputs from the *ASSP* webserver for the CMP kinase protein (PDB entry 1cke chain *A*). (*a*) Text representation of the SSE assignment performed by *ASSP*; (*b*) images assigned to each structural state; (*c*) 'wiring diagram' representation of the *ASSP* assignment; (*d*) helical wheel representation of the α-helix Lys18–Leu29 and colours assigned to the amino-acid residues (single-letter codes); (*e*) φ–ψ map for the residues and twist *versus* rise plot.

coordinates of the protein structure in Brookhaven PDB file format. The uploaded file name will be shortened to 11 characters (including the extension) if this is not already the case, and hyphens or underscores will be removed if present. The web server has an added advantage over the stand-alone program because it provides additional options (Fig. 5). A successful run of *ASSP* will lead to a page providing a link to the results page consisting of graphical and textual representation of the assignments. Additional options available on the results page are (i) running *HELANAL-Plus* for the *ASSP*-assigned SSEs, (ii) generating a helical wheel for the SSEs or any defined protein segment and (iii) downloading the output files. If a PDB code is provided, the header section will also be displayed, which includes general information about the structure [author, compound, resolution (if applicable) *etc.*].

The graphical representation page contains a cartoon representation of SSEs along with the corresponding amino-acid residues, similar to the 'wiring diagram' of PDBsum (Laskowski *et al.*, 1997) or the *STRIDE* web server (Heinig & Frishman, 2004). The representation is generated as an image by parsing the *ASSP*, *DSSP* and *STRIDE* SSE assignments and an image is assigned to each structural state, which can be downloaded by right-clicking on it. The corresponding input protein will be displayed as a *Jmol* applet in ribbon representation and residues will be coloured according to the colour of the image assigned to each structural state as mentioned at the top of the page. Users are provided with an option for comparing the SSE assignments of *ASSP* with those from *DSSP* or *STRIDE*. At the same time, a φ–ψ plot and a twist *versus* rise plot for each chain can be visualized by clicking on the appropriate button given at the top right. A detailed description of the residue (SSE assignment by *ASSP*, φ–ψ plot *etc.*) can also be found by letting the cursor hover over the corresponding wire diagram.

The textual representation page displays the *ASSP* assignments in text format. If the PDB identifier is given as input, then an external link will be provided leading to the corresponding entry in different databases such as PDB, SCOP, CATH, PDBsum and PDBe (Gutmanas *et al.*, 2014). The graphical representation page can also be accessed from the textual representation page and *vice versa* by clicking on the appropriate button at the top left.

Detailed helical parameters and geometry of the *ASSP*-assigned helices can be obtained by clicking on the 'HELANAL-Plus' button. An option is also provided for generating the 'helical wheel' (Schiffer & Edmundson, 1967) for any selected SSE or segment of the input protein structure. The helical wheel is generated by using the local twist values with the first C^{α} corresponding to 0° twist; subsequent C^{α} atoms are placed relative to it when looking down the helix axis. A user manual listing the various available options and their usage is provided in the help section.

The current version of *ASSP* is available as a prebuilt binary for Linux/Unix as well as Windows and can be downloaded from the download section. The downloaded program should be kept in the working directory along with the protein files; there are no any other prerequisites for its installation.

4. Conclusions

In addition to the use of (φ, ψ) torsion angles, regions of regular SSE in a protein chain can also be identified based on local geometric parameters calculated using the path traversed by the C^{α} atoms. In general, the α-helices identified by *ASSP* match those assigned by *DSSP* and *STRIDE*. In addition, *ASSP* was found to identify 3_{10}-helical and π-helical segments that were assigned as part of an α-helix by other algorithms. Hence, it can provide a better understanding of the finer nuances of helices in proteins. The identification of left-handed α-helices, 3_{10}-helices, π-helices and PPII helices, along with the other commonly observed SSEs, makes *ASSP* more versatile. The unassigned uniform stretches could lead to the identification of new SSEs. The scope of *ASSP* usage will grow with the increase in the number of low-resolution structures, structures solved by EM or NMR or structures with only C^{α} atoms. We believe that *ASSP* could make an important contribution towards a better understanding of comparatively less frequently occurring structural motifs and their sequence specificity, which will lead to a better understanding of their role in protein function.

5. Related literature

The following references are cited in the Supporting Information for this article: Levitt & Greer (1977) and Taylor (2001).

Acknowledgements

PK is a DBT-BINC, India fellow. MB is the recipient of a J. C. Bose National Fellowship of the Department of Science and Technology, Government of India. We thank Professor N. V. Joshi for helping in the cluster analysis of twist and rise data.

References

Ballesteros, J. A., Deupi, X., Olivella, M., Haaksma, E. E. & Pardo, L. (2000). *Biophys. J.* **79**, 2754–2760.
Bansal, M., Kumart, S. & Velavan, R. (2000). *J. Biomol. Struct. Dyn.* **17**, 811–819.
Barlow, D. J. & Thornton, J. M. (1988). *J. Mol. Biol.* **201**, 601–619.
Blake, C. C. F., Koenig, D. F., Mair, G. A., North, A. C. T., Phillips, D. C. & Sarma, V. R. (1965). *Nature (London)*, **206**, 757–761.
Blundell, T., Barlow, D., Borkakoti, N. & Thornton, J. (1983). *Nature (London)*, **306**, 281–283.
Brenner, S. E., Koehl, P. & Levitt, M. (2000). *Nucleic Acids Res.* **28**, 254–256.
Carter, P., Andersen, C. A. & Rost, B. (2003). *Nucleic Acids Res.* **31**, 3293–3295.
Chakrabarti, P., Bernard, M. & Rees, D. C. (1986). *Biopolymers*, **25**, 1087–1093.
Cowan, P. M. & McGavin, S. (1955). *Nature (London)*, **176**, 501–503.
Cubellis, M. V., Cailliez, F. & Lovell, S. C. (2005). *BMC Bioinformatics*, **6**, Suppl. 4, S8.
Deupi, X., Olivella, M., Govaerts, C., Ballesteros, J. A., Campillo, M. & Pardo, L. (2004). *Biophys. J.* **86**, 105–115.
Donohue, J. (1953). *Proc. Natl Acad. Sci. USA*, **39**, 470–478.

Fodje, M. N. & Al-Karadaghi, S. (2002). *Protein Eng.* **15**, 353–358.
Frishman, D. & Argos, P. (1995). *Proteins*, **23**, 566–579.
Gibrat, J.-F., Madej, T. & Bryant, S. H. (1996). *Curr. Opin. Struct. Biol.* **6**, 377–385.
Gutmanas, A. *et al.* (2014). *Nucleic Acids Res.* **42**, D285–D291.
Heinig, M. & Frishman, D. (2004). *Nucleic Acids Res.* **32**, W500–W502.
Humphrey, W., Dalke, A. & Schulten, K. (1996). *J. Mol. Graph.* **14**, 33–38.
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–i105.
Kumar, P. & Bansal, M. (2012). *J. Biomol. Struct. Dyn.* **30**, 773–783.
Kumar, P., Kailasam, S., Chakraborty, S. & Bansal, M. (2014). *J. Appl. Cryst.* **47**, 1772–1776.
Kumar, S. & Bansal, M. (1996). *Biophys. J.* **71**, 1574–1586.
Kumar, S. & Bansal, M. (1998). *Biophys. J.* **75**, 1935–1944.
Labesse, G., Colloc'h, N., Pothier, J. & Mornon, J.-P. (1997). *Comput. Appl. Biosci.* **13**, 291–295.
Laskowski, R. A., Hutchinson, E. G., Michie, A. D., Wallace, A. C., Jones, M. L. & Thornton, J. M. (1997). *Trends Biochem. Sci.* **22**, 488–490.
Levitt, M. & Greer, J. (1977). *J. Mol. Biol.* **114**, 181–239.
Low, B. W. & Grenville-Wells, H. J. (1953). *Proc. Natl Acad. Sci. USA*, **39**, 785–801.
MacArthur, M. W. & Thornton, J. M. (1996). *J. Mol. Biol.* **264**, 1180–1195.
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.
McDonald, I. K. & Thornton, J. M. (1994). *J. Mol. Biol.* **238**, 777–793.
Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). *J. Mol. Biol.* **247**, 536–540.
Novotny, M. & Kleywegt, G. J. (2005). *J. Mol. Biol.* **347**, 231–241.
Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. M. (1997). *Structure*, **5**, 1093–1108.
Pal, L., Chakrabarti, P. & Basu, G. (2003). *J. Mol. Biol.* **326**, 273–291.
Pauling, L. & Corey, R. B. (1951). *Proc. Natl Acad. Sci. USA*, **37**, 251–256.
Pauling, L., Corey, R. B. & Branson, H. R. (1951). *Proc. Natl Acad. Sci. USA*, **37**, 205–211.
Perutz, M. F. (1951). *Nature (London)*, **167**, 1053–1054.
Porter, L. L. & Rose, G. D. (2011). *Proc. Natl Acad. Sci. USA*, **108**, 109–113.
Ramachandran, G. N. & Sasisekharan, V. (1968). *Adv. Protein Chem.* **23**, 283–438.
Richards, F. M. & Kundrot, C. E. (1988). *Proteins*, **3**, 71–84.
Sankararamakrishnan, R. & Vishveshwara, S. (1990). *Biopolymers*, **30**, 287–298.
Sayle, R. A. & Milner-White, E. J. (1995). *Trends Biochem. Sci.* **20**, 374–376.
Schiffer, M. & Edmundson, A. B. (1967). *Biophys. J.* **7**, 121–135.
Shelar, A., Kumar, P. & Bansal, M. (2013). *Biomolecular Forms and Functions*, edited by M. Bansal & N. Srinivasan, pp. 116–127. Bangalore: World Scientific. doi:10.1142/9789814449144_0009.
Sklenar, H., Etchebest, C. & Lavery, R. (1989). *Proteins*, **6**, 46–60.
Srinivasan, R. & Rose, G. D. (1999). *Proc. Natl Acad. Sci. USA*, **96**, 14258–14263.
Sugeta, H. & Miyazawa, T. (1967). *Biopolymers*, **5**, 673–679.
Taylor, W. R. (2001). *J. Mol. Biol.* **310**, 1135–1150.