Subject Section

SCOT: Rethinking the Classification of Secondary Structure Elements

Tobias Brinkjost [1,2,*,†], Christiane Ehrt [1,2,†], Oliver Koch [2] and Petra Mutzel [3]

[1] Department of Computer Science, TU Dortmund University, Dortmund, Germany
[2] Faculty of Chemistry and Chemical Biology, TU Dortmund University, Dortmund, Germany
[3] Institute of Computer Science, University of Bonn, Bonn, Germany
This work was done while the author was a member of the Computer Science Department of TU Dortmund University

[*] To whom correspondence should be addressed.
[†] These authors contributed equally to this work.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Secondary structure classification is one of the most important issues in structure-based analyses due to its impact on secondary structure prediction, structural alignment, and protein visualization. There are still open challenges concerning helix and sheet assignments which are currently not addressed by a single multi-purpose software.

Results: We introduce SCOT (Secondary structure Classification On Turns) as a novel secondary structure element assignment software which supports the assignment of turns, right-handed α-, 3_{10}-, and π-helices, left-handed α- and 3_{10}-helices, 2.2_7- and polyproline II helices, β-sheets, and kinks. We demonstrate that the introduction of helix Purity values enables a clear differentiation between helix classes.

SCOT's unique strengths are highlighted by comparing it to six state-of-the-art methods (DSSP, STRIDE, ASSP, SEGNO, DISICL, and SHAFT). The assignment approaches were compared concerning geometric consistency, protein structure quality and flexibility dependency, and their impact on secondary structure element-based structural alignments. We show that only SCOT's combination of hydrogen bonds, geometric criteria, and dihedral angles enables robust assignments independent of to structure quality and flexibility. We demonstrate that this combination and the elaborate kink detection lead to SCOT's clear superiority for protein alignments. As the resulting helices and strands are provided in a PDB conform output format, they can immediately be used for structure alignment algorithms.

Taken together, the application of our new method and the straight-forward visualization using the accompanying PyMOL scripts enable the comprehensive analysis of regular backbone geometries in proteins.

Availability: https://this-group.rocks

Contact: tobias.brinkjost@tu.dortmund.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The impact of automated secondary structure assignment methods (SSAMs) reaches from secondary structure prediction to secondary structure-based protein alignment to the assignment of protein domains. Additionally, the secondary structure element (SSE) information plays a key role in the visualization of protein structures. The high number of published algorithms for the classification of helices, strands, and turns in proteins points toward the most challenging issue: we cannot strive for the correct answer (Tyagi *et al.*, 2009). Despite the overwhelming number of available SSAMs, DSSP (Kabsch and Sander, 1983) and STRIDE (Frishman and Argos, 1995) are the commonly used methods (Supplementary Table S1). DSSP is still actively developed and one of the predominantly applied methodologies for many structure-based approaches. Comparing the SSAMs' citation counts in Web of Science v.5.30 (https://apps.webofknowledge.com), a clear superiority in citation count per year of both methods is evident. Various modified versions exist as part of visualization tools (Pettersen *et al.*, 2004; Schrödinger, LLC, 2015).

So, what are the criteria that prompt scientists to prefer one tool over another? The definition of termini, kinks, and the impact on SSE-based protein alignments are only a few. Different approaches have been developed to tackle some of these problems. We set out to combine different aspects of SSE assignments to address many challenges by a single method. These aspects are covered by dedicated results sections concerning the assignment of right- and left-handed helices including rare helix classes, extended conformations, structure quality and flexibility, and the SSE-based alignment quality.

All published SSAMs can be divided based on the data utilized for the assignment: hydrogen-bonding patterns, dihedral angles, and/or geometric properties.

While purely hydrogen bond-based methods identify less regular geometrical stretches of backbone conformations and do not apply to SSEs that are not stabilized by main-chain hydrogen bonds, geometry- and dihedral angle-based methods overestimate the stability of backbone conformations. Our method SCOT (Secondary structure Classification On Turns) was developed to find an acceptable compromise between these approaches.

Repetitive stretches of hydrogen-bonded and open turns of different lengths, which are classified based on their dihedral angles and geometric distance criteria, were used to develop a novel alternative for a reliable and consistent assignment of SSEs. Inspired by the work of previous work (Koch and Cole, 2011), we incorporate the knowledge of hydrogen-bonded turns, non-hydrogen-bonded turns, and geometry to assign helices and strands and identify SSE irregularities.

Our results point toward the unique features of SCOT as multi-purpose SSAM with optimal performance for numerous challenges: support of commonly observed and rare SSEs, comprehensive assignment of turn types, kink detection, geometric consistency, applicability toward molecular dynamics (MD) data, robustness with respect to structure quality and flexibility, and suitability for SSE-based protein structure alignments. For the analysis of SSEs, the use of multiple tools is often necessary, e.g., assigning helices with STRIDE and subsequently detecting kinks with Kink Finder (Wilman *et al.*, 2014). In contrast, SCOT is the first method that enables extensive analyses of SSEs, kinks, and turns in proteins in a single step and is freely available to the scientific community (Supplementary Table S1).

2 Methods

2.1 The SCOT Methodology

2.1.1 Turns

Turns in the protein backbone are the basis for all assignments by SCOT. To distinguish between hydrogen-bonded and non-hydrogen bonded turns, we introduce a novel hydrogen-bond assignment based on a previously published criterion (Dahiyat *et al.*, 1997) to circumvent the assignment of geometrically unfavorable hydrogen bonds. We distinguish *normal* (hydrogen bond from the backbone carbonyl O (C–O) of residue r_i to the backbone nitrogen H (N–H) of residue r_{i+k}), *reverse* (hydrogen bond from N–H of residue r_i to C–O of residue r_{i+k}) and *open* turns (no hydrogen bond according to our hydrogen-bond criterion, but Cα–Cα distances between 4 Å and 8 Å) of different length, e.g., *normal-5*. Examples for *normal*, *reverse*, and *open* turns are given in Figure 1a. All assigned turns are further subdivided based on their dihedral angle ranges, e.g., *normal-5 1* (see Supplementary Table S2 for the characteristics of the turn classes). SCOT assigns SSEs based on a distinct set of turn classes, hydrogen-bonding patterns, and geometric criteria, which are outlined in the following.

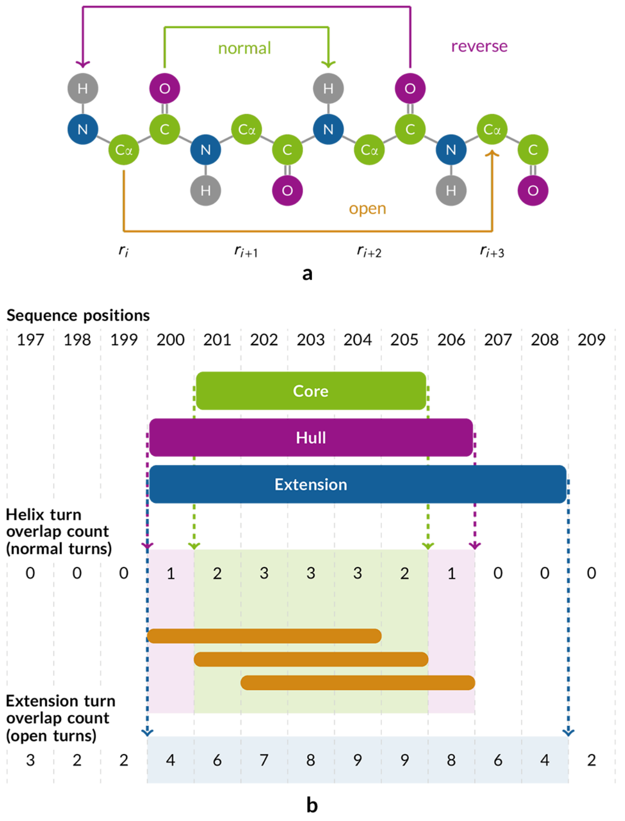

Fig. 1. The basics of SCOT a, Schematic visualization of a normal-3, a reverse-3, and an open-4 turn. b, Visualization of the core helix layer definition based on an α-core helix at residues 193–209 of 1lg7A@pdb. The core (green) is based on three normal-5 1 turns (highlighted in orange) which lead to a helix turn overlap of at least 2 at residues 201–205. The hull (purple) requires at least one normal-5 1 turn. The extension shown in this example is based on open-5 1 and open-6 4 turns. The required turn overlap of at least 3 is given at residues 200–208.

2.1.2 Helices

SCOT detects helices and assigns helix classes based on consecutively overlapping *normal* turns of lengths 3 (2.2_7), 4 (3_{10}), 5 (α), and 6 (π). Due to the presence of non-hydrogen-bonded (with respect to the backbone) helix capping motifs, we extend helices based on overlaps of the corresponding *open* turns (see Figure 1b). Helices are not uniform and are often spanned by different turn classes. Therefore, we merge overlapping α- and 3_{10}-helices and assign Purity values which reflect the individual turn class fractions of all SCOT-assigned helices. In the case of equal fractions, we assign the class mixed. Helices that are assigned based on hydrogen bonds and dihedral angles are often prone to irregularities. Our helix splitting procedure using 4-residue Cα–Cα distances allows the assignment of geometrically uniform helices. The handedness of helices is determined based on the dihedral angles of the underlying turns. Polyproline II (PPII) helices are solely classified based on *open* turn overlaps.

Additionally, SCOT provides Purity values for all assigned helix classes. The Purity for a helix spanning the residues from sequence position f to b is exemplarily defined for an α-helix in Equation 1. It puts the sum of turn overlaps of each α-helix residue in relation to the overlaps of the turns of the group of right-handed helices. $N_\alpha(f,b)$ (Equation 2) is the sum

Table 1. Secondary structure color coding used in all figures and for the PyMOL (Schrödinger, LLC, 2015) export scripts. The handedness of helices is given by RH (right-handed) and LH (left-handed).

Helices									Sheets	Kinks
RH				LH						
α	3_{10}	π	Mixed	α	3_{10}	PPII	2.2_7	Termini		
●	●	●	●	●	●	●	●	●	●	●

of the turn overlaps for the α-helix-specific turns, i.e., *normal-5 1*. The functions for the other helix classes, e.g., N_π for π-helices, are defined analogously based on their respective turns (e.g., *normal-6 2*). We use $T_{n,5,1}$ for the set of all *normal-5 1* turns for a given protein chain. $N(i, T)$ counts the number of turns that span a sequence position i.

$$\text{Purity}_\alpha(f, b) := \frac{N_a(f, b)}{N_a(f, b) + N_{3_{10}}(f, b) + N_\pi(f, b)} \tag{1}$$

$$N_a(f, b) := \sum_{i=f}^{b} N(i, T_{n,5,1}) \tag{2}$$

$$N(i, T) := |\{t | t \in T \wedge t.\mathit{front} \leq i \leq t.\mathit{back}\}| \tag{3}$$

2.1.3 Sheets

SCOT assigns parallel and anti-parallel β-sheets based on the standard hydrogen-bonding patterns (Supplementary Notes S1). Isolated strands are not covered by this definition. Therefore, all SCOT-assigned strands are grouped in β-sheets. We avoid highly flexible strand termini and geometric irregularities by a shrinking procedure and an optional splitting based on kinks. The co-occurrence of PPII helices and strands is avoided by preferring hydrogen bond-stabilized strands over PPII helices.

2.1.4 Input, Output, and Implementation Details

SCOT requires files in the RCSB Protein DataBank (PDB) format. The SCOT output includes PDB files with the SSE assignments and optional PyMOL files to visualize different classes and geometric features of the assigned SSEs using the coloring scheme given in Table 1.

The implementation details of the SCOT SSE assignment can be found in Supplementary Notes S1.

2.2 Datasets

For the turn classification and the general analysis of SSEs, we used all protein structures available from the PDB. We created datasets for the analysis of SSAMs with respect to different goals which are listed in Table 2. Further details regarding the datasets and the production of the MD trajectories for the solution NMR structures with the PDB IDs 1klp@pdb, 2c34@pdb, and 2k3i@pdb can be found in Supplementary Notes S2.

2.3 Analysis of SSAMs

2.3.1 Scaled B-Factor

A scaling procedure was applied to the B-factor of the protein backbone atoms to enable the comparison between the flexibility of the residue backbone atoms for all protein structures under investigation. A normalization (Carugo and Argos, 1998) was used to obtain a scaled B-factor distribution around the mean of 0 with unit variance.

2.3.2 Geometric Characteristics

The geometric SSE characteristics Twist, Rise, Radius, virtual torsion angle (Vtor), and bending angle (BDA) were calculated (Kumar and Bansal, 2015b). We adapted the BDA calculation for strands by determining the BDA between the pseudo-bonds defined by the backbone carbonyl carbon atom (C) of residue r_{i-2} and the backbone nitrogen atom (N) of residue r_i, and the N of residue r_i and the C of residue r_{i+2}.

2.3.3 Conformational Parameter

The conformational parameter P of the residues within different SSE classes was calculated according to (Chou and Fasman, 1974). The d-test with a $0.1\,\%$ boundary (Wilmot and Thornton, 1988) was applied to gain insights into significantly over- ($d > 3.3$) and underrepresented ($d < -3.3$) residue types per SSE class.

2.3.4 Consensus

The Consensus between the SSAMs for different helix classes and definitions of extended conformations was calculated based on a binary fingerprint representation. The Tanimoto coefficient ($\mathit{Tanimoto} : (F_1, F_2) \to [0, 1]$) between two fingerprints F_1 and F_2, one per protein structure and each of length n, was used as the measure of the Consensus.

2.3.5 Consistency

The Consistency of SSAMs was calculated using a fingerprint-based representation of the assigned SSEs. Let $F_1, \ldots, F_K$ be the fingerprints for the K structures of an ensemble annotated by an SSAM. A Weighted Tanimoto coefficient ($\mathit{Tanimoto}_W : (F_1, \ldots, F_K) \to [0.5, 1]$, see Equations 4–6) reflects the Consistency of one helix class or class of extended conformations considering all structures of one protein ensemble. We used boxplots to present the Consistency distribution for all ensembles of one dataset. We determine U (see Equation 5) based on the fingerprints of all analyzed SSAMs. This ensures that the Consistency for the same ensemble for each SSAM is calculated relative to the same divisor which results in a unified evaluation independent of the SSE lengths. For the dividend, the majority of the number of bits set to 0 or 1 is considered which is at least $K/2$ here.

$$\begin{aligned} C_i(F_1, \ldots, F_K) = \\ |\{k \in \{1, \ldots, K\} | b_{k,i} = 1, b_{k,i} \in F_k\}| \end{aligned} \tag{4}$$

$$\begin{aligned} U(F_1, \ldots, F_A) = \\ |\{i \in \{1, \ldots, n\} | \exists a \in \{1, \ldots, A\} : b_{a,i} = 1, b_{a,i} \in F_a\}| \end{aligned} \tag{5}$$

$$\begin{aligned} \mathit{Tanimoto}_W(F_1, \ldots, F_K) = \\ \frac{\sum_{i=1}^{n} \max(C_i(F_1, \ldots, F_K), 1 - C_i(F_1, \ldots, F_K))/K}{U} \end{aligned} \tag{6}$$

2.3.6 Structural Alignment

The described analyses were performed using SNOT (Brinkjost and Ehrt, unpublished).

Structural alignments of the protein structures of the same CATH topology and superfamily were performed using LOCK2 (Shapiro and Brutlag, 2004). To this end, the CATH domains of the corresponding PDB structures were extracted and the SSEs were assigned using the SSAMs discussed in this study.

LOCK2-based structure alignments for the topology and the superfamily dataset were obtained using default settings. The resulting alignment scores (see (Singh and Brutlag, 1997) for a detailed description) were normalized by the number of matched SSEs per alignment.

For SCOT, different SSE assignments of the domains were used. Apart from the default assignments, we omitted π-helices for the alignment, split strands based on the strand kink data, and split both helices and strands based on the corresponding kinks. For all these settings, we evaluated the alignment performance for both datasets to obtain optimal settings for SCOT-based SSE assignments, i.e., settings which result in low root-mean-square deviation (RMSD) values and high per-residue-SSE scores. These settings can be applied to successfully superpose topologically similar protein structures.

2.4 Application of SSAMs

The application of the SSAMs and the analysis of their results is described in Supplementary Notes S3.

3 Results

To investigate the general applicability of SCOT and the reliability of the assigned SSEs, we picked the of DSSP (Kabsch and Sander, 1983) reimplementation MKDSSP (based on hydrogen bonds), STRIDE (Frishman and Argos, 1995) and SHAFT (Koch and Cole, 2011) (both based on hydrogen bonds and dihedral angles), ASSP (Kumar and Bansal, 2015b) (geometry-based), DISICL (Nagy and Oostenbrink, 2014) (based on dihedral angles), and SEGNO (Cubellis *et al.*, 2005) (based on geometry and dihedral angles) to put the results for the analyzed quality criteria into perspective. An overview is given in Table 3. These SSAMs were selected as they cover most of the SSE classes that are also assigned by SCOT. They are briefly introduced in Supplementary Notes S4.

3.1 The SCOT Secondary Structure Assignment

3.1.1 Right-Handed α- and 3_{10}-Helices

α- and 3_{10}-helices are assigned based on overlapping *normal-5 1* and *normal-4 1* turns, and extended by *open-5 1*, *open-4 2*, and *open-6 4* turns. The helix class depends on the number of per-residue overlaps of the helix-constituting *normal* turn classes.

A post-processing step is applied to avoid kinked helical structures (Figure 2). We calculate the Cα–Cα distances of all 4-residue segments. Regular helical regions are characterized by optima for these distances (Supplementary Figures S1a–b). The introduction of distance cut-offs to split α- and 3_{10}-helices leads to the assignment of rarely bent helices (Figure 2b). Plotting the 4-residue segment Cα-Cα distances against the BDA highlights the good correlation between both indicators of helical irregularities (Supplementary Figure S2).

A complete overview of all parameters of the assigned SSEs by different SSAMs for the X-ray representatives which are discussed in the following sections (4-residue segment Cα–Cα distance, Twist, Rise, Radius, BDA, φ, ψ, ω, scaled B-factor, length, Purity, and number of assigned SSEs) can be found in Supplementary Table S3. With respect to α- and 3_{10}-helices, the high fluctuations of the φ and ψ angles for DISICL, SHAFT, and the PDB classification are in line with high BDAs. Additionally, the scaled B-factors for the helical residues are significantly higher. Together with ASSP, DISICL, MKDSSP, and SEGNO, SCOT assigns α-helices with the most stable geometric parameters. SHAFT-assigned helices are considerably longer, while DISICL assigned helices are often very short. The helix termini assignments show huge discrepancies (Figure 2a), underpinning the findings of Tyagi and colleagues (Tyagi *et al.*, 2009). The Consensus for α- and 3_{10}-helices is highest for SCOT and MKDSSP (Supplementary Tables S4 and S5).

SCOT assigns the geometrically most consistent 3_{10}-helices (see Supplementary Table S3 for the standard deviations of the analyzed geometric parameters). The scaled B-factor for 3_{10}-helices is higher than that for α-helices which is in line with earlier studies (Enkhbayar *et al.*, 2006). They characterized 3_{10}-helices as para-helices the high variances of their geometric parameters. In contrast to geometry-based methods, SCOT does not rely on uniform φ and ψ angles which are inappropriate to classify 3_{10}-helices (Enkhbayar *et al.*, 2006). *Normal-4 2* turns used for 3_{10}-helix assignment have lower hydrogen bond energies as compared to *normal-5 1* turns. Moreover, Pro residues are overrepresented in this helix class according to all methods, but ASSP and DISICL (Supplementary Table S6).

Overlapping α- and 3_{10}-helices are the only SSE classes that are by definition merged in SCOT. The class of the final helix is defined based on the Purity measure which reflects the fraction of helix class-defining turns in the helix sequence segment. In consequence, right-handed helices are characterized by a Purity value for each helix class. In case of equal maximal helix class Purities, we assign the class mixed. Figures 2c–e give examples of predominantly α-helical, 3_{10}-helical, and mixed class backbone segments and their three-dimensional structure. The fraction of mixed helix residues for the X-ray representatives dataset is 10.4% for SEGNO (Supplementary Table S7), but only 0.1% for SCOT. The Consensus between the SCOT- and SEGNO-assigned mixed helices is only 0.0028. For SCOT-derived mixed helices, the geometric characteristics lie between those of α- and 3_{10}-helices underlining the difficulty of a unique assignment. For our representative dataset, α-helices show a Purity of 0.87 while 3_{10}-helices are characterized by a Purity of 0.92. The slightly lower Purity of α-helical structures can be attributed to the occurrence of additional $r_i \rightarrow r_{i+3}$ but especially $r_i \rightarrow r_{i+5}$ hydrogen bonds. This is discussed in the next section.

As described above, SCOT-derived helices are characterized by lower BDAs compared to MKDSSP and STRIDE. SCOT identifies remaining kinks in the assigned helices by an additional hydrogen bond criterion. An analysis of known helix kinks (Meruelo *et al.*, 2011) revealed that they are often characterized by a non-consecutive sequence of turn overlaps. Looking for regions with missing hydrogen-bonded turns assists in the

Table 2. All created datasets for the analysis of SSAMs, the experimental method, their total number of protein structures, their relative amount of membrane protein structures, and their benchmark scopes.

Dataset Name	Experimental Method	Structures	% Membrane Protein Structures	Benchmark Scope
X-ray representatives	X-ray	3,597	1.0	Geometric characteristics
Non-redundant set of structures with left-handed helices	X-ray	55	1.8	Geometric characteristics
Quality dependency	X-ray	30	0.0	Quality dependency
NMR ensembles	NMR	2,856	1.5	Consistency
X-ray ensembles	X-ray	1,584	0.0	Consistency
MD simulation-derived structures	NMR	3	0.0	Applicability to MD snapshots
CATH topology	X-ray	786	1.4	Structure alignment quality
CATH superfamily	X-ray	2,304	1.7	Structure alignment quality

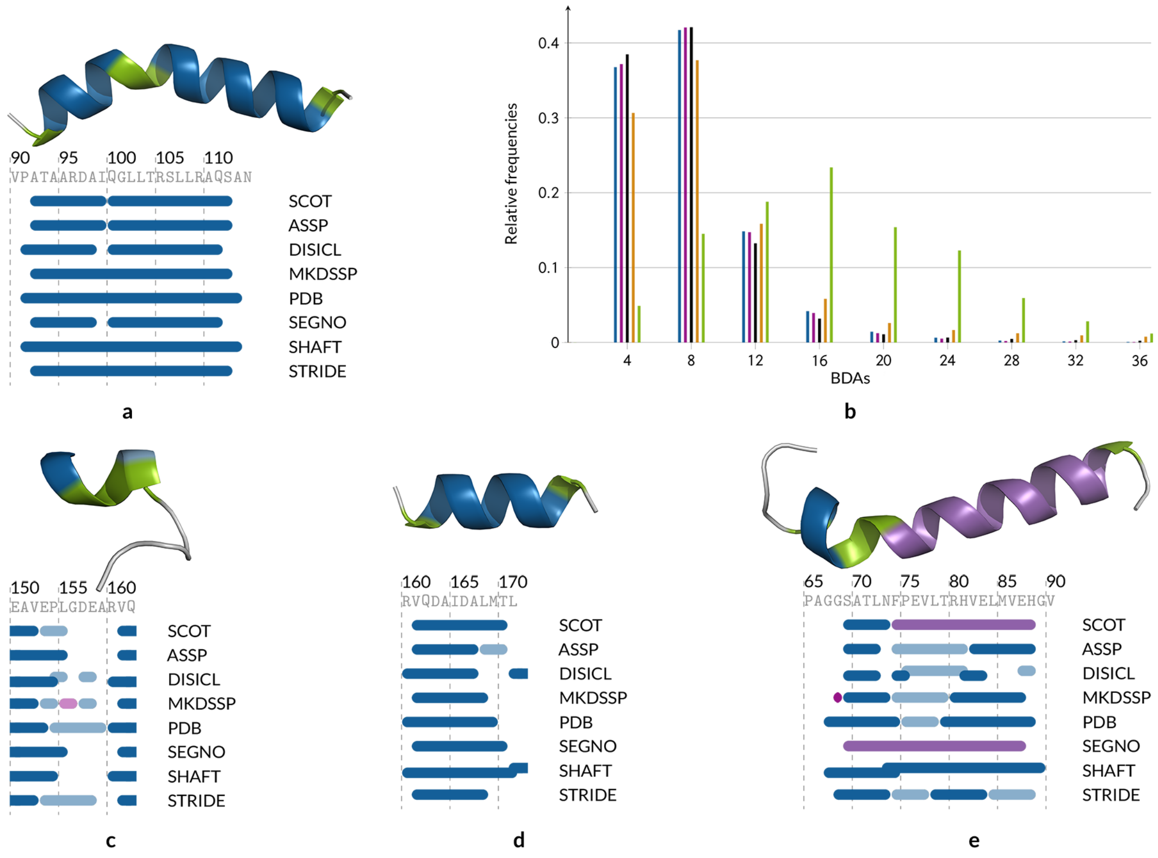

Fig. 2. Examples of helical structures and bending angle (BDA) distributions of α-helices in the X-ray representatives dataset. a, Different α-helix assignments for residues 90–114 of chain A of the structure 3rxy@pdb. The Purities of the SCOT-assigned helices are α: 0.893 , 3_{10}: 0.107 and α: 0.898, 3_{10}: 0.102. Residues 98–101 show BDAs of $28.3°$, $44.7°$, $48.1°$, and $29.2°$. b, BDA (in degrees) histograms for α-helices assigned by SCOT (blue), SCOT$_{\text{kinked}}$ (purple), ASSP (black), SHAFT (orange), and for the SCOT-assigned kink residues (green). c, Example of the assignment of a 3_{10}-helix (Purity α: 0.143 and 3_{10}: 0.857). d, Example of the assignment of an α-helix (Purity α: 0.641 and 3_{10}: 0.359). e, Example of the assignment of a mixed helix (Purity α: 0.500 and 3_{10}: 0.500). We use the following SSE-coloring scheme: right-handed α-helices (●), right-handed 3_{10}-helices (●), right-handed mixed helices (●), left-handed α-helices (●), left-handed 3_{10}-helices (●), and termini (●). Structure figures were generated with PyMOL (Schrödinger, LLC, 2015).

Table 3. The evaluated SSAMs grouped by their underlying methodology: dihedral angles (DH), hydrogen bond (HB), or geometry (GO). For our analyses of the method DSSP (Kabsch and Sander, 1983), we used the reimplementation MKDSSP.

Method	**DH**	**HB**	**GO**
SCOT (this publication)	●	●	●
SHAFT (Koch and Cole, 2011)	●	●	
STRIDE (Frishman and Argos, 1995)	●	●	
SEGNO (Cubellis *et al.*, 2005)	●		●
DISICL (Nagy and Oostenbrink, 2014)	●		
DSSP (Kabsch and Sander, 1983)		●	
ASSP (Kumar and Bansal, 2015b)			●

identification of bent regions in helices (Figure 2b). The preferred residue at SCOT-defined kinks is Leu (Supplementary Table S8). A preference for Pro at position r_{k+1} is in line with previous studies (Langelaan *et al.*, 2010). Intriguingly, only $3\,\%$ of the hydrogen bond-based kinks have BDAs above $30°$ which is an indicator for kink regions (Kumar and Bansal, 2012). Consequently, the residues assigned by SCOT are kink-prone, i.e., hot spots for kink formation.

3.1.2 π-Helices

π-helices are frequently part of other helix classes (Cooley *et al.*, 2010) and are also referred to as α-bulges (van der Kant and Vriend, 2014). In SCOT, π-helices are regarded as a special class and are not merged with other helices. They are assigned if at least two consecutive *normal-6 2* turns are detected. Most of the detected π-helices overlap with other right-handed helices. This is also reflected by the low average π-helix Purity of 0.56 with a maximum of 0.91. We identified three π-helices with Purities above 0.8 (Figure 3a) which contradicts the hypothesis that π-helices do not occur independently (Hollingsworth *et al.*, 2009). Due to the stabilizing effects of predominantly hydrophobic and aromatic residues (Kumar and Bansal, 2015a) and the co-occurrence of *normal-5* and *normal-6* turns, π-helical stretches are characterized by lower scaled B-factors than α- and 3_{10}-helices. For most assignment methods, Val, Leu, Ile, Phe, and Tyr residues are preferentially found (Supplementary Table S6). The assignments of 572 π-helices by SCOT and $1{,}053$ π-helices by MKDSSP are most similar (Supplementary Table S9). SHAFT and STRIDE assign only 36 and 47 π-helices. The main reason is the hierarchy underlying SHAFT and STRIDE which prefer α- and 3_{10}-helix over π-helix assignments. Both tools were excluded from the following analyses, as the assigned π-helices show comparatively huge deviations from the mean geometry which hints at the improper π-helix definition. Both methods underestimate the π-helical content, as the SCOT-derived Purity

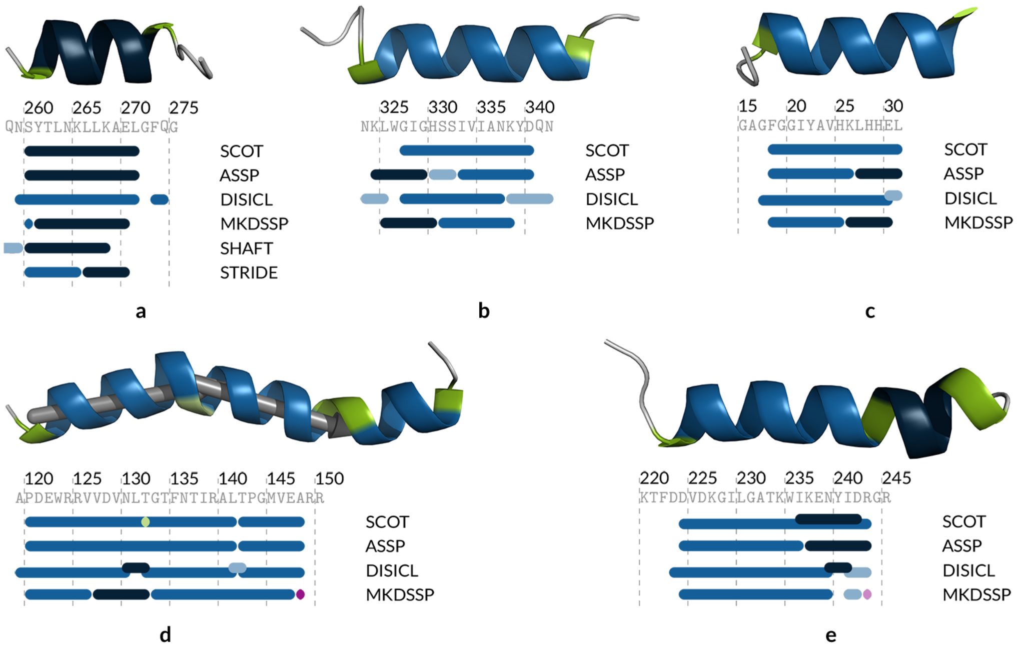

Fig. 3. Differences in the assignment of π-helices by different SSAMs. a, The π-helix in 4iic@pdb (chain A) is characterized by predominantly occurring $r_i \rightarrow r_{i+5}$ hydrogen bonds and does not overlap with other helix classes (Purity of α: 0.087 and π: 0.913). b, In contrast, the structures 3w36@pdb (chain A) (Purity of α: 0.765, π: 0.059, and 3_{10}: 0.176) and, c, 4rg3@pdb (chain A) (Purity of α: 0.867 and π: 0.133) contain α-helices which are N- and C-terminally capped by $r_i \rightarrow r_{i+5}$ hydrogen bonds. d, An α-bulge which leads to a kinked (highlighted in green) α-helical structure in 4nbr@pdb (chain A) (Purity of α: 0.891 and π: 0.109). The gray arrows indicate the helix axis of the helix parts separated by the kink. e, A π-helix of overlapping normal-6 2 which can be found inside an α-helix in 4q2w@pdb (chain A) (Purity of α: 0.469, π: 0.375, and 3_{10}: 0.156). We use the following SSE-coloring scheme: right-handed α-helices (●), right-handed 3_{10}-helices (●), right-handed π-helices (●), right-handed mixed helices (●), left-handed α-helices (●), left-handed 3_{10}-helices (●), termini (●), and kinks (●). Structure figures were generated with PyMOL (Schrödinger, LLC, 2015).

values clearly show that the SCOT-assigned π-helices are predominantly based on $r_i \rightarrow r_{i+5}$ backbone hydrogen bonds.

We analyzed potential reasons for the low number of SCOT-assigned π-helices compared to ASSP, DISICL, and MKDSSP (Supplementary Table S3). Single π-helical turns, which are predominantly assigned at the termini of α-helices (Figs. 3b and 3c), are not classified by SCOT. In the case of N-terminally unclassified π-helices, the common overlaps are between the last or the first residue of the π-helix and the first of α- and 3_{10}-helices. These structures are capping motifs consisting of a *normal-5 1* turn and a succeeding *normal-6 2* turn. The N- and C-terminal residues of the corresponding *normal-6* turns are mainly hydrophobic and aromatic and probably lead to a stabilization of the helix terminus.

Additionally, α-bulges consisting of one $r_i \rightarrow r_{i+5}$ hydrogen bond are not classified as π-helices by SCOT. In these cases, the helix is split due to an increased 4-residue segment Cα–Cα distance (Fig. 3d and Supplementary Figure S1c). For some of the SCOT-assigned π-helical stretches, the other SSAMs highly disagree concerning their classifications. They are especially interesting as they are characterized by overlapping $r_i \rightarrow r_{i+3}$, $r_i \rightarrow r_{i+4}$, and $r_i \rightarrow r_{i+5}$ hydrogen bonds (Fig. 3e).

Consequently, only the combination of hydrogen bond and geometric criteria provides a reliable classification of helices.

3.1.3 Left-Handed Helices

Left-handed helices are rarely occurring SSEs with restricted lengths (Novotny and Kleywegt, 2005). We identified a *normal-4* and *normal-5* turn class whose dihedral angles agree with those of left-handed helical conformations. However, we did not identify a *normal-6* turn class with the dihedral angles of left-handed π-helices. We used the dataset of Novotny and Kleywegt (Novotny and Kleywegt, 2005) to compare SCOT to ASSP, DISICL, and MKDSSP. DISICL assigns no left-handed classes. MKDSSP classifies the handedness (chirality) based on the parameter Vtor, which leads to the classification of one residue long left-handed helices at the C-terminus of right-handed helices in the X-ray representatives dataset (6,035 of the 6,136 left-handed α-helices are one residue long).

All methods show discrepancies in the reliable classification of left-handed helices. Although the results of SCOT are comparable to those of ASSP and DISICL, we find some striking differences (Supplementary Table S10).

Given the sparsity of left-handed helices in the X-ray representatives dataset, we searched the PDB for left-handed helices with SCOT. The set of structures with at least 4-residue long left-handed helices (see Figure 4 for examples) was used to generate and analyze a non-redundant set of structures.

The Consensus between ASSP, DISICL, and SCOT for this set is highest (Supplementary Table S11). However, the class assignments by ASSP and SCOT differ, leading to a low Consensus for the helix classes (Supplementary Tables S12 and S13). A huge proportion of ASSP-assigned α-helices is classified as 3_{10}-helices by our method which is based on the $r_i \rightarrow r_{i+3}$ hydrogen-bonding pattern. While the geometric parameters of SCOT-assigned left-handed α-helices are more robust, these of the ASSP-assigned 3_{10}-helices are more stable and show lower B-factors. SCOT-assigned right- and left-handed 3_{10}-helices have higher scaled B-factors which is in accordance with the lower hydrogen bond energies of the underlying turn type (*normal-4 3*) (Supplementary Table S2). The dihedral angles of the SCOT-derived left-handed helix classes conform to their right-handed counterparts. In contrast, the dihedral angles of ASSP-assigned left-handed α-helical segments tend toward those of 3_{10}-helices, underlining that a purely geometry-based assignment is not sufficient for distinct class assignments.

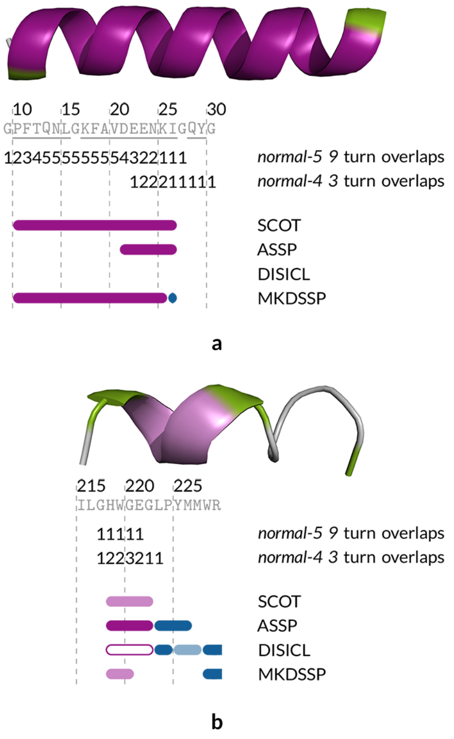

Fig. 4. Examples of left-handed helices which occur in protein structures. a, Example of a left-handed α-helix partially composed of D-amino acids in 2q33@pdb (chain B) (Purity of left-handed α: 0.887 and left-handed 3_{10}: 0.113). b, Example of a left-handed 3_{10}-helix classified in the structure 2dvt@pdb (chain A) (Purity of left-handed 3_{10}: 0.714 and left-handed α: 0.286). D-amino acids are underlined. The overlaps of the corresponding normal turns are given to illustrate the respective class assignments. DISICL does not assign left-handed classes (pink outline). We use the following SSE-coloring scheme: right-handed α-helices (●), right-handed 3_{10}-helices (●), left-handed α-helices (●), left-handed 3_{10}-helices (●), and termini (●). Structure figures were generated with PyMOL (Schrödinger, LLC, 2015).

3.1.4 Polyproline II Helices

Another class of left-handed helices in proteins are PPII helices (Cowan and McGavin, 1955). They are characterized by an extended conformation and predominantly occurring Pro residues. Their role in different biological processes was shown by several investigations (Adzhubei *et al.*, 2013; Narwani *et al.*, 2017). SCOT uses repetitive *open-4 9* turns to assign PPII helices. ASSP, DISICL, and SEGNO assign PPII helices based on the backbone geometry and dihedral angles. The Consensus between SCOT and these methods is below 20%. This is in accordance with other analyses (Mansiaux *et al.*, 2011; Chebrek *et al.*, 2014). SCOT assignments are most similar to those of SEGNO (Supplementary Table S14). The fraction of residues assigned as part of PPII helices is 1.4% for ASSP, 8.9% for DISICL, 1.3% for SCOT, and 2.9% for SEGNO (Supplementary Table S7). The stability of the geometric helical descriptors for our PPII helices is between that of DISICL and that of ASSP and SEGNO. The latter ones assign the geometrically most stable PPII helices with the lowest B-factors.

1,133 and 933 of the ASSP- and SEGNO-defined PPII helices overlap with the SCOT assigned β-strands. SCOT assignments are designed to exclude any helices occurring within hydrogen bond-stabilized strands. Enabling PPII assignment independent of strands leads to an increase from 2,752 to 4,559 PPII helices. The insufficient differentiation between PPII helices and β-strands by geometry-based methods explains the comparatively high mean B-factor of SCOT-assigned PPII helices. The mean B-factor decreases from 0.31 to 0.04 if overlapping β-strands and PPII helices are allowed.

3.1.5 Sheets

The most extended backbone conformation is found in β-strands. Our strand assignment is based on the analysis of the underlying hydrogen-bonding patterns and strand-splitting turns.

The most similar method to SCOT regarding the strand assignment is MKDSSP (Supplementary Table S15). The BDA distributions of MKDSSP, STRIDE, and SCOT are broader than those of ASSP, DISICL, and SEGNO. To address this problem, we use 4-residue segment Cα–Cα distances to split strands (distance below 8.5 Å based on Supplementary Figure S1e). High BDAs (59.7°) are found at these residue positions (Supplementary Figure S3). The splitting at the kink positions leads to a decrease in the mean BDA from 23.8° to 18.6° which is similar to that of DISICL and SEGNO. Gly (and also Ser and Val) residues are overrepresented at the kink positions (Supplementary Table S8). The assigned kinks can be used to obtain geometrically uniform strands without influencing the residue the Ncap and Ccap residue preferences (Supplementary Table S16). In contrast to geometry-based methods, SCOT assignments focus on finding stable extended conformations as revealed by the significantly lower average scaled B-factor.

The significant amino acid preferences for strands are similar for all methods (Supplementary Table S17). However, ASSP, DISICL, and SEGNO find Pro as overrepresented residue at strand termini (Supplementary Table S16). This is in line with the subsequent findings.

3.1.6 Disagreements in Assigning Extended Conformations

The assignment of β-strands and PPII helices is highly different for ASSP, DISICL, SEGNO, and SCOT (Supplementary Tables S14, S15, and S18). Figure 5 gives one example of highly different per-residue assignments. The Consensus between MKDSSP, SCOT, and STRIDE concerning hydrogen bond-based strands is high, whereas ASSP and DISICL define isolated β-strand structures that overlap with PPII helices assigned by other geometry-based methods. 24%, 48%, and 36% of the SCOT-derived PPII residues are classified as strand residues by ASSP, DISICL, and SEGNO, respectively. Vice versa, many of the PPII helices assigned by these three tools correspond to SCOT-assigned strands. Consequently, there is no clear differentiation between extended PPII helix and strand conformation underlining the inapplicability of backbone geometry and dihedral angles alone for the differentiation between both SSEs. Our hydrogen bond-based β-sheet classification and the exclusion of helices in strands enables a unique assignment of PPII helices.

3.1.7 Rare Helix Classes

2.2_7-helices are rarely observed in proteins. SHAFT assigns 2.2_7-helices as right-handed γ-helices based on inverse γ-turns (*normal-3 1*). Intriguingly, most of them are found within β-strands. Their assignment can be attributed to the non-restrictiveness of the hydrogen bond assignment according to Kabsch and Sander (Kabsch and Sander, 1983). The geometry of these hydrogen bonds highly deviates from the optimum (Bruno *et al.*, 1997; Liu *et al.*, 2008). The original definition of γ-helices relates to the first investigations concerning SSEs (Pauling *et al.*, 1951) who coined the term γ-helix as an alternative to the α-helix (Ramachandran and Sasisekharan, 1968). To prevent any confusion, we will name helices composed of *normal-3* turns as 2.2_7-helices (Donohue, 1953).

This helix class exists in globular protein structures (Tsunemi *et al.*, 1996; Yang *et al.*, 2007) and its left-handed form is characterized by φ- and ψ-angles of approximately $-80°$ and $60°$ in the left-handed form (Ramachandran and Chandrasekaran, 1972). These dihedral angles correspond to the dihedral angle space of *normal-3 1* turns (mean φ =

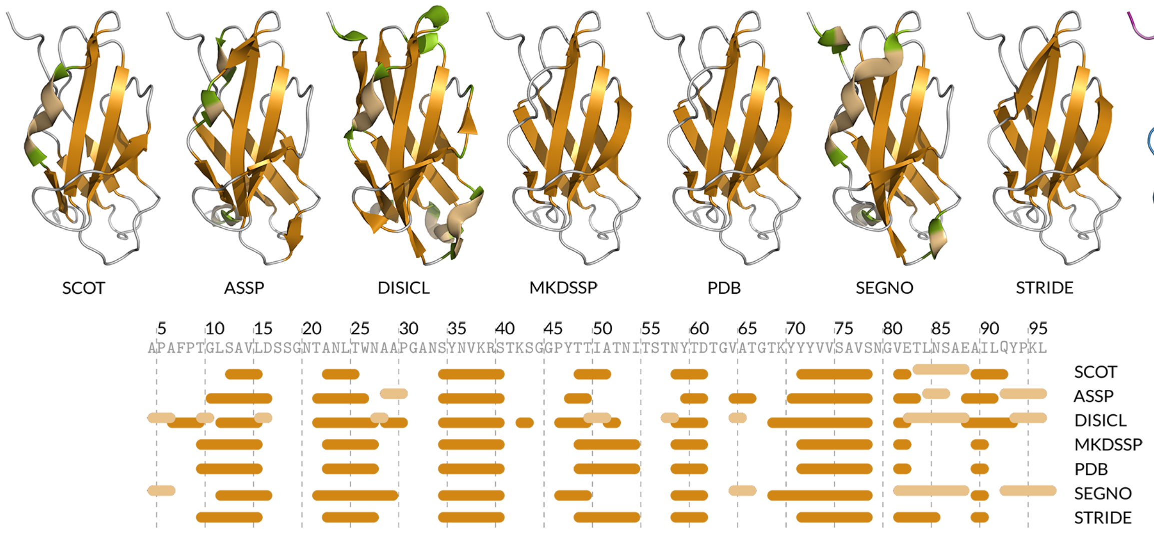

Fig. 5. Strand and PPII assignments by different SSAMs. The assignments by different SSAMs are given for chain A of 3mpc@pdb. The three-dimensional representations at the top were generated using PyMOL (Schrödinger, LLC, 2015). MKDSSP, PDB, and STRIDE do not support the assignment of PPII helices. The ribbon structure is colored according to the B-factor of the proteins' Cα atoms with a gradient ranging from purple for the highest to white to blue for the lowest value. We use the following SSE-coloring scheme: β-strands (●), PPII helices (●), and termini (●).

$-82.3°$, $\psi = 61.1°$). *Normal-3 2* turns (mean $\varphi = 74.3°$, $\psi = -52.0°$) are used to assign right-handed 2.2_7-helices. Due to their sparsity, 2.2_7-helices were classified for the structures in the PDB (2018/03/27) (Supplementary Table S19). The helices assigned by SCOT are characterized by an extended conformation and should be recognized as ribbon structures. The right-handed class can only be found at flexible N-termini of NMR ensembles and their meaning as regular SSE is questionable. In contrast, left-handed 2.2_7-helices are worthwhile further investigations as they are located in stable parts of the proteins.

ω-helices can also be found in proteins (Enkhbayar *et al.*, 2010). The average φ-, ψ-, and ω-angles of $-75°$, $-34°$, and $-177.6°$ of ω-helices lie well within the standard deviations of those of SCOT-assigned α-helices and differentiation is therefore infeasible (see Supplementary Table S3 for the dihedral angles of α-helices).

The potential occurrence of the rare α-sheets in proteins (Armen *et al.*, 2004) can also be analyzed by SCOT using an alternating pattern of *open-4 5* and *open-4 6* turns. Although we find regions with the corresponding pattern of φ- and ψ-angles in some structures of the PDB, their meaning is questionable as they lack a typical sheet-like character as originally described in the literature (Pauling *et al.*, 1951).

3.2 Impact of Structure Quality on SSE Assignments

First, we addressed the SSAMs' ability to reliably assign SSEs independent of the structure quality. This evaluation is crucial for the analysis of SCOT as the more restrictive hydrogen bond criterion might lead to inconsistencies. A correlation between resolution and SSE assignment was already observed for other SSAMs (Martin *et al.*, 2005). We modified the dataset of Konagurthu and colleagues to calculate the Consensus for pairs of X-ray structures with high and low resolution (Konagurthu *et al.*, 2012). Table 4 presents the results for this analysis.

The highest mean Consensus was observed for STRIDE, MKDSSP, and SCOT underlining their independence from structure quality. The low Consensus for SCOT and SHAFT for the structure pair 4k20@pdb and 5cna@pdb results from the fact that the protein mainly consists of β-strands. Two short helices (3 to 6 residues) in the structure bias the results if β-strands are not considered. Obviously, (completely or partially) hydrogen bond-based assignment methods except for SHAFT are the most robust ones regarding structure quality.

3.3 Consistency of SSE Assignment

Next, we analyzed the Consistency of the assigned SSEs within structural ensembles. This issue was already discussed for DSSP (Andersen *et al.*, 2002) and DSSPcont was introduced as a more robust method (Carter *et al.*, 2003). One possibility to assess the impact of protein flexibility on the Consistency of SSAMs is the analysis of the assignments for multiple models of NMR solution structures. We calculated the Weighted Tanimoto coefficient per ensemble to assess the robustness of the secondary structure assignments. The boxplots summarize the Consistency for right-handed, left-handed helices and extended conformations (Figure 6).

Overall, SCOT assignments are highly consistent. The assignment of extended conformations is less consistent than that of MKDSSP and STRIDE, which can be attributed to the additional assignment of PPII helices. The Weighted Tanimoto coefficients for the SSE classes are summarized in Supplementary Figure S4.

As it is difficult to assess the structural quality of models derived from NMR studies, we used a second dataset of X-ray structures to validate conclusions. Similar trends are observed for this dataset (Figure 6 and Supplementary Figure S4).

As in the previous analysis, SCOT shares the robustness of MKDSSP and STRIDE underpinning the benefits of hydrogen bond-based assignments which consider the backbone geometry and the stability of backbone segments. However, SCOT will never outperform MKDSSP and STRIDE regarding the Consistency due to the more restrictive hydrogen bond criterion in favor of geometrically more regular SSEs. In contrast, the hydrogen bond-based method SHAFT is less consistent due to its inconsistent helix terminus assignment (see Supplementary Table S16 for the conformational parameter of SSE terminal residues). Including geometry and dihedral angle data as realized by STRIDE and SCOT improves the geometric regularity of SSEs without neglecting backbone flexibility. Both methods are the best choices for the high-wire act of obtaining consistent, but also geometrically regular SSE assignments.

When the methods are ranked according to the mean Weighted Tanimoto per SSE class, SCOT assignments yield ranks 1 to 4 for the

Table 4. Given is the Consensus of right-handed helices and extended conformations for the structure pairs with low and high resolution, the corresponding UniProt accession codes (UniProt) and PDB IDs of the structures, and the means and standard deviations (σ) for all structures. The structure pairs are sorted with ascending difference between the resolution of the low- and high-quality structures. For SCOT* and SHAFT assignments, solely right-handed helices were considered as SHAFT does not support the assignment of β-sheets.

	Low Resolution (R)		**High Resolution (R)**									
UniProt	**PDB ID**	**R/Å**	**PDB ID**	**R/Å**	**SCOT**	**ASSP**	**DISICL**	**MKDSSP**	**SEGNO**	**STRIDE**	**SCOT***	**SHAFT**
P17802	1wef	1.9	1kg2	1.2	0.986	0.976	0.987	0.996	0.918	0.997	0.973	0.983
Q9REU3	2awc	2.2	3agt	1.4	0.987	0.964	0.991	0.967	1.000	0.982	0.973	0.943
Q9AMP1	3aro	2.2	3arx	1.16	0.997	0.946	0.946	0.984	0.917	0.979	0.994	0.962
P95339	3zow	2.4	4jcg	1.6	0.988	0.932	0.945	0.939	0.941	0.976	0.975	0.891
P24295	1aup	2.5	1bgv	1.9	0.993	0.964	0.935	0.986	0.763	0.991	0.987	0.974
P62157	1yru	2.5	1fw4	1.7	0.988	0.988	0.951	0.961	0.855	0.988	0.976	1.000
P32396	1ld3	2.6	1doz	1.8	0.994	0.975	0.958	0.981	0.885	0.991	0.988	1.000
P02213	1nwn	2.8	3sdh	1.4	0.996	0.992	0.982	0.980	0.927	0.998	0.992	1.000
P01009	1psi	2.9	3ne4	1.8	0.983	0.975	0.897	0.973	0.750	0.975	0.966	1.000
P10933	4af6	2.9	3mhp	1.7	0.972	0.967	0.908	0.968	0.852	0.991	0.944	0.928
P00390	1grh	3.0	3grs	1.5	1.000	0.977	0.986	0.993	0.987	0.997	1.000	0.970
P00512	1mto	3.2	4i7e	2.0	0.956	0.938	0.879	0.934	0.638	0.955	0.912	0.879
Q9WFX3	4gxu	3.3	4gxv	1.5	1.000	0.941	0.859	0.838	0.864	1.000	1.000	0.756
P02866	4k20	3.4	5cna	2.0	0.615	0.643	0.659	1.000	0.875	0.958	0.231	0.227
P00698	1bhz	3.9	2zq3	1.6	0.945	0.950	0.918	0.943	0.960	0.933	0.891	0.943
Mean					0.960	0.942	0.920	0.963	0.875	0.981	0.920	0.897
σ					0.097	0.085	0.083	0.040	0.097	0.019	0.193	0.196

NMR ensembles dataset and ranks 1 to 3 for the X-ray ensembles dataset, with 3_{10}-helices as the only exception (rank 5). Even rare SCOT-assigned SSEs (e.g., 2.2_7-helices or left-handed helices) show a high Consistency, underlining the reliability of the SCOT-assigned SSEs.

3.4 MD Simulations

Another field of application for SSAMs is the analysis of MD simulation data. While the assignment consistency is crucial for the analysis of MD snapshots, it is furthermore of interest to visualize changes in the SSE classes, the geometry, and the stability of SSEs. Six SSAMs were used to analyze the MD simulation data for three structures (mainly α, mainly β, mixed α/β proteins, see Supplementary Figure S5 for the RMSD plots of the simulations) to evaluate the consistency of the assignments and the sensitivity with respect to changes in the SSE patterns. SHAFT was omitted from this analysis as its applicability is restricted to helical structures. As a general trend, the geometry-based methods are highly sensitive toward minor conformational changes. This is reflected by the low Consistency of DISICL and SEGNO (Table 5). Additionally, ASSP fails to detect β-strand structures throughout the MD simulations whereas SEGNO fails to assign stable helices in the simulation of the mainly α-helical protein structure (see Supplementary Figure S6). The Consistency of SCOT is comparable to that of MKDSSP and STRIDE. However, the main benefit of using SCOT for the analysis of MD simulation data is its unique ability to reliably detect the helical handedness and variations in the helix classes. Additionally, the assignment of the PPII conformation and sheets enables the observation of a slow transition from a coiled structure to a PPII conformation and finally to an additional β-strand (see Supplementary Figure S7). In summary, SCOT offers a unique means of analyzing changes in the SSE composition over MD trajectories but also enables a consistent assignment of stable SSEs.

Table 5. Consistency of the SSE assignments for MD simulation snapshots. The given SSE classes are: right-handed helices (RH) (●), left-handed helices (LH) (●), right-handed mixed helices (RH mixed) (●), PPII helices (●), β-sheet (●), and extended conformations (β-sheet and PPII helices) (●). SSE classes that are not supported by a method are assigned n/s. If the SSE was not found in the snapshots, this is highlighted by n/f.

Mainly α	**SCOT**	**ASSP**	**DISICL**	**MKDSSP**	**SEGNO**	**STRIDE**
RH	0.927	0.907	0.875	0.942	0.851	0.930
LH	n/f	n/f	n/f	0.913	n/s	n/s
RH mixed	0.996	n/s	n/s	n/s	0.805	n/s
PPII	0.993	0.959	0.854	n/s	0.981	n/s
β-sheet	n/f	n/s	0.876	0.889	0.936	0.889
Extended	0.994	0.962	0.837	0.949	0.949	0.927

Mainly β	**SCOT**	**ASSP**	**DISICL**	**MKDSSP**	**SEGNO**	**STRIDE**
RH	0.930	0.945	0.832	0.935	0.913	0.906
LH	n/f	n/f	n/f	0.967	n/s	n/s
PPII	0.988	0.989	0.882	n/s	0.982	n/s
β-sheet	0.945	0.976	0.914	0.955	0.917	0.958
Extended	0.941	0.972	0.890	0.956	0.912	0.959

Mixed α/β	**SCOT**	**ASSP**	**DISICL**	**MKDSSP**	**SEGNO**	**STRIDE**
RH	0.965	0.982	0.904	0.972	0.861	0.972
LH	n/f	n/f	n/f	0.964	n/s	n/s
RH mixed	n/f	n/s	n/f	n/s	0.812	n/s
PPII	0.992	0.989	0.943	n/s	0.983	n/s
β-sheet	0.977	1.000	0.917	0.971	0.925	0.982
Extended	0.976	0.997	0.898	0.973	0.925	0.983

3.5 Impact of SSE Assignments on Alignment Quality

We applied LOCK2 (Shapiro and Brutlag, 2004) to assess the impact of secondary structure definitions on a tool that uses a vector-based SSE representation to align protein structures. This method reports the RMSD values, an Alignment-score (see (Singh and Brutlag, 1997) for a detailed description), and the fraction of matched SSEs.

We analyzed different versions of SCOT assignments. Besides the original assignment, we omitted π-helices as they usually overlap with

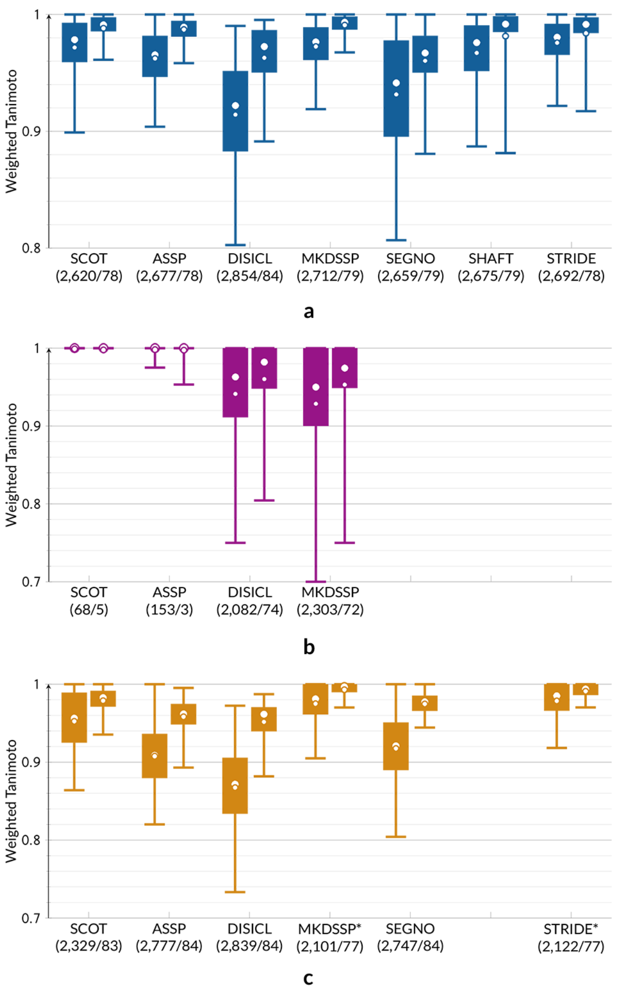

Fig. 6. Boxplots showing the Consistency. Boxplots showing the Consistency of different SSAMs for right-handed helices (a), left-handed helices (b), and PPII helices and β-strands (c) based on the Weighted Tanimoto coefficient for the NMR (left, 2,856 ensembles) and X-ray (right, 84 ensembles) ensembles datasets. The median is indicated by a big and the mean by a small white dot. SSAMs that do not support the assignment of left-handed helices or extended conformations are omitted in the boxplots (Supplementary Table S1). The numbers of ensembles in which SSEs were classified by each SSAM are given in parentheses. * The classification of PPII helices is not supported by the SSAM.

other helices. The SCOT helices and strands were split at the kink positions. All these changes led to alignments with a lower RMSD and higher scores per matched SSE pair retaining the fraction of matched SSEs as compared to SCOT using default settings (Supplementary Table S20).

To identify the best-suited SSAM for an SSE-based structure alignment, we compared the normalized scores per matched SSE using different SSAMs. The results for the CATH topology dataset are given in Table 6. SCOT assignments led to the best scores for the largest number of matches. We calculated the SSAMs' differences in the scores per match. For per-SSE-score differences above 5, the SSAM with the lower score was declared to be outperformed. ASSP, DISICL, SEGNO, and SHAFT are most frequently outperformed by other methods. Additionally, a lower fraction of SSEs is matched for their assignments (Supplementary Figure S8). In contrast, MKDSSP, SCOT, and STRIDE assignments lead to a high fraction of matched SSEs per topology pair. They are rarely outperformed by other methods. Nevertheless, there is one notable difference. MKDSSP and STRIDE are more often outperformed by geometry-based methods than SCOT showing that SCOT is a promising compromise between hydrogen bond- and geometry-based approaches to enable the most reliable SSE-based comparisons. This holds also true for the CATH superfamily pairs (Table 6). Even PPII helices and left-handed helices can be used to obtain trustworthy alignments of related domain pairs. Alignment examples underline the benefits of the SCOT assignments (Figure 7 and Supplementary Figure S9). LOCK2 alignments which led to good scores with geometry-based, but not hydrogen bond-based SSE assignments and vice versa, showed high scores using SCOT-assigned SSEs.

4 Discussion

This paper introduces SCOT for the assignment of numerous SSEs classes, which enables investigations on rarely occurring SSEs, provides additional information about helices and strands, reports irregularities therein, and assigns distinct sets of turn conformations according to the underlying dihedral angles. This information is provided in the established and widely supported PDB file format.

As we cannot provide *correct answers* for the assignment of SSEs, we present multiple criteria to evaluate SSAMs. The geometric uniformity of different SSE classes, the dependency of the assignment on structure quality, the consistency throughout structural ensembles, the applicability to MD snapshots, and the impact of the assignment on the quality of SSE-based structural alignments were evaluated. We compared SCOT to the widely applied methods MKDSSP and STRIDE, and a range of geometry- and dihedral angle-based methods. The latter methods are superior concerning the geometric regularity of the SSEs which is following other studies (Kumar and Bansal, 2015b; Martin *et al.*, 2005). However, the robustness of ASSP, DISICL, and SEGNO concerning structure quality and flexibility is lower. The hydrogen bond-based methods MKDSSP and STRIDE provide the most robust classifications with respect to structure quality and flexibility, and are well suited for SSE-based structure comparisons but show high BDAs in the assigned helices and strands. SHAFT, as an alternative hydrogen bond-based helix assignment approach, is also characterized by geometric inconsistencies, and restricted to the assignment of right-handed helices. In contrast, SCOT bridges the benefits of geometry- and hydrogen bond-based methods to gain insights into the structural space of proteins. Its dual character enables the robust classification of SSEs without significant influence on the regularity of the assigned SSEs. SCOT is perfectly suited to assign SSEs for the analysis of MD snapshots and subsequent SSE-based alignments with methods such as LOCK2 (Shapiro and Brutlag, 2004). Remaining challenges are the π- and PPII helix assignments, the classification of left-handed helices, and the reliable differentiation between β-strands and PPII helices.

In summary, SCOT represents a novel, easily applicable, and comprehensive method for SSE assignments which can be visualized by the accompanying PyMOL scripts. The rigorous preprocessing steps ensure the reliable processing of PDB files including modified residues, D-amino acids, insertion codes, and alternate locations. The output of a PDB file with the novel assignments ensures the immediate use of the SSEs for further analyses. Bridging the gap between geometric irregularities of hydrogen bond-based assignments and the missing robustness of geometry-based methods, SCOT is not restricted to single application domains and facilitates the reliable characterization of backbone geometries covering a broad spectrum of user requirements.

Acknowledgements

We wish to thank the CCDC for the Relibase+ license which enabled us to assign SHAFT-based helices. We are grateful to Chris Oostenbrink for

Table 6. The impact of different SSAMs on the per-SSE-scores of LOCK2 alignments for the CATH topology and superfamily pairs. We counted the number of times, the per-SSE-score obtained with one method was at least 5 higher (columns) or lower (rows) than that of the LOCK2 alignment using different SSAMs. Additionally, we counted the number of times an assignment method led to the best (columns) and worst (rows) alignments in terms of the per-SSE-scores.

	Score at least 5 higher →									
Topology	**SCOT**	**ASSP**	**DISICL**	**MKDSSP**	**PDB**	**SEGNO**	**SHAFT**	**STRIDE**	**Mean**	**Best**
SCOT	0	59	58	38	51	50	53	46	44.375	87
ASSP	23	0	41	36	46	37	48	39	33.750	49
DISICL	34	45	0	39	56	37	59	42	39.000	55
MKDSSP	29	55	49	0	34	49	46	33	36.875	52
PDB	21	46	39	14	0	40	23	28	26.375	35
SEGNO	26	40	36	33	48	0	49	40	34.000	41
SHAFT	26	46	46	21	22	47	0	31	29.875	40
STRIDE	30	49	49	26	38	45	42	0	34.875	42
Mean	23.625	42.500	39.750	25.875	36.875	38.125	40.000	32.375		
Worst	30	71	69	28	52	63	57	42		

Score at least 5 lower ↓

	Score at least 5 higher →									
Superfamily	**SCOT**	**ASSP**	**DISICL**	**MKDSSP**	**PDB**	**SEGNO**	**SHAFT**	**STRIDE**	**Mean**	**Best**
SCOT	0	87	86	37	58	49	68	62	55.875	265
ASSP	25	0	64	46	61	38	60	62	44.500	95
DISICL	24	59	0	33	56	30	60	43	38.125	102
MKDSSP	13	78	78	0	27	41	41	43	40.125	141
PDB	20	71	77	17	0	43	24	47	37.375	121
SEGNO	19	60	57	32	54	0	47	51	40.000	147
SHAFT	31	70	81	27	24	48	0	46	40.875	164
STRIDE	22	82	59	26	44	43	43	0	39.875	130
Mean	19.250	63.375	62.750	27.250	40.500	36.500	42.875	44.250		
Worst	64	330	247	69	123	116	111	128		

Score at least 5 lower ↓

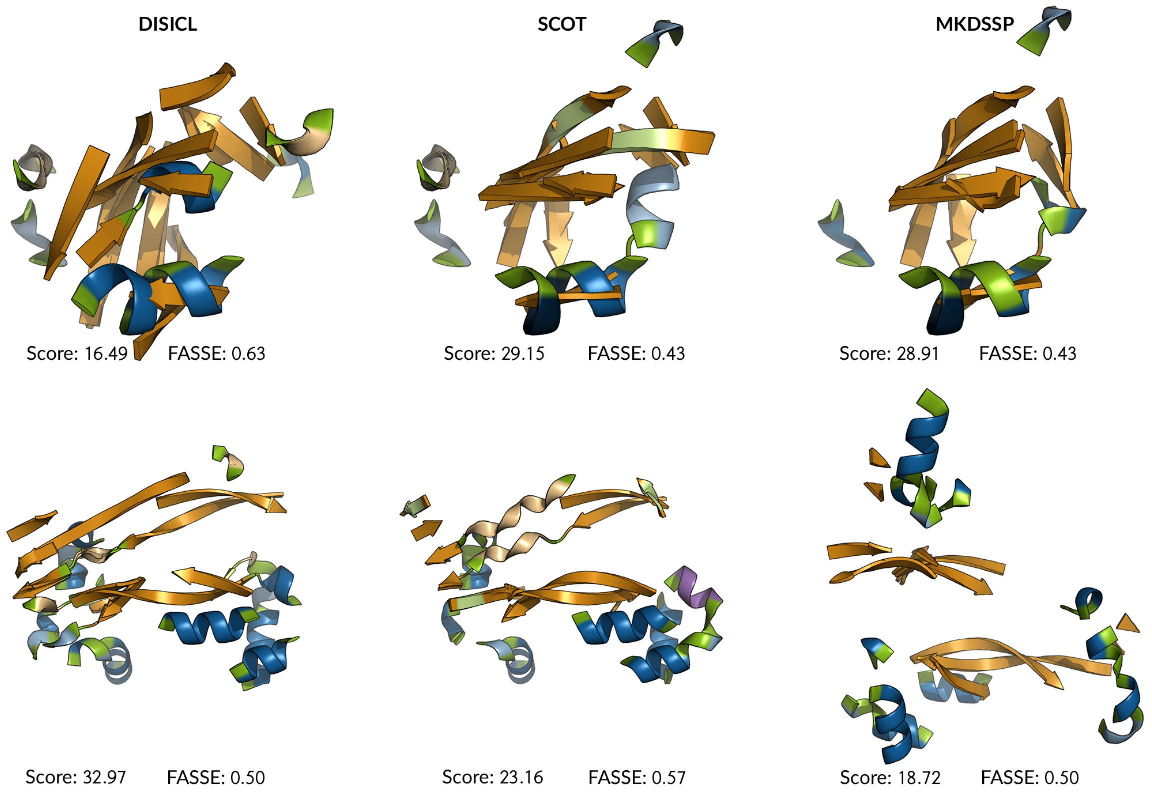

Fig. 7. LOCK2 alignments of the CATH topology domain pairs. Alignments of Barwin-like endoglucanases (5b6cA02@cath (3.10.330.10) and 1o54A01@cath (3.10.330.20), top), and Penicillin-binding protein 2a, domain 2 (4mnrA01@cath (3.90.1310.10) and 3oc2A01@cath (3.90.1310.30), bottom) based on different SSE classifications. The alignments are given for each SSAM together with the corresponding scores for matched SSEs and the fractions of aligned SSEs (FASSE). We use the following SSE-coloring scheme: right-handed α-helices (●), 3_{10}-helices (●), π-helices (●), mixed helices (●), PPII helices (●), β-strands (●), and termini (●). Figures were generated with PyMOL (Schrödinger, LLC, 2015).

providing us with the the software DISICL and supporting us concerning usage details. We also thank Simon Lovell for giving us the opportunity to analyze the method SEGNO. Stefan Kast, Daniel Rauh, Lina Humbeck, and Julia Jasper have given useful suggestions and advice for the manuscript and we acknowledge this help.

Funding

C.E. was funded by the Kekule Mobility Fellowship of the Chemical Industry Fund (VCI). O.K. acknowledges the funding by the German Federal Ministry for Education and Research (BMBF, Medizinische Chemie in Dortmund, TU Dortmund University, Grant BMBF 1316053).

References

Adzhubei, A. A. *et al.* (2013). Polyproline-ii helix in proteins: structure and function. *J. Mol. Biol.*, **425**(12), 2100–2132.

Andersen, C. A. *et al.* (2002). Continuum secondary structure captures protein flexibility. *Structure*, **10**(2), 175–184.

Armen, R. S. *et al.* (2004). Conversion of β-sheet to α-sheet structure in transthyretin at acidic ph. *Structure*, **12**(10), 1847–1863.

Bruno, I. J. *et al.* (1997). Isostar: A library of information about nonbonded interactions. *J. Comput. Aided Mol. Des.*, **11**(6), 525–537.

Carter, P. *et al.* (2003). Dsspcont: continuous secondary structure assignments for proteins. *Nucleic Acids Res.*, **31**(13), 3293–3295.

Carugo, O. and Argos, P. (1998). Accessibility to internal cavities and ligand binding sites monitored by protein crystallographic thermal factors. *Proteins*, **31**(2), 201–213.

Chebrek, R. *et al.* (2014). Polypronline: polyproline helix ii and secondary structure assignment database. *Database (Oxford)*, **2014**, bau102.

Chou, P. Y. and Fasman, G. D. (1974). Conformational parameters for amino acids in helical, β-sheet, and random coil regions calculated from proteins. *Biochemistry*, **13**(2), 211–222.

Cooley, R. B. *et al.* (2010). Evolutionary origin of a secondary structure: π-helices as cryptic but widespread insertional variations of α-helices that enhance protein functionality. *J. Mol. Biol.*, **404**(2), 232–246.

Cowan, P. M. and McGavin, S. (1955). Structure of poly-l-proline. *Nature*, **176**, 501–503.

Cubellis, M. V. *et al.* (2005). Secondary structure assignment that accurately reflects physical and evolutionary characteristics. *BMC Bioinf.*, **6**(Suppl 4).

Dahiyat, B. I. *et al.* (1997). Automated design of the surface positions of protein helices. *Protein Sci.*, **6**(6), 1333–1337.

Donohue, J. (1953). Hydrogen bonded helical configurations of the polypeptide chain. *Proc. Natl. Acad. Sci. U. S. A.*, **39**(6), 470–478.

Enkhbayar, P. *et al.* (2006). 3_{10}-helices in proteins are parahelices. *Proteins*, **64**(3), 691–699.

Enkhbayar, P. *et al.* (2010). ω-helices in proteins. *Protein J.*, **29**(4), 242–249.

Frishman, D. and Argos, P. (1995). Knowledge-based protein secondary structure assignment. *Proteins*, **23**(4), 566–579.

Hollingsworth, S. A. *et al.* (2009). On the occurrence of linear groups in proteins. *Protein Sci.*, **18**(6), 1321–1325.

Kabsch, W. and Sander, C. (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**(12), 2577–2637.

Koch, O. and Cole, J. (2011). An automated method for consistent helix assignment using turn information. *Proteins*, **79**(5), 1416–1426.

Konagurthu, A. S. *et al.* (2012). Minimum message length inference of secondary structure from protein coordinate data. *Bioinformatics*, **28**(12), i97–i105.

Kumar, P. and Bansal, M. (2012). Helanal-plus: a web server for analysis of helix geometry in protein structures. *J. Biomol. Struct. Dyn.*, **30**(6), 773–783.

Kumar, P. and Bansal, M. (2015a). Dissecting π-helices: sequence, structure and function. *FEBS J.*, **282**(22), 4415–4432.

Kumar, P. and Bansal, M. (2015b). Identification of local variations within secondary structures of proteins. *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, **71**(5), 1077–1086.

Langelaan, D. N. *et al.* (2010). Improved helix and kink characterization in membrane proteins allows evaluation of kink sequence predictors. *J. Chem. Inf. Model.*, **50**(12), 2213–2220.

Liu, Z. *et al.* (2008). Geometrical preferences of the hydrogen bonds on protein–ligand binding interface derived from statistical surveys and quantum mechanics calculations. *J. Chem. Theory Comput.*, **4**(11), 1959–1973.

Mansiaux, Y. *et al.* (2011). Assignment of polyproline ii conformation and analysis of sequence – structure relationship. *PLoS One*, **6**(3), 1–15.

Martin, J. *et al.* (2005). Protein secondary structure assignment revisited: a detailed analysis of different assignment methods. *BMC Struct. Biol.*, **5**(17).

Meruelo, A. D. *et al.* (2011). Tmkink: A method to predict transmembrane helix kinks. *Protein Sci.*, **20**(7), 1256–1264.

Nagy, G. and Oostenbrink, C. (2014). Dihedral-based segment identification and classification of biopolymers i: Proteins. *J. Chem. Inf. Model.*, **54**(1), 266–277.

Narwani, T. J. *et al.* (2017). Recent advances on polyproline ii. *Amino Acids*, **49**(4), 705–713.

Novotny, M. and Kleywegt, G. J. (2005). A survey of left-handed helices in protein structures. *J. Mol. Biol.*, **347**(2), 231–241.

Pauling, L. *et al.* (1951). The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. *Proc. Natl. Acad. Sci. U. S. A.*, **37**(4), 205–211.

Pettersen, E. F. *et al.* (2004). Ucsf chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.*, **25**(13), 1605–1612.

Ramachandran, G. and Chandrasekaran, R. (1972). Conformation of peptide chains containing both l- and d-residues. i. helical structures with alternating l- and d-residues with special reference to the ld-ribbon and the ld-helices. *Indian J. Biochem. Biophys.*, **9**(1), 1–11.

Ramachandran, G. and Sasisekharan, V. (1968). Conformation of polypeptides and proteins. **23**, 283–437.

Schrödinger, LLC (2015). The PyMOL molecular graphics system, version 1.8.

Shapiro, J. and Brutlag, D. L. (2004). Foldminer and lock 2: protein structure comparison and motif discovery on the web. *Nucleic Acids Res.*, **32**(suppl_2), W536–W541.

Singh, A. P. and Brutlag, D. L. (1997). Hierarchical protein structure superposition using both secondary structure and atomic representations. In *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, volume 5, pages 284–293.

Tsunemi, M. *et al.* (1996). Crystal structure of an elastase-specific inhibitor elafin complexed with porcine pancreatic elastase determined at 1.9 å resolution. *Biochemistry*, **35**(36), 11570–11576.

Tyagi, M. *et al.* (2009). Analysis of loop boundaries using different local structure assignment methods. *Protein Sci.*, **18**(9), 1869–1881.

van der Kant, R. and Vriend, G. (2014). α-bulges in g protein-coupled receptors. *Int. J. Mol. Sci.*, **15**(5), 7841–7864.

Wilman, H. R. *et al.* (2014). Helix kinks are equally prevalent in soluble and membrane proteins. *Proteins*, **82**(9), 1960–1970.

Wilmot, C. and Thornton, J. (1988). Analysis and prediction of the different types of β-turn in proteins. *J. Mol. Biol.*, **203**(1), 221–232.

Yang, M. *et al.* (2007). Structural basis of histone demethylation by lsd1 revealed by suicide inactivation. *Nat. Struct. Mol. Biol.*, **14**(6), 535–539.