JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

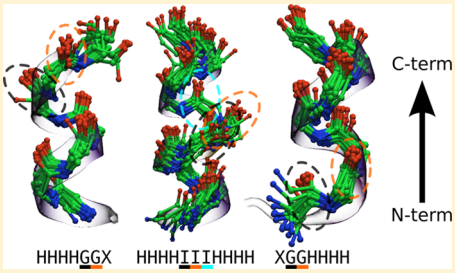
Protein Secondary Structure Classification Revisited: Processing DSSP Information with PSSC

Jan Zacharias and Ernst-Walter Knapp*

Fachbereich Biologie, Chemie, Pharmazie/Institute of Chemistry and Biochemistry, Freie Universität Berlin, Fabeckstrasse 36A, 14195 Berlin, Germany

Supporting Information

ABSTRACT: A first step toward three-dimensional protein structure description is the characterization of secondary structure. The most widely used program for secondary structure assignment remains DSSP, introduced in 1983, with currently more than 400 citations per year. DSSP output is in a one-letter representation, where much of the information on DSSP's internal description is lost. Recently it became evident that DSSP overlooks most π-helical structures, which are more prevalent and important than anticipated before. We introduce an alternative concept, representing the internal structure characterization of DSSP as an eight-character string that is human-interpretable and easy to parse by software. We demonstrate how our protein secondary structure characterization (PSSC) code allows for inspection of complicated structural features. It recognizes ten times more π-helical residues than does the standard DSSP. The plausibility of introduced changes in interpreting DSSP information is demonstrated by better clustering of secondary structures in (φ, ψ) dihedral angle space. With a sliding sequence window (SSW), helical assignments with PSSC remain invariant compared with an assignment based on the complete structure. In contrast, assignment with DSSP can be changed by residues in the neighborhood that are in fact not interacting with the residue under consideration. We demonstrate how one can easily define new secondary structure classification schemes with PSSC and perform the classifications. Our approach works without changing the DSSP source code and allows for more detailed protein characterization.



INTRODUCTION

The secondary structure of a protein reflects sequences of regular, mainly local structural pattern. Traditionally, a rough characterization of protein secondary structure involves only three structure classes, which are helix, strand, and "other", where the latter contains the more irregular, coil-like structures. More detailed discrimination involves three helix classes (α-helix, 3_{10}-helix, and π-helix) and several classes belonging to the coil moiety.[1−3] Secondary structure content is used to classify proteins in structural families.[4,5] It is also an essential component for protein structure prediction approaches.[6] Two commonly used programs that classify protein secondary structures are DSSP[7−9] and stride,[10] which we compare in the present study. DSSP is used by RASMOL;[11] both DSSP and stride are used in the Protein Data Bank (PDB)[12] to characterize protein secondary structure. Stride is used in VMD,[13] which displays the three-dimensional structure of proteins highlighting the secondary structure elements. PyMOL[14] uses a tool called "dss" to define secondary structure based on protein backbone geometry and H-bond pattern. Alternatively, DSSP and stride can be used by PyMOL[14] via an additional plugin. Software libraries dealing with protein structures such as MDAnalysis[15] or Biopython[16,17] usually possess means of processing DSSP and stride output. The latter library was used intensively for this work.

The compact form of native three-dimensional protein structures is stabilized by hydrophobic interactions forcing the nonpolar, hydrophobic residues into the interior of the protein volume, while charged and polar hydrophilic residues are located mainly on the protein surface that is solvent exposed. To lower the free energy of such compact protein structures, charged and polar groups which were forced into the interior part of the native protein structure form salt bridges (of oppositely charged groups) and hydrogen bonds, respectively. In this context the polar CO- and NH-groups of the protein backbone often form H-bonds among each other. The regular architecture of the peptide backbone favors repetition of the same local structures to saturate the polar backbone groups by H-bonds leading to regular helix and β-sheet structures. The formation of these H-bonds is largely responsible for the native protein structure.

The majority of protein secondary structure classification concepts are based on H-bond patterns using energetic[7,18−20] and/or geometric[10,21−27] criteria. An interesting extension of DSSP is presented with DSSPcont[18] that applies the DSSP algorithm with different energy thresholds for H-bond definition. For reviews about the different approaches see refs 28 and 29. DSSP uses the coordinates of the four polar backbone atoms (N, H, C′, O) involved in pairwise backbone H-bonds. Accordingly, H-bonds formed between NH- and CO-groups belonging to pairs of C_α atoms at sequence positions (i,

Received: February 11, 2014

   dx.doi.org/10.1021/ci5000856 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

i+3), (i, i+4), and (i, i+5) are considered as hallmarks of 3_{10}-, α-, and π-helical structures respectively, if this condition holds for at least two subsequent C_α-atoms in the polypeptide sequence. Structures are classified as β-sheet, if one pair of backbone CO- and NH-groups belonging to C_α-atoms at alternating sequence positions i and i+2 form H-bonds with backbone CO- and NH-groups of a second polypeptide strand also localized at alternating C_α-atom positions forming a parallel or antiparallel β-sheet structure. The second strand may belong to a different chain or the same chain with a sufficiently large distance between the participating residues.

Amino acids (AA) belong to coil structures, if the backbone NH- and CO-groups are involved in nonrepetitive irregular H-bond pattern or possess no H-bonds. Coil structures connect regular helix and strand structures. Besides truly irregular coil structures a small number of them exhibit specific motifs but may appear to be irregular, since the structural variations within the same motif are larger than for the regular helix and strand structures.

The majority of secondary structure classification schemes uses sharp discrimination criteria between different classes, albeit in a few works[19,30] AAs were considered to belong to more than one secondary structure motif. Helix structures often show mixtures of different helix types, such as α- and 3_{10}- or α- and π-helices.[20,31−33] An example of such mixed helix structures is shown in Figure 1. In the Supporting Information (SI) we show more of such examples of mixed helix structures where the assignment with DSSP differs from the one performed with the protein secondary structure characterization (PSSC) program introduced in the present study (SI Figure S1). The occurrence of such overlapping helix motifs has motivated us to design an alternative secondary structure classification scheme, in which such mixed structural motives are considered explicitly using structural profiles. The present approach is motivated by the pioneering work of the DSSP developers. DSSP was created 30 years ago and was therefore based on the limited number of only 62 protein structures.[7] Since then a number of different approaches to characterize protein secondary structure have been developed. The second most prominent tool to characterize secondary structures is stride[10] which was developed nearly 20 years ago again using a rather limited set of 226 protein structures. The present approach uses the same strategy as DSSP to characterize the secondary structure of proteins based on the analysis of the backbone H-bond pattern.

The purpose of the present study is to offer a modernized PSSC tool, which is based on Astral40,[34] a representative large body of more than 10000 nonredundant protein structures now available in the PDB.[12] PSSC uses the backbone H-bond pattern obtained by the analysis with DSSP as the main criterion for secondary structure classification. Some of the problems and ambiguities connected with DSSP are avoided in PSSC. The local secondary structure (LSS) information given by PSSC is rich and easily interpretable. This facilitates the introduction of new LSS classes. It also opens new avenues to improve the quality of secondary structure prediction, which with the present tools seems to be limited to an accuracy of slightly more than 80% for a three-class classification scheme.[35−37]

To demonstrate results in the present study all structures from the Astral40[34] database (version 1.75) are used, including more than 1.8×10^6 residues in 10569 different protein domain structures. To make a connection with the three-class secondary structure classification scheme derived from DSSP[7] data, a filter can be applied which merges related classes and assigns mixed classes to the appropriate closest pure class.

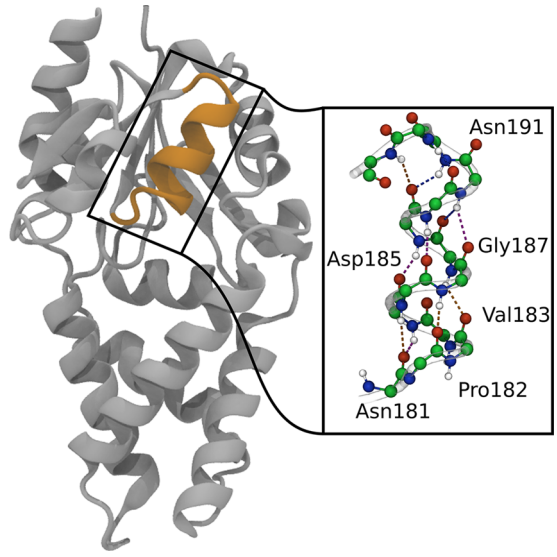


Figure 1. Protein structure of the haloacid dehalogenase PH0459 (PDB id, 1X42[48]). The dotted lines in the close-up on the right side denote H-bonds as defined by DSSP. Close-up: Segment with mixed helix types. DSSP detects three π-helix-type H-bonds with acceptors Asn181, Pro182, and Val183 and respective donors Cys186, Gly187, and Gly188. This would lead to a characterization as π-helical for the residues Pro182 to Gly187 because these residues are bridged by π-helix-type H-bonds. On the other hand the residues Asp185, Cys186, and Gly187 are involved in H-bond acceptors of α-helix type with the respective donors Ser189, Lys190, and Asn191. These α-helix-type H-bonds bridge the residues Cys186 to Lys190 which are thus considered as belonging to an α-helix using the former DSSP version 2.0.4.[7] Since in this version α-helices are assigned first and thus favored, the minimal length of five residues required for the π-helix is no longer satisfied and the residues Pro182 to Asp185 are marked as turn (T). The new DSSP version 2.2.1[8] assigns the π-helix first, characterizing residues Pro182 to Gly187 as π-helix. In the new DSSP version the residues Gly188 to Lys190 are still α-helical. Hence, a minimal length restriction of four residues for the α-helix motif does not apply. In both DSSP versions the information that residues Cys186 and Gly187 could be considered π-helical as well as α-helical is not given. Residues Gly188, Ser189, and Lys190 are involved in H-bond acceptors of 3_{10}-helix type with the respective H-bond donors Asn191, Leu192, and Gly193. Residues Ser189 to Leu192 are bridged by these H-bonds and should thus be considered as 3_{10}-helical. But even with the new DSSP version the α-helix is assigned before possible 3_{10}-helices are considered. Since residues Ser189 and Lys190 are already marked as belonging to an α-helix, only residues Asn191 and Leu192 are left for the 3_{10}-helical motif. As DSSP only assigns 3_{10}-helices with a minimal length of three, both residues Asn191 and Leu192 are put in the turn class. See also Table 1

RESULTS

Secondary Structure Assignment of Proteins with DSSP. We start with a description of the structural features used by DSSP for LSS characterization, since they are the basis for PSSC. DSSP classifies protein secondary structures mainly by backbone H-bonds.[7] However, it also provides information on C_α-pseudodihedral angles and C_α-pseudobond angles from which only the latter is actually used by DSSP to specify the LSS of a residue. Repeating H-bond patterns of the same type lead to classification as helix or strand while nonrepetitive H-bonds are classified as turns or β-bridges. As the relative

dx.doi.org/10.1021/ci5000856 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

orientations of the backbone oxygen and nitrogen atoms are reflected in the respective (φ, ψ) backbone torsion angles, residues belonging to the same secondary structure type are reasonably well clustered in a Ramachandran plot.[38,39]

DSSP provides protein secondary structure information on two levels. The upper level classifies the LSS of a residue by a one-character secondary structure information (1CSSI) code summarizing the DSSP secondary structure analysis that is mainly based on the H-bond pattern of the protein backbone in eight classes. These classes are denoted by the seven letters E, B, G, H, I, T, and S and the space character (_), which for better readability is replaced by the letter "C" in the present study (see first position denoted "full" in column 2 of Table 1).

Table 1. Mixed Helical Motif (Residues 180−194) in the Protein PH0459 (PDB Code 1X42) from the Haloacid Dehalogenase Family (For the Three-Dimensional Structure, See Figure 1), Where the Last Residue, Mse194, Is a Selenomethionine[a]

residue id	1CSSI from DSSP		1CSSI from Stride		1CSSI from PSSC		8CSSI from PSSC	7CSSI from DSSP
	full	SSW	full	SSW	full	SSW		
Asp180	S	S	C	C	S	S	␣␣␣␣␣rSD	␣␣␣S+␣␣
Asn181	C	C	T	T	C	C	␣␣␣␣␣o␣n	␣>>␣-␣␣
Pro182	T	I	T	I	I	I	␣hI␣␣RSP	␣4>S+␣␣
Val183	T	I	T	I	I	I	␣hI␣␣RSV	␣4>S+␣␣
Lys184	T	I	H	I	I	I	␣hI␣␣RSK	␣45S+␣␣
Asp185	T	T	H	H	I	I	␣␣I␣␣oSD	␣X5S+␣␣
Cys186	H	H	H	H	H	H	␣HI␣␣RSC	␣><S+␣␣
Gly187	H	H	H	H	H	H	␣HI␣␣RSG	␣><S+␣␣
Gly188	H	H	H	H	H	H	␣H␣␣␣RSG	>4<S+␣␣
Ser189	H	H	H	H	H	H	GH␣␣␣RSS	><>S+␣␣
Lys190	H	H	H	H	H	H	GHi␣␣RSK	><5S+␣␣
Asn191	T	T	H	H	G	G	G␣i␣␣RSN	<<5S+␣␣
Leu192	T	G	C	G	G	G	G␣i␣␣RSL	<␣5S-␣␣
Gly193	T	T	C	C	T	T	␣␣i␣␣LSG	<␣5S+␣␣
Mse194	C	C	C	C	C	C	␣␣␣␣␣L␣M	␣␣<␣-␣␣

[a]Columns 2, 3, and 4 show a comparison of the 1CSSI code based on DSSP,[7] stride,[10] and PSSC of this work, respectively. The second position in these three columns shows the classification results for a sliding sequence window (SSW) approach, where the respective secondary structure classification program uses only structure information of five residues up- and down-stream in sequence relative to the residue under consideration. The 1CSSI code of PSSC was defined to be as close as possible to the one of DSSP assigning first the α-helical class (see text). Column 5 (6) shows the eight-character (8CSSI) {seven-character (7CSSI)} secondary structure information of PSSC (DSSP). The 7CSSI code from DSSP and the 8CSSI code from PSSC both remain unchanged for the SSW approach compared to the full protein approach. The new DSSP (version 2.2.1) assigns residues 182−187 to π-helix (class I) but still considers residues 191−193 as belonging to the turn class T.

Most users employ only the 1CSSI code of DSSP. However, DSSP also provides more detailed structure information on a deeper level using a seven-character string (7CSSI) as described in the next section.

The seven letters of the 1CSSI code characterize the LSS of a residue and refer to "E" for extended β-strand involving a minimum of two neighbor residues; "B", for isolated β-bridge (shortest possible strand motif, which does not contain a repetitive H-bond pattern). H-bonds connecting two sequential neighbor residues at positions i (with C═O) and $i+n$ (with N−H) are called n-turn in DSSP, where n = 3, 4, 5. They are denoted as "T" in 1CSSI if they are isolated. If two turns of the same type are direct neighbors, DSSP assigns the residues located between the residues involved in turn-forming H-bonds to a helix of the same type (if not assigned yet), namely, 3_{10} for n = 3, α for n = 4, and π for n = 5, denoted by the letters "G"′, "H", and "I", respectively (see Table 1, column 1CSSI). Residues may simultaneously possess H-bond patterns of different turn types (see for instance Figure 1, closeup). Until recently DSSP assigned first α-helical LSS. As a consequence, residues in short patches of 3_{10}- or π-helices may no longer be identified and thus classified as isolated turn motifs denoted by the letter "T".[15,31] However, from version 2.1.0 upward DSSP starts helix assignments with the π-helix and thus prefers now the π-helix versus the other two helix types.

For each residue (i) DSSP determines the C_α-pseudobond angle, which is the angle between the vectors $C_\alpha(i) - C_\alpha(i-2)$ and $C_\alpha(i) - C_\alpha(i+2)$. For residues not assigned to a helix, strand, or turn, the summary class S of bends is used if this angle is smaller than 110° corresponding to a strongly bent geometry without characteristic backbone H-bonds. If none of the above conditions applies, DSSP sets a space, which we mark by the letter "C" for better discrimination. Such residues do not involve backbone H-bonds relevant for secondary structure formation and belong to a relatively straight region of protein backbone structure.

Detailed Secondary Structure Information from DSSP.

Besides the commonly known 1CSSI code described above, DSSP also provides a more detailed seven-character secondary structure information (7CSSI) code (see Scheme 1 and, for an example, column 6 of Table 1) under the column "structure" in the detailed DSSP output file. While the 7CSSI code of DSSP refers solely to the H-bond pattern involving the residue under consideration, the 1CSSI code also uses information involving the H-bond pattern of residues in the neighborhood that are needed for a classification in helices or strand. Hence, the 1CSSI code contains additional nonlocal H-bond information.

Scheme 1. Seven-Character Secondary Structure Information (7CSSI) Of DSSP[a]

```
[>,<,X,3] [>,<,X,4] [>,<,X,5] [S] [+,-] [ω,Ω] [ω,Ω]
|   1   |   2   |   3   | 4 | 5 | 6 | 7 |
```

[a]For each position the set of possible characters is shown in the rectangular brackets above. The space character "_" is set, if none of the conditions listed in the rectangular brackets apply for this position. In the original DSSP the position remains empty in this case. The Greek letters ω or Ω are place holders that for strand geometries identify the parallel or antiparallel partner strands, respectively. More details and the meaning of the other characters are given in the text. An example is shown in Table 1; a more detailed example can be found in Supporting Information Table S1.

The first three character positions in the 7CSSI code of DSSP (see Scheme 1 and Table 1 column 6) are used to specify the involvement of the considered residue in H-bonds corresponding to 3-turn, 4-turn and 5-turn, respectively. The character ">" or "<" in position 1, 2, or 3 denotes that the residue is involved in the corresponding turn structure by forming an H-bond with its backbone donor (-NH) or acceptor group (-CO), respectively. If donor and acceptor groups of the residue both form H-bonds resulting in turns of the same type, the residue is marked with X in the position of the corresponding type of turn. The number 3, 4, or 5 in the

dx.doi.org/10.1021/ci5000856 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

respective positions indicates that although the residue is not directly involved in a turn-forming H-bond, it is "bracketed" by a corresponding turn-like motif. For a more detailed explanation see Figure 2.

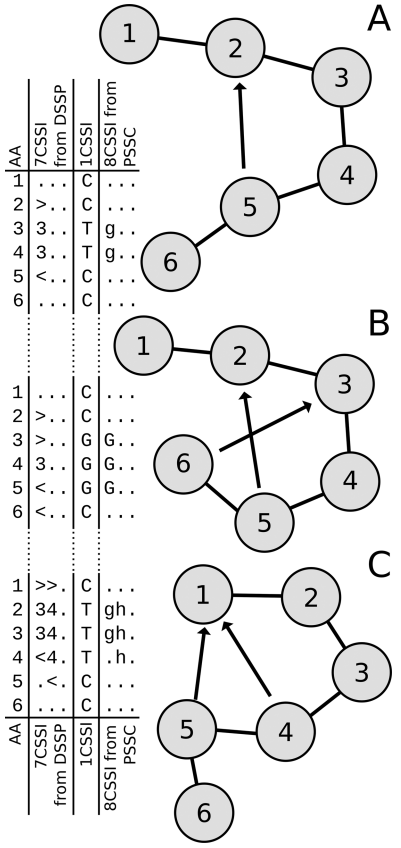


Figure 2. Three examples of schematic turn- and helix-type H-bond pattern of a formal sequence of six amino acids (AA number, first column). The second column contains the first three characters of the 7CSSI code of DSSP; the last column contains the corresponding characters in the 8CSSI code from PSSC. The third column contains the 1CSSI code, whose content is, for this example, the same if based on DSSP or PSSC. (A) Singular 3_{10}-type H-bond from residue 5 to residue 2 leads to a characterization as the turn (T) of the two "bridged" residues 3 and 4. (B) Here two 3_{10}-type H-bonds occur with residues 5 and 6 as donors. Hence, the three residues 3, 4, and 5 are classified as 3_{10}-helix (G). (C) Two different turn types (n = 3, 4) follow one after other; therefore, residues 2, 3, and 4 are considered to be turns (T) and not helices.

The letter "S" at position 4 in the 7CSSI string has the same meaning as in the 1CSSI code corresponding to bend geometry. Additionally, for each residue (i) DSSP also determines the C_α-pseudodihedral angle, built by the four consecutive C_α-atoms $C_\alpha(i-1)$, $C_\alpha(i)$, $C_\alpha(i+1)$, and $C_\alpha(i+2)$ and sets a "+" ("−") in position 5 if this angle is positive (negative). The last two positions (6 and 7) are used for a strand residue to denote contacts with other strands. Since contacts on either side are possible, two such positions are needed. DSSP labels strand sequentially by letters of the Latin alphabet. The corresponding letter is lowercase for parallel geometry and capital for antiparallel strand geometry. If none of these criteria apply, a space character "_" is set at the corresponding positions in the 7CSSI string.

In the 1CSSI code of DSSP, information from the 7CSSI code at position 5 denoting the sense of bending is ignored. Information from positions 6 and 7 are only used to identify strands. Positions 1, 2, and 3 are used to identify turn or a specific helix type. However, the very detailed information where an H-bond emerges from a residue, denoted by the characters "⟨", "⟩", and "X", are not used explicitly.

Characterizing Protein Secondary Structure with PSSC. *General Considerations.* One problem in the strategy of DSSP characterizing protein secondary structure is that on the lower level of description the 7CSSI code of one specific residue is not sufficient to generate the one character code 1CSSI, since the latter requires also information from neighboring residues. This additional information is not explicitly given but needs to be collected from the 7CSSI code of neighbor residues. A second problem, which is partly related to the first problem, is that mixed secondary structures, such as the simultaneous occurrence of two different helix-type H-bonds, are not easily identifiable and are lost in the 1CSSI code. This applies to mixed helix motives as well as to turn markers appearing in the first three positions of the 7CSSI string (see for instance Table 1). In such cases, the decision of DSSP in favor of one secondary structure type depends critically on the order with which the assignment for the different classes is performed. It leads to preferences for secondary structure classes which are filled first. To overcome these problems PSSC uses an eight-character secondary structure information (8CSSI) code shown in Scheme 2. The values in the rectangular brackets in Scheme 2 indicate the allowed characters at the respective position in the 8CSSI string. An example is shown in Table 1 column 5.

Scheme 2. Eight-Character Secondary Structure Information (8CSSI) Code of PSSC[a]

[g,G] [h,H] [i,I] [a,A,p,P] [a,A,p,P] [R,r,o,l,L] [S] [A…Y,a…y]
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |

[a]For each position the set of possible characters is shown in the rectangular brackets above. The space character "_" is set if none of the conditions listed in the rectangular brackets apply for this position. A detailed explanation is given in the text. An example is shown in Table 1; a more detailed example can be found in Supporting Information Table S1.

The 8CSSI string should not be considered as a replacement of the well-established 1CSSI code of the DSSP characterization scheme into the eight LSS classes but rather as an alternative for secondary structure characterization on the lower level of description. Formally, the eight-character string of PSSC allows for $3^3 \times 37 \times 5 \times 2 \times (21 \times 2) = 419580$ different combinations, although many of them are by definition impossible or highly unlikely. The 8CSSI string provides practically all of the information necessary to recreate the complete 1CSSI code of DSSP and offers insight and additional information on LSS whose type is ambiguous or mixed. Relying on a nomenclature that is close to that of DSSP, the 8CSSI is human-readable and simultaneously suitable to be parsed by computer programs.

Detailed Description of the 8CSSI Code of PSSC. Analogous to the 7CSSI code of DSSP, the first three characters in 8CSSI of PSSC describe the helical characteristics of a residue. However, we refrain from characterizing the directions of H-bonds explicitly. The uppercase letters "G", "H", and "I" have

dx.doi.org/10.1021/ci5000856 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

the same meaning as in the 1CSSI code of DSSP, namely, that two H-bonds of the same helix type occur at two neighboring residues. However, the 8CSSI code of PSSC also allows for the characterization of mixed helical structures in which a residue is simultaneously part of two different types of helices (for an example, see Figure 2 and Table 1). In rare cases, a residue can even belong to all three types of helices. The 7CSSI code of DSSP contains such information only implicitly since it also requires information from the 7CSSI code of neighboring residues. Until recently the 1CSSI code of DSSP[7] started with the assignment of α-helices and then the 3_{10}-helices, while π-helices were assigned last, resulting in a overrepresentation of α-helices.[20,31] In the current version (2.2.1) DSSP starts with π-helices leading now to a preference of them.

A lowercase letter in the first three positions of the 8CSSI code indicates the presence of an isolated n-turn (g, h, i, for n = 3, 4, 5, respectively) that, in the absence of an uppercase letter, is converted to the turn class T in the 1CSSI code of the PSSC model. A realization of this code is demonstrated in Table 1 and Figure 2. The characters in positions 4 and 5 of the 8CSSI string indicate a parallel (P or p) or antiparallel (A or a) strand neighborhood, allowing, for instance, a residue to be marked as being part of a sandwiched β-strand with two antiparallel neighbor strands denoted by A in positions 4 and 5. No attention is paid to the order of parallel and antiparallel markers. Hence, "AP" and "PA" are equivalent. In the same spirit as for helices and turns, lowercase letters mark nonrepetitive H-bonds patterns; i.e., residues belonging to an isolated β-bridge are marked with "a" or "p". In the absence of an uppercase letter in the 8CSSI string, this leads to "B" (denoting a β-bridge) in the 1CSSI code instead of "E".

In DSSP imperfections in the hydrogen bonding pattern of a β-sheet, so-called β-bulges in which up to four residues in a sequence are not involved in the H-bond pattern of β-bridges, are still considered to belong to the same β-strand.[7] In contrast PSSC makes a β-strand assignment only for a single residue not involved in a β-bridge-type H-bond if its predecessor and successor belong to the same strand. In this case, the parallel (P) and/or antiparallel (A) strand markers from the direct neighbor residues are assigned to the residue in focus.

The handedness of a residue is indicated in position 6 of the 8CSSI code of PSSC. It is defined by the sign of ϑ, the rotation angle per residue, that is computed as described in Methods.[38] Absolute values of ϑ larger than 150° and values of d (rise per residue as described in Methods) smaller than 1 Å do not provide a significant handedness signal. In this case, a lower case "o" is set in position 6. Positive values of ϑ correspond to right-, negative values to left-handedness, which are denoted by the letters (R, r) or (L, l) in position 6, respectively. Capital letters R or L are set if a direct neighbor residue up- or downstream in the sequence possesses the same handedness; otherwise the corresponding lowercase letters are set. In DSSP information on handedness is contained in the sign of the C_α-pseudodihedral angle given in position 5 of the 7CSSI code. The character in position 7 ("S") is identical to the respective character in position 4 of the 7CSSI code of DSSP. "S" marks a residue as having bent geometry if the absolute value of the C_α-pseudobond angle is smaller than 110°.

The AA type may also be of interest and is used by one of our more specialized secondary structure models proposed below. Thus, we include the one-letter AA name in position 8 of the 8CSSI code of the PSSC model. For a residue preceding an isolated proline (i.e., the residue has a proline as a neighbor residue on the C-terminal side), its local conformation is preferentially in the β-strand region.[40] To denote such influences a lowercase letter is used for the one-letter code; otherwise an uppercase letter is used. For nonstandard residues the character "X" or "x" is used accordingly.

Generation of the Eight-Character Secondary Structure Information on PSSC. The 8CSSI code of PSSC is generated by exclusively parsing the 7CSSI code of DSSP with a Python script. No information from the 1CSSI code of DSSP was used, and no changes in the source code of the DSSP program were made. However, to classify helix structures appropriately, we use the information from the 7CSSI code of DSSP for the residue under consideration and its five neighbor residues up- and downstream in the sequence. Our 1CSSI code is created as a precise summary of the 8CSSI code of PSSC for a single residue. Hence, the creation of the 1CSSI code is considerably simplified and can be done without considering the 8CSSI code of neighbor residues.

DISCUSSION

Differences between DSSP and PSSC Demonstrated with an Example. To demonstrate differences between DSSP[7,8] and PSSC in characterizing protein secondary structure, we have chosen as an example the protein PH0459 (PDB code 1X42; Figure 1) from the haloacid dehalogenase family. This protein comprises all eight secondary structure types of the 1CSSI code from the PSSC model. An especially interesting feature of this protein is a short helical motif (residues Asp180 to Mse194; see close-up in Figure 1 and Table 1) involving all three helix types, classified as H, G, and I. In the 1CSSI code of DSSP[7,8] (version 2.0.4) these π- and 3_{10}-helix components are not recognized, since the α-helical part in the center obscures the surrounding π- and 3_{10}-helix components which are classified as turns (T) instead (see Table 1). The classification results obtained with stride[10] are similar.

If a sliding sequence window of only 11 residues centered at the residue in focus is considered, classification results of helical structures obtained with DSSP[7,8] and stride[10] approach those obtained with PSSC, but become not identical. On the other hand, results obtained with PSSC do not change with the application of the SSW, since it strictly uses only the information on the 7CSSI code of DSSP, which is within the 11 residues sequence window (Table 1).

The SSW size was chosen large enough to include π-helical H-bonds that contain the residue at the window's center as donor or acceptor. (Note that for strand structures the application of SSW does not make sense.) However, even with the SSW approach the 1CSSI code of DSSP can still differ from the corresponding PSSC results (Table 1). This indicates that for the generation of the 1CSSI code DSSP employs effectively also structure information outside of the SSW of 11 residues. Hence, the information on all helix types is available in DSSP[7,8] using the 7CSSI code but is overwritten by neighboring α-helical structures.

Next we consider two residues in more detail. For Lys190 the 8CSSI code from PSSC reads "GHi_ _RSK" (Table 1), indicating a valid α-helix turn as well as a valid 3_{10}-helix turn and a singular (i.e., nonvalid) π-helix turn. R indicates the handedness of the helical structure to be right. Since α-helical structures are assigned first, the corresponding 1CSSI code is α-helix (H). For Gly188, the first three characters of the 7CSSI string read ">4<" indicating that this residue is part of a starting

dx.doi.org/10.1021/ci5000856 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

3-turn, an ending 5-turn and is bridged by a 4-turn. Without knowing the LSS for the residues preceding and following Lys190 it is not possible to determine whether this residue will be classified by DSSP as helical or not. However, since the residue Gly188 is embedded in an α-helix, it is assigned to this class.

Comparing the Two DSSP Versions 2.0.4 and 2.2.1. A useful measure of the quality of LSS classifications are the roots of weighted variances (RWVAR) in (φ, ψ)-space defined in Methods, eq 8. Smaller RWVAR values indicate better clustering in the (φ, ψ)-space. The first two entries in Table 2 show the RWVAR values from DSSP 2.0.4[7] valid until recently and from the updated version 2.2.1,[8] yielding 62.5° (58.1° if the structurally diverse coil class is not considered) for both DSSP models. Note that these RWVAR values are much smaller than the value of 130.2° obtained if assignment with eight classes is performed for an even distribution in the (φ, ψ)-space with equal magnitude of the individual classes. The two DSSP models differ only with respect to the order of assigning LSS to the three helix classes. Originally DSSP (version 2.0.4) filled the three helix classes starting with α-helix (H), then 3_{10}-helix (G), and finally π-helix (I) (in short, HGI). The new DSSP version 2.2.1[8] reverses the order of helix assignments according to IHG (i.e., π-, α-, and 3_{10}-helices). With the assignment order of HGI α-helical structures are preferred, while classification of π-helical structures is disfavored.[7,39] Although it is useful to emphasize the recognition of π-helices, the new procedure (IHG) in DSSP may now overestimate the occurrences of π-helices. Considering the Astral40 v1.75 database DSSP 2.0.4 classifies only 289 residues to be in the π-helix class, while this number increases to 10129, when DSSP 2.2.1 is used. Since these numbers are very small compared to the total number of 1.8×10^6 residues in the Astral40 v1.75 database, the corresponding RWVAR values are practically identical for the two DSSP versions (Table 2).

Table 2. Root of Weighted Variances (RWVAR), Equation 8, in (φ, ψ)-Space in Units of degrees for Different Eight-Class Local Secondary Structure (LSS) Classification Models Based on DSSP and PSSC[a]

LSS model	total RWVAR	RWVAR without coil
DSSP 2.0.4[c], HGI	62.478(0.047)	58.150(0.054)
DSSP 2.2.1[d], IHG	62.413(0.046)	58.063(0.051)
PSSC[e], HGI	62.144(0.051)	57.556(0.063)
PSSC[e], IHG	62.147(0.046)	57.560(0.056)
PSSC[e], IGH	62.170(0.041)	57.591(0.049)
PSSC[e], GHI	62.167(0.046)	57.587(0.060)
PSSC[e], GIH	62.169(0.039)	57.591(0.056)

[a]Standard deviations are given in parentheses. As a reference value, the root of the variance in (φ, ψ)-space is formally for a single class model 107.5°. The RWVAR are also given without the contribution from the very diverse coil class C (right column). [b]The standard deviation is computed as follows. 10^6 residues are selected at random 50 times out of $N_{total} = 1.8027 \times 10^6$ residues of the Astral40[34] database (version 1.75) of protein domains resulting in a Gaussian-like distribution. The standard deviation is computed as the square root of the variance of these distributions and downscaled by the factor $1/\sqrt{(1.8027)}$ to yield the corresponding standard deviation of the full data set of 1.8027×10^6 residues. [c]DSSP version 2.0.4,[7] where the assignment to helix classes follows the sequence HGI, i.e., α-, 3_{10}-, and π-helices, preferring α-helix. [d]DSSP version 2.2.1,[8] where the assignment to helix classes follows the sequence IHG, i.e., π-, α-, and 3_{10}-helices, preferring π-helix. [e]PSSC using the same LSS classes as DSSP. The assignments to the helix classes are made according the listed sequence of classes.

Comparing Secondary Structure Classification of DSSP and PSSC. Using PSSC with settings that should reproduce DSSP 2.0.4 results, we obtain 98.34% agreement. The small albeit significant discrepancy is due to several assignment problems inherent in DSSP which are avoided in PSSC. This small difference leads to RWVAR values that are 0.3−0.4° lower than the values obtained with DSSP 2.0.4. This decrease is even larger if the relatively diverse coil class is not considered (right column in Table 2) indicating that improvements in LSS classification using PSSC are mainly made with respect to the other classes. It is interesting to note that the slightly better performance of LSS classification with PSSC measured in RWVAR values is practically independent from the order in which the three helix classes are filled.

A comparison of the square roots of variances of the individual classes is shown in Table 3. As can be inferred from the size of the variances the α-helix class possesses the most compact cluster in (φ, ψ)-space followed by clusters of 3_{10}-helices, β-strands, and β-bridge classes, which all three have similar variance sizes. The three most diverse clusters belong to classes describing bend (S) or straight (C) geometries without a specific H-bond pattern and the turn class (T).

The main effect leading to total variances (see Table 2) which are smaller for the LSS models based on PSSC than on DSSP is due to re-distribution of residues in clusters with lower variances. This is particularly the case for the turn class whose size shrinks by about 10% by applying PSSC instead of DSSP classification (see Table 3). Most of the residues removed from the turn class appear in the 3_{10}-helix class (G) which forms a cluster of considerably smaller variance.

The confusion matrix shown in Table 4 compares the LSS classification results of the 1CSSI code from DSSP (HGI)

Table 3. Square Roots of Variances According to Equation 7 of the Eight Individual LSS Classes in (φ, ψ)-Space in Units of Degrees for Classification Models Based on DSSP and PSSC[a]

LSS classes	C	S	T	E	H	G	B	I
DSSP HGI	77.80	112.20	90.08	46.32	17.07	44.11	48.91	60.48
DSSP IHG	77.80	112.20	90.10	46.32	16.98	44.11	48.91	34.32
PSSC HGI	78.03	112.36	92.57	43.59	17.06	43.64	48.72	50.58
PSSC IHG	78.03	112.36	92.57	43.59	16.97	43.66	48.72	35.06
PSSC IGH	78.03	112.36	92.57	43.59	38.66	16.63	48.72	35.06
PSSC GHI	78.03	112.36	92.57	43.59	38.65	16.73	48.72	50.58
PSSC GIH	78.03	112.36	92.57	43.59	38.65	16.63	48.72	35.09

[a]The LSS models are the same as those shown in Table 2. Data are generated using the 1.8027 × 106 residues of the Astral40 database (version 1.75) of protein domains. Standard deviations are of the same size as in Table 2.

dx.doi.org/10.1021/ci5000856 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 4. Confusion Matrix of Eight LSS Classes[a]

1CSSI from	PSSC E	PSSC B	PSSC H	PSSC G	PSSC I	PSSC T	PSSC S	PSSC C	sum
DSSP E	21.01	0.12	0.00	0.00	0.00	0.05	0.11	0.24	21.53
DSSP B	0.00	1.07	0.00	0.00	0.00	0.00	0.00	0.00	1.07
DSSP H	0.00	0.00	33.71	0.00	0.00	0.00	0.00	0.00	33.71
DSSP G	0.00	0.00	0.00	3.69	0.00	0.00	0.00	0.00	3.69
DSSP I	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.02
DSSP T	0.00	0.00	0.00	0.98	0.16	10.25	0.00	0.00	11.39
DSSP S	0.00	0.00	0.00	0.00	0.00	0.00	9.04	0.00	9.04
DSSP C	0.00	0.00	0.00	0.00	0.00	0.00	0.00	19.55	19.55
sum	21.01	1.19	33.71	4.67	0.18	10.30	9.15	19.78	100.00

[a]1CSSI code from DSSP version 2.0.4 versus PSSC with helix assignment order of HGI. Values are given in percent. Data are generated using the 1.8027×10^6 residues of the Astral40[34] database (version 1.75) of protein domains.

version 2.0.4 with results of the 1CSSI code from PSSC (HGI), where HGI indicates the order in which the three helix classes are assigned. The largest differences occur for 3_{10}-helices (G), turns (T), and β-strands (E). The size of the 3_{10}-helix class G increases at the expense of the turn class by nearly 1% referring to the total number of considered residues, while the relative increase in the size of class G is more than 26%. The β-strand class of DSSP diminishes by 0.5%, corresponding to a relative size reduction of about 2.3%. The 0.5% residues that are removed from the β-strand class if PSSC is used appear in classes C, B, and S, with 0.24%, 0.12%, and 0.11%, respectively. Although the size of the π-helix class (I) is very small, regardless of whether DSSP or PSSC is used, the relative increase from 0.02% to 0.18%—the latter obtained with PSSC—is dramatic. Since the PSSC model, applied for the generation of the confusion matrix, assigns the helix classes in the same order (HGI) as does DSSP, an increase in π-helix class cannot come from the α-helix class. The new DSSP, version 2.2.1,[8] assigns the π-helix first, leading to a much larger size of the π-helix class. Nevertheless, such a classification scheme may overestimate the occurrence of π-helices as will be discussed in the next section.

3_{10}-helix-like turns often appear at the end of an α-helix. With DSSP, turn motives in such an environment are assigned to the turn class (T), but PSSC recognizes the helical environment of such turns and therefore assigns the corresponding residue to helix class G. The discrepancies in the β-strand assignment are related to the fact that DSSP allows for quite large β-bulges (involving up to four residues). They are imperfections in a strand-like motive that do not possess a β-bridge and are commonly found in the non-β-strand region of the Ramachandran plot. A recent study[41] found no significant conservation of β-bulges among structural homologues in contrast to earlier studies with less data.[42,43] For β-bulges larger than one residue, PSSC refrains from an assignment to β-strands. Hence, PSSC is more generous in helix assignment but more stringent in strand assignment than is DSSP.

Occurrences of π-Helices. The occurrence of π-helices in protein structures was for many years considered to be very rare. In part this was due to the assignment strategy of DSSP[7−9] and stride,[10] which both favored α-helices in cases with ambiguities. The work of Fodje and Al-Karadaghi[20] corrected this view by analyzing the H-bond energies of the respective helix types and required that at least two subsequent H-bonds are of the π-helix type (i−1, i+4 and i, i+5) to assign the residues in the corresponding window of seven residues, i.e., i−1 and i+5, to a π-helix. This assignment strategy differs from PSSC and DSSP, which both are not assigning the initial and final residues involved in π-helix H-bonds to the π-helix class. Thus, for the minimum size window of seven residues from i−1 to i+5, the residues i−1 and i+5 would not belong to the π-helix class. Fodje and Al-Karadaghi[20] considered 932 polypeptide chains and identified a total of 104 π-helices. Unfortunately, the program SECSTR used for this study is not available any more.

More recently, using the same criteria, a much more extensive study on the occurrence of π-helices in protein structures was performed by Karplus et al.[31] on a large subset of the PDB[12] involving 14197 polypeptide chains with less than 90% sequence identity. With their program π-HUNT they found a total of 2897 π-helix segments. This number is slightly lower than the 2967 π-helices given by the authors, since we merged several interconnected pairs of π-helix segments. Compared to PSSC(HIG) using the classification sequence α-helix, π-helix, and 3_{10}-helix, we obtain 3663 π-helices. From these, 2786 are related (have at least one residue in common) to the π-helices found with π-HUNT; 877 were not found by π-HUNT; 111 were not found by PSSC. Using PSSC with assignment sequence (IHG) and comparing again with π-HUNT, these numbers change to 3958, 2786, 1064, and 3, respectively. The latter results are very close to the ones obtained with the new DSSP version 2.2.1[8] which pays attention to assign helical residues to the π-helix class. But, for the same reason 3_{10}-helices will still be underrepresented even by the version of DSSP.

We recommend for PSSC the assignment sequence (HIG) as used in the classical DSSP up to version 2.0.4. PSSC(HIG) fills first the α-helix class assigning residues of mixed helix types (α- and π-helices or α- and 3_{10}-helices) to the α-helix class analogue to the classical DSSP2.0.4. But, it differs from DSSP, since PSSC considers also residues with partial π-helix character to enlarge the necessary window size of at least seven π-helical residues to assign a residue with pure π-helix character to the π-helix class. As a consequence even small patches of one to four residues can be π-helical, if the neighbor residues, assigned to the α-helix class are of mixed α- and π-helix character. π-HUNT does not consider the mixed helix character of some residues. Instead it enforces assignments to specific helix classes using the DSSP energy function for H-bonds. As a consequence, it does not recognize such small π-helical patches, which is the main reason for the smaller number of π-helices found by π-HUNT as compared to PSSC. DSSP2.0.4 assigns such residues to the turn class although these residues may form a bulky helical turn in the middle of an intact α-helix.

dx.doi.org/10.1021/ci5000856 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Comparing the results of π-HUNT with PSSC(HIG) on the basis of the number of residues, π-HUNT classifies 21248 residues to be π-helical, while with PSSC(HIG) the number is only 6643. Since in PSSC initial and final residues involved in the H-bond scheme of a helix are not considered to belong to the helix, we need to increase the 6643 π-helical residues found with PSSC(HIG) by twice the number of π-helices, i.e., $2 \times 3663 = 7326$, before we can perform an essentially quantitative comparison. The corrected number of π-helical residues to be compared with 21248 π-helical residues obtained with π-HUNT is then 13969. Hence, the energy criterion used by π-HUNT yields about 50% more π-helical residues than does PSSC(HIG).

Clustering the π-Helices. All results shown in the following section refer to the Astral40[34] data set of 10569 different protein domain structures. The cluster of the π-helix class (I) plays a special role (SI Figure S1, center). Its RWVAR value is relatively large if it is assigned after the α-helix class, while the RWVAR value is much smaller corresponding to a relatively compact cluster if assigned before the α-helix class, which is valid for DSSP and PSSC (Table 3). In the latter case the π-helix class may contain also elements of the α-helix class, which yield a more compact cluster. However, since the total number of residues belonging to the π-helix class is very small, 3238 (10637) if the LSS model PSSC with helix order HGI (IHG) is used, its influence on the value of the weighted variances (RWVAR) of all eight classes shown in Table 2 is small.

Interestingly, the number of residues assigned to π-helix class is nearly the same for PSSC and DSSP, if π-helices are assigned first, namely, 10637 and 10129, respectively. These numbers differ considerably if π-helices are assigned last, namely, 3238 and 289, respectively. The difference of the number of π-helix residues using the assignment strategy IHG versus HGI in PSSC is 7399. These are precisely the residues with mixed α- and π-helix character (see also the Venn diagram, Figure 4).

Direct Transitions between Helices and Strands. In the Astral40[34] database (version 1.75) there are less than 100 cases, where a residue belongs to an H-bond pattern that is simultaneously helical and strand-like. On the other hand a direct transition between strand and helix without a coil residue in between can be found quite often in the Astral40 database. For a β-strand followed directly by a 3_{10}-, α-, or π-helix, we found 2069, 1779, and 10 cases, respectively, while there are 1814, 304, and 13 examples of helix residues of the respective type followed directly by a β-strand residue. These numbers should be compared with the number of helical segments and β-strands rather than with the number of residues in such motifs. Since there are around 72910 β-strands and 36691, 56051, and 2004 3_{10}-, α-, and π-helices in the Astral40[34] database, we calculated that more than 8% of all β-strands are directly preceded or followed by a helical residue and more than 6% of all helices are directly connected to a β-strand. In agreement with ref 44, we found that α-helices are only rarely followed by β-strand residues. The 304 examples that we found correspond to 0.4% of strand motifs. For both cases, strands followed by helices or vice versa, 20 examples are shown in Figure 3.

Influence of the Order of Assignment of LSS Classes. Since the LSS assignment strategies of DSSP and of PSSC are nearly exclusively based on the H-bond pattern of the polypeptide backbone, LSS with a dominant H-bond pattern should be assigned first. Hence, the LSS assignment should begin with helix and strand structures. We have demonstrated above that there is practically no ambiguity between α-helix and β-strand residues, such that the assignment order between strands and helices can have no significant influence on the result of an LSS classification. The same argument applies to the β-bridge class (B), which exhibits an isolated β-strand-like H-bond pattern. Changing the assignment order between the three helix classes and β-bridge class yields practically the same results. Hence, the LSS assignment of DSSP and PSSC starts with β-strand, followed by β-bridge and then the helix classes. In the preceding section we have shown that the order of assignment in LSS classes is critical for the three helix classes. In particular there is a large difference in the number of occurrences and (φ, ψ)-space variance (see Table 3) of the π-helix if the assignment of the π-helix class is performed before or after the α-helix class.

Based on the relevance of the H-bond pattern, the three turn classes are assigned next. They involve a helix-type H-bond but are isolated and are therefore not assigned in one of the three helix classes. Regarding the H-bond priority the β-bridge class is of the same level as the turn classes. Hence, the question arises whether the assignment order between the β-bridge and turn classes has an influence on the results. If the β-bridge class (B) is assigned before the turn class (T), which practically corresponds to the suggested assignment order, the numbers

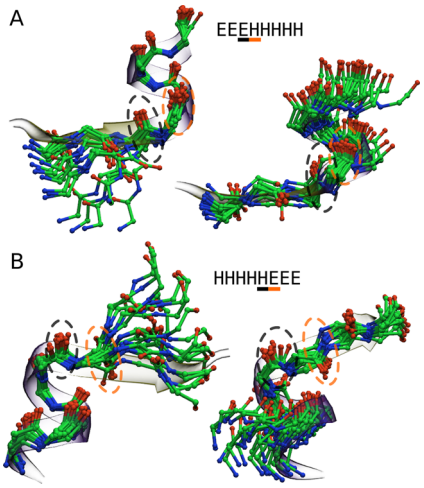


Figure 3. Examples of strand to α-helix (part A) and α-helix to strand (part B) transitions found in the Astral40 v1.75 database (looking from N- to C-terminus). The black and orange dashed cycles mark the two residues, which are on either side of the transition point. Left and right sides of each part show the same 20 randomly selected example structures from a total of 1224 and 157 occurrences, respectively. The Astral ids of the displayed example structures are listed in Supporting Information Table S2. The structures on the left side are structure-aligned with respect to the helix part. The structures on the right side are structure-aligned with respect to the strand part. Part A: The structure motif of type EEEHHHHH has 1224 occurrences. The large majority of 1137 of these possess dihedral angles whose values are in line with the PSSC assignment. Only in 52 cases the first helical residue does not possess dihedral angles corresponding to a helical conformation, which are not among the 20 displayed examples. Part B: The structure motif of type HHHHHEEE has 157 occurrences. Also here the majority of 111 structures possess dihedral angles whose values are in line with the PSSC assignment. In 37 cases the dihedral angles of the second strand residue are closer to a helical than to an extended strand-like conformation—even though the first strand residue is stretched. These are not included in the set of 20 displayed structures.

dx.doi.org/10.1021/ci5000856 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

of occurrences in class B and class T are 21300 and 185764, respectively. If the order of assignment is reversed the corresponding numbers are 20658 and 186406. Since the magnitude of the two classes is not too large and the variation with the order of assignment is not dramatic, the following assignment order can be maintained: β-strand; β-bridge (the latter is placed here, since the criteria for assignment of β-strand and β-bridge are practically the same, which facilitates classification); helices; turns. In DSSP the classes S (bent geometry without backbone H-bonds) and C (unspecific structure, so far unclassified) follow. Here, alternative classes may be assigned according to other criteria, such as AA type. The DSSP and all PSSC models for secondary structure characterization agree with this order of assignment.

Introducing New LSS Classes. We now demonstrate with several examples how one can generate specific classification models of LSS. In structural biology the traditional LSS scheme with three classes—helix, strand, and other (or coil)—may be detailed enough. Nevertheless, the precise definition of these classes can be critical for residues which are in a transition region between two different LSS classes. To characterize such ambiguous, mixed LSS areas, the introduction of additional classes may be helpful. The 8CSSI code of PSSC can be understood as an LSS profile code in a sense similar to that of sequence profiles[45] which have boosted the quality of LSS prediction considerably.[37] A similar and additional effect can be expected from an LSS profile. The 8CSSI code of PSSC opens the possibility of re-defining established LSS classes and of introducing new LSS classes which can be used in different classification schemes. There are a number of ways of doing this, as we discuss next.

We can introduce new types of LSS classes, which comprise mixtures of conventional LSS classes. Besides the helix, strand, and coil classes (H, E, and C) we may introduce the corresponding mixed classes HC, HE, and EC. More specifically, we can introduce mixed helix classes HG and HI as transitions between the pure helix classes H, G, and I for the α-helix, 3_{10}-helix, and π-helix, respectively. Even when bifurcated H-bonds are neglected, each residue can take part in one in- and one outgoing H-bond involving the backbone groups C−O and N−H, respectively. Those two H-bonds may or may not be of the same type (turn, bridge, or unclassified). Hence, based on the evaluation of the H-bond pattern, a residue may be assigned to two different LSS motifs. The decision of which motif to choose is to some degree arbitrary. This is of special importance when looking at residues in a helix-type structure. They may possess one α-helical and one π- or 3_{10}-helical H-bond with its backbone groups. As can be seen in Figure 4, residues that are helical but not α-helical are quite often mixed according to PSSC. For example, a residue that belongs to classes G and H could be considered as being part of a 3_{10}- or α-helix by DSSP, if only the respective local helix segment would be examined. Different applications may lead to different preferences when a residue lies in the intermediate region of two helical motifs. By using our proposed PSSC notation this decision is postponed and can be done easily as the last step of secondary and/or supersecondary structure assignment.

The occurrence of mixed helix structures is displayed in the Venn diagram, Figure 4, part A, based on a classification with PSSC. For this special classification model only the first three characters of the 8CSSI code are used. Accordingly, a residue with one or more uppercase letters G, H, or I is assigned to the

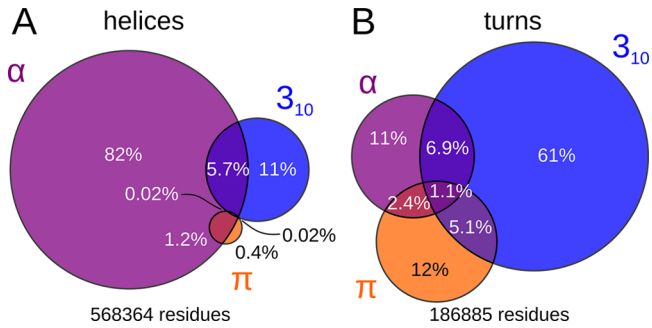


Figure 4. Venn diagram demonstrating the mixing of different helix types (3_{10}, α, π; part A) and mixing of the respective turn types (part B) characterized with PSSC. The Astral40[34] database (version 1.75) involving a total of 1.8027×10^6 residues is used. For more details see text.

responding helix classes 3_{10}, α, or π. A residue with one or multiple lowercase letters (g, h, or i) is handled in the same way but assigned to one of the turn classes, displayed in part B. Residues with mixed upper- and lowercase letters are not considered in Figure 4. Accordingly, 82% of all helix-type residues are pure α-helical, 11% are pure 3_{10}-helical, and 5.7% are mixed α- and 3_{10}-helical. From the 1.6% residues with π-helix character 0.4% are pure π-helical, while 1.2% are mixed with α-helical character. Only a very small percentage (0.02%) is mixed π- and 3_{10}-helical. Based on these considerations it may be more useful to introduce a mixed α- and 3_{10}-helical class (HG) than a pure π-helix class (I). Only a very small fraction of helical residues, namely, 0.02% (86 of 568364) belong to all three helix classes simultaneously.

A similar analysis can be performed for the turn-like residues, whose results are shown in Figure 4 part B. A definition of the three turn-types and an explanation of when they are used is given in Figure 2 and the corresponding caption. Here, the turn-type g corresponding to the H-bond pattern of 3_{10}-helices dominates with 61% of all turn-like residues. The turn-types i and h corresponding to an H-bond pattern of π- and α-helices involve 12% and 11% of all turn-type residues, respectively. More than 1% of all turn-like residues belong to all three turn types simultaneously. Mixed turn-types among turns occur more often than mixed helix types among helices. These considerations show that a partitioning of the turn class (T) into several classes may be very useful.

Using PSSC, the strand class (E) can be partitioned in E_P(15%), E_A(51%), describing pairs of parallel or antiparallel strands as well as the classes E_{PP}(8.2%), E_{AA}(21%), E_{PA}(4.8%) describing strands, which are flanked by two parallel or two antiparallel strands or one of each type, respectively. Besides pure structural features AA types can also serve to characterize new types of LSS classes although this is primarily sequence and not structure information. Good candidates are glycine and proline, which both can have a strong influence on LSS of the polypeptide backbone of a residue.

Two Examples for Alternative LSS Characterizations Using Eight Classes. The 8CSSI code of PSSC introduced in this work is suitable for creating new classification schemes for protein secondary structure. We demonstrate this here with two examples involving eight classes in analogy to the DSSP[7] approach. They are listed in Table 5, where we also show the resulting RWVAR value in the (φ, ψ)-space.

In the first example (alt1 in Tables 5 and 6) the S, B, and I class of DSSP are deleted. The classes S and B (bent geometry

dx.doi.org/10.1021/ci5000856 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 5. Root of Weighted Variances (RWVAR), Equation 8, in (φ, ψ)-Space in Units of degrees for Two Eight-Class LSS Classification Models Based on PSSC[a]

LSS model	total RWVAR	RWVAR without coil
PSSC alt1[c], E,H,G,TR,TL,CL,Co,CR	59.772(0.050)	46.949(0.045)
PSSC alt2[c], E,H,h,T,S,C,Gly,pre	61.418(0.049)	57.204(0.069)

[a]Standard deviations are given in parentheses.[b] The RWVAR are also given without the contribution from the very diverse coil class C (right column). [b]The standard deviations are computed as explained in the footnote to Table 2. [c]PSSC, using alternative classes: CL, Co, CR for coil structures with left, ambivalent, right handedness; TL, TR for left, right handed turn structures; h for the joint 3_{10}- and π-helix classes; Gly for glycine and pre, if the following residue is proline.

without H-bonds and β-bridge) are merged in three new coil classes (CL, Co, CR), which are defined with the handedness information (L, o, R) given in position 6 of the 8CSSI code of PSSC. Similarly the turn class is split into two classes, TL and TR, accounting for handedness. The variance of the larger of the two turn classes, TR, is now smaller than of the former single turn class T. Deleting S class and re-distributing its members are beneficial since this class possesses by far the largest variance (see Table 3). However, now the variances of the two coil classes (CL and CR involving a larger part of the former S class) are larger (Table 6). The π-helix class (I) is skipped since it is very small and is merged with the α-helix class. The resulting RWVAR of 59.8° is considerably smaller than that obtained with the traditional eight class scheme of DSSP (see Table 2). One reason for this reduction in variance may be due to the fact that the handedness information is used which, according to eqs 1 and 2, relies on the (φ, ψ)-angles. On the other hand it should be noted that the variances in (φ, ψ)-space are not necessarily the only quality criterion for an appropriate secondary structure classification scheme.

In the second example for an eight-class classification scheme of secondary structure (alt2 in Tables 5 and 6) class B is given up. Residues from the former class B are assigned to the strand class E if they do not possess a bent marker "S" at the same time. The two minority helix classes G and I are merged in the new helix class h. The five classes (C, E, S, T, H) of the traditional 8 class scheme are kept. In addition to class h, two new classes are introduced which use AA information, the glycine class "Gly" and the class "pre". The pre class denotes whether a proline residue follows the residue under consideration. It is known that in such a case the backbone geometry of a residue is considerably constrained.[46] For LSS classification the AA information (Gly and pre) has lower priority than the backbone H-bond information involved in helix, β-strand, β-bridge, and turn. However, it has higher priority than the structure information on the classes S and C. Hence, for the alt2 LSS model the assignment of classes follows the sequence E, H, h, T, Gly, pre, S, C. For the alt2 LSS model, the T and S classes have considerably smaller variances and the C class has a slightly larger variance compared with the traditional classification scheme at the expense of the new class Gly which has a rather large variance. In this case the RWVAR is also smaller than for the traditional classification scheme (compare the weighted variances listed in Tables 2 and 5.

LSS Characterization Using Three Classes. In most applications a three-class characterization [strand, helix, and other (E, H, C)] of protein secondary structures is used. For DSSP based LSS models the eight classes E, B, H, G, I, T, S, C are usually converted to three classes by merging the three helix classes (H, G, I), keeping the β-strand class (E) alone and merging the remaining four classes, β-bridge, turn, bend and coil (B, T, S, C), to the corresponding united coil class resulting in the three united classes H, E(≡E), C. The same merger is applied for the PSSC models. As a consequence the corresponding RWVAR values in (φ, ψ)-space are nearly the same (Table 7), although there is a minor reduction of the RWVAR values for the PSSC based classification. This reduction is due to the different treatment of large β-bulges which PSSC does not include in the β-strand class.

Table 7. RWVAR Values in (φ, ψ)-Space after Merging the Corresponding Eight Class Models Considered in Tables 2 and 5 to Three United Classes E, H, and C Whose Definition Are Given in Text[a]

LSS model	total RWVAR
DSSP 2.0.4, HGI	71.247(0.039)
DSSP 2.2.1, IHG	71.119(0.038)
PSSC, HGI	70.879(0.039)
PSSC, IHG	70.879(0.042)
PSSC, IGH	70.879(0.052)
PSSC, GHI	70.879(0.056)
PSSC, GIH	70.879(0.043)
PSSC, alt1	71.077(0.044)
PSSC, alt2	70.609(0.037)

[a]Standard deviations are given in parentheses. They are computed as explained in the footnote to Table 2. The same notation as in Table 2 is used.

The three-class models based on the alternative eight-class models alt1 and alt2 of PSSC (defined with Table 5) are obtained as follows. For both LSS classification models the

Table 6. Square Roots of Variances (SQVAR) According to Equation 7 of the Eight Individual LSS Classes in (φ, ψ)-Space in Units of degrees[a]

model alt1	E	H	G	TR	TL	CL	Co	CR
SQVAR	43.59	17.07	43.64	45.39	91.95	98.73	72.25	102.77
rel size, %	21.01	33.71	4.67	6.24	3.62	19.16	8.65	2.92
model alt2	E	H	h	T	Gly	pre	S	C
SQVAR	43.53	16.64	37.91	73.72	130.83	63.10	104.57	78.78
rel size, %	21.87	31.19	7.07	7.62	4.68	3.66	6.88	17.04

[a]The eight classes correspond to two classification models alt1 and alt2 based on PSSC as introduced in Table 5. Data are generated using the 1.8027×10^6 residues of the Astral40[34] database (version 1.75) of protein domains. Standard deviations are of the same size as those in Table 5. The listing of the eight classes corresponds to the order in which the assignment was performed for the different classes.

dx.doi.org/10.1021/ci5000856 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

strand class (E) remains unchanged consisting of β-strands only. Also the united helix class (H) contains all three helix types resulting from a merger of H and G for alt1 and of H and h for alt2. For both models (alt1 and alt2) the third class (C) involves the five remaining classes. Also for these two LSS models the RWVAR shows nearly no difference to the other LSS models (Table 7). Individual square roots of variances (SQVAR) according to eq 7 are given in Table S1 of the Supporting Information.

METHODS

Handedness of Secondary Structure. In PSSC the handedness of a residue is defined by the sign of ϑ, the rotation angle per residue, that is computed according to[38]

$$\cos\left(\frac{\vartheta}{2}\right) = -0.8235 \sin\left(\frac{\psi+\varphi}{2}\right) - 0.0222 \sin\left(\frac{\psi-\varphi}{2}\right) \quad (1)$$

$$d \sin\left(\frac{\vartheta}{2}\right) = 2.999 \cos\left(\frac{\psi+\varphi}{2}\right) - 0.657 \cos\left(\frac{\psi-\varphi}{2}\right) \quad (2)$$

In eqs 1 and 2, φ and ψ are the backbone dihedral angles and d is the helical rise per residue parameter (given in angstroms). The factors in front of the trigonometric functions account for the rigid-bond geometry of the polypeptide backbone. A graphical representation of a (d, ϑ)-plot resulting from eqs 1 and 2 is given in Figure 5.

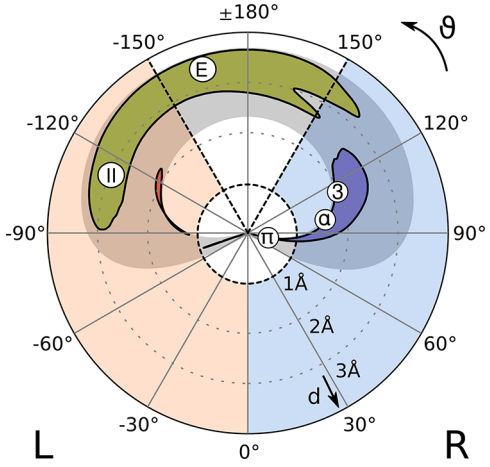


Figure 5. (φ, ψ)-backbone dihedral angles are used to calculate the generic helix parameters ϑ (angular step per residue) and d (rise per residue). See eqs 1 and 2. Shown are the commonly populated regions of this polar representation of the (d, ϑ)-plot[38] with the reference LSS motifs 3_{10}-, α-, and π-helices, extended strand (E), and the polyproline helix (II). The light blue area denotes significant right-handedness; the light red area, significant left-handedness. Areas with no significant handedness are in white background color.

Comparing the Quality of Different LSS Classification Schemes. The detailed characterization of LSS inherent in DSSP or PSSC can be used to create a large number of different classification schemes. However, it is not obvious how one can compare the quality of these classification schemes. Certainly one needs to consider the same number of classes. However, it is also important to define an unbiased and objective quality measure for secondary structure similarity and dissimilarity. The measure that we have chosen here is the sum of weighted variances of the clusters in the (φ, ψ)-space of the backbone dihedral angles, where the classes can graphically be represented by a Ramachandran plot.[39] For fair comparisons of different LSS classification schemes using this criterion, we must ensure sure that no direct information on the (φ, ψ)-angles is used.

To describe the weighted variance computation, we consider N_{total} residues which should be assigned to K_c different LSS classes. The (φ, ψ)-angles of a residue i in class k are denoted as $\vec{\alpha}_i^k = [\alpha_i^k(1), \alpha_i^k(2)] \equiv [\varphi_i^k, \psi_i^k]$. To compute the averages of (φ, ψ)-angles of residues in class k one needs to consider the periodicity appropriately. This can be achieved by transforming the angles in Cartesian coordinates on a circle with unit radius yielding

$$\vec{r}_i^k(j) = [x_i^k(j), y_i^k(j)] = [\cos(\alpha_i^k(j)), \sin(\alpha_i^k(j))], \quad j = 1, 2 \quad (3)$$

The average of $\vec{r}_i^k(j)$ within class k containing N_k residues is

$$\langle \vec{r}^k(j) \rangle = \frac{1}{N_k} \sum_{i=1}^{N_k} \vec{r}_i^k(j), \quad j = 1, 2 \quad (4)$$

After normalization of the vector $\langle \overline{\vec{r}}^k(j) \rangle = \langle \vec{r}^k(j) \rangle / |\langle \vec{r}^k(j) \rangle|$ we can regain the corresponding average angle $\langle \alpha^k(j) \rangle$, $j = 1, 2$, by inverting relation 3. To obtain the variance of this angle in class k, we subtract the average from the computed value $\Delta\alpha_i^k(j) = \alpha_i^k(j) - \langle \alpha^k(j) \rangle$ and account for periodicity as follows

$$\Delta\hat{\alpha}_i^k(j) = \min\{\Delta\alpha_i^k(j), \Delta\alpha_i^k(j) - \mathrm{sign}[\Delta\alpha_i^k(j)] \times 2\pi\}, \quad j = 1, 2 \quad (5)$$

The variance with respect to the angle $\alpha(j)$ can now simply be computed as the second moment

$$\langle [\Delta\hat{\alpha}^k(j)]^2 \rangle = \frac{1}{N_k} \sum_{i=1}^{N_k} [\Delta\hat{\alpha}_i^k(j)]^2, \quad j = 1, 2 \quad (6)$$

A measure of the total variance of class k is

$$\langle [\Delta\hat{\alpha}^k]^2 \rangle = \langle [\Delta\hat{\alpha}^k(1)]^2 \rangle + \langle [\Delta\hat{\alpha}^k(2)]^2 \rangle \quad (7)$$

Using the above sum of one-dimensional variances for the two-dimensional (φ, ψ)-space may overestimate the variance of clusters which have a large extension along one of the diagonals in (φ, ψ)-space. The resulting RWVAR of a classification result with K_C classes is

$$\langle [\Delta\hat{\alpha}]^2 \rangle^{1/2} = \left(\frac{1}{N_{total}} \sum_{k=1}^{K_c} N_k \langle [\Delta\hat{\alpha}^k]^2 \rangle \right)^{1/2} \quad (8)$$

Evaluating expression 8 for an even distribution of values in the (φ, ψ)-space, we obtain for the case of a single class a root of the variance of 368.4°. For 8 classes the RWVAR is reduced by the factor $8^{1/2}$ to 130.2°. When all residues in the Astral40 database version 1.75[34] are assigned to the same class, the RWVAR is 107.5°.

Database Used for Data Mining and Application. For the study presented in this work we used the well-established Astral40 database v1.75[34] involving 10569 different protein domains. The "pdbstyle" protein structure files of the Astral40 database were cleaned using the tool "pdbcur" from the CCP4[47] software suite, which performs the following tasks: (i) chain termination entries were added when missing and atom serial numbers were ensured to be consecutive; (ii) solvent

dx.doi.org/10.1021/ci5000856 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

residues were moved to the end of the structure files; (iii) in case of multiple occupancies, the most probable conformation was used; (iv) for PDB structure files based on NMR spectroscopy that usually contain several model structures, only the first was taken. Due to ill-formatted data, six protein structure files could not be processed by pdbcur and seven where skipped as they contain only C_α-coordinates. This resulted in 10556 protein domain structure files that were analyzed with DSSP[7,8] and stride.[10] The results files of DSSP and stride were subsequently processed by Biopython.[16,17] Residues with incomplete assigned data, which is for instance the case for residues at the beginning and end of a polypeptide chain were excluded. This led to a total of N_{total} = 1802758 residues. According to DSSP 2.0.4[7] these residues consist of 37% of helices, to 23% of strands (including 1% of singular β-bridges), and to 40% of other (coil class).

SUMMARY

With PSSC we have introduced a detailed eight-character secondary structure information (8CSSI) code to characterize protein structures. The conversion to specific one-letter codes using different numbers of classes is straightforward. The scope of information contained in the 8CSSI code of PSSC allows for creation of different models for secondary structure characterization. Such models may use different numbers and different types of secondary structure classes or even secondary structure profiles which allow mixed classification when unique class assignment is not appropriate. Hence, the recreation of the traditional 1CSSI code of DSSP is just one specific application to process the information from the 8CSSI code of PSSC (see Table 1, column 5).

PSSC uses essentially all information created by DSSP on a lower hierarchical level. An exception is the sign (+,−) of the pseudo-dihedral angle defined by C_α-atoms in DSSP. In PSSC this information is replaced by the handedness of local secondary structure given by the sign of ϑ, the rotation angle per residue, which is computed from the (φ, ψ)-angles; see eqs 1 and 2. Nearly all information used by DSSP is derived from the H-bond pattern. PSSC uses this as well, but it does not include information on the directionality of the H-bonds in its 8CSSI code. Since this information is not used by DSSP for secondary structure classification, PSSC can provide essentially the same classification results when used with appropriate settings. However, there are more significant differences between the detailed seven-character (7CSSI) code of DSSP and the corresponding 8CSSI code of PSSC. (i) For each individual residue the corresponding 8CSSI code of PSSC contains all information to generate the traditional eight-class model of DSSP directly, while DSSP needs to refer also to the 7CSSI code of the neighbor residues. This feature of PSSC facilitates the creation of alternative classification models for protein secondary structures. (ii) In DSSP, residues consisting of two β-strand segments connected by a β-bulge of up to four residues are considered as a single β-strand unit. PSSC merges two such strand segments only if they are interrupted by at most one residue. For larger interruptions PSSC classifies such a structure as two independent β-strands and assigns residues belonging to a β-bulge to the next appropriate class. (iii) Regarding helical structures, DSSP assigns turn residues of one type embedded in or attached to a helix of a different type to the turn class. In PSSC such a turn residue is assigned to the appropriate helix class. In this sense DSSP is more generous filling the β-strand class while PSSC is more generous filling the helix classes.

We have shown that residues assigned to a β-strand (E) can be immediately followed by residues marked as helical (H) and vice versa. These EH and HE motifs are accompanied by interwoven hydrogen bonding patterns, which are simultaneously of strand and helix type. Still, a sharp transition between helix and strand as performed by DSSP is often justified according to the values of the backbone (φ, ψ)-angles.

For helical residues we demonstrated that a significant part of them is not purely α-helical but can be a mixture of two different types of helices. This ambiguity is clearly visible in the full 8CSSI code of PSSC. We have shown that the preference of DSSP for α- over π-helices is an artifact of the order used for the assignment procedure. With PSSC a balanced assignment of α- and π-helices is obtained by removing the restraint of minimal length of helical segments while keeping the requirement of two consecutive H-bonds of the same 5-turn type, if the π-helix motif is isolated and not attached to another helix. This procedure alone enlarges the occurrence of π-helices by a factor of 10, even if the assignment of α-helices is done first. If on the other hand the assignment of π-helices is done first, the occurrence of them increases by another factor of 3.

Different LSS assignment models are compared using a weighted variance defined in (φ, ψ)-space, eq 8. It should however be noted that such a quantity is not necessarily the only legitimate quality measure for LSS characterization. One quality measure is certainly the acceptance of the proposed secondary structure characterizations by the community of structural biologists. Another quality criterion is the performance of secondary structure prediction devices using various secondary structure characterization tools. In a recent contribution using Sparrow[37] it was shown that for a 3-class LSS assignment the prediction performance increases from 80.46% to 82.51% if the helix class contains α-helices only and the π- and 3_{10}-helices are merged with the coil class.[37] This significant improvement is not due to a moderate shrinking of the size of the helix class from 37% to 34% of all considered residues. Rather this increase in prediction accuracy indicates that there is space for further improvement if different and more precisely defined secondary structure classes or even secondary structure profiles are used for learning and prediction.

We hope that PSSC will be beneficial for structural biologists and useful as a basis to predict protein secondary structure. A python script which generates the 8CSSI code of PSSC from DSSP is available for download at the webpage http://agknapp.chemie.fu-berlin.de/secsass. A DSSP version that directly generates PSSC 8CSSI code and calculates the handedness of residues is available, too.

ASSOCIATED CONTENT

Supporting Information

Figures showing mixed helices, chain V of protein 1T0T, a schematic of the H-bond pattern in the protein domain, a specific helix and strand dependent Ramachandran plot, and the haloacid dehalogenase family protein, text describing direct transitions between helices and strands and accompanying references, and tables listing comparisons of different secondary structures for protein 1X42 and astral ids of examples shown in Figure 3. This material is available free of charge via the Internet at http://pubs.acs.org.

dx.doi.org/10.1021/ci5000856 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

AUTHOR INFORMATION

Corresponding Author

*E-mail: knapp@chemie.fu-berlin.de.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors thank Dr. Nadia Elghobashi-Meinhardt for proofreading the manuscript.

ABBREVIATIONS:

1CSSI, one-character secondary structure information; 7CSSI, seven-character secondary structure information; 8CSSI, eight-character secondary structure information; AA, amino acid; DSSP, define secondary structure of proteins; HGI, helix assignment order: α-helix, 3_{10}-helix, and π-helix; IGH, helix assignment order: π-helix, 3_{10}-helix, and α-helix; LSS, local secondary structure; PSSC, protein secondary structure characterization; RWVAR, roots of weighted variance; SSW, sliding sequence window

REFERENCES

(1) Adzhubei, A. A.; Sternberg, M. J. Left-Handed Polyproline II Helices Commonly Occur in Globular Proteins. *J. Mol. Biol.* **1993**, *229*, 472−493.

(2) Toniolo, C. Intramolecularly Hydrogen-Bonded Peptide Conformations. *CRC Crit. Rev. Biochem.* **1980**, *9*, 1−44.

(3) Martin, J.; Letellier, G.; Marin, A.; Taly, J.-F.; de Brevern, A. G.; Gibrat, J.-F. Protein Secondary Structure Assignment Revisited: A Detailed Analysis of Different Assignment Methods. *BMC Struct. Biol.* **2005**, *5*, 17.

(4) Andreeva, A.; Howorth, D.; Brenner, S. E.; Hubbard, T. J. P.; Chothia, C.; Murzin, A. G. SCOP Database in 2004: Refinements Integrate Structure and Sequence Family Data. *Nucleic Acids Res.* **2004**, *32*, D226−D229.

(5) Orengo, C.; Michie, A.; Jones, S.; Jones, D.; Swindells, M.; Thornton, J. CATH—A Hierarchic Classification of Protein Domain Structures. *Structure* **1997**, *5*, 1093−1109.

(6) Garnier, J. Protein Structure Prediction. *Biochimie* **1990**, *72*, 513−524.

(7) Kabsch, W.; Sander, C. Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers* **1983**, *22*, 2577−2637.

(8) Hekkelman, M.; Kabsch, W.; Sander, C. http://swift.cmbi.ru.nl/gv/dssp/ (accessed Apr. 19, 2014).

(9) Joosten, R. P.; te Beek, T. A. H.; Krieger, E.; Hekkelman, M. L.; Hooft, R. W. W.; Schneider, R.; Sander, C.; Vriend, G. A Series of PDB Related Databases for Everyday Needs. *Nucleic Acids Res.* **2011**, *39*, D411−D419.

(10) Frishman, D.; Argos, P. Knowledge-Based Protein Secondary Structure Assignment. *Proteins* **1995**, *23*, 566−579.

(11) Sayle, R. A.; Milner-White, E. J. RASMOL: Biomolecular Graphics for All. *Trends Biochem. Sci.* **1995**, *20*, 374.

(12) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.

(13) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J. Mol. Graphics* **1996**, *14*, 33−38.

(14) Schrödinger, L. *The PyMOL Molecular Graphics System*; Schrödinger: Portland, OR, USA, 2010.

(15) Michaud-Agrawal, N.; Denning, E. J.; Woolf, T. B.; Beckstein, O. MDAnalysis: A Toolkit for the Analysis of Molecular Dynamics Simulations. *J. Comput. Chem.* **2011**, *32*, 2319−2327.

(16) Cock, P. J. A.; Antao, T.; Chang, J. T.; Chapman, B. A.; Cox, C. J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; de Hoon, M. J. L. Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics. *Bioinformatics* **2009**, *25*, 1422−1423.

(17) Hamelryck, T.; Manderick, B. PDB File Parser and Structure Class Implemented in Python. *Bioinformatics* **2003**, *19*, 2308−2310.

(18) Carter, P.; Andersen, C. A. F.; Rost, B. DSSPcont: Continuous Secondary Structure Assignments for Proteins. *Nucleic Acids Res.* **2003**, *31*, 3293−3295.

(19) King, S. M.; Johnson, W. C. Assigning Secondary Structure from Protein Coordinate Data. *Proteins* **1999**, *35*, 313−320.

(20) Fodje, M. N.; Al-Karadaghi, S. Occurrence, Conformational Features and Amino Acid Propensities for the Pi-Helix. *Protein Eng.* **2002**, *15*, 353−358.

(21) Cubellis, M. V.; Cailliez, F.; Lovell, S. C. Secondary Structure Assignment That Accurately Reflects Physical and Evolutionary Characteristics. *BMC Bioinf.* **2005**, *6* (Suppl 4), S8.

(22) Park, S. Y.; Yoo, M.-J.; Shin, J.; Cho, K.-H. SABA (Secondary Structure Assignment Program Based on Only Alpha Carbons): A Novel Pseudo Center Geometrical Criterion for Accurate Assignment of Protein Secondary Structures. *BMB Rep.* **2011**, *44*, 118−122.

(23) Richards, F. M.; Kundrot, C. E. Identification of Structural Motifs from Protein Coordinate Data: Secondary Structure and First-Level Supersecondary Structure. *Proteins* **1988**, *3*, 71−84.

(24) Labesse, G.; Colloc'h, N.; Pothier, J.; Mornon, J. P. P-SEA: A New Efficient Assignment of Secondary Structure from C Alpha Trace of Proteins. *Comput. Appl. Biosci.* **1997**, *13*, 291−295.

(25) Taylor, W. R. Defining Linear Segments in Protein Structure. *J. Mol. Biol.* **2001**, *310*, 1135−1150.

(26) Srinivasan, R.; Rose, G. D. A Physical Basis for Protein Secondary Structure. *Proc. Natl. Acad. Sci. U. S. A.* **1999**, *96*, 14258−14263.

(27) Konagurthu, A. S.; Lesk, A. M.; Allison, L. Minimum Message Length Inference of Secondary Structure from Protein Coordinate Data. *Bioinformatics* **2012**, *28*, i97−i105.

(28) Tyagi, M.; Bornot, A.; Offmann, B.; de Brevern, A. G. Analysis of Loop Boundaries Using Different Local Structure Assignment Methods. *Protein Sci.* **2009**, *18*, 1869−1881.

(29) Andersen, C. A. F.; Rost, B. Secondary Structure Assignment. *Methods Biochem. Anal.* **2003**, *44*, 341−363.

(30) Majumdar, I.; Krishna, S. S.; Grishin, N. V. PALSSE: A Program to Delineate Linear Secondary Structural Elements from Protein Structures. *BMC Bioinf.* **2005**, *6*, 202.

(31) Cooley, R. B.; Arp, D. J.; Karplus, P. A. Evolutionary Origin of a Secondary Structure: Π-Helices as Cryptic but Widespread Insertional Variations of A-Helices That Enhance Protein Functionality. *J. Mol. Biol.* **2010**, *404*, 232−246.

(32) Pal, L.; Basu, G. Novel Protein Structural Motifs Containing Two-Turn and Longer 310-Helices. *Protein Eng., Des. Sel.* **1999**, *12*, 811−814.

(33) Baker, E. N.; Hubbard, R. E. Hydrogen Bonding in Globular Proteins. *Prog. Biophys. Mol. Biol.* **1984**, *44*, 97−179.

(34) Chandonia, J.-M.; Hon, G.; Walker, N. S.; Lo Conte, L.; Koehl, P.; Levitt, M.; Brenner, S. E. The ASTRAL Compendium in 2004. *Nucleic Acids Res.* **2004**, *32*, D189−D192.

(35) Rost, B.; Sander, C. Improved Prediction of Protein Secondary Structure by Use of Sequence Profiles and Neural Networks. *Proc. Natl. Acad. Sci. U. S. A.* **1993**, *90*, 7558−7562.

(36) Rost, B. Rising Accuracy of Protein Secondary Structure Prediction. In *Protein Structure: Determination, Analysis, and Applications for Drug Discovery*; Chasman, D., Ed.; CRC Press: New York, 2003; pp 207−249.

(37) Bettella, F.; Rasinski, D.; Knapp, E. W. Protein Secondary Structure Prediction with SPARROW. *J. Chem. Inf. Model.* **2012**, *52*, 545−556.

(38) Zacharias, J.; Knapp, E. W. Geometry Motivated Alternative View on Local Protein Backbone Structures. *Protein Sci.* **2013**, *22*, 1669−1674.

(39) Ramachandran, G. N.; Ramakrishnan, C.; Sasisekharan, V. Stereochemistry of Polypeptide Chain Configurations. *J. Mol. Biol.* **1963**, *7*, 95−99.

(40) MacArthur, M. W.; Thornton, J. M. Influence of Proline Residues on Protein Conformation. *J. Mol. Biol.* **1991**, *218*, 397−412.
(41) Craveur, P.; Joseph, A. P.; Rebehmed, J.; de Brevern, A. G. B-Bulges: Extensive Structural Analyses of B-Sheets Irregularities. *Protein Sci.* **2013**, *22*, 1366−1378.
(42) Chan, A. W.; Hutchinson, E. G.; Harris, D.; Thornton, J. M. Identification, Classification, and Analysis of Beta-Bulges in Proteins. *Protein Sci.* **1993**, *2*, 1574−1590.
(43) Chen, P. Y.; Gopalacushina, B. G.; Yang, C. C.; Chan, S. I.; Evans, P. A. The Role of a Beta-Bulge in the Folding of the Beta-Hairpin Structure in Ubiquitin. *Protein Sci.* **2001**, *10*, 2063−2074.
(44) Fitzkee, N. C.; Rose, G. D. Steric Restrictions in Protein Folding: An Alpha-Helix Cannot Be Followed by a Contiguous Beta-Strand. *Protein Sci.* **2004**, *13*, 633−639.
(45) Altschul, S. F.; Madden, T. L.; Schäffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs. *Nucleic Acids Res.* **1997**, *25*, 3389−3402.
(46) Lovell, S. C.; Davis, I. W.; Arendall, W. B.; de Bakker, P. I. W.; Word, J. M.; Prisant, M. G.; Richardson, J. S.; Richardson, D. C. Structure Validation by Calpha Geometry: Phi,psi and Cbeta Deviation. *Proteins* **2003**, *50*, 437−450.
(47) Winn, M. D.; Ballard, C. C.; Cowtan, K. D.; Dodson, E. J.; Emsley, P.; Evans, P. R.; Keegan, R. M.; Krissinel, E. B.; Leslie, A. G. W.; McCoy, A.; McNicholas, S. J.; Murshudov, G. N.; Pannu, N. S.; Potterton, E. A.; Powell, H. R.; Read, R. J.; Vagin, A.; Wilson, K. S. Overview of the CCP4 Suite and Current Developments. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2011**, *67*, 235−242.
(48) Arai, R.; Kukimoto-Niino, M.; Kuroishi, C.; Bessho, Y.; Shirouzu, M.; Yokoyama, S. Crystal Structure of the Probable Haloacid Dehalogenase PH0459 from Pyrococcus Horikoshii OT3. *Protein Sci.* **2006**, *15*, 373−377.