JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

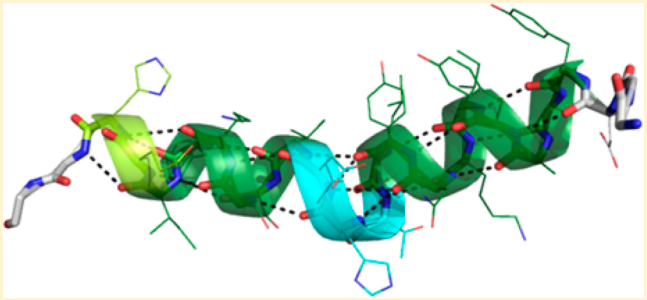
Dihedral-Based Segment Identification and Classification of Biopolymers I: Proteins

Gabor Nagy and Chris Oostenbrink*

University of Natural Resources and Life Sciences, Institute for Molecular Modeling and Simulation, Muthgasse 18, 1190 Vienna, Austria

Supporting Information

ABSTRACT: A new structure classification scheme for biopolymers is introduced, which is solely based on main-chain dihedral angles. It is shown that by dividing a biopolymer into segments containing two central residues, a local classification can be performed. The method is referred to as DISICL, short for Dihedral-based Segment Identification and Classification. Compared to other popular secondary structure classification programs, DISICL is more detailed as it offers 18 distinct structural classes, which may be simplified into a classification in terms of seven more general classes. It was designed with an eye to analyzing subtle structural changes as observed in molecular dynamics simulations of biomolecular systems. Here, the DISICL algorithm is used to classify two databases of protein structures, jointly containing more than 10 million segments. The data is compared to two alternative approaches in terms of the amount of classified residues, average occurrence and length of structural elements, and pair wise matches of the classifications by the different programs. In an accompanying paper (Nagy, G.; Oostenbrink, C. Dihedral-based segment identification and classification of biopolymers II: Polynucleotides. *J. Chem. Inf. Model.* 2013, DOI: 10.1021/ci400542n), the analysis of polynucleotides is described and applied. Overall, DISICL represents a potentially useful tool to analyze biopolymer structures at a high level of detail.



INTRODUCTION

Biopolymers like proteins and DNA are essential building blocks of all living organisms and understanding how they fulfill their biological functions is one of the most important tasks in life sciences. It is a widely accepted fact that many of the biopolymers (like proteins, RNA, oligosaccharides) form complex three-dimensional structures, and this structure is essential for their biological function.[1−3] To understand how structure defines function, it is often segmented into smaller parts and grouped together into structural classes based on common properties. Good examples are proteins of which individual functions are tied to separate domains, which may be recognized by a number of properly organized, smaller, secondary structure elements (helices, strands, turns, and coils). The structure of a biopolymer is usually classified based on even smaller segments and properties that are readily calculated from the 3D structure, like backbone hydrogen-bonding and backbone dihedrals.

The most basic secondary structure elements were originally derived from the structure of α-keratin and β-keratin[4] during the early 1950s, and were named α-helix[5] and β-sheet,[6] respectively. The highly repetitive nature and regular patterns of both hydrogen bonding and backbone dihedrals of these structure elements allowed identification even at low resolutions. The classical treatment of the protein secondary structure (α-helix, β-sheet, and random coil) already led to the discovery of key concepts in protein architecture (such as domains, and folds), but the technical advancements in X-ray crystallography and spectroscopic methods soon revealed new, less regular structural elements. While most of the newly discovered structural elements were nonrepetitive, comparative studies of protein structures showed that they are well-defined and not random at all. The most important group of these structural elements can be defined as tight turns. While defined in various ways since 1968,[7−11] the importance of the turn structures (especially β-turns) is to connect the strands of the β-sheets and α-helices and to allow the formation of folds and domains. The second group of new structural elements can be defined as distortions of classical α-helices (like the 3_{10}-helix[12] and the π-helix[13]) or β-sheets (like the β-bulges[14]). Structural elements of this second group are important both because of the functional role of the distortion and because of the structural stress they can remove from the overall fold of the proteins.

Structure-based classification programs are best established for protein analysis, and the most widespread method is called DSSP (Dictionary of Secondary Structure for Proteins).[15] This protocol uses the hydrogen bonding patterns along the protein backbone for classification and is quite robust in discriminating α-helices from β-strand structures, while also providing insight on the presence of turns. In recent years, protocols to improve

Received: September 18, 2013
Published: December 24, 2013

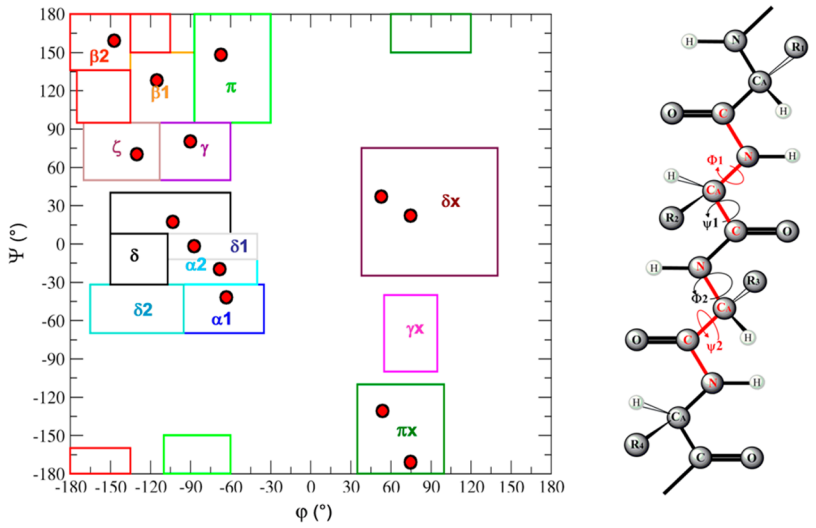


Figure 1. Representation of region definitions used for protein classification (on the left) based on subsequent (φ,ψ) values within a tetrapeptide segment (on the right). Colored rectangles show the boundaries of regions marked with Greek letters. Red dots show cluster centers of Hollingsworth et.al. Atoms and bonds that define φ1 and ψ2 are marked with red.

the protein structure classification have been proposed, such as the STRIDE algorithm[16] (STRuctural IDEntification), which combines hydrogen bonding and backbone dihedrals φ and ψ to provide further details. While (φ,ψ) backbone dihedrals are well known to be characteristic for the protein shape—and are often used by crystallographers to refine X-ray and NMR models—to our knowledge no purely dihedral-based classification tool exists for detailed protein structure classifications. However, Hollingsworth et al. recently provided an in depth 4D clustering study based on the (φ,ψ) pairs of tetrapeptide segments within proteins[17] (Figure 1). This work suggests that tetrapeptide segments are already characteristic for secondary structure and that a purely dihedral-based approach is possible for classification.

Here, we introduce a new segment-based structure classification protocol, which classifies biopolymers based on their backbone dihedral angles. We refer to it as DIhedral based Segment Identification and Classification or in short DISICL. The DISICL protocol is designed for the detailed comparison of multiple similar structures or to monitor dynamic changes of a biopolymer during molecular simulations. To demonstrate the potential of the approach, we perform a large-scale analysis on two databases of proteins downloaded from the Brookhaven protein database[18] and an analysis on a set of selected protein simulations.[19] Classifications are compared to the results of the already well-established analysis tools DSSP and STRIDE. The aim of this newly introduced classification method is to provide an alternative way to interpret the structural information stored in the 3D models and simulations. The quantitative comparison with different classification methods should help to decide on the most suitable approach for specific needs and not to judge the correctness of one approach. As classifications are always a matter of definitions, we emphasize that there is no "correct" or "wrong" answer. In the accompanying paper,[20] we extend the DISICL approach to the classification of polynucleotides and perform a similar analysis of DNA and RNA. The first application of a preliminary version of the DISICL algorithm was used to study the fine structural differences of Cytochrome C molecular dynamics in the oxidized and reduced state. Observations were correlated to a combined IR/Raman correlation spectroscopy approach to monitor the protein structural changes upon oxidation.[21]

METHODS

Data Sets. For the purpose of testing and comparing different classification algorithms, two large-scale protein data sets were obtained from the Brookhaven protein databank (PDB, http://www.rscb.org).[18] Both data sets were selected from all PDB entries available on October 23, 2012, using the following criteria. (1) Entries show at most 30% sequence identity. (2) Entries contain only one type of biopolymer. (3) Entries obtained from X-ray crystallography have a resolution of 0.8−2.0 Å.

One data set contained structures obtained from X-ray crystallography (Prot_Xr) and another one from nuclear magnetic resonance experiments (Prot_NMR). The resolution range for X-ray structures was chosen such that the backbone dihedrals can be reliably determined, but the number of alternative locations for groups of atoms in the data set is kept low. Prior to the analysis, alternative locations, nonstandard residues, cofactors, and nonbiopolymer elements were discarded. Multiple chains and multimeric structures were retained, but residues were renumbered to avoid identical residue numbers from different chains. Only those models were considered for which all three programs could classify at least one residue; the others were discarded. While this approach decreased the number of analyzed residues, combined NMR and X-ray data sets still provided about 10 million applicable segments of four residues. Further details are provided in the protein data set section of Table 1.

DSSP Algorithm. For comparison purposes, a publicly available version of the original algorithm of Kabsch and Sander[15] was downloaded (http://swift.cmbi.ru.nl/gv/dssp). DSSP is based on a hierarchical classification of hydrogen-bonded patterns along the protein backbone. First, the presence of hydrogen bonds is determined by an energy function, which allows for some deviation from the ideal backbone atom distances and bond angles. Second, local hydrogen bonding (within five residues) is recognized as 3-, 4-, or 5-turns, while hydrogen bonds further away are classified as bridges. In the third

dx.doi.org/10.1021/ci400541d | J. Chem. Inf. Model. 2014, 54, 266−277

Table 1. Summary of Analyzed Data Sets, Classification Efficiency of Various Algorithms, and Agreement between These Algorithms

protein data set			
database	Prot_Xr	Prot_NMR	combined set
file number	8,064	4,234	12,298
model number	7,592	74,530	82,122
total residues	3,218,726	6,826,846	10,045,572
multiplicity	0.9	17.6	6.7
ave. length	424.0	91.6	122.3
methods performance			
method	DSSP	DISICL	STRIDE
data set size	8,238,693	9,671,468	10,041,485
completeness (%)	82.0	96.3	99.96
classification ratio (%)	73.7	73.7	79.0
total efficiency (%)	60.4	71.0	79.0
methods agreement			
methods	DISICL/DSSP	DISICL/STRIDE	DSSP/STRIDE
helical match (%)	80.5	88.7	96.1
beta strand match (%)	74.5	83.7	96.9
turn struct. match (%)	33.1	55.6	68.2
overall match (%)	68.6	77.9	81.6

step, consecutive turn patterns are recognized as helices (3-, 4-, or 5-helix), and consecutive bridges are classified as β-strands. Finally, if no well-defined pattern can be found, but the Cα atoms around the residue shows a local curvature of more than 70°, the structure is classified as a bend. The DSSP algorithm provides seven classes (3-helix, 4-helix, 5-helix, β-strand, β-bridge, turn, and bend), which gives a clear separation between α-helices and β-strands and readily maps the connection between interconnected β-strands. Differentiation of parallel and antiparallel beta sheets is not performed by the algorithm but should be straightforward based on hydrogen-bonding patterns. The DSSP algorithm slightly favors certain classes (such as the 4-helix) because of its hierarchical nature, and finer details of the structure (like different turn types, left or right handed structures) are often lost. The downloaded algorithm only handled the first chain of each PDB entry, which led to a significant reduction of the analyzed structures for the X-ray protein data set.

STRIDE Algorithm. The program STRIDE is based on a similar approach as the DSSP algorithm but uses backbone dihedral angles as weighting parameters to sharpen the separation of α- and β-structures and gain additional information about turn structures. STRIDE was optimized to reproduce the visual assignments of well-trained crystallographers and tends to assign classes to ends of secondary structure elements, where DSSP is less robust. The STRIDE algorithm[16] is freely available through a Web server (webclu.bio.wzw.tum.de/stride), and on request, the source code is also provided. The standard output of STRIDE contains seven classes (3_{10} helix, α-helix, π-helix, β-strand, β-bridge, turn, and coil), which can be directly compared to the results of DSSP; the only significant difference lies in the coil structure, which contains everything that does not fit into any other class. In addition, STRIDE provides a more detailed classification of turns based on four amino acid segments, which contain various turn types (β turn types I, IV, VIII, Schellman turn, Gamma turns, etc.).

DISICL. This section describes the basic approach of our new dihedral angle-based classification tool, DISICL. The basic idea behind the algorithm presented in the current work is that the dihedral angles of a biopolymer backbone are characteristic for its shape, and if sufficiently long segments are taken, these alone can describe its shape and structure. DISICL was inspired by the work of Hollingsworth et al., who performed a large-scale 4D clustering study based on 76,000 ordered tetrapeptide segments.[17] Interestingly, a similar tetrapeptide segmentation is used in STRIDE to obtain more detailed classification of turn structures. More than 100 observed clusters were reported,[17] of which many had strong preferences toward specific secondary structure elements. The four coordinates of clustering were the two (φ,ψ) dihedral angle pairs of the central residues of the segment (the peptide bonds of the flanking residues are required for a complete definition of these angles, see Figure 1), which are the very same angles used in Ramachandran plots. On the basis of the observed cluster centers and the density map of this two-dimensional dihedral space, we defined 13 "Ramachandran" regions (shown in Figure 1 and Table S1, Supporting Information), which can be used to classify the segments into 18 different structural classes. The region definitions are given by rectangular areas grouped together to distinguish most of the cluster centers specific or selective for secondary structures, and at the same time, they cover the most densely populated areas of the dihedral angle space. Individual regions do not have overlaps in the (φ,ψ) space, and any combination of two subsequent pairs of (φ,ψ) angles that fall into specific regions leads to a single secondary structure assignment for the segment.

The classification by DISICL is performed in the following way: (1) Calculate the appropriate dihedral angles for the given biopolymer segment. (2) Assign central residues to the regions in the dihedral angle space. (3) Classify segments based on the regions assigned to the central residues. (4) Move on to the next segment.

Figure 2 shows a flowchart representation of the DISICL approach. Most of the region (Figure 1 and Table S1, Supporting Information) and class definitions (Table 2) can be directly derived from the clusters described by Hollingsworth et al.[17] with two exceptions. No cluster centers were defined for regions δ2 and γx, even though they show a moderate population in the (φ,ψ) dihedral angle space. On the basis of the position of these

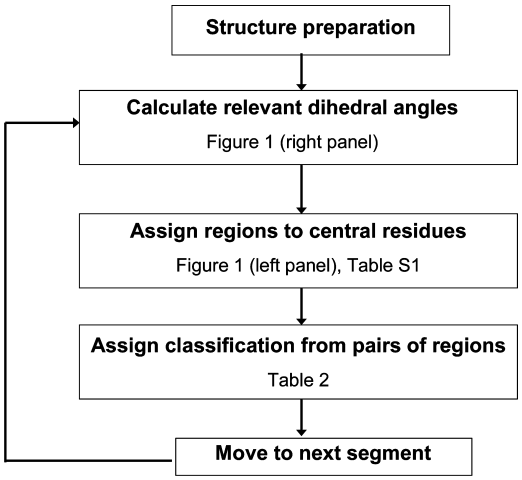


Figure 2. Flowchart of the DISICL algorithm to assign structural classes to biopolymer segments. For more details, see text.

regions and preliminary tests of the classification libraries, these were associated with the π-helix and inverse γ-turn, respectively. While the 18 defined classes provide a very good resolution on the change of the backbone shape, it may sometimes be preferable to summarize the structure with less detail, for example, to compare to DSSP. Hence, we grouped similar classes together to make a simplified classification library with only seven classes. The detailed and simplified protein classes of DISICL are shown (along with their abundance and average length) in Table 3. A more detailed description of the newly introduced, or less known DISICL classes, and the logic behind the grouping for the simplified library can be found in the Results and Discussion section.

Implementation. Currently, the DISICL algorithm exists as a number of independent python scripts, which can carry out the classification of individual structures or simulation trajectories in standard pdb 1.0 format. The standard output of these modules include the time series of residues for each class and a statistics file containing the residence time in all classes for each analyzed residue. This output information can be easily processed further by other programs. In addition, a script was written that allows direct visualization of the classification in Pymol[22] (images in Figures 3−5 were made using this script). These modules are combined into a package that can be automated or used independently as modules and which can be downloaded at http://disicl.boku.ac.at. Furthermore, DISICL will be integrated into the GROMOS++ analysis package[23] in the near future.

Comparison Studies. All structural models were analyzed separately by the applicable classification algorithms. As the different programs produced output in different formats, all results were ordered into identically formatted data series. The data series contained the name of the class along with all the residues in the model that belonged to that class, segment-based classifications were assigned to the first central residue. Second, the data series of all models were collected and combined into a single data set for each of the individual algorithms, containing elements a_{nj}, which was assigned the value 1 if residue n was member of the class j. For segment based assignments a value of 0.5 was assigned to both central residues, if it was compared with a residue-based method. Tables 3-4 show the abundance (occ_j), and average length (L_j) of each structural element, which were calculated based on the number of residues in the class (N_j), the number of interruptions (N_j^{int}), and the total number of residues (N_{sum}) according to eqs 1-4.

$$N_j = \sum_{n=1}^{N_{\mathrm{sum}}} a_{nj} \tag{1}$$

$$\mathrm{occ}_j = \frac{N_j}{N_{\mathrm{sum}}} \times 100 \tag{2}$$

$$L_j = N_j/N_j^{\mathrm{int}} \quad \text{for residue-based classification} \tag{3}$$

Table 2. Definitions for DISICL Protein Classification[a]

structure class	code	segment definitions
3/10-helix	3H	$\alpha1.\delta1$, $\alpha2.\alpha2$, $\alpha2.\delta$
turn type 1	T1	$\alpha1.\delta$, $\alpha2.\delta1$, $\delta.\delta$,
		$\delta1.\delta$, $\delta1.\delta1$, $\delta1.\alpha2$
turn-cap	TC	$\beta1.\alpha2$, $\delta.\alpha1$, $\delta.\delta1$, $\delta.\alpha2$,
		$\delta1.\alpha1$, $\delta2.\alpha2$, $\delta2.\delta1$, $\delta x.\alpha1$, $\delta\xi.\alpha2$, $\zeta.\alpha2$,
α-helix	αH	$\alpha1.\alpha1$, $\alpha1.\alpha2$, $\alpha2.\alpha1$
π-helix	πH	$\alpha1.\delta2$, $\delta2.\delta2$, $\delta2.\alpha1$, $\alpha2.\delta2$
helix-cap	HC	$\alpha2.\delta2$, $\delta.\delta2$, $\delta1.\delta2$, $\delta2.\delta$,
		$\beta1.\alpha1$, $\beta2.\alpha1$, $\beta2.\alpha2$, $\pi.\alpha1$, $\pi.\alpha2$,
ext. β-strand	EβS	$\beta2.\beta2$
normal β-strand	NβS	$\beta1.\beta1$, $\beta1.\beta2$, $\beta2.\beta1$
β-cap	BC	$\beta1.\pi$, $\beta2.\pi$, $\pi.\beta1$, $\pi.\beta2$
PP helical	PP	$\pi.\pi$
β bulge	BU	$\pi.\delta$, $\alpha1.\beta2$, $\delta.\beta2$, $\beta2.\delta x$
turn type 2	T2	$\pi.\delta x$
turn type 8	T8	$\delta.\zeta$, $\delta1.\zeta$, $\alpha2.\zeta$, $\alpha2.\beta1$, $\delta.\gamma x$
γ turns	GXT	$\pi.\gamma x$, $\pi x.\gamma$, $\gamma.\pi x$, $\gamma.\delta x$
Schellman turn	SCH	$\delta.\delta x$, $\delta1.\delta x$, $\delta x.\beta2$, $\delta x.\pi$
hairpin 2:2	HP	$\beta1.\delta x$, $\beta1.\pi x$, $\delta x.\beta1$
left turn 2	LT2	$\pi x.\alpha2$, $\pi x.\delta$, $\pi x.\delta1$
left-handed helix	LHH	$\delta x.\delta x$

[a]Segments are assigned to a class if their central residues fall into regions separated by a dot in the segment definitions (on the right).

dx.doi.org/10.1021/ci400541d | J. Chem. Inf. Model. 2014, 54, 266−277

Table 3. Detailed and Simplified DISICL Classes for Protein Classification and Their Abbreviations (code)[a]

DISICL detailed classes				DISICL simple classes			
structure class	code	occ. (%)	length	structure class	code	occ. (%)	length
3/10-helix	3H	3.8	2.2				
turn type 1	T1	2.8	2.2	3-helical turns	3HT	9.0	2.5
turn-cap	TC	2.3	2.0				
α-helix	αH	27.2	7.2				
π-helix	πH	0.4	2.1	α helical	HEL	32.2	5.4
helix-cap	HC	4.6	2.0				
ext. β-strand	EβS	2.7	2.4				
normal β-strand	NβS	10.2	3.3	β-strand	BS	21.3	3.9
β-cap	BC	8.4	2.4				
PP helical	PP	2.8	2.3	irreg. β struct.	IRB	4.7	2.2
β bulge	BU	1.9	2.1				
turn type 2	T2	0.8	2.0	β turns	BT	1.4	2.0
turn type 8	T8	0.6	2.0				
γ turns	GXT	1.0	2.1				
Schellman turn	SCH	2.6	2.2	other tight turns	OTT	4.5	2.2
hairpin 2:2	HP	1.0	2.0				
left turn 2	LT2	0.2	2.0	left-handed turns	LHT	0.6	2.0
left-handed helix	LHH	0.4	2.1				
unclassified	UC	26.3	3.3	unclassified	UC	26.3	3.2

[a]Occurrence (occ.) and average structure element length are calculated for combined X-ray and NMR data sets.

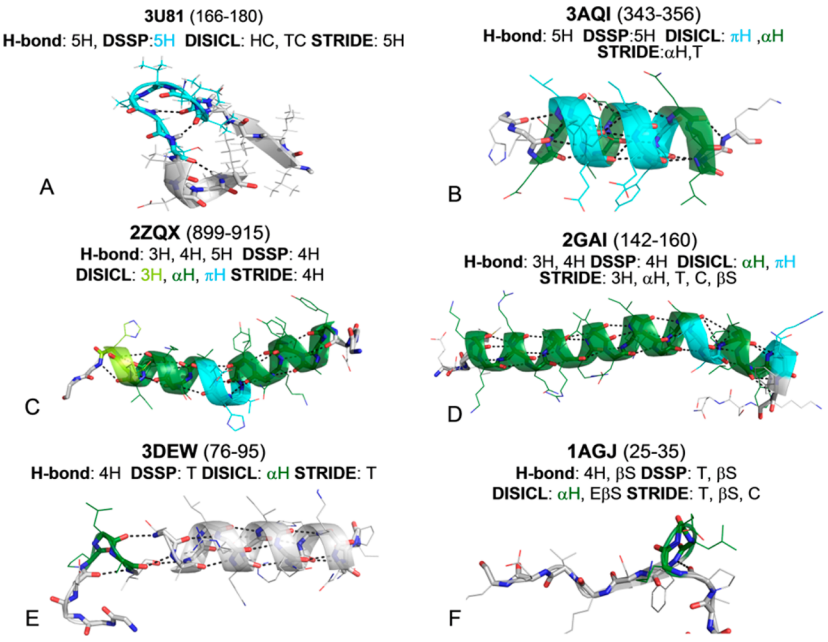


Figure 3. Six examples of classification of helical structures for which DSSP, DISICL, and STRIDE disagree. Titles of the panels show the PDB code and residue numbers of the displayed protein fragment. Also indicated are the observed hydrogen bonds and abbreviations of the classifications defined in Tables 3 and 4. Highlighted areas are colored to match their respective structures.

$$L_j = \left(\frac{N_j}{N_j^{\text{int}}}\right) + 1 \quad \text{for segment-based classification} \tag{4}$$

Note that for segment-based classifications, the average length was increased by 1, so that the first and last residues with a value of 0.5 are fully counted. To compare the various classification algorithms, the correlation matrices of algorithms were calculated, containing the correlation scores C_{ij} where i and j

dx.doi.org/10.1021/ci400541d | J. Chem. Inf. Model. 2014, 54, 266−277

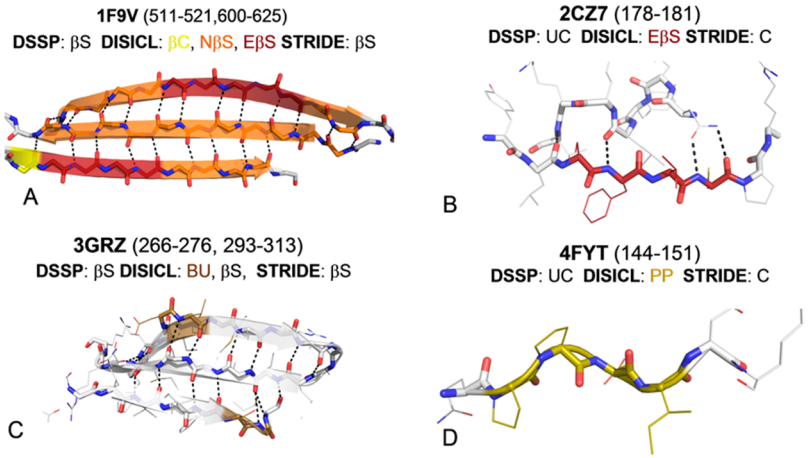


Figure 4. Four examples of β-structure classification by the programs DSSP, DISICL, and STRIDE. Titles of the panels show the PDB code and residue numbers of the displayed protein fragment. Also indicated are the observed hydrogen bonds and abbreviations of the classifications defined in Tables 3 and 4. Highlighted areas are colored to match their respective structures.

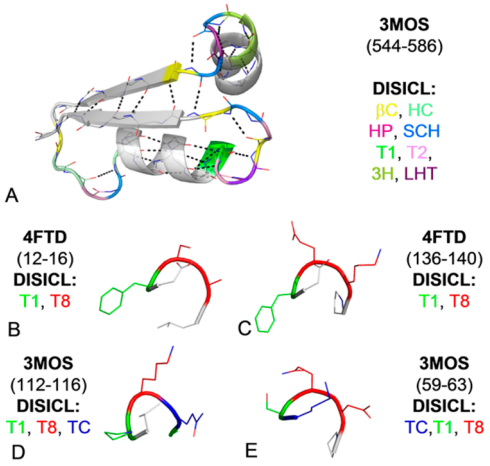


Figure 5. Five examples of turn classification by the program DISICL. Titles of the panels show the PDB code and residue numbers of the displayed protein fragment. Also indicated are the observed hydrogen bonds and abbreviations of the classifications defined in Table 3. Highlighted areas are colored to match their respective structures.

Table 4. Classes Defined for DSSP and STRIDE for Protein Classification and Their Abbreviations (code)[a]

method		DSSP		stride	
structure class	code	occ. (%)	length	occ. (%)	length
3-helix	3H	2.3	3.3	2.7	3.3
4-helix	4H	29.3	10.9	31.6	12.2
5-helix	5H	0.02	5.1	0.01	5.0
bend/coil	B/C	12.3	1.7	21.0	2.9
beta-bridge	BB	1.0	1.0	0.9	1.0
beta-strand	BS	18.1	5.1	20.0	5.3
turn	T	10.7	2.2	23.8	4.0
unclassified	UC	26.3	2.2	0.04	1.0

[a]Occurrence (occ.) and average structure element length are calculated for combined X-ray and NMR data sets.

marks the i^{th} class of the first algorithm and the jth class of the second algorithm, respectively. Three types of correlation scores were used: Pearson correlation (R_{ij}), match score (M_{ij}), and scaled match score (M^s_{ij}). The Pearson correlation (R_{ij}) is calculated from eq 5, where $\bar{a}_{ni}$ is the average occurrence of class i ($\bar{a}_{ni} = N_i/N_{sum}$).

$$R_{ij} = \frac{\sum_{n=1}^{N_{sum}} (a_{ni} - \bar{a}_{ni})(a_{nj} - \bar{a}_{nj})}{\sqrt{\sum_{n=1}^{N_{sum}} (a_{ni} - \bar{a}_{ni})^2 \sum_{n=1}^{N_{sum}} (a_{nj} - \bar{a}_{nj})^2}} \quad (5)$$

While the R-score drops quickly with the amount of mismatches (or different average occurrences of classes i and j), a large positive R-score is still a good measure to determine correspondence of different algorithm classes. The unscaled match score (M_{ij}) is calculated using eq 6 and represents the absolute number of residues assigned to class i in one algorithm, and to class j in the other algorithm.

$$M_{ij} = \sum_{n=1}^{N_{sum}} (a_{ni} \times a_{nj}) \quad (6)$$

The M-score is additive, which makes it possible to group classes or track distributions of correlations for one class. The scaled match scores (M^s_{ij}) provides a better comparison between algorithms and is calculated by eq 7.

$$M^s_{ij} = \frac{M_{ij}}{M_{max}} \times 100 \quad (7)$$

In words, the scaled match score is obtained by dividing the observed match (M_{ij}) between two classes with the maximal theoretical match (M_{max}). For comparison of two residue-based methods or two segment-based methods, M_{max} is equal to the size of the smaller data set.

$$M_{max} = \min\{N_i, N_j\} \quad (8)$$

However, when comparing a segment-based method with a residue-based one (mixed comparison), the maximal match is defined as

$$M_{max} = \min\left\{N_i, N_j, \frac{1}{2}\left(\frac{N_i(L_i - 2)}{L_i - 1} + N_j\right)\right\} \quad (9)$$

The mixed comparison analyses include (1) DISICL classes vs DSSP or simplified STRIDE classes and (2) detailed STRIDE classes vs simplified STRIDE or DSSP classes.

dx.doi.org/10.1021/ci400541d | J. Chem. Inf. Model. 2014, 54, 266−277

To summarize comparisons, the weighted average of the scaled match scores were calculated for helical classes, β-strand classes, and turn classes (see Table 1, methods agreement). Additionally, the weighted average of all these superclasses and the scaled match score for unclassified residues was calculated to obtain an overall match between methods. The grouping for superclasses is provided in Table S2 of the Supporting Information.

RESULTS AND DISCUSSION

DISICL Protein Classes. Most of the newly introduced structural classes are connected to transitory areas apart from more conventional structural elements (α-helix, β-strands, and β-turns). The helix-cap class (HC), for instance, is based on a collection of clusters in the study of Hollingsworth et al. that are specifically found prior and sometimes after helical structures—most often α-helices—and possibly play a role in the formation of such structures. The clustering study also revealed typical backbone elements next to β-strands and β-turns (grouped together as β-caps, BC) and the β-turns type I and III (turn-caps, TC). While most of the cap structures typically appear when less ordered protein sections turn into more ordered structural elements, the bulge class (BU) marks a residue with α-helical dihedral angles, which is inserted into an ordered β-strand or reversed.

There are eight types of β-turns (I, II, III, IV, VI, VIII, I′, II′) having similar i − (i + 3) hydrogen bonding, differentiated mainly by their backbone dihedrals.[24] Six of the β-turns are covered by the detailed DISICL classes. β-turn type III is identical with the 3_{10}-helix (3H), while type I′ and II′ are left handed structures, shown as left-handed helix (LH) and left turn II (LHT). The missing β-turns are type VI, which is not considered because it requires a cis-proline, and the turn type IV, for which no typical dihedrals could be defined as it represents every turn that does not fit into any of the other turn classes. It remains difficult to differentiate between β-turn type I and III (or 3_{10} helix), as their clusters are close and not distinctly differentiated. In the simplified classification, these turn types are grouped together.

Additionally, a number of other tight turns were found to have selective clusters in the (φ,ψ) dihedral space. These are the normal and inverse γ turns (GT), which have typically i − (i + 2) hydrogen bonding, the Schellman turn (SCH), which is known to terminate α-helixes, and a 2:2 hairpin (HP), which is a tight turn usually connecting β-strands.

The PP helical class represents the π-region of the dihedral angle space, which borders the area associated with the normal beta strand (NBS). This class was defined to be representative for the polyproline helix (PP), although it is not highly selective, and typically also appears in areas where a regular β-strand is broken. It was grouped together with the bulge to form the irregular β-structures.

Classification Comparisons. The classification study of proteins was carried out on the Prot_Xr and Prot_NMR data sets by three different algorithms DSSP, STRIDE, and DISICL. Table 1 shows the summary of results for this study. Despite the fact that X-ray structures were about four times longer on average, two-thirds of the analyzed residues originated from the NMR data set because of the high number of models per database entry (multiplicity). The multiplicity of the Prot_Xr data set is slightly below 1.0 (0.94) because most X-ray PDB entries contained only one model and some had unrecognized formatting, so no models were found in them. In terms of structural elements, the two data sets were not highly different, although the slightly longer average length of the secondary structure elements and the lower percentage of unclassified or coil structures (17% vs 26%) show that X-ray models on average are more ordered. Because this difference can be easily explained by the differences in the experimental methods, the data sets were combined for the further analysis of the classifications.

The "methods performance" section of Table 1 shows the number of residues that could be handled by each algorithm (data set size) and its percentage with respect to the overall data set (completeness). Only those models were considered for which all three programs can classify at least one residue. The algorithms assigned a meaningful structural classification to 73–80% of the handled residues (classification ratio), and the rest of residues were marked as unclassified (DISICL), coil (STRIDE), or were left out from the results (DSSP). Considering every factor, STRIDE was the most effective in assigning classifications with 79% of the total data set, while DISICL and DSSP classified 71% and 60%, respectively. The lower percentage for DSSP is largely explained by the fact that it cannot always handle multiple chain models. A brief overview of the agreement between algorithms for the indicated superclasses is also provided in Table 1 under the section "methods agreement". The table contains the

Table 5. DISICL Classification Results for Trajectories of Six Proteins[a]

class/protein[b]	CM	Colds	Fox	HEWL	ProtG	SAC
3H %	3.8	1.7	2.1	4.0	2.1	2.0
T1 %	2.0	0.6	0.9	2.3	0.7	0.5
TC %	2.1	0.5	1.0	2.9	0.5	0.4
αH %	68.9	5.2	16.9	31.9	27.9	17.0
πH %	0.2	0.1	0.2	1.4	0.4	0.4
HC %	3.6	6.5	5.9	7.3	5.1	6.0
NβS %	0.2	13.2	7.5	2.1	20.0	11.8
EβS %	0.1	3.8	2.1	0.2	3.9	2.5
BC %	2.8	19.9	11.9	4.9	9.6	15.1
PP %	4.6	6.7	7.3	3.0	1.5	6.5
BU %	0.1	3.3	2.1	0.8	1.1	1.3
T2 %	0.3	1.3	1.8	2.2	0.0	0.5
T8 %	0.4	0.4	0.5	0.3	0.1	0.5
GXT %	0.1	2.0	3.7	0.7	1.4	4.3
SCH %	2.4	4.2	2.4	4.8	2.6	0.3
HP %	0.5	0.4	0.6	2.1	2.3	0.9
LT2 %	0.0	0.4	0.0	0.5	0.0	0.7
LHH %	0.0	0.0	0.2	1.4	0.0	0.4
UC %	7.9	29.7	32.9	27.3	20.7	29.0
class/protein[b]	CM	Colds	Fox	HEWL	ProtG	SAC
3HT %	7.8	2.7	4.0	9.2	3.4	2.9
HEL %	72.7	11.8	23.0	40.6	33.4	23.3
BS %	3.2	36.8	21.5	7.1	33.5	29.5
IRB %	4.7	10.0	9.3	3.8	2.6	7.8
BT %	0.7	1.8	2.2	2.5	0.1	1.0
OTT %	3.0	6.7	6.8	7.6	6.3	5.5
LHT %	0.0	0.4	0.2	1.9	0.0	1.1
UC %	7.9	29.7	32.9	27.3	20.7	29.0

[a]Upper part of table shows the result for detailed analysis; bottom parts shows the simplified analysis. Abbreviations for the classes are displayed in Table 3. [b]CM: chorismate mutase from *Mycobacterium tuberculosis*. Colds: major cold shock protein CspA from *Escherichia coli*. Fox: RNA binding domain of the Fox-1 protein. HEWL: hen egg white lysozyme. ProtG: B1 immunoglobulin-binding domain of streptococcal protein G. SAC: hyperthermophilic protein Sac7d from *Sulfolobus acidocaldarius*.

dx.doi.org/10.1021/ci400541d | J. Chem. Inf. Model. 2014, 54, 266−277

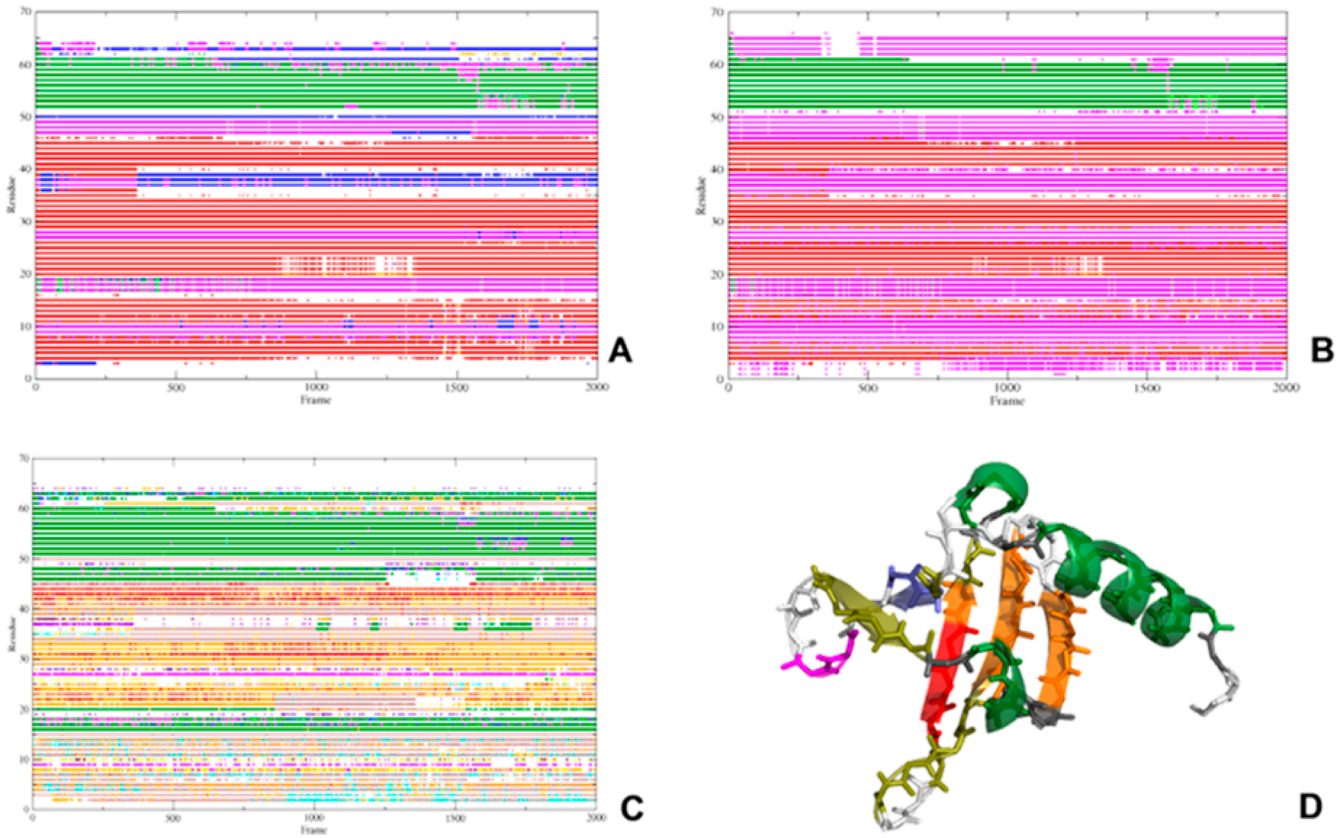


Figure 6. Occurrence of secondary structure elements during a 20 ns simulation of the SAC protein. (A) DSSP, (B) STRIDE, and (C) DISICL detailed. The green color represents helical structures. Red, orange, and brown colors represent β structures. Purple and blue colors represent turn and bend structures, respectively. (D) Structure of the protein, colored according to the DISICL classification. The detailed color scheme for structural classes is provided in Figure 7.

weighted averages of the scaled match scores between corresponding classes. The precise grouping of superclasses is provided in Table S2 of the Supporting Information. The helical match was calculated based on the match between 3-helical turns/3-helix classes, α-helical/4-helix classes, and α-helical/5-helix classes. Helical structures are most well ordered and well described both in terms of hydrogen bonding and of backbone dihedrals, which is reflected in the remarkable agreement between the classification algorithms, where even the worst average agreement amounts to more than 80% of the theoretical maximal agreement. The beta-strand match was calculated based on the agreement of β-strand classes and agreement between β-bridge classes for STRIDE and DSSP. Agreement between the algorithms ranges from 74% to 97% with slightly less agreement between DISICL and the other two programs as for helical classes. The agreement between Turn structures ranges from 33% to 69%, displaying that turn classification is the most challenging task for proteins and also represents the greatest difference between the DSSP algorithm and STRIDE. The overall match summarizes agreement between algorithms, calculated from the weighted average of helical match, beta-strand match, and turn match, plus the scaled match score between unclassified/coil classes. STRIDE and DSSP show the strongest agreement (82%), which is not surprising considering the shared principle work mechanism. Also unsurprisingly, DISICL agrees more with STRIDE (78%), which also takes backbone dihedrals into account, than with DSSP (69%), which is mostly based on hydrogen bonds. It is important to mention that some classes in the algorithms had no analogs and were not explicitly represented in this summary. For instance, the irregular β-structures of DISICL, consisting of the bulge and the polyproline region classes, do not show distinctive hydrogen bond patterns and agree best with unclassified/coil structures of DSSP and STRIDE. Similarly the β-bridge class agreed reasonably well between STRIDE and DSSP (65%) but had no real correlation in terms of dihedrals. The bend structure of DSSP shows significant correlations with the various turns of DISICL and STRIDE but also strong correlations with unclassified structures. On the basis of the R-score correlations, we assigned the DSSP bend structures to the turns in the DSSP–STRIDE comparison but did not include them in the final comparison between DSSP and DISICL. In the following, we describe the detailed correlations between helical, strand, and turn structures, as well as the differences between occurrences and average structure lengths.

Helical Structures. Here, we compare the abundance, average length, and agreement scores of helical structural elements of the different algorithms (3-helix, 4-helix, 5-helix of DSSP and STRIDE; 3_{10}-helix, α-helix, π-helix for DISICL; or 3-helical turns and α-helical class in the simplified library of DISICL). Table 3 displays the classification results for DISICL (both with the detailed and simplified library), and Table 4 shows the same results for DSSP and the simplified STRIDE algorithm. Interestingly, the amount of α-helix residues are very similar in the different algorithms, but the average helix lengths differ significantly (most obvious for the α-helix, where average class lengths are 7.1 (detailed) and 5.4 (simple) in DISICL vs 10.9 and 12.2 in DSSP and STRIDE, respectively). For the detailed

dx.doi.org/10.1021/ci400541d | J. Chem. Inf. Model. 2014, 54, 266–277

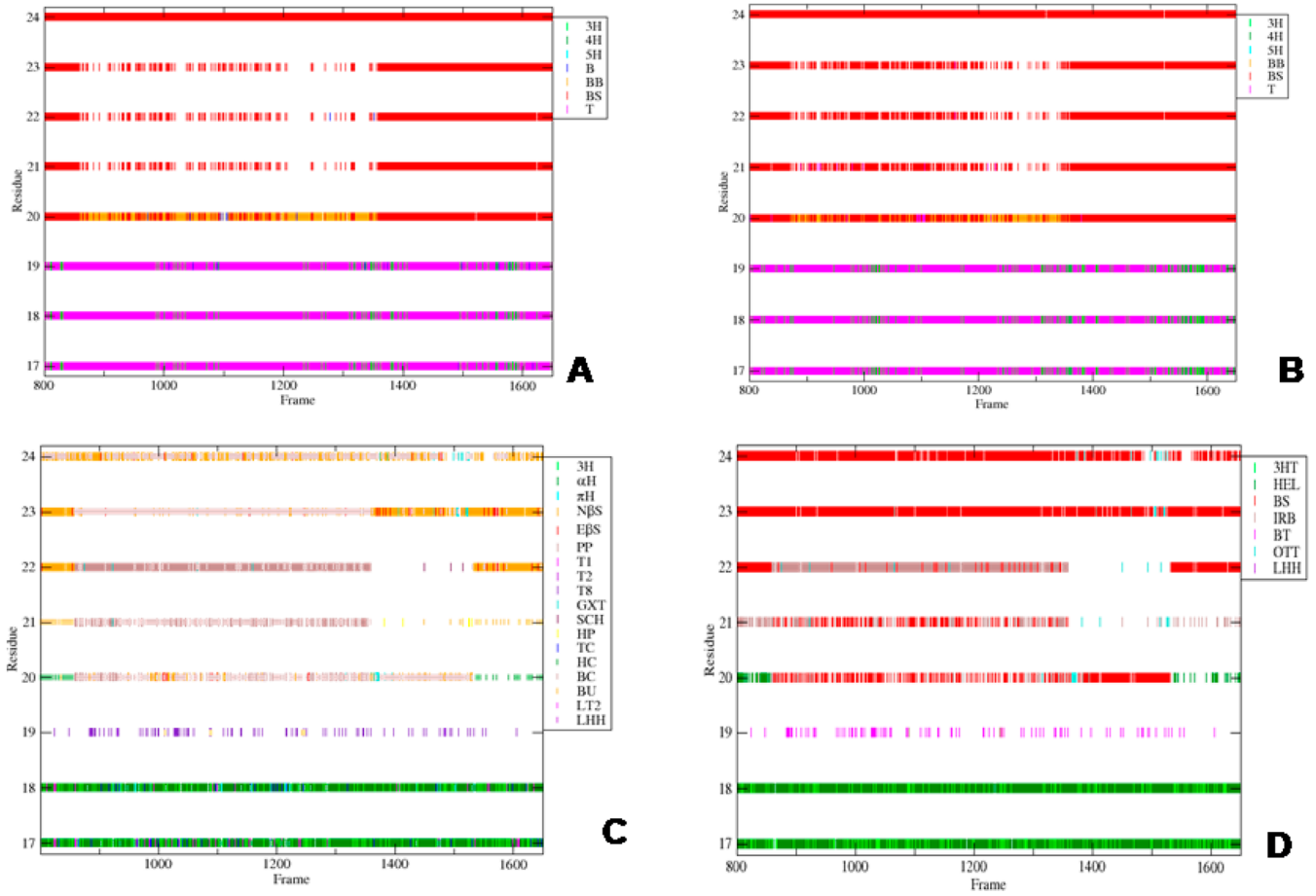


Figure 7. Close-up of the occurrence of secondary structure classification of a reversibly unfolding β-strand, observed with DSSP (A), STRIDE (B), DISICL [detailed and simplified libraries in (C and D), respectively]. Abbreviations for the classes can be found in Tables 3 and 4.

DISICL library, this is observed because many regular α-helices contain kinks that are usually classified as other helix types but are often not detected by the other two algorithms (examples for this are shown in panels C and D of Figure 3). For the simplified DISICL classes, the reason of the shorter average length comes from the very short but often occurring cap structures, which are grouped together with the more regular and longer structural elements.

There is also a significant difference in the occurrence and average class length of the other two helix types. The reason for this is again the hierarchical nature of hydrogen bond-based algorithms, as they slightly prefer the α-helix at the expense of other helix types (3_{10}-helix and π-helix). The abundance of the 3_{10}-helix is about 1.5 times higher for DISICL (3.8% vs 2.5%), as kinks and deformations in the middle of standard α-helices as well as at the ends are often classified as a 3_{10}-helix. A similar but even more striking difference is observed for the π-helix—which is still the rarest type of secondary structure element for all algorithms—where the difference in abundance is at least one order of magnitude (0.4% vs 0.02% and 0.01% for DISICL, DSSP, and STRIDE, respectively). Additionally, the average length of the π-helix is around five residues (not counting the two flanking amino acids) in STRIDE and DSSP, while it is 2.3 residues in DISICL.

In terms of agreement scores (Tables S3−S8, Supporting Information), helical classes are most robust over all algorithms, especially the α-helix. The scaled match scores between DSSP and DISICL amount to 85% for α-helices, and STRIDE agrees with both algorithms in more than 90% of the cases (as it is more abundant than both). The R-scores range from 0.6 to 0.9 (smaller data set on X-ray models for DSSP results in lower Pearson correlation scores for both comparisons). The 3-helix and 3_{10}-helix classes share a moderate correlation with M^s scores ranging from 40% to 75% percent and R-scores ranging from 0.25 to 0.65. The 3-helix also shares significant correlations with the α-helix and certain turns (especially the turn type I) and to a lesser extent the π-helix. The 3_{10}-helix of DISICL is shared almost evenly between 3-helix and 4-helix structures of the other two algorithms, while also showing some correlation with turn classes. More surprisingly, the correlation between the π-helix in DISICL and 5-helix structures in DSSP and STRIDE is very low, amounting to only 10% of 5-helix. π-helix residues in DISICL are usually interpreted as 4-helix, bend, or turn structures in the other algorithms, while 5-helix residues of STRIDE and DSSP were mainly unclassified, α-helix, or caps according to DISICL. It is possible that the DISICL definitions for the 3_{10}-helix and π-helix are ill placed; however, visual checks on the differently classified protein fragments often confirmed the existence of the $i-(i+3)$ or $i-(i+5)$ hydrogen bonds (many times in coexistence with the $i+4$ hydrogen bonding) along with deformed helical structures. Recent literature[13,25] also suggests a connection between helix deformations and the π-helix (and to some extent the 3_{10}-helix), as well as their evolutionary and functional importance, and also presents the dihedral angle distribution for 2-, 3-, 4-, and 5-hydrogen-bonded distortions in α-helices. These

dx.doi.org/10.1021/ci400541d | J. Chem. Inf. Model. 2014, 54, 266−277

distributions indeed agree well with the region definitions of 3_{10}-helix and π-helix in DISICL.

The correlation scores show to what extent classifications agree, but as any classification depends on definitions, its correctness can only be interpreted through examples and how well they meet those definitions. Figure 3 shows six example structures where classification algorithms gave different answers. Panel A of Figure 3 shows a structure that was assigned as a 5-helix by both DSSP and STRIDE. While this protein fragment clearly shows a nonhelical backbone, visual checks indeed reveal two consecutive hydrogen bonds between flanking residues normally found in 5-helices. On the other hand, a true 5-helix structure is shown in panel B, which shows a completely uniform $(i + 5)$ hydrogen bonding. It is known that 5-helices have a nonuniform backbone dihedral distribution, which is reflected in DISICL by an alternating pattern of α- and π-helix segments. The examples in panel C and D show distortions in regular α-helices, which DISICL defines as a π-helix; both of these structures feature a complex hydrogen bonding pattern. In light of the examples shown in panels B, C, and D of Figure 3, the present DISICL definitions of different helical classes are successfully identifying changes and distortions in the 3D structure of helical protein elements, but further optimization and/or visual checks might be required as the definition of different helix types may overlap. Panels E and F of Figure 3 show the minimal structure of DISICL for an α-helix. Panel E shows a complete α-helix, where a single irregular residue was deleted during the structure preparation step. The tetrapeptide segment marked on the left retained its α-helical dihedrals, but was only left with one $(i + 4)$ hydrogen bond flanking it and as such was classified as turn by both STRIDE and DSSP. While this example might be called artificial, a very similar segment is shown in panel F, which also shows a single flanking hydrogen bond and the preferred backbone dihedral angles of an α-helix. Despite the discrepancies that were mentioned above, which are usually due to the different priorities in the classification algorithms, the major proportion of helical protein elements is identified correctly by all three algorithms.

Beta Structures. The second major type of secondary structural elements is formed by the β-structures, mostly consisting of β-strands and β-sheets. β-strands are well known for their distinct distribution in the (φ,ψ) dihedral space, as well as their regular backbone hydrogen bonding connecting the individual strands into β-sheets, frequently playing critical functional roles in proteins. As shown in Tables 3 and 4, β-strands take up roughly 20% of the residues in our protein data set and have an average length of 4−5 amino acids according to all classification algorithms. The area for β-structures in the (φ,ψ) distributions could be differentiated further to separate the normal β-strands from distorted structures and turns. On the basis of the cluster centers reported by Hollingsworth et al., DISICL divides the classical β-strand definition into a normal β-strand and the extended β-strand classes related to the $\beta1$ and $\beta2$ regions. Besides the β-strand classes, there were certain turn definitions (γ-turns, β-turns, tight hairpin, etc.), which are connected to this area of the dihedral space, as well as the bulge and polyproline-like classes. Dividing the β-strand into two separate classes also decreased the average class length in the detailed DISICL algorithm, while for the simplified library the presence of individually occurring β-caps decreased the class length to some extent compared to the DSSP and STRIDE β-strand structures. The irregular β-structures took up about 5% of the residues, with a short average length of 2.2 residues. In terms of correlations, the β-strand classes show a good correlation with R-scores ranging from 0.55 to 0.9 and M^s-scores of 75−98% (Tables S3−S8, Supporting Information). The β-bridge classes in DSSP and STRIDE are moderately correlated with the DISICL β-strand in terms of the M^s-scores (∼35% of β-bridges), and about 28% of the DSSP β-bridges were recognized as strand by STRIDE. The DISICL irregular β structures show moderate correlations with the unclassified/coil classes in the other two algorithms and also with turn classes to a lesser extent. The polyproline-helical class shows weaker correlations with bend structures in DSSP, while the bulge class showed a similar correlation with β-strand structures of both STRIDE and DSSP.

Examples of β-structures are shown in panels A−D of Figure 4. Panel A shows a β-sheet structure, where colorings mark the detailed DISICL classification. While all three β-strands have a certain twist in the backbone, extended β-strand segments give the two strands on the edges an extra curvature not observed in the middle strand. Panel B shows a protein fragment that was unclassified by DSSP or STRIDE but was considered as an extended β-strand by DISICL. While this segment lacks the proper backbone hydrogen bonds with another β-strand, its linear structure is partially stabilized by side-chain interactions. Bulge class elements are usually deformations in β-strands; panel C shows two examples (at the end of the strand and in the middle). The bulge segment within the β-strand did indeed protrude from the regular plane of the strand and changed the direction of it without breaking the hydrogen bond pattern. The protein fragment shown in panel D was found by searching the polyproline-helical class in DISICL. While this stretch indeed featured two prolines and a helical structure, the class definition is not highly selective and contains a large set of different conformations often flanking β-strands. However, this class also often showed helical characteristics and was indeed highly enriched in proline residues.

Summarizing the observations on β-structures described above, classification of DISICL differs slightly from the results of DSSP and STRIDE. While correlation of β-strand elements is still very high, DISICL effectively identifies distortions in β-strand structures, while also pointing at several special structural elements in the β dihedral region.

Turn Structures. The third type of secondary structure elements shows the widest variety of hydrogen bonding and dihedral angle patterns, building loop structures that connect the linear structural elements, ultimately playing a very important role in the fold and functionality of enzymes. While loops are deemed generally flexible, less ordered, and structurally less important than α-helices and β-sheets, there are many examples in which a small modification on the loop structure can compromise the fold of the full protein or when loops have functionally important roles (such as kinase loops, antibody variable regions, HNH activation loop[26−28]). To fulfill their roles in the protein, loops can have their own shape-stabilizing backbone and side-chain interactions including hydrogen bonds. While these interactions are usually more complex than those of the more linear structure elements, loops may be broken down into smaller structural segments (such as turns, caps, etc.). Turn structures were originally defined by the hydrogen bonding patterns as well [β-turns typically have $i - (i + 3)$ hydrogen bonding, for instance], but the importance of the backbone shape was also realized and described by the dihedral angles of the turn structures. Six β-turn definitions are described by DISICL (see above), which correspond to broadened turn definitions of Wilmot and Thornton.[8] Approximately 5% of the residues were

classified as β-turns by the DISICL algorithm (not including the 3_{10} helix), with average class lengths usually only slightly above two residues (or one segment). Additionally, the detailed DISICL library contains definitions for the sharper γ-turn and inverse γ-turn (grouped together in γ-turns class), the sharp 2:2 hairpin structure, and the Schellman turn, which were grouped together in the "other tight-turns" class in the simplified library (consisting of another 4.5%). The Schellman motif often appears as terminator of α-helices and contains very characteristic segments that were grouped together to form the Schellman turn class. The full motif starts from an α-helix with a turn type I or 3_{10} helix segment, followed by two of the four Schellman turn segments, also represented in the average 2.6 residue length of the Schellman turn class.

Turn classes of DISICL have a relatively low level of correlation with the DSSP turn class, usually with an R-score of 0.13−0.3 and M^s score of 20−60% (Tables S3−S8, Supporting Information). Turn type I shows a smaller correlation (20%) with the 4-helix, while the rest was distributed evenly between the 3-helix and turn classes. Some turns also show some correlation with the DSSP bend (generally M^s score around 15%), while the γ-turn was mostly considered as unclassified [as it should have an $i - (i + 2)$ hydrogen bond, which is not considered by DSSP]. The definition of the turn class is significantly different between the DSSP and the STRIDE algorithm, which is reflected in the abundance of the class (10% vs 24%, respectively). Match scores reveal a significantly higher agreement between the DISICL and STRIDE turn classes. For certain classes (like β-turn type II), the agreement can be as high as 90% of DISICL residues, but for the two most abundant turn types (Schellman turn and β-turn type I), M^s scores remain around 45%, resulting in lower overall agreement. The STRIDE turn class shows the highest correlation with the DISICL turn classes, but 66% was still unclassified in the DISICL algorithm. Additionally, the STRIDE turn shows significant correlation with helix-cap (50%) and turn-cap (40%) classes (while the β-caps were mostly considered as part of β-strands), π-helix (40%), bulge (35%), and polyproline-like and 3_{10}-helix (both 25%) classes. The correlation between the STRIDE turn class and the DISICL caps can be explained easily from the fact that caps are special turn structures found next to more common structural elements. When compared to DSSP classes, 70% of DSSP turn residues were also considered turns in the STRIDE algorithm, along with most of the bend (66%) and 5-helix (52%) residues, as well as a significant proportion of the unclassified (25%) and 3-helix (20%) classes.

Similar to the Schellman motif, it is often observed that loops consist of consecutive segments of turns and caps, such as shown in panel A of Figure 5, depicting two α-helices and two β-strands connected by three loops. The figure also shows β-cap segments that are indeed introducing the β-strands of this structure. Panel B−E of Figure 5 shows four different loop segments from two (nonmultimeric) proteins, featuring a turn I−turn VIII motif, which nicely demonstrates how backbone dihedrals can describe the shape of loops. While the structure of the loop can change significantly if we add further turn segments, all four loops are very similar in the part where the turn I−turn VIII motif occurs, despite the fact that they do not share an identical amino acid sequence. On the basis of the examples shown above, the detailed turn class definitions of DISICL are useful in monitoring loop structures of proteins, as well as to compare similar loop elements in different proteins.

Analysis of Simulation Trajectories. To validate the performance of DISICL when following structural changes during molecular simulations, we reanalyzed trajectories of MD simulations for six distinct proteins, formerly used to validate the 54A8 GROMOS parameter set.[19] The analysis was performed with both libraries of DISICL (summarized in Table 5), as well as DSSP and STRIDE (Tables S9 and S10, Supporting Information). While differing greatly for individual proteins, the overall content of structural elements was similar to the analysis of the PDB data sets for all three algorithms and the same holds for their correlations.

As a case study, we chose the analysis of the hyperthermophilic protein Sac7d of *Sulfolobus acidocaldarius* (SAC), as it contained both α- and β-structural elements, and it showed significant structural change during the simulations. Figure 6 shows the occurrence of the secondary structure elements as defined by all three algorithms based on 2000 snapshots, sampled at 1 ps intervals from a 20 ns trajectory. The general features of the SAC protein appear in all three structure classification plots, namely, the three-stranded β-sheet in the middle of the sequence followed by an α-helix, which partially unfolds at the C-terminus. All three algorithms show instability in the first of the three β-strands during the third quarter of the simulation trajectory, as detailed in Figure 7. While STRIDE shows a smaller change in stability of the strand, the β-strand of DSSP disappears over 50% of this time period. The DISICL algorithm shows a change in the backbone conformation, as the residues are classified mainly as β-cap or polyproline-like structures (both are associated with β-structures) before the structure refolds into a regular β-strand.

Considering the details of the N-terminal part of the protein, DSSP classifies a short but stable β-sheet based on the hydrogen bonding. In this region, STRIDE shows a similar but less stable sheet with an increased proportion of turns, while DISICL classifies the structure predominantly as mixture of β-caps, polyproline-like structures, and γ-turns. Visual checks confirm the hydrogen bonds as well as the extremely distorted nature of these β-strands (Figure 6D).

CONCLUSIONS AND OUTLOOK

We have introduced a new structure classification algorithm, DISICL, which performs the classification of short biopolymer segments based on dihedral angles within the segment. We demonstrated the potential of the algorithm by performing a large-scale classification of protein models found in the Brookhaven Protein Databank and a comparative analysis for a set of six simulation trajectories using DISICL and two already established algorithms (DSSP and STRIDE). The comparison included the amount of handled and classified residues and average occurrence and length of structural elements represented by the classes, as well as pairwise matches of the classes between algorithms. The analysis provided useful information and general visualization of the DISICL classes and showed that the algorithm stood its ground against similar methods in terms of classification efficiency, while providing a higher level of structural detail. We propose the DISICL algorithm as a useful tool for molecular simulations, where this higher level of detail might provide a better insight on the dynamics and interactions of biopolymers and a better comparison to structural information obtained from advanced spectroscopic methods.

ASSOCIATED CONTENT

Supporting Information

Detailed information on the definitions of DISICL regions, super classes, and correlation analysis results. This material is available free of charge via the Internet at http://pubs.acs.org. The

dx.doi.org/10.1021/ci400541d | J. Chem. Inf. Model. 2014, 54, 266−277

DISICL libraries and programs are available via the Internet at http://disicl.boku.ac.at.

AUTHOR INFORMATION

Corresponding Author

*E-mail: chris.oostenbrink@boku.ac.at.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors thank Maria Reif for making the simulation trajectories available. Financial support from Grant LS08-QM03 of the Vienna Science and Technology Fund (WWTF), Grant 260408 of the European Research Council (ERC), and the PhD. Programme "BioTop—biomolecular technology of proteins" (Austrian Science Fund, FWF Project W1224) is gratefully acknowledged.

REFERENCES

(1) Liberles, D. A.; Teichmann, S. A.; Bahar, I.; Bastolla, U.; Bloom, J.; Bornberg-Bauer, E.; Colwell, L. J.; De Koning, A. P.; Dokholyan, N. V.; Echave, J. The interface of protein structure, protein biophysics, and molecular evolution. *Protein Sci.* **2012**, *21*, 769−785.

(2) Uversky, V. N.; Dunker, A. K. Understanding protein non-folding. *Biochim. Biophys. Acta* **2010**, *1804*, 1231−1264.

(3) Redfern, O. C.; Dessailly, B.; Orengo, C. A. Exploring the structure and function paradigm. *Curr. Opin. Struct. Biol.* **2008**, *18*, 394−402.

(4) Pauling, L.; Corey, R. The structure of hair, muscle, and related proteins. *Proc. Natl. Acad. Sci. U.S.A.* **1951**, *37*, 261−271.

(5) Pauling, L.; Corey, R.; Branson, H. The structure of proteins: 2 Hydrogen-bonded helical configurations of the polypeptide chain. *Proc. Natl. Acad. Sci. U.S.A.* **1951**, *37*, 205−211.

(6) Richardson, J. Beta-sheet topology and relatedness of proteins. *Nature* **1977**, *268*, 495−500.

(7) Venkatachalam, C. M. Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units. *Biopolymers* **1968**, *6*, 1425−1436.

(8) Wilmot, C.; Thornton, J. Beta-turns and their distortions: A proposed new nomenclature. *Protein Eng.* **1990**, *3*, 479−493.

(9) Lewis, P.; Momany, F.; Scheraga, H. Chain reversals in proteins. *Biochim. Biophys. Acta* **1973**, *303*, 211−229.

(10) Nemethy, G.; Printz, M. Gamma-turn, a possible folded conformation of polypeptide chain: Comparison with beta-turn. *Macromolecules* **1972**, *5*, 755−&.

(11) Schellman, C. Alpha-L conformation at the ends of helices. *Hoppe-Seyler's Z. Physiol. Chem.* **1979**, *360*, 1014−1015.

(12) Nemethy, G.; Phillips, D.; Leach, S.; Scheraga, H. A second right-handed helical structure with parameters of pauling-corey alpha-helix. *Nature* **1967**, *214*, 363−&.

(13) Weaver, T. M. The π-helix translates structure into function. *Protein Sci.* **2000**, *9*, 201−206.

(14) Richardson, J.; Getzoff, E.; Richardson, D. Beta-bulge: Small unit of non-repetitive protein-structure. *Biophys. J.* **1978**, *21*, A144−A144.

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

(16) Frishman, D.; Argos, P. Knowledge based protein secondary structure assignment. *Proteins* **1995**, *23*, 566−579.

(17) Hollingsworth, S. A.; Lewis, M. C.; Berkholz, D. S.; Wong, W.-K.; Karplus, P. A. $(\varphi,\psi)_2$ motifs: a purely conformation-based fine-grained enumeration of protein parts at the two-residue level. *J. Mol. Biol.* **2012**, *416*, 78−93.

(18) 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.

(19) Reif, M. M.; Winger, M.; Oostenbrink, C. Testing of the GROMOS force-field parameter set 54A8: Structural properties of electrolyte solutions, lipid bilayers, and proteins. *J. Chem. Theory Comput.* **2013**, *9*, 1247−1264.

(20) Nagy, G.; Oostenbrink, C. Dihedral-based segment identification and classification of biolpolymers II: Nucleotides. *J. Chem. Inf. Model.* **2013**, DOI: 10.1021/ci400542n.

(21) Zou, C.; Larisika, M.; Nagy, G.; Srajer, J.; Oostenbrink, C.; Chen, X.; Knoll, W.; Liedberg, B.; Nowak, C. Two-dimensional heterospectral correlation analysis of the redox-induced conformational transition in cytochrome C using surface-enhanced raman and infrared absorption spectroscopies on a two-layer gold surface. *J. Phys. Chem. B* **2013**, *117*, 9606−9614.

(22) Schrödinger, L. L. C. *The PyMOL Molecular Graphics System*.

(23) Eichenberger, A. P.; Allison, J. R.; Dolenc, J.; Geerke, D. P.; Horta, B. A. C.; Meier, K.; Oostenbrink, C.; Schmid, N.; Steiner, D.; Wang, D.; Van Gunsteren, W. F. GROMOS++ software for the analysis of biomolecular simulation trajectories. *J. Chem. Theory Comput.* **2011**, *7*, 3379−3390.

(24) Richardson, J. S. The anatomy and taxonomy of protein structure. *Adv. Prot. Chem.* **1981**, *34*, 167−340.

(25) Cooley, R. B.; Arp, D. J.; Karplus, P. A. Evolutionary origin of a secondary structure: π-Helices as cryptic but widespread insertional variations of α-helices that enhance protein functionality. *J. Mol. Biol.* **2010**, *404*, 232−246.

(26) Narciso, J. E. T.; Uy, I. D. C.; Cabang, A. B.; Chavez, J. F. C.; Pablo, J. L. B.; Padilla-Concepcion, G. P.; Padlan, E. A. Analysis of the antibody structure based on high-resolution crystallographic studies. *New Biotechnol.* **2011**, *28*, 435−447.

(27) Arencibia, J. M.; Pastor-Flores, D.; Bauer, A. F.; Schulze, J. O.; Biondi, R. M. AGC protein kinases: From structural mechanism of regulation to allosteric drug development for the treatment of human diseases. *Biochim. Biophys. Acta* **2013**, *1834*, 1302−1321.

(28) Czene, A.; Németh, E.; Zóka, I. G.; Jakab-Simon, N. I.; Körtvélyesi, T.; Nagata, K.; Christensen, H. E.; Gyurcsik, B. The role of the N-terminal loop in the function of the colicin E7 nuclease domain. *J. Biol. Inorg. Chem.* **2013**, 1−13.