JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

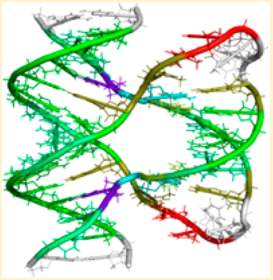
Dihedral-Based Segment Identification and Classification of Biopolymers II: Polynucleotides

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: In an accompanying paper (Nagy, G.; Oostenbrink, C. Dihedral-based segment identification and classification of biopolymers I: Proteins. *J. Chem. Inf. Model.* 2013, DOI: 10.1021/ci400541d), we introduce a new algorithm for structure classification of biopolymeric structures based on main-chain dihedral angles. The DISICL algorithm (short for DIhedral-based Segment Identification and CLassification) classifies segments of structures containing two central residues. Here, we introduce the DISICL library for polynucleotides, which is based on the dihedral angles ε, ζ, and χ for the two central residues of a three-nucleotide segment of a single strand. Seventeen distinct structural classes are defined for nucleotide structures, some of which—to our knowledge—were not described previously in other structure classification algorithms. In particular, DISICL also classifies noncanonical single-stranded structural elements. DISICL is applied to databases of DNA and RNA structures containing 80,000 and 180,000 segments, respectively. The classifications according to DISICL are compared to those of another popular classification scheme in terms of the amount of classified nucleotides, average occurrence and length of structural elements, and pairwise matches of the classifications. While the detailed classification of DISICL adds sensitivity to a structure analysis, it can be readily reduced to eight simplified classes providing a more general overview of the secondary structure in polynucleotides.



INTRODUCTION

Since the first elucidation of three-dimensional models of protein structures and polynucleotides, a wealth of structural information has become available. For proteins, different secondary structure elements have been described, and also for DNA, two different helices were proposed very early on.[1,2] While proteins are readily classified in terms of helices, sheets, and various kinds of turns, the full structural diversity of DNA and RNA is only recently becoming clear. On the basis of our previous work on the classification of protein structures using an algorithm called DISICL (Dihedral-based Segment Identification and CLassification),[3] we here propose a definition of polynucleotide structural elements based on two dihedral angles of the nucleotide backbone complemented by the dihedral angle linking the sugar and the base.

While studies of backbone dihedral angles are available for polynucleotides,[4] secondary structure prediction and classification of DNA and RNA models are more often based on sequence- or knowledge-based approaches such as sequence alignments,[5−7] context free grammar, and machine learning[8,9] or empirical energy functions and dynamic programming[10,11] to determine the most stable secondary structure or an ensemble of structures with a central member. There has been significant effort to combine these methods to predict and construct both the secondary and tertiary structure of RNA molecules.[12−14] Structure-based analysis and classification methods on the other hand rely on three-dimensional models to evaluate the shape and intramolecular and intermolecular interactions between DNA, RNA, and other molecules such as proteins.[15−18] Established structure-based analysis methods for polynucleotides are commonly based on complex helical parameters, such as in the program SCHNAAP[19] (structure and conformation of helical nucleic acids: analysis program). The X3DNA analysis tool[20] also relies on helical parameters but performs a local DNA classification based on phosphate coordinates of dinucleotide base pair steps. Another very useful and effective package, Curves,[21] can analyze global helical curvature and local base pair parameters, as well as groove dimensions. While the recently reimplemented Curves+ program can effectively analyze molecular dynamics trajectories, its intrabase and interbase pair parameters, which can be used to assign local structure information, are almost identical to those reported by X3DNA.[22] Almost all of the structure-based approaches for DNA or RNA classification require double helical structures (originating from a single or from multiple strands) and seem relatively limited in terms of the diversity of the structural classes that are being considered.

In the current work, we define an extensive library for the classification of nucleotide sequences and apply this classification on databases containing 260,000 trinucleotide segments. After a description of the data sets to be analyzed, we shortly review the X3DNA tools and introduce the DISICL algorithm. The suggested classifications are discussed in more detail and demonstrated by selected examples of DNA and RNA structures obtained from the Brookhaven Protein

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

Table 1. Summary of Analyzed Polynucleotide Data Sets, Classification Efficiency of DISICL and X3DNA Algorithms, and Agreement between These Algorithms

nucleotide database		
database	DNA	RNA
file number	1,871	900
model number	8,044	5,109
total data set	94,080	187,602
ave. multiplicitiy	5.1	5.7
ave. model length	14.4	32.0
base pairing (%)	52.9	54.9
methods performance (DNA)		
method	DISICL	X3DNA
data set size	63,741	82,202
completeness (%)	67.8	87.4
classification ratio (%)	86.0	35.9
total efficiency (%)	58.3	31.4
methods performance (RNA)		
method	DISICL	X3DNA
data set size	164,453	152,405
completeness (%)	87.7	81.2
classification ratio (%)	84.4	57.9
total efficiency (%)	74.0	47.0
methods agreement		
DBSSP/X3DNA	DNA	RNA
A-helix match (%)	64.2	66.3
B-helix match (%)	59.3	5.9
transition match (%)	18.4	38.8
overall match (%)	57.3	59.6

Databank.[23] The performance of DISICL is compared to that of X3DNA, and finally, some conclusions are drawn.

METHODS

Data Sets. For the purpose of testing and comparing different classification algorithms, two large scale polynucleotide data sets were obtained from the Brookhaven protein databank (PDB, www.rscb.org).[23] 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.

Separate DNA and RNA data sets were defined (DNA_comb and RNA_comb, respectively), containing structural models determined by both X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy, because the number of entries for polynucleotides was considered too small for further partitioning. The resolution range for X-ray structures was chosen such that the relevant dihedral angles 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 nucleotides, 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. If any of the classification algorithms failed on a particular entry, it was completely discarded from the full analysis to ease the comparison of methods. This approach yielded approximately 80,000 and 180,000 segments for DNA and RNA models, respectively. Further details are provided in the database summary section of Table 1, where the number of downloaded pdb entries (file number), number of extracted individual structures (model number), and total number of classified nucleotides/residues (total data set) is provided, along with the average number of structures per pdb entry (ave. multiplicity), average number of nucleotides per structure (ave. model length), and amount of base-paired nucleotides (base pairing).

X3DNA Tools. 3DNA version 1.5 by Xiang Jun Lu et al. X3DNA[20] is a freely available analysis, reconstruction, and visualization tool for DNA modeling (www.x3dna.org). It has many smaller modules that can be used to produce idealized DNA models based on their sequence and the required helix type, as well as to analyze existing DNA and RNA structures. The analysis package can determine base pairing and obtain helical parameters (such as roll, twist, displacement, and groove dimensions) based on simple geometric calculations, and it also features a dinucleotide segment-based local helix classification algorithm. This classification takes the mean phosphorus atom Z coordinates and helix inclinations (Zp and ZpH, respectively) of A-DNA[24] to distinguish A-DNA, B-DNA, and transitory TA-DNA forms (often found in TATA boxes). This particular algorithm does not recognize Z-DNA forms (unless the full helix is left handed), and the classification should still be verified by other helical parameters also printed by the analysis program. Other structure-based programs focus on full helical descriptions of DNA sequences[21] and global hydrogen bonding,[5] while the X3DNA analysis program performs more localized dinucleotide segment classification, which is better suited to capture the structural diversity of, for example, molecular simulations.

dx.doi.org/10.1021/ci400542n | J. Chem. Inf. Model. 2014, 54, 278−288

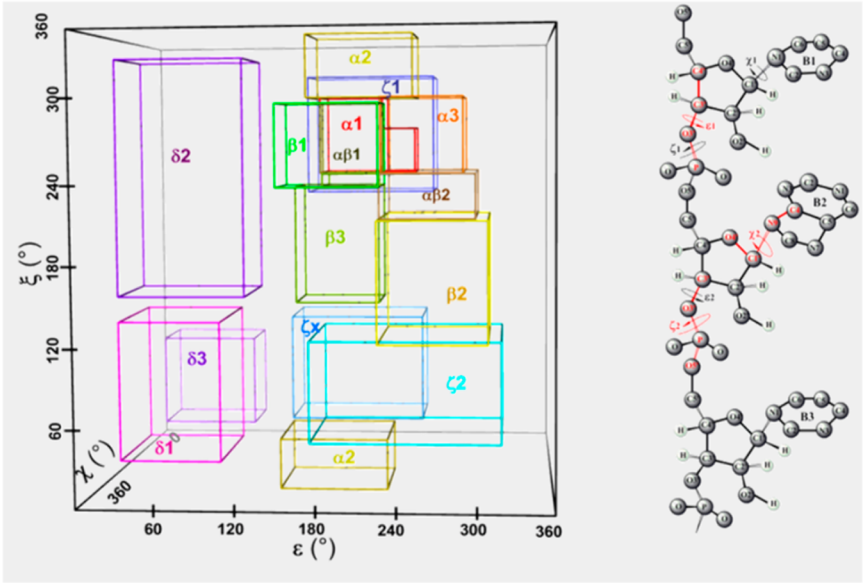


Figure 1. Representation of region definitions used for polynucleotide classification (on the left) based on subsequent (ε, ζ, χ) values within a trinucleotide segment (on the right). Colored rectangles show the boundaries of regions marked with Greek letters. Atoms and bonds that define $\varepsilon 1$, $\zeta 2$, and $\chi 2$ are marked in red.

Table 2. Definitions for DISICL Polynucleotide Classification[a]

DISICL polynucleotide classes		
class	code	segment definition
BI-helix	BI	$\beta 1.\beta 1$, $\beta 1$.ab1
BII-helix	BII	$\beta 1.\beta 2$, $\beta 2.\beta 1$
BIII-helix	BIII	$\beta 3.\beta 3$
B-loop	BL	$\beta 1.\beta 3$, $\beta 3.\beta 1$, $\beta 2.\beta 2$, $\beta 3.\beta 2$, $\beta 2.\beta 3$, $\beta 3.\alpha\beta 1$, ab1.$\beta 3$, $\beta 2$.ab1, ab1.$\beta 2$
A-helix	AH	$\alpha 1.\alpha 1$, $\alpha 1$.ab1
A-loop	AL	ab1.$\alpha 3$, $\alpha 3$.ab1, $\alpha 3.\alpha 3$, $\alpha 1.\alpha 3$, $\alpha 1.\alpha 2$
Z-helix	ZH	$\zeta 1.\zeta 2$, $\zeta 2.\zeta 1$, $\zeta 2.\zeta 3$, $\zeta 3.\zeta 2$
quad loop	QL	$\zeta 1.\zeta 1$, $\zeta 1.\zeta 3$, $\zeta 3.\zeta 1$, $\delta 1.\delta 1$, $\delta 3.\delta 3$, $\delta 1.\delta 3$, $\delta 3.\delta 1$, $\delta 3.\delta 2$, $\delta 2.\delta 3$, ab1.$\zeta 1$, $\zeta 1$.ab1, $\zeta 1.\beta 1$, $\beta 1.\zeta 1$, $\delta 3.\beta 1$, $\zeta 1.\beta 3$, $\beta 3.\zeta 1$, $\beta 1.\zeta 2$, $\zeta 1.\delta 1$, $\zeta 1.\alpha 3$, $\zeta 1.\beta 2$,
sharp turns	ST	$\zeta 2.\zeta 2$, $\alpha 3.\beta 3$, $\delta 2.\delta 2$, $\delta 2.\delta 1$, $\zeta 2.\alpha 3$, $\alpha 3.\beta 1$, $\delta 2.\beta 2$, $\zeta 2$.ab1, $\zeta 2.\beta 3$, $\zeta 2.\alpha 1$, ab2.$\beta 2$, $\zeta 2.\alpha 2$, $\zeta 2.\beta 1$, $\zeta 2.\beta 2$
tetraloop B	TL	$\alpha 2.\beta 2$, $\delta 1.\delta 2$, $\alpha 3.\alpha 1$, $\alpha 2.\alpha 2$, $\alpha 2.\alpha 1$, $\alpha 2.\beta 3$, $\alpha 2.\alpha 3$, $\alpha 3.\alpha 2$, $\alpha 3.\zeta 2$, $\alpha 3.\beta 2$, $\alpha 1.\zeta 2$ ab1.$\zeta 2$, $\delta 1.\zeta 1$, $\alpha 2$.ab2, $\alpha 2.\zeta 2$
AB trans.	AB	ab1.ab1, ab1.$\alpha 1$, ab1.$\beta 1$, $\alpha 1.\beta 1$, $\alpha 1.\beta 3$, $\beta 1.\alpha 1$, $\alpha 1$.ab2, $\beta 1$.ab2
AB2 trans.	AB2	$\beta 2.\alpha 3$, $\beta 3.\alpha 1$, $\beta 3.\alpha 3$, ab2.ab2, $\beta 3.\alpha 2$, $\alpha 1.\beta 2$, $\beta 2.\alpha 1$, ab2.$\alpha 1$, ab2.$\alpha 3$, ab2.$\beta 1$, ab2.d1, $\delta 1$.ab2,
AZ trans.	AZ	$\alpha 1.\zeta 1$, $\alpha 1.\zeta 3$, $\zeta 1.\alpha 1$, $\zeta 3.\alpha 1$, $\alpha 2.\zeta 1$, $\alpha 2.\zeta 2$, $\alpha 2.\zeta 3$, $\zeta 1.\alpha 2$, $\zeta 3.\alpha 2$, $\alpha 3.\zeta 1$, $\alpha 3.\zeta 3$, $\zeta 3.\alpha 3$
ZD trans.	ZD	$\zeta 1.\delta 2$, $\zeta 1.\delta 3$, $\delta 2.\zeta 1$, $\delta 3.\zeta 1$, $\zeta 2.\delta 1$, $\zeta 2.\delta 2$, $\zeta 2.\delta 3$, $\delta 1.\zeta 2$, $\delta 2.\zeta 2$, $\delta 3.\zeta 2$, $\zeta 3.\delta 1$ $\zeta 3.\delta 2$, $\zeta 3.\delta 3$, $\delta 1.\zeta 3$, $\delta 2.\zeta 3$, $\delta 3.\zeta 3$,
ZB trans.	ZB	$\beta 1.\zeta 3$, $\beta 2.\zeta 1$, $\beta 2.\zeta 2$, $\beta 2.\zeta 3$, $\beta 3.\zeta 2$, $\beta 3.\zeta 3$, $\zeta 3.\beta 1$, $\zeta 3.\beta 2$, $\zeta 3.\beta 3$,
BD trans.	BD	$\beta 1.\delta 1$, $\beta 1.\delta 2$, $\beta 1.\delta 3$, $\delta 1.\beta 1$, $\delta 2.\beta 1$ $\beta 2.\delta 1$, $\beta 2.\delta 3$, $\delta 1.\beta 2$, $\delta 3.\beta 2$, $\beta 3.\delta 1$, $\beta 3.\delta 2$ $\beta 3.\delta 3$, $\delta 1.\beta 3$, $\delta 2.\beta 3$, $\delta 3.\beta 3$, $\delta 1.\alpha\beta 1$,
AD trans.	AD	$\alpha 1.\delta 1$, $\alpha 1.\delta 2$, $\alpha 1.\delta 3$, $\alpha 3.\delta 3$, $\delta 1.\alpha 1$, $\delta 2.\alpha 1$, $\delta 3.\alpha 1$, $\delta 3.\alpha 3$,

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

DISICL for Polynucleotides. The DISICL algorithm for protein structure classification is described in more detail in the accompanying paper.[3] In short, DISICL is based on the classification of segments of biopolymers. First, relevant (backbone) dihedral angles are calculated and attributed to regions in the dihedral angle space. The pair of regions occupied by the two central residues of the segment determines the structural class to which the segment is assigned. While the DISICL algorithm was originally designed for protein analysis, the purely dihedral-based classification can be applied to other biopolymers as well. Using a study of dihedral angles in selected DNA backbones, we prepared region and class definitions for polynucleotides as well. Schneider et al. published one- and two-dimensional distributions on eight backbone dihedrals of DNA oligomers crystallized in different helical structures (A, B, Z).[4] This work provided an excellent base for our classifications, while other papers[25−29] confirm that helical structures and their subconformations are important factors when DNA interacts with proteins and drug-like molecules. Finally, RNA molecules, which often fold into complex structures, also have a tendency to form DNA-like helical segments.[20,24,30] On the basis of the two-dimensional distributions, we chose three dihedral angles that can characterize helical structures of nucleotides: backbone dihedrals ε and ζ and base torsion angle χ. While some helical parameters (like groove dimensions) can be more easily measured for full helix turns (4−5 base pair segments), a pair of the triplets (ε, ζ, χ) provided by a base triplet has a sufficiently high resolution to separate the polynucleotide helices. The backbone dihedral angle definitions and the 14

dx.doi.org/10.1021/ci400542n | J. Chem. Inf. Model. 2014, 54, 278−288

region definitions of the dihedral angle space are shown in Figure 1. Similar to the first and last amino acid of the protein segments of DISICL,[3] the third nucleotide only provides one atom for the calculations, and as such, the first two nucleotides were used as central residues for the comparison studies. On the basis of the $(\varepsilon, \zeta, \chi)^2$ definitions, 17 detailed (Table 3) and eight simplified (Table 4) classes were defined for DNA and RNA structures. Their region mapping and precise region definitions are shown in Table 2 and Table S1 of the Supporting Information, respectively.

The region definitions of the DNA classes are not as straightforward as the protein region definitions, so a summary is provided here. The 14 region definitions can be divided into five groups marked by Greek letters. α Regions (α1, α2, α3) have high density in RNA structures, with α1 containing the most densely populated area associated with the A-helix. β Regions (β1, β2, β3) are dominant in the DNA data set and are associated with the different forms of the B-helix. The experimentally derived subconformations of B-DNA, and BI and BII,[4,25,31] fall in the β1 and β2 regions, respectively. Three ζ regions (ζ1, ζ2, ζ3) contain local density peaks normally found in Z-DNA, although ζ1 residues are also regularly found in DNA quadruplexes, and ζ2 residues regularly appear in sharp backbone turns. δ Regions (δ1, δ2, δ3) have populations comparable to the ζ regions and are separated from the α, β, and ζ regions by their low ε value. The δ1 region appears in distorted A-helices, and the δ2 region regularly appears in backbone turns, while the δ3 region is almost exclusive found in DNA quadruplexes. The fifth group represents A/B transitions and contains the regions ab1 and ab2. The ab1 region represents the intersect volume of the α1 and β1 regions (which contain the maximal peak densities in the RNA and DNA data sets, respectively) and is densely populated for both RNA and DNA structures. Region ab2 is a moderate density volume surrounded by the regions α1, α3, β1, β2, and β3. Regions α1, β1, β2, ζ1, ζ2, and ζ3 are based on the angular distributions of Schneider et al.[4] but were modified to better fit the density distribution of both DNA and RNA data sets. This procedure was based on the classification of a data set containing 150,000 nucleotides consisting of approximately equal amounts of DNA and RNA. The rectangular regions were adjusted to include ~75% of the data points including the nearest local density maxima. Afterward, selected subsets of structures with common structural elements (DNA quadruplexes, junctions, RNA tetraloops, riboswitches, etc.) were analyzed. If structurally important nucleotides were repeatedly observed near unassigned peaks in the dihedral angle space, additional regions were assigned to those areas (resulting in regions α2, α3, β3, δ1, δ2, and δ3). On the basis of chemical intuition, visual checks concerning the shape of the backbone, directionality of the bases, and the annotation provided in the PDB entries, segment definitions (pairs of regions) were associated with structural classes. Associated segment definitions were assigned to a class after a careful visual analysis of 20−100 randomly picked structures. Segment definitions were assigned if at least 50% of the examples were of one particular class. Finally, the borders of neighboring regions were fine-tuned in an iterative process, where the effect of shifting the border was determined by performing structural analyses of structures containing the segments that were reassigned to a different class.

Comparison Studies. All structural models were analyzed separately by both 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 segments in the model that belonged to that class. 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 nucleotide n was classified to belong to class j. Tables 3 and 4 show the abundance (occ_j) and average length (L_j) of each structural element (in nucleotides), which were calculated based on the number of residues in the class (N_j), number of interruptions (N_j^{int}), and total number of residues (N_{sum}) according to eqs 1−3. The number of interruptions was increased by one whenever a gap was found in a continuous chain of dinucleotide segments of class j.

Table 3. Detailed DISICL Classes for Polynucleotide Classification, and Their Abbreviations (code), Occurrence (occ.), and Average Structure Element Lengths in the DNA and RNA Data Sets (top)[a]

method		DISICL		DISICL	
database		DNA		RNA	
class	code	occ. (%)	length	occ. (%)	length
BI-helix	BI	35.6	3.3	0.4	2.2
BII-helix	BII	3.4	2.3	0.1	2.0
BIII-helix	BIII	2.4	2.4	0.1	2.1
B-loop	BL	15.9	2.6	0.8	2.1
A-helix	AH	2.2	2.9	51.3	4.6
A-loop	AL	0.4	2.1	7.1	2.1
Z-helix	ZH	1.0	2.5	0.4	2.1
quad loop	QL	3.6	2.3	0.4	2.1
sharp turns	ST	3.1	2.2	2.1	2.1
tetraloop B	TL	1.6	2.0	8.6	2.1
AB trans.	AB	11.1	2.2	6.6	2.3
AB2 trans.	AB2	1.4	2.1	3.2	2.1
AZ trans.	AZ	0.3	2.1	1.0	2.1
ZB trans.	ZB	0.9	2.0	0.4	2.0
AD trans.	AD	0.3	2.1	1.0	2.3
BD trans.	BD	2.3	2.3	0.4	2.1
ZD trans.	ZD	0.6	2.2	0.2	2.1
unclassified	UC	14.0	3.1	16.0	3.3
method		X3DNA		X3DNA	
database		DNA		RNA	
class	code	occ. (%)	length	occ. (%)	length
A-helix	A	1.2	4.5	56.7	5.6
B-helix	B	34.2	5.0	0.02	2.8
transition	TA	0.5	3.0	1.2	3.7
unclassified	UC	64.1	4.5	42.1	3.1

[a]The same data is given for the X3DNA classes at the bottom of the table.

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

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

$$L_j = \left(\frac{N_j}{N_j^{\mathrm{int}}}\right) + 1 \tag{3}$$

dx.doi.org/10.1021/ci400542n | J. Chem. Inf. Model. 2014, 54, 278−288

Table 4. Simplified DISICL Classes for Polynucleotide Classification and Detailed Classes of Which They Are Formed, Occurrence (occ.), and Average Structure Element Lengths in DNA and RNA Data Sets

simplified class	detailed class	DNA		RNA	
name	code	occ (%)	length	occ (%)	length
B-helix	BI	35.6	3.3	0.4	2.2
irregular B	BII, BIII, BL	21.7	3.0	1.0	2.2
A-helix	AH	2.2	2.9	51.3	4.6
irregular A	AL, TL	2.1	2.1	16.0	2.8
Z-helix	ZH	1.0	2.5	0.4	2.1
quad loop	QL	3.6	2.3	0.4	2.1
AB transition	AB	11.1	2.5	6.6	2.3
transitory	AB2, ST, AZ, BZ, ZD, AD, BD	8.8	2.2	8.0	2.3
unclassified	unclassified	14.0	3.1	16.0	3.3

To compare the classification algorithms, the correlation matrices of algorithms were calculated containing the correlation scores C_{ij} where i and j mark the i^{th} class of the first algorithm and the j^{th} 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 4, where $\bar{a}_{ni}$ is the average occurrence of the class i ($\bar{a}_{ni} = N_i/N_{sum}$).

$$R_{ij} = \frac{\sum_{n=1}^{N_{sum}} (a_{ni} - \bar{a}_{ni})(a_{nj} - \bar{a}_{nj})}{\sqrt{\sum_{n=1}^{N_{sum}} (a_{ni} - \bar{a}_{ni})^2 \sum_{n=1}^{N_{sum}} (a_{nj} - \bar{a}_{nj})^2}} \quad (4)$$

While the R-score drops quickly with the amount of mismatches (or different occurrences of classes i and j), a large positive R-score is still a good measure to determine agreement between algorithm classes. The unscaled match score (M_{ij}) is calculated using eq 5 and represents the absolute number of residues assigned to class i in one algorithm and to class j in the other algorithm.

$$M_{ij} = \sum_{n=1}^{N_{sum}} (a_{ni} a_{nj}) \quad (5)$$

The M-score is additive, which makes it possible to group classes or track distributions of correlations for one class. The scaled match score (M^s_{ij}) provides a better comparison between algorithms and is calculated by eq 6.

$$M^s_{ij} = \frac{M_{ij}}{M_{max}} \times 100 \quad (6)$$

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}). Here, M_{max} is equal to size of the smaller data set.

$$M_{max} = \min\{N_i, N_j\} \quad (7)$$

To summarize comparisons, the weighted average of the scaled match scores were calculated for A-helical, B-helical, and transitory DNA or RNA forms for nucleotides (Table 1, methods agreement). Additionally, the weighted average of all these superclasses and the scaled match score for unclassified residues were calculated to obtain an overall match between methods. The grouping for superclasses is provided in Table S2 of the Supporting Information.

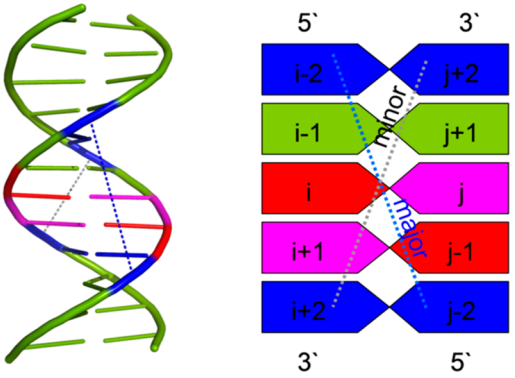


Figure 2. Schematic representation of the calculation of groove dimensions in double-stranded DNA helices. Groove dimensions are calculated as distances of phosphorus atoms in the indicated nucleotides. See the corresponding part of the Methods section for further information.

For DNA classifications, DNA groove dimensions were measured with a simple algorithm using a similar basic idea as used in X3DNA[32] (see Figure 2 for a schematic representation of relevant nucleotides for this calculation). Because the full turn of the B-DNA structure consists of approximately five (base-paired) nucleotides on each of the two strands, helical fragments of a given classification with five consecutive base pairs identified were used to determine the groove dimensions. Groove dimensions were assigned to the paired segments S_i–S_{j-1}, paired with central residues i–j and $(i + 1)$–$(j - 1)$, such that base pair i–j was the middle of the helical turn. As a rough estimate for major and minor groove widths, the distance between phosphorus atoms $P_{(i-2)}$ and $P_{(j-2)}$ yields the major groove width, while the distance between atoms $P_{(i+2)}$ and $P_{(j+2)}$ provides the minor groove width. Groove depths were estimated by the distance between the midpoint of the vector defining the width and the midpoint of the vector P_i–P_j.

RESULTS AND DISCUSSION

DISICL Nucleotide Classes. The classification of polynucleotides was performed on two data sets containing models of DNA molecules (DNA_comb) and RNA molecules (RNA_comb) using two segment-based algorithms, namely, X3DNA and DISICL.

DISICL defines 17 detailed classes, which can be grouped into classical helical structures (A-, BI-, BII-, BIII-, and Z-helices), special loops and turns (A-loop, tetraloop bulge, B-loop, sharp turns, and quadruplex loops), and transitory classes (AB, AB2, AD, AZ, BD, BZ, and ZD) (see Table 3 for the average occurrences and lengths of these structural elements). In the simplified version of the nucleotide DISICL library (Table 4), this is reduced to eight classes (A-, B-, Z-helices, irregular A and irregular B structures, quadruplex loops, AB transitions, and other transitory segments).

Helical Classes. The majority of DNA and RNA molecules in the databases assume double helical structures. DNA under physiological conditions assumes a right-handed double helical form usually referred to as B-DNA. The nucleotides in the B-DNA form have two identified subconformations (BI and BII) mostly differing in their ε and ζ angle. Under certain salt concentrations, RNA and DNA can form a different helix, normally referred to as the A-form, while some DNA structures

dx.doi.org/10.1021/ci400542n | J. Chem. Inf. Model. 2014, 54, 278−288

Table 5. Average Groove Dimensions for Various DNA Double Helices Observed in the DNA Data Set[a]

sorted groove dimensions (DNA)		MGW		MGD		mgW		mgD	
structure	occurrence	mean	rmsf	mean	rmsf	mean	rmsf	mean	rmsf
BI-helix/BI-helix	2511	17.5	2.8	9.4	1.2	12.9	2.4	8.3	1.1
BI-helix/BII-helix	185	19.1	2.9	8.9	2.0	13.4	2.6	8.2	1.0
BI-helix/BIII-helix	144	18.2	3.1	9.5	1.5	12.0	2.8	8.3	1.0
BI-helix/B-loop	1217	18.5	3.1	9.3	1.7	13.4	3.0	7.9	1.8
BI-helix/A-helix	19	20.4	3.7	10.5	1.5	12.5	2.3	8.6	1.5
BI-helix/Z-helix	5	21.3	0.6	3.0	0.5	13.4	0.1	8.5	0.4
BI-helix/AB	938	18.1	3.1	9.6	1.5	13.0	2.4	8.2	1.1
BII-helix/BII-helix	85	21.0	3.5	8.7	1.1	13.3	3.1	8.6	1.1
BII-helix/BIII-helix	26	17.7	3.7	8.8	1.1	12.4	3.2	8.9	1.2
BII-helix/B-loop	131	18.9	3.0	9.2	1.5	11.9	2.6	8.4	1.1
BII-helix/A-helix	3	16.3	2.4	7.7	0.7	15.6	4.7	7.9	2.4
BII-helix/AB	132	18.4	2.0	8.8	1.5	12.0	2.6	8.5	1.1
BIII-helix/BIII-helix	42	20.5	3.1	8.6	1.1	11.3	3.0	8.7	0.5
BIII-helix/B-loop	213	18.7	2.9	9.1	1.2	12.1	2.7	8.5	0.8
BIII-helix/A-helix	3	18.7	2.8	9.5	1.0	13.6	0.3	7.2	0.2
BIII-helix/AB	47	19.9	4.5	9.0	2.0	11.8	2.6	8.4	1.1
B-loop/B-loop	617	19.3	3.1	9.1	1.4	12.3	2.4	8.5	1.2
B-loop/A-helix	17	23.1	5.7	10.3	2.1	12.7	1.7	8.3	2.0
B-loop/AB	429	20.1	4.0	9.0	1.7	12.5	2.4	8.1	1.3
A-helix/A-helix	147	15.2	2.4	10.0	0.5	17.2	1.0	6.0	0.8
A-helix/AB	43	19.6	4.8	10.5	1.2	13.8	2.9	7.5	1.6
Z-helix/Z-helix	1	21.0	0.0	5.9	0.0	13.3	0.0	5.9	0.0
AB/AB	274	18.8	3.7	9.8	1.3	12.5	2.1	8.2	1.1
overall average	7229	18.3	3.2	9.4	1.5	12.9	2.6	8.2	1.3

[a]Helices are sorted based on the assigned DISICL classification for the central segment of the helix turn on both strands. Groove dimensions are given as averages (mean) and root-mean-square fluctuation (rmsf) in Å. MGW: major groove width. MGD: major groove depth. mgW: minor groove width. mgD: minor groove depth.

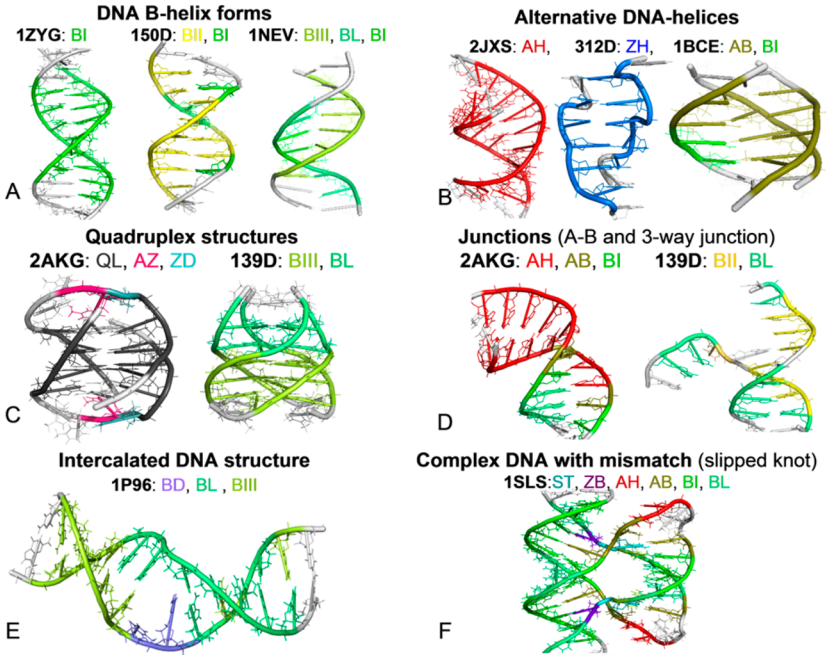


Figure 3. Examples of DNA structures and structure classification by DISICL. For each model, the PDB identification code is given followed by the abbreviation of classes according to Table 3, which are color coded to match the structures they mark.

can assume left-handed helices, normally referred to as the Z-form. Our helical classes represent these structures.

The BI class (BI) contains the DNA (ε, ζ, χ) density maximum associated with continuous repeats of the BI subconformation (located in the region $\beta1$). Occurrence of longer stretches of the BI class on both strands forms the classical B-helix, which makes up 35% of all DNA nucleotides. The BII class (BII) contains definitions for ($\beta1-\beta2$) alternating segments, which are an alternate form of the B-helix (3.5% of DNA segments). While the BII class rarely appears in longer

dx.doi.org/10.1021/ci400542n | J. Chem. Inf. Model. 2014, 54, 278−288

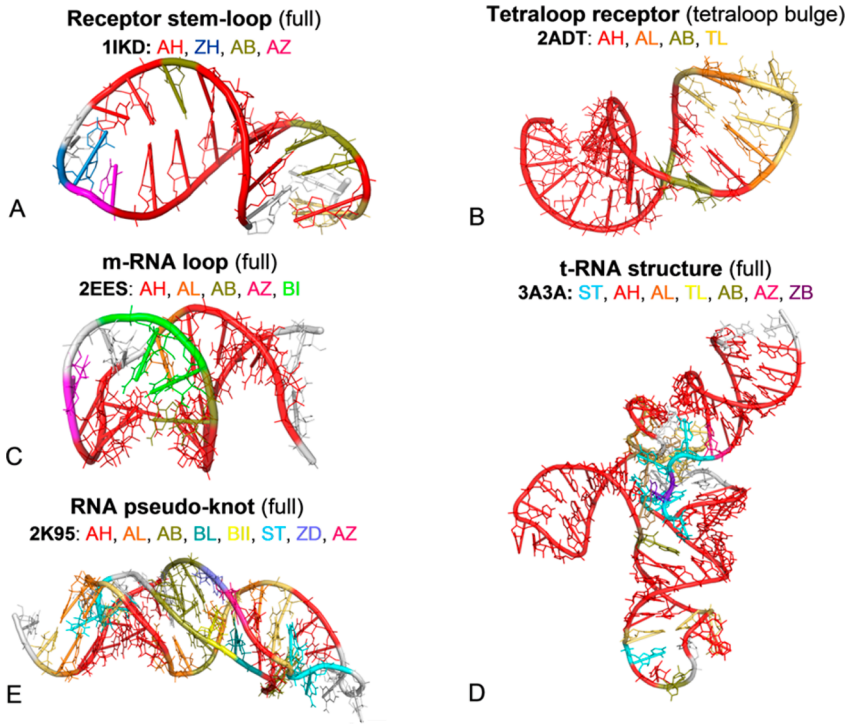


Figure 4. Examples of RNA structures and structure classification by DISICL. For each model, the PDB identification code is given followed by the abbreviation of classes according to Table 3, which are color coded to match the structures they mark.

stretches in both strands, BII-rich areas of DNA form helices that are more varied in their groove dimensions and on average have wider and more shallow grooves (Table 5), which might be important for DNA−protein and DNA−drug interactions. The BII class in longer stretches also appears in single strands for DNA loops and three-way junctions. The BIII class (BIII) was defined for pure $\beta 3$ segments, which occur in 2.5% of the analyzed nucleotides. BIII segments (often accompanied by B-loop segments) distort the B-DNA helix leading to a wider major groove but narrower minor groove. Examples of BI-, BII-, and BIII-rich DNA models are depicted in panel A of Figure 3.

The A-helix class (AH) relates to the bent A-form helix of DNA (2%), and it is the predominant form of ribonucleic acids (50%). The class is defined by segments with pure $\alpha 1$ conformation, which usually appear in fully A-helical models or prior to turns in more complex DNA structures. Examples are shown in panels B and F of Figure 3.

The Z-helix class (ZH) appears predominantly in Z-helical DNA structures and consists of definitions with an alternating pattern of either $\zeta 1-\zeta 2$ or $\zeta 1-\zeta 3$. It is the least common of the helices (occurrence is 1%). The Z-helix class appears consecutively only in Z-helical DNA models (see, for example, panel B of Figure 3), but it is observed isolated in segments of DNA loops and quadruplexes.

In RNA structures, the predominant class by far is the A-helix, containing over 50% of RNA residues, building up the helices and stem loops that form the majority of the more complex structures. Segments which are classified as B-helix appear with less than 2% occurrence and mostly at isolated positions. These segments sometimes have a backbone shape different from the normal helical forms appearing in DNA, resembling more the sharp turn class (this is especially common for segments of the B2 class). Z-helical segments also appear at isolated positions in RNA structures, mostly at the end of stem-loops with receptor functions, suggesting an important functional role (as shown in Figure 4 panel A).

Loops and Sharp Turns. Apart from the classes associated with the classical DNA helices, a number of special classes were defined for functionally important segments mostly found in more complex RNA and DNA structures. While the classes defined here help to monitor possible structurally important parts of polynucleotide structures, these structures do not separate sharply in the $(\varepsilon, \zeta, \chi)^2$ space, leading to a lower (40−70%) selectivity for individual definitions.

The quadruplex loop class contains definitions highly specific for DNA quadruplexes, which typically appear at the ends of chromosomes. The quadruplex loops rarely appear in longer stretches than three residues and instead are connected by sharp turns and transitory structures to form repeats. As quadruplexes are mainly formed in DNA structures, the occurrence of this class is significantly higher in the DNA data set (3.5%) than in the RNA set (0.5%) While the quadruplex loop class is highly selective for quadruplex structures (especially for quadruplexes made from one or two strands), quadruplexes formed by multiple strands of DNA can exist with one, two, or all four parts built from B-helical segments (see examples in panel C of Figure 3).

The tetraloop bulge class (TL) was defined for a special bulged loop structure, which appears often in RNA loops. The model structure of this class derived from tetraloop receptors, where the loop contains at least one ~90° turn in the backbone, with a base facing outward from the loop to interact with bases further away in the RNA sequence, possibly playing an important structural role. The occurrence of the class is 9% in RNA. However, based on visual checks, it is only moderately selective for the required shape, and many segments belong to the more general A-loop class. A bulged loop from a tetraloop receptor is shown in Panel B of Figure 4.

dx.doi.org/10.1021/ci400542n | J. Chem. Inf. Model. 2014, 54, 278−288

Table 6. Scaled Match Scores for Comparison of Secondary Structure Classifications by DISICL (simple) and X3DNA on the Combined DNA and RNA Data Set[a]

class	XDNA	A-helix	B-helix	TA trans.	unclassified
DISICL	%	37.5	12.19	0.9	49.3
B-helix	10.2	0.1	47.0	7.9	46.5
irregular B	6.8	1.3	38.7	2.8	48.0
A-helix	37.6	66.3	0.3	21.7	27.5
irregular A	12.1	29.5	0.2	7.3	44.1
Z-helix	0.6	0.8	0.0	0.0	43.9
Quad loop	1.3	1.6	3.8	0.1	74.2
AB transition	8.2	11.1	4.2	1.8	44.4
transitory	7.9	20.5	13.6	33.7	49.0
unclassified	15.4	11.8	4.2	9.8	42.6

[a]For both algorithms, the occurrence of each class is displayed in the first row or column, respectively.

The sharp turn class (ST) collects definitions, which are enriched in segments with a more than 90° turn in the backbone and/or the torsion of their bases (defined by the atoms $C1_i$, $C1'_i$, $C1'_{i+1}$, $C1_{i+1}$). Sharp turn segments typically appear where the bases of the stem loops are connected, at the end of certain riboswitch and aptamer RNA loops and in DNA and RNA knot structures. The occurrence of the sharp turn class is less than 2% in both RNA and DNA data sets, and sharp turns typically appear as isolated segments. For examples of the sharp turn, see panels D and E of Figure 4.

The A-loop class (AL) contains definitions of the α region, which were not found to be highly selective for any of the special classes, and they are typically not forming a perfect A-helix either. A-loop structures appear often in and between RNA-stem loops, connecting the classical A-helical segments with each other or with TL and ST segments. The A-loop class takes up about 10% of all RNA structures, but it is rarely found in DNA. The AL residues can form longer stretches as well, but these stretches are often single stranded or have significant distortions compared to A-helix structures. Examples of A-loop segments in different RNA structures are shown in panels B, C, and D of Figure 4.

The B-loop class (BL) contains the atypical definitions of β-regions. Similar to the AL class, B-loop segments usually connect the B-helical parts of DNA models and are often found in different junctions (Holliday junctions, kissing complexes, etc.), DNA-loop structures, and at sites where small molecules are intercalated into a DNA helix. Longer helical stretches of B-loop structures also appear in single strands, typically complemented by pure BIII segments on the other strand. The average occurrence of B-loop segments in DNA is 16% and around 1% in RNA models. Examples of B-loop class segments are depicted in Figure 3.

Transitory Structures. The AB class collects definitions for segments with a transition from the density maxima of RNA and DNA structures (α1 and β1 region, respectively). The volume bridging these two peaks is also highly populated in both data sets (around 10%) and was suggested to have a functional structure of its own.[20,24] We found that the AB class is often observed in helical structures in three functional roles: (1) Isolated or short AB segments often serve as junctions for A-helical and B-helical parts of both DNA and RNA (as shown on the left side of panel D in Figure 3). (2) Short stretches of AB segments temper the bending of A-helices in RNA stem-loops allowing for less strained loop structures (panel A in Figure 4). (3) Longer stretches of AB segments (especially pure ab1 stretches) are often found in three stranded structures, like DNA triplexes (right side of panel B in Figure 3) and RNA pseudoknots (panel E in Figure 4).

The AB2 class collects definitions typically transiting between the α3 and β2 regions. Unlike the AB class that mostly looks helical, AB2 segments typically appear more linear as the backbone dihedrals are close to 180° with bases looking well aligned or pointing away from each other. This nonideal position for stacking interactions agrees well with the observation of the AB2 class near unpaired or mismatched nucleotides and interaction sites of more bulky drug molecules. The remaining five transitory classes, namely, the AD, AZ, BD, BZ, and ZD, were defined based on the major areas that their segments connect. No particular selectivity for any of the previous classes was detected for the definitions of which they are comprised. These classes contain 3% of nucleotides in both data sets and have a similar role as the different turn definitions in the protein classification libraries.

Simplified Nucleotide Library. The simplified DISICL library for nucleotides is designed to provide an easier comparison to CD spectroscopy, where A-, B-, Z-, and quadruplex forms of DNA can be distinguished from each other. For this reason, the detailed DISICL classes BI-helix (renamed to B-helix or BH), A-helix, Z-helix, and quadruplex loop remain as separate classes in the simplified library. As the BII-helix, BIII-helix, and B-loop classes relate to distorted but mostly B-helical forms of the DNA, they were grouped together into the irregular B (IB) class. The A-loop and tetraloop bulge classes usually appear in RNA-stem loops and are not sharply separated in the $(\varepsilon, \zeta, \chi)^2$ dihedral angle space. They are grouped together to form the irregular A (IA) class in the simplified classification. Although the AB class is a transitory class, we decided to keep it as a separate class in the simplified classification because of its high abundance, enrichment in special helical segments, and definition through its own region (ab1). The remaining seven classes in the detailed classification (ST, AB2, AZ, AD, BD, ZB, and ZD) typically stand for nonhelical segments, which connect helical parts in stem loops and other complex polynucleotide structures grouped together in the transitory (TR) class in the simplified classification. The average structure element length and occurrence of the simplified DISICL classes for nucleotides is shown in Table 4, along with the codes of the detailed classes grouped together in each simplified class.

Correlation Analysis. Full correlation matrices (Pearson scores and scaled match scores) for the comparison of DISICL and X3DNA are given in Tables S3−S6 of the Supporting Information. Here, we provide an overview of the overall

dx.doi.org/10.1021/ci400542n | J. Chem. Inf. Model. 2014, 54, 278−288

correlation analysis in Table 6. The abundance and average lengths of A- and B-helix structures are similar for the two algorithms (Table 3). The average helix length of DISICL is shorter due to the more detailed classification and the fact that DISICL can classify one residue less for fully base-paired chains (as X3DNA requires a base-paired dinucleotide step for classification, while DISICL uses a segment of three nucleotides in one strand).

Correlation analysis reveals that the assigned helical structures show only a partial overlap between the algorithms. A-helix classes of DISICL and X3DNA show the best agreement both in the DNA and the RNA data sets amounting to 65% of DISICL residues and Pearson correlation scores close to 0.6. In the DNA data set, the B-helical classes as determined by DISICL, show a comparable amount of correlation with the B-helix in X3DNA. Highest correlations are observed for the B-helix class or BI-helix (M^s score 48% and R-score 0.4), closely followed by the irregular B class, for which 43% of the residues are also classified as B-form by X3DNA (with similar values for B2, B3, and BL classes), but its lower abundance decreases the R-score to 0.3. For the RNA data set, the amount of segments assigned by X3DNA as B-helix is extremely low (0.04%) and shows little or no correlations with the DISICL B-helical classes (or any other class), leading to a combined agreement amounting to 6% the X3DNA class. The abundance of B-helical segments in RNA according to DISICL is 1.5% (mostly due to the B-loop class). Visual checks reveal that X3DNA B-helix segments do not show the shape normally associated with DISICL B-helical segments, while DISICL B-helix segments appear in RNA mostly as bulges in A-helices or at the end of stem-loops (an example is shown in Panel C of Figure 4). We found no models in the RNA data set, with hydrogen-bonded base pairs for which DISICL classified both strands as B-helix, which partially explains the low correlation with X3DNA as this program monitors paired bases only. While the Z-helix class in DISICL has a low abundance (1% and 0.4% in DNA and RNA, respectively), it has no overlap with any of the X3DNA classes in the DNA data set and a minimal overlap with the A-helix class in RNA (due to the isolated Z-helix segments in RNA stem loops). This shows that X3DNA very rarely mistakes Z-helical segments for A- and B-form segments, even though not explicitly making the classification (except for full Z-helices in DNA).

Considering transitory and special classes, the TA-transitory class of X3DNA shows moderate correlations (M^s scores) with the BI-helix (32%), AB (13%), and B-loop (12%) classes of DISICL in DNA and with the AB (38%) and A-helix (26%) classes in RNA, showing that the peak of its density distribution falls in the ab1 region. The correlation might be low for DNA because the TA class was based on special DNA segments meant for interacting with polymerase enzymes, and protein−nucleotide complexes were filtered out from our data sets. About one-third of the AB class in DISICL was considered as B-helix in DNA (34%) and A-helix in RNA (33%) by X3DNA. Additionally the BD class shows a moderate correlation (30%) with X3DNA B-helix in DNA, while AD (15%) and AB2 (14%) classes correlate weakly. In RNA models, moderate agreement with the X3DNA A-helix is also observed for the A-loop (39%) and AD (38%) classes (often found in distorted A-helices) and the tetraloop bulge class (25%). The rest of the DISICL classes remained mainly unclassified by X3DNA with no significant correlations. A summary of the correlation analysis is shown in Table 1 (methods agreement), which reveals an overall agreement between X3DNA and DISICL slightly below 60% for both the RNA and DNA data sets, which is slightly lower than the agreement between protein classification algorithms.

Protein−Nucleotide Complexes. The analysis of the DNA and RNA data sets provides a solid basis to define and characterize the DISICL nucleotide structure classes, and their correlations with the classification of X3DNA. The higher level of detail in DISICL allows us to monitor the structural effects of interactions of nucleotides with small molecules and proteins as well. Two examples for RNA−protein and DNA−protein complexes are shown in Figure 5. Panels A and B show a model

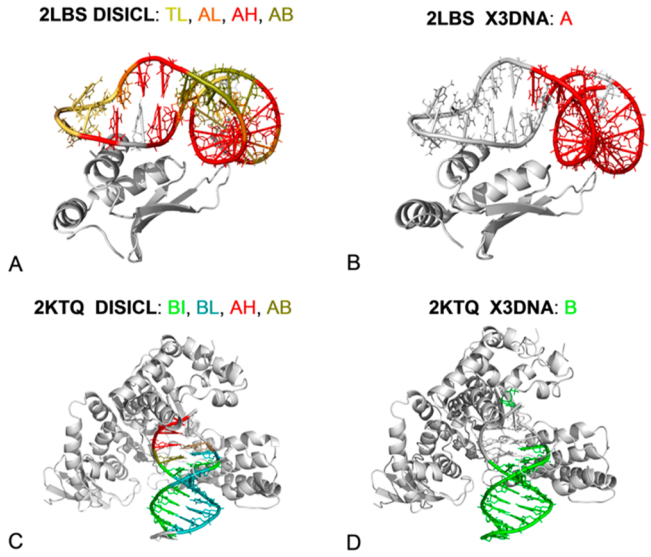


Figure 5. Examples of DNA/RNA−protein complexes classified by DISICL and X3DNA. For each model, the PDB identification code is given, followed by the method of classification and the abbreviation of structural classes according to Table 3. Abbreviations are color coded to match the structures they mark.

of an AAUG tetraloop hairpin in complex with a yeast RNase binding domain (PDB code 2LBS). The bulk of the interactions take place between a short α-helix of the protein at the end of the tetraloop hairpin. While X3DNA steadily recognizes the A-helical conformation at the base of the hairpin, the interaction site remains unclassified. DISICL assigns a classification for 70% of the nucleotides over the NMR solution models, mainly to the tetraloop bulge or AB2 class. Another interesting example is the ternary complex of double-stranded DNA and a protein fragment of the polymerase I from *T. aquaticus* (PDB code 2KTQ). In this case—shown in panels C and D of Figure 5—the longer template strand of the mainly B-form DNA is bent by the protein, recognized as an A-helical stretch in DISICL. As in the previous example, X3DNA readily recognizes the B-helical nature of the double-stranded part but leaves the DNA at the interaction site unclassified. The examples suggest that fine structural changes might be revealed by DISICL, yielding additional information on interactions between nucleotides, proteins, and small molecules.

CONCLUSIONS

The DISICL algorithm for dihedral-based structure classification was extended to allow for the classification of nucleotide structures. Starting from previously published distributions of dihedral angles, three dihedral angles (ε, ζ, χ) were selected to perform the classifications. Fourteen distinct regions were

dx.doi.org/10.1021/ci400542n | J. Chem. Inf. Model. 2014, 54, 278−288

defined in the resulting three-dimensional dihedral angle space. A classification is performed based on the assignment of the two central nucleotides in a trinucleotide segment, first to their regions and as a pair to one of 17 structural classes. Apart from helical structures, we define loop regions, turns, and transitory structural elements, and examples of these were given with DNA and RNA models from the Brookhaven PDB. Newly suggested structural classes include the quadruplex loop, sharp turn, and tetraloop bulge, as well as a number of transitory elements. The detailed classification was simplified into eight more general classes and were compared to the classification in X3DNA. Overall, DISICL seems a very powerful tool for the detailed structural analysis of both proteins and polynucleotides.

Studies of practical applications for the DISICL algorithm are currently the focus of our attention. Additionally, a new application of DISICL for carbohydrate structures is under consideration.

■ ASSOCIATED CONTENT

Supporting Information

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

■ AUTHOR INFORMATION

Corresponding Author

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

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

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) Watson, J.; Crick, F. Genetical implications of the structure of deoxyribonucleic acid. *Nature* **1953**, *171*, 964−967.

(2) Watson, J.; Crick, F. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature* **1953**, *171*, 737−738.

(3) Nagy, G.; Oostenbrink, C. Dihedral-based segment identification and classification of biolpolymers I: Proteins. *J. Chem. Inf. Model* **2013**, DOI: 10.1021/ci400541d.

(4) Schneider, B.; Neidle, S.; Berman, H. M. Conformations of the sugar-phosphate backbone in helical DNA crystal structures. *Biopolymers* **1997**, *42*, 113−124.

(5) Shapiro, B. An algorithm for comparing multiple RNA secondary structures. *Comput. Appl. Biosci.* **1988**, *4*, 387−393.

(6) Pedersen, J. S.; Bejerano, G.; Siepel, A.; Rosenbloom, K.; Lindblad-Toh, K.; Lander, E. S.; Kent, J.; Miller, W.; Haussler, D. Identification and classification of conserved RNA secondary structures in the human genome. *PloS Comp. Biol.* **2006**, *2*, e33.

(7) Washietl, S.; Pedersen, J. S.; Korbel, J. O.; Stocsits, C.; Gruber, A. R.; Hackermüller, J.; Hertel, J.; Lindemeyer, M.; Reiche, K.; Tanzer, A. Structured RNAs in the ENCODE selected regions of the human genome. *Genome Res.* **2007**, *17*, 852−864.

(8) Knudsen, B.; Hein, J. RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics* **1999**, *15*, 446−454.

(9) Van Batenburg, F. H. D.; Gultyaev, A. P.; Pleij, C. W. An APL-programmed genetic algorithm for the prediction of RNA secondary structure. *J. Theor. Biol.* **1995**, *174*, 269−280.

(10) Zuker, M.; Stiegler, P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.* **1981**, *9*, 133−148.

(11) Ren, J.; Rastegari, B.; Condon, A.; Hoos, H. H. HotKnots: Heuristic prediction of RNA secondary structures including pseudoknots. *RNA* **2005**, *11*, 1494−1504.

(12) Lorenz, R.; Bernhart, S. H.; Zu Siederdissen, C. H.; Tafer, H.; Flamm, C.; Stadler, P. F.; Hofacker, I. L. ViennaRNA Package 2.0. *Algorithms Mol. Biol.* **2011**, *6*, 26.

(13) Jossinet, F.; Ludwig, T. E.; Westhof, E. Assemble: An interactive graphical tool to analyze and build RNA architectures at the 2D and 3D levels. *Bioinformatics* **2010**, *26*, 2057−2059.

(14) Reuter, J. S.; Mathews, D. H. RNAstructure: Software for RNA secondary structure prediction and analysis. *BMC Bioinf.* **2010**, *11*, 129.

(15) Bélanger, F.; Léger, M.; Saraiya, A. A.; Cunningham, P. R.; Brakier-Gingras, L. Functional studies of the 900 tetraloop capping helix 27 of 16S ribosomal RNA. *J. Mol. Biol.* **2002**, *320*, 979−989.

(16) Siggers, T. W.; Silkov, A.; Honig, B. Structural alignment of protein−DNA interfaces: Insights into the determinants of binding specificity. *J. Mol. Biol.* **2005**, *345*, 1027−1045.

(17) Liu, Z.-P.; Wu, L.-Y.; Wang, Y.; Zhang, X.-S.; Chen, L. Prediction of protein−RNA binding sites by a random forest method with combined features. *Bioinformatics* **2010**, *26*, 1616−1622.

(18) Kim, R.; Guo, J. PDA: An automatic and comprehensive analysis program for protein−DNA complex structures. *BMC genomics* **2009**, *10*, S13.

(19) Lu, X. J.; ElHassan, M. A.; Hunter, C. A. Structure and conformation of helical nucleic acids: Analysis program (SCHNAaP). *J. Mol. Biol.* **1997**, *273*, 668−680.

(20) Lu, X.-J.; Olson, W. K. 3DNA: A software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res.* **2003**, *31*, 5108−5121.

(21) Lavery, R.; Sklenar, H. The definition of generalized helicoidal parameters and of axis curvature for irregular nucleic acids. *J. Biomol. Struct. Dynam.* **1988**, *6*, 63−91.

(22) Lavery, R.; Moakher, M.; Maddocks, J. H.; Petkeviciute, D.; Zakrzewska, K. Conformational analysis of nucleic acids revisited: Curves+. *Nucleic Acids Res.* **2009**, *37*, 5917−5929.

(23) 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.

(24) Lu, X. J.; Shakked, Z.; Olson, W. K. A-form conformational motifs in ligand-bound DNA structures. *J. Mol. Biol.* **2000**, *300*, 819−840.

(25) Oguey, C.; Foloppe, N.; Hartmann, B. Understanding the sequence-dependence of DNA groove dimensions: implications for DNA interactions. *PloS one* **2010**, *5*, e15931.

(26) Rohs, R.; West, S. M.; Sosinsky, A.; Liu, P.; Mann, R. S.; Honig, B. The role of DNA shape in protein−DNA recognition. *Nature* **2009**, *461*, 1248−U81.

(27) Peberdy, J. C.; Malina, J.; Khalid, S.; Hannon, M. J.; Rodger, A. Influence of surface shape on DNA binding of bimetallo helicates. *J. Inorg. Biochem.* **2007**, *101*, 1937−1945.

(28) Spitzer, G. M.; Wellenzohn, B.; Markt, P.; Kirchmair, J.; Langer, T.; Liedl, K. R. Hydrogen-bonding patterns of minor groove-binder−DNA complexes reveal criteria for discovery of new scaffolds. *J. Chem Inf. Model.* **2009**, *49*, 1063−1069.

(29) Fuchs, J. E.; Spitzer, G. M.; Javed, A.; Biela, A.; Kreutz, C.; Wellenzohn, B.; Liedl, K. R. Minor groove binders and drugs targeting proteins cover complementary regions in chemical shape space. *J. Chem Inf. Model.* **2011**, *51*, 2223−2232.

(30) Subbotin, S. A.; Sturhan, D.; Vovlas, N.; Castillo, P.; Tambe, J. T.; Moens, M.; Baldwin, J. G. Application of the secondary structure model of rRNA for phylogeny: D2−D3 expansion segments of the

LSU gene of plant-parasitic nematodes from the family Hoplolaimidae Filipjev, 1934. *Mol. Phylogen. Evol.* **2007**, *43*, 881−890.
(31) Maehigashi, T.; Hsiao, C.; Woods, K. K.; Moulaei, T.; Hud, N. V.; Williams, L. D. B-DNA structure is intrinsically polymorphic: Even at the level of base pair positions. *Nucleic Acids Res.* **2012**, *40*, 3714−3722.
(32) El Hassan, M. A.; Calladine, C. R. Two distinct modes of protein-induced bending in DNA. *J. Mol. Biol.* **1998**, *282*, 331−343.