

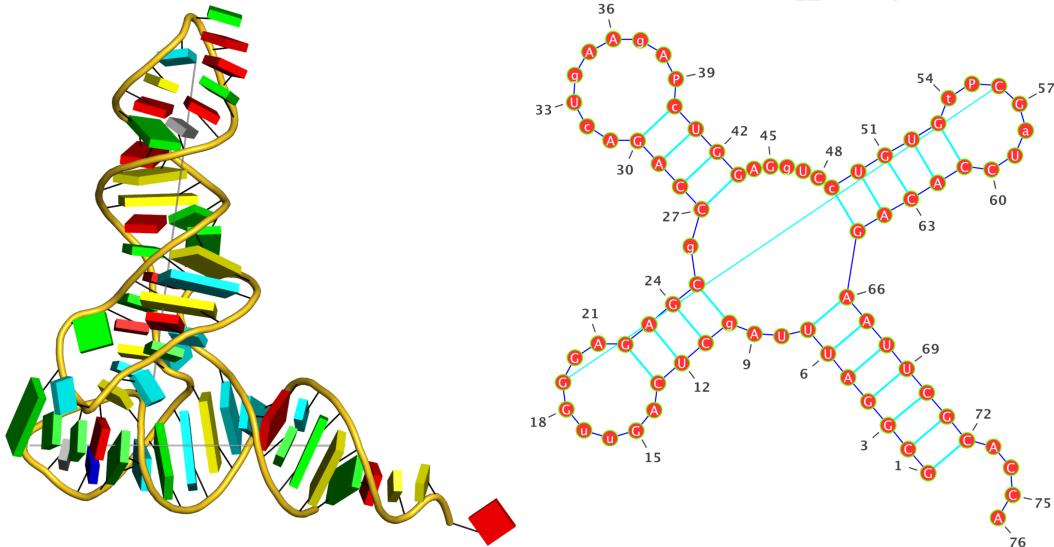


# DSSR: an integrated software tool for Dissecting the Spatial Structure of RNA

## User Manual for DSSR 2.2 (Basic Version)

*for noncommercial use only*

Last updated: January 12, 2021



modified nucleotide, non-canonical base pair, helix, stem, coaxial stacking  
hairpin/internal/junction loop, kink turn, G-quadruplex, i-motif, pseudoknot  
cartoon-block innovative schematics in PyMOL, SQL-like feature queries in Jmol  
*in silico* base mutations, regular helical models, customized rebuilding

by Dr. Xiang-Jun Lu (律祥俊)

Department of Biological Sciences  
Columbia University, New York



- DSSR is licensed by Columbia University.
- DSSR is distributed in two variants:
  1. The DSSR basic covers features described in the three DSSR papers<sup>1–3</sup> to ensure that reported results can be *fully* reproduced. It is free for academic use.
  2. The DSSR Pro includes advanced capabilities for model building, additional features in structural analyses, annotations, and visualizations. It is for paid users only.
- This user manual documents the basic DSSR. The basic manual PDF is available free of charge.
- The Pro manual documents the full-featured DSSR. It is available to paid users only.
- A brief summary and background information of DSSR can be found in [the overview PDF](#).

# Contents

|  |   |          |
|--|---|----------|
| <b>1</b>   | <b>Introduction</b>   | <b>4</b> |
| <b>2</b>   | <b>Download and installation</b>  | <b>8</b> |
| <b>3</b>   | <b>Analysis and annotation</b>  | <b>8</b> |
| 3.1  | Command-line help . . . . .   | 8        |
| 3.2  | Default run on PDB entry 1msy: detailed explanations . . . . .                | 10       |
| 3.2.1  | Summary section . . . . .   | 11       |
| 3.2.2  | Base pairs . . . . .  | 13       |
| 3.2.3  | Multiplets (higher-order coplanar base associations) . . . . .                | 19       |
| 3.2.4  | Helices . . . . .   | 21       |
| 3.2.5  | Stems (canonical pairs with continuous backbones) . . . . .                   | 22       |
| 3.2.6  | Isolated canonical pairs . . . . .  | 23       |
| 3.2.7  | Base stacks . . . . .   | 23       |
| 3.2.8  | Atom-base capping interactions . . . . .                                      | 24       |
| 3.2.9  | Various loops . . . . .   | 24       |
| 3.2.10   | Single-stranded fragments . . . . .   | 26       |
| 3.2.11   | 2D structure in dot-bracket notation . . . . .                                | 26       |
| 3.2.12   | Structural features per nucleotide . . . . .                                  | 27       |
| 3.2.13   | Backbone torsion angles and suite names . . . . .                             | 28       |
| Main chain conformational parameters               | . . . . .   | 28       |
| Virtual torsion angles                             | . . . . .   | 29       |
| Sugar conformational parameters                    | . . . . .   | 30       |
| Assignment of sugar-phosphate backbone suite names | . . . . .   | 31       |
| 3.3  | Default run on PDB entry 1ehz (tRNA <sup>Phe</sup> ): summary notes . . . . . | 32       |
| 3.3.1  | Brief summary . . . . .   | 33       |
| 3.3.2  | Modified nucleotides . . . . .  | 33       |
| 3.3.3  | The four triplets . . . . .   | 34       |
| 3.3.4  | Relationship between helices and stems . . . . .                              | 34       |
| 3.3.5  | Three hairpin loops . . . . .   | 34       |
| 3.3.6  | One four-way junction loop . . . . .  | 35       |
| 3.3.7  | Pseudoknot . . . . .  | 35       |

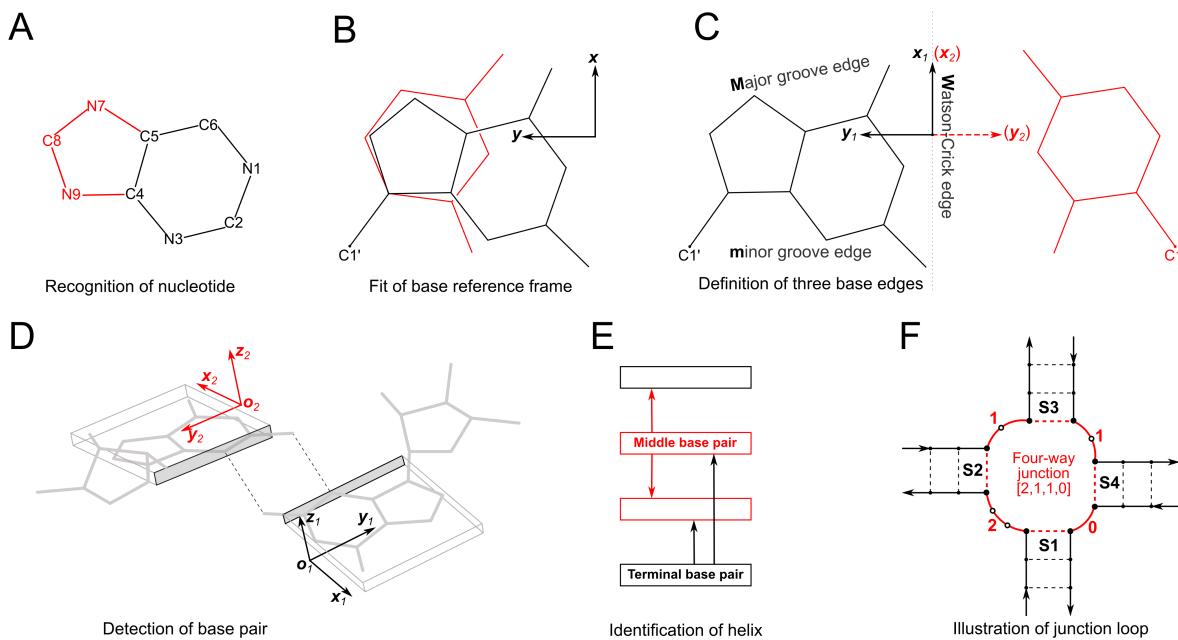
|          |   |           |
|----------|---|-----------|
| 3.4      | Default run on PDB entry 1jj2: four auto-checked motifs . . . . . | 36        |
| 3.4.1    | Kissing loops . . . . .   | 37        |
| 3.4.2    | A-minor motifs . . . . .  | 38        |
| 3.4.3    | Ribose zippers . . . . .  | 40        |
| 3.4.4    | Kink turns . . . . .  | 41        |
| 3.5      | Identification and characterization of G-quadruplexes . . . . .   | 42        |
| 3.6      | Pseudoknots detection and removal . . . . .                       | 43        |
| 3.6.1    | Higher-order pseudoknots . . . . .                                | 43        |
| 3.6.2    | Pseudoknot removal . . . . .                                      | 46        |
| 3.7      | The <code>--more</code> option . . . . .                          | 46        |
| 3.7.1    | Extra characterization of base pairs . . . . .                    | 47        |
| 3.7.2    | Orientation of helices/stems . . . . .                            | 48        |
| 3.7.3    | Base-pair morphology parameters for helices/stems . . . . .       | 49        |
| 3.8      | The <code>--non-pair</code> option . . . . .                      | 50        |
| 3.9      | The <code>--json</code> option . . . . .                          | 51        |
| 3.10     | The <code>--pair-only</code> option . . . . .                     | 52        |
| 3.11     | The <code>--nmr</code> option . . . . .                           | 53        |
| 3.12     | The <code>--get-hbond</code> option . . . . .                     | 54        |
| 3.13     | The <code>--idstr</code> option . . . . .                         | 56        |
| 3.14     | The <code>--symmetry</code> option . . . . .                      | 56        |
| 3.15     | The <code>--prefix</code> option . . . . .                        | 57        |
| <b>4</b> | <b>Visualization features</b> . . . . .                           | <b>57</b> |
| 4.1      | DSSR-Jmol integration . . . . .                                   | 57        |
| 4.2      | DSSR-PyMOL integration . . . . .                                  | 59        |
| <b>5</b> | <b>Modeling capabilities</b> . . . . .                            | <b>59</b> |
| <b>6</b> | <b>Frequently asked questions</b> . . . . .                       | <b>59</b> |
| 6.1      | What does DSSR stand for? . . . . .                               | 59        |
| 6.2      | How does DSSR compare with other tools? . . . . .                 | 59        |
| 6.3      | How is DSSR related to 3DNA? . . . . .                            | 60        |
| 6.4      | What are the differences between DSSR basic and pro? . . . . .    | 60        |
| <b>7</b> | <b>Acknowledgements</b> . . . . .                                 | <b>61</b> |

## List of Figures

|    |   |    |
|----|---|----|
| 1  | Definitions of key nucleic acid structural components in DSSR . . . . .             | 5  |
| 2  | Three- and two-dimensional structure images of PDB entry 1msy . . . . .             | 12 |
| 3  | Parallel M+N and antiparallel M–N orientations between two nucleobases .            | 14 |
| 4  | Lenotis-Westhof definition of three edges and the <i>cis/trans</i> orientations . . | 16 |
| 5  | Issues with the three edges in the Leontis-Westhof definition . . . . .             | 17 |
| 6  | Six types of G·A pairs involving the minor-groove edge of G . . . . .               | 19 |
| 7  | Pictorial definitions of rigid-body base morphology parameters . . . . .            | 20 |
| 8  | Base triplet (GUA) identified in PDB entry 1msy . . . . .                           | 21 |
| 9  | Common secondary structural elements of RNA . . . . .                               | 25 |
| 10 | Base pentaplet (AUUAG) identified in PDB entry 1jj2 . . . . .                       | 37 |
| 11 | Kissing-loop motif identified in PDB entry 1jj2 . . . . .                           | 39 |
| 12 | Types I and II A-minor motifs identified in PDB entry 1jj2 . . . . .                | 40 |
| 13 | Ribose zipper identified in PDB entry 1jj2 . . . . .                                | 41 |
| 14 | Normal k-turn identified in PDB entry 1jj2 . . . . .                                | 43 |
| 15 | Schematic representation of G-quadruplexes . . . . .                                | 44 |
| 16 | Topological descriptors and V-shaped loops of G-quadruplexes . . . . .              | 45 |
| 17 | Additional parameters calculated with the --more option . . . . .                   | 48 |
| 18 | DSSR-Jmol featured in the cover image of NAR . . . . .                              | 58 |

## 1 Introduction

As the number of experimentally solved RNA-containing structures grows, it is becoming increasingly important to characterize the geometric features of the molecules consistently and efficiently. Existing RNA bioinformatics tools are fragmented, and suffer in either scope or usability. DSSR<sup>1</sup> is an integrated software tool for **D**issecting the **S**patial **S**tructure of **R**NA, designed from ground up to streamline the analyses/annotation of 3D RNA structures. Figure 1 outlines some key algorithms underlying the DSSR program.



**Figure 1:** Definitions of key nucleic acid structural components in DSSR (Reproduced from Figure 1 of the DSSR paper<sup>1</sup>). (A) Nucleotides are recognized using standard atom names and base planarity. This method works for both the standard bases (A, C, G, T and U), and those of modified nucleotides, regardless of their tautomeric or protonation states. (B) Bases are assigned a standard reference frame<sup>4</sup> that is independent of sequence identity: purines and pyrimidines are symmetrically placed with respect to the sugar. (C) The standard base frame is derived from an idealized Watson-Crick base pair, and defines three base edges (Watson-Crick, minor groove, and Major groove) that are used to classify pairing interactions. (D) Base pairs are identified from the coplanarity of base rings and the occurrence of hydrogen bonds. This geometric algorithm can find canonical (Watson-Crick and G-U wobble) as well as non-canonical pairs. Higher-order (three or more) coplanar base associations, termed multiplets, are also detected. (E) Helices are defined by stacking interactions of base pairs, regardless of pairing type (canonical or otherwise) or backbone connectivity (covalently connected or broken). A helix consists of at least two base pairs. The same algorithm is applied to identify continuous base stacks that are outside of helical regions, by using bases instead of pairs as the assembly unit. Nucleotides not involved in base-stacking interactions are collected into one separate group. A stem is defined as a special type of helix, made up of canonical pairs and with a continuous backbone along each strand. Coaxial stacking is defined by the presence of two or more stems within one helix. An isolated canonical pair is one that is not contained within a stem. (F) ‘Closed’ loops of various types (hairpin, bulge, internal, and junction loops) are delineated by stems or isolated pairs, and specified by the lengths of the intervening, consecutive nucleotide segments.

Starting from an RNA structure in Protein Data Bank (PDB) or PDBx/mmCIF format, DSSR uses standard atom names and base planarity to detect nucleotides, including modified ones (Figure 1A). It employs the standard base reference frame (Figure 1B, C)<sup>4</sup> and a set of simple geometric criteria (Figure 1D) to identify all existent base pairs (bp): either canonical Watson-Crick (WC) and wobble pairs, or non-canonical pairs with at least one hydrogen bond (H-bond). The latter pairs may include normal or modified bases, regardless of tautomeric or protonation state. DSSR uses the six standard rigid-body bp parameters (shear, stretch, stagger, propeller, buckle, and opening) to rigorously quantify the spatial disposition of any two interacting bases. Where applicable, the program also denotes a bp by common names (e.g., WC, reverse WC, Hoogsteen A+U, reverse Hoogsteen A-U, wobble G-U, sheared G-A, imino G-A, Calcutta U-U, dinucleotide platform), the Saenger classification scheme<sup>5</sup> of 28 H-bonding types, and the Leontis-Westhof nomenclature<sup>6</sup> of 12 basic geometric classes.

DSSR detects multiplets (triplets or higher-order base associations) by searching horizontally in the plane of the associated bp for further H-bonding interactions. The program determines double-helical regions (Figure 1E) by exploring vertically in the neighborhood of selected bps for base-stacking interactions, regardless of backbone connection (e.g., coaxial stacking of helices). DSSR then identifies hairpin loops, bulges, internal loops, and multi-branch (junction) loops (Figure 1F). The program outputs RNA secondary structure in three commonly used formats—dot-bracket notation (dbn), connectivity table (.ct), and bp sequence (.bpseq)—that can be fed directly into visualization tools such as VARNA<sup>7</sup>. DSSR derives proper dbn for RNA with higher-order pseudoknots, and it can also produce pseudoknot-free secondary structures.

In DSSR, each helix/stem is characterized by a least-squares fitted helical axis, and dinucleotide steps are classified into the most common A-, B- or Z-form double helices (where appropriate) and quantified by helical parameters. DSSR calculates commonly used backbone torsion angles (including the virtual  $\eta/\theta$  torsions), classifies the backbone into BI/BII conformations and the sugar into C2'/C3'-endo-like puckers, and assigns the consensus RNA backbone suite names<sup>8</sup>. The program also identifies A-minor interactions, splayed-apart dinucleotide conformations, atom-base capping interactions, ribose zippers, G quadruplexes, i-motifs, kissing loops, U-turns, and k-turns, etc. Furthermore, DSSR reports non-pairing interactions (H-bonding or base-stacking) between two nucleotides, and contacts involving phosphate groups.

DSSR has been integrated to Jmol, a widely used Java/JavaScript viewer for 3D structures. The DSSR-Jmol integration<sup>2</sup> bridges the DSSR command-line analyzing tool and the Jmol

molecular viewer together via a simple [JSON](#) interface and a powerful query language. Users can now select DSSR-derived RNA structural features (such as base pairs, double helices, and various loops) as easily as they can select protein  $\alpha$ -helices and  $\beta$ -strands. Moreover, fine-grained characteristics of these features can be queried via Jmol SQL for DSSR (see the [DSSR-Jmol manual](#)). Notably, the novel representation styles (step diagram and base blocks) and coloring schemes bring RNA visualization to an entirely new level (see [Figure 3 of the DSSR-Jmol paper](#)).

DSSR supersedes the 3DNA suite of programs<sup>9-11</sup> for the analysis, rebuilding, and visualization of 3D nucleic acid structures. It has been created from scratch by employing my extensive experience in supporting 3DNA, increased knowledge of RNA structures, and improved C programming skills. The software has a set of unique features not available elsewhere (to the best of my knowledge). By connecting dots in RNA structural bioinformatics, it makes many common tasks simple and advanced applications feasible. DSSR is efficient and robust, with sensible default settings and an intuitive output; thus it is accessible to a much broader audience than the classic 3DNA distribution (up to v2.4).

There is actually more to DSSR than meets the eye. This basic version does not contain modeling features, or advanced variations of documented options herein. Please support DSSR by obtaining a paid [license from Columbia University](#) for the DSSR Pro version and the corresponding User Manual.

DSSR is being actively maintained and developed. As always, I greatly appreciate user feedback. So far, all reported bugs have always been promptly responded and fixed where appropriate. The program has been checked using all nucleic acid-containing entries in the PDB, without any known issues. Simply put, I strive to make DSSR a practical tool that the user community can count on.

When reporting bugs or submitting feature requests, please provide the specific DSSR version you are using (with `-v` or `--version`). For example, the version that the manual is based upon corresponds to (`x3dna-dssr -v`):

```
DSSR: an Integrated Software Tool for
Dissecting the Spatial Structure of RNA
v2.2.1-2021jan12 by xiangjun@x3dna.org
```

```
As of version 2, DSSR may be LICENSED from Columbia University.
DSSR basic is FREE for academic uses, with ABSOLUTELY NO WARRANTY.
DSSR Pro is available for paid academic or commercial users only,
and it is actively maintained and continuously improved.
```

```
This is DSSR *basic*. Advanced features are available in DSSR Pro.
```

## 2 Download and installation

As of version 2.0, DSSR has been licensed by Columbia University for all uses. Please visit the [Columbia Technology Ventures \(CTV\) website on DSSR](#) for details.

The basic DSSR software remains free for academic uses, and it comes with ABSOLUTELY NO WARRANTY. For long-term stability of the project, the DSSR Pro and its corresponding User Manual (in PDF) are available for *paid users* only. The DSSR Pro manual documents (new and enhanced) features not available elsewhere, including this basic manual.

DSSR is stable in terms of basic functionality and main output format. Where possible, I will try to maintain backward compatibility in future releases. If you only need fundamental features, you may not need to bother with frequent updates. DSSR is designed with simplicity and robustness in mind: get the job done, and then stay out of the way. On the other hand, DSSR is being actively maintained and developed, with new features added and known bugs fixed. Users are advised to keep up to date: it is as simple as downloading DSSR again from the [CTV website](#) to replace (i.e., overwrite) the old copy.

## 3 Analysis and annotation

Although the above introduction summarizes DSSR's major functionality in dry, 'abstract' technical terms, using DSSR effectively is very straightforward. This is best illustrated with concrete examples.

### 3.1 Command-line help

As is the norm of Linux/Unix command-line tools, running DSSR with -h (or --help) provides help information, as shown below (x3dna-dssr -h):

```
Usage: x3dna-dssr [options]

Each option is specified via --key[=val] (or -key[=val] or key[=val];
i.e., two/one/zero preceding dashes are all accepted), where 'key' can
be in either lower, UPPER or MiXed case. Options can be in any order.

Options:
  --help          Print this command-line help information (-h)
  --version       Print version number and exit (-v)
  --citation      Print preferred citation(s) and exit
  --input=file    Specify a PDB/mmCIF file for analysis (-i=file)
  --output=file   Designate the main DSSR output file (-o=file)
  --more          Report detailed bp and step/helical parameters
  --non-pair      Check non-pairing H-bonds/stacking interactions
  --pair-only     Output just base-pairing information [10x faster]
```

```
--json           Generate output in JSON format for easy parsing
--nmr            Process an ensemble of NMR structures
--blocview       Generate cartoon/blocks, in extended view, in PyMOL
--frame          Reorient a structure with specified base/pair frame
--get-hbond      Report a list of H-bonds within nucleic acids

Examples:
x3dna-dssr -i=1msy.pdb
x3dna-dssr -i=1msy.cif --more --non-pair
x3dna-dssr -i=1ehz.pdb -o=1ehz.out # yeast phenylalanine tRNA
x3dna-dssr -i=1jj2.pdb -o=1jj2.out # large ribosomal subunit
x3dna-dssr -i=5afii.cif -o=5afii.out # ribosome (only in .cif)
x3dna-dssr -i=5afii.cif -o=5afii-pairs.out --PAIR-ONLY # 10x faster
x3dna-dssr -i=1msy.pdb --json -o=1msy.json
# x3dna-dssr -i=1msy.pdb --json | jq .pairs
x3dna-dssr -i=5afii.cif --json --pair-only
x3dna-dssr -i=2n2d.pdb --nmr --json -o=2n2d-dssr.json
x3dna-dssr -i=1msy.pdb --blocview --block-file=wc -o=1msy.pml
# load '1msy.pml' into PyMOL for interactive visualization
x3dna-dssr -i=1msy.pdb --frame='A.2658-wc-minor' -o=1msy-bp.pdb
# set 1msy into the minor-groove view of the C2658-G2663 pair
x3dna-dssr -i=1oct.pdb --get-hbond # H-bonds in DNA (default)
x3dna-dssr -i=1oct.pdb --get-hbond-protein # plus H-bonds in proteins
x3dna-dssr snap -h # help for the SNAP module on the interactions
#   of DNA-protein or RNA-protein complexes
```

The above help message should be sufficient to get most users started using DSSR. It is worth noting that the command-line interface is consistent on all operating systems, including the command shell on Windows.

DSSR introduces a consistent and flexible way to process command-line options. Here, each option can be specified via a `--key[=value]` pair, or `-key[=value]` or `key[=value]`; i.e., two/one/zero preceding dashes are all acceptable. The key can be in lower, UPPER or MiXed case, and the value is optional for Boolean switches. Moreover, options can be put in any order; if the same key is repeated more than once, the value specified with the last key prevails. Some typical use cases are given below:

```
#1 analyze the PDB entry '1msy', with default output to stdout
x3dna-dssr --input=1msy.pdb

#2 same as #1, with output directed to file '1msy.out'
x3dna-dssr --input=1msy.pdb --output=1msy.out

#3-6, same as #2
x3dna-dssr --output=1msy.out --input=1msy.pdb
x3dna-dssr --OUTPUT=1msy.out --Input=1msy.pdb
x3dna-dssr -output=1msy.out input=1msy.pdb
x3dna-dssr output=1msy.out --input=1msy.pdb

#7 the value '1ehz.pdb' overwrites '1msy.pdb'
x3dna-dssr --input=1msy.pdb input=1ehz.pdb

#8-12 with the switch --more set to true
x3dna-dssr -input=1msy.pdb --more
x3dna-dssr -input=1msy.pdb --more=true
```

```
x3dna-dssr -input=1msy.pdb --more=yes
x3dna-dssr -input=1msy.pdb --more=on
x3dna-dssr -input=1msy.pdb --more=1

#13 same as without specifying --more,
#      or with values set to false/no/0
x3dna-dssr -input=1msy.pdb --more=off

#14 shorthand forms for --input and --output
x3dna-dssr -i=1msy.pdb -o=1msy.out

#15 it can also be more verbose
x3dna-dssr --input-pdb-file=1msy.pdb

#16-19 within a key, separator dash(-) and underscore (_) 
#      are treated the same, and can be omitted
x3dna-dssr -i=1msy.pdb -non-pair
x3dna-dssr -i=1msy.pdb -non_pair
x3dna-dssr -i=1msy.pdb -nonpair
x3dna-dssr -i=1msy.pdb -nonPair
```

By allowing for two/one/zero dashes to precede each key and a dash/underscore/none to separate words within a key, DSSR provides great versatility in specifying command-line options to fit into a user-preferred style.

### 3.2 Default run on PDB entry 1msy: detailed explanations

The PDB entry 1msy is a small RNA fragment with 27 nucleotides, containing a GUAA tetraloop (GNRA-type) mutant of the sarcin/ricin domain from *E. Coli* 23S rRNA<sup>12</sup> that includes a bulged-G motif (the GpU dinucleotide platform<sup>13</sup>). This structure contains several features that nicely illustrate fundamental aspects of DSSR.

Let the PDB-formatted 3D coordinate file be called 1msy.pdb (as shown below, the PDBx/mmCIF version 1msy.cif works as well), the output file 1msy.out, and leave all other options in their default settings. Then simply run the command:

```
x3dna-dssr -i=1msy.pdb -o=1msy.out
x3dna-dssr -i=1msy.cif -o=1msy.out # gives the same results as above
```

The screen output provides a brief summary of the run, as shown below. Note that DSSR starts from a raw .pdb or .cif file as directly downloaded from the PDB, and takes little time to analyze small to mid-size (non-ribosomal) structures.

```
total number of nucleotides: 27
total number of base pairs: 13
total number of multiplets: 1
total number of helices: 1
total number of stems: 1
```

```

total number of isolated WC/wobble pairs: 1
total number of atom-base capping interactions: 2
total number of splayed-apart dinucleotides: 1
total number of hairpin loops: 1
total number of internal loops: 1
total number of non-loop single-stranded segments: 2

Time used: 00:00:00:00

```

For easy reference, [Figure 2](#) shows the 3D structure and the corresponding secondary (2D) structure of 1msy. The schematic 3D structure was automatically generated in schematic `blocview` representation. The 2D diagram was rendered with VARNA<sup>7</sup>, using DSSR-generated 2D structure in connectivity table format (.ct, see below).

The main output file (`1msy.out`) contains many sections. We will go over them one by one, along the way, explaining the notations used therein. Additionally, DSSR generates the following auxiliary files (named with prefix `dssr-` by default):

```

List of 13 additional files
1 dssr-pairs.pdb -- an ensemble of base pairs
2 dssr-multiplets.pdb -- an ensemble of multiplets
3 dssr-stems.pdb -- an ensemble of stems
4 dssr-helices.pdb -- an ensemble of helices (coaxial stacking)
5 dssr-hairpins.pdb -- an ensemble of hairpin loops
6 dssr-iloops.pdb -- an ensemble of internal loops
7 dssr-2ndstrs.bpseq -- secondary structure in bpseq format
8 dssr-2ndstrs.ct -- secondary structure in connectivity table format
9 dssr-2ndstrs.dbn -- secondary structure in dot-bracket notation
10 dssr-torsions.txt -- backbone torsion angles and suite names
11 dssr-splays.pdb -- an ensemble of splayed-apart units
12 dssr-stacks.pdb -- an ensemble of base stacks
13 dssr-atom2bases.pdb -- an ensemble of atom-base stacking interactions

```

### 3.2.1 Summary section

Note: By default, each nucleotide is identified by `chainId.name#`. So a common case would be `B.A1689`, meaning adenosine #1689 on chain B. One-letter base names for modified nucleotides are put in lower case (e.g., 'c' for 5MC). For further information about the output notation, please refer to the DSSR User Manual.

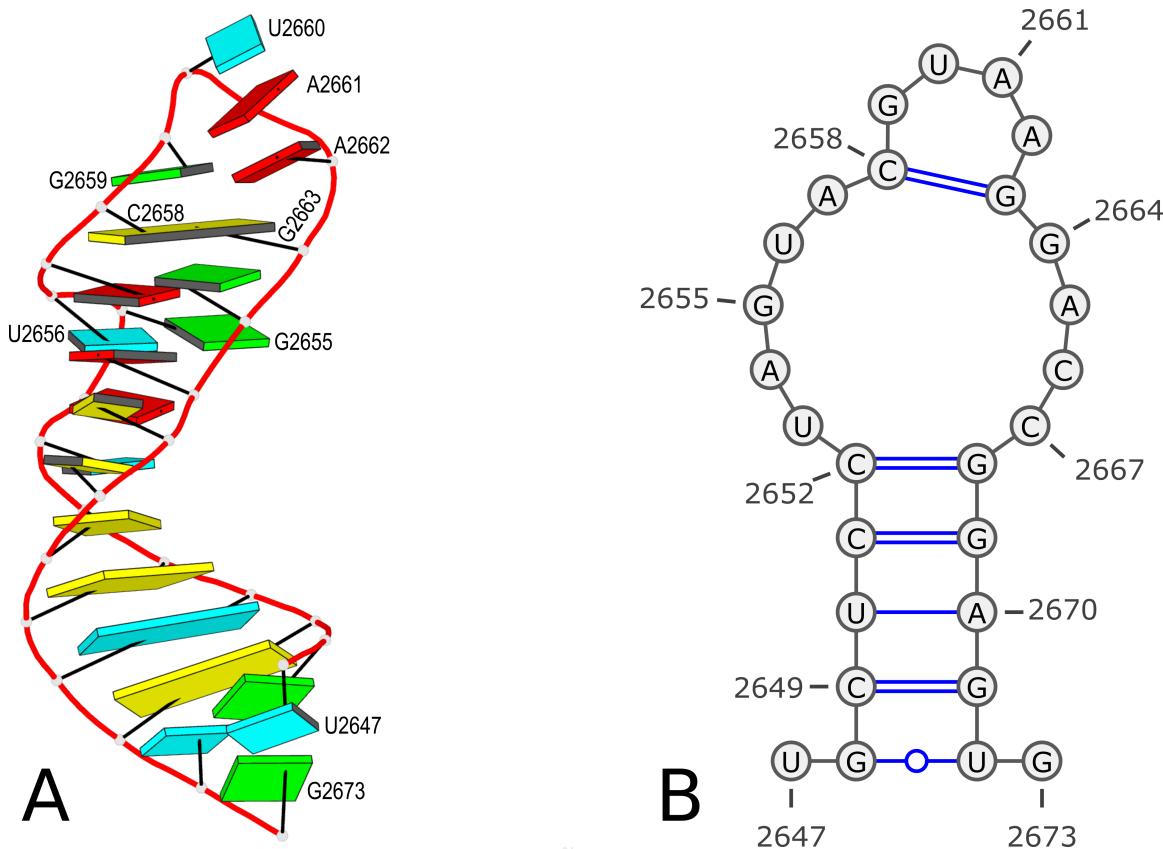
Questions and suggestions are \*always\* welcome on the 3DNA Forum.

```

Command: x3dna-dssr -i=1msy.pdb -o=1msy.out
File name: 1msy.pdb
no. of DNA/RNA chains: 1 [A=27]
no. of nucleotides: 27
no. of atoms: 685
no. of waters: 109
no. of metals: 0

```

The note in the summary section first explains how a nucleotide (nt) is specified in



**Figure 2:** Images of PDB entry 1msy. (A) Schematic 3D structure in block presentation<sup>3</sup>. (B) 2D diagram produced with VARNA<sup>7</sup> using DSSR-derived secondary structure information in .ct format.

the output file. Typically, a chain id, residue name, and sequence number are sufficient to unambiguously specify a nt. For example, A.G19 means guanosine no. 19 on chain A. In addition to the standard three-letter residue names commonly adopted (as in the PDB), DSSR also uses a one-letter shorthand symbol. For RNA, the four standard nts in three-letter form—LUU A, LUU C, LUU G and LUU U (where L stands for a space)—are shortened to A, C, G, and U, respectively. For DNA, the three-letter (one-letter) nts are LDA (A), LDC (C), LDG (G), and LDT (T), respectively. The one-letter shorthand forms for modified nts, which occur frequently in RNA (e.g., tRNA), are mapped to their canonical counterparts, but put in lower case letters (e.g., c for 5MC). Note that **pseudouridine**, the most prevalent modified nt in RNA, is denoted P in DSSR and the small case p is reserved for potential modified pseudouridines.

The following sections provide more information on notations that are used consistently in DSSR output files. It is worth the time and effort to become familiar with them. In practice,

a normal user should have little difficulty following the convention, after going over a few examples.

The remaining lines in the summary section should be self-explanatory. The **Command:** line provides the running DSSR command with relevant options so that the results can be reproduced. The meaning of the **Date and time:** line is obvious. The **File name:** line lists the data file analyzed by DSSR, followed by a list of numbers: DNA/RNA chains (here chain A with 27 nts), nts (27), atoms (685), waters (109), and metals (0). In cases with more than one chain or type of metal, the output would be as below (using 1jj2 as an example).

```
no. of DNA/RNA chains: 2 [0=2754,9=122]
no. of metals: 210 [Na=86,Mg=117,K=2,Cd=5]
```

Here the first line means that 1jj2 contains two RNA chains: 0 and 9, with 2754 and 122 nts, respectively. The second line indicates the PDB entry includes 210 metal atoms: 86 Na, 117 Mg, 2 K, and 5 Cd.

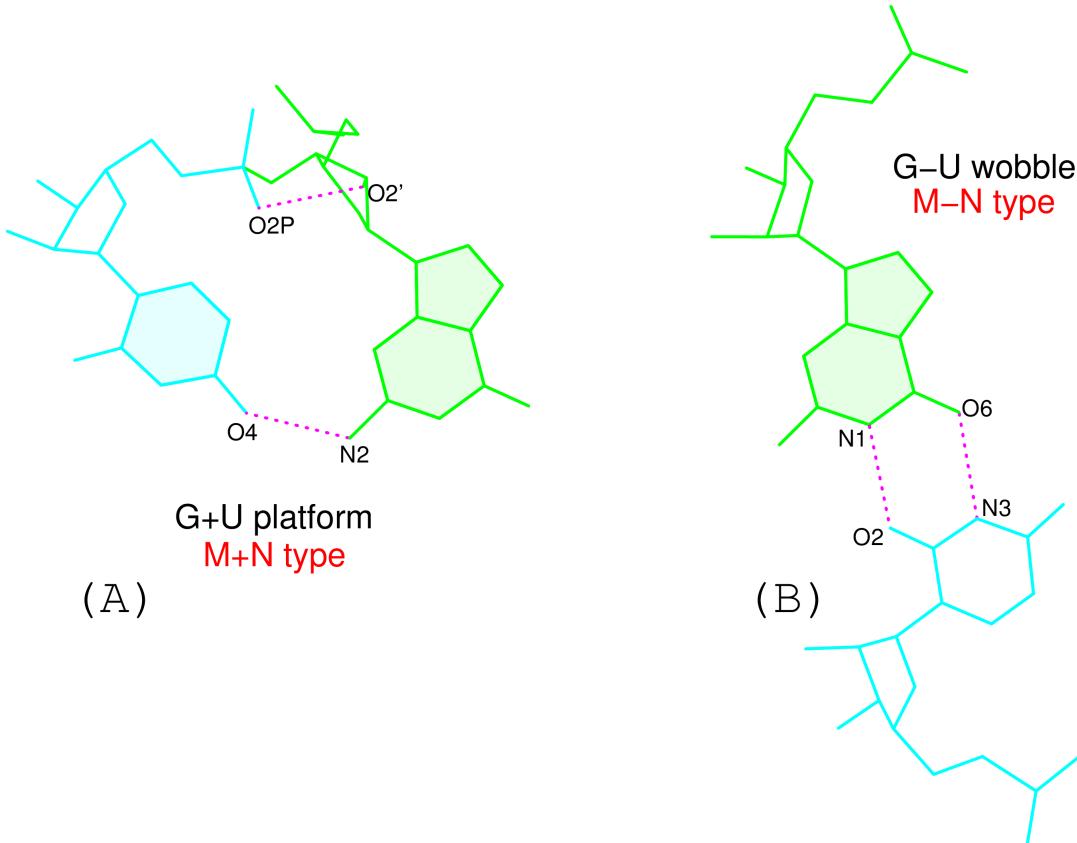
### 3.2.2 Base pairs

DSSR identifies a total of 13 bps, including not only the canonical Watson-Crick (WC) and wobble pairs, but also non-canonical bps (a dinucleotide platform, reverse Hoogsteen, sheared G–A, etc.).

| List of 13 base pairs |         |         |     |            |           |     |      |
|-----------------------|---------|---------|-----|------------|-----------|-----|------|
|                       | nt1     | nt2     | bp  | name       | Saenger   | LW  | DSSR |
| 1                     | A.U2647 | A.G2673 | U-G | --         | --        | cWW | cW-W |
| 2                     | A.G2648 | A.U2672 | G-U | Wobble     | 28-XXVIII | cWW | cW-W |
| 3                     | A.C2649 | A.G2671 | C-G | WC         | 19-XIX    | cWW | cW-W |
| 4                     | A.U2650 | A.A2670 | U-A | WC         | 20-XX     | cWW | cW-W |
| 5                     | A.C2651 | A.G2669 | C-G | WC         | 19-XIX    | cWW | cW-W |
| 6                     | A.C2652 | A.G2668 | C-G | WC         | 19-XIX    | cWW | cW-W |
| 7                     | A.U2653 | A.C2667 | U-C | --         | --        | tW. | tW-. |
| 8                     | A.A2654 | A.C2666 | A+C | --         | --        | tHH | tM+M |
| 9                     | A.G2655 | A.U2656 | G+U | Platform   | --        | cSH | cm+M |
| 10                    | A.U2656 | A.A2665 | U-A | rHoogsteen | 24-XXIV   | tWH | tW-M |
| 11                    | A.A2657 | A.G2664 | A-G | Sheared    | 11-XI     | tHS | tM-m |
| 12                    | A.C2658 | A.G2663 | C-G | WC         | 19-XIX    | cWW | cW-W |
| 13                    | A.G2659 | A.A2662 | G-A | Sheared    | 11-XI     | tSH | tm-M |

As the header line (the 2nd in the listing) shows, each bp is characterized by the two constituent nts (nt1 and nt2), and an abbreviated bp type of the form M±N. Here M and N are the one-letter shorthand symbol of the nts, which can be in upper or lower cases (for modified nts), as noted above. The symbol ± reflects the two possible relative ‘face’ orientations between the two planar base moieties: the sign is normally – as for canonical

WC/wobble bps with opposite faces, and + if one of the two base planes is *flipped*. Figure 3 provides further details on the rationale of the M±N convention<sup>13</sup> that is used consistently in 3DNA<sup>9,10</sup> and DSSR<sup>1</sup>.



**Figure 3:** Comparison of the parallel M+N and antiparallel M–N arrangements of G and U bases found in (A) the G+U platform and (B) the G–U wobble pair, respectively. Here G is depicted in green, and U in cyan. For easy comparison, the two pairs are oriented in the standard reference frame of the guanine base, with the shading used to specify the face from which the positive base normal, i.e., z-axis, emanates. The U shares the same face as G in the G+U dinucleotide platform (A); the z-axes of the two bases are parallel, forming a positive dot product expected of M+N pairing. By contrast, the U points in the opposite direction from G in the G–U wobble pair (B); the z-axes of the two bases are anti-parallel, forming a negative dot product expected of M–N pairing. Note that this M–N nomenclature follows the convention expected for the canonical, antiparallel Watson-Crick base pairs: A–U/A–T and G–C.

The name column gives the common names of the corresponding bps, where appropriate. Currently, the list includes WC, reverse WC (rWC), Hoogsteen, reverse Hoogsteen (rHoog-

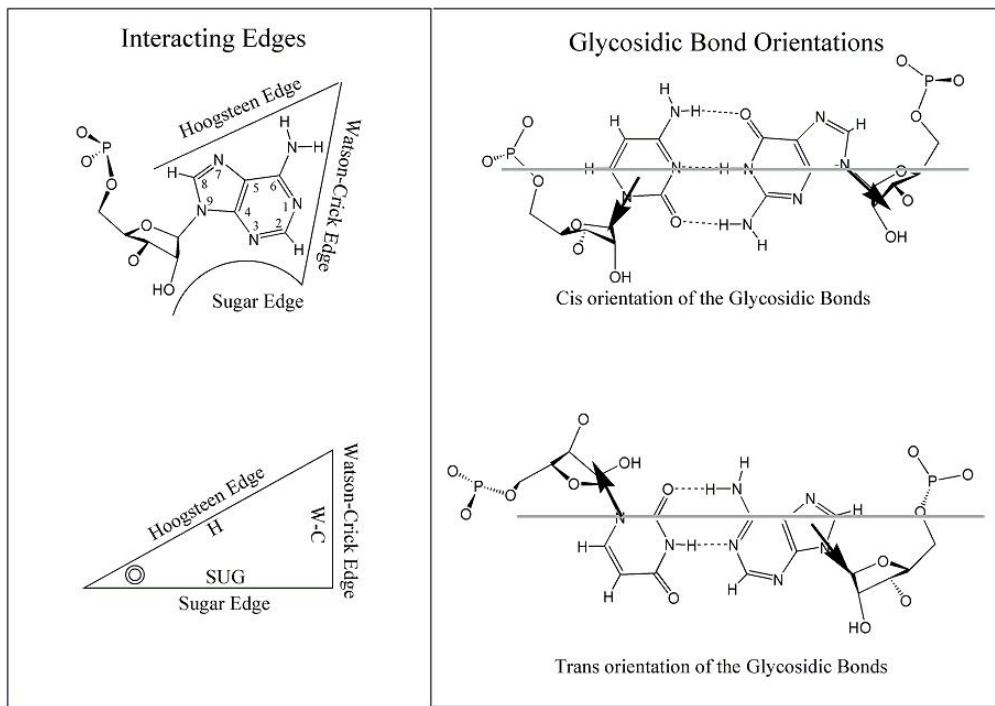
steen), wobble, sheared, imino, Calcutta, and dinucleotide platform, and unnamed bps are designated --. Other than WC and wobble pairs, the PDB entry 1msy also contains a GpU-pA/GpA miniduplex<sup>13</sup> characterized by three common bps (nos. 9 to 11): a platform (G+U), a reverse Hoogsteen (U–A), and a sheared (A–G). However, it is difficult to connect those bp structures with their familiar names, even for experts. With DSSR-assigned names, one can immediately see common bps and their types within a structure. As it turns out, the name column is also handy to quickly delineate stretches of canonical pairs (e.g., nos 2 to 6 in the listing on [Page 13](#)) that form double-helical regions, termed ‘stem’ in DSSR (see [Page 22](#)).

The Saenger column provides the classification of 29 bp types with at least two H-bonds between base atoms. This list, [initially compiled by Saenger for 28 bps](#)<sup>5</sup>, includes an extra G+C bp added to the list [by Burkard et al.](#) in an appendix to the second edition of “The RNA World” book<sup>14</sup>. Note that in addition to the Roman numerals (I to XXIX) originally used by Saenger, the DSSR output also includes the corresponding Arabic numerals (01 to 29), which may be easier for some users to recognize (at least for myself). If a pair cannot be categorized into one of the 29 known types, the symbol -- is assigned.

The LW column in the list of base pairs (on [Page 13](#)) gives Leontis-Westhof classification of bps<sup>6</sup>. As illustrated in [Figure 4](#) (see below), this classification is based on the three edges of each base with potential for H-bonding interactions (Watson-Crick, Hoogsteen, and Sugar), and the two orientations (*cis* or *trans*) of the interacting bases with respect to the glycosidic bonds. The combinations of edges and orientations “gives rise to 12 basic geometric types with at least two H bonds connecting the bases”<sup>6</sup>. This geometry-based method captures salient features of bp interactions and strikes a balance between simplicity and expressiveness. The LW scheme is more generally applicable than the Saenger classification, and it is easy to grasp by biologists. As a result, the LW bp classification has become standard in RNA structural bioinformatics.

Strictly speaking, however, the RNA-centric LW classification has its limitations ([Figure 5](#)). For one thing, the Sugar edge explicitly includes the 2'-hydroxyl group, rendering it less applicable to DNA structures. Moreover, while the aromatic base can be taken as a rigid body with three fixed edges, the  $\chi$  (chi) torsion angle characterizes [the internal freedom between base and sugar \(anti/syn\)](#). When  $\chi$  is in the relatively rare but by no means uncommon *syn* conformation, the Sugar edge, defined by the common *anti* conformation, no longer seems to exist.

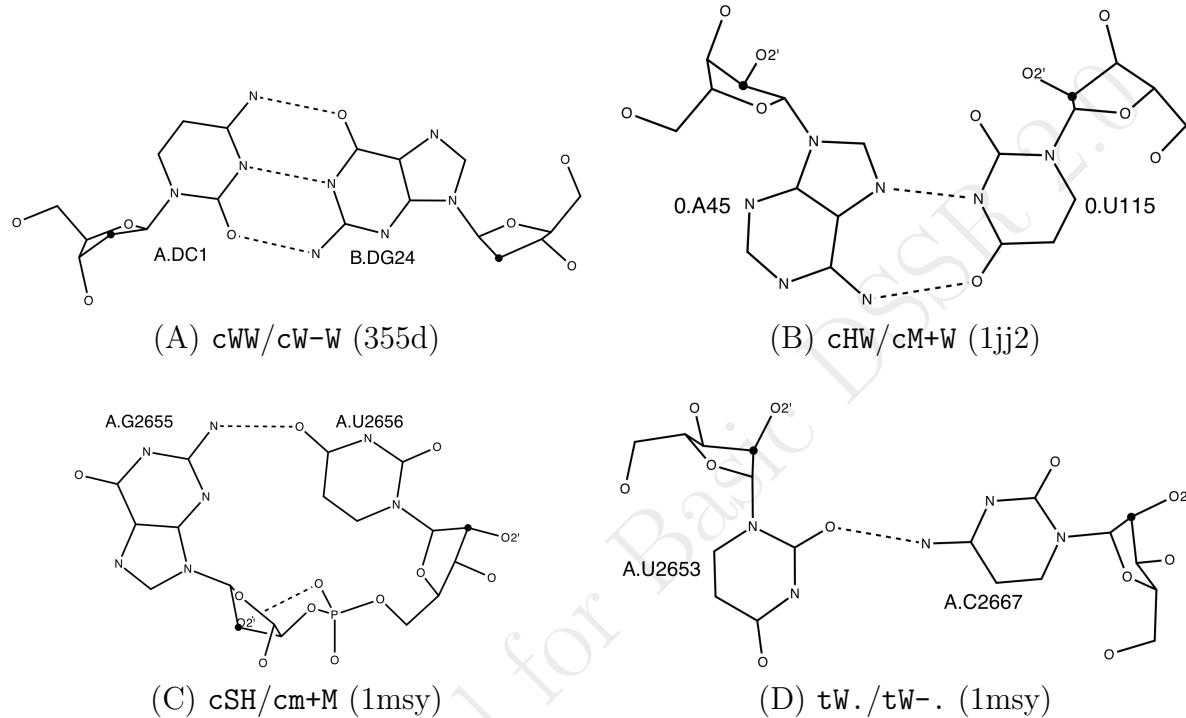
The rich variety of RNA bps extends beyond the 12 basic LW types. There are numerous pairs in RNA with only one H-bond or with bifurcated H-bonds, at boundary locations where



**Figure 4:** Base edges and base-pair geometric isomerism. (Upper left) An adenine showing the three base edges (Watson-Crick, Hoogsteen, and Sugar-edge) available for H-bonding interactions. (Lower left) Representation of an RNA base as a triangle. The position of the ribose is indicated with a circle in the corner defined by the Hoogsteen and Sugar edges. (Right) *Cis* and *trans* base-pairing geometries, illustrated for two bases interacting with Watson-Crick edges<sup>6</sup>.

the LW classification does not strictly apply. Lemieux and Major<sup>15</sup> were the first to extend the LW nomenclature. We noted the importance of the out-of-plane ‘backbone edge’ formed by an RNA-specific H-bond between O2'(G) and OP2(U)<sup>13</sup>. The RNA 3D Hub website, hosted by the Leontis-Zirbel co-lab, lists non-standard bp interactions ncSW, ntSH, and ntHH for the 1msy entry.

The DSSR bp classification is presented in the last column (e.g., cW-W for canonical bps, and cm+M for the G+U dinucleotide platform). Overall, the DSSR scheme follows the pattern of [ct] [WMm] ± [WMm] (see Figure 1C). Here, [ct] stands for *cis/trans*, defined by the positioning of the glycosidic bonds related to a line connecting the centers of base-ring atoms; [WMm] for the interacting edges defined with reference to an idealized WC bp: W for the Watson-Crick edge, M for the Major-groove edge, m for the minor-groove edge; ± for normal (–) vs. flipped (+) base orientations, as noted above. The dot symbol denotes cases



**Figure 5:** Sample cases where the Leontis-Westhof definition of the three edges does not strictly apply. The results of LW/DSSR classifications are those implemented in the DSSR program. In each image, H-bonds are shown as dashed lines, C2' atoms in black dots, and O2' atoms labeled. (A) A Watson-Crick C-G pair in 355d, a standard B-DNA molecule. The sugar-edge for C or G does not have an O2' atom. (B) The Hoogsteen A+U pair in 1jj2. Here, nucleotide 0.A45 is in the *syn* conformation; thus the O2' atom is pointing away from (instead of towards) the minor-groove edge of the base. (C) The G+U platform (in 1msy) has an out-of-plane 'backbone edge'. (D) The U-C pair (in 1msy) has only one boundary H-bond. Thus, the interacting edge for A.C2667 cannot be unambiguously assigned.

where edges or orientations cannot be defined (see bp no. 7 in the list of base pairs on [Page 13](#), and [Figure 5\(D\)](#)).

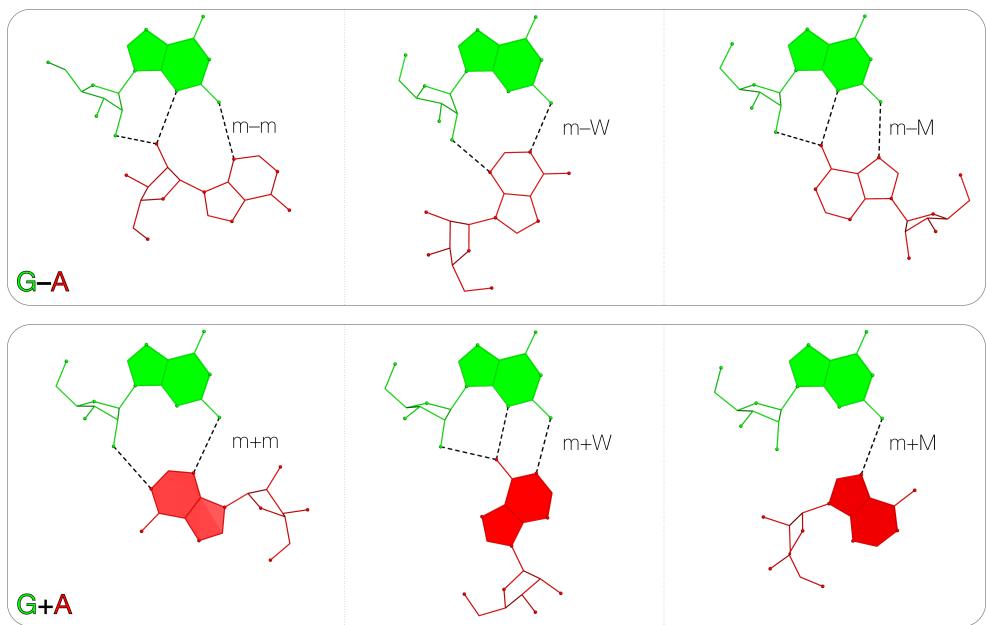
In general, the M in DSSR scheme corresponds to the Hoogsteen/CH-edge (H) in the LW notation, and the m to the LW Sugar-edge (S) if  $\chi$  is in the *anti* conformation. In the DSSR implementation, we assumes direct correspondences between M/H, and m/S, regardless of the *anti/syn* base-sugar orientations. Moreover, the LW *cis/trans* is assigned in the same way as for the DSSR scheme. So the LW and DSSR notations are strictly parallel, except for the extra  $\pm$  character in the DSSR classification.

The  $M \pm N$  bp notation has been used consistently in 3DNA for over a decade, and a whole section titled ‘Base pair parameters’ is included in the 2003 3DNA publication<sup>9</sup>. The orientation, combined with the six standard bp parameters (shear, stretch, stagger, buckle, propeller, and opening; see [Figure 7](#)) derived from the 3DNA `analyze` program, can unambiguously characterize any pair. Conversely, given the  $M \pm N$  notation with six corresponding bp parameters, the 3DNA `rebuild` program can rigorously reconstruct the spatial disposition of the two interacting bases. This reversibility is one of the unique features of 3DNA and can be applied to bps in both DNA and RNA.

Even with the increasing popularity of 3DNA, the value of using  $M \pm N$  with six parameters to classify bps has never received the attention it deserves, especially in the RNA structural bioinformatics community. The incorporation of bp classifications in DSSR provides an opportunity to emphasize the strength of this approach. In DSSR, the three interacting edges strictly center on a base, which can be taken as a *rigid* body. The standard base reference frame<sup>4</sup> has distinct geometric features to allow for easy identification of the edges ([Figure 1C](#)). As is also clear from [Figure 1C](#), for WC pairs, the distinction of the minor-groove vs. the major-groove edges is simply a consequence of the asymmetric glycosidic linkage between the base and sugar moiety. Moreover, the terminology of minor/major grooves is widely used, even in the RNA structure literature: the most obvious being the A-minor motif<sup>16</sup> (see [Section 3.4.2](#)).

Our recent paper “Effects of noncanonical base pairing on RNA folding: structural context and spatial arrangements of G·A pairs”<sup>17</sup> in the ACS *Biochemistry* journal highlights the advantages enabled by DSSR to uncover previously unrecognized patterns. The key features are summarized in the abstract (see [Figure 6](#)), and excerpted below:

“The rigorous descriptions of base-pair geometry that we employ facilitate characterization of recurrent geometric motifs and the structural settings in which these arrangements occur. Moreover, the numerical parameters hint at the natural



**Figure 6:** Six types of G·A pairs, uncovered and characterized by DSSR, that involve the minor-groove edge of G. Top row: three examples in the G·A category with types m-m, m-W, and m-M, respectively; Bottom row: three cases in the G+A category with types m+m, m+W, and m+M, respectively.

motions of the interacting bases and the pathways likely to connect different spatial forms. We draw attention to higher-order multiplexes involving two or more G·A pairs and the roles these associations appear to play in bridging different secondary structural units.”

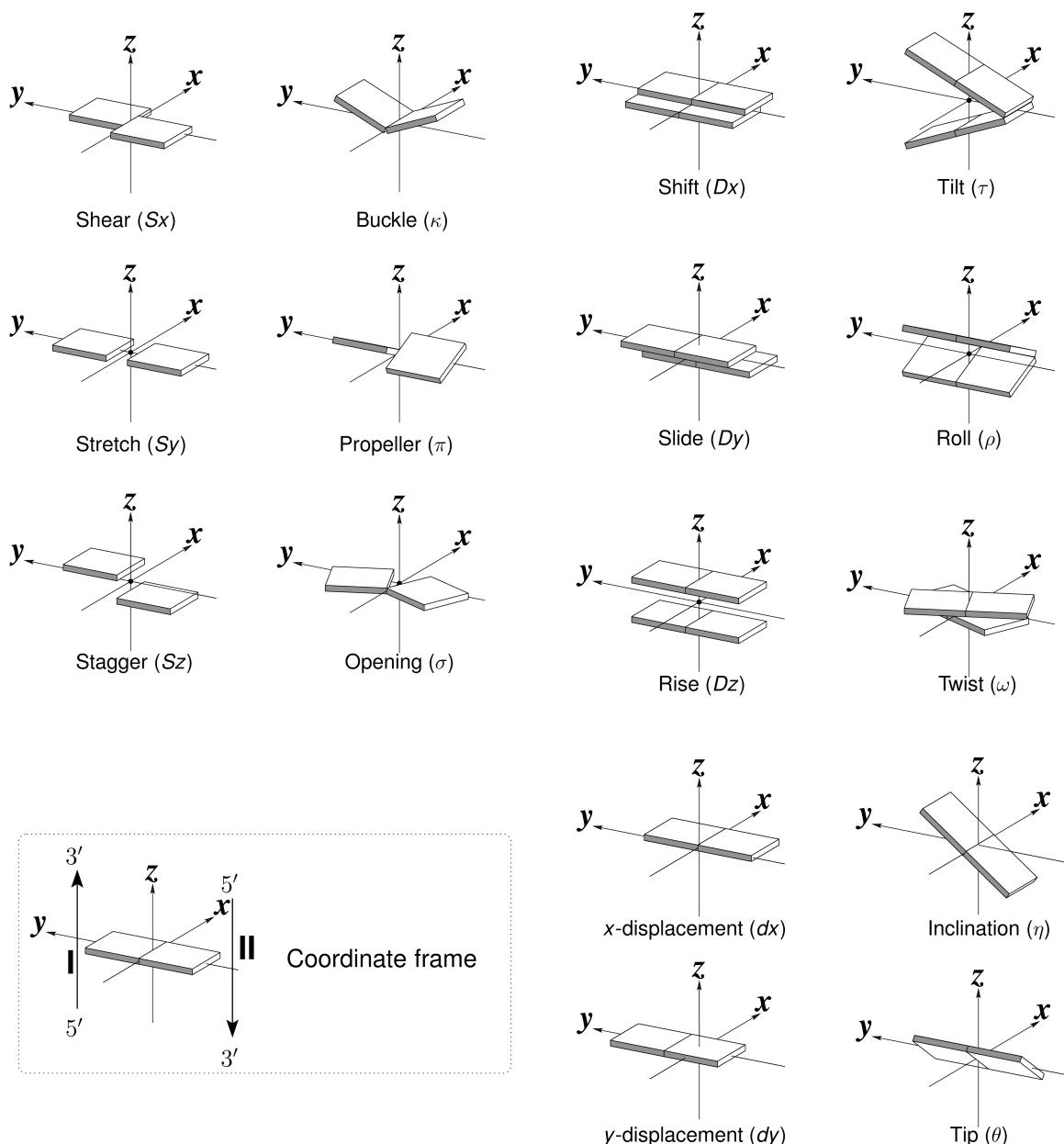
DSSR also produces a file named `dssr-pairs.pdb` that contains an ensemble of MOD-EL/ENDMDL delineated atomic coordinates for identified bps. Each bp is expressed in its own reference frame for easy and consistent visualization.

### 3.2.3 Multiplets (higher-order coplanar base associations)

There is only one multiplet, a triplet, in 1msy, as shown below.

```
List of 1 multiplet
1 nts=3 GUA A.G2655,A.U2656,A.A2665
```

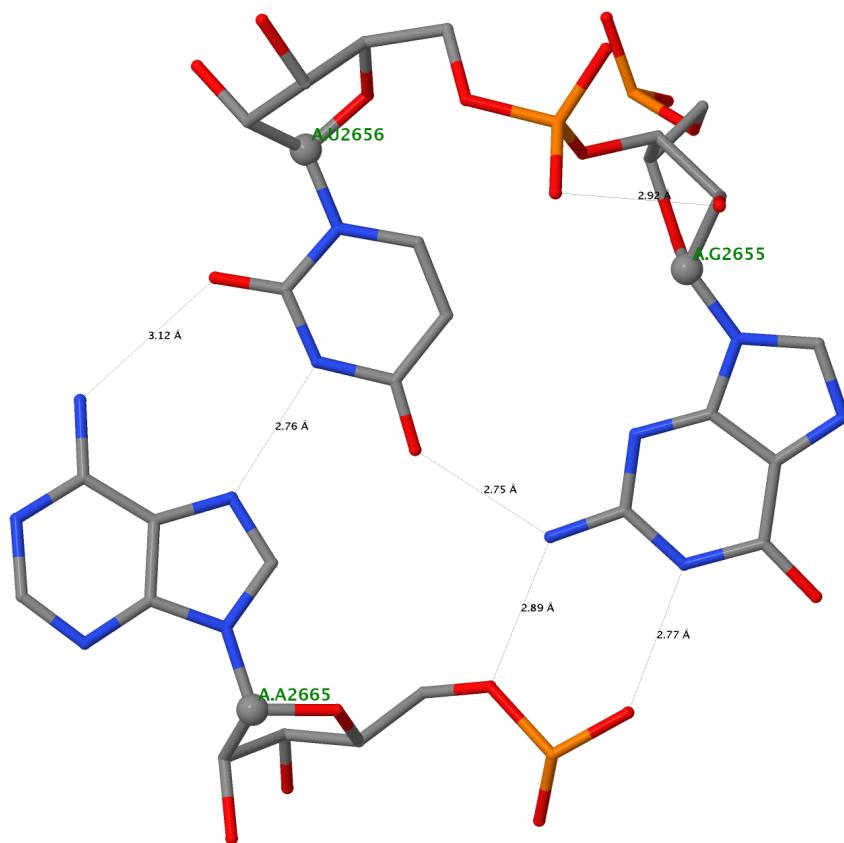
For each identified multiplet, the serial number (1) is followed by the number of nts (3), the base sequence in one-letter shorthand form (GUA), and a comma-delimited list of the



**Figure 7:** Pictorial definitions of rigid-body parameters used to describe the geometry of complementary (or non-complementary) base pairs and sequential base pair steps. The base-pair reference frame (lower left) is constructed such that the  $x$ -axis points away from the (shaded) minor groove edge of a base or base pair and the  $y$ -axis points toward the sequence strand (I). The relative position and orientation of successive base pair planes are described with respect to both a dimer reference frame (upper right) and a local helical frame (lower right). Images illustrate positive values of the designated parameters.

corresponding nts. The G-tetrad motif, which forms the **G-quadruplexes**, is also detected with DSSR (although not in 1msy); it is simply a special multiplet with four Gs.

As for bps, DSSR also generates a file (named `dssr-multiplets.pdb` by default) for MODEL/ENDMDL delineated multiplets. In the file, each multiplet is set to the most extended view with respect to the base rings. The triplet in 1msy is shown in [Figure 8](#). Note the extensive network of H-bonding interactions between the three nts<sup>13</sup>.



**Figure 8:** The GUA triplet identified in 1msy. Here G2655 and U2656 form a G+U dinucleotide platform; U2656 and A2665 form a reverse Hoogsteen pair. The phosphate group of A2665 interacts with the Watson-Crick edge of G2655 with two sequence-specific H-bonds. The image was produced with Jmol.

### 3.2.4 Helices

DSSR identifies one helix with 12 bps in 1msy (see below). This section starts with a note explaining the definition of a helix, and other related information. By referring to [Figure 2\(A\)](#),

one can immediately see the duplex formed by base-stacking interactions, starting from the very bottom of the structure all the way up to the sheared G–A pair at the top (part of the GUAA tetraloop). Note that by definition, a helix is composed of at least two stacked bps.

```
List of 1 helix
Note: a helix is defined by base-stacking interactions, regardless of bp
      type and backbone connectivity, and may contain more than one stem.
helix#number[stems-contained] bps=number-of-base-pairs in the helix
bp-type: '||' for a canonical WC/wobble pair, '.' otherwise
helix-form: classification of a dinucleotide step comprising the bp
      above the given designation and the bp that follows it. Types
      include 'A', 'B' or 'Z' for the common A-, B- and Z-form helices,
      '.' for an unclassified step, and 'x' for a step without a
      continuous backbone.

-----
helix#1[1] bps=12
strand-1 5'-UGCUCCUAUACG-3'
  bp-type .|||||.|.
strand-2 3'-GUGAGGCCAGGA-5'
  helix-form ..AAA..x...
1 A.U2647     A.G2673      U-G --          --      cWW  cW-W
2 A.G2648     A.U2672      G-U Wobble    28-XXVIII cWW  cW-W
3 A.C2649     A.G2671      C-G WC        19-XIX   cWW  cW-W
4 A.U2650     A.A2670      U-A WC        20-XX    cWW  cW-W
5 A.C2651     A.G2669      C-G WC        19-XIX   cWW  cW-W
6 A.C2652     A.G2668      C-G WC        19-XIX   cWW  cW-W
7 A.U2653     A.C2667      U-C --          --      tW.  tW-
8 A.A2654     A.C2666      A+C --          --      tHH  tM+M
9 A.U2656     A.A2665      U-A rHoogsteen 24-XXIV  tWH  tW-M
10 A.A2657    A.G2664      A-G Sheared  11-XI    tHS  tM-m
11 A.C2658    A.G2663      C-G WC        19-XIX   cWW  cW-W
12 A.G2659    A.A2662      G-A Sheared  11-XI    tSH  tm-M
```

Of the 13 bps listed on [Page 13](#), only the G+U dinucleotide platform formed by G2655 and U2656 is not contained in the helix. Since U2656 also forms a reverse Hoogsteen U–A pair with A2665 (as part of the triplet, see [Figure 8](#)), and the U–A bp is included in the helix, only the so-called bulged G (i.e., G2655) is excluded.

The helix-form subsection (lines with `strand-1`, `by-type`, `strand-2`, and `helix-form`, respectively) gives a quick summary of the stacked bps: base sequences, bp types (canonical or otherwise), and helix conformations of the dinucleotide steps. Needless to say, the A-form helix is the most common type for RNA. However, since DSSR can be equally applied to DNA, the B- and Z-forms would also be included in the classification.

### 3.2.5 Stems (canonical pairs with continuous backbones)

In DSSR, a stem is defined as a helix consisting of only canonical WC/wobble pairs and possessing a continuous backbone along each strand. The literature does not appear to be consistent as to what constitutes a helix, stem, or arm. Hopefully DSSR will help to clarify

the confusion in the field. The default requirement for canonical bps in DSSR follows the convention widely adopted for RNA secondary structures, as in the **mfold/UNAFold** software and the **ViennaRNA Package**. The PDB entry 1msy contains one stem with five canonical bps, as shown below (see also [Figure 2\(B\)](#)). Notice how the bp names and the helix-form sub-section on [Page 21](#) facilitate a quick visual identification of the stem in the helix.

```
List of 1 stem
Note: a stem is defined as a helix consisting of only canonical WC/wobble
      pairs, with a continuous backbone.
stem#number[#helix-number containing this stem]
Other terms are defined as in the above Helix section.
-----
stem#1[#1] bps=5
strand-1 5'-GCUCC-3'
bp-type    | | | |
strand-2 3'-UGAGG-5'
helix-form .AAA
1 A.G2648     A.U2672      G-U Wobble   28-XXVIII cWW  cW-W
2 A.C2649     A.G2671      C-G WC       19-XIX   cWW  cW-W
3 A.U2650     A.A2670      U-A WC       20-XX    cWW  cW-W
4 A.C2651     A.G2669      C-G WC       19-XIX   cWW  cW-W
5 A.C2652     A.G2668      C-G WC       19-XIX   cWW  cW-W
```

### 3.2.6 Isolated canonical pairs

DSSR defines an isolated canonical bp as the one that is not part of a stem. In 1msy, A.C2658 and A.G2663 form such an isolated C–G WC pair, as detailed below. The [#1] at the beginning means this pair is part of helix #1. Note also the negative indices for isolated canonical bps vs. the positive values for stems. The significance of the distinction will become obvious later on, as internal and hairpin loops are specified.

```
List of 1 isolated WC/wobble pair
Note: isolated WC/wobble pairs are assigned negative indices to
      differentiate them from the stem numbers, which are positive.
-----
[#1] -1 A.C2658     A.G2663      C-G WC       19-XIX   cWW  cW-W
```

### 3.2.7 Base stacks

DSSR defined a base stack as an ordered list of nucleotides assembled together via base-stacking interactions, regardless of backbone connectivity or pairing involvement. Stacking interactions within a stem are **not** included by default. In 1msy, note the UAA stack (no. 2 in the listing) in the GUAA tetraloop, and the GGGG stack (no. 4). See [Figure 2](#). Nucleotides

not involved in stacking interactions are also listed, if any.

```
List of 5 base stacks
Note: a stack is an ordered list of nucleotides assembled together via
      base-stacking interactions, regardless of backbone connectivity.
      Stacking interactions within a stem are *not* included.
1 nts=2 GG A.G2648,A.G2673
2 nts=3 UAA A.U2660,A.A2661,A.A2662
3 nts=4 CUAU A.C2652,A.U2653,A.A2654,A.U2656
4 nts=4 GGGG A.G2655,A.G2664,A.G2663,A.G2659
5 nts=6 CAACCG A.C2658,A.A2657,A.A2665,A.C2666,A.C2667,A.G2668
```

### 3.2.8 Atom-base capping interactions

First described by Quigley and Rich in the tRNA<sup>Phe</sup> structure<sup>18</sup>, the phosphate group (actually the exocyclic OP2 atom) can stack over a base ring to cap a helix. Atom-base capping interactions are often observed in other structural motifs, including U-turns or GNRA tetraloops.

In DSSR, the stacking atoms also includes oxygen from water or the sugar moiety. The output for 1msy is given below. Check also the associated PDB file `dssr-a2bases.pdb` to visualize the findings.

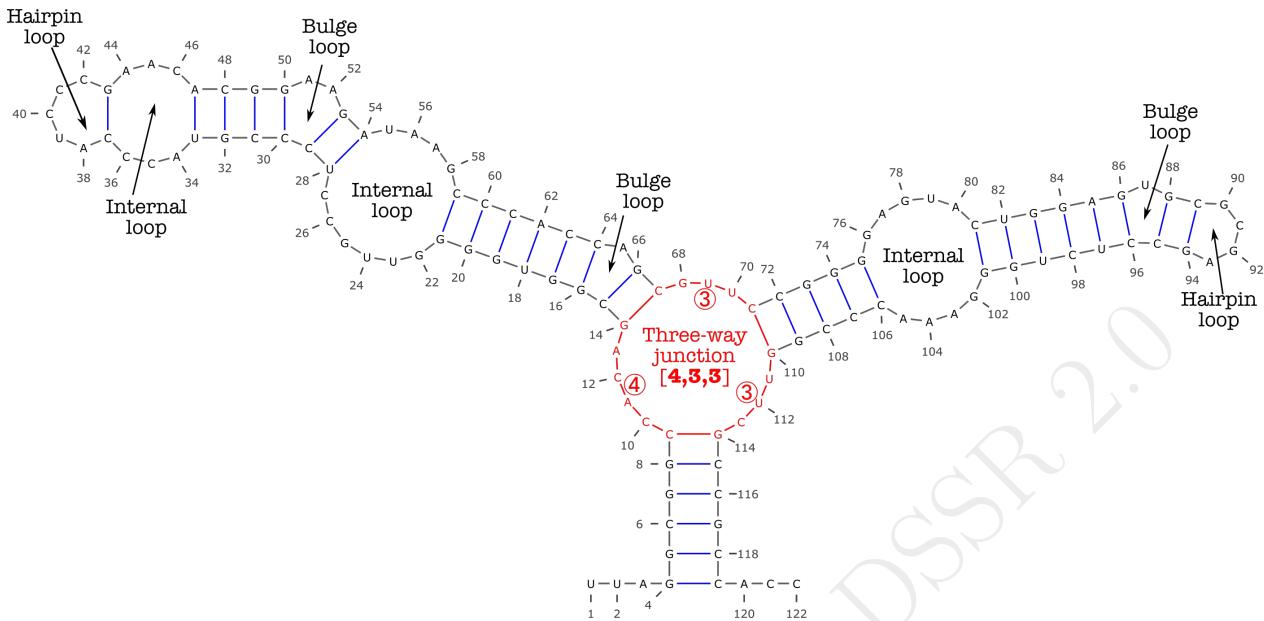
```
List of 2 atom-base capping interactions
dv: vertical distance of the atom above the nucleotide base
-----
type      atom          nt        dv
1 phosphate OP2@A.A2661    A.G2659   3.04
2 sugar     O4'@A.G2664    A.G2663   3.48
```

Here the `type` column can be ‘phosphate’, ‘sugar’, or ‘other’. The meaning of the `atom` and `nt` columns should be obvious. As noted on the 2nd line, the `dv` column gives vertical distance (in Å) of the atom from the nucleotide base. Thus, the first case shows that the OP2 atom of the A2661 phosphate group stacks 3.04 Å over the G2659 base ring.

### 3.2.9 Various loops

Commonly occurring RNA loops are illustrated in [Figure 9](#). DSSR identifies all of these different types and further distinguishes symmetric vs. asymmetric internal loops.

DSSR finds one hairpin loop and an asymmetric internal loop in 1msy, as detailed below. Note that the list of the 13 nts in the internal loop is too long to fit on the page width. The list is actually written on the same line as `nts=13 CUAGUACGGACCG`.



**Figure 9:** Common 2D structural elements of RNA. Single-stranded fragments at the 5' and 3' ends, and double-helical stems are obvious. 'Closed' loops of various types (hairpin, bulge, internal, and junction loops) are labeled.

Note: for the various types of loops listed below, numbers within the first set of brackets are the number of loop nts, and numbers in the second set of brackets are the identities of the stems (positive number) or isolated WC/wobble pairs (negative numbers) to which they are linked.

```
*****
List of 1 hairpin loop
 1 hairpin loop: nts=6; [4]; linked by [#-1]
  summary: [1] 4 [A.2658 A.2663] 1
  nts=6 CGUAAG A.C2658,A.G2659,A.U2660,A.A2661,A.A2662,A.G2663
  nts=4 GUAA A.G2659,A.U2660,A.A2661,A.A2662
*****
List of 1 internal loop
 1 asymmetric internal loop: nts=13; [5,4]; linked by [#1,#-1]
  summary: [2] 5 4 [A.2652 A.2668 A.2658 A.2663] 5 1
  nts=13 CUAGUACGGACCG A.C2652,A.U2653,A.A2654,A.G2655,A.U2656,A.A2657,A.C2658,A.G2663,A.
    ↪ G2664,A.A2665,A.C2666,A.C2667,A.G2668
  nts=5 UAGUA A.U2653,A.A2654,A.G2655,A.U2656,A.A2657
  nts=4 GACC A.G2664,A.A2665,A.C2666,A.C2667
```

With the note at the top and by referring to Figure 2(B), it should be straightforward to understand the meaning of most of the items in this section. Nevertheless, the two sets of matched brackets are worth further explanation. The hairpin loop contains a total of 6 nts, including the closing canonical pair: the [4] means 4 nts, i.e., a tetraloop; the [#-1] indicates

that the tetraloop is closed by the first isolated pair. The asymmetric internal loop contains 13 nts in total: the [5,4] means 5 nts along one strand, and 4 nts on the other; [#1,#-1] indicates that the internal loop is linked by the first stem and the first isolated pair.

### 3.2.10 Single-stranded fragments

DSSR also characterizes single-stranded fragments not included in various loop regions. For 1msy, there are two, each consisting of one terminal nucleotide (see also Figure 2(B)).

```
List of 2 non-loop single-stranded segments
1 nts=1 U A.U2647
2 nts=1 G A.G2673
```

### 3.2.11 2D structure in dot-bracket notation

The DSSR-derived 2D structure is written in an extended dot-bracket notation (dbn) with information for pseudoknots (as matched [], {}, or <> pairs, etc) and chain breaks (as &s), which can be fed directly into VARNA<sup>7</sup>. The dbn is written in FASTA format where the title line (with a > on the first column) is followed by base sequence and the 2D structure, each on a separate line. The dbn for 1msy is as follows.

```
Secondary structures in dot-bracket notation (dbn) as a whole and per chain
>1msy nts=27 [whole]
UGCUCUAGUACGUAGGACCGGAGUG
.((((.....(....)....))))..
>1msy-A #1 nts=27 0.30(2.47) [chain] RNA
UGCUCUAGUACGUAGGACCGGAGUG
.((((.....(....)....))))..
```

Since 1msy contains only a single continuous chain, the dbn contents for the whole and per chain are the same. For the Dickerson DNA dodecamer<sup>19</sup> structure 355d, the difference in contents between the whole structure and each chain is obvious from the directions of the brackets (see below).

```
Secondary structures in dot-bracket notation (dbn) as a whole and per chain
>355d nts=24 [whole]
CGCGAATTCTCGCG&CGCGAATTCTCGCG
((((((((((&))))))))))
>355d-A #1 nts=12 3.21(0.51) [chain] DNA
CGCGAATTCTCGCG
((((((((((
>355d-B #2 nts=12 3.37(0.50) [chain] DNA
CGCGAATTCTCGCG
))))))))))
```

DSSR also generates a file named `dssr-2ndstrs.dbn` containing the 2D structure for the whole molecule in dbn notation. For better connection to third-party tools, DSSR produces an additional file named `dssr-2ndstrs.ct`, which expresses the 2D structure in connectivity table (.ct) format. First introduced by the `mfold` program, the .ct format is one of the most commonly used RNA 2D structure formats. The .ct output file for 1msy is listed below. Moreover, DSSR also generates a 2D structure representation in the .bpseq format in a file named `dssr-2ndstrs.bpseq`.

```

27 ENERGY = 0.0 [1msy] -- secondary structure derived by DSSR
 1 U      0      2      0  2647 # name=A.U2647
 2 G      1      3     26  2648 # name=A.G2648, pairedNt=A.U2672
 3 C      2      4     25  2649 # name=A.C2649, pairedNt=A.G2671
 4 U      3      5     24  2650 # name=A.U2650, pairedNt=A.A2670
 5 C      4      6     23  2651 # name=A.C2651, pairedNt=A.G2669
 6 C      5      7     22  2652 # name=A.C2652, pairedNt=A.G2668
 7 U      6      8      0  2653 # name=A.U2653
 8 A      7      9      0  2654 # name=A.A2654
 9 G      8     10      0  2655 # name=A.G2655
10 U     9     11      0  2656 # name=A.U2656
11 A    10     12      0  2657 # name=A.A2657
12 C    11     13     17  2658 # name=A.C2658, pairedNt=A.G2663
13 G    12     14      0  2659 # name=A.G2659
14 U    13     15      0  2660 # name=A.U2660
15 A    14     16      0  2661 # name=A.A2661
16 A    15     17      0  2662 # name=A.A2662
17 G    16     18     12  2663 # name=A.G2663, pairedNt=A.C2658
18 G    17     19      0  2664 # name=A.G2664
19 A    18     20      0  2665 # name=A.A2665
20 C    19     21      0  2666 # name=A.C2666
21 C    20     22      0  2667 # name=A.C2667
22 G    21     23      6  2668 # name=A.G2668, pairedNt=A.C2652
23 G    22     24      5  2669 # name=A.G2669, pairedNt=A.C2651
24 A    23     25      4  2670 # name=A.A2670, pairedNt=A.U2650
25 G    24     26      3  2671 # name=A.G2671, pairedNt=A.C2649
26 U    25     27      2  2672 # name=A.U2672, pairedNt=A.G2648
27 G    26      0      0  2673 # name=A.G2673

```

### 3.2.12 Structural features per nucleotide

This section summarizes structural features that each nucleotide possesses or is part of. For each identified nucleotide, the output includes one-letter shorthand name, dbn, residue identifier, the rmsd of the base ring atoms, and a comma-separated list of features: *anti* or *syn* ( $\chi$ ) orientation, C2' or C3'-endo sugar puckering, BI or BII backbone, modified or not, if involved in canonical or non-canonical pair, multiplet, helix, stem, or various loops, etc.

```

Summary of structural features of 27 nucleotides
Note: the first five columns are: (1) serial number, (2) one-letter
shorthand name, (3) dbn, (4) id string, (5) rmsd (~zero) of base
ring atoms fitted against those in a standard base reference

```

```

frame. The sixth (last) column contains a comma-separated list of
features whose meanings are mostly self-explanatory, except for:
  turn: angle C1'(i-1)--C1'(i)--C1'(i+1) < 90 degrees
  break: no backbone linkage between O3'(i-1) and P(i)
  1 U . A.U2647 0.011 anti,~C3'-endo,non-canonical,non-pair-contact,helix-end,ss-non-loop
  2 G ( A.G2648 0.012 anti,~C3'-endo,BI,canonical,non-pair-contact,helix,stem-end
  3 C ( A.C2649 0.019 anti,~C3'-endo,BI,canonical,non-pair-contact,helix,stem
  4 U ( A.U2650 0.019 anti,~C3'-endo,BI,canonical,non-pair-contact,helix,stem
  5 C ( A.C2651 0.024 anti,~C3'-endo,BI,canonical,non-pair-contact,helix,stem
  6 C ( A.C2652 0.032 anti,~C3'-endo,BI,canonical,non-pair-contact,helix,stem-end,internal-loop
  7 U . A.U2653 0.019 anti,~C3'-endo,non-canonical,non-pair-contact,helix,internal-loop,phosphate
  8 A . A.A2654 0.019 anti,~C2'-endo,BII,non-canonical,non-pair-contact,helix,internal-loop
  9 G . A.G2655 0.022 turn,anti,~C2'-endo,non-canonical,non-pair-contact,multiplet,internal-loop
 10 U . A.U2656 0.020 anti,~C3'-endo,BI,non-canonical,non-pair-contact,helix,multiplet,internal-loop
    ↪ ,phosphate
 11 A . A.A2657 0.023 anti,~C3'-endo,BI,non-canonical,non-pair-contact,helix,internal-loop
 12 C ( A.C2658 0.013 anti,~C3'-endo,BI,isolated-canonical,non-pair-contact,helix,hairpin-loop,
    ↪ internal-loop
 13 G . A.G2659 0.033 u-turn,anti,~C3'-endo,BI,non-canonical,non-pair-contact,helix-end,hairpin-loop
    ↪ ,cap-acceptor,splayed-apart
 14 U . A.U2660 0.020 turn,u-turn,anti,~C3'-endo,non-pair-contact,hairpin-loop,splayed-apart
 15 A . A.A2661 0.015 u-turn,anti,~C3'-endo,BI,non-pair-contact,hairpin-loop,cap-donor,phosphate
 16 A . A.A2662 0.010 u-turn,anti,~C3'-endo,BI,non-canonical,non-pair-contact,helix-end,hairpin-loop
    ↪ ,phosphate
 17 G ) A.G2663 0.019 anti,~C3'-endo,BI,isolated-canonical,non-pair-contact,helix,hairpin-loop,
    ↪ internal-loop,cap-acceptor
 18 G . A.G2664 0.014 anti,~C3'-endo,BI,non-canonical,non-pair-contact,helix,internal-loop,cap-donor
 19 A . A.A2665 0.014 anti,~C3'-endo,BI,non-canonical,non-pair-contact,helix,multiplet,internal-loop
    ↪ ,phosphate
 20 C . A.C2666 0.016 anti,~C3'-endo,BI,non-canonical,non-pair-contact,helix,internal-loop,phosphate
 21 C . A.C2667 0.029 anti,~C3'-endo,BI,non-canonical,non-pair-contact,helix,internal-loop
 22 G ) A.G2668 0.012 anti,~C3'-endo,BI,canonical,non-pair-contact,helix,stem-end,internal-loop
 23 G ) A.G2669 0.020 anti,~C3'-endo,BI,canonical,non-pair-contact,helix,stem
 24 A ) A.A2670 0.019 anti,~C3'-endo,BI,canonical,non-pair-contact,helix,stem
 25 G ) A.G2671 0.023 anti,~C3'-endo,BI,canonical,non-pair-contact,helix,stem
 26 U ) A.U2672 0.024 anti,~C3'-endo,BI,canonical,non-pair-contact,helix,stem-end
 27 G . A.G2673 0.010 anti,~C3'-endo,non-canonical,non-pair-contact,helix-end,ss-non-loop

```

### 3.2.13 Backbone torsion angles and suite names

The auto-generated output file named `dssr-torsions.txt` contains many commonly used backbone parameters, including torsion angles, sugar puckers, and suite names<sup>8</sup>. Given below is a summary of the sections contained in the file so that users can better understand what DSSR has to offer.

**Main chain conformational parameters** Here the single-stranded Zp (ssZp) parameter is an extension to the original 3DNA Zp for distinguishing different types of duplexes<sup>20</sup>. Its addition to 3DNA/DSSR has been inspired by the work of Richardson *et al.*<sup>21</sup>, who observed a correlation between the sugar pucker and the perpendicular distance from the 3'-phosphate to the glycosidic bond vector: >2.9 Å for C3'-endo and <2.9 Å for C2'-endo sugars.

Note also the classifications of the backbone into BI/BII forms,  $\chi$  into *anti/syn* conformations, and sugar into C2'-endo/C3'-endo puckers. The collection of relevant information should prove convenient for identification of key backbone features of potential interest.

```

alpha:  O3'(i-1)-P-05'-C5'
beta:   P-05'-C5'-C4'
gamma:  O5'-C5'-C4'-C3'
delta:  C5'-C4'-C3'-O3'
epsilon: C4'-C3'-O3'-P(i+1)
zeta:   C3'-O3'-P(i+1)-O5'(i+1)
e-z:    epsilon-zeta (BI/BII backbone classification)

chi for pyrimidines(Y): 04'-C1'-N1-C2; purines(R): 04'-C1'-N9-C4
Range [170, -50(310)] is assigned to anti, and [50, 90] to syn

phase-angle: the phase angle of pseudorotation and puckering
sugar-type: ~C2'-endo for C2'-endo like conformation, or
            ~C3'-endo for C3'-endo like conformation
            Note the ONE column offset (for easy visual distinction)

ssZp: single-stranded Zp, defined as the z-coordinate of the 3' phosphorus atom
      (P) expressed in the standard reference frame of the 5' base; the value is
      POSITIVE when P lies on the +z-axis side (base in anti conformation);
      NEGATIVE if P is on the -z-axis side (base in syn conformation)
Dp: perpendicular distance of the 3' P atom to the glycosidic bond
[Ref: Chen et al. (2010): "MolProbity: all-atom structure
validation for macromolecular crystallography."
Acta Crystallogr D Biol Crystallogr, 66(1):12-21]
splay: angle between the bridging P to the two base-origins of a dinucleotide.

```

There are too many parameters to fit into the page width. The following shows a selected portion of nucleotides in 1msy, with parameters split into two sections.

|    | nt        | alpha        | beta            | gamma      | delta | epsilon | zeta  | e-z      |
|----|-----------|--------------|-----------------|------------|-------|---------|-------|----------|
| 5  | C A.C2651 | -62.3        | 173.8           | 46.2       | 84.3  | -154.6  | -71.3 | -83(BI)  |
| 6  | C A.C2652 | -67.5        | 175.2           | 58.1       | 78.0  | -154.0  | -67.0 | -87(BI)  |
| 7  | U A.U2653 | -61.8        | 167.4           | 55.9       | 82.4  | -149.4  | 55.3  | 155(..)  |
| 8  | A A.A2654 | 165.2        | 133.8           | 56.7       | 149.4 | -98.3   | 161.4 | 100(BII) |
| 9  | G A.G2655 | -96.2        | 91.7            | 179.3      | 149.3 | -167.2  | 141.4 | 51(..)   |
| 10 | U A.U2656 | -72.5        | 157.8           | 37.9       | 91.6  | -141.0  | -65.6 | -75(BI)  |
|    | nt        | chi          | phase-angle     | sugar-type | ssZp  | Dp      | splay |          |
| 5  | C A.C2651 | -159.0(anti) | 12.9(C3'-endo)  | ~C3'-endo  | 4.29  | 4.58    | 24.06 |          |
| 6  | C A.C2652 | -158.8(anti) | 15.5(C3'-endo)  | ~C3'-endo  | 4.28  | 4.53    | 22.84 |          |
| 7  | U A.U2653 | -151.0(anti) | 15.7(C3'-endo)  | ~C3'-endo  | 4.00  | 4.62    | 22.57 |          |
| 8  | A A.A2654 | -145.3(anti) | 151.0(C2'-endo) | ~C2'-endo  | 0.91  | 0.92    | 43.08 |          |
| 9  | G A.G2655 | -93.5(anti)  | 151.5(C2'-endo) | ~C2'-endo  | 2.15  | 2.14    | 54.21 |          |
| 10 | U A.U2656 | -173.9(anti) | 0.4(C3'-endo)   | ~C3'-endo  | 4.35  | 4.42    | 27.39 |          |

**Virtual torsion angles** Three sets of virtual torsion angles are calculated: one is the most commonly used  $\eta/\theta$  pair pioneered by Olson<sup>22</sup>; the second is its  $\eta'/\theta'$  variant (using the C1' atom instead of C4') later introduced by Pyle *et al.*<sup>23</sup>; and the third (termed  $\eta''/\theta''$ ) takes

advantage of the origin of the base reference frame<sup>4</sup> in place of the C4' or C1' atom. The set of base-phosphorus virtual torsions is unique to 3DNA/DSSR.

```

eta:    C4'(i-1)-P(i)-C4'(i)-P(i+1)
theta:   P(i)-C4'(i)-P(i+1)-C4'(i+1)
[Ref: Olson (1980): "Configurational statistics of polynucleotide chains.
An updated virtual bond model to treat effects of base stacking."
Macromolecules, 13(3):721-728]

eta':   C1'(i-1)-P(i)-C1'(i)-P(i+1)
theta': P(i)-C1'(i)-P(i+1)-C1'(i+1)
[Ref: Keating et al. (2011): "A new way to see RNA." Quarterly Reviews
of Biophysics, 44(4):433-466]

eta":   base(i-1)-P(i)-base(i)-P(i+1)
theta": P(i)-base(i)-P(i+1)-base(i+1)

      nt          eta     theta     eta'     theta'     eta"     theta"
1    U A.U2647      ---      ---      ---      ---      ---      ---
2    G A.G2648     172.1   -133.3   -163.5   -131.2   -97.6   -93.1
3    C A.C2649     162.7   -140.1   -178.0   -138.9   -120.5  -104.8
4    U A.U2650     167.0   -148.3   -174.5   -150.1   -122.6  -127.9
5    C A.C2651     165.4   -147.2   -177.6   -145.6   -144.3  -123.9
6    C A.C2652     171.3   -142.0   -176.3   -139.3   -142.0  -91.5
7    U A.U2653     172.2   -18.3    -170.0   -63.3    -107.9  -99.1
8    A A.A2654      46.3    172.0    120.5    135.6    126.4   150.3
9    G A.G2655     -44.2    24.9    -82.3    59.0    -98.6   91.3
10   U A.U2656     170.9   -121.7   163.1    -122.7   161.9   -94.8
11   A A.A2657     162.1   -127.4   -177.7   -123.2   -126.9  -67.0
12   C A.C2658     159.4   -135.3   -176.0   -135.4   -98.2   -106.5
13   G A.G2659     167.6   -117.7   -179.4   -160.6   -134.1  156.9
14   U A.U2660      15.1    -126.1    43.5    -124.9    21.0   -65.2
15   A A.A2661     160.4   -132.0   -169.4   -138.7   -95.8   -93.3
16   A A.A2662     167.0   -83.0    -174.8   -81.7   -117.0  -64.2
17   G A.G2663     172.6   -154.0   -148.2   -148.6   -101.2  -97.0
18   G A.G2664     166.2    166.9   -168.9   147.4    -95.5   137.9
19   A A.A2665     -155.6   141.6    175.0   164.3    154.7   -178.6
20   C A.C2666     -178.4   -125.3   -169.0   -123.0   -153.9  -74.8
21   C A.C2667     164.6   -120.7   -172.9   -116.0   -101.6  -71.6
22   G A.G2668     164.9   -150.0   -168.4   -145.8   -97.4   -120.2
23   G A.G2669     171.3   -139.8   -172.8   -139.5   -131.7  -111.7
24   A A.A2670     170.6   -153.2   -173.8   -152.7   -127.2  -132.0
25   G A.G2671     170.4   -134.4   -180.0   -133.1   -147.1  -94.7
26   U A.U2672     172.2   -167.9   -166.6   -170.6   -110.6  -171.6
27   G A.G2673      ---      ---      ---      ---      ---      ---
```

**Sugar conformational parameters** By default, the sugar pucker analysis follows the work of Altona and Sundaralingam<sup>24</sup>. The phase angle (in the range of 0° to 360°) of pseudorotation is divided equally into ten 36° regions; the two most frequent sugar pucker modes are the C3'-endo [0°, 36°) as in canonical RNA and A-form DNA, and the C2'-endo [144°, 180°) as in standard B-form DNA. Where appropriate, each sugar pucker is assigned into either ~C3'-endo or ~C2'-endo (see above in [Section 3.2.13](#) under ‘sugar-type’) by matching against corresponding fiber models.

|    | nt        | v0    | v1    | v2    | v3    | v4   | tm   | P     | Puckering |
|----|-----------|-------|-------|-------|-------|------|------|-------|-----------|
| 1  | U A.U2647 | 7.5   | -34.5 | 46.2  | -43.8 | 23.1 | 46.9 | 9.9   | C3'-endo  |
| 2  | G A.G2648 | 9.5   | -31.5 | 39.4  | -35.5 | 16.8 | 39.6 | 5.3   | C3'-endo  |
| 3  | C A.C2649 | 4.0   | -28.3 | 39.9  | -38.4 | 21.9 | 40.9 | 12.9  | C3'-endo  |
| 4  | U A.U2650 | -2.4  | -25.5 | 41.9  | -44.0 | 29.4 | 44.9 | 21.3  | C3'-endo  |
| 5  | C A.C2651 | 4.8   | -32.5 | 45.7  | -44.6 | 25.0 | 46.9 | 12.9  | C3'-endo  |
| 6  | C A.C2652 | 2.9   | -29.7 | 43.8  | -44.0 | 25.9 | 45.4 | 15.5  | C3'-endo  |
| 7  | U A.U2653 | 1.6   | -28.2 | 42.3  | -41.3 | 25.2 | 44.0 | 15.7  | C3'-endo  |
| 8  | A A.A2654 | -33.2 | 44.3  | -38.7 | 20.1  | 8.5  | 44.3 | 151.0 | C2'-endo  |
| 9  | G A.G2655 | -37.3 | 50.1  | -43.9 | 22.9  | 8.9  | 50.0 | 151.5 | C2'-endo  |
| 10 | U A.U2656 | 12.7  | -32.9 | 39.6  | -33.2 | 13.3 | 39.6 | 0.4   | C3'-endo  |
| 11 | A A.A2657 | -6.4  | -21.7 | 39.9  | -44.5 | 32.0 | 44.6 | 26.5  | C3'-endo  |
| 12 | C A.C2658 | -0.0  | -28.5 | 44.6  | -44.4 | 28.4 | 46.9 | 17.9  | C3'-endo  |
| 13 | G A.G2659 | -9.4  | -20.6 | 40.1  | -45.3 | 35.5 | 46.1 | 29.4  | C3'-endo  |
| 14 | U A.U2660 | -8.9  | -14.6 | 31.3  | -37.4 | 29.0 | 37.0 | 32.3  | C3'-endo  |
| 15 | A A.A2661 | 7.2   | -28.7 | 38.3  | -35.0 | 17.5 | 38.6 | 8.0   | C3'-endo  |
| 16 | A A.A2662 | 0.0   | -21.1 | 32.9  | -33.9 | 21.4 | 34.8 | 18.7  | C3'-endo  |
| 17 | G A.G2663 | -7.9  | -22.7 | 42.3  | -47.3 | 35.7 | 47.8 | 27.7  | C3'-endo  |
| 18 | G A.G2664 | 7.4   | -33.6 | 45.9  | -43.1 | 22.1 | 46.5 | 9.7   | C3'-endo  |
| 19 | A A.A2665 | 11.2  | -35.6 | 45.2  | -39.6 | 18.1 | 45.4 | 4.5   | C3'-endo  |
| 20 | C A.C2666 | 2.0   | -27.6 | 41.2  | -41.2 | 24.6 | 42.8 | 16.0  | C3'-endo  |
| 21 | C A.C2667 | 1.9   | -30.7 | 45.7  | -46.0 | 28.2 | 47.6 | 16.5  | C3'-endo  |
| 22 | G A.G2668 | 9.1   | -34.4 | 44.4  | -40.9 | 20.0 | 44.7 | 7.3   | C3'-endo  |
| 23 | G A.G2669 | 4.5   | -31.3 | 45.1  | -43.5 | 24.4 | 46.3 | 13.0  | C3'-endo  |
| 24 | A A.A2670 | 5.3   | -32.9 | 46.7  | -45.3 | 25.1 | 47.9 | 12.6  | C3'-endo  |
| 25 | G A.G2671 | 7.6   | -33.2 | 44.6  | -41.4 | 21.7 | 45.2 | 9.2   | C3'-endo  |
| 26 | U A.U2672 | -0.3  | -23.0 | 35.9  | -37.6 | 23.6 | 38.0 | 19.2  | C3'-endo  |
| 27 | G A.G2673 | 17.1  | -39.6 | 45.4  | -37.1 | 12.9 | 45.4 | 357.2 | C2'-exo   |

**Assignment of sugar-phosphate backbone suite names** According to Richardson *et al.*<sup>8</sup>, the backbone suite is defined as the sugar-to-sugar version of a nucleotide (in contrast to the traditional definition as a phosphate-to-phosphate unit). A total of 53 backbone conformer bins have been defined and expressed in mnemonic 2-letter names (e.g., 1a, 5z, with outliers signified by !!).

```
bin: name of the 12 bins based on [delta(i-1), delta, gamma], where
      delta(i-1) and delta can be either 3 (for C3'-endo sugar) or 2
      (for C2'-endo) and gamma can be p/t/m (for gauche+/trans/gauche-
      conformations, respectively) (2x2x3=12 combinations: 33p, 33t,
      ... 22m); 'inc' refers to incomplete cases (i.e., with missing
      torsions), and 'trig' to triages (i.e., with torsion angle
      outliers)
```

```
cluster: 2-char suite name, for one of 53 reported clusters (46
certain and 7 wannabes), '--' for incomplete cases, and
'!!!' for outliers
suiteness: measure of conformer-match quality (low to high in range 0 to 1)
```

[Ref: Richardson et al. (2008): "RNA backbone: consensus all-angle conformers and modular string nomenclature (an RNA Ontology Consortium contribution)." RNA, 14(3):465-481]

|    | nt        | bin  | cluster | suiteness |
|----|-----------|------|---------|-----------|
| 1  | U A.U2647 | inc  | --      | 0         |
| 2  | G A.G2648 | 33p  | 1a      | 0.052     |
| 3  | C A.C2649 | 33p  | 1a      | 0.665     |
| 4  | U A.U2650 | 33p  | 1a      | 0.875     |
| 5  | C A.C2651 | 33p  | 1a      | 0.871     |
| 6  | C A.C2652 | 33p  | 1a      | 0.919     |
| 7  | U A.U2653 | 33p  | 1a      | 0.929     |
| 8  | A A.A2654 | 32p  | 5z      | 0.849     |
| 9  | G A.G2655 | 22t  | 4s      | 0.730     |
| 10 | U A.U2656 | 23p  | #a      | 0.842     |
| 11 | A A.A2657 | 33p  | 1a      | 0.693     |
| 12 | C A.C2658 | 33p  | 1a      | 0.884     |
| 13 | G A.G2659 | 33p  | 1a      | 0.894     |
| 14 | U A.U2660 | 33p  | 1g      | 0.736     |
| 15 | A A.A2661 | 33p  | 1L      | 0.688     |
| 16 | A A.A2662 | 33p  | 1a      | 0.692     |
| 17 | G A.G2663 | 33t  | 1c      | 0.321     |
| 18 | G A.G2664 | 33p  | 1a      | 0.878     |
| 19 | A A.A2665 | 33t  | 1e      | 0.875     |
| 20 | C A.C2666 | 33p  | 1a      | 0.891     |
| 21 | C A.C2667 | 33p  | 1a      | 0.887     |
| 22 | G A.G2668 | 33p  | 1a      | 0.756     |
| 23 | G A.G2669 | 33p  | 1a      | 0.625     |
| 24 | A A.A2670 | 33p  | 1a      | 0.914     |
| 25 | G A.G2671 | 33p  | 1a      | 0.878     |
| 26 | U A.U2672 | 33p  | 1a      | 0.912     |
| 27 | G A.G2673 | trig | !!      | 0         |

Concatenated suite string per chain. To avoid confusion of lower case modified nucleotide name (e.g., 'a') with suite cluster (e.g., '1a'), use --suite-delimiter to add delimiters (matched ')' by default).

```
1 A RNA nts=27 U1aG1aC1aU1aC1aC1aU5zA4sG#
 ↳ aU1aA1aC1aG1gU1LA1aA1cG1aG1eA1aC1aC1aG1aA1aG1aU!!G
```

Note that in assigning suite names, the  $\chi$  torsion angle, which characterizes the relative sugar-base orientation, is not taken into consideration. Moreover, the 53 defined backbone conformer bins are RNA-centric: even for the classic B-DNA Dickerson dodecamer ([355d](#)), 16 out of 22 suites (~73%) are classified as outliers (!!).

### 3.3 Default run on PDB entry 1ehz (tRNA<sup>Phe</sup>): summary notes

The cloverleaf 2D structure of tRNA has become an iconic image in structural biology. The four stems (the acceptor stem, the D stem, the anti-codon stem, and the T stem) form two halves of the L-shaped tertiary structure through coaxial stacking of two pairs of stems.

Yet, other than DSSR, there appear to be no alternative software programs that can neatly delineate the L-shaped 3D vs. the cloverleaf 2D structures of a tRNA molecule from atomic coordinates. The cover image of the manual illustrates DSSR's capability to solve this basic problem, using [1ehz](#), the crystal structure of yeast phenylalanine tRNA, as an example.

### 3.3.1 Brief summary

The screen output of a DSSR run on [1ehz](#) (assuming the 3D coordinates file is called [1ehz.pdb](#)) is shown below. Note that DSSR correctly identifies two helices, four stems, three hairpin loops, and one four-way junction loop, among other things.

```

total number of nucleotides: 76
modified nucleotides: 14
total number of base pairs: 34
total number of multiplets: 4
total number of helices: 2
total number of stems: 4
total number of isolated WC/wobble pairs: 1
total number of atom-base capping interactions: 4
total number of splayed-apart dinucleotides: 9
consolidated into units: 6
total number of hairpin loops: 3
total number of junctions: 1
total number of non-loop single-stranded segments: 1
total number of kissing loops: 1

```

Time used: 00:00:00:00

### 3.3.2 Modified nucleotides

The 27-nt long RNA fragment [1msy](#) detailed in [Section 3.2](#) contains only canonical nts (A, C, G, and U). However, 14 out of the 76 nts in tRNA<sup>Phe</sup> [1ehz](#) are modified (of 11 different types), as listed below.

| List of 11 types of 14 modified nucleotides |       |       |                   |
|---|-------|-------|-------------------|
|   | nt    | count | list              |
| 1   | 1MA-a | 1     | A.1MA58           |
| 2   | 2MG-g | 1     | A.2MG10           |
| 3   | 5MC-c | 2     | A.5MC40 , A.5MC49 |
| 4   | 5MU-t | 1     | A.5MU54           |
| 5   | 7MG-g | 1     | A.7MG46           |
| 6   | H2U-u | 2     | A.H2U16 , A.H2U17 |
| 7   | M2G-g | 1     | A.M2G26           |
| 8   | OMC-c | 1     | A.OMC32           |
| 9   | OMG-g | 1     | A.OMG34           |
| 10  | PSU-P | 2     | A.PSU39 , A.PSU55 |
| 11  | YYG-g | 1     | A.YYG37           |

Details about each of the 11 types of modified nts are: a 3-letter residue name followed by its 1-letter shorthand form (under column **nt**), its frequency (under column **count**), and a comma separated enumeration of its occurrences (under column **list**). For example, 5MC (no. 3) is found twice in 1ehz, with residue numbers 40 and 49, respectively.

From its 3-letter residue name, further information about a modified nt can be obtained via RCSB Ligand Explorer. For example, click the link to check [H2U \(5,6-dihydrouridine-5'-monophosphate\)](#).

### 3.3.3 The four triplets

DSSR detects four base triplets as listed below. Use Jmol/PyMOL to visualize the output file **dssr-multiplets.pdb** for verification.

```
List of 4 multiplets
1 nts=3 UAA A.U8,A.A14,A.A21
2 nts=3 AUA A.A9,A.U12,A.A23
3 nts=3 gCG A.2MG10,A.C25,A.G45
4 nts=3 CGg A.C13,A.G22,A.7MG46
```

### 3.3.4 Relationship between helices and stems

The connection between the two helices and the four stems is available from the main output file, with related portions excerpted below. The meaning of each section should be easy to follow, especially in connection with the tRNA<sup>Phe</sup> (1ehz) 2D structure image shown on the cover of this manual.

```
helix#1[2] bps=15
helix#2[2] bps=15

stem#1[#1] bps=7
stem#2[#2] bps=4
stem#3[#2] bps=4
stem#4[#1] bps=5

List of 2 coaxial stacks
1 Helix#1 contains 2 stems: [#1,#4]
2 Helix#2 contains 2 stems: [#3,#2]
```

### 3.3.5 Three hairpin loops

As expected, three hairpin loops are identified, with details listed below.

```
List of 3 hairpin loops
1 hairpin loop: nts=10; [8]; linked by [#2]
  summary: [1] 8 [A.13 A.22] 4
  nts=10 CAGuuGGGAG A.C13,A.A14,A.G15,A.H2U16,A.H2U17,A.G18,A.G19,A.G20,A.A21,A.G22
    nts=8 AGuuGGGA A.A14,A.G15,A.H2U16,A.H2U17,A.G18,A.G19,A.G20,A.A21
2 hairpin loop: nts=11; [9]; linked by [#3]
  summary: [1] 9 [A.30 A.40] 4
  nts=11 GAcUgAAGAPc A.G30,A.A31,A.OMC32,A.U33,A.OMG34,A.A35,A.A36,A.YYG37,A.A38,A.PSU39,
    ↪ A.5MC40
    nts=9 AcUgAAGAP A.A31,A.OMC32,A.U33,A.OMG34,A.A35,A.A36,A.YYG37,A.A38,A.PSU39
3 hairpin loop: nts=9; [7]; linked by [#4]
  summary: [1] 7 [A.53 A.61] 5
  nts=9 GtPCGaUCC A.G53,A.5MU54,A.PSU55,A.C56,A.G57,A.1MA58,A.U59,A.C60,A.C61
    nts=7 tPCGaUC A.5MU54,A.PSU55,A.C56,A.G57,A.1MA58,A.U59,A.C60
```

### 3.3.6 One four-way junction loop

The four-way junction loop is delineated by the four stems ([#1,#2,#3,#4]), with [2,1,5,0] nucleotides connecting each consecutive stems. This loop is well-documented in literature, forming the core of the cloverleaf: see the cover image on the right, and also the 2015 DSSR paper<sup>1</sup>. Note that the junction loop contains four modified nucleotides (2MG10, M2G26, 7MG46, 5MC49), which pose a problem for some RNA structural analysis programs.

```
List of 1 junction
1 4-way junction: nts=16; [2,1,5,0]; linked by [#1,#2,#3,#4]
  summary: [4] 2 1 5 0 [A.7 A.66 A.10 A.25 A.27 A.43 A.49 A.65] 7 4 4 5
  nts=16 UUAgCgCGAGgUCcGA A.U7,A.U8,A.A9,A.2MG10,A.C25,A.M2G26,A.C27,A.G43,A.A44,A.G45,A
    ↪ .7MG46,A.U47,A.C48,A.5MC49,A.G65,A.A66
  nts=2 UA A.U8,A.A9
  nts=1 g A.M2G26
  nts=5 AGgUC A.A44,A.G45,A.7MG46,A.U47,A.C48
  nts=0
```

### 3.3.7 Pseudoknot

The tRNA<sup>Phe</sup> 1ehz contains a pseudoknot, due to the formation of the canonical G–C pair between G19 and C56 (see the long cyan line in the 2D representation on the cover image). This pair is conserved and located at the elbow of the L-shaped tertiary structure that brings the D-loop and the TΨC-loop together, via essentially a simple kissing loop interaction. The elbow G–C pair has been shown to stack against a base triplet of the T-box riboswitch<sup>25</sup>, apparently playing an import role in forming the [stem-I-tRNA complex \(41ck\)](#).

In DSSR dbn output, the bps in first-order pseudoknots are designated by matched square brackets, as shown below for 1ehz.

```
>1ehz-A #1 nts=76 0.09(2.86) [chain] RNA
GCGGAUUUAgCUCAGuuGGGAGAGCgCCAGAcUgAAGAPcUGGAGgUCcUGUGtPCGaUCCACAGAAUUCGCACCA
((((((..(((.....[..))).((((.....))))....((((..]....))))))))....
```

For bps in higher-order pseudoknots, matched curly brackets {}, angle brackets <>, or upper-lower case letters (Aa, Bb, Cc etc.) are used. Please refer to [Section 3.6](#) for an example.

### 3.4 Default run on PDB entry 1jj2: four auto-checked motifs

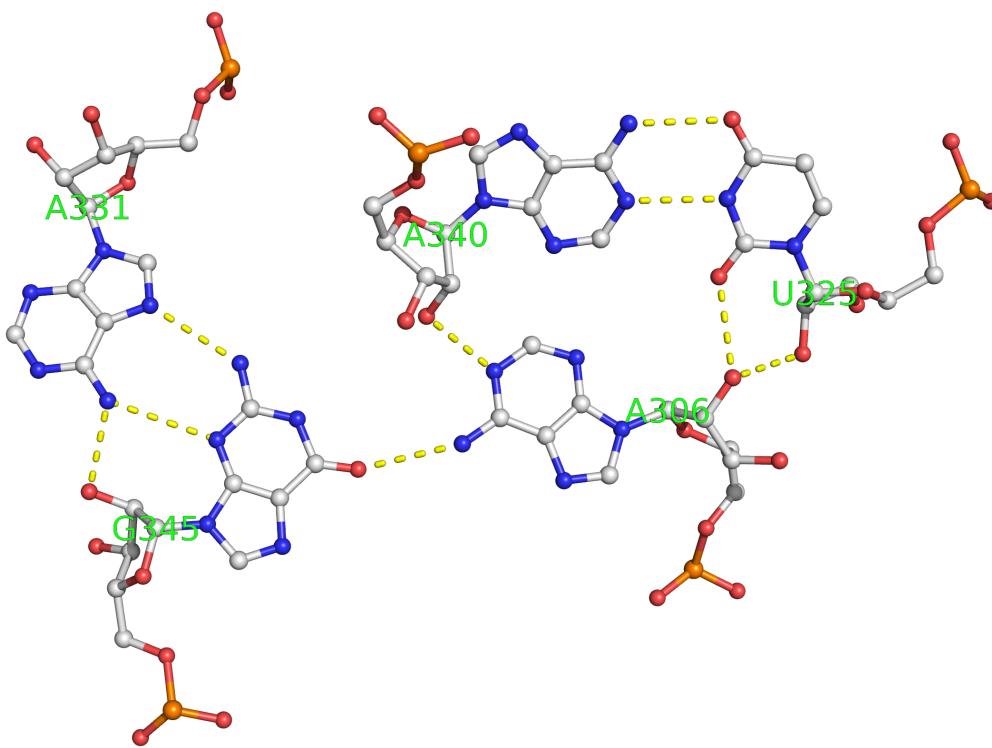
The crystal structure of the *Haloarcula marismortui* large ribosomal subunit ([1jj2](#)) serves as an example on how DSSR can analyze complicated RNA structures as easily as smaller ones (such as [1msy](#) and [1ehz](#) discussed in previous sections). Let the PDB 3D atomic coordinates file be called `1jj2.pdb`, the screen output is shown below. The running time will obviously depend on hardware configurations.

```
total number of nucleotides: 2876
total number of amino acids: 3701
total number of base pairs: 1470
total number of multiplets: 249
total number of helices: 87
total number of stems: 179
total number of isolated WC/wobble pairs: 53
total number of atom-base capping interactions: 239
total number of splayed-apart dinucleotides: 409
    consolidated into units: 257
total number of hairpin loops: 68
total number of bulges: 39
total number of internal loops: 66
total number of junctions: 37
total number of non-loop single-stranded segments: 30
total number of kissing loops: 5
total number of A-minor (types I and II) motifs: 107
total number of eXtended A-minor (type X) motifs: 49
total number of ribose zippers: 46
total number of kink turns: 8
```

Time used: 00:00:00:31

Notice the large numbers of bps, multiplets, loops (of various types), and four additional motifs—kissing loops, A-minor motifs, ribose zippers, and k-turns—that will be discussed in detail in the following sections.

As a side note, DSSR detects an additional base pentaplet (AUAAG, see [Figure 10](#)) which was missed using 3DNA<sup>9</sup>. The five nts (A306, U325, A331, A340, and G345), all on chain 0, form five base-base H-bonds as well as four additional H-bonding interactions involving O2' atoms. Overall, in default settings, DSSR is more sophisticated than the 3DNA `find_pair` program in identifying bps, multiplets, and double helical regions.



**Figure 10:** The additional base pentaplet (AUUAG) identified by DSSR but missed using 3DNA<sup>9</sup>. Here all five nts are derived from chain 0 of 1jj2. The image was produced with PyMOL.

### 3.4.1 Kissing loops

The kissing-loop motif is characterized by canonical base-pairing interactions (forming a stem) between two hairpin loops. Five such motifs are identified in 1jj2 as listed below, and no. 5 is illustrated in Figure 11.

```
List of 5 kissing loop interactions
 1 isolated-pair #42 between hairpin loops #51 and #53
 2 isolated-pair #6 between hairpin loops #6 and #8
 3 stem #8 between hairpin loops #1 and #3
 4 stem #9 between hairpin loops #1 and #3
 5 stem #29 between hairpin loops #14 and #57
```

The stem (#29) referred to above is:

|                       |         |        |        |          |
|-----------------------|---------|--------|--------|----------|
| stem#29[#15] bps=6    |         |        |        |          |
| strand-1 5'-CAUCGA-3' |         |        |        |          |
| bp-type               |         |        |        |          |
| strand-2 3'-GUAGUU-5' |         |        |        |          |
| helix-form AAA..      |         |        |        |          |
| 1 O.C418              | 0.G2449 | C-G WC | 19-XIX | cWW cW-W |

|          |         |            |           |     |      |
|----------|---------|------------|-----------|-----|------|
| 2 O.A419 | O.U2448 | A-U WC     | 20-XX     | cWW | cW-W |
| 3 O.U420 | O.A2447 | U-A WC     | 20-XX     | cWW | cW-W |
| 4 O.C421 | O.G2446 | C-G WC     | 19-XIX    | cWW | cW-W |
| 5 O.G422 | O.U2445 | G-U Wobble | 28-XXVIII | cWW | cW-W |
| 6 O.A423 | O.U2444 | A-U WC     | 20-XX     | cWW | cW-W |

The two interacting hairpin loops (#14 and #57) are:

```
14 hairpin loop: nts=9; [7]; linked by [#28]
  summary: [1] 7 [0.416 0.424] 5
  nts=9 GGCAUCGAC 0.G416 ,0.G417 ,0.C418 ,0.A419 ,0.U420 ,0.C421 ,0.G422 ,0.A423 ,0.C424
  nts=7 GCAUCGA 0.G417 ,0.C418 ,0.A419 ,0.U420 ,0.C421 ,0.G422 ,0.A423

57 hairpin loop: nts=9; [7]; linked by [#145]
  summary: [1] 7 [0.2442 0.2450] 5
  nts=9 GCUUGAUGC 0.G2442 ,0.C2443 ,0.U2444 ,0.U2445 ,0.G2446 ,0.A2447 ,0.U2448 ,0.G2449 ,0.C2450
  nts=7 CUUGAUG 0.C2443 ,0.U2444 ,0.U2445 ,0.G2446 ,0.A2447 ,0.U2448 ,0.G2449
```

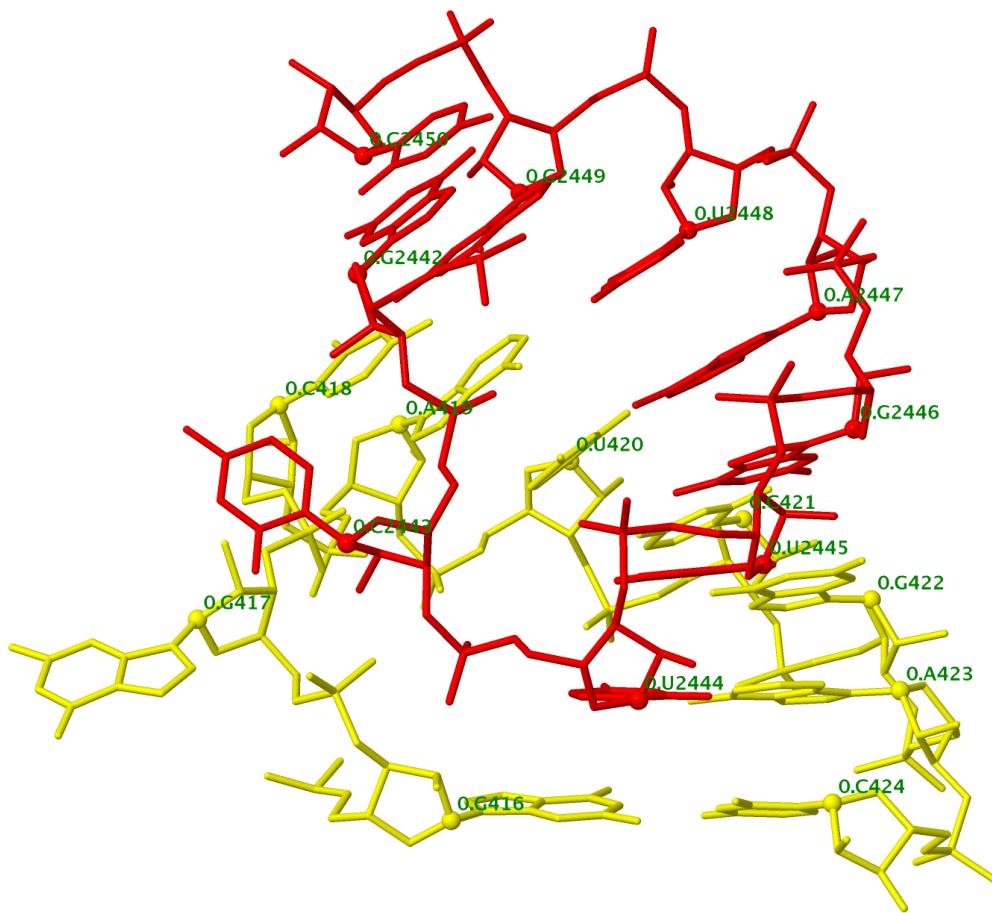
### 3.4.2 A-minor motifs

The interaction of the minor groove edge of an adenine with the minor groove side of a canonical pair is defined as the A-minor motif<sup>16</sup>. This abundant structural motif stabilizes RNA tertiary structures. Depending on the position of the adenine with respect to the interacting pair, the A-minor motif has been further divided into four subtypes. Of these, only two types (I and II) are believed to be adenine-specific, and they are identified by DSSR (see [Figure 12](#)).

A new type, designated X (eXtended, for cases other than the classic types I and II), was also identified by DSSR. In types I and II A-minor motifs, the adenine has its minor groove edge facing the minor groove of a canonical pair, and the O2' atom of adenine is involved in H-bonding interactions with the pair. In the miscellaneous type X, the adenine uses its Watson-Crick edge or major-groove edge to interact with the minor groove of a canonical pair, without resorting to the O2' atom. This type of A-minor interactions was reported in the crystal structure of the self-cleaving Pistol ribozyme<sup>26</sup>.

DSSR is unique in using standard base reference frames ([Figure 1B](#)) and bp parameters to characterize A-minor motifs. When applied to 1jj2, a total of 156 A-minor motifs were identified (stored in file `dssr-Aminors.pdb`), including type X cases. Three sample entries, each for types I, II, and X, are listed below: no. 4 for type I, no. 7 for type II (see also [Figure 12](#)), and no. 11 for type X.

```
List of 156 A-minor motifs (types I, II, or X)
 4  type=I A|G-C          0.A69|0.G54,0.C65 WC
```



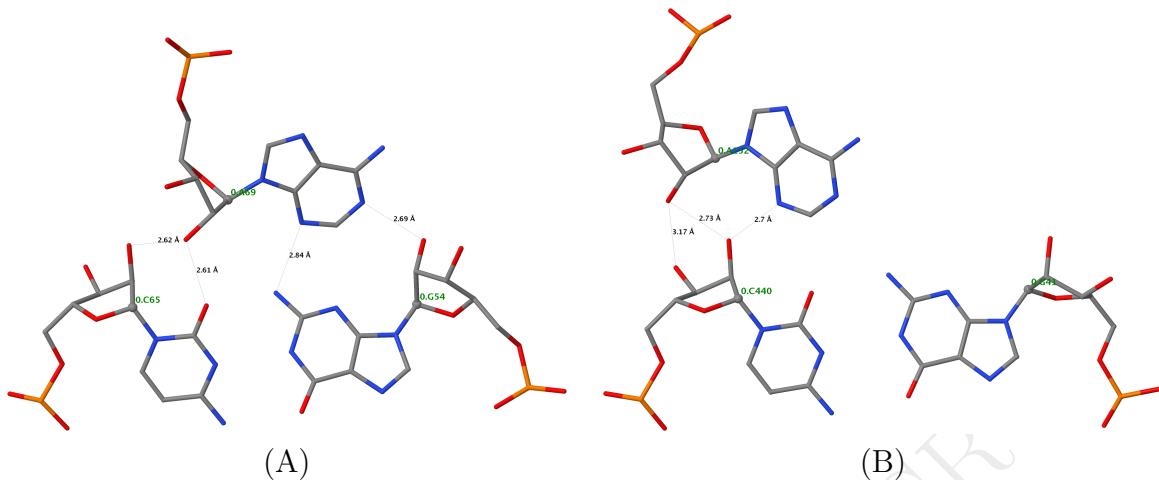
**Figure 11:** One kissing-loop motif identified in 1jj2. Here, one hairpin loop is colored yellow (#14, with nts from 416 to 424 on chain 0), and the other red (#57, with nts from 2442 to 2450 on the same chain). Six base pairs formed between the two loop regions constitute stem #29. The image was produced with Jmol.

```
+0.G54 H-bonds [2]: "N1-O2'(hydroxyl) [2.69], N3-N2(aminoo) [2.84]"
-0.C65 H-bonds [2]: "O2'(hydroxyl)-O2'(hydroxyl) [2.62], O2'(hydroxyl)-O2'(carbonyl)
→ [2.61]"

7 type=II A|G-C      0.A152|0.G41,0.C440 WC
+0.G41 H-bonds [0]: ""
-0.C440 H-bonds [3]: "O2'(hydroxyl)-O3' [3.17], O2'(hydroxyl)-O2'(hydroxyl) [2.73], N3-O2
→ '(hydroxyl) [2.70]"

11 type=X A|U-G      0.A166|0.U919,0.G924 Wobble
-0.U919 H-bonds [0]: ""
+0.G924 H-bonds [3]: "N6(aminoo)-O2'(hydroxyl) [2.90], N6(aminoo)-N3 [3.15], N1-N2(aminoo)
→ [2.98]"
```

For each entry, the type (I, II, or X) is followed by the A-minor motif identity first in



**Figure 12:** Two types of A-minor motifs presumably to be specific to adenine. (A) Type I, where the O2' and N3 atoms of adenine lie inside the minor-groove edge of the canonical base pair; (B) Type II, where the O2' of adenine lies outside but N3 remains inside the minor-groove edge. The images were produced with Jmol.

one-letter shorthand notation (e.g., A|G-C) and then the corresponding full specification of the interacting nts (e.g., 0.A69|0.G54,0.C65). The canonical bp name (WC or wobble) is added at the end. The following two lines list the relative orientation (+ or -)<sup>9</sup> and the H-bonding interactions that each nt of the canonical pairs has with adenine, respectively. Thus for the type I A-minor motif no. 4 listed above, +0.G54 means that 0.G54 and 0.A69 have similar faces (i.e., the dot product of the z-axes of their base reference frames is positive). Conversely, 0.A69 and 0.C65 have opposite faces, so -0.C65 is placed in the next line.

As one can see from the above listing, in type I A-minor motifs the adenine interacts with both nucleotides of the pair. In type II or X, on the other hand, the adenine interacts with only one of paired nucleotides.

### 3.4.3 Ribose zippers

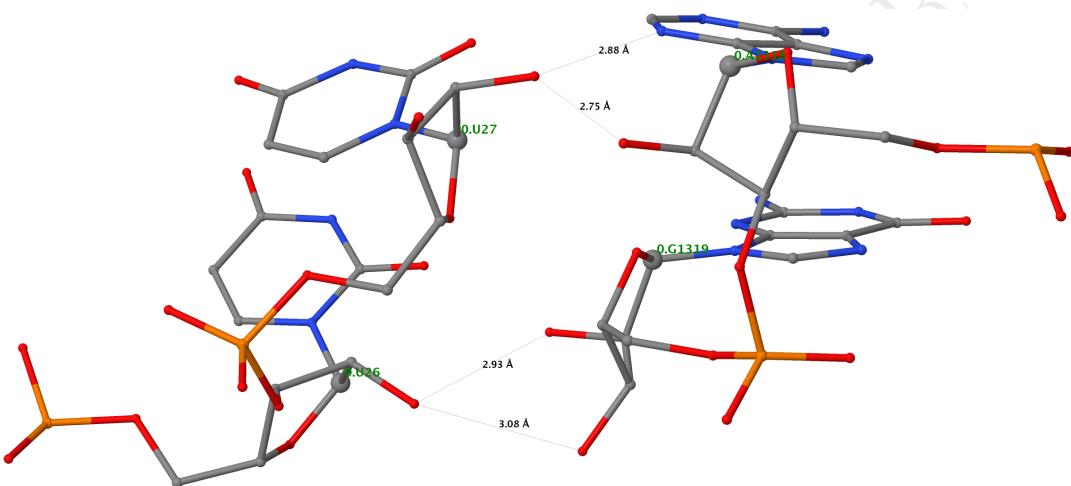
First described in the P4-P6 domain of a group I intron<sup>27</sup>, the ribose zipper is a tertiary interaction that is important for RNA packing. In DSSR, a ribose-zipper motif is defined by two or more (very rare) consecutive H-bonds between ribose 2'-hydroxyl groups from two RNA fragments (see Figure 13).

In 1jj2, DSSR detects a total of 46 ribose zippers, all consisting of  $2 \times 2$  nts. The first twelve zippers are listed below, and the top one is illustrated in Figure 13.

```

List of 46 ribose zippers
 1 nts=4 UUAG O.U26 ,O.U27 ,O.A1318 ,O.G1319
 2 nts=4 ACAC O.A152 ,O.C153 ,O.A439 ,O.C440
 3 nts=4 AACC O.A160 ,O.A161 ,O.C769 ,O.C770
 4 nts=4 AGAU O.A189 ,O.G190 ,O.A204 ,O.U205
 5 nts=4 CGAA O.C208 ,O.G209 ,O.A665 ,O.A666
 6 nts=4 UAAA O.U233 ,O.A234 ,O.A436 ,O.A437
 7 nts=4 AACC O.A242 ,O.A243 ,O.C376 ,O.C377
 8 nts=4 AAUG O.A305 ,O.A306 ,O.U325 ,O.G326
 9 nts=4 GAAC O.G471 ,O.A472 ,O.A773 ,O.C774
10 nts=4 AACC O.A520 ,O.A521 ,O.C637 ,O.C638
11 nts=4 AACC O.A551 ,O.A552 ,O.C1334 ,O.C1335
12 nts=4 AACU O.A565 ,O.A566 ,O.C1263 ,O.U1264

```



**Figure 13:** A sample canonical ribose zipper identified in 1jj2. This motif consists of the U26–U27 dinucleotide in one strand, and A1318–G1319 in the other, both on chain 0. The images were produced with Jmol.

### 3.4.4 Kink turns

The kink-turn (k-turn), first characterized in the large ribosomal subunit of *H. marismortui*<sup>28</sup>, is a widespread structural motif in RNA. The motif contains a sharp kink in the RNA helix, with an asymmetric internal loop flanked by C–G bps on one side and sheared G–A bps on the other. The Lilley laboratory established a systematic nomenclature for nucleotides<sup>29</sup> and maintains a [dedicated database](#) for k-turns.

By default, DSSR defines a k-turn motif as an asymmetric internal loop with at least one sheared G–A pair and a large bending angle in the helical axis, among other criteria. The program detects three types of k-turns: normal, reverse, or else (for possible but undecided

cases). The normal type includes simple (standard or non-standard) and complex k-turns as defined by Lilley<sup>30</sup>, as long as they involve asymmetric internal loops. DSSR finds a total of eight k-turns in 1jj2, the first of which (commonly known as *H. marismortui Kt-7*) is listed below and depicted in Figure 14.

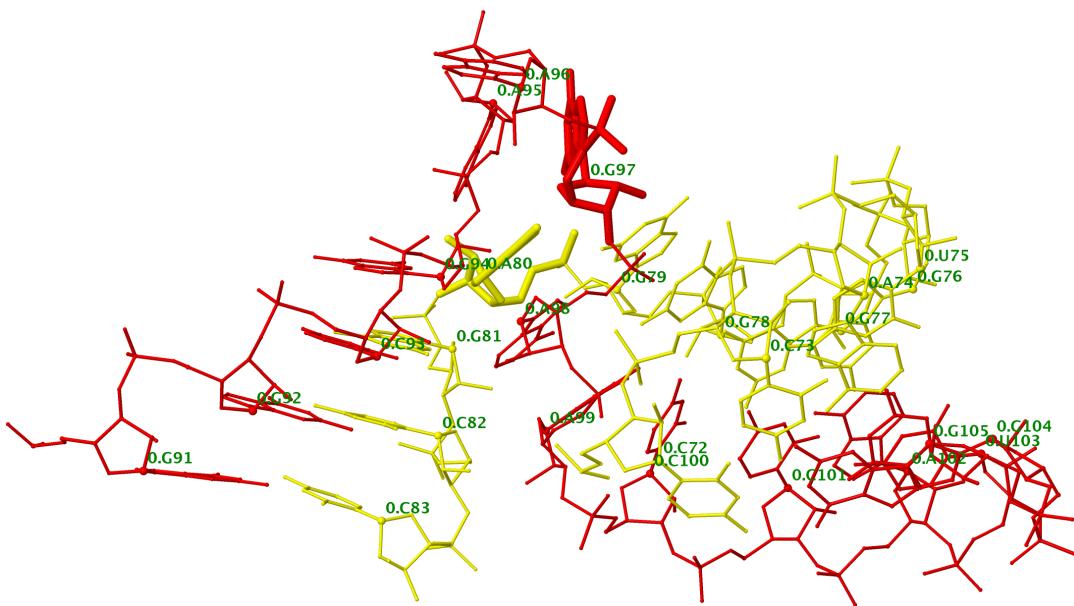
```
1 Normal k-turn; iloop#43; between stems [#11,#10]; bending-angle=54
C#11[CG 0.C93,0.G81] [GA 0.G97,0.A80] NC#10[CG 0.C100,0.G77]
strand1 nts=15; GGCAGAACCAUGG 0.G91,0.G92,0.C93,0.G94,0.A95,0.A96,0.G97,0.A98,0.A99
    ↵ ,0.C100,0.C101,0.A102,0.U103,0.G104,0.G105
strand2 nts=12; CCAUGGGGAGCC 0.C72,0.C73,0.A74,0.U75,0.G76,0.G77,0.G78,0.G79,0.A80,0.
    ↵ G81,0.C82,0.C83
```

- The 1st line means that this is a normal k-turn, derived from internal loop #43 (which is delineated by stems #11 and #10). The bending angle between the two stems is 54°.
- The 2nd line shows that the canonical helix consists of stem #11, with a C–G pair (0.C93 and 0.G81) closing the internal loop at one end. The non-canonical helix contains stem #10, with a C–G pair (0.C100 and 0.G77) closing the internal loop at the other end. The crucial sheared G–A pair is formed by 0.G97 and 0.A80 (highlighted with thick lines in Figure 14). See reference<sup>29</sup> for nomenclature of C and NC helices in k-turns.
- The 3rd and 4th lines list the two strands composing the k-turn.

### 3.5 Identification and characterization of G-quadruplexes

G-quadruplexes (hereafter referred to as G4) are a common type of higher-order DNA and RNA structures formed from G-rich sequences<sup>31–33</sup>. The building block of G4 is a tetrad of guanines in a cyclic ‘planar’ alignment, with four G+G pairs (cW+M type, see Figure 15). A G4 is formed by stacking of G-tetrads and stabilized by cations. G4 structures are polymorphic: the four strands can be parallel or anti-parallel, and loops connecting them can be of different types: lateral (edgewise), diagonal, or propeller (double-chain reversal). Moreover, G4 structures can be intra- or intermolecular, and even contain bulges<sup>34</sup>. Overall, G4 can take a large variety of topologies.

DSSR has a dedicated module that streamlines the analysis, annotations, and visualization of G4 structures. See Figures 15 and 16, and <http://g4.x3dna.org/> (DSSR-G4DB: G-quadruplexes auto-curated with DSSR from the PDB).



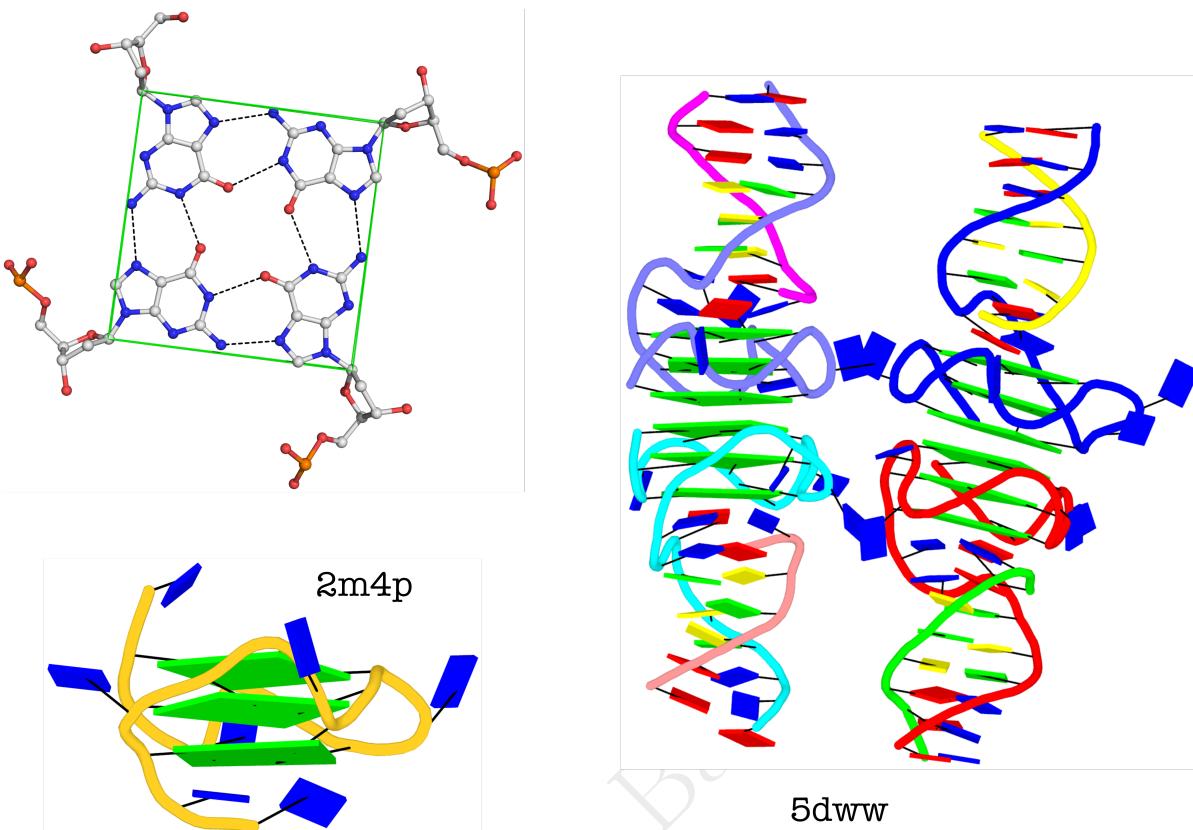
**Figure 14:** A normal k-turn identified in 1jj2, commonly known as *H. marismortui* Kt-7. The two strands are colored yellow and red, respectively, for easy visualization of the backbone trajectory. The crucial sheared G–A pair is highlighted with thick lines. The image was produced with Jmol.

### 3.6 Pseudoknots detection and removal

RNA pseudoknot is defined by WC pairing interactions between bases in a hairpin loop and those outside the corresponding stem. Intuitively, WC pairs in pseudoknotted RNAs are not fully nested, but crossed in 2D notation. Pseudoknots are abundant in RNA structures in the PDB, and are known to play essential functional roles<sup>36</sup>. However, they possess a challenge to many RNA computational tools, e.g., dynamic programming based prediction of 2D structures. DSSR can characterize pseudoknots of arbitrary complexity in a 3D RNA structure. It also has the capability to remove pseudoknots to produce a nested 2D structure in dbn notation.

#### 3.6.1 Higher-order pseudoknots

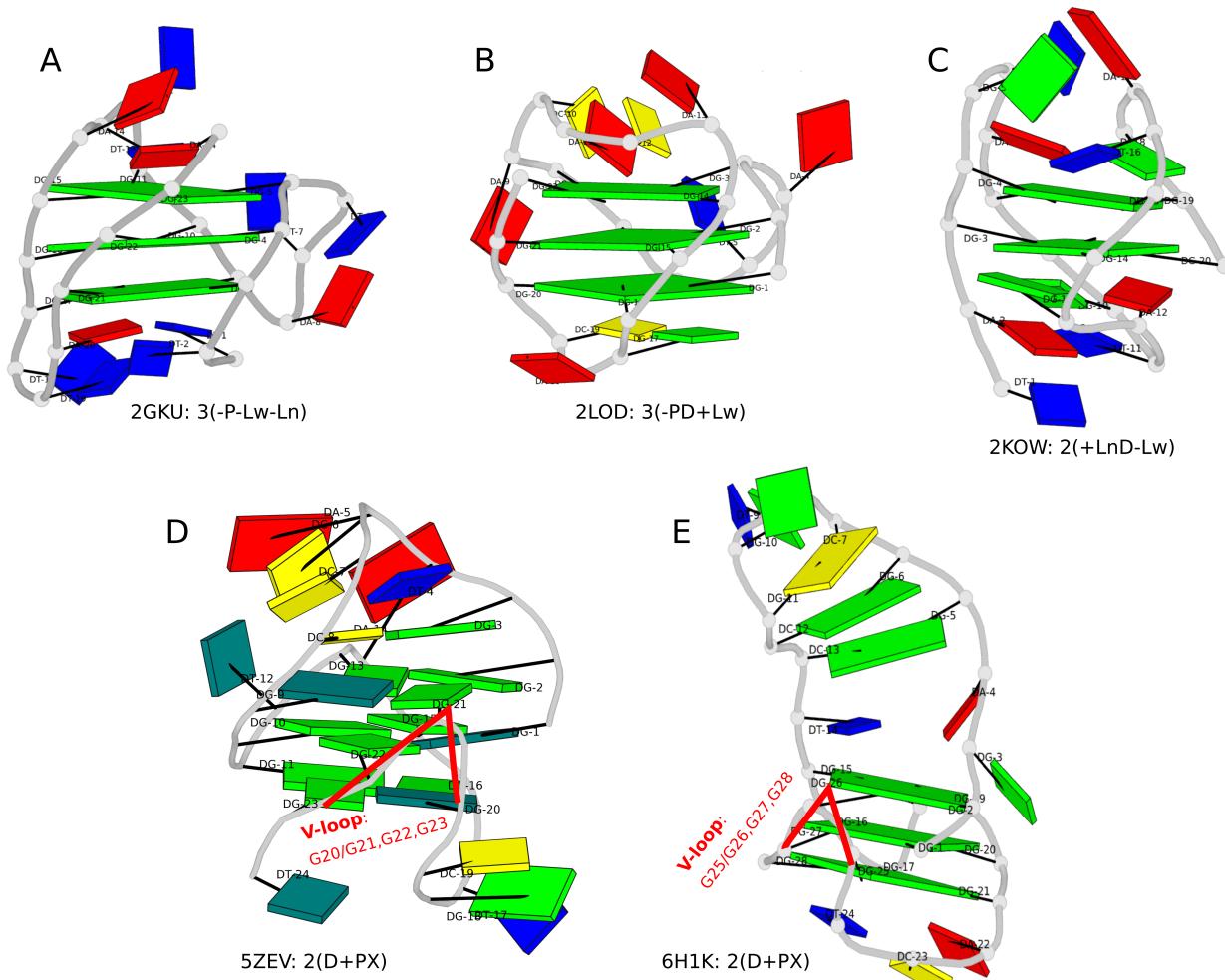
Using PDB entry 1ddy<sup>37</sup> as an example, here is the relevant DSSR output. Note that 1ddy contains four RNA chains (A, C, E, and G), and DSSR handles each correctly. Here WC pairs in the 2nd order pseudoknots are represented as matched curly brackets {}. For even higher-order pseudoknots, the WC pairs are represented by matched angle brackets <>, and upper-lower case letters (Aa, Bb, Cc, etc.)



**Figure 15:** Representative G4 structures. Upper left: atomic structure of G-tetrad, the building block of G4 structures. Here the green 'square' is created by connecting the C1' atoms of the guanosines, and it is used to simplify the representation of G4 structures of PDB entries 2m4p (lower left) and 5dww (right). Note that the asymmetric unit of 5dww contains four biological units, which are coaxially stacked into two columns.

```
This structure contains 2-order pseudoknot
o You may want to run DSSR again with the '--nested' option which removes
pseudoknots to get a fully nested secondary structure representation.

Secondary structures in dot-bracket notation (dbn) as a whole and per chain
>1ddy nts=140 [whole]
GGAACCGGGUGCGCAUAACCACCUUCAGUGCGAGCAA&GGAACCGGGUGCGCAUAACCACCUUCAGUGCGAGCAA&
  ↵ GGAACCGGGUGCGCAUAACCACCUUCAGUGCGAGCAA&GGAACCGGGUGCGCAUAACCACCUUCAGUGCGAGCAA
.....((.{{[[...[[[.]]]..&.....((.{{[[...[[.]]]..&.....((.{{[[...[[.]]]..&.....((.{{[[...[[.]]]..
  ↵ ))....].].]]..&.....((.{{[[...[[.]]]....].].]]..
>1ddy-A #1 nts=35 0.46(2.84) [chain] RNA
GGAACCGGGUGCGCAUAACCACCUUCAGUGCGAGCAA
.....((.{{[[...[[.]]]..].].]]..
>1ddy-C #2 nts=35 0.66(2.88) [chain] RNA
GGAACCGGGUGCGCAUAACCACCUUCAGUGCGAGCAA
.....((.{{[[...[[.]]]..].].]]..
>1ddy-E #3 nts=35 0.72(2.88) [chain] RNA
GGAACCGGGUGCGCAUAACCACCUUCAGUGCGAGCAA
```



**Figure 16:** Five G4 structures with revised topological descriptors<sup>35</sup>. It is worth noting that the descriptors for PDB entries 2gku and 2lod as originally reported in Figure 1 of Dvorkin *et al.*<sup>35</sup> are incorrect. Applying DSSR consistently can correct errors from leading experts who proposed the descriptors in the first place. The structural features were automatically derived by DSSR. The images were rendered with PyMOL.

```
.....((.{{[...[[()]]}})...].).]..  
>1ddy-G #4 nts=35 0.54(2.87) [chain] RNA  
GGAACCGGUGCGCAUAACCACCUUCAGUGCGAGCAA  
.....((.{{[...[[()]]}})...].).]..
```

### 3.6.2 Pseudoknot removal

A closely related issue is pseudoknot removal, a topic nicely summarized by Smit *et al*<sup>38</sup>. The **--nested** (or abbreviated form **--nest**) option can be used to remove pseudoknots to get a nested 2d structure in dbn. Using PDB entry 1ddy as an example, the relevant DSSR output with option **--nested** is as follows:

```
This structure contains 2-order pseudoknot
o You have chosen to remove the pseudoknots with the '--nested' option so
only a fully nested secondary structure representation remains.

Secondary structures in dot-bracket notation (dbn) as a whole and per chain
>1ddy nts=140 [whole]
GGAACCGGUGCGCAUAACCACCUUCAGUGCGAGCAA&GGAACCGGUGCGCAUAACCACCUUCAGUGCGAGCAA&
    ↪ GGAACCGGUGCGCAUAACCACCUUCAGUGCGAGCAA&GGAACCGGUGCGCAUAACCACCUUCAGUGCGAGCAA
.....(((.....))).....&.....(((.....))).....&.....(((.....)))
    ↪ )). ....&.....(((.....))). .....
>1ddy-A #1 nts=35 0.46(2.84) [chain] RNA
GGAACCGGUGCGCAUAACCACCUUCAGUGCGAGCAA
.....(((.....))). .....
>1ddy-C #2 nts=35 0.66(2.88) [chain] RNA
GGAACCGGUGCGCAUAACCACCUUCAGUGCGAGCAA
.....(((.....))). .....
>1ddy-E #3 nts=35 0.72(2.88) [chain] RNA
GGAACCGGUGCGCAUAACCACCUUCAGUGCGAGCAA
.....(((.....))). .....
>1ddy-G #4 nts=35 0.54(2.87) [chain] RNA
GGAACCGGUGCGCAUAACCACCUUCAGUGCGAGCAA
.....(((.....))). .....
```

### 3.7 The --more option

The `--more` option triggers additional parameters for several sections, including base pairs, helices, and stems. Since helices/stems share the same format for the added parameters, only an example from stems is shown. In what follows, the excerpted portions are based on PDB entry [1msy](#) with the following DSSR command:

```
x3dna-dssr -i=1msy.pdb --more -o=1msy-more.out
```

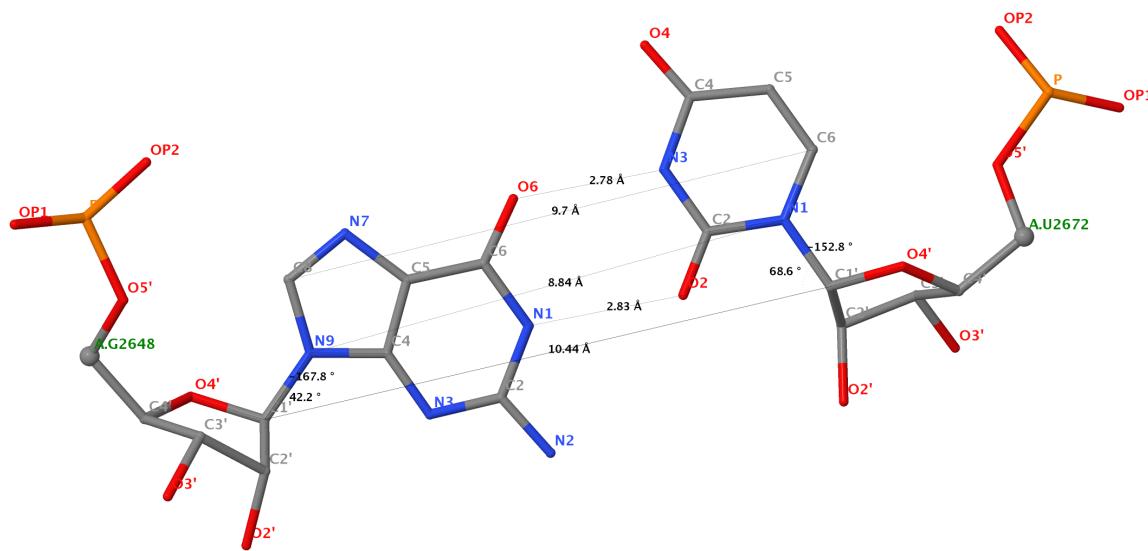
### 3.7.1 Extra characterization of base pairs

The G–U wobble pair formed by A.G2648 and A.U2672 (the second bp on [Page 13](#)) is used as an example of the additional information provided about bp geometry and H-bonding with the `--more` option (see [Figure 17](#)).

```

1   2 A.G2648      A.U2672      G-U Wobble      28-XXVIII cWW  cW-W
2   [-167.8(anti) ~C3'-endo lambda=42.1] [-152.8(anti) ~C3'-endo lambda=68.6]
3   d(C1'-C1')=10.44 d(N1-N9)=8.84 d(C6-C8)=9.70 tor(C1'-N1-N9-C1')=-8.1
4   H-bonds [2]: "O6(carbonyl)-N3(imino)[2.78], N1(imino)-O2(carbonyl)[2.83]"
5   interBase-angle=9  Simple-bpParams: Shear=-2.44 Stretch=-0.01 Buckle=2.7 Propeller
6   ← =-8.6
bp-pars: [-2.37    -0.60    0.11     4.67     -7.75    -2.95]
```

- Line no. 1: specification of the bp, as shown previously on [Page 13](#).
- Line no. 2: the first bracket `[-167.8(anti) C3'-endo lambda=42.1]` corresponds to A.G2648, and it contains three items:  $-167.8(\text{anti})$  is the  $\chi$  torsion angle formed by  $O4'-C1'-N9-C4$ ,  $C3'$ -endo is the sugar pucker (as is the norm for RNA), and  $\lambda$  (lambda) is the angle  $N9-C1'-C1'$  (A.U2672). The second bracket corresponds to A.U2672: `[-152.8(anti) C3'-endo lambda=68.6]` with similar meanings for parameters, except that  $\chi$  is defined by  $O4'-C1'-N1-C2$ , and  $\lambda$  is the angle  $N1-C1'-C1'$  (A.G2648).
- Line no. 3: lengths of three virtual bonds ( $C1'-C1'$ ,  $N1-N9$ ,  $C6-C8$ ), and the virtual torsion angle ( $C1'-N1-N9-C1'$ ). Note here N1 and N9 are general terms, referring to either N1 of pyrimidines (Y: C/U/T) or N9 of purines (R: A/G) as appropriate. Similar conventions apply for the labeling of C6/C8.
- Line no. 4: detailed H-bonding information (atom names, types, and H-bond distances in square brackets).
- Line no. 5: inter-base-angle ( $9^\circ$ ), and a set of four ‘simple’ bp parameters<sup>11</sup> (shear, stretch, buckle and propeller) which are easier to understand than the six rigorous rigid-body parameters (listed below) for non-canonical pairs, especially when opening is  $\sim 180^\circ$ .
- Line no. 6: the six rigid-body bp parameters in the order of shear, stretch, stagger, buckle, propeller, and opening (see [Figure 7](#)).



**Figure 17:** Molecular image of the G–U wobble pair formed by A.G2648 and A.U2672 in 1msy, labeled with additional parameters (bond lengths, angles, and torsions). The image was produced with Jmol.

### 3.7.2 Orientation of helices/stems

The additional output contains information about the best-fitted linear helical axis of a helix/stem, derived using a combination of equivalent C1' and RN9/YN1 atom pairs along each strand<sup>39</sup>. The result for the 1msy stem listed on [Page 22](#) is shown below.

```

helical-rise: 2.60(0.18)
helical-radius: 9.12(0.79)
helical-axis: -0.776 -0.167 -0.608
  point-one: 24.637 21.051 22.830
  point-two: 16.686 19.344 16.602

```

- The **helical-rise** line, with numbers 2.60(0.18), represents the average helical rise (2.60) between successive nucleotides and its standard deviation (sd, 0.18) in Å. For a perfectly regular DNA/RNA helix, the sd would be zero. RNA models generated with the DSSR fiber module, for example, are characterized by numbers 2.55(0.00). In DSSR, the sd is used to determine if a helix/stem is strongly curved, with a default cutoff of 0.6 Å. If the sd for a helix/stem is over the cutoff (as for the 1msy helix listed on [Page 21](#)), a \* is appended, serving as a reminder that the best-fitted linear helical axis may not be meaningful.

- The **helical-radius** line, with numbers  $9.12(0.79)$ , gives the average and sd of the perpendicular distances from phosphorus atoms (of both strands) to the helical axis. Typically, the mean radius is around 9.2 Å for double helical DNA or RNA.
- The **helical-axis** line, with three numbers  $-0.776 -0.167 -0.608$ , provides the normalized helical axis vector (in the original coordinate frame). This vector can be used to calculate DNA bending angles (as in DNA-protein complexes), or to quantify the relative orientation between any two fairly straight helices/stems<sup>10</sup>.
- The lines starting with **point-one** and **point-two** designate the end points (in the original coordinate frame) of the helical axis of the helix/stem. The two points can be added to the original PDB coordinates file for visualization of the helical axis or for rendering of helical regions as cylinders<sup>1,10</sup>.

### 3.7.3 Base-pair morphology parameters for helices/stems

The 2nd dinucleotide step consisting of pairs C–G (A.C2649 with A.G2671) and U–A (A.U2650 with A.A2670) in the stems section of 1msy is used here as an example. Six extra lines are enabled by the **--more** option. For the definitions of base and step parameters, see Figure 7.

|   |            |         |        |             |       |               |
|---|------------|---------|--------|-------------|-------|---------------|
| 2 | A.C2649    | A.G2671 | C-G WC | 19-XIX      | cWW   | cW-W          |
|   | bp1-pars:  | [0.09   | -0.17  | 0.02        | 9.60  | -15.93 -2.31] |
|   | step-pars: | [0.79   | -1.42  | 3.25        | -0.33 | 7.26 33.48]   |
|   | heli-pars: | [-3.49  | -1.39  | 2.88        | 12.42 | 0.57 34.24]   |
|   | bp2-pars:  | [0.11   | -0.11  | 0.25        | 4.92  | -16.45 6.33]  |
|   | C1'-based: |         |        | rise=3.25   |       | twist=33.33   |
|   | C1'-based: |         |        | h-rise=2.87 |       | h-twist=34.09 |
| 3 | A.U2650    | A.A2670 | U-A WC | 20-XX       | cWW   | cW-W          |

- The **bp1-pars** line lists the six bp parameters in the order of shear, stretch, stagger, buckle, propeller, and opening for the C–G pair (A.C2649 with A.G2671).
- The **step-pars** line lists the six step parameters in the order of shift, slide, rise, tilt, roll, and twist for the dinucleotide step between pairs C–G (A.C2649 with A.G2671) and U–A (A.U2650 with A.A2670), calculated based on a middle-step frame<sup>9</sup>.
- The **heli-pars** line lists the six helical parameters in the order of *x*-displacement, *y*-displacement, helical rise, inclination, tip, and helical twist for the aforementioned dinucleotide step, calculated based on a middle-helical frame<sup>9</sup>.

- The **bp2-pars** line lists the six bp parameters in the order of shear, stretch, stagger, buckle, propeller, and opening for the U–A pair (A.U2650 with A.A2670).
- The 1st C1'-based line provides rise and twist derived using the two consecutive C1'-C1' vectors, each defined by a pair in the dinucleotide step. These two parameters are related to the middle-step frame used to calculate the six step parameters listed on the **step-pars** line.
- The 2nd C1'-based line provides helical rise and helical twist derived using the two consecutive C1'-C1' vectors, each defined by a pair in the dinucleotide step. These two parameters are related to the middle-helical frame used to calculate the six helical parameters listed on **heli-pars** line.

## 3.8 The --non-pair option

With the **--non-pair** option, DSSR identifies H-bonding and base-stacking interactions between two nts, excluding those duos that already form a pair. For 1msy, the running command and relevant results are given below.

```
# ~/Luxes/src/academic/DSSR2-src/x3dna-dssr -i=1msy.pdb --non-pair -o=1msy-nonpair.out

List of 30 non-pairing interactions
 1 A.U2647  A.G2648  stacking: 1.0(0.5)--pm(>,forward) interBase-angle=6 connected min-baseDist=3.26
 2 A.G2648  A.C2649  stacking: 7.3(4.6)--pm(>,forward) interBase-angle=5 connected min-baseDist=3.30
 3 A.G2648  A.G2673  stacking: 2.0(0.2)--mm(<,outward) interBase-angle=2 min-baseDist=3.28
 4 A.C2649  A.U2650  stacking: 2.8(1.1)--pm(>,forward) interBase-angle=9 connected min-baseDist=3.09
 5 A.U2650  A.C2651  stacking: 0.6(0.0)--pm(>,forward) interBase-angle=7 connected min-baseDist=3.30
 6 A.C2651  A.C2652  stacking: 0.5(0.1)--pm(>,forward) interBase-angle=12 connected min-baseDist=3.30
 7 A.C2652  A.U2653  stacking: 5.2(2.6)--pm(>,forward) interBase-angle=13 connected min-baseDist=3.43
 8 A.C2652  A.G2669  stacking: 0.2(0.0)--mm(<,outward) interBase-angle=7 min-baseDist=3.22
 9 A.U2653  A.A2654  stacking: 3.3(2.0)--pp(<,inward) interBase-angle=13 H-bonds[1]: "OP2-O2'(hydroxyl)[2.62]"
    ↪ connected min-baseDist=3.23
10 A.A2654  A.U2656  stacking: 3.7(1.1)--mm(<,outward) interBase-angle=1 H-bonds[1]: "O4'*O4'[3.05]" min-baseDist=3.45
11 A.G2655  A.G2664  stacking: 4.4(2.2)--pp(<,inward) interBase-angle=10 H-bonds[2]: "O2'(hydroxyl)-O6(carbonyl)"
    ↪ [3.09], O2'(hydroxyl)-N1(imino)[3.34]" min-baseDist=3.37
12 A.G2655  A.A2665  interBase-angle=21 H-bonds[2]: "N1(imino)-OP2[2.77], N2(amino)-O5'[2.89]" min-baseDist=4.79
13 A.U2656  A.G2664  interBase-angle=7 H-bonds[2]: "OP2-N1(imino)[3.04], OP2-N2(amino)[2.94]" min-baseDist=3.36
14 A.A2657  A.C2658  stacking: 6.7(2.6)--pm(>,forward) interBase-angle=4 connected min-baseDist=3.46
15 A.A2657  A.A2665  stacking: 3.7(3.3)--mm(<,outward) interBase-angle=11 min-baseDist=3.29
16 A.C2658  A.G2659  stacking: 0.4(0.1)--pm(>,forward) interBase-angle=10 connected min-baseDist=3.34
17 A.G2659  A.A2661  interBase-angle=31 H-bonds[2]: "O2'(hydroxyl)-N7[2.60], O2'(hydroxyl)-N6(amino)[3.26]" min-baseDist
    ↪ =3.97
18 A.G2659  A.G2663  stacking: 3.9(1.2)--mm(<,outward) interBase-angle=4 min-baseDist=3.35
19 A.U2660  A.A2661  stacking: 7.5(4.2)--pm(>,forward) interBase-angle=17 connected min-baseDist=3.26
20 A.A2661  A.A2662  stacking: 6.3(4.4)--pm(>,forward) interBase-angle=19 connected min-baseDist=3.38
21 A.G2663  A.G2664  stacking: 2.7(0.6)--pm(>,forward) interBase-angle=8 connected min-baseDist=3.38
22 A.G2664  A.A2665  interBase-angle=14 H-bonds[1]: "O2'(hydroxyl)-O4'[2.75]" connected min-baseDist=5.83
23 A.A2665  A.C2666  stacking: 1.6(1.1)--pm(>,forward) interBase-angle=10 connected min-baseDist=3.18
24 A.C2666  A.C2667  stacking: 4.3(2.1)--pm(>,forward) interBase-angle=8 connected min-baseDist=3.35
25 A.C2667  A.G2668  stacking: 3.1(1.0)--pm(>,forward) interBase-angle=7 connected min-baseDist=3.38
26 A.G2668  A.G2669  stacking: 4.3(3.0)--pm(>,forward) interBase-angle=4 connected min-baseDist=3.28
```

```

27 A.G2669  A.A2670  stacking: 4.3(2.9)--pm(>,forward) interBase-angle=4 connected min-baseDist=3.29
28 A.A2670  A.G2671  stacking: 1.5(1.5)--pm(>,forward) interBase-angle=6 connected min-baseDist=3.24
29 A.G2671  A.U2672  stacking: 7.4(4.0)--pm(>,forward) interBase-angle=10 connected min-baseDist=3.22
30 A.U2672  A.G2673  interBase-angle=11 H-bonds [1]: "O2'(hydroxyl)-O4'[3.37]" connected min-baseDist=3.61

```

Following 3DNA<sup>9</sup>, DSSR quantifies base-stacking interactions by the area (in Å<sup>2</sup>) of the overlapped polygon defined by the two bases of the interacting nts, where the base atoms are projected onto the mean base plane. In the output file, values in parentheses measure the overlap areas of base ring atoms only, and those outside parentheses include exocyclic atoms.

In DSSR, base-stacking interactions are classified into one of the following four categories: (i) `pm(>,forward)`, (ii) `mp(<,backward)`, (iii) `mm(<,outward)`, and (iv) `pp(><,inward)`. Here `p` and `m` represent the plus and minus faces of the base ring, as defined by the direction of the *z*-axis of the standard base reference frame<sup>4</sup> (see [Figure 1](#)). The symbols (`>`, `<`, `<`, and `><`) follow [Major et al.](#), except that `pm(>)` is called *forward* instead of upward, and `mp(<)` is named *backward* instead of downward<sup>40</sup>.

The term `interBase-angle` represents the inter-base angle; a value close to zero means that the two bases are nearly parallel. The term `connected`, if present, means that the two nucleotides are connected by a covalent phosphodiester linkage. The term `min-baseDist` gives the minimum distance between base atoms.

The H-bonding information should be self-explanatory, except for a note on convention: when a pair of donor/donor or acceptor/acceptor atoms fulfills the geometry-based H-bond definition in DSSR, the `*` symbol is used instead of `-` to connect the atoms (e.g., entry no. 10: `O4'*O4'[3.05]`).

### 3.9 The --json option

The `--json` option directs DSSR to output results in the standard [JSON](#) data exchange format. The single JSON file contains numerous DSSR-derived structural features, including those in the default output, backbone torsions, and H-bonds.

The JSON output makes DSSR readily interoperable with other (bioinformatics) tools. Using tRNA<sup>Phe</sup> 1ehz as an example, let's go over some use cases with the `jq` command-line JSON processor.

```

x3dna-dssr -i=1ehz.pdb --json -o=1ehz-dssr.json
jq . 1ehz-dssr.json # reformatted for pretty output
x3dna-dssr -i=1ehz.pdb --json | jq . # the above two steps combined

```

With file `1ehz-dssr.json` in hand, one can easily extract DSSR-derived structural features of interest:

```
jq .pairs 1ehz-dssr.json    # list of 34 pairs
jq .multiplets 1ehz-dssr.json  # list of 4 base triplets
jq .hbonds 1ehz-dssr.json   # list of hydrogen bonds
jq .helices 1ehz-dssr.json
jq .stems 1ehz-dssr.json
  # list of nucleotide parameters, including torsion angles and suites
jq .nts 1ehz-dssr.json
  # list of 14 modified nucleotides
jq '.nts[] | select(.is_modified)' 1ehz-dssr.json
  # select nucleotide id, delta torsion, sugar puckering and cluster of suite name
jq '.nts[] | {nt_id, delta, puckering, cluster}' 1ehz-dssr.json
  # same selection as above, but in 'comma-separated-values' (csv) format
jq -r '.nts[] | [.nt_id, .delta, .puckering, .cluster] | @csv' 1ehz-dssr.json
```

Here is the result of running `jq` to select multiplets:

```
# jq .multiplets 1ehz-dssr.json
[
  {
    "index": 1,
    "num_nts": 3,
    "nts_short": "UAA",
    "nts_long": "A.U8,A.A14,A.A21"
  },
  {
    "index": 2,
    "num_nts": 3,
    "nts_short": "AUA",
    "nts_long": "A.A9,A.U12,A.A23"
  },
  {
    "index": 3,
    "num_nts": 3,
    "nts_short": "gCG",
    "nts_long": "A.2MG10,A.C25,A.G45"
  },
  {
    "index": 4,
    "num_nts": 3,
    "nts_short": "CGg",
    "nts_long": "A.C13,A.G22,A.7MG46"
  }
]
```

### 3.10 The --pair-only option

DSSR provides far more features than a typical user normally need. The `--pair-only` option directs DSSR to output only base-pairing information, the most fundamental aspect in DNA/RNA structural bioinformatics. It can be combined with the `--more` or `--json` option.

DSSR runs approximately 10 times faster with the `--pair-only` option than the default.

### 3.11 The `--nmr` option

The `--nmr` option has been introduced for the analysis of an ensemble of NMR structures, as deposited in the PDB. The input file can be in the `.pdb` format where each model is delineated with MODEL/ENDMDL, or the `.cif` format where each ATOM/HETATM record is associated with a model number.

Using PDB entry [2n2d](#) as an example, two simple usages are as follows:

```
x3dna-dssr -i=2n2d.pdb --nmr -o=2n2d-model.out
x3dna-dssr -i=2n2d.pdb --nmr --json -o=2n2d.json
jq '.models[].parameters.num_Gtetrad' 2n2d.json
```

The top-level skeleton of the JSON output is shown below. Note that each member of the `models` array contains three items: an auto-incremental `index` (1 to the number of models), the actual `model` number, and the `parameters` object which corresponds to the JSON output if the model is analyzed separately. Normally, `index` and `model` match each other, as is the case for 2n2d. However, it is conceivable that models do not start from one, or the numbers are not continuous. In such cases, they will no longer have the same value for each entry.

```
{
  "input_file": "2n2d.pdb",
  "num_models": 10,
  "models": [
    {
      "index": 1,
      "model": 1,
      "parameters": { ... }
    },
    ...
    {
      "index": 10,
      "model": 10,
      "parameters": { ... }
    }
  ],
  "start_at": "Thu Aug 20 00:06:03 2020",
  "finish_at": "Thu Aug 20 00:06:04 2020",
  "time_used": "00:00:00:01"
}
```

The `--json` option makes it easy to parse the output of multiple models pragmatically. In addition to NMR structures, trajectories from molecular dynamics (MD) simulations can also be processed. As noted above, DSSR takes only the standard `.pdb` or `.cif` format for

input. Thus, proprietary binary formats for trajectories from popular MD packages need to be converted to a standard format. The combination of `--nmr` and `--json` renders DSSR easily accessible to the MD community. The *pro* version of DSSR has a far more efficient engine than the basic one for pragmatic analyses of MD trajectories.

### 3.12 The `--get-hbond` option

H-bonding interactions are crucial for defining RNA secondary and tertiary structures. DSSR contains a geometrically based algorithm for identifying H-bonds in nucleic-acid or protein structures specified in `.pdb` or `.cif` format. Over the years, the method has been continuously refined, and it has served its purpose quite well. This functionality is also directly available through the `--get-hbond` option. By default, the output only includes H-bonds involving nucleotides.

```
x3dna-dssr -i=1msy.pdb --get-hbond -o=1msy-hbonds.txt
```

Running the above command, the output for 1msy is as listed below. The 1st line is the header (“# H-bonds in ‘1msy.pdb’ identified by DSSR …”). The 2nd line provides the total number of H-bonds (40) identified in the structure. Afterwards, each line consists of eight space-delimited columns used to characterize a specific H-bond. Using the first H-bond as an example, the meaning of each of the eight columns is detailed below:

1. The serial number (15), as denoted in the `.pdb` or `.cif` file, of the 1st atom of the H-bond.
2. The serial number (578) of the 2nd atom of the H-bond.
3. The H-bond index (#1), a serial number from 1 to the total number of H-bonds.
4. A one-letter symbol showing the atom-pair type (p) of the H-bond. It is p for a donor-acceptor atom pair; o for a donor/acceptor (such as the 2'-hydroxyl oxygen) with any other atom; x for a donor-donor or acceptor-acceptor pair (as in #17); ? if the donor/acceptor status of any H-bond atom is unknown.
5. Distance between donor/acceptor atoms in Å (2.768).
6. Elemental symbols of the two atoms involved in the H-bond (O:N).

7. Identifier of the 1st H-bonded atom (O4@A.U2647).

8. Identifier of the 2nd H-bonded atom (N1@A.G2673).

```
# H-bonds in '1msy.pdb' identified by DSSR version 2.0
40
 15  578 #1      p    2.768 O:N 04@A.U2647 N1@A.G2673
 35  555 #2      p    2.776 O:N 06@A.G2648 N3@A.U2672
 36  554 #3      p    2.826 N:O N1@A.G2648 O2@A.U2672
 55  537 #4      p    2.965 O:N 02@A.C2649 N2@A.G2671
 56  535 #5      p    2.836 N:N N3@A.C2649 N1@A.G2671
 58  534 #6      p    2.769 N:O N4@A.C2649 O6@A.G2671
 76  513 #7      p    2.806 N:N N3@A.U2650 N1@A.A2670
 78  512 #8      p    3.129 O:N 04@A.U2650 N6@A.A2670
 95  492 #9      p    2.703 O:N 02@A.C2651 N2@A.G2669
 96  490 #10     p    2.853 N:N N3@A.C2651 N1@A.G2669
 98  489 #11     p    2.987 N:O N4@A.C2651 O6@A.G2669
115  466 #12     p    2.817 O:N 02@A.C2652 N2@A.G2668
116  464 #13     p    2.907 N:N N3@A.C2652 N1@A.G2668
118  463 #14     p    2.897 N:O N4@A.C2652 O6@A.G2668
123  151 #15     o    2.622 O:O OP2@A.U2653 O2'@A.A2654
135  443 #16     p    2.898 O:N 02@A.U2653 N4@A.C2667
147  192 #17     x    3.054 O:O O4'@A.A2654 O4'@A.U2656
158  408 #18     p    2.960 N:O N6@A.A2654 OP2@A.C2666
173  188 #19     o    2.923 O:O O2'@A.G2655 OP2@A.U2656
173  378 #20     o    3.093 O:O O2'@A.G2655 O6@A.G2664
173  379 #21     o    3.343 O:N O2'@A.G2655 N1@A.G2664
181  386 #22     p    2.768 N:O N1@A.G2655 OP2@A.A2665
183  203 #23     p    2.754 N:O N2@A.G2655 O4@A.U2656
183  387 #24     p    2.887 N:O N2@A.G2655 O5'@A.A2665
188  379 #25     p    3.044 O:N OP2@A.U2656 N1@A.G2664
188  381 #26     p    2.944 O:N OP2@A.U2656 N2@A.G2664
200  401 #27     p    3.122 O:N 02@A.U2656 N6@A.A2665
201  398 #28     p    2.759 N:N N3@A.U2656 N7@A.A2665
220  381 #29     p    3.035 N:N N7@A.A2657 N2@A.G2664
223  371 #30     o    2.963 N:O N6@A.A2657 O2'@A.G2664
223  382 #31     p    3.039 N:N N6@A.A2657 N3@A.G2664
242  358 #32     p    2.821 O:N 02@A.C2658 N2@A.G2663
243  356 #33     p    2.890 N:N N3@A.C2658 N1@A.G2663
245  355 #34     p    2.887 N:O N4@A.C2658 O6@A.G2663
258  305 #35     o    2.604 O:N O2'@A.G2659 N7@A.A2661
258  308 #36     o    3.264 O:N O2'@A.G2659 N6@A.A2661
268  315 #37     p    2.973 N:O N2@A.G2659 OP2@A.A2662
268  327 #38     p    2.864 N:N N2@A.G2659 N7@A.A2662
371  390 #39     o    2.751 O:O O2'@A.G2664 O4'@A.A2665
550  566 #40     o    3.372 O:O O2'@A.U2672 O4'@A.G2673
```

In its default settings, DSSR detects 117 H-bonds for 1ehz and 5,809 H-bonds for 1jj2 among nucleotides, respectively. By including `protein` (or `amino`, `aa`) in the option, as in `--get-hbond=with-aa`, the program also outputs H-bonds in proteins or at the RNA/protein interfaces. For 1jj2, DSSR now detects a total of 10,583 H-bonds.

Overall, the H-bonding identification method in DSSR is robust and efficient, extensively tested with nucleic-acid-containing structures in the PDB. While there exist dedicated tools for finding H-bonds, such as HBPLUS or HBexplore, DSSR may well serve as a pragmatic

tool for most applications.

### 3.13 The --idstr option

By default, the identifier for nts in DSSR looks like A.U2647. DSSR offers two additional variations, `--idstr=long` and `--idstr=short`, to serve different purposes. Shown below are the listings for a 1msy pair, with three `--idstr` variations.

```
# without the --idstr option (default)
 1 A.U2647      A.G2673      U-G --
                   --          cWW  cW-W
# with the --idstr=long option
 1 ..A.U.2647.  ..A.G.2673.  U-G --
                   --          cWW  cW-W
# with the --idstr=short option
 1 U2647        G2673       U-G --
                   --          cWW  cW-W
```

The long id string consists of six components, which are put in strict order, as given below. A complete example—with model number 1, no segid, chain id C, residue name A (for adenosine), sequence number 24, and insertion code L—is 1..C.A.24.L. The corresponding short form is A24.

```
model-number.segid.chain-id.residue-name.residue-number.insertion-code
```

### 3.14 The --symmetry option

By default, DSSR only analyzes the first structure (model) in a given .pdb for .cif file. For x-ray crystal structures, the asymmetric unit may contain only a fraction of the biological unit. As an example, the asymmetric unit of PDB entry 4ms9 is single-stranded (4ms9.pdb), while the biological unit is a double helix (4ms9.pdb1).

Running DSSR on 4ms9.pdb, or 4ms9.pdb1 with default settings, finds no base pairs. In such cases, one needs to use the `--symmetry` (short form `--symm`) option as shown below for the desired result:

```
x3dna-dssr -i=4ms9.pdb1 --symm
```

DSSR just reads ATOM/HETATM records, as provided in the coordinates file. It does not perform auto-expansion of an asymmetric unit into biological unit based on crystallographic symmetric information that may exist in the .pdb or .cif file. Users must supply the biological unit file (4ms9.pdb1) and the specify the `--symm` option for the desired result.

### 3.15 The --prefix option

By default, DSSR auxiliary output files are prefixed with `dssr`, as in `dssr-pairs.pdb`. The fixed generic names are overwritten on repeated DSSR runs in a directory. With the `--prefix=text` option, the auxiliary output files will be prefixed by `text`. For example, with the following command, the auxiliary output files will be named `1ehz-pairs.pdb`, etc.

```
x3dna-dssr -i=1ehz.pdb -o=1ehz.out --prefix=1ehz
```

## 4 Visualization features

DSSR can be easily incorporated into other structural bioinformatics pipelines. Over the years, I have been fortunate to collaborate with Robert Hanson and Thomas Holder to connect DSSR to Jmol and PyMOL, respectively. These integrations led to two peer-reviewed articles, “DSSR-enhanced visualization of nucleic acid structures in Jmol”<sup>2</sup> and “DSSR-enabled innovative schematics of 3D nucleic acid structures with PyMOL”<sup>3</sup>, both published in *Nucleic Acids Research*. They also exemplify the critical roles that a domain-specific analysis engine may play in general-purpose molecular visualization tools.

The DSSR-Jmol integration excels in its SQL-like, flexible searching capability of structural features. The DSSR-PyMOL integration, on the other hand, stands out for the appealing cartoon-block schematics it brings.

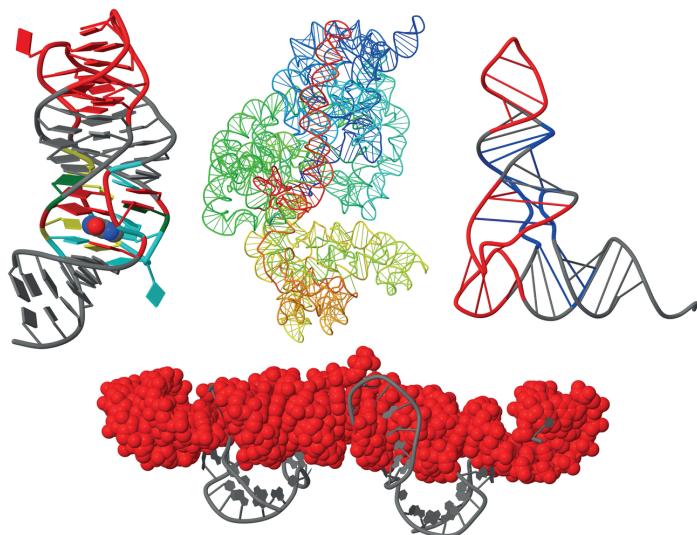
### 4.1 DSSR-Jmol integration

The “DSSR-Jmol integration” section on the 3DNA Forum contains scripts and data files for reproducing reported results, including the graphical abstract and the cover image (Figure 18).

Overall, the DSSR-Jmol integration makes salient features of DSSR readily accessible via Jmol/JSmol, as demonstrated at the website <http://jmol.x3dna.org>. This work fills a gap in RNA structural bioinformatics. It enables deep analyses and SQL-like queries of RNA structural characteristics, interactively.

PRINT ISSN: 0305-1048  
ONLINE ISSN: 1362-4962

# Nucleic Acids Research

VOLUME 45 WEB SERVER ISSUE JULY 3, 2017  
<https://academic.oup.com/nar>OXFORD  
UNIVERSITY PRESS

Open Access

No barriers to access – all articles freely available online



**Figure 18:** 3D interactive visualization of selected RNA structural features enabled by the DSSR-Jmol integration (<http://jmol.x3dna.org>). Clockwise from upper left: Structure of the xpt-pbuX guanine riboswitch in complex with hypoxanthine (PDB entry 4fe5) in 'base blocks' representation. The three-way junction loop encompassing the metabolite (in space-filling representation) is color-coded by base identity: A, red; C, yellow; G, green; U, cyan. The loop-loop interaction at the top is highlighted in red. Structure of the *T. thermophilus* 30S ribosomal subunit in complex with antibiotics (PDB entry 1fjg) in step diagram. The 16S rRNA is color-coded in spectrum with the 5'-end in blue and the 3'-end in red (upper middle). Structure of the classic L-shaped yeast tRNA<sup>Phe</sup> (PDB entry 1ehz) in step diagram, with the three hairpin loops highlighted in red and the four-way junction loop in blue (upper right corner). Structure of the Pistol self-cleaving ribozyme (PDB entry 5ktj), showcasing (in red) the horizontal helix in space-filling representation. The helix is composed of six short stems stabilized via coaxial stacking interactions (bottom).

## 4.2 DSSR-PyMOL integration

The DSSR-PyMOL integration brings unprecedented visual clarity to 3D nucleic acid structures, especially for G-quadruplexes (see [Figures 15](#) and [16](#)). These features can be accessed via four interfaces: the command-line interface, the DSSR plugin for PyMOL, the web application, and the web application programming interface. The easiest way to get started and quickly benefit from this work is via the web application at <http://skmatic.x3dna.org>. Please refer to the paper<sup>3</sup> and the corresponding supplemental PDF for details.

The website <http://skmatic.x3dna.org> has been recommended in Faculty Opinions as “simple and effective”, and classified as “Good for Teaching”.

## 5 Modeling capabilities

The DSSR Pro comes with three modeling modules: (i) `mutate` for *in silico* (base) mutations, (ii) `fiber` for generating regular helical models, and (iii) `rebuild` for creating customized structures. They replace the 3DNA `mutate_bases`, `fiber`, and `rebuild` programs, with significantly enhanced features and dramatically improved usability.

These features are documented in the DSSR Pro User Manual (licensed to paid users only).

## 6 Frequently asked questions

### 6.1 What does DSSR stand for?

DSSR stands for Dissecting the Spatial Structure of RNA. It could also mean Defining the Secondary Structure of RNA. Note that DSSR has far more to offer than just defines RNA 2D structures as DSSP does for proteins. The acronym may have other interpretations.

### 6.2 How does DSSR compare with other tools?

Dozens of software programs and online resources are currently in use for nucleic acid structural bioinformatics. It is fair to say that each has its unique features and no two tools are identical. Comparative studies as seen in the literature are useful for general understanding of a topic. Without going into technical details, however, such comparisons are mostly superficial

and seldomly convincing. See my early paper on “Resolving the discrepancies among nucleic acid conformational analyses”<sup>41</sup>.

DSSR is an integrated and automated computational tool designed from the bottom up to streamline RNA structural bioinformatics. It possesses a combined set of functionalities well beyond the scope of any other software tools in the field. As a simple example, users are encouraged to compare DSSR with any other tools on the classic yeast tRNA<sup>Phe</sup> (1ehz, see [Section 3.3](#)). Pay close attention to the fact that 1ehz contains 14 modified nucleotides. Among other features, yeast tRNA<sup>Phe</sup> has four base triplets, two helices corresponding to the L-shaped tertiary structure, four stems matching the cloverleaf 2D structure, three hairpin loops, and a [2,1,5,0] four-way junction loop.

### 6.3 How is DSSR related to 3DNA?

From a historical perspective, DSSR is built upon the 3DNA suite of software programs for the analysis, rebuilding, and visualization of 3D nucleic acid structures<sup>9-11</sup>. 3DNA takes advantage of standard base reference frame<sup>4</sup>, enabled by our contributions on resolving the discrepancies among nucleic acid conformational analyses<sup>41,42</sup>. 3DNA had also benefitted greatly from the SCHNAAp and SCHNArP pair of programs for rigorous analysis and reversible rebuilding of double-helical nucleic acid structures<sup>39,43</sup>. Specifically, the algorithms that underpinned SCHNAAp/SCHNArP laid the foundation of `analyze/rebuild`, two core components of the 3DNA suite.

Just as 3DNA has replaced SCHNAAp/SCHNArP in functionality and real-world applications, DSSR has superseded 3DNA. Key 3DNA features for analysis, visualization, modeling, and utilities have been integrated into DSSR, with vastly enhanced functionality and significantly improved usability. Notably, the `mutate/fiber/rebuild` modules in DSSR completely supersedes the `mutate_bases`, `fiber`, and `rebuild` programs distributed with 3DNA v2.4. In short, 3DNA is becoming the past; DSSR is the future.

### 6.4 What are the differences between DSSR basic and pro?

DSSR Pro contains more features than documented here for the basic version. DSSR Pro has advanced modeling modules that are missing in the basic version. DSSR Pro comes with a comprehensive user manual that is currently over 230 pages. DSSR Pro can be taken as X3DNA since it completely supersedes 3DNA v2.4. If you find a feature missing in the basic version,

it is likely to be available in DSSR Pro. Moreover, user support, software maintenance and further development are devoted to DSSR Pro.

## 7 Acknowledgements

The development of DSSR has been made possible by the NIH grant R01GM096889. I'd like to thank Drs. Wilma Olson and Harmen Bussemaker for helpful discussions. I appreciate the user community for providing feedback.

## References

1. Lu, X.-J., Bussemaker, H. J., and Olson, W. K. (2015) DSSR: An integrated software tool for dissecting the spatial structure of RNA. *Nucleic Acids Res.*, **43**(21), e142.
2. Hanson, R. M. and Lu, X.-J. (2017) DSSR-enhanced visualization of nucleic acid structures in Jmol. *Nucleic Acids Res.*, **45**(W1), W528–W533.
3. Lu, X.-J. (2020) DSSR-enabled innovative schematics of 3D nucleic acid structures with PyMOL. *Nucleic Acids Res.*, **48**(13), e74.
4. Olson, W. K., Bansal, M., Burley, S. K., Dickerson, R. E., Gerstein, M., Harvey, S. C., Heinemann, U., Lu, X.-J., Neidle, S., Shakked, Z., Sklenar, H., Suzuki, M., Tung, C.-S., Westhof, E., Wolberger, C., and Berman, H. M. (2001) A standard reference frame for the description of nucleic acid base-pair geometry. *J. Mol. Biol.*, **313**(1), 229–237.
5. Saenger, W. (1984) Principles of nucleic acid structure, Springer advanced texts in chemistrySpringer-Verlag, New York first edition.
6. Leontis, N. B. and Westhof, E. (2001) Geometric nomenclature and classification of RNA base pairs. *RNA*, **7**(4), 499–512.
7. Darty, K., Denise, A., and Ponty, Y. (2009) VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics*, **25**(15), 1974–5.
8. Richardson, J. S., Schneider, B., Murray, L. W., Kapral, G. J., Immormino, R. M., Headd, J. J., Richardson, D. C., Ham, D., Hershkovits, E., Williams, L. D., Keating, K. S., Pyle,

- A. M., Micallef, D., Westbrook, J., and Berman, H. M. (2008) RNA backbone: consensus all-angle conformers and modular string nomenclature. *RNA*, **14**(3), 465–81.
9. Lu, X.-J. and Olson, W. K. (2003) 3DNA: A software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res.*, **31**(17), 5108–5121.
10. Lu, X.-J. and Olson, W. K. (2008) 3DNA: A versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. *Nat. Protoc.*, **3**(7), 1213–1227.
11. Li, S., Olson, W. K., and Lu, X.-J. (2019) Web 3DNA 2.0 for the analysis, visualization, and modeling of 3D nucleic acid structures. *Nucleic Acids Res.*, **47**(W1), W26–W34.
12. Correll, C. C., Beneken, J., Plantinga, M. J., Lubbers, M., and Chan, Y.-L. (2003) The common and the distinctive features of the bulged-G motif based on a 1.04 Å resolution RNA structure. *Nucleic Acids Res.*, **31**(23), 6806–6818.
13. Lu, X. J., Olson, W. K., and Bussemaker, H. J. (2010) The RNA backbone plays a crucial role in mediating the intrinsic stability of the GpU dinucleotide platform and the GpUpA/GpA miniduplex. *Nucleic Acids Res.*, **38**(14), 4868–76.
14. Gesteland, R. F., Cech, T., and Atkins, J. F. (1999) The RNA world: the nature of modern RNA suggests a prebiotic RNA world, Number 37 in Cold Spring Harbor monograph seriesCold Spring Harbor Laboratory Press, Cold Spring Harbor, NY second edition.
15. Lemieux, S. and Major, F. (2002) RNA canonical and non-canonical base pairing types: a recognition method and complete repertoire. *Nucleic Acids Res.*, **30**(19), 4250–63.
16. Nissen, P., Ippolito, J. A., Ban, N., Moore, P. B., and Steitz, T. A. (2001) RNA tertiary interactions in the large ribosomal subunit: the A-minor motif. *Proc. Natl. Acad. Sci.*, **98**(9), 4899–903.
17. Olson, W. K., Li, S., Kaukonen, T., Colasanti, A. V., Xin, Y., and Lu, X.-J. (2019) Effects of noncanonical base pairing on RNA folding: structural context and spatial arrangements of G·A pairs. *Biochemistry*, **58**(20), 2474–2487.
18. Quigley, G. and Rich, A. (1976) Structural domains of transfer RNA molecules. *Science*, **194**(4267), 796–806.

19. Drew, H. R., Wing, R. M., Takano, T., Broka, C., Tanaka, S., Itakura, K., and Dickerson, R. E. (1981) Structure of a B-DNA dodecamer: conformation and dynamics. *Proc Natl Acad Sci U A*, **78**(4), 2179–83.
20. Lu, X. J., Shakked, Z., and Olson, W. K. (2000) A-form conformational motifs in ligand-bound DNA structures. *J. Mol. Biol.*, **300**(4), 819–40.
21. Chen, V. B., Arendall, W. B., Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S., and Richardson, D. C. (2010) MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D Biol. Crystallogr.*, **66**(1), 12–21.
22. Olson, W. K. (1980) Configurational statistics of polynucleotide chains. An updated virtual bond model to treat effects of base stacking. *Macromolecules*, **13**(3), 721–728.
23. Keating, K. S., Humphris, E. L., and Pyle, A. M. (2011) A new way to see RNA. *Q. Rev. Biophys.*, **44**(4), 433–66.
24. Altona, C. and Sundaralingam, M. (1972) Conformational analysis of the sugar ring in nucleosides and nucleotides. A new description using the concept of pseudorotation. *J Am Chem Soc*, **94**(23), 8205–12.
25. Zhang, H., Endrizzi, J. A., Shu, Y., Haque, F., Sauter, C., Shlyakhtenko, L. S., Lyubchenko, Y., Guo, P., and Chi, Y.-I. (2013) Crystal structure of 3WJ core revealing divalent ion-promoted thermostability and assembly of the phi29 hexameric motor pRNA. *RNA*, **19**(9), 1226–1237.
26. Nguyen, L. A., Wang, J., and Steitz, T. A. (2017) Crystal structure of Pistol, a class of self-cleaving ribozyme. *Proc Natl Acad Sci USA*, **114**(5), 1021–1026.
27. Cate, J. H., Gooding, A. R., Podell, E., Zhou, K., Golden, B. L., Kundrot, C. E., Cech, T. R., and Doudna, J. A. (1996) Crystal structure of a group I ribozyme domain: principles of RNA packing. *Science*, **273**(5282), 1678–85.
28. Klein, D. J., Schmeing, T. M., Moore, P. B., and Steitz, T. A. (2001) The kink-turn: a new RNA secondary structure motif. *EMBO J.*, **20**(15), 4214–21.
29. Lilley, D. M. (2012) The structure and folding of kink turns in RNA. *WIREs RNA*, **3**(6), 797–805.

30. Wang, J., Daldrop, P., Huang, L., and Lilley, D. M. J. (2014) The k-junction motif in RNA structure. *Nucleic Acids Res.*, **42**(8), 5322–5331.
31. Burge, S., Parkinson, G. N., Hazel, P., Todd, A. K., and Neidle, S. (2006) Quadruplex DNA: Sequence, topology and structure. *Nucleic Acids Res.*, **34**(19), 5402–5415.
32. Rhodes, D. and Lipps, H. J. (2015) G-quadruplexes and their regulatory roles in biology. *Nucleic Acids Res.*, **43**(18), 8627–8637.
33. Lightfoot, H. L., Hagen, T., Tatum, N. J., and Hall, J. (2019) The diverse structural landscape of quadruplexes. *FEBS Lett.*, **593**(16), 2083–2102.
34. Meier, M., Moya-Torres, A., Krahn, N. J., McDougall, M. D., Orriss, G. L., McRae, E. K. S., Booy, E. P., McEleney, K., Patel, T. R., McKenna, S. A., and Stetefeld, J. (2018) Structure and hydrodynamics of a DNA G-quadruplex with a cytosine bulge. *Nucleic Acids Res.*, **46**(10), 5319–5331.
35. Dvorkin, S. A., Karsisiotis, A. I., and Webba da Silva, M. (2018) Encoding canonical DNA quadruplex structure. *Sci. Adv.*, **4**(8), eaat3007.
36. Theimer, C. A., Blois, C. A., and Feigon, J. (2005) Structure of the human telomerase RNA pseudoknot reveals conserved tertiary interactions essential for function. *Mol. Cell*, **17**(5), 671–682.
37. Sussman, D., Nix, J. C., and Wilson, C. (2000) The structural basis for molecular recognition by the vitamin B 12 RNA aptamer. *Nat. Struct. Biol.*, **7**(1), 53–57.
38. Smit, S., Rother, K., Heringa, J., and Knight, R. (2008) From knotted to nested RNA structures: A variety of computational methods for pseudoknot removal. *RNA*, **14**(3), 410–416.
39. Lu, X.-J., El Hassan, M., and Hunter, C. (1997) Structure and conformation of helical nucleic acids: Analysis program (SCHNAAP). *J. Mol. Biol.*, **273**(3), 668–680.
40. Parisien, M., Cruz, J. A., Westhof, E., and Major, F. (2009) New metrics for comparing and assessing discrepancies between RNA 3D structures and models. *RNA*, **15**(10), 1875–1885.
41. Lu, X. J. and Olson, W. K. (1999) Resolving the discrepancies among nucleic acid conformational analyses. *J. Mol. Biol.*, **285**(4), 1563–75.

42. Lu, X.-J., Babcock, M. S., and Olson, W. K. (1999) Overview of Nucleic Acid Analysis Programs. *J. Biomol. Struct. Dyn.*, **16**(4), 833–843.
43. Lu, X.-J., El Hassan, M., and Hunter, C. (1997) Structure and conformation of helical nucleic acids: Rebuilding program (SCHNArP). *J. Mol. Biol.*, **273**(3), 681–691.

User Manual for Basic DSSR 2.0