

Lost in missing residues

Challenges while working with RNA
structures from the Protein Data Base

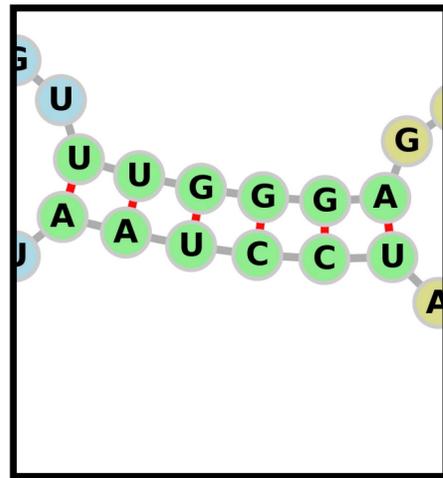
Bernhard Thiel, Bled 2018

What is RNA?

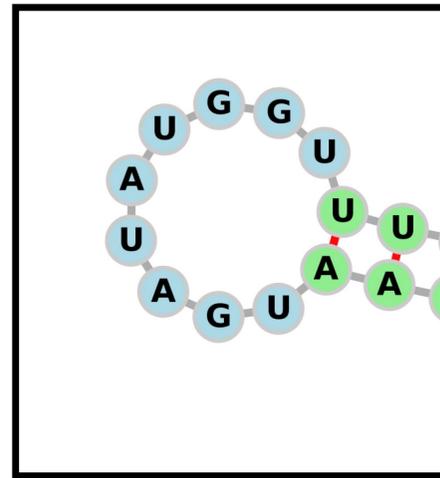
*) Polymer-chain of Adenine, Uracil, Guanine and Cytosine

AUAGUGCUGACUGACUGACUCGAUCGUCAGUCAGCA

The **forgi** representation of RNA?

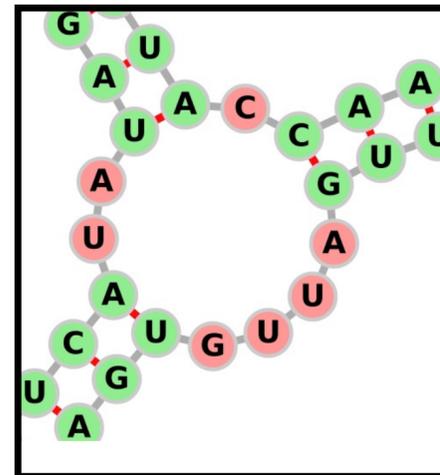
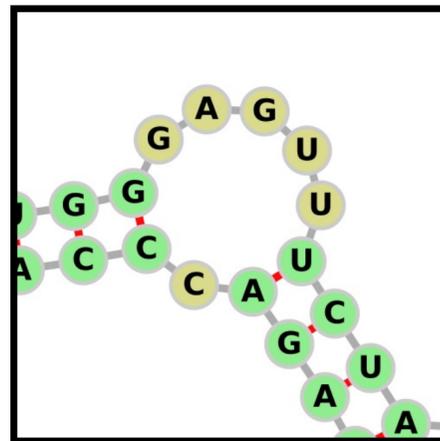


Stem

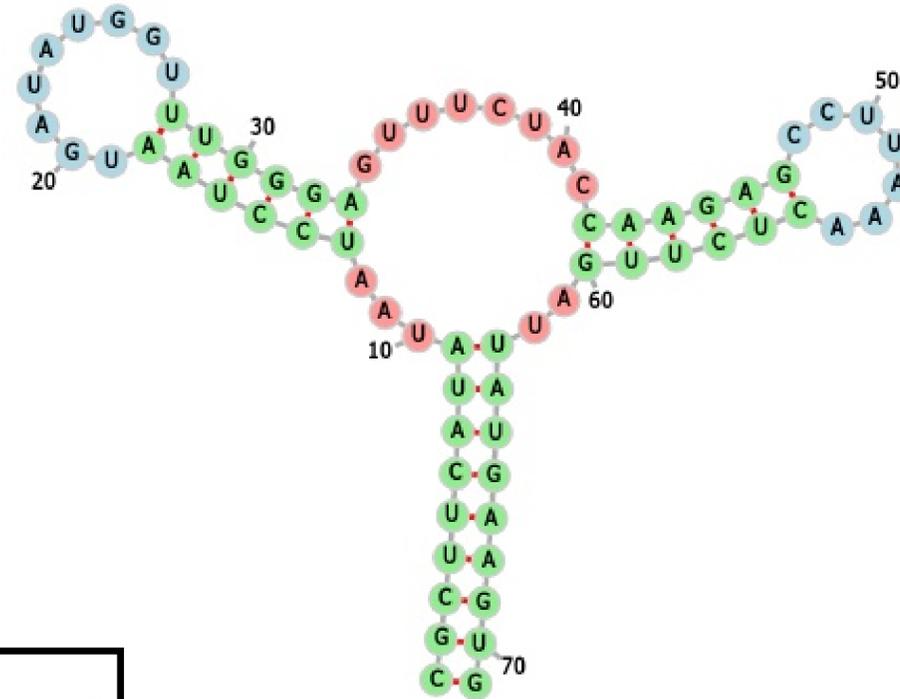


Hairpin

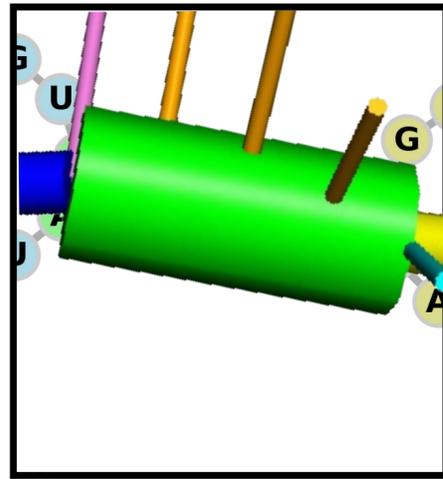
Interior loop



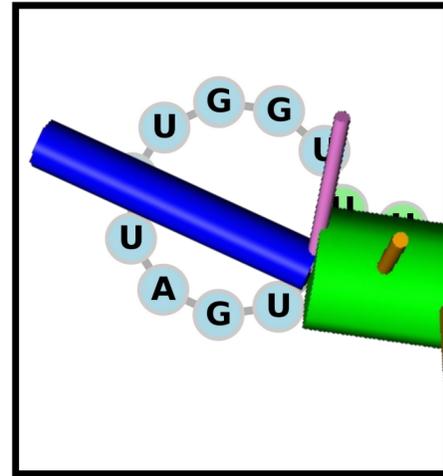
Multiloop



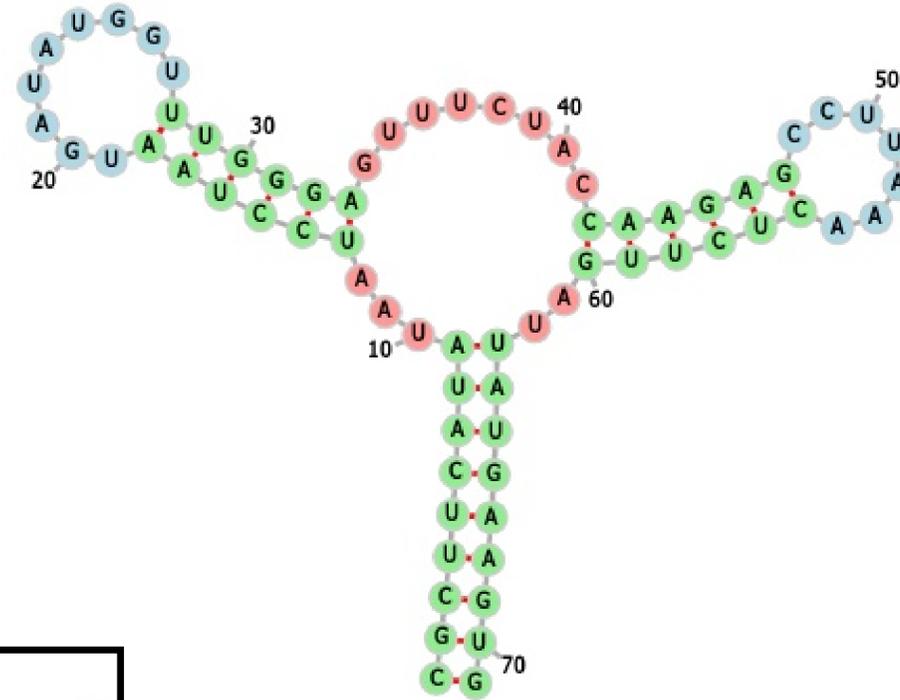
The **forgi** representation of RNA?



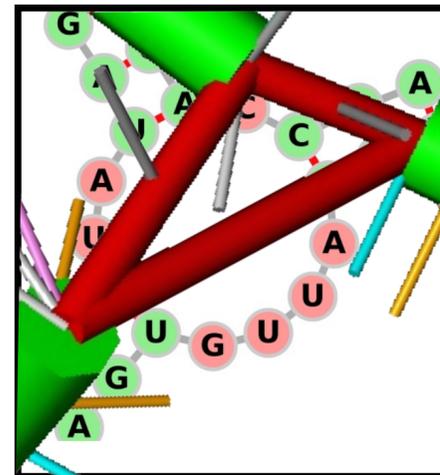
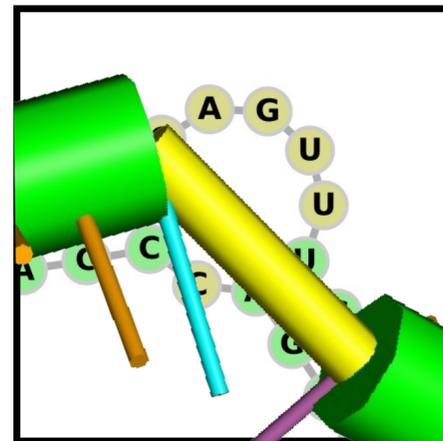
Stem



Hairpin

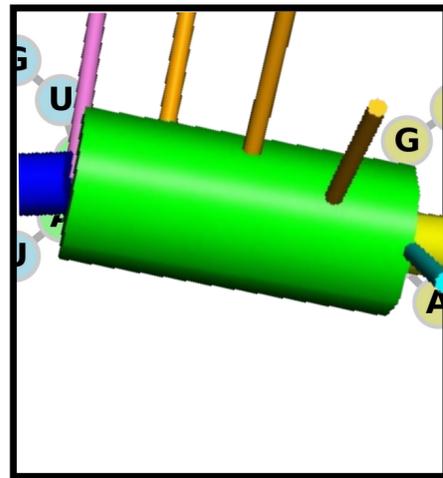


Interior loop

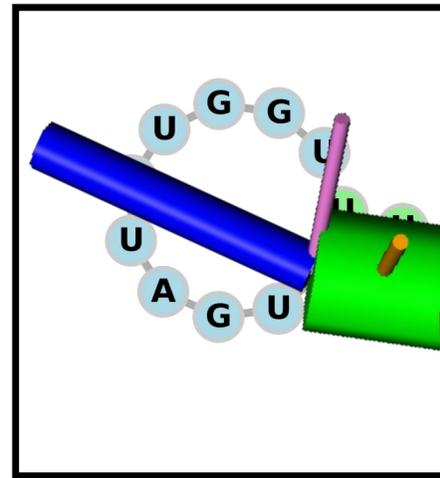


Multiloop

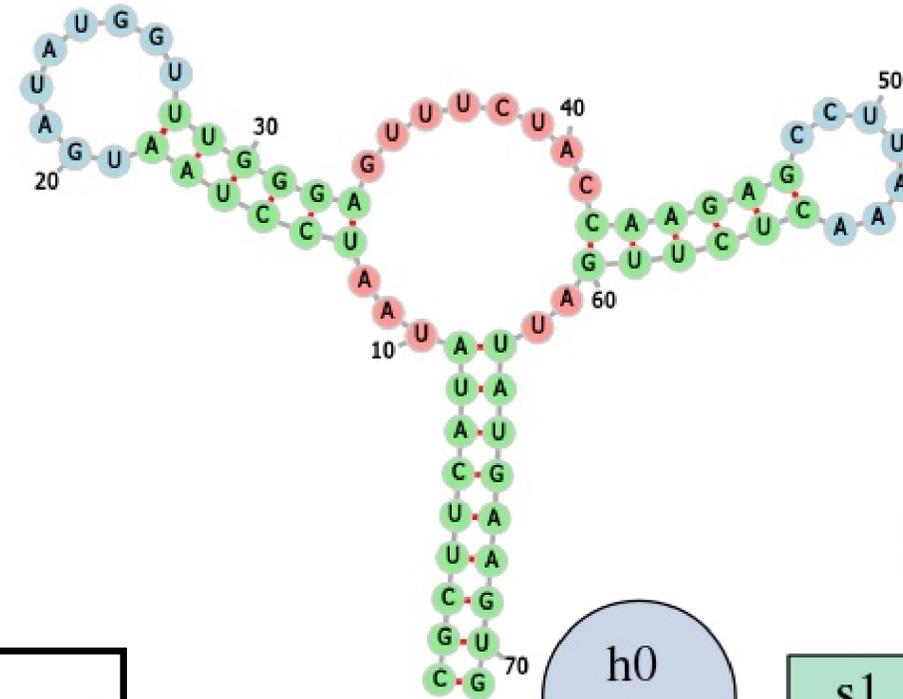
The **forgi** representation of RNA?



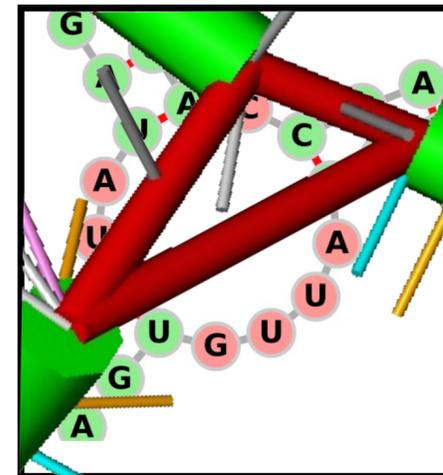
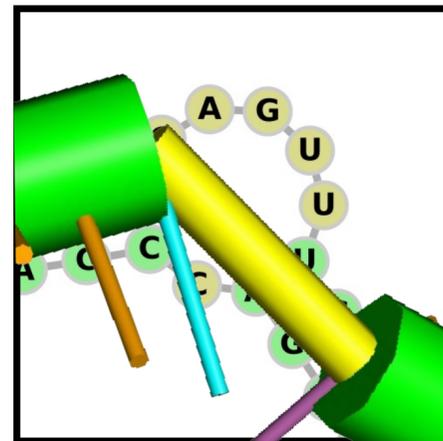
Stem



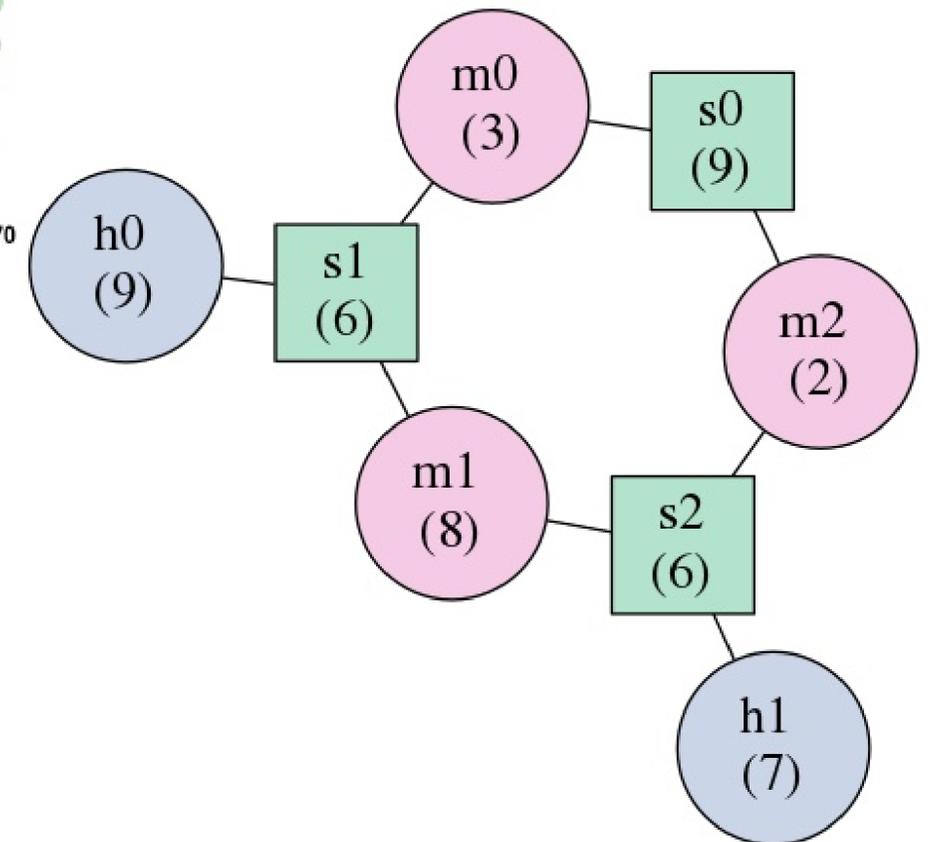
Hairpin



Interior loop



Multiloop



What is the PDB?

- *) Experimentally solved structures
- XRAY, NMR, Cryo-EM

The screenshot shows the RCSB PDB website interface. At the top, there is a navigation bar with links for Deposit, Search, Visualize, Analyze, Download, Learn, and More. A search bar is prominently displayed with the text "Search by PDB ID, author, macromolecule, sequence, or ligands". Below the search bar, there are several featured sections:

- Welcome:** A vertical sidebar with icons for Deposit, Search, Visualize, Analyze, Download, and Learn.
- A Structural View of Biology:** A text-based section explaining the PDB's mission: "This resource is powered by the Protein Data Bank archive-information about the 3D shapes of proteins, nucleic acids, and complex assemblies that helps students and researchers understand all aspects of biomedicine and agriculture, from protein synthesis to health and disease." It also mentions that the RCSB PDB curates and annotates PDB data as a member of the wwPDB.
- February Molecule of the Month:** A section featuring a 3D molecular model of a protein structure, titled "February Molecule of the Month" and "EPSP Synthase and Weedkillers".
- New Video: What is a Protein?:** A video thumbnail with the text "VIDEO WHAT IS A PROTEIN?" and a "PDB-101" logo.
- Latest Entries:** A section titled "As of Tuesday Feb 06 2018" showing a 3D molecular model of a protein structure.
- Features & Highlights:** A section with several news items, including "New Architecture and Services Enable Faster Access to More Information" and "Implementation of PDB Entry Versioning and Better Revision History to Improve PDB Archive Management".
- News & Publications:** A section with a "News" tab and a "Publications" tab, featuring a "Making Life Visible: Art, Biology and Visualization Paintings from Molecule of the Month creator David Goodsell" exhibit at Fau.

PDB-File Format

Some header information and ATOM records

For each atom, the following is stored:

- * the residue name, e.g. A for Adenine
- * the number in the sequence
- * the Atom type and its coordinates

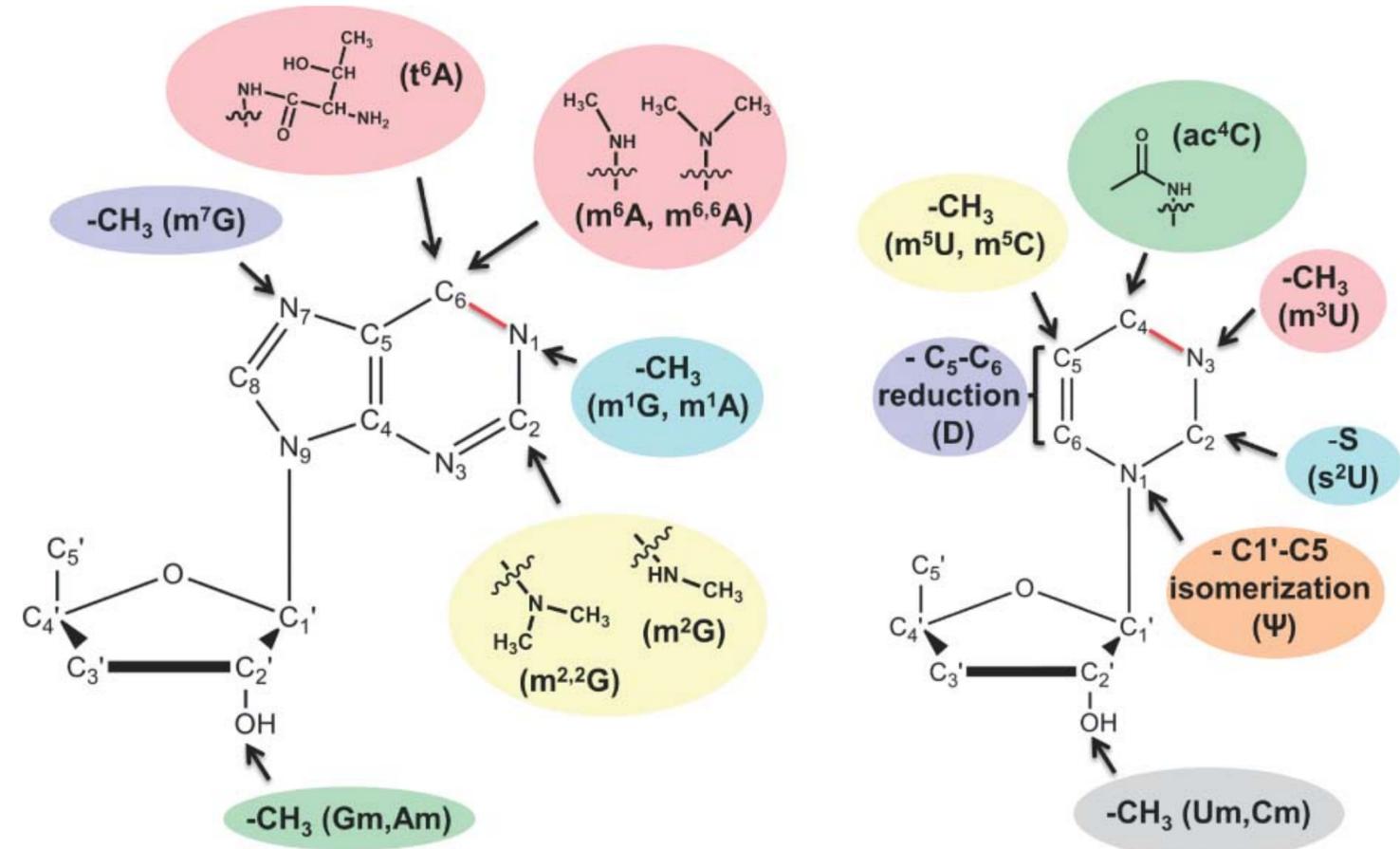
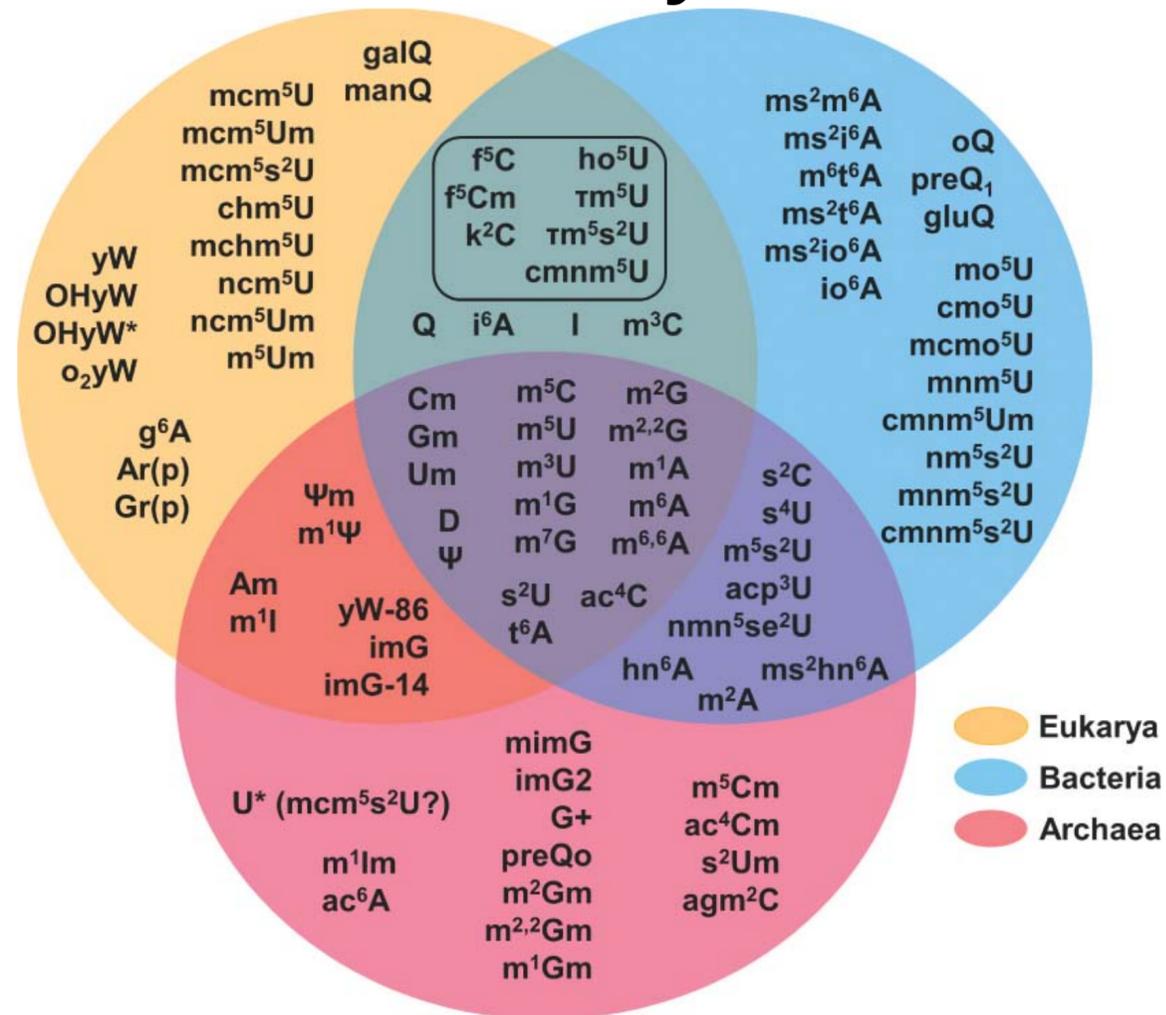
```

      1         2         3         4         5         6         7         8
1234567890123456789012345678901234567890123456789012345678901234567890
MODEL          1
ATOM           1  N   ALA  A   1      11.104   6.134  -6.504   1.00   0.00           N
ATOM           2  CA  ALA  A   1      11.639   6.071  -5.147   1.00   0.00           C
...
...
...
ATOM          293 1HG  GLU  A   18     -14.861  -4.847   0.361   1.00   0.00           H
ATOM          294 2HG  GLU  A   18     -13.518  -3.769   0.084   1.00   0.00           H
TER           295          GLU  A   18

```

RNA Modifications

tRNA is heavily modified



Images:

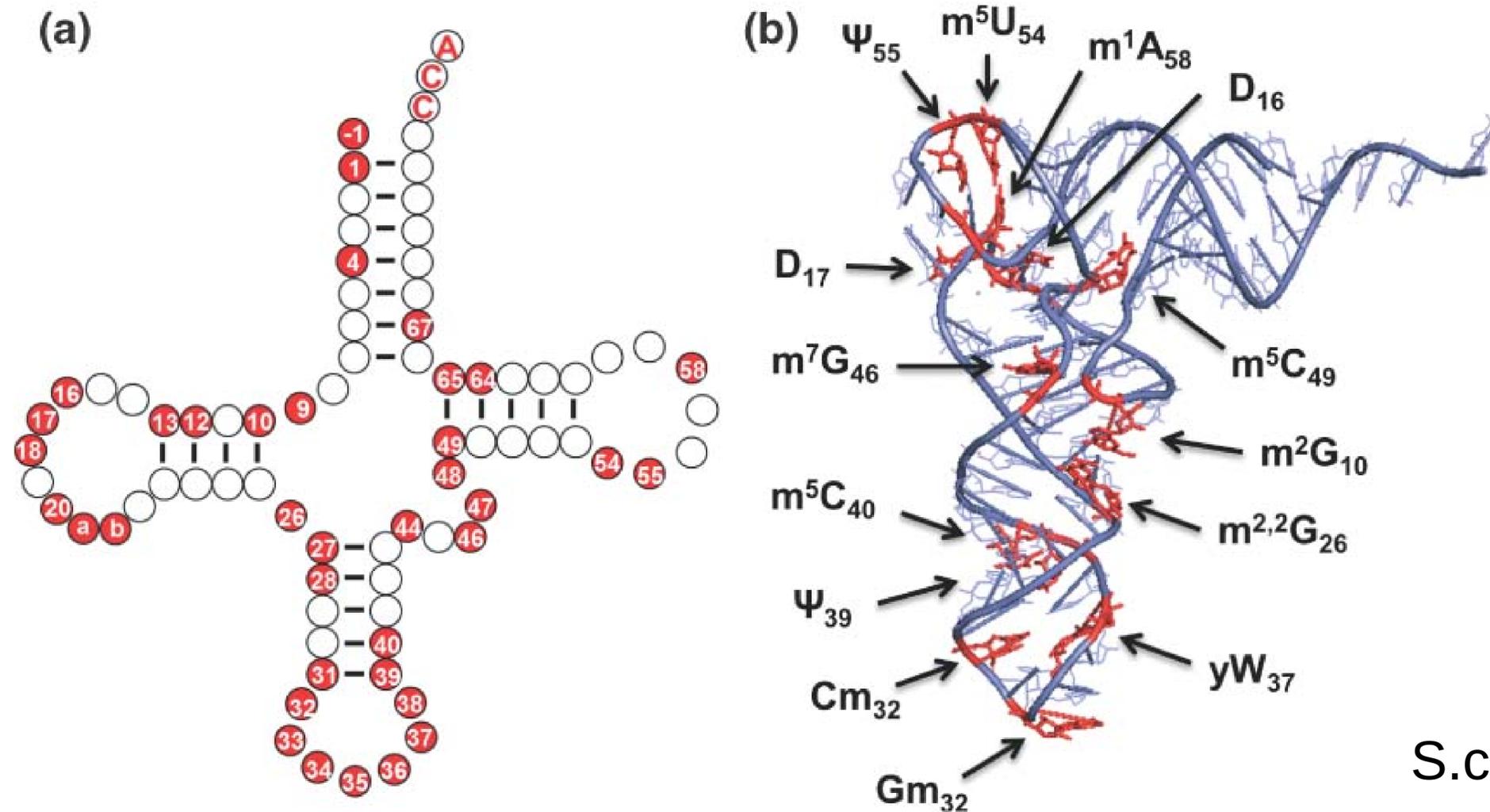
Transfer RNA modifications: nature's combinatorial chemistry playground

Wiley Interdisciplinary Reviews: RNA

Volume 4, Issue 1, pages 35-48, 8 NOV 2012 DOI: 10.1002/wrna.1144

RNA Modifications

tRNA is heavily modified



S.cerevisiae tRNA

Image:

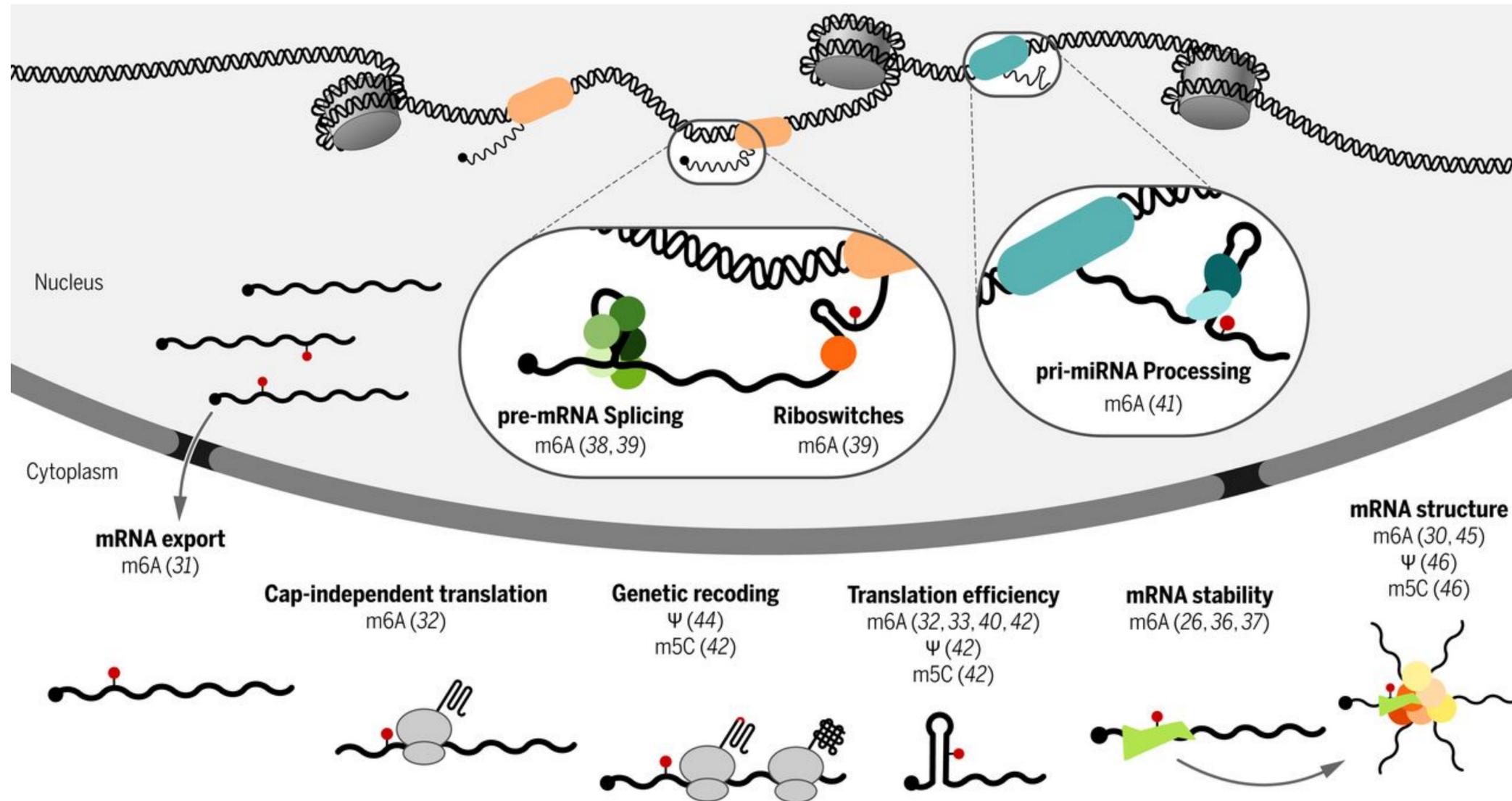
Transfer RNA modifications: nature's combinatorial chemistry playground

Wiley Interdisciplinary Reviews: RNA

Volume 4, Issue 1, pages 35-48, 8 NOV 2012 DOI: 10.1002/wrna.1144

RNA Modifications

"Widespread sparse modification of messenger RNAs"¹



Quote and image:

Gilbert et al, 2016, DOI: 10.1126/science.aad8711 ([1]: Quote from Abstract)

RNA Modifications

A-to-I editing

*) in double stranded RNA

*) by enzyme ADAR

*) I base-pairs like G

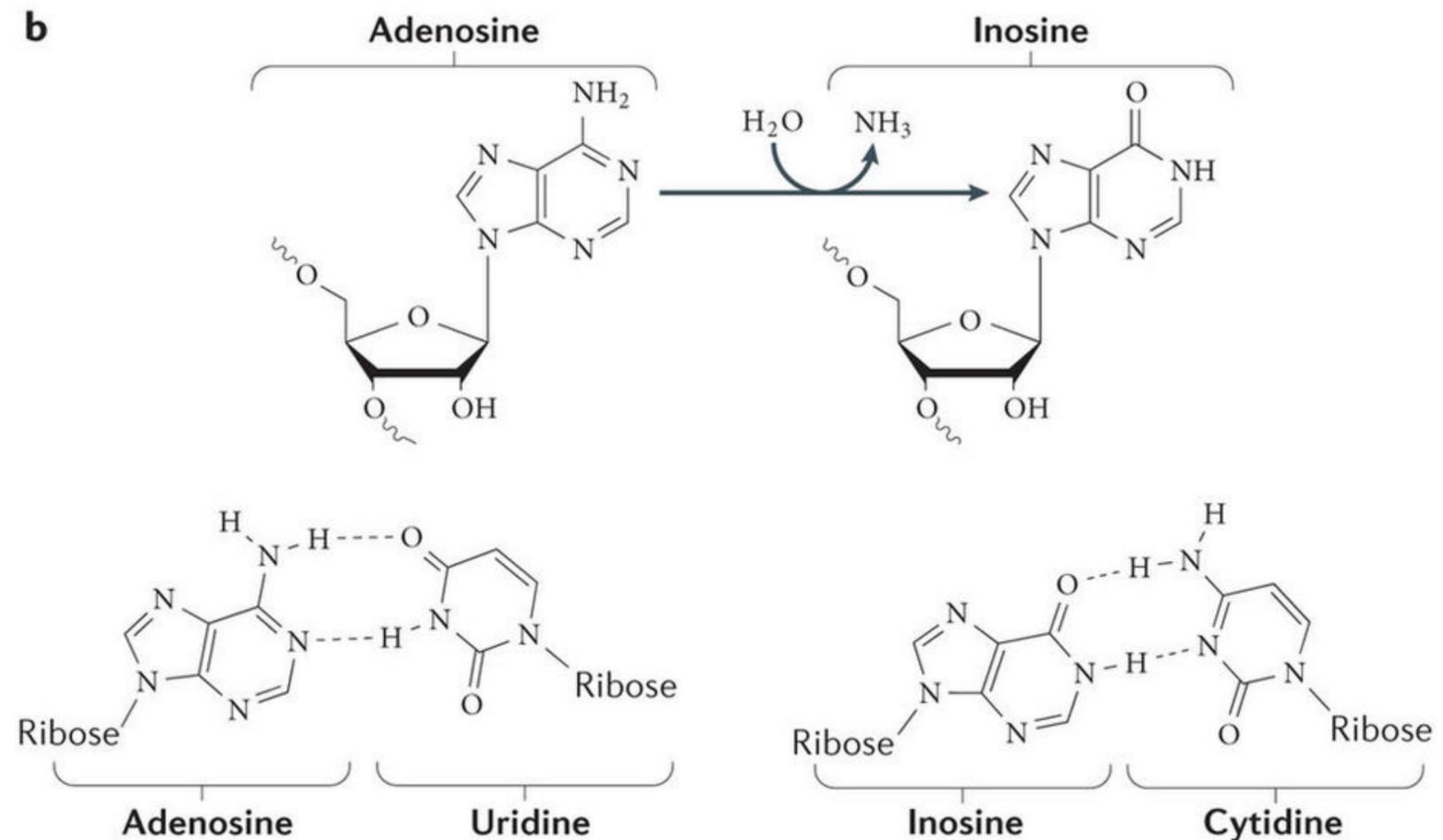
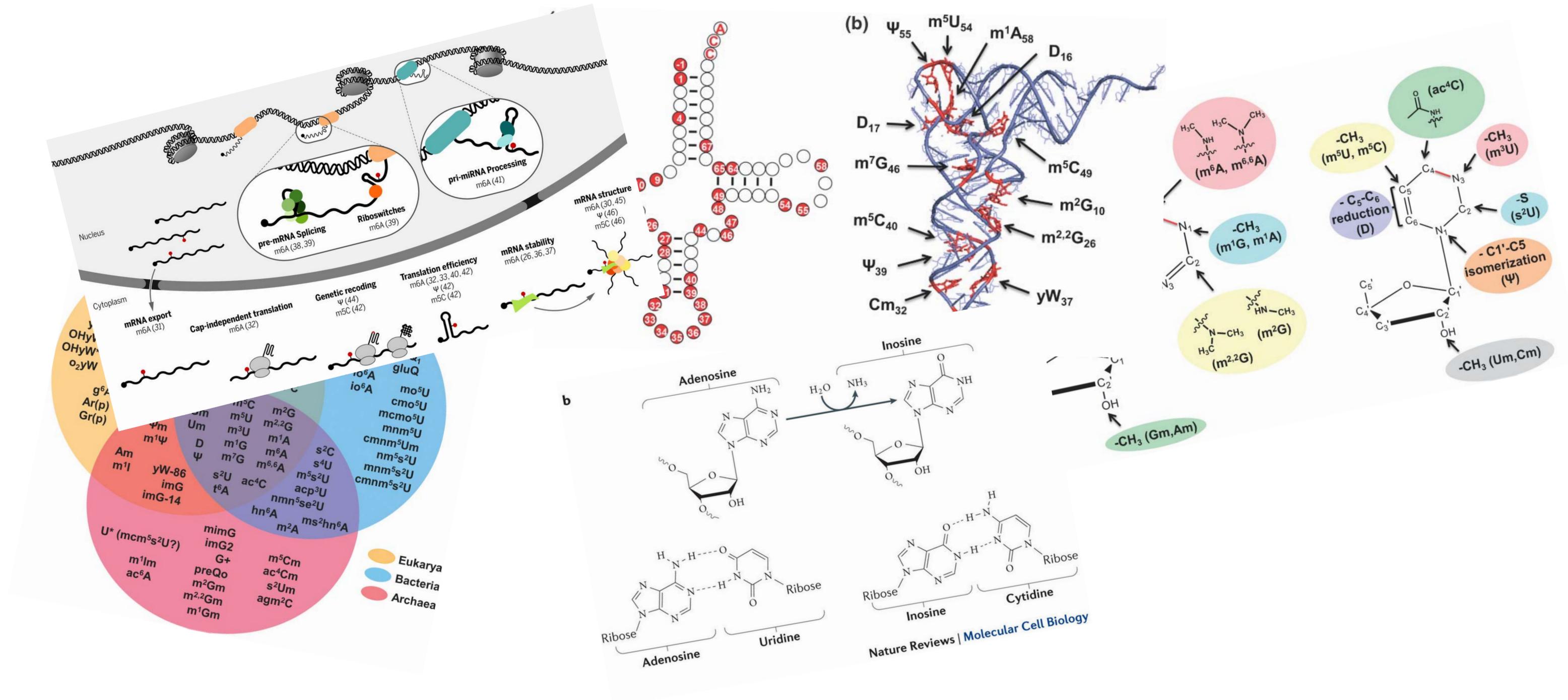


Image:

Nishikura, 2016, DOI: 10.1038/nrm.2015.4

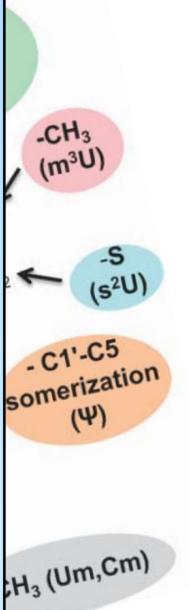
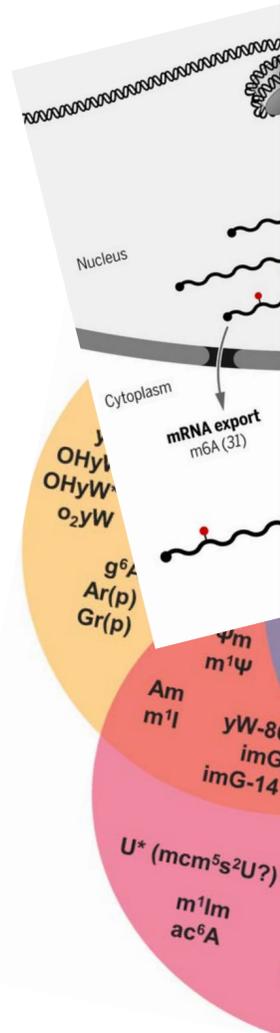
RNA Modifications



RNA Modifications

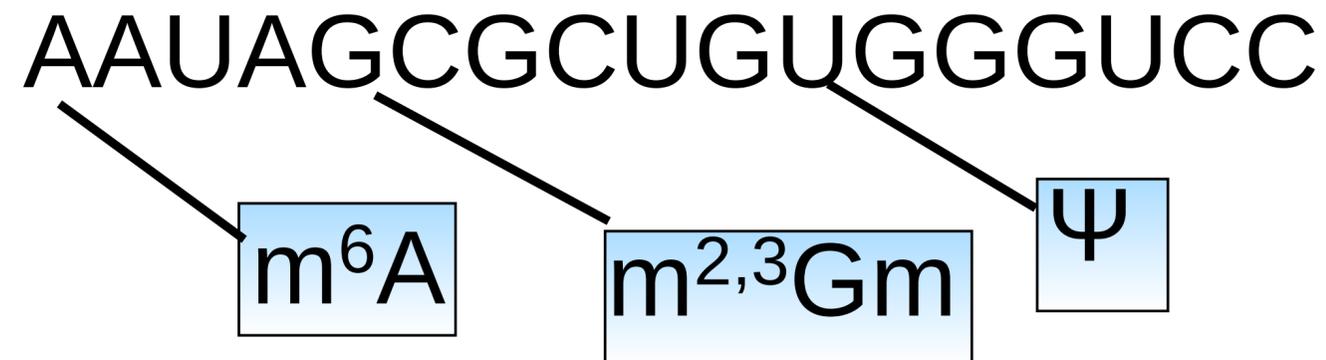
FORGI 2.0

- Load PDB files with modifications
- For each modified nt:
 - * Query PDBeChem
 - * Retrieve canonical parent
 - * Distinguish ligand and backbone molecules
- Sequence class
 - * a string of letters AUGC
 - * also stores modifications
 - * convenient access to modifications



What is RNA?

*) Polymer-chain of Adenine, Uracil, Guanine and Cytosine with annotations for modifications



www.rcsb.org/pdb/explore/remediatedSequence.do?structureId=3CGP

RCSB PDB Deposit Search Visualize Analyze Download Learn More

Structure Summary 3D View Experiment

3CGP

X-ray structure of a pseudouridine-containing yeast spliceosomal U2 snRNA-intron branch site duplex bound to iodide ions

Display Files Download Files

Sequence Display for the Entities in PDB 3CGP

The graphical representation below shows this entry's sequences as reported in UniProtKB, in the sample (SEQRES), or as observed in the experiment (ATOM). Different 3rd party annotations can be graphically mapped on the sequence and displayed in the Jmol viewer. Read more about the [sequence display on our help pages](#).

The structure **3CGP** has in total **2** chains. These are represented by **2** sequence-unique entities.

Display Options
Currently viewing **unique chains** only

Sequence & Structure Relationships. Enable Jmol to view annotations in 3D.

Redundancy Reduction and Sequence Clustering. View the [clustering results](#) for 3CGP.

Chain A: RNA (5'-R>(*GP*CP*GP*CP*GP*(PSU)P*AP*GP*UP*AP*GP*C)-3')

Chain Downloadable Files	Chain Info	Display Parameters
<input type="button" value="Download FASTA File"/> <input type="button" value="View Sequence & DSSP Image"/> <input type="button" value="Download Sequence Chain Image"/>	Polymer: 1 Length: 12 residues Chain Type: polyribonucleotide	No parameters are available for this sequence

Mouse over an annotation to see more details. Click on any annotation to enable Jmol.

Annotations Details

Sequence Chain View

PDB G C G C G U A G U A G C
PDB 1 10 12

Modified Residue: PSU 6 PSEUDOURIDINE-5'-MONOPHOSPHATE C9 H13 N2 O9 P (parent: U)

Chain B: RNA (5'-R(*CP*GP*CP*UP*AP*CP*UP*AP*AP*CP*GP*CP*G)-3')

Residue number

Author provided

Number 1 is usually biologically meaningful,
but can be anywhere

Numbers can be missing

-) Deletions

-) No coordinates determined

Nucleotides can be inserted

Nucleotides can have negative numbers

Residue number

A -3
U -2
G 1
C 2
U 2B
A 2C
A 3
G 7
U 8
G 9

Residue number

A -3
U -2
G 1
C 2
U 2B
A 2C
A 3
G 7
U 8
G 9

FORGI 2.0

Sequence class supports 2 kinds of indices:

*) consecutive, 1-based

*) corresponding to PDB

Indices are distinguished by type

Missing residues

Residues present in experiment, for which no coordinates were determined.

PDB-File contains information at 3 places:

SEQRES: The sequence of the molecule used for the experiment, without numbers. -> parse with Bio.SeqIO

ATOM: atom-coordinates, nucleotide number, nt-name

REMARK 465: nt-number and nt-name for residues without coordinates

Missing residues

PDB-File contains information at 3 places:
SEQRES, ATOM, REMARK 465

```

1      2      3      4      5      6      7      8
1234567890123456789012345678901234567890123456789012345678901234567890
MODEL      1
ATOM       1  N   ALA A  1      11.104  6.134 -6.504  1.00  0.00      N
ATOM       2  CA  ALA A  1      11.639  6.071 -5.147  1.00  0.00      C
...
...
...
ATOM      293 1HG  GLU A  18      -14.861 -4.847  0.361  1.00  0.00      H
ATOM      294 2HG  GLU A  18      -13.518 -3.769  0.084  1.00  0.00      H
TER       295      GLU A  18

```

```

SEQRES  10 C  129  ILE PRO SER ALA ILE ALA ALA ASN SER GLY ILE TYR
SEQRES  1 R   14   U  C  G  C  C  A  A  C  A  G  G  C  G
SEQRES  2 R   14   G
SEQRES  1 S   14   U  C  G  C  C  A  A  C  A  G  G  C  G
SEQRES  2 S   14   G
FORMUL  6 HOH *90(H2 O)

```

```

REMARK 350 BIOMT3 60 0.934172 -0.127323 -0.333334 0.000000
REMARK 465
REMARK 465 MISSING RESIDUES
REMARK 465 THE FOLLOWING RESIDUES WERE NOT LOCATED IN THE
REMARK 465 EXPERIMENT. (M=MODEL NUMBER; RES=RESIDUE NAME; C=CHAIN
REMARK 465 IDENTIFIER; SSSEQ=SEQUENCE NUMBER; I=INSERTION CODE.)
REMARK 465
REMARK 465 M RES C SSSEQI
REMARK 465 U S 1
REMARK 465 C S 2
REMARK 465 G S 14
REMARK 470
REMARK 470 MISSING ATOM
REMARK 470 THE FOLLOWING RESIDUES HAVE MISSING ATOMS(M=MODEL NUMBER;

```

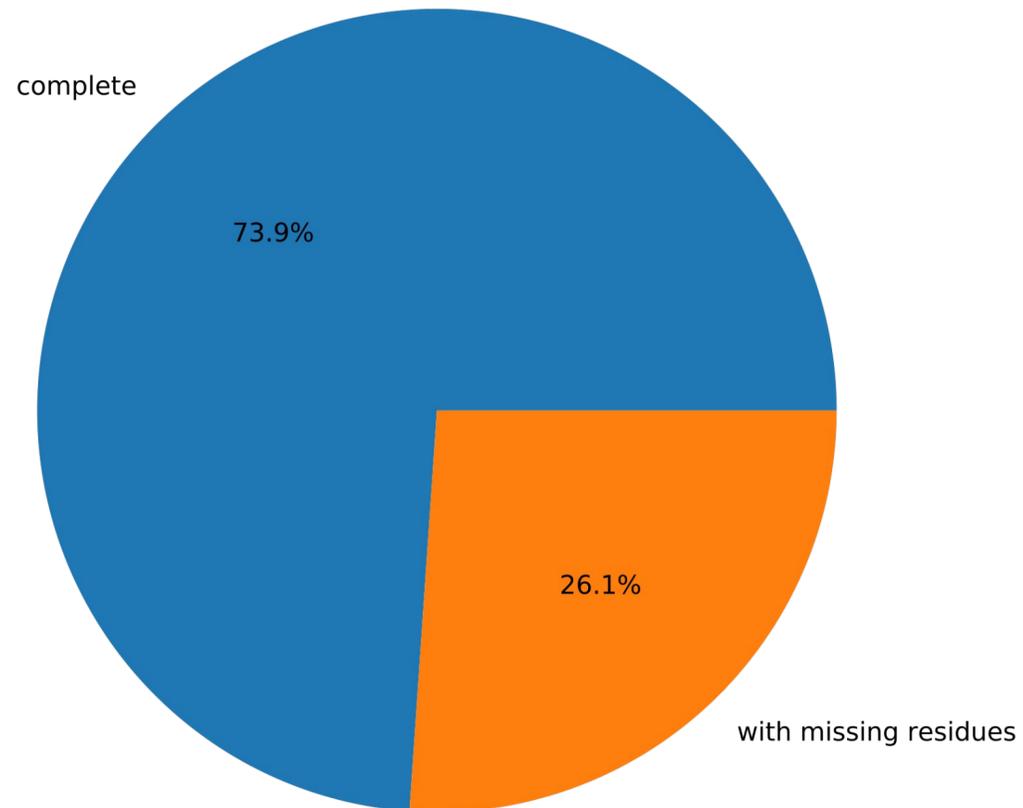
Missing residues

FORGI 2.0

- * Parse REMARK 465 (Biopython PR)
- * Get (sub-) sequences with or without missing residues.
- * Identify coarse-grained element (stem/...) which contain missing residues.

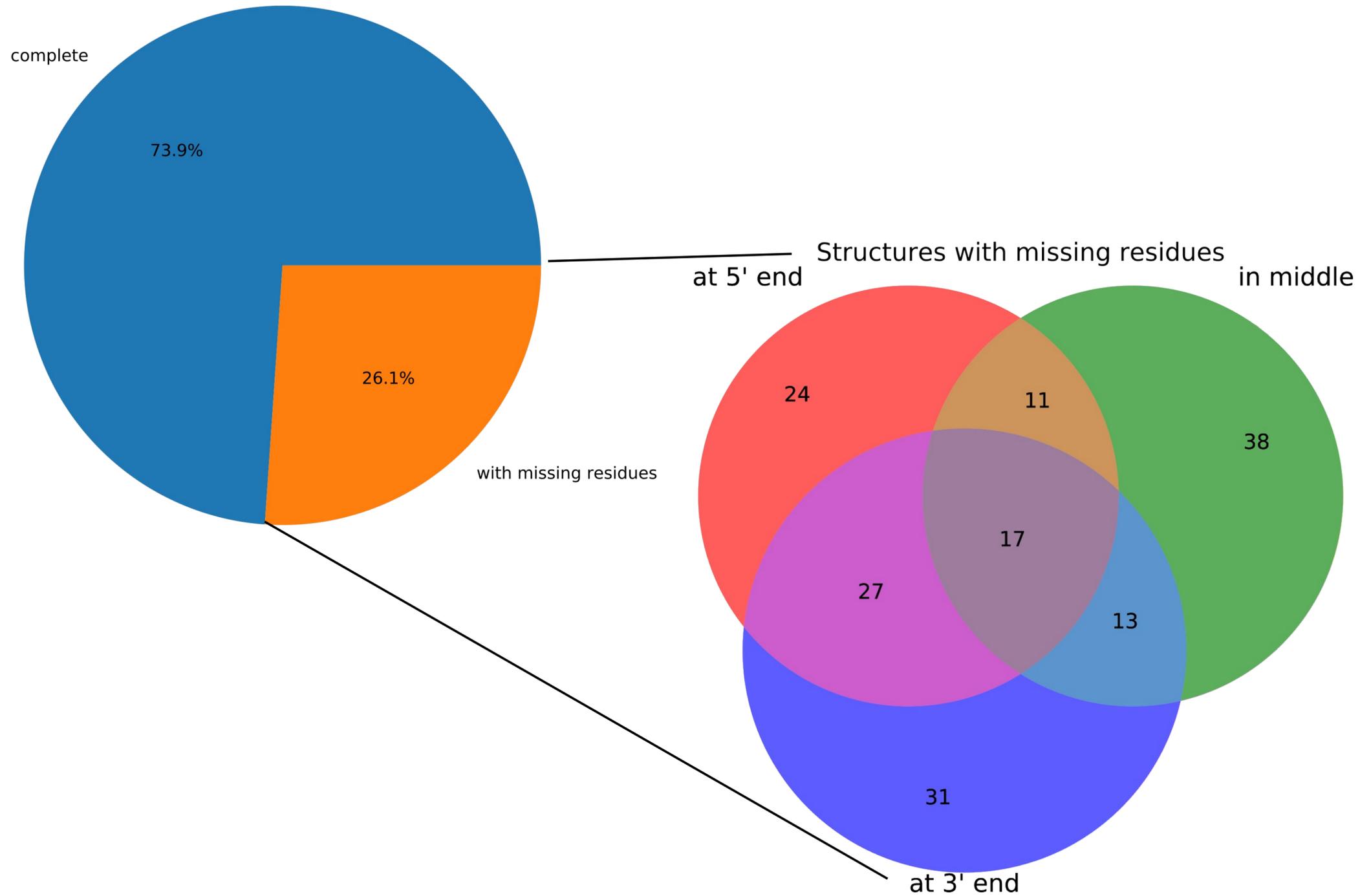
Missing residues

Large PDB structures in representative set



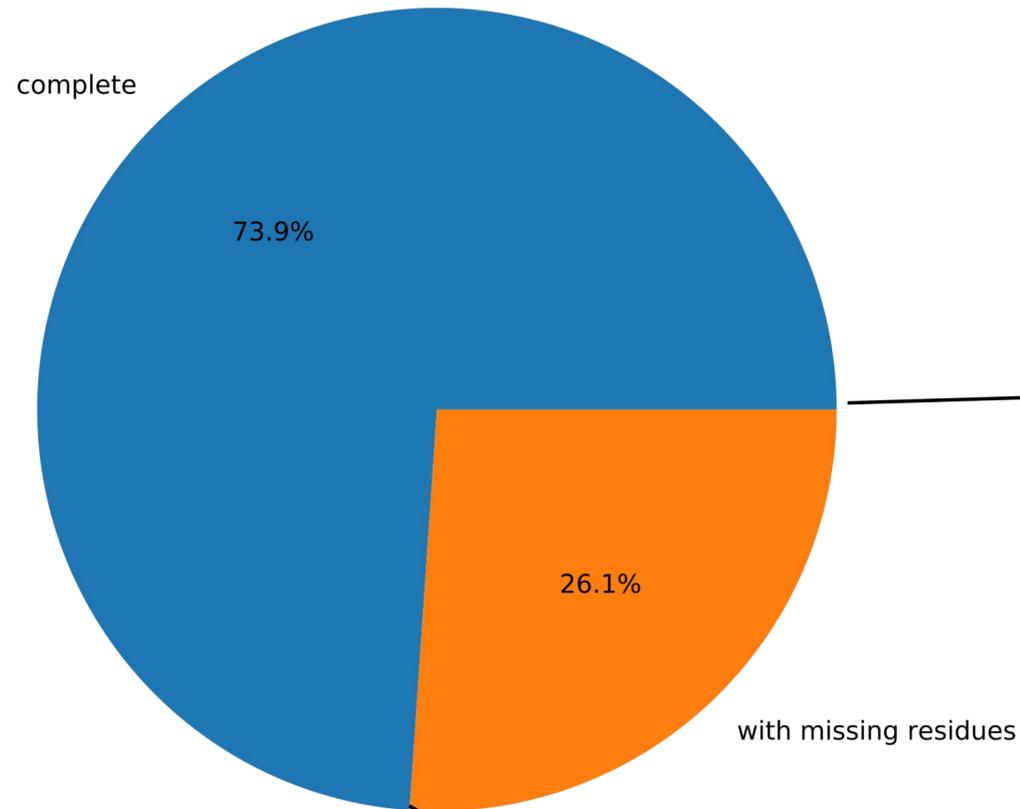
Missing residues

Large PDB structures in representative set

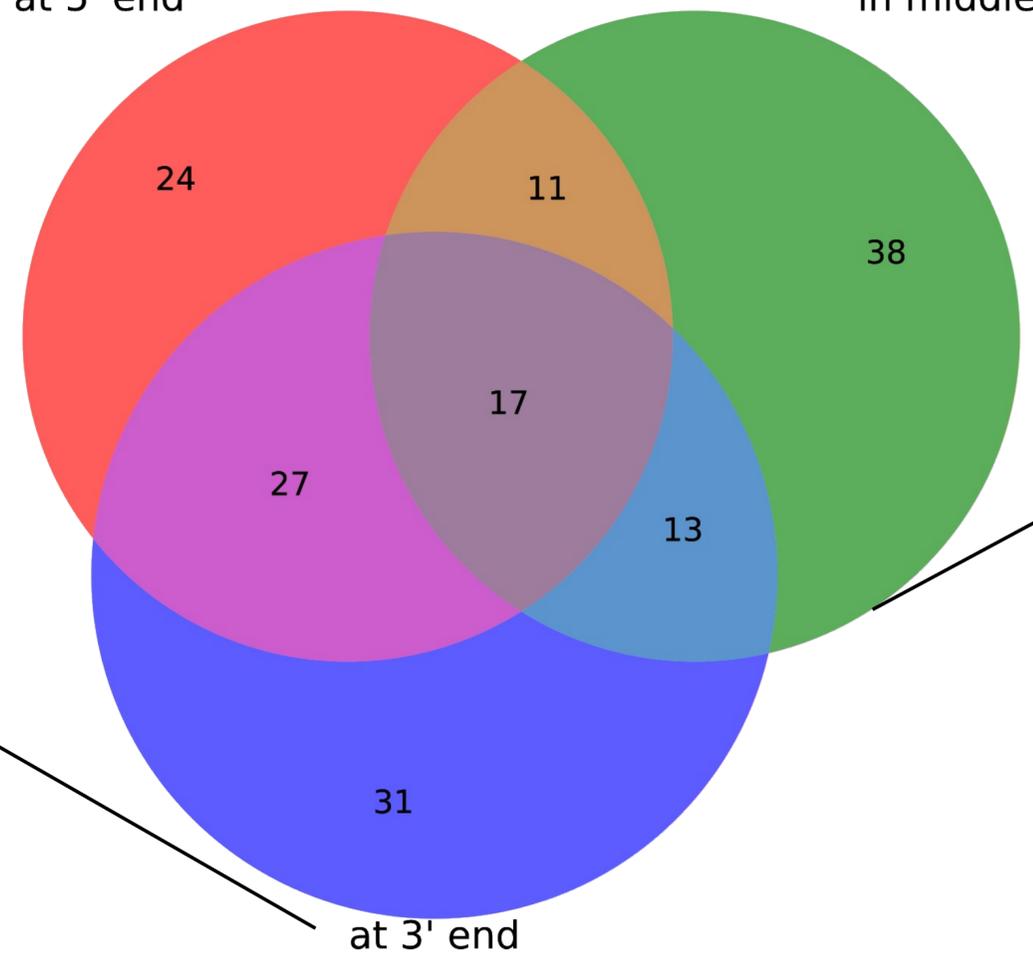


Missing residues

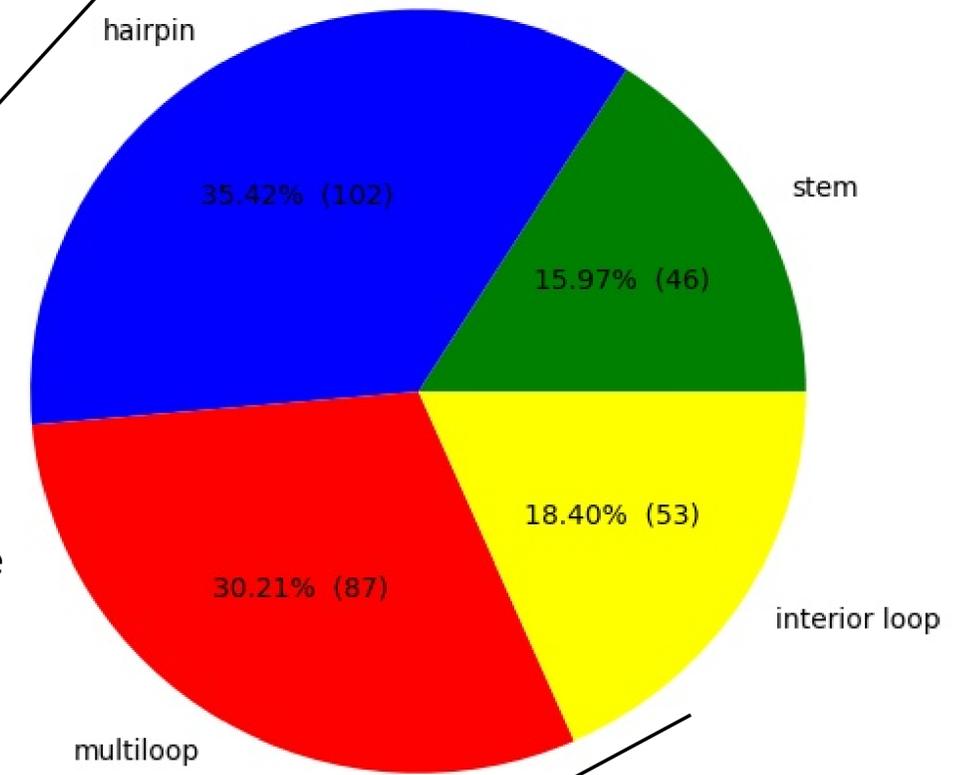
Large PDB structures in representative set



Structures with missing residues



Elements containing missing residues



Co- and Multi-fold structures

Base-pairs between multiple chains



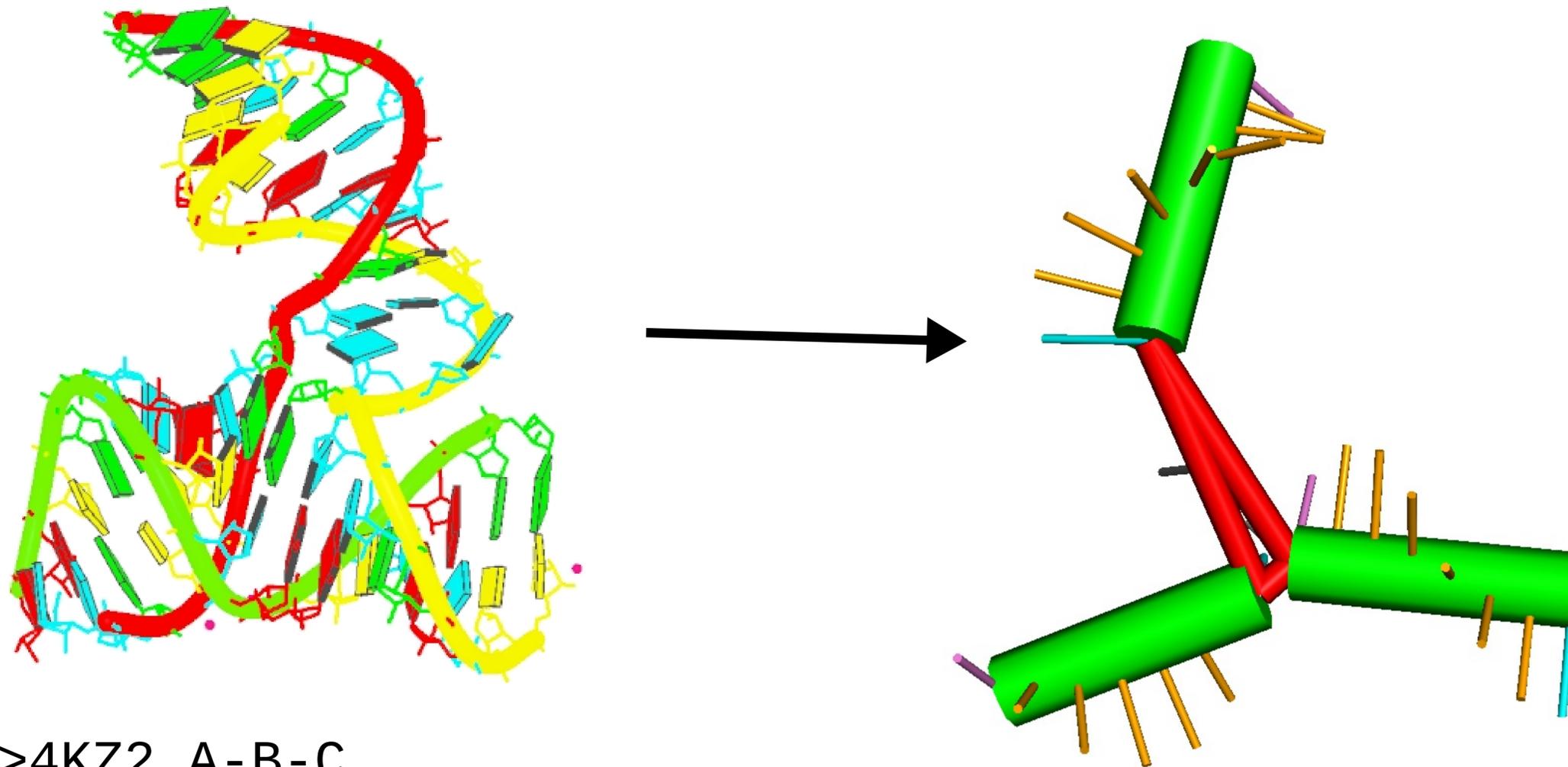
Co- and Multi-fold structures

FORGI 2.0

- * Arbitrary number of cutpoints ('&') in sequence
- * Indexing and slicing works
- * Each Bulge-Graph object represents one connected component

Co- and Multi-fold structures

Base-pairs between multiple chains

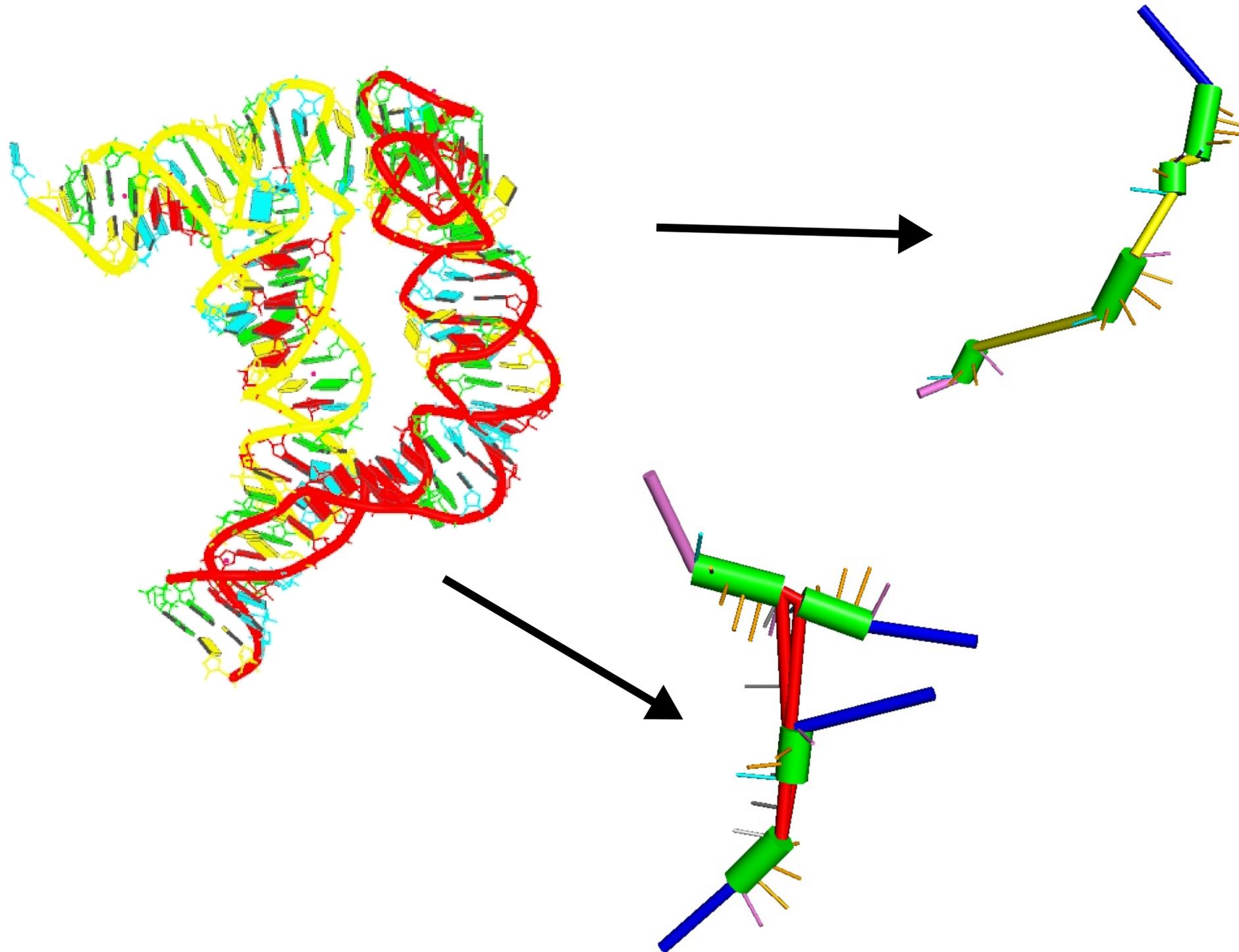


>4KZ2_A-B-C

UUGCCAUGUGUAUGUGGG&CCCACAUACUUUGUUGAUCC&GGAUCAAUCAUGGCAA

((((((((. ((((((((&))))))))) . . . ((((((((&)))))))))

Multiple connected components



FORGI 2.0

* 1 connected component per BulgeGraph object

Other improvements

- *) All scripts use the same commandline arguments for RNA parsing.
- *) Choose between MC-Annotate and DSSR
- *) Full support of MMCIF file format (coming soon)

What can forgi do for YOU?

FORGI 2.0

- * Convert between RNA file formats (bpseq, fasta, pdb, dotplot)
- * Simplified 3D structure operations:
RMSD, angle between stems, ...
- * Select secondary structure elements: bulges, interior loops, ...

Thank you

Peter Kerpedjiev
Gregor Entzian
Irene Beckmann
Ivo Hofacker
TBI



FORGI 2.0

www.github.com/ViennaRNA/forgi