RNA Structures Description Standards

Jacek Śmietański
Institute of Computer Science,
Jagiellonian University
ul. Łojasiewicza 6, 30-348, Kraków, Poland

Secondary structure:

a) Text formats:

- dot-brackets, one-line
- dot-brackets, multi-line
- pair list
- CT (connection table)
- BPSEQ (base-pairs sequence)
- RNAML (XML)

Dot-brackets, one-line

```
AGUCGCAUUACAACAU
.(...).([(..].))
```

Dot-brackets, multi-line

```
AGUCGCAUUACAACAU
.(...).(.(....))
.........(...)...
```

Pair list

AGUCGCAUUACAACAU

{(2,6),(8,16),(9,13),(10,15)}

CT

```
>first_seq.ct
1 A 0 2 0 1
2 G 1 3 10 2
3 U 2 4 0 3
4 C 3 5 12 4
5 G 4 6 0 5
6 C 5 7 14 6
7 A 6 8 0 7
8 U 7 9 16 8
9 G 8 10 0 9
10 C 9 11 2 10
11 A 10 12 0 11
12 U 11 13 4 12
13 G 12 14 0 13
14 C 13 15 6 14
15 A 14 16 0 15
16 G 15 17 8 16
17 C 16 18 0 17
```

BPSEQ

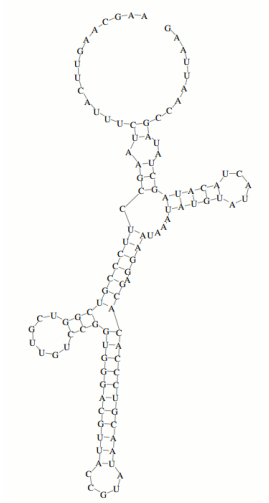
```
>first_seq.bpseq
1 A 0
2 G 10
3 U 0
4 C 12
5 G 0
6 C 14
7 A 0
8 U 16
9 G 0
10 C 2
11 A 0
12 U 4
13 G 0
14 C 6
15 A 0
16 G 8
17 C 0
```

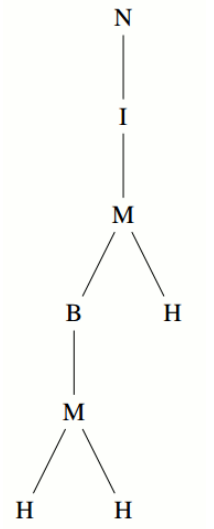
RNAML

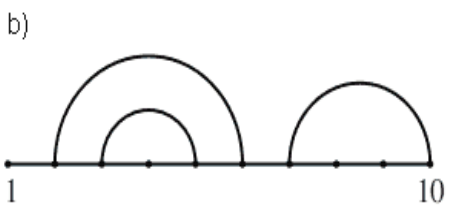
```
<rnaml>
    <molecule>
        <identity>
            <name>seq_name</name>
        </identity>
        <sequence length="17">
            <seq-data>AGUCGCAUGCAUGCAGC</seq-data>
        </sequence>
        <structure>
            <base-pair>
                <base-id-p5>
                    <base-id>
                        <position>2</position>
                    </base-id>
                </base-id-p5>
                <base-id-p3>
                    <base-id>
                        <position>10</position>
                    </base-id>
                </base-id-p3>
            </base-pair>
            <base-pair>
                <base-id-p5>
                    <base-id>
                        <position>4</position>
                    </base-id>
                </base-id-p5>
```

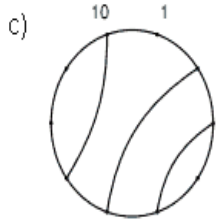
b) Visualisation:

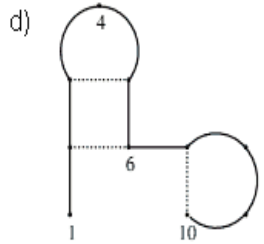
- plain
- tree
- arc
- dome
- circle
- graph











Non-canonical interactions:

a) Classification standards

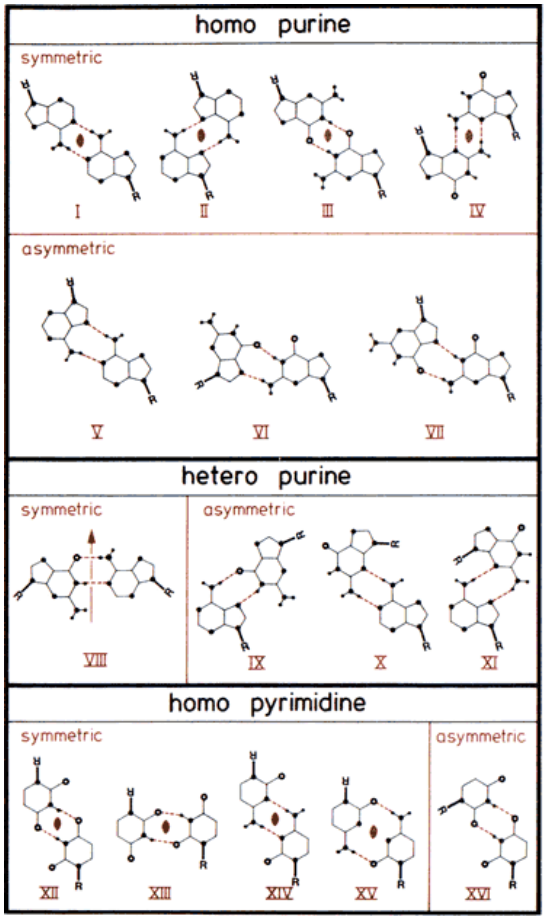
- Saenger: 28 types selected according to composition and symmetry, consisting of only purine, only pyrimidine, or mixed purine/pyrimidine pairs and asymmetric or symmetric base-pairs.

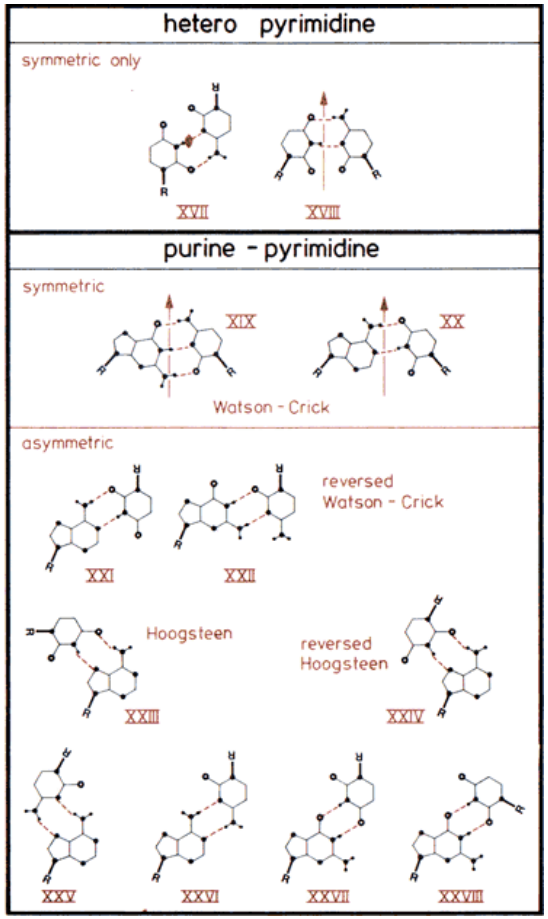
- Leontis-Westhof: 12 classes constructed on the basis of the planar edge-to-edge interactions that involve one of three distinct edges: the Watson-Crick edge, the Hoogsteen edge, and the Sugar edge. Bases can interact in either of two orientations with respect to the glycosidic bonds, cis or trans relative to the hydrogen bonds. This gives 12 basic geometric types.

b) File formats:

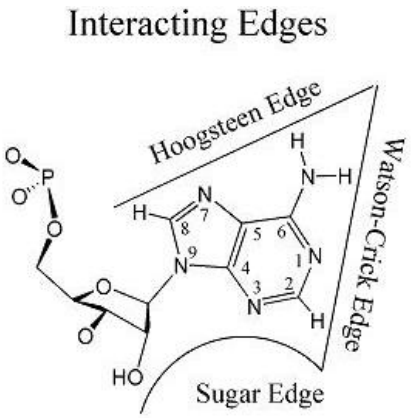
Each program (eg. FR3D, RNA View, MCAnnotate) has its own format. Non compatible with each other and hard to automatic processing.

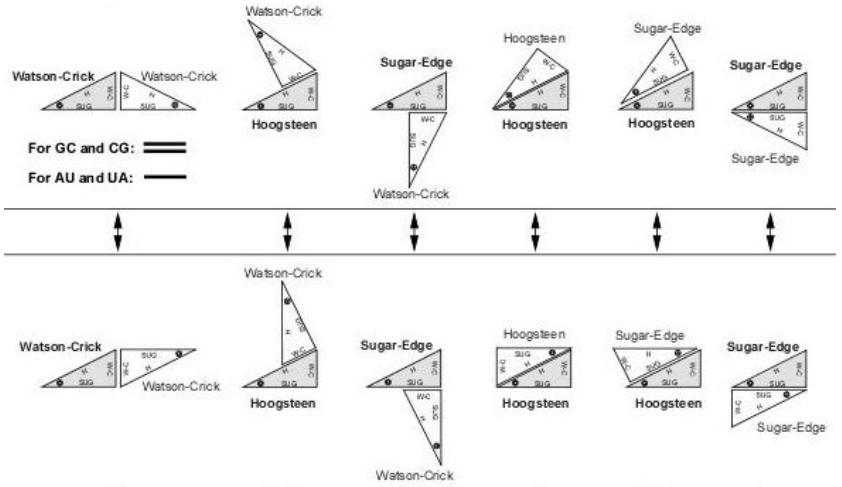
Saenger notation





Leontis-Westhof notation





FR3D

```
 1 G    1(A) - C    8(B) -    cWW  -    0
 2 C    2(A) - G    7(B) -    cWW  -    0
 3 G    3(A) - G    6(B) -    cWH  -    0
 4 G    4(A) - C    5(B) -    cWW  -    0
 5 C    5(A) - G    4(B) -    cWW  -    0
 6 G    6(A) - G    3(B) -    cHW  -    0
 7 G    7(A) - C    2(B) -    cWW  -    0
 8 C    8(A) - G    1(B) -    cWW  -    0
 9 G    1(B) - C    8(A) -    cWW  -    0
10 C    2(B) - G    7(A) -    cWW  -    0
11 G    3(B) - G    6(A) -    cWH  -    0
12 G    4(B) - C    5(A) -    cWW  -    0
13 C    5(B) - G    4(A) -    cWW  -    0
14 G    6(B) - G    3(A) -    cHW  -    0
```

RNA View

```
BEGIN_base-pair
    1_27, A:    1 G-C    27 A: +/+ cis       XIX
    2_26, A:    2 G-U    26 A: W/W cis       XXVIII
    3_25, A:    3 C-G    25 A: +/+ cis       XIX
    4_24, A:    4 U-A    24 A: -/- cis       XX
   13_18, A:   13 U-U    18 A: W/W cis       XVI
   10_21, A:   10 U-U    21 A: W/W cis       !1H(b_b)
   14_16, A:   14 U-A    16 A: S/S tran      !(s_s)
END_base-pair
The total base pairs =   9 (from   27 bases)
```

MCAnnotate

```
Base-pairs ----------------------------------------------------
A1-A469 : G-C Bs/O2P pairing
A4-A5 : C-G O2P/Hh adjacent_5p pairing
A5-A386 : G-A Ss/C8 pairing parallel cis one_hbond
A5-A469 : G-C O2P/Ww O2P/Bh pairing
A17-A382 : U-A Ww/Ww pairing antiparallel cis XX
A18-A381 : G-C Ww/Ww pairing antiparallel cis XIX
A20-A379 : U-A Ww/Ww pairing antiparallel cis XX
```

Three dimensional structure:

- **PDB:** old, deprecated, but still de-facto standard; plain text format with many limitations; cannot store large structures;

many variations and inconsistences with official specification exists;

- PDBx/mmCif: recommended by Protein Data Bank but still not so popular;

most of applications and databases cannot cope with it;

- PDBML: XML format for automatic management purposes;

- MMTF: new format with high level of binary compression;

Non human readable but save a lot of time and transfer when downloading and save disk space in local repository.

PDB

```
MODEL        1
ATOM      1  P      G A   1      18.588   2.270   2.042  1.00  0.00           P
ATOM      2  OP1    G A   1      20.026   2.544   1.825  1.00  0.00           O
ATOM      3  OP2    G A   1      17.796   3.297   2.754  1.00  0.00           O
ATOM      4  O5'    G A   1      17.910   2.026   0.613  1.00  0.00           O
ATOM      5  C5'    G A   1      18.042   0.761  -0.032  1.00  0.00           C
ATOM      6  C4'    G A   1      16.679   0.221  -0.402  1.00  0.00           C
ATOM      7  O4'    G A   1      15.992  -0.639   0.072  1.00  0.00           O
ATOM      8  C3'    G A   1      15.710   1.270  -0.939  1.00  0.00           C
ATOM      9  O3'    G A   1      16.128   1.681  -2.238  1.00  0.00           O
ATOM     10  C2'    G A   1      14.369   0.589  -0.980  1.00  0.00           C
ATOM     11  O2'    G A   1      14.275   0.173  -2.323  1.00  0.00           O
```

mmCif

```
_atom_site.auth_asym_id
_atom_site.auth_atom_id
_atom_site.pdbx_PDB_model_num
ATOM 1     P P      . G A 1 1  ? 18.588 2.270  2.042   1.00 0.00 ? ? ? ? ? ? 1  G A P      1
ATOM 2     O OP1    . G A 1 1  ? 20.026 2.544  1.825   1.00 0.00 ? ? ? ? ? ? 1  G A OP1    1
ATOM 3     O OP2    . G A 1 1  ? 17.796 3.297  2.754   1.00 0.00 ? ? ? ? ? ? 1  G A OP2    1
ATOM 4     O "O5'"  . G A 1 1  ? 17.910 2.026  0.613   1.00 0.00 ? ? ? ? ? ? 1  G A "O5'"  1
ATOM 5     C "C5'"  . G A 1 1  ? 18.042 0.761  -0.032  1.00 0.00 ? ? ? ? ? ? 1  G A "C5'"  1
ATOM 6     C "C4'"  . G A 1 1  ? 16.679 0.221  -0.402  1.00 0.00 ? ? ? ? ? ? 1  G A "C4'"  1
ATOM 7     O "O4'"  . G A 1 1  ? 15.992 -0.639 0.072   1.00 0.00 ? ? ? ? ? ? 1  G A "O4'"  1
ATOM 8     C "C3'"  . G A 1 1  ? 15.710 1.270  -0.939  1.00 0.00 ? ? ? ? ? ? 1  G A "C3'"  1
ATOM 9     O "O3'"  . G A 1 1  ? 16.128 1.681  -2.238  1.00 0.00 ? ? ? ? ? ? 1  G A "O3'"  1
ATOM 10    C "C2'"  . G A 1 1  ? 14.369 0.589  -0.980  1.00 0.00 ? ? ? ? ? ? 1  G A "C2'"  1
ATOM 11    O "O2'"  . G A 1 1  ? 14.275 0.173  -2.323  1.00 0.00 ? ? ? ? ? ? 1  G A "O2'"  1
```

PDBML

```
<PDBx:atom_siteCategory>
   <PDBx:atom_site id="1">
      <PDBx:B_iso_or_equiv>0.00</PDBx:B_iso_or_equiv>
      <PDBx:Cartn_x>18.588</PDBx:Cartn_x>
      <PDBx:Cartn_y>2.270</PDBx:Cartn_y>
      <PDBx:Cartn_z>2.042</PDBx:Cartn_z>
      <PDBx:auth_asym_id>A</PDBx:auth_asym_id>
      <PDBx:auth_atom_id>P</PDBx:auth_atom_id>
      <PDBx:auth_comp_id>G</PDBx:auth_comp_id>
      <PDBx:auth_seq_id>1</PDBx:auth_seq_id>
      <PDBx:group_PDB>ATOM</PDBx:group_PDB>
      <PDBx:label_alt_id xsi:nil="true" />
      <PDBx:label_asym_id>A</PDBx:label_asym_id>
      <PDBx:label_atom_id>P</PDBx:label_atom_id>
      <PDBx:label_comp_id>G</PDBx:label_comp_id>
      <PDBx:label_entity_id>1</PDBx:label_entity_id>
      <PDBx:label_seq_id>1</PDBx:label_seq_id>
      <PDBx:occupancy>1.00</PDBx:occupancy>
      <PDBx:pdbx_PDB_model_num>1</PDBx:pdbx_PDB_model_num>
      <PDBx:type_symbol>P</PDBx:type_symbol>
   </PDBx:atom_site>
```

MMTF