

## Notes on RnaO.owl

Ribonucleic acid (RNA) molecules consists of heterocyclic nucleobases covalently bonded to ribose rings. The ribose rings are connected covalently to other ribose rings through phosphate groups. Thus the primary unit is the nucleotide, an assembly of a nucleobase, a ribose and a phosphate. The combination of base and ribose is called a nucleoside.

Under physiological conditions, single strands of RNA consisting of tens to hundreds or even thousands of nucleotides may fold in part or as a whole to form specific three-dimensional structures that depend on the solution conditions and temperature. The folding of the RNA chain brings together pairs of short sequence segments that are Watson-Crick complementary to form anti-parallel double helices consisting of Watson-Crick basepairs stacked one each other. These are the simplest and most regular kinds of RNA 3D motifs. The set of Watson-Crick paired helices comprise the secondary structure of the RNA. Some RNA molecules can form more than one secondary structure and can be induced by appropriate perturbations to switch between them. In RNAO, we treat basepairing as a binary relation as explained below.

The looping of the chain forms other motifs called hairpin loops. There are many kinds of hairpin loops, some of which are structured by specific sets of interactions, including base-stacking and base-pairing as well as base-phosphate interactions. The basepairs in hairpin loops often involve hydrogen-bonding between Hoogsteen or Sugar edges of the bases in addition to the Watson-Crick edge. Basepairs that involve one or more non-Watson-Crick edge (i.e. Hoogsteen or Sugar edges) are called non-Watson-Crick basepairs. Segments of sequence joining two helices can also form structured motifs comprising non-Watson-Crick basepairs. These are called internal loops. Finally, junction loops result when three or more helical segments are joined together. Junction loops frequently exhibit stacking between pairs of helices across the junction. An example of a three-way junction is the hammerhead ribozyme. The junction is the catalytically active center of the molecule. RNA 3D motifs frequently comprise functional centers of RNA molecules, that can catalyze reactions, as in the case of the hammerhead ribozyme junction motif, bind small molecules or proteins or mediate RNA tertiary interactions. The hammerhead ribozyme also comprises two loop motifs at the ends of two of its helical arms that form a tertiary interaction to force the three-way junction into the enzymatically active conformation.

RNA 3D motifs can recur time and again in different RNA molecules encoded by genes from different families in very different organisms. Further, some of these motifs have different functions on account of their three-dimensional structure. Motifs combine to define characteristic RNA folds or domains.

The plan of the RNA Ontology Consortium is to describe the three-dimensional structures of RNA molecules in an ontology. One promising route is to start with the basepair classification proposed by (Leontis and Westhof, 2001). The initial focus is to convert this classification into OWL by performing a formal ontological analysis of RNA basepairing. This approach will be extended to other pairwise interactions (i.e. *binary relations*) between nucleotides or their constituents, for example, base-stacking and base-phosphate interactions. Base-phosphate interactions are energetically favorable interactions between specific edges of the nucleobases and the oxygen atoms of the phosphate of the same or a different nucleotide.

### Boundaries

The first ontological problem is that of boundaries. We may identify in a nucleotide firstly a *bona fide* boundary which can be drawn at some contour of the electron density of the unpaired nucleotide, and secondly a pair of *fiat* boundaries which serve to divide the nucleotide from its immediate neighbors in the covalent RNA chain. Our chief interest is in the *bona fide* boundaries of the unpaired nucleotide, as it is along these that hydrogen atoms and heteroatoms in spatially adjacent nucleotides undergo hydrogen bonding. Leontis and Westhof treat the nucleoside as a triangle with three edges, the Watson-Crick edge, along the DNA analogue of which base-pairing takes place in a DNA double helix, the Hoogsteen edge and the Sugar edge. These three are *fiat* parts of the *bona fide* boundary.

### Covalent bonding relations

The most fundamental relations are those along the backbone, across the *fiat* boundaries that separate

nucleotides. By convention, the sense of the chain is 5'-to-3' (which is the direction of enzymatic chain synthesis), that is to say that the phosphate groups connect the 5' and 3' carbon atoms of the ribose rings of successive nucleotides in RNA chains. (Note that each phosphate connects 3' carbon of a nucleotide to the 5' carbon of the next nucleotide in the chain.) We therefore require relations expressing the covalent connectivities of nucleotides along the chain to define motifs. We anticipate that the "covalentlyBondedTo" relation will be defined in ChEBI and so we can specialize it to RNA to cover all possible relations. The most common are X covalently\_bonded\_5'\_to\_3' Y and the inverse relation Y covalently\_bonded\_3'\_to\_5' X. These relations are exclusive in the sense that (X covalently\_bonded\_5'\_to\_3' Y) AND (X covalently\_bonded\_5'\_to\_3' Z) IMPLIES Y equal\_to Z. The domain and range of these relations is nucleotides, although the relation may have to be extended to include abasic sites -- positions in RNA or DNA molecules that have lost the nucleobase, although the sugar and phosphate remain, at least temporarily. So perhaps the domain and range should be entities that include at least a sugar and a phosphate. The "covalently\_bonded\_to" relation can be extended to allow for 5' to 2' bonding as occurs during intron splicing or 3' to 2' bonding as occurs during some forms of strand cleavage.

This approach is based on that of (Villanueva-Rosales and Dumontier, 2007), who classify simple molecules into functional classes (halides, carboxylic acids and so forth) according to the atoms they contain and the connectivity of those atoms.

### Basepair relations

The model we choose for the RNA ontology is the following: a nucleotide is a *fiat* part of an RNA molecule that has itself two further *fiat* parts, the nucleoside and the phosphate group. The nucleoside, in turn, has two *fiat* parts, the ribose ring and the nucleobase. Each edge of the nucleoside can interact with edges of other nucleosides by forming hydrogen bonds. According to the LW scheme, to a first approximation:

(1) *each edge of a nucleoside may interact only with a single edge of a different nucleoside*

This is an oversimplification, as there are special cases in which a single edge of a base can interact simultaneously with two different nucleobases. One case involves the Sugar edge of a purine (usually A) interacting with the sugar edges of two other bases which themselves form a cWW basepair. In this case we can say: If X pairsWithTSS Y and X pairsWithCSS Z then (Y pairsWithCWW Z). These special cases will be added in version 2.0 of RNAO.

With six different combinations of edge interaction (WC-WC, H-H, S-S, WC-H, WC-S and H-S), and two relative orientations (*cis* and *trans*) for the interaction of the nucleosides/nucleobases, there result twelve basepairing classes in the LW scheme.

Three of these edge interactions are symmetric (WC-WC, H-H, S-S) and with two relative orientations that gives us six symmetric relations, which are self-inverse. There are three asymmetric interactions, each of which has an inverse, and all six of these can occur with either relative orientation, resulting in a grand total of eighteen basepairing relations.

We can express statement (1) above in OWL terms by declaring each of these relations to be *disjoint* from other relations. If we declare the relations pairsWithCWH and pairsWithCWW to be disjoint then this means that

if X pairsWithCWH Y then there is no Z such that X pairsWithCWW Z

In practice we define pairsWithCWH to be disjoint with all other pairing relations where the first base in the relationship is presenting a Watson-Crick edge to the second.

1	pairsWithCWW	self-inverse	disjoint with other W-first pairings disjoint with other W-second pairings
2	pairsWithTWW	self-inverse	disjoint with other W-first pairings disjoint with other W-second pairings
3	pairsWithCWH	inverse of pairsWithCHW	disjoint with other W-first pairings disjoint with other H-second pairings
4	pairsWithTWH	inverse of pairsWithTHW	disjoint with other W-first pairings disjoint with other H-second pairings
5	pairsWithCHW	inverse of pairsWithCWH	disjoint with other H-first pairings disjoint with other W-second pairings
6	pairsWithTHW	inverse of pairsWithTWH	disjoint with other H-first pairings disjoint with other W-second pairings
7	pairsWithCHH	self-inverse	disjoint with other H-first pairings disjoint with other H-second pairings
8	pairsWithTHH	self-inverse	disjoint with other H-first pairings disjoint with other H-second pairings
9	pairsWithCHS	inverse of pairsWithCSH	disjoint with other H-first pairings disjoint with other S-second pairings
10	pairsWithCSH	inverse of pairsWithCHS	disjoint with other S-first pairings disjoint with other H-second pairings
11	pairsWithTHS	inverse of pairsWithTSH	disjoint with other H-first pairings disjoint with other S-second pairings
12	pairsWithTSH	inverse of pairsWithTHS	disjoint with other S-first pairings disjoint with other H-second pairings
13	pairsWithCSS	self-inverse	disjoint with other S-first pairings disjoint with other S-second pairings (this will change in v2)
14	pairsWithTSS	self-inverse	disjoint with other S-first pairings disjoint with other S-second pairings (this will change in v2)
15	pairsWithCSW	inverse of pairsWithCWS	disjoint with other S-first pairings disjoint with other W-second pairings
16	pairsWithCWS	inverse of pairsWithCSW	disjoint with other W-first pairings disjoint with other S-second pairings
17	pairsWithTSW	inverse of pairsWithTWS	disjoint with other S-first pairings disjoint with other W-second pairings
18	pairsWithTWS	inverse of pairsWithTSW	disjoint with other W-first pairings disjoint with other S-second pairings

They are not transitive (if  $X R Y$  and  $X R Z$  then  $X \text{ not } R Z$ ), and only the self-inverse ones are symmetric (if  $X R Y$  then necessarily  $Y R X$ ). Where appropriate we define them to be *functional relations*, in the

OWL sense that if  $X R Y$  and  $X R Z$ , then  $Y = Z$ .

The test for the ontology correctly classifying basepairings according to the LW scheme is as follows:

1. We define necessary and sufficient conditions for the occurrence of a basepair in one of the LW classes (currently these are written in terms of nucleobases, but nucleotides contain nucleobases, so all is well). These rely on the eighteen relations (*pairsWithCWW*, etc.) above. For example, here written in the English-like Manchester syntax for OWL (Horridge *et al.* 2006):

Family1BasePair = hasPart only (Nucleobase and pairsWithCWW some Nucleobase)

2. We then create test classes, which contain, for example an A (which is a child of nucleobase) bound in a particular way to a U, and then run the reasoner (in this case FaCT++ as built into Protege).

3. In the test file we explicitly define classes, for example CWWAGBasePair is defined as follows:

CWWAGBasePair = hasPart only (A and pairsWithCWW some G)

On running the classifier we see that CWWAGBasePair is correctly defined as a Family1BasePair (because A is a Nucleobase and G is a Nucleobase and two *cis*-WC-WC-paired nucleobases constitute a Family1BasePair). It is not classified as a CanonicalBasePair, because it fails to satisfy any of the four conditions for being a CanonicalBasePair.

## Motifs

On the basis of the covalentlyBondedTo relations and the pairsWith relations it is possible to create rudimentary definitions of most motifs. Motifs can be understood as collections of nucleotides that form highly connected networks of binary relations, including covalent connectivity, base-pairing, base-stacking and base-phosphate relations.

The next step is to define a motif. For this we need relations along the backbone, across the *fiat* boundaries that separate nucleotides. The usual sense of the chain is 5'-to-3', that is to say that the phosphate groups connect the 5' and 3' carbon atoms of the ribose rings of successive nucleotides in RNA chains.

In the current RNAO.owl we simply assert that a Motif is a FiatObjectPart and defer the precise definition to a later version. As a 'straw man' model of a GNRA tetraloop motif (parts of an RNA molecule that are crystallographically a GNRA tetraloop may contain more than four nucleotides, some of which are bulged out), we define it as follows:

*GNRA TetraloopMotif* = hasPart some (Nucleobase and fivePrimeTo some (G and fivePrimeTo some (Nucleobase and fivePrimeTo some (Nucleobase and fivePrimeTo some (A and fivePrimeTo some (Nucleobase and pairsWithCWW some Nucleobase) and pairsWithTHS some G))) and pairsWithTSH some A) and pairsWithCWW some Nucleobase)

all of which is meant to capture that the specific GNRA tetraloop motif is bounded by a standard *cis*-Watson-Crick pair, and has a *trans*-Hoogsteen-sugar pair between the G and the A. Then as a minimal test we take a known structure that contains a tetraloop motif. We also include a tertiary interaction, a *cis*-Sugar-WC pairing in order to demonstrate that the reasoner is not distracted by additional bases. This is the GAAA hairpin loop starting at base 1363 in the *T. thermophilus* 23S rRNA (PDB file 2j01):

*GAAA TetraloopTthermophilus* = hasPart some (C1363 and fivePrimeTo some (G1364 and fivePrimeTo some (A1365 and fivePrimeTo some (A1366 and fivePrimeTo some (A1367 and fivePrimeTo some (G1368 and pairsWithCWW some C1363) and pairsWithTHS some G1364))) and pairsWithCSW some G187) and pairsWithTSH some A1367) and pairsWithCWW some G1368)

and run the classifier. Fortunately this is indeed classified as a GNRA tetraloop motif. But the next stage is to convert some real data into OWL and look for false positives.

While above we have written definitions in the English-like Manchester syntax, OWL itself is written in RDF/XML, so it is straightforward to generate RNAO-specific OWL representations of a given RNA structure from a plain text file using for example Perl or XSLT.

## Backbone conformers

The backbone in RNA molecules is a chain of covalently-bonded atoms, some of which are parts of the phosphate group of nucleotide (O5', P, O3') and others of which are parts of the ribose rings (C3'-C4'-C5'). We are interested in backbone conformation in RNA for two reasons: (1) particular RNA motifs are often statistically associated with particular backbone conformations, and (2) they provide sites for interaction with ions, proteins, small molecules, proteins, and other nucleic acids or segments of the same RNA.

We therefore tentatively outline a method for representing backbone conformers in the RNAO. We could continue the treatment above where interactions between bases are represented by relations. We take as an example GNRA tetraloop motif written in terms of the Richardson backbone configuration notation (Richardson *et al.* 2008):

N1aG1gN1aR1aA1cN1a

The notations, 1a, 1g, *etc.*, are a shorthand for the dihedral angles in a 'suite'. A 'suite' consists of two half-nucleotides, and as such is an alternative way of dividing up the RNA molecule. Each classification, 1a, 1g, *etc.*, indicates a cluster in seven-dimensional space of the seven dihedral angles. The first character is determined by the C4'-C3', C3'-O3' and O3'-P dihedral angles, whereas the second is determined by the P-O5', O5'-C5' and C5'-C4' angles.

The most natural thing to do is to treat the different suite conformations in terms of specializations of the covalent bonding relations, for example `fivePrimeTo` can be specialized in this case into `fivePrimeTo1a`, `fivePrimeTo1c` and `fivePrimeTo1g`. There are hence 46 children of `fivePrimeTo`, though depending on the application it may be useful to include intermediate terms which might include `fivePrimeToC2PrimeEndoPuckerC3PrimeEndoPucker` or `fivePrimeToC3PrimeEndoPuckerC3PrimeEndoPucker`. We defer the question of the ontological status of individual dihedral angles and conformational isomers more generally.

A backbone site, therefore, would be defined as a site that has at least one physical (*bona fide*) boundary, a length of backbone. Hence a `OneLsqbIntercalationSite` would be any site which has as its *bona fide* boundary two objects that are linked by a `fivePrimeTo1Lsqb` (**1**[ ]) relation. However the ontological status of suites and heminucleotides, as an alternative exhaustive partition of the RNA molecule, needs further work.

## Conclusions and next steps

We have presented a rudimentary version of RNAO which can classify basepairings into the twelve categories of Leontis and Westhof and can distinguish parts of RNA structures that contain a given motif (here the GNRA tetraloop) from at least some parts of RNA structures that don't. We have also suggested a tentative method for incorporating backbone configurations into the ontology.

The most pressing next step is to test the ontology as it stands on more data, both 'real' data from the PDB and further motifs, and work out whether OWL-DL is powerful enough for what we want to do. We also need to flesh out the tentative approach to backbones mentioned above. We have not yet considered base-base stacking and base-phosphate relations, and we intend to do so in the near future. Lastly, while the version of RNAO we have at present contains formal, computable, machine-readable definitions, it has no human-readable definitions, and writing these is a must if RNAO is to be accepted by the RNA bioinformatics community.

## Links:

Evolution Ontology workshop:

[http://bioontology.org/wiki/index.php/Evolutionary\\_Biology\\_and\\_Ontologies](http://bioontology.org/wiki/index.php/Evolutionary_Biology_and_Ontologies)

**For the latest version of RnaO see the OWL file:**

<http://code.google.com/p/rnao/downloads/list>

## References

M. Horridge, N. Drummond, J. Goodwin, A. Rector, R. Stevens and H. H. Wang. 2006. The Manchester OWL Syntax, in *OWL Experiences and Directions Workshop, 2006*.

N. B. Leontis and E. Westhof. 2001. Geometric nomenclature and classification of RNA base pairs, *RNA*, 2001, **7**, 499-512.

J. S. Richardson, B. Schneider, L. W. Murray, G. J. Kapral, R. M. Immormino, J. J. Headd, D. C. Richardson, D. Ham, E. Hershkovits, L. D. Williams, K. S. Keating, A. M. Pyle, D. Micallef, J. Westbrook and H. M. Berman. 2008. RNA backbone: Consensus all-angle conformers and modular string nomenclature (an RNA Ontology Consortium contribution), *RNA*, 2008, **14**, 465-481.

N. Villanueva-Rosales and M. Dumontier. 2007. Describing chemical functional groups in OWL-DL for the classification of chemical compounds, in *OWL: Experiences and Directions (OWLED 2007)*, co-located with European Semantic Web Conference (ESWC2007), Innsbruck, Austria.