RNA Secondary Structure Prediction

Roger Sayle and Noel O'Boyle
NextMove Software, Cambridge

Abstract
This white paper considers the application of RNA secondary structure algorithms to the prediction and assignment of hydrogen bonding base pairs to RNA sequences. The secondary structures of both RNAs and peptides are implicitly determined by their primary sequences; and whilst a scientist can specify the sequence they synthesize [or express], the shape and fold that this adopts is driven by physics. The use of secondary structure algorithms allows the internal hydrogen bonding of a sequence to annotated (if previously unassigned) or verified (if previously specified).

Use Cases
The two motivating use cases we'll are:
[1] The automatic annotation of the hydrogen bonding layer to RNA in Pistoia's HELM.
[2] The graphical depiction/display of RNA sequences showing such hydrogen bonding.

Intermediate File Formats
A particularly useful intermediate file format in this work is the Dot-Bracket Notation (or DBN). DBN format represents the base-pairing of a nucleic acid sequence using ".", "(" and ")" characters. This "de facto" standard allows different secondary structure prediction algorithms to be evaluated and used interchangeably. Extensions to the original DBN notation allow the use of braces ("{" and "}") and/or square brackets ("[" and "]") for representing RNA pseudo-knots.

The equivalence between DBN format/notation and the hydrogen-bonding layer of the Pistoia Alliance's HELM notation suggests the potential value of a interconversion between one and the other. RNA HELM should be convertible into FASTA+DBN, and likewise FASTA+DBN should be convertible into RNA HELM. NextMove Software are currently investigating supporting this functionality in Sugar & Splice. It should be noted that interconverting just the RNA sequence is already supported in Sugar & Splice, RDKit and the HELM Java libraries; the challenge is (preserving) the hydrogen bonding layer.

Prediction Algorithms
For our initial evaluation, we investigated the use of the ViennaRNA package, version 2.3.5. This suite of tools includes the "RNAfold" and "RNAcofold" programs.
https://www.tbi.univie.ac.at/RNA/

Case Study 1
As an example of a moderately complex RNA, we'll use the structure to the Phe-tRNA in PDB 4TNA. This molecule contains both non-standard bases and an interesting clover-leaf secondary structure.

The Sugar & Splice IUPAC condensed notation for this molecule is:
P-rGuo-P-rCyd-P-rGuo-P-rGuo-P-rAdo-P-rUrd-P-rUrd-P-rUrd-P-rAdo-P-m2Gua-Ribf-P-rCyd-P-rUrd-P-rCyd-P-rAdo-P-rGuo-P-hUra-Ribf-P-hUra-Ribf-P-rGuo-P-rGuo-P-rGuo-P-rAdo-P-rGuo-P-rAdo-P-rGuo-P-rCyd-P-m22Gua-Ribf-P-rCyd-P-rCyd-P-rAdo-P-rGuo-P-rAdo-P-Cyt-Ribf2Me-P-rUrd-P-Gua-Ribf2Me-P-rAdo-P-rAdo-P-Wyb-Ribf-P-rAdo-P-rYrd-P-m5Cyt-Ribf-P-rUrd-P-rGuo-P-rGuo-P-rAdo-P-rGuo-P-m7Gua-Ribf-P-rUrd-P-rCyd-P-m5Cyt-Ribf-P-rUrd-P-rGuo-P-rUrd-P-rGuo-P-rThd-P-rYrd-P-rCyd-P-rGuo-P-m1Ade-Ribf-P-rUrd-P-rCyd-P-rCyd-P-rAdo-P-rCyd-P-rAdo-P-rGuo-P-rAdo-P-rAdo-P-rUrd-P-rUrd-P-rCyd-P-rGuo-P-rCyd-P-rAdo-P-rCyd-P-rCyd-P-

rAdo

For which the RSEQ (strict) and NSEQ (permissive) sequences are respectively:
GCGGAUUUANCUCAGNNGGGAGAGCNCCAGANUNAANANNUGGAGNUCNUGUGTNCGNUCCACAGAAUUCGCACCA
GCGGAUUUAGCUCAGUUGGGAGAGCGCCAGACUGAAGAUCUGGAGGUCCUGUGTUCGAUCCACAGAAUUCGCACCA

The prediction produced by RNAfold v2.3.5 looks like:

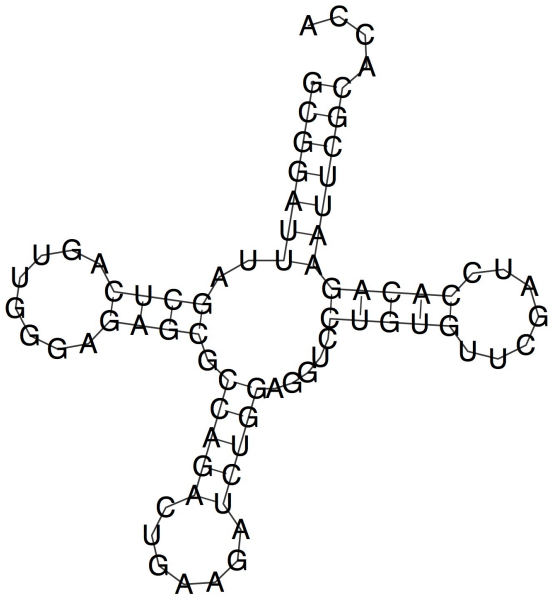
```
gcggauuuagcucaguugggagagcgccagacugaagaucuggagguccuguguucgauccacagaauucgcacca
(((((((..((((........))))).(((((.......)))))).....(((((.......)))))))))).... 
```
length = 76, minimum free energy = -22.40 kcal/mol (@37 deg C)

The proposed RNA depiction for this sequence would be:

5'-GCGGAUUUAGCUCAGUUGGGAGAGCGCCAGACUGAAGAUCUGGAGGUCCUGUGTUCGAUCCACAGAAUUCGCACCA

This uses colored backgrounds to indicate complementary base-pairing (hydrogen bonding) and bold/underline characters to indicate non-standard residues [modified bases and/or sugars]. Variants using characters above/below the main sequence and residue count/position markers are being investigated/evaluated. The major benefit of the above depiction/representation is that it allows several related sequences to be presented to the user at the same time (much like a multiple sequence alignment or a database search result list) and the differences between entries easily identifier. The choice of visualization depends on whether one is presenting one, ten or one hundred sequences to the user, and whether each of these are one, ten or one hundred bases long.

The (default) RNAfold produced depiction looks like:



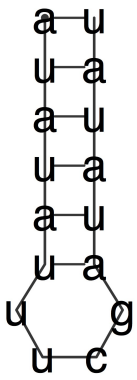
Case Study 2
The second case study 16 base sequence intended to represent a hairpin or shRNA, with the sequence AUAUAUUUCGAUAUAU.

The (Sugar & Splice) IUPAC condensed line notation for this sequence is rAdo-P-rUrd-P-rAdo-P-rUrd-P-rAdo-P-rUrd-P-rUrd-P-rUrd-P-rCyd-P-rGuo-P-rAdo-P-rUrd-P-rAdo-P-rUrd-P-rAdo-P-rUrd.

The proposed RNA depiction for this hairpin sequence would be:
5'-AUAUAUUUCGAUAUAU

Unfortunately, the RNAfold v2.3.5 prediction for this molecule is that it wouldn't form a hairpin at the default temperature of 37 degrees celsius. The short complementary sequences are only 6 pairs long of weak A-U bonds. This hairpin can be found by lowering the temperature (just a single degree is sufficient). In practice, shRNAs are 20-30 bases long, so this example may have been just a fragment. Indeed, showing a set of related sequences where this one doesn't have a colored background, would highlight the potential problem with the thermodynamics.

The RNAfold produced depiction (using --temp=24.0) looks like:



Case Study 3
The third case study is the 20 base sequence of the drug mipomersen. This example is particularly challenging as mipomersen uses non-standard thiophosphate linkages, numerous non-standard base and sugar modification and a combination of ribose and deoxy ribose sugars!

Note: Version 2.1.9h of th ViennaRNA package contained proof-of-concept support for RNA/DNA hybridization, but this functionality hasn't yet been included in the main (trunk) release.

The NSEQ sequence of mipomersen is GCCUCAGTCTGCTTCGCACC

The prediction from RNAfold v2.3.5 is

```
GCCUCAGUCUGCUUCGCACC
.........(((...)))..
```
minimum free energy = -0.30 (@37 deg C)

The proposed RNA depiction for this sequence would be:
5'-**GCCUCAGTCTGCTTCGCACC**

Self Dimerization and Reverse Complementarity
Following on from the above analysis, we investigated whether the ViennaRNA package's RNAcofold program could be use be used to predict duplex structures, by providing two copies of each molecule to the secondary structure prediction algorithm. The input file format is a variant of the FASTA file format, simply using an "&" to separate each component.

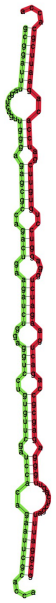
First we investigated case study 2, the shRNA which appeared to be predicted much more stable as a dimer (even at body temperature).

```
auauauuucgauauau&auauauuucgauauau
```

```
..(((((.((((.((((&..))))).)))).)))) ( -6.50)
```

However, a control experiment of confirming that Phe-tRNA (case study 1) was more stable as two clover-leaf monomeric subunits came to the confounding conclusion that this too had a lower energy as a dimer.

```
>4tna_dimer
gcggauuuagcucaguugggagagcgccagacugaagaucuggagguccuguguucgauccacagaauucgcacca&gcg
gauuuagcucaguugggagagcgccagacugaagaucuggagguccuguguucgauccacagaauucgcacca
(((((((((.......((((.((((((..((((..((((((...((((..((((((..((((..
(((((((((....&)))))))).......)))).))))))..)))))..))))))...)))))..))))))..))))..)))
)))))).... (-57.00)
```



Hence, it appears, RNAcofold (with default settings) isn't useful for duplex recognition.

Bibliography

- Ronny Lorenz, Stephan H. Bernhart, Christian Honer zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F. Stadler and Ivo L. Hofacker, “ViennaRNA Package 2.0”, Algorithms for Molecular Biology, Vol. 6, No. 1, p. 26, 2011.
- D.H. Mathews, D.H. Turner, “Prediction of RNA Secondary Structure by Free Energy Minimization”, Current Opinion in Structural Biology, Vol. 16, No. 3, pp. 270-278, June 2006.
- M. Zucker and P. Steigler, “Optimal Computer Folding of Large RNA Sequences using Thermodynamics and Auxilliary Information”, Nucleic Acids Research, Vol. 9, No. 1, pp.

133-148, January 1981.