# Comp 598: Inferring RNA ancestors

## 1 RNA families

Rfam is a database of RNA families available at : `http://rfam.sanger.ac.uk`. Each family is represented by an alignment and a consensus structure. This information can be retrieved in the data stored using the Stockholm format (`http://en.wikipedia.org/wiki/Stockholm_format`). For instance, the alignment of the *Alfalfa mosaic virus RNA 1 5' UTR stem-loop* (`http://rfam.sanger.ac.uk/family/RF00196`) is:

```
# STOCKHOLM 1.0
#=GF ID    AMV_RNA1_SL
#=GF AC    RF00196
#=GF SQ    6

#=GS Alfalfa_mosaic_virus.1 AC    M35975.1/2-36
#=GS Alfalfa_mosaic_virus.2 AC    V00044.1/1-35
#=GS Alfalfa_mosaic_virus.3 AC    M36391.1/2-36
#=GS Alfalfa_mosaic_virus.4 AC    L00163.1/1-35
#=GS Alfalfa_mosaic_virus.5 AC    X00819.1/1-35
#=GS Alfalfa_mosaic_virus.6 AC    M28375.1/1-35

Alfalfa_mosaic_virus.1             GUUUUUAUCUUACACACGCUUGUGCAAGAUAGUUA
Alfalfa_mosaic_virus.2             GUUUUUAUCUUACACACGCUUGUGUAAGAUAGUUA
Alfalfa_mosaic_virus.3             GUUUUCAUCUUACACACGCUUGUGCAAGAUAGUUA
Alfalfa_mosaic_virus.4             GUUUUUAUCUUACACACGCUUGUGUAAGAUAGUUA
Alfalfa_mosaic_virus.5             GUUUUCAUCUUACACACGCUUGUGCAAGAUAGUUA
Alfalfa_mosaic_virus.6             GUUUUUAUCUUAUACACGCUUGUGUAAGAUAGUUA
#=GC SS_cons                       ....<<<<<<<<<<<<....>>>>>>>>>>>>...
#=GC RF                            GuuuuuauCuuacaCacGcuuGuguaaGauaguuA
//
```

The consensus secondary structure is indicated after the tag `#=GC SS_cons`.

## 2 RNA secondary structure

RNA molecule can fold onto themselves and create complex structures. Secondary structure is the simplest level of description of these structures. A RNA secondary structure is the set of base pairs $(i, j)$ found in a single stranded RNA. Allowed base pairs are `C-G`, `A-U` (Watson-Crick) and `G-U` (Wobble). In our cases, crossing base pairs are not allowed (i.e. there is no base pairs $(i, j)$ and $(k, l)$ s.t. $i < k < j < l$). RNA secondary structures can be conveniently represented using a well-bracketed expression, where matching parenthesis indicate the base pairs. An example of such structure is given below.

```
ACUUGAAACGGU
((.((...))))
```

Here, the base pairs are $(1, 12)$, $(2, 11)$, $(4, 10)$, and $(5, 9)$.

## 3 Vienna RNA package

The Vienna RNA package is a collection of program to predict and analyze RNA secondary structures. The software suite is available at `http://www.tbi.univie.ac.at/RNA/`. In this project you will use the `RNAfold` program that has been developed to predict RNA structures from the sequence data, and the `RNAdistance` program that compare secondary structures. Read the documentation to learn more about `RNAfold` and `RNAdistance`.

## 4 Objectives

1. Write a parser that read a Stockholm file, extract the sequence alignment and the consensus secondary structure.

2. Implement the Sankoff algorithm (See `http://www.cs.mcgill.ca/~jeromew/comp561/docs/`). This is a version of Fitch with a slightly more sophisticated scoring scheme. Here you will use a penalty of $-1$ for mutations that conserve purines or pyrimidines, and -2 otherwise.

3. Expand your algorithm to account for gaps using a fifth character. You will use a linear gap penalty of -2 for gaps.

4. Apply your algorithm on a Rfam family of your choice (size of sequences must vary between $50$ and $150$ nucleotides) and calculate ancestor sequences. Trees can be found in Rfam entries. You are allowed to restrict your analysis to a subset of sequences.

5. Calculate the secondary structure of each ancestor sequence with `RNAfold` and calculate their distance from the consensus sequence using `RNAdistance`.

6. Propose an extension of the algorithm that accounts for the base pairing dependencies of the consensus structure (i.e. base pairing properties must be conserved in ancestor sequences).

7. Redo steps $4$ and $5$ with your new algorithm.

8. Apply your algorithm on more Rfam families and identify families where the consensus structure is the most stable on each ancestors (i.e. the frequency of the MFE return by `RNAfold` with the option `-p`). Compare you results with the length, number of sequences and GC-content of the families. Discuss.