Exercise: Structure from multiple RNA sequences

Stefan E. Seemann seemann@rth.dk
Jan Gorodkin gorodkin@rth.dk

1. In the first part of the exercise we will look into the tetrahydrofolate riboswitch example from RNAbook chapter 1 Fig. 2.

(a) Copy single sequence from RNAbook chapter 1 Fig. 2
AGCAGAGUAAGUGCCUACGCGUUAAGUGCCGGAGUACGGGGAGUUGACAUCUGGACGAAA
GCCUUCGGGCUGCGGUGUAAGCAUUGCAUCCCGCUGCU
and run the RNAfold webserver (`http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi`) with default parameters. Compare the predicted RNA structure and the RNA structure in the RNAbook chapter 1 Fig. 2b. How similar are they?

(b) Now go to the RNA annotation database Rfam (`http://rfam.xfam.org/`) and find the corresponding RNA family (RF01831)! Get the first ten sequences of the multiple alignment of the RNA family (Hint: Go to 'Alignment' and download the FASTA alignment in gapped format). Run RNA consensus secondary structure prediction with RNAalifold (`http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAalifold.cgi`). Again use default parameters. How well does the predicted structure agrees with the annotated RNA structure in Rfam (and the RNAbook chapter)? Explain why you see the differences (if any) between the single sequence based structure predictor and the alignment based predictor (Hint: Consider the conservation annotation of RNAalifold).

2. In the second part we work with the output of the WAR server (`http://nar.oxfordjournals.org/content/36/suppl_2/W79.full`). The WAR server runs 14 different RNA alignment and structure prediction methods. These methods use the mutual information content in homologous sequences for RNA secondary structure predictions. From the results of these methods a consensus prediction is made. We are investigating in the following homologous sequences:

```
>seq1
AGAGAGGAGUGAAUAAGGUUGUUCAUAUAAAUUGCAAAUUUAUACAUUUAGGGUUCGAUUCCCUCUUCUCUC
>seq2
ACGAUAGAAACAUGUAUUGGUUCAUGUACUUGCUUUGGGUGUGAGAGUUUGUUAGUUCGAAUCUAACCUAUCCGA
>seq3
UUAAUUGAAGCCAAAAAGAGGCGUAUCACUGUUAAUGAUAUAAUUGAGUAUAAACUCCAAUUAAG
>seq4
AUUCAAAUAGCUUAUAUUUAGAGUAUGACACUGAAGAUGUCAUGGAGAUUAAUUAAUCUUUGAAUA
>seq5
GGAAGCGUGCCUGAAAGUUAAGGACCUCCUUGAUAGGGAGGCUUAUAGGGGUUCAAACCCCCUCACUUCCU
>seq6
CAUUAAGAAGCUAUGCACCAGCACUAGCCUUUUAAGCUAGAGAGAGGGGACACCCUCCCCCCUUAAUGA
```

(a) Which RNA family are we looking at? You may want to align one of the sequences to a

non-redundant nucleotide collection (for example by using nucleotide blast of NCBI BLAST `http://blast.ncbi.nlm.nih.gov/Blast.cgi`).

(b) We have already submitted the sequences to the WAR server. You can examine the results on the following page: `http://genome.ku.dk/resources/war/pages/results.php?id=war_53a9243c1c71d`. Lets investigate the results:

(i) For each program the Matthew's Correlation Coefficient (MCC) of its prediction to the corresponding Rfam annotation is shown in the most right column. The MCC is defined in the RNAbook chapter 1 page 23. What is the MCC of the two methods which start by aligning the sequences using ClustalW? How long time does it take to run these methods?

(ii) What is the MCC of the two methods which start by aligning the sequences using MAFFT? How long time does it take to run these methods?

(iii) What does this tell you about RNAalifold's and Pfold's dependency on the input alignment? Click on the *ClustalW-RNAalifold* and *MAFFT-RNAalifold* programs. Can you by just looking at the alignment see which of the methods makes the best predictions?

(iv) When would you use these kind of methods? If you are working with sequences where the reference alignment/structure is not known how would you determine if these methods could be used?

(c) The Sankoff-style structural alignment methods are slower then the sequence-based alignment methods. Sankoff-style alignment methods are FoldalignM, LocARNA, and Murlet.

(i) What is the MCC of these three methods? How long time did it take to run them?

(ii) When would you use these kinds of methods?

(d) Alignment-based RNA structure prediction methods are taking covariation (compensatory base pair changes) into account. The power of these methods is dependent on the mutual information content of the alignment.

(i) Why are we considering covariation in RNA structure predictions? How is the covariance correlated to the mutual information content of a base pair?

(ii) Click on the *Consensus* program. The inner base pair of the red stem of the consensus alignment calculated with `T-coffee` forms a canonical base pair in all 6 sequences. The base pairs in the 6 sequences are `AU`, `CG`, `AU`, `AU`, `CG` and `CG`. Calculate the mutual information content of this conserved base pair by hand! What would be the mutual information content if all 6 sequences had exactly the same conserved base pair.