Structural Bioinformatics Final Exam

RNA part report

Antonio Ortega Jiménez

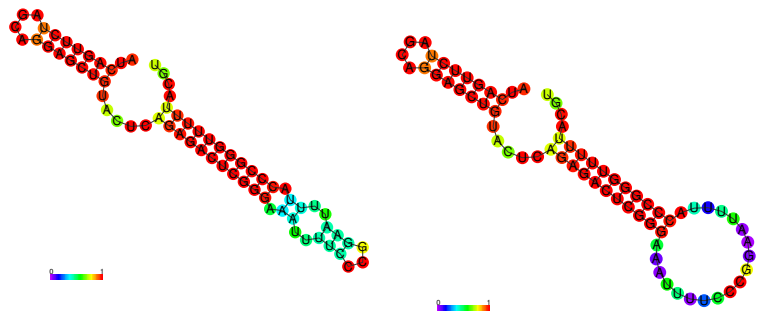
January 27, 2017

Exercise 1: RNAfold prediction on seq1

```
AUCAGUUCUAGCAGGAGCUGUACUCAGAGACUCGGGAAAUUUUCCCGGAAUUUUACCCGGGUUUUUACGU
..(((((((....)))))))......((((((((((.((..(((...))))..)).))))))))))....
```

a) The obtained Minimum Free Energy (MFE) secondary structure consists of a multi-loop with two closed components. The left arm contains a hairpin, whereas the right arm contains another hairpin with a couple of internal loops. The centroid version discards these internal loops and instead shows a simple hairpin with a huge loop. This is due to the low probability exhibited by the base pairs in the terminal region of the hairpin [**?**].

b) RNAfold predicts 18 base pairs with a probability higher than 0.8.



(a) Minimum Free Energy (MFE) structure

(b) Centroid structure

Figure 1: Structures predicted by RNAFold for the query sequence. The color code shows base pair probabilities.

Exercise 2: RNAfold prediction on seq2

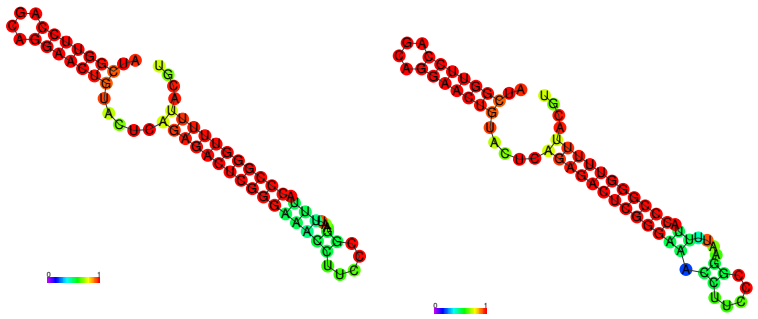
```
AUCGGUUCCAGCAGGAACUGUACUCGGGGGCUCGGGAAACCCUCCCGGGGUUUUACCCGGGUUUUUACGU
..(((((((....))))))).(((((((..(((((((.....)))))))......)))))))........
```

a) *A priori*, this sequence has a similar structure. Base pairs are distributed differently in the right arm, and this has triggered an increase in the hairpin stability. This can be easily noticed:

- The centroid structure now includes the hairpin, because there are more structures in the ensemble that now feature the same hairpin.

- Instead of a highly stable duplex and a highly unstable loop, stability is uniformly distributed across the whole hairpin (see second bulge in right violin from figure 4). This enables the formation of the most interior bonds in the duplex, right before the hairpin's loop, which therefore grows smaller.
- Thermodynamical ensemble prediction parameters show that the structural diversity is reduced and the frequency of the MFE structure increases.

b) RNAfold predicts 20 base pairs with a probability higher than 0.8.



(a) Minimum Free Energy (MFE) structure

(b) Centroid structure

Figure 2: Structure predicted by RNAFold for the query sequence. The color code shows base pair probabilities.

Exercise 3: Hamming distance

Hamming distance between structures is 22, and just 11 at the sequence level.

Exercise 4: Base pair distance

Base pair distance between structures is 30.

Exercise 5: Discussion

Both sequences have a high identity, with a Hamming distance of just 11 at the sequence level. Nevertheless, these differences are not uniformly distributed across the sequence, and thus, they exert different impacts on each arm.

- In spite of the 3 mutations suffered by the left arm, on positions 3 10 and 18, the structure is not changed. This is due to the consistent nature of these mutations, that preserve the affected base pairs:

$$\texttt{A-U} \rightarrow \texttt{G-U} \,|\texttt{U-G} \rightarrow \texttt{C-G} \,|\texttt{U-G} \rightarrow \texttt{U-A}$$

The dotplots (figure 3) support the idea that the left arm has not changed: no significant differences are observed in the left arm area (top left corner), and no alternative structures are observed within the probability space.

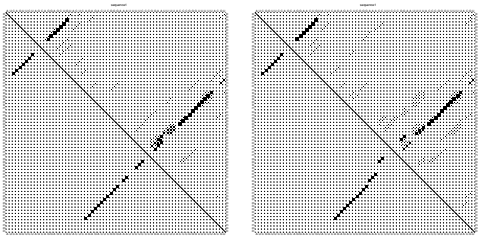


Figure 3: Base pair probability dot plot for seq1 (left) and seq2 (right)

- The remaining 8 mutations are located on the right arm, and though the structural motif is preserved, its internal organization has changed totally.

  On the one hand, base pairs cluster into 2 blocks of the same length (7 base pairs each, 14 in total), instead of 3 blocks of length 11, 2 and 3 (16 in total). This organization leads to a unique internal loop in the duplex center. Since internal loops and bulges destabilize duplexes, as stated in [] it follows that a reorganization that reduces the number of internal loops is bound to make the structure more stable.

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

left arm **right arm**

  On the other hand, this reorganization leads to a redistribution of base pairing probability toward inner regions of the duplex, that become more stable. In other words, in exchange for the loss of 2 base pairs, the remaining ones become more stable and trigger the stabilization of the hairpin's duplex. As a result, the whole hairpin in the right arm is more probable. This phenomenon can be easily observed in figure 4.

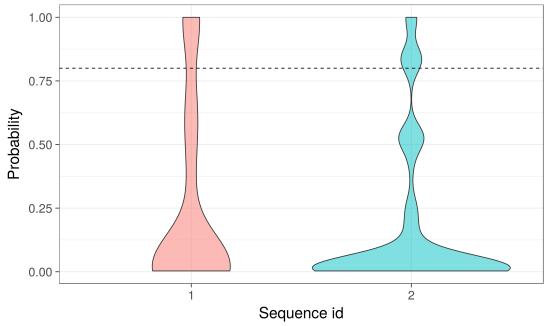


Figure 4: Violin plot showing the base pair probabilities distribution. In sequence 1, we observe a linear gradient of base pair probabilities. In sequence 2, the gradient is broken, and instead a central probability value, around 0.5, becomes very frequent. This represents the uniform base pair probability value newly acquired across the whole right arm.

	Hamming	Base pair
seq1	22	30
seq2	0	0

Table 1: Hamming distance and base pair distance at the structure level between each individual sequence and the RNAalifold consensus.

Exercise 6: RNAalifold

The sequence returned by the RNAalifold server when queried with the proposed alignment of seq1 and seq2 is exactly the same as seq2. Indeed, both the Hamming distance and the base pair distance between the consensus and seq2 are 0.

Consensus returned by RNAalifold

```
AUCGGUUCCAGCAGGAACUGUACUCGGGGGCUCGGGAAACCCUCCCGGGGUUUUACCCGGGUUUUUACGU
..(((((((....)))))))..(((((((..(((((((.....)))))))......)))))))........
```

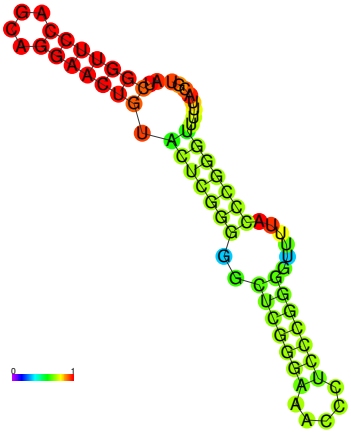


Figure 5: Consensus structure predicted by RNAalifold for the queried alignment of seq1 and seq2. The color code shows base pair probabilities.

This is probably due to the fact that the ensemble of structures associated to sequence 2 is smaller than that of sequence 1, which makes each individual structure more probable and stable. In turn, structures coming from sequence 2 become more relevant than those from 1, and so are are more relevant to the consensus.

A closer look to the changes between seq1 and seq2 (figure 6) reveals that all base pairs in the left arms are conserved. Indeed, all 3 mutations in this are are consistent changes, which consist of single mutation but that do not entail base pair loss. On the contrary, all base pairs in the right arm are unique to each structure. 16 are unique to the seq1 right arm, and 14 to the consensus. This indicates how deeply the right arm's inner organization has changed.

The fact that all remaining base pairs are unique implies that there are no more consistent/compensatory mutations. To sum up, there are three consistent mutations and no compensatory ones.

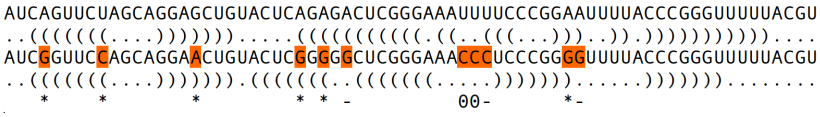


Figure 6: Seq1 and consensus (seq2) sequence and structure. Mutations are shown in red. The bottom annotation distinguishes mutations in three groups. 1) Those that entail a base pair loss (-). 2) A base pair prevails, even if it's identity is not (*). 3) Mutations in positions with no base pair in either of the structures (0).