Exercise: Measures to detect the effect of SNPs on RNA secondary structure

(mainly made by Sabarinathan Radhakrishnan)

Genome-wide association studies (GWAS) often identify that the SNPs associated to diseases or phenotypic traits are present in the non-coding regions of the human genome. It is estimated that up to 75% of the human genome is transcribed, suggesting that these SNPs in non-coding regions can be transferred to the transcriptome (Morton, 2008). In such instances, the SNPs carried in the transcript may affect the structure of a ncRNA or an RNA regulatory structure in an UTR sequence and thereby affect its function. Since many RNA secondary structure prediction programs are available via open source, the question is how we can use them to predict the effect of SNPs on ncRNA transcripts?

The effect of SNPs on RNA secondary structure can be predicted by comparing the structures of wild-type and mutant (with SNP) RNA. The structures being compared may be either optimal (MFE) structure or ensemble structure (obtained from partition function).

While comparing the optimal structure between wild-type and mutant, the structures can be considered as two distinct strings and the difference at the pure string level be used to measure how divergent they are. This strategy is employed by the `RNAmute` program (Churkin and Barash, 2006):

1. Hamming distance - The number of position at which the corresponding symbols are different.
2. base-pair distance - The total number of base pairs that are different between two structures

This can be examplified by:

```
WT    GCGGGCCCCGC
      ((((...))))
MUT   ACGGGCCCCGC
      .(((...))).
```

for which the Hamming distance is 2 and the base-pair distance is 1.

1. Predict structural effect on MFE structure. Consider the following two cases

   Case 1:

   ```
   WT    CAAUCCCGGCUGCGUCCCAGUUGGAUUUAUCCAGCUGGUUCGUGCUGGUU
         .....(((((.(((..(((((((((....)))))))))..))))))))..
   MUT   CAAUCCCGGCUGCGUCCCAGUUGGAUUUAUCCAGCUGGUUCGUGGUGGUU
         ......(.((((((..(((((((((....)))))))))..)))))).)..
   ```

   and

   Case 2:

   ```
   WT    AGCGGGGGAGACAUAUAUCACAGCCUGUCUCGUGCCCGACCCCGCUGGUU
         (((((((((((((............))))))........)))))))....
   MUT   AGCGGGGGAGAGAUAUAUCACAGCCUGUCUCGUGCCCGACCCCGCUGGUU
         (((((((..((((((..........))))))........)))))))....
   ```

   (i) Compute both the Hamming and base pair distance. (ii) Compare the two results and argue which one you find most suitable for the comparison.

2. Predict structural effect on ensemble structures.

   Goto to the RNAsnp webserver available at `http://rth.dk/resources/rnasnp/` and copy/paste the corresponding sequence / SNP information listed below into the input page.

   Input sequence:
   ```
   GCCUGUAUCCUAGGCUACACACUGAGGACUCUGUUCCUCCCCUUUCCGCCUAGGGGAAAGUCCCCGGACCU
   CGGGCAGAGAGUGCCACGUGCAUACGCACGUAGACAUUCCCCGCUUCCCACUCCAAAGUCCGCCAAGAAGC
   GUAUCCCGCUGAGCGGCGUGGCGCGGGGGCGUCAUCCGUCAGCUCCCUCUAGUUACGCAGGCAGUGCGUGU
   CCGCGCACCAACCACACGGGGCUCAUUCUCAGCGCGGCU
   ```

   SNP detail:
   ```
   A201G
   ```

   (i) Comment whether the SNP effect is local or global on the ensemble structures?

   Download all results files (available in zip format) from the RNAsnp output page.

   (ii) From the txt files inside subdirectory "*StructureDetails/secStr/*.txt" you can extract the dot bracket structure assignment of the two respective MFE structures (WT and MT). Compute the Hamming distance between them and discuss how amount and locations contributing to the Hamming distance relates to the difference and to what was obtained from RNAsnp in terms of considering base pair probability differences.

(iii) Using the base pair probability information of the affected region (available as txt file[1] in the subdirectory "*StructureDetails/bpp/*dpp.txt") find the base pair in the wild type sequence with the highest probability and compare it to the probability of the same base pair position in the mutant version.

(iv) In the local region with altered RNA structure find the number of base pairs with pair probabilities higher then 0.5 and 0.8. Do you see a consistent pattern? Discuss what information you obtain by this in contrast to considering the Hamming distance only?

[1] Similar to dot plot, the txt file contains the values in the matrix format where the upper triangle contains the information about wild-type and the lower triangle about mutant.