Comp 598: Assignment 1

RNA bioinformatics (Incomplete assignment)

Due on March 9th, 2014.

- To some extent, collaborations are allowed, but you must indicate the name of all collaborators (including instructors) on your answers. Uncredited collaborations will be penalized.
- Unless specified, all answers must be justified.
- Partial answers will receive credits.

Exercise 1 (10 points) Dangling ends are (unpaired) nucleotides that stack on the ends of helices. In secondary structures, they occur in multi-branch and exterior loops. In class, we have seen that we can enumerate the complete RNA secondary structures conformational landscape using the recursive decompositions detailed in Figure 1.

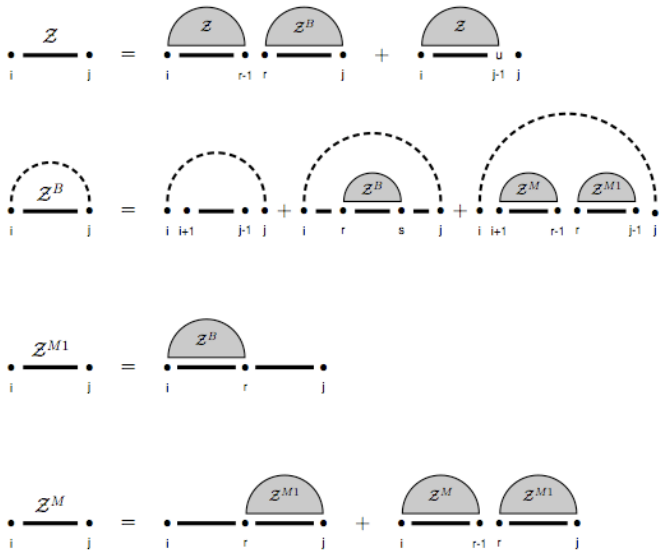

Figure 1: Feynman diagrams of the recursive decomposition used to enumerate all RNA secondary structures without pseudo-knots.

Unfortunately, these specific recursions do not allow to incorporate the energetic contribution of dangling ends in RNA secondary structures (I.e. There are ambiguities in the decomposition that prevents us to decide if a dangling end occurs or not). Explain how it can happen and point at the cases that are concerned in Figure 1. Suggest a modification to overcome this limitation.

Exercise 2 (10 points) We want to compute consensus RNA secondary structure of a given RNA sequence from a set of samples generated with the stochastic backtracking procedure. A base pair belongs to the consensus structure if it occurs with a frequency higher than 0.5 in the sample set. The following example illustrates this method.

```
sample 1 : ((.((.....))))
sample 2 : ((..((...)).))
sample 3 : (((.(.....))))
-------------------------
Consensus: ((..(.....).))
```

The sampled structures do not contain pseudo-knots. Prove that this is also necessarily the case for the consensus secondary structure.

Exercise 3 (20 points) The algorithm of D. Sankoff (1985) performs a simultaneous alignment and folding of 2 RNA sequences with unknown structures. Now, lets assume that we know the secondary structures $\mathcal{S}_1$ and $\mathcal{S}_2$ (without pseudo-knots) of 2 RNA sequences ω_1 and ω_2.

Propose an algorithm that aligns ω_1 and ω_2 with their secondary structure ω_1 and ω_2.

Exercise 4 (20 points) In 1999, E. Rivas and S. Eddy proposed an algorithm that extends the recursions from Zuker & Stiegler (1981) to predict RNA secondary structures with pseudo-knots. Four new dynamic programming tables (named here vhx, zhx, yhx and whx) were introduced. The Feynman diagrams of these tables are shown in Figure 2.

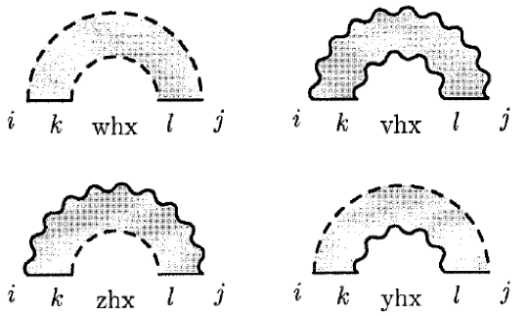

Figure 2: Feynman diagrams of the extended dynamic programming tables for pseudo-knot prediction.

vhx(i, j; k, l)	: (i,j) paired, (k,l) paired
zhx(i, j; k, l)	: (i,j) paired, (k,l) unknown
yhx(i, j; k, l)	: (i,j) unknown, (k,l) paired
whx(i, j; k, l)	: (i,j) unknown, (k,l) unknown

The recursive decompositions used to compute values of the tables vhx and zhx are detailed in Figure 3.

(a) vhx (b) zhx

Figure 3: Recursive decompositions of arrays vhx and zhx.

Using Feynman diagrams, write the decomposition for the tables yhx and whx. Describe each case with one sentence.

Exercise 5 (10 points) Let ω_1 =GGGCCGCCGACCCGGCCC and ω_2 =CCCGGCGGUCGGGCCGGG be two RNA sequences. We want to decide if and how these two molecules can interact. To this end, we propose to use the `RNAcofold` and `RNAup` programs of the Vienna RNA package.

Run both programs and describe their predictions. Why are the predictions different? Which one would you trust more? Why? Explain the limitations and advantages of both programs.

N.B.: You can run Vienna RNA software suite on SOCS servers or using the Vienna RNA web server accessible at `http://rna.tbi.univie.ac.at/`

Exercise 6 (30 points) We want to study evolution of RNA using an evolutionary reactor. You will start by implementing the reactor using `RNAfold` and `RNAdistance` from the Vienna RNA package (these program are installed on SOCS servers but you can also install the package available at `http://www.tbi.univie.ac.at/RNA/`.

1. Fix a target secondary structure T of size L.
2. Generate a starting population of size N random RNA sequences of length L sampled uniformly.
3. Calculate the MFE structure S_i of each sequence ω_i in the current population using `RNAfold`.
4. Estimate the fitness d of the MFE structures S_i with the target secondary structure T using the base pair distance implemented in `RNAdistance`.
5. Determine the reproduction rate R of sequence ω_i as $R(\omega_i) = \frac{e^{-\beta d(S_i,T)}}{Z}$, where $d(S_i, T)$ is the base pair distance between S_i and T, β denotes the selection pressure (here we will take $\beta = \frac{2}{L}$), and Z_i the Boltzmann partition function defined as $Z = \sum_i e^{-\beta\beta d(S_i,T)}$.
6. Replicate sequences i from the current population with probability $R(i)$ and an error rate $\mu = 0.02$ (i.e. 2% of mutations per nucleotide). Replace the old population with the new one. Keep the size of the population fixed.
7. Iterate from 3.

Simulate the evolution of RNA populations of size $N = 100$ over 500 generations with mutations rates $\mu = 0.01, 0.02, 0.05, 0.1$ and the following target structures:

T_1 `(((((((((....)))))))))`
T_2 `(((((..(((....)))))))))`
T_3 `(((....)))(((....)))`

For each target structure, plot a graph showing the average distance of the population to the target structure $\bar{d} = \frac{\sum_i d(S_i,T)}{N}$ vs. the generation. Each graph will feature 4 curve corresponding for the 4 proposed mutation rates μ.