Comp 598: Assignment 1

RNA bioinformatics

Due on October 31st, 2016.

- 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.
- You can run the programs from the Vienna RNA package from SOCS or install the package available at `http://www.tbi.univie.ac.at/RNA/`. You should use the latest release (2.2.10) of the software suite.

Exercise 1 (10 points) We propose the following grammar $G = \langle \Sigma, V_N, P, S \rangle$ to describe RNA secondary structures.

$$\Sigma = \{\ (\ ,\ )\ ,\ \bullet\ \}$$

$$V_N = \{\ S\ ,\ T\ \}$$

$$P = \left\{ \begin{array}{l} S \to (\ T\ )\ S \mid \bullet\ S \mid \epsilon \\ T \to (\ T\ )\ S \mid \bullet\ T \mid \bullet \end{array} \right\}$$

Where Σ is the set of terminals, V_N the set of non-terminals, P is the set of production (re-writing) rules, and S is the axiom. Draw a derivation tree representing a sequence of re-writing rules that generates the following word (i.e. secondary structure).

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

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 (i.e. crossing base pairs). Prove that this is also necessarily the case for the consensus secondary structure.

Exercise 3 (20 points) Rational sampling of RNA secondary structure enables us to compute a picture of the structures present at the equilibrium (i.e. a sample set of secondary structures). From this set, we can estimate frequencies and distributions of complex structural features such as base pairing probabilities.

The Vienna RNA package has a program named `RNAsubopt` that draws secondary structures with probabilities equal to their Boltzmann weights in the low energy ensemble. We will use `RNAsubopt` to generate $k = \{100, 1000, 10000\}$ secondary structures for the sequence available at:
`http://cs.mcgill.ca/~jeromew/docs/comp598/HW1Q3.fasta`
For this question, you will have to use the Python template available at:
`http://cs.mcgill.ca/~jeromew/docs/comp598/HW1Q3.txt`
In particular, you will have to find all occurrences of the "@TO_STUDENT" marks that indicate parts that are you are required to complete. You should also replace the suffix ".txt" from the filename by ".py" (i.e. rename `HW1Q3.txt` as `HW1Q3.py`).

1. Implement `get_answer_Q3_1` method in template file. This program will estimate contact probabilities from a sample set of secondary structures. It should return a list of lists (as shown in template) with $[i,\ j,\ freq(i,j)]$, where (i,j) is a base pair such that $i < j$, and $freq(i,j)$ is the frequency of the base pair (i,j) in this set.

2. The program `RNAfold` can compute directly base pair probabilities, and store them by default in a postscript file named "dot.ps". Implement `get_answer_Q3_2` of the template file mentioned above. It will compare the base pairing frequencies estimated from the sample set with your implementation `get_answer_Q3_1`, with the values calculated by `RNAfold`, and return the value:

$$error = \sqrt{\sum_{\substack{(i,j)\\ i<j}} \left(P^{RNAfold}(i,j) - freq(i,j)\right)^2}$$

3. apply your programs 10 times for each value of $k = \{100, 1000, 10000\}$, and compute the mean and standard deviation of the score $error$. Estimate a value k such that $error < 0.001$.

Exercise 4 (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 $\mathcal{S}_1$ and $\mathcal{S}_2$.

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.

Exercise 6 (30 points) We want to study evolution of RNA using an evolutionary reactor. You will start by implementing the reactor using `RNAfold` (secondary structure prediction) and `RNAdistance` (comparison of structures) from the Vienna RNA package.

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 $Z = \sum_i e^{-\beta d(S_i,T)}$.

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.