COMP564 Assignment 1

Anne-Sophie Fratzscher
ID: 260705446
March 9th, 2021

1.

NOTES: L = length of sequence,
n = number of sequences,
$\lambda = 1$,
'bp' is shorthand for base pair,
Let there be m structures: $s_1, s_2, \ldots s_m$

Nussinov-Jacobson variant

```
$\mathcal{L}(s_i)$ holds the highest energy among all sequences $\omega_i$ for structure $s_i$
For each structure $s_i$ do
    // initialize nxn matrix called M
    $M_{i,j} = -\infty$ for all i, j // because want to maximize and energies are negative
    For L = $\lambda$ to N do
            // excludes bp that don't meet min number of nucleotides in hairpin
        For i = 0 to N-L do
            j = i+L
            maxval = M(i, j-1) // means no bp at j
            for k = i to j- $\lambda$ do
                for each sequence ω do
                    maxval = max(maxval, M(i, k-1) + M(k+1, j-1) + δ(ω[k], ω [j])
                    // maxval = least negative energy (to get highest energy of sequence
                        folding into structure)
            M(i,j) = maxval
    $\mathcal{L}$ ($s_i$) = M(0, L) // when done, set $\mathcal{L}$ ($s_i$) to highest energy for full sequence (i. e. from 0 to L)

// pick structure with minimum L(s)
return structure $s_i$ with min $\mathcal{L}$(s)
```

Scoring function

$$\delta(x,y) = \begin{cases} -\infty \; if\ not\ bp\ in\ structure \\ -4 \; if\ G - C\ or\ C - G \\ -2 \; if\ A - U\ or\ U - A \\ -1 \; if\ U - G\ or\ G - U \\ \infty \; if\ not\ valid\ bp \end{cases}$$

// $-\infty$ if no base pairing because shouldn't penalize for not being at base pair yet (e. g. if the structure = ((.)) but are at index k = 2, j = 3, then have .), which is not a bp)

// ∞ if not a valid base pair (e. g. x = A and y = C, then should not base pair) -> is to remove structures that are incompatible with a sequence because this will always be the biggest value, so when do min(L(s)), this structure will never be selected

2.

Let us assume that the consensus sequence has a pseudoknot. Additionally, we are told that the sampled structures do not contain pseudoknots. First of all, we cannot use the typical bracket notation, as it assumes there are no crossing interactions (i. e. if there were crossing interactions, it would be impossible to tell where they are using the traditional bracket notation using '(', ')', and '.'). To show crossing interactions, the professor suggested using a different type of parenthesis (e.g. {}).

For this question, assume that the structures have some notation to account for crossing interactions. From the definition, the frequency of each base pair in the consensus sequence occurs with a frequency higher than 0.5 in the sample set. This means that the base pair (or base pairs, if multiple) involved in the pseudoknot must occur with at least 50% frequency in the sample set. This means that at least 50% of the sample structures have a crossing interaction. However, by definition, none of the sample structures contain pseudoknots. This is a contradiction. Therefore, if the sampled structures do not contain pseudo-knots, the consensus secondary structure also will not contain pseudoknots. ■

3.

The code for this program can be found in HW1Q3.py (parts 1 and 2) and HW1Q3_3.py (part 3). The following commands are executed in the files to get the suboptimal secondary structures and the dot.ps file for the HW1Q3.fasta sequence:

To get k subopt secondary structures sequence in HW1Q3.fasta:

"RNAsubopt -p 100 < HW1Q3.fasta > subopt100.txt" for k = 100

To get the dot.ps file:

"RNAfold -p --MEA HW1Q3.fasta >/dev/null 2>&1"

Note that the >/dev/null 2>&1 prevents printing of the results of RNAfold.

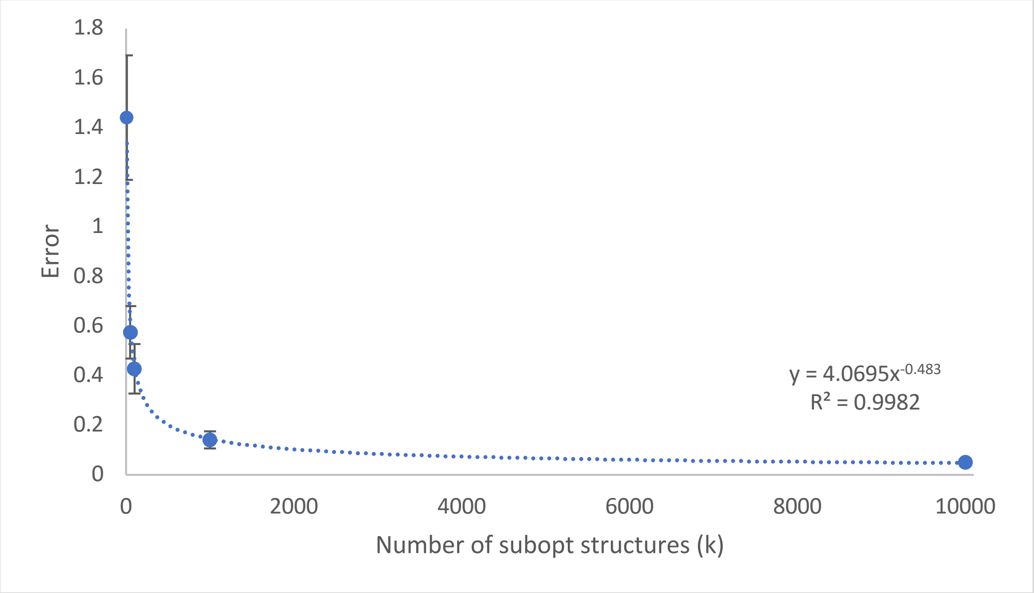
For part 2, the following errors were found:

```
Error for k = 10: 1.1590391749081743
Error for k = 50: 0.4735709153962975
Error for k = 100: 0.4299029249897637
Error for k = 1000: 0.10341308768323006
Error for k = 10000: 0.047455002040801476
```

Part 3 was run by executing the HW1Q3_3.py file. The results for part 3 are shown below:

Run	k=10	k=50	k=100	k=1000	k=10000
1	1.3240286851628436	0.6858810584805938	0.3117008656094568	0.16457964992946542	0.057665273610095
2	1.7399197927163021	0.673006760423676	0.3390055991203789	0.12644938890876437	0.04350390252071703
3	1.328988451084026	0.6086799420090924	0.5145420462673237	0.1395650252039824	0.04994893374234512
4	1.1185753809978247	0.4325430896077708	0.4962829395276018	0.17313727937992462	0.032277235290871426
5	1.636573794346515	0.6270858505911641	0.5226116976906566	0.10950340567378246	0.03832464604224701
6	1.5401642106775413	0.6649656840008141	0.5537619434501337	0.09454878186775262	0.0656804125120934
7	1.0853569952708784	0.5581473838112836	0.3446850011549101	0.170202200785169	0.05136104082749465
8	1.2470006764305166	0.4110945154232797	0.3590514167191367	0.20157880575916656	0.037766943248802615
9	1.7530619979517572	0.45009665649931013	0.319783145905404	0.11632921406750059	0.04261133247191879
10	1.6368243903855368	0.6373035966912646	0.5167766477102023	0.11534072147751588	0.07334802884531706
mean	1.4410494375023741	0.5748804537538249	0.42782013031552046	0.14112344730530238	0.04924877491119021
stdev	0.25110195916509803	0.10581244579826586	0.09980657445367618	0.03454753451728491	0.01309293067243131

Plotting the error for the different k values gives the graph shown below. The data fits a power curve with the equation $y = 4.0695x^{-0.483}$ with a R^2 value of 0.9982. Using this equation, we find that y = 0.01 when x = 252800. Therefore, k > 252800 gives an error < 0.01.



4.

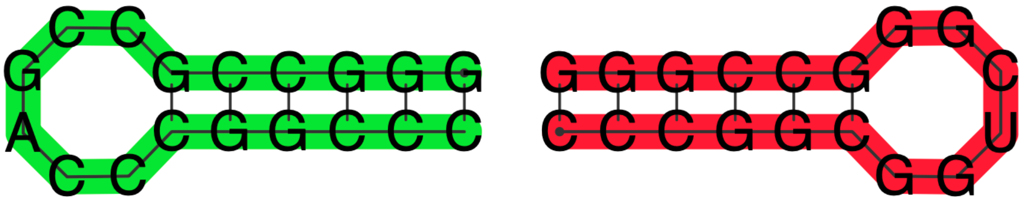
The two sequences were put into the file HW1Q4.fasta as follows:

>HW1Q4
GGGCCGCCGACCCGGCCC&CCCGGCGGUCGGGCCGGG

RNAcofold was run using the following command: “RNAcofold < HW1Q4.fasta”. The following was given as the output:

```
        ((((((......))))))&((((((......)))))) (-23.70)
```

As well as the following in the HW1Q4_ss.ps file:



RNAup was run using the following command: “RNAup -b < HW1Q4.fasta”. The following was produced as output:

```
((((((&))))))   7,12  :   7,12  (-12.00 = -12.00 + 0.00 + 0.00)
CCGACC&GGUCGG
```

Additionally, the file RNA_w25_u1.out was produced:

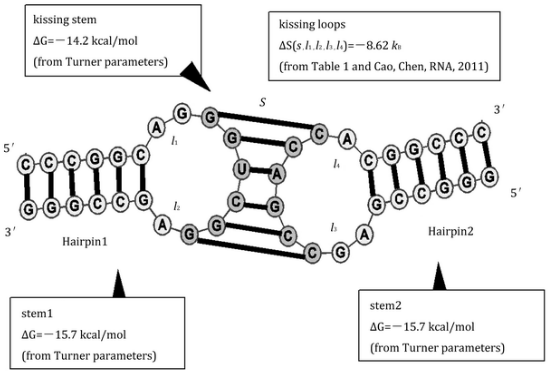
```
# RNAup --include_both
# 18
# GGGCCGCCGACCCGGCCC
# 18
# CCCGGCGGUCGGGCCGGG
#    pos     u4S      dG
      1       NA    0.000
      2       NA   -0.134
      3       NA   -0.134
      4    6.167   -0.134
      5    6.166   -0.134
      6    8.799   -1.614
      7    7.611  -11.998
      8    6.355  -11.998
      9    1.739  -11.998
     10    0.000  -11.998
     11    0.000  -11.998
     12    0.001  -11.998
     13    1.739   -7.290
     14    6.514   -2.117
     15    9.173   -2.117
     16    9.143   -2.117
     17    7.141   -2.117
     18    7.141    0.000
#    pos     u4S
      1       NA
      2       NA
      3       NA
      4    8.920
      5    8.931
      6   10.880
      7   10.045
      8    7.422
      9    4.792
     10    0.000
     11    0.000
     12    0.001
     13    4.799
     14    8.687
     15   11.014
     16    9.376
     17    7.176
     18    7.175
```

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

RNAcofold concatenates two molecules and uses modified energies for the loop containing the cut to compute a MFE secondary structure. RNAup, on the other hand, predicts interaction by computing the free energy needed to open a potential binding site and combining it with the interaction energy to get the total binding energy. These different calculations for energy lead to different predictions.[1]

Both algorithms have limitations and advantages. RNAcofold can generate many binding sites, whereas RNAup can only predict a single interaction site. However, RNAup predictions can have a complex pseudoknotted configuration, whereas RNAcofold predictions do not allow pseudoknotted configurations. [2] Additionally, RNAcofold also disregards kissing hairpin complexes. [3] Generally, RNAup is used more for predicting binding of regulatory RNA to target RNAs (such as binding of miRNA to mRNA), whereas RNAcofold is used to compute concentrations of monomer and dimer species for different concentrations of monomers (since binding is concentration dependent). [4]

For this example (ω_1 = GGGCCGCCGACCCGGCCC and ω_2 = CCCGGCGGUCGGGCCGGG), RNAcofold predicted a structure with energy = -23.70 kcal/mol. RNAup predicted an interaction between positions 7-12 of ω_1 and positions 7-12 on ω_2,with a total free energy of -12 kcal/mol (-12 kcal/mol from duplex formation, 0 kcal/mol for opening energy for both sequences). I believe the results from RNAup are to be trusted more. The opening energy of 0 kcal/mol for both implies that both structures are hairpins (as we saw in the output file from RNAcofold) that are not opened to create a binding site, and the interaction between positions 7-12 of ω_1 and positions 7-12 on ω_2 suggests a kissing stem-loop structure (example shown below). As mentioned earlier, RNAcofold cannot predict pseudoknotted or kissing hairpin complexes, explaining the difference in prediction.



Link for image: https://link.springer.com/article/10.1007/s41048-015-0001-4

[1] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3319429/

[2] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3319429/

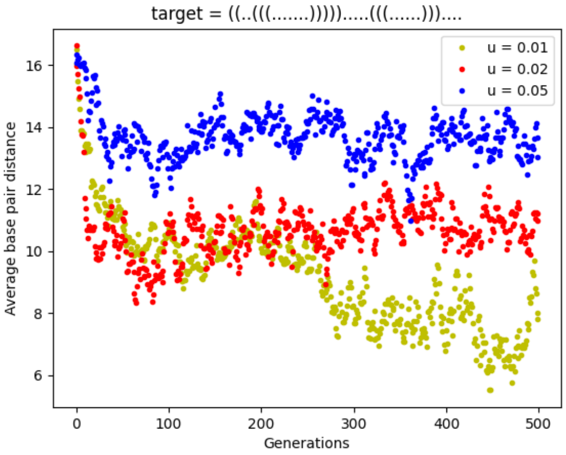
[3] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1459172/

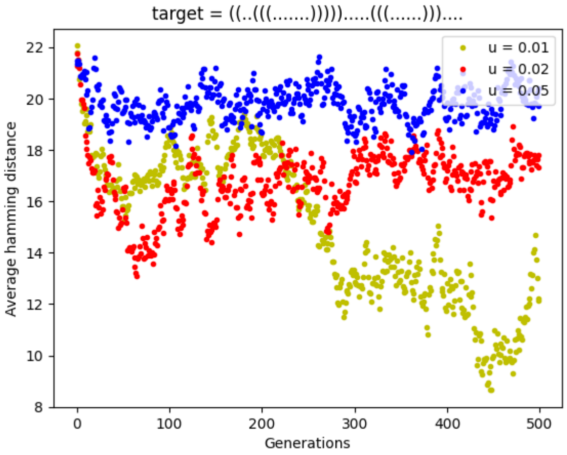
[4] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1459172/

5.

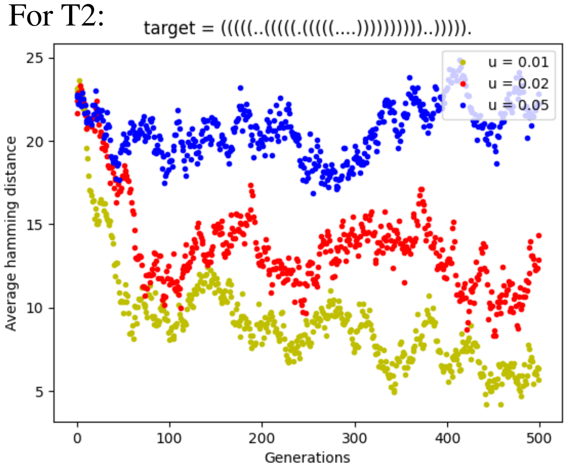
The code for this question can be found in HW1Q5.py. The code was based off of the code by Carlos Oliver that can be found here: https://github.com/cgoliver/RNA-Popgen-Notebook/blob/master/Population_Genetics.ipynb.

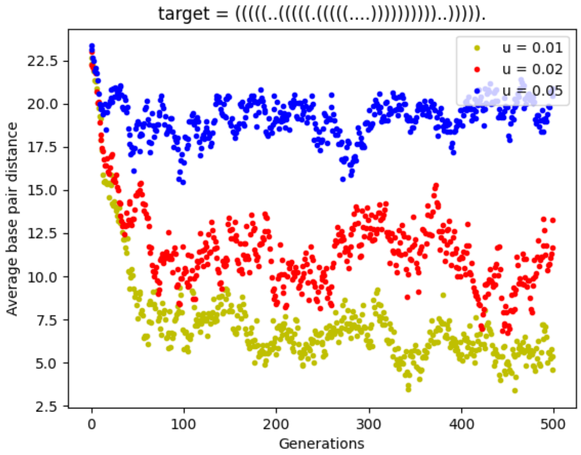
For T1:



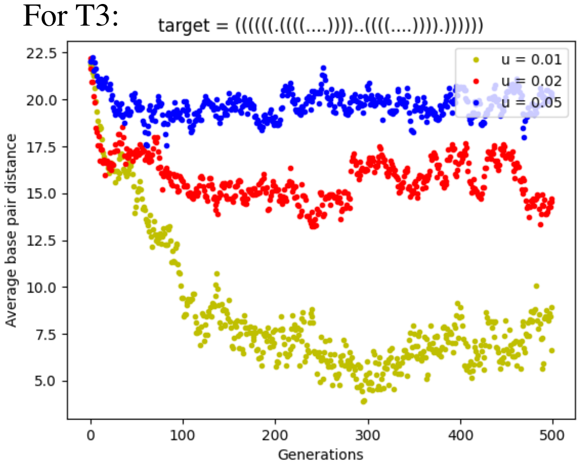


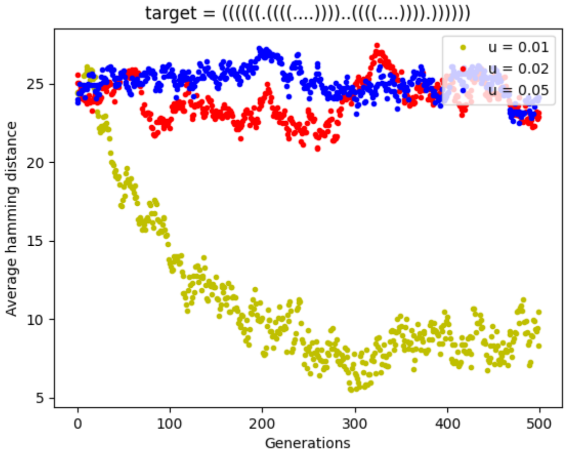
For T2:





For T3:





As a reminder, the base pair distance was calculated using RNAdistance on the target structure and the MFE structure predicted by RNAfold. Base pair distance is the number of base pairs that have to be removed or added to get from the target to the MFE structure. The Hamming distance is the number of positions at which the symbols are different for the secondary structures. In class, we were told that the base pair distance is the standard for evaluating structure similarities.

For all three target structures, we see that the start distance (both hamming and base pair) is the same for all mutation rates, but that, over the course of multiple generations, lower mutation rates have lower average distance (rate of 0.05 has the highest average base pair distance, 0.02 has lower distance, and 0.01 has lowest distance). For all the mutation rates, we can see that the distance fluctuates (i. e. is not constant or strictly decreasing every generation), which is due to mutations being either beneficial (lower distance because is closer to the target) or deleterious (e. g. added mutation that led to worse distance). Hamming and base pair distance should be similar.

The graphs show the "survival of the flattest" (meaning that, although the distance is higher/ fitness is worse, these sequences are less sensitive to deleterious mutations, so there is less change in distance) for the higher mutation rate (0.05), as the distance lowers to a certain value and then stays around there. The lower mutation rate (0.01) seems to be dictated by fitness instead, as we had learned in class, with the average distance continuing to decrease for up to 500 generations. It should be noted that the initial population was randomly generated, which may explain why mutation rates of 0.01 and 0.02 for T1 yield similar distances up until generation 250. This is because the initial population may have consisted of sequences that were not sensitive to mutation (e. g. if the base pair was A-U, mutating A to G would lead to the same secondary structure, so the base pair distance and hamming distance would not be affected).