ChE/BE 163 Fall 2020
Problem Set #1
Due 1pm Thursday, October 22

Problem 1 (Equilibrium analysis of a complex of interacting nucleic acid strands, 15 pts)**.** Consider the hybridization chain reaction (HCR) mechanism of Figure 1 (*Proc Natl Acad Sci USA* **101**, 15275–15278, 2004), in which metastable DNA hairpins H1 and H2 polymerize upon exposure to initiator sequence I1. Note that each hairpin is intended to have a 6-nt toehold, an 18-bp stem, and a 6-nt loop.

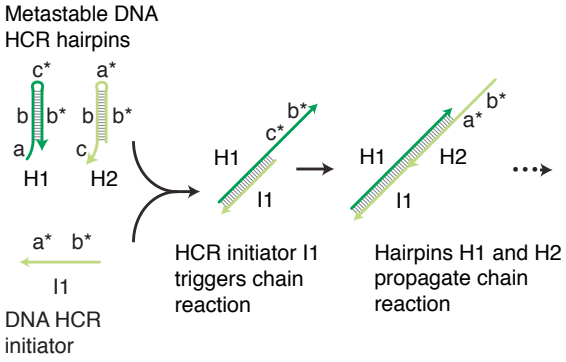

Figure 1: Conditional self-assembly via hybridization chain reaction (HCR).

Consider two different sequence designs for HCR that are intended to detect initiator:

I1: `CCACACCACAACAACCACATCTCG`

HCR Design 1

H1: `CGAGATGTGGTTGTTGTGGTGTGGATACAACCACACCACAACAACCAC`

H2: `CCACACCACAACAACCACATCTCGGTGGTTGTTGTGGTGTGGTTGTAT`

HCR Design 2

H1: `CGAGATGTGGTTGTTGTGGTGTGGGTGGTTCCACACCACAACAACCAC`

H2: `CCACACCACAACAACCACATCTCGGTGGTTGTTGTGGTGTGGAACCAC`

Use the Utilities page of the NUPACK web application (nupack.org) to analyze these two different designs at 23 °C by examining the equilibrium base-pairing properties of the reactant and product complexes (each complex comprising one or more interacting strands) for two elementary steps in the HCR reaction pathway:

Step 1: I1 + H1 → I1·H1

Step 2: I1·H1 + H2 → I1·H1·H2

Use MFE structures and equilibrium base-pairing probabilities to explain which design you prefer.

Problem 2 (Ensemble size via dynamic programming, 20 pts)**.**
In this problem, you will adapt the dynamic programming algorithm for computing the partition function of a single strand over the ensemble of unpseudoknotted secondary structures, Γ, to the simpler problem of computing the number of possible unpseudoknotted secondary structures, $|\Gamma|$. The primary reference for this problem is Dirks and Pierce, *J. Comp. Chem.*, **24**, 1664 (2003), with an emphasis on the $O(N^4)$ Algorithm (equations [7] and [8], Figures 3, 4 and 5, and the pseudocode in Figure 6).

a) (3 pts) The partition function algorithm enables you to efficiently calculate

$$Q(\phi) = \sum_{s \in \Gamma} e^{-\Delta G(\phi,s)/k_B T}$$

for a strand with sequence ϕ. Here, we wish to calculate the size of the ensemble

$$|\Gamma(\phi)| = \sum_{s \in \Gamma(\phi)} 1 \tag{1}$$

where each structure $s \in \Gamma(\phi)$ contains only Watson-Crick pairs (i.e., no wobble pairs) and we note that two bases cannot pair if $j - i < 4$ (due to steric constraints). How can you make a simple change to the energy $\Delta G(\phi, s)$ to use the partition function algorithm to calculate $|\Gamma(\phi)|$?

b) (10 pts) Consider the DNA sequence `CGAGATACCTCGATCACGCG`. Write a Python script to execute the dynamic program and calculate $|\Gamma(\phi)|$. Include your script and show the final state of the matrices Q, Q^b and Q^m. Hint: *What is the value of $Q^b_{i,j}$ if i and j cannot base-pair (either because they are not Watson-Crick complements or because they are sterically prevented from pairing)?* Hint: *Do not use NUPACK's* `ensemble_size` *command to check your script. It will give you different answers because it includes wobble pairs. You can test your program with the sequence* `ATAGTTTTCTCGAAAACGAT`, *which has 2349 unpseudoknotted secondary structures.*

c) (7 pts) Generate 5 random sequences for each length $N \in \{10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}$ nt (select the sequence of each nucleotide from a uniform distribution over {`A,C,G,T`}). Use your script to calculate $|\Gamma(\phi)|$ for each sequence; display $|\Gamma(\phi)|$ vs N as a scatter plot to observe the blowup in ensemble size as N increases. Suppose $|\Gamma(\phi)| = ae^{bN}$ so that taking log of both sides yields $\log|\Gamma(\phi)| = \log a + bN$. Display $\log|\Gamma(\phi)|$ vs N as a scatter plot and use a built-in least squares fitting procedure (e.g., `polyfit` in NumPy) to estimate a and b; display the line of best fit in your plot.

Problem 3 (Objective function and nucleic acid sequence design, 50 pts)**.**
In this problem you will design sequences intended to form an "RNA stick figure" at equilibrium (Figure 2).

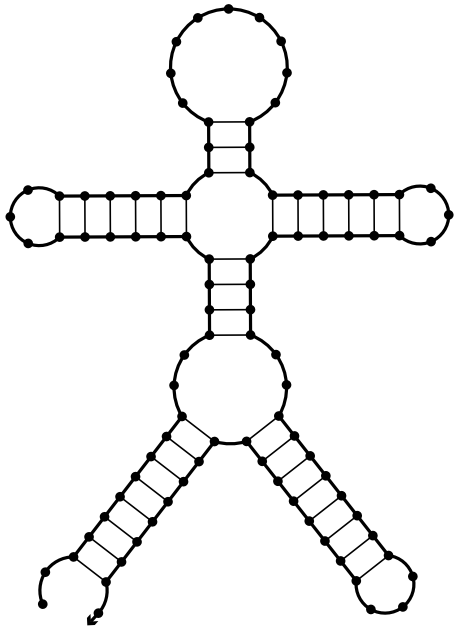

Figure 2: RNA stick figure.

We specify the secondary structure using *dot-parens* notation. Each unpaired base is represented by a dot and each base pair by matching parentheses. If a given base has a dot, it is unpaired. If it has an open parenthesis, it is paired with the parenthesis that closes it (e.g., "`.(((...)))`" represents a secondary structure where the second base is paired with the last, the third with the ninth, and the fourth with the eighth). The stick figure is denoted below:

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

For unpseudoknotted structures, the dot-parens representation of the secondary structure has the same information as the drawing above, the corresponding polymer graph (not shown), and the secondary structure matrix S. The dot-parens representation is a convenient way to represent the structure in a computer program because it is a simple string, and you will use it in your script.

In doing the design calculations, you will make use of Utilities commands in the NUPACK Python module (but not the Design commands). The NUPACK User Guide provides a useful online reference describing the NUPACK python module. Additionally, you will need to use other Python-based software, such as NumPy. You can use a computing resource on the NUPACK server cluster that we have set up for you that already has NUPACK 4.0 installed, along with Python wrappers to aid in its use, and other Python packages you will need. For instructions on how to launch and use a NUPACK server, see the Computing section of the Course Info page.

Unless specifically instructed to use the NUPACK web application, you will write your own Python script that uses commands from the NUPACK Python module (again, with the obvious exception of

the design commands). You are encouraged to work with your classmates, but your script must be your own. Please include both your script and output when submitting the assignment.

We will formulate sequence design as an optimization problem with the goal of reducing an objective function below a user-defined stop condition. For a given target secondary structure, s, with N nucleotides and sequence ϕ, consider two objective functions for sequence design:

- Minimum free energy (MFE) defect

$$\mu(\phi, s) = N - \sum_{\substack{1 \le i \le N \\ 1 \le j \le N+1}} S_{ij}\left(s^{\mathrm{MFE}}(\phi)\right) S_{ij}(s),$$

  with stop condition $\mu(\phi, s) \le N/100$.

- Ensemble defect

$$n(\phi, s) = N - \sum_{\substack{1 \le i \le N \\ 1 \le j \le N+1}} P_{ij}(\phi) S_{ij}(s)$$

  with stop condition $n(\phi, s) \le N/100$.

To optimize either objective function, start from a random initial sequence that satisfies the base-pairing requirements of the target structure using Watson-Crick pairs. Attempt random mutations of either an unpaired base (one base at a time) or a base pair (two bases at a time), accepting a mutation if it does not increase the objective function. Terminate the optimization if the stop condition is satisfied or if you reach a maximum (large) number of mutation attempts without a mutation that lowers the objective function.

For parts (a)-(e), s is the stick figure and $T = 23\ ^\circ\mathrm{C}$ (use the `rna1995` parameters and `ensemble='stacking'` for your physical model). Time your design calculations.

a) (5 pts) Generate five random sequences that satisfy the base-pairing requirements of the stick figure using Watson-Crick base pairs.

b) (5 pts) Perform MFE defect optimization to design five sequences that satisfy the MFE defect stop condition.

c) (5 pts) Perform ensemble defect optimization to design five sequences that satisfy the ensemble defect stop condition.

d) (15 pts) Calculate $\Delta G(\phi, s)$, $p(\phi, s)$, $\mu(\phi, s)$, and $n(\phi, s)$ and GC content for all of the sequences from parts (a), (b) and (c). Comment on the effectiveness and speed of the different design strategies, as well as on implications for the importance of positive and negative design.

e) (10 pts) Analyze one of your best designs using the Analysis page on the NUPACK website. Show the pair probability matrix at 23 °C and comment on it. Check for pseudoknots and dimers (e.g., at a concentration of 1 µM). Compute melt curves with the temperature ranging from 20°C to 95°C in 2.5°C increments. Comment on the shape of the melt curve and how it relates to base pairing.

f) (10 pts) Define an objective function to design a sequence that adopts two different target structures,

$$s_1 = \texttt{((((((((((.....)))))))))).....}$$

$$s_2 = \texttt{......((((((((((((........))))))))))))},$$

each with probability 0.5 at equilibrium. Modify your design script to use this objective function and optimize a sequence to satisfy a user-specified stop condition for these two structures. Use the Utilities page to quantify the degree to which you succeeded for each target structure.

Problem 4 (Equilibrium analysis of a test tube of interacting nucleic acid strands, 15 pts)**.**
Run the Analysis demo on the NUPACK website. Observe that there are two transitions in the melt profile. Use the ensemble pair fractions for the dilute solution, as well as the equilibrium concentrations, MFE structure and/or equilibrium pair probabilities of the various ordered complexes to figure out what is happening in the test tube to produce the two melting transitions. Use relevant plots and graphics to justify your claims. Don't simply include a large number of plots—carefully select the most illuminating visuals and make a coherent argument as to what is occurring in the test tube.