IE 400: Principles of Engineering Management Project

Spring 2022-2023

Due Date: 25 April 2023

In living organisms, Ribonucleic acid (RNA) is an important matter that has essential roles. RNA molecules play an active role in different processes such as encoding and decoding. Subsequently, RNA molecules fold themselves to fit inside cells and perform specific functions in order to achieve these processes. Accordingly, finding the secondary structure by using primary structure helps to determine the function of RNA. Further, RNA is said to be highly important for diagnostic and therapeutic design. Hence, it can be said that the findings of RNA studies could provide a better understanding to develop treatments for RNA-related diseases [1]. Within this perspective, RNA folding studies aim to clarify the relationship among sequence, tridimensional structure, and biological function [2]. In particular, the purpose of RNA folding problem is to predict the secondary structure of an RNA molecule while only its nucleotide sequence is known.

To understand the concept of RNA folding problem, we will denote s as a string of n characters consisting of nucleotides: Adenine (A), Guanine (G), Uracil (U) and Cytosine (C). For instance,

$$s = ACGUCCAUGCAG.$$

Moreover, it is averted that the stability of the RNA is measured by the number of bonds. Consequently, the most stable structure is the one with

[1] January 15, 2021 — B. A. M. (2023, February 22). New videos show RNA as it's never been seen. *Northwestern Now.*

[2] Shaw, E., St-Pierre, P., McCluskey, K., Lafontaine, D. A., & Penedo, J. C. (2014). Using SM-fret and denaturants to reveal folding landscapes. *Methods in Enzymology*, 313–341.

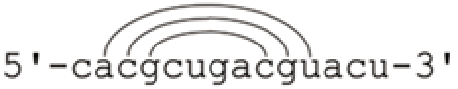
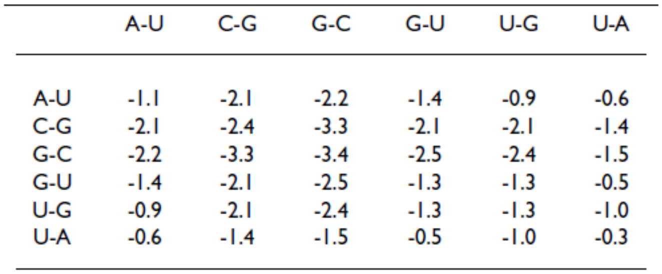


Figure 1: Arc representation of pairings

Table 1: Energy levels of stacking pairs



	A-U	C-G	G-C	G-U	U-G	U-A
A-U	-1.1	-2.1	-2.2	-1.4	-0.9	-0.6
C-G	-2.1	-2.4	-3.3	-2.1	-2.1	-1.4
G-C	-2.2	-3.3	-3.4	-2.5	-2.4	-1.5
G-U	-1.4	-2.1	-2.5	-1.3	-1.3	-0.5
U-G	-0.9	-2.1	-2.4	-1.3	-1.3	-1.0
U-A	-0.6	-1.4	-1.5	-0.5	-1.0	-0.3

the maximum binding strength; i.e., the largest number of bonds. Having all the above information, there are also some important conditions that should be taken into consideration while dealing with the RNA folding problem.

- Nucleotide A must be paired with only U in which C must be paired with only G (or vice versa).
- There are 2 hydrogen bonds between A and U, 3 hydrogen bonds between G and C. Therefore, binding strength differs from pair to pair.
- Each nucleotide must be paired with at most 1 nucleotide.
- The pairings are not allowed to cross each other. For example, let $i < i' < j < j'$, then (i, j) and (i', j') cannot be paired at the same time.

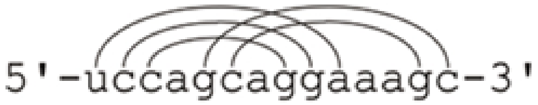


Figure 2: Arc representation of cross pairings (pseudoknots)

- There is a distance limitation that close nucleotides cannot be paired; i.e., the nucleotide cannot pair with any nucleotide that is less than 4 positions away from it on the sequence s.

We try to find an integer programming (IP) model that solves the following questions while considering the above conditions.

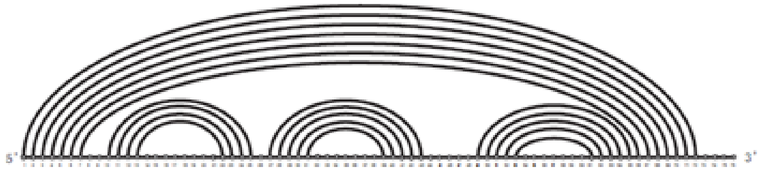
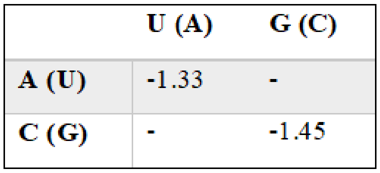


Figure 3: A line representation of a nested pairing

(a) Given the nucleotide sequence s (each group will have different sequence s; provided in **data.xlsx**) of an RNA molecule, find a nested pairing that achieves the maximum number of pairs.

Table 2: Energy level of matched pairs



	U (A)	**G (C)**
A (U)	-1.33	-
C (G)	-	-1.45

(b) Now consider that we try to find a nested pairing that gives the minimum total free energy associated with pairs in the sequence (please use Table 2 for energy levels of matched pairs in your model). Make necessary changes in the IP model and compare your results with part (a). What are the differences in terms of solution time and the number of pairs?

(c) Now consider that there must be at least 7 nucleotides (rather than 4) between pairing ones while keeping the total free energy at minimum as in part (b). What are the differences in terms of solution time and the number of pairs?

(d) A matched pair (i, j) in a nested pairing is defined as a *stacked pair* if either $(i + 1, j - 1)$ or $(i - 1, j + 1)$ is also a matched pair in the nested pairing. In other words, there is a stacked pair **if and only if** both (i, j) and $(i + 1, j - 1)$ are in the nested pairing. It is also

known that a stack with k matched pairs adds more to stability than k individual matched pairs, which do not belong to any stacks. The stack pair correspond to an energy level, which is shown in Table 1. For instance, if the structure of the sequence s is $s = ACGAAAACGU$, then the stacked pairs' energy level will be (-2.1) +(− 2.2) = - 4.3, since we have A-U, C-G, and G-C as nested pairs. That is to say, there has to be at least 2 stacked pairs to lead the energy level. If there is one pair, it is not included in the energy level calculation. We now try to incorporate a count of the number of stacked pairings to obtain more stable structure. Include the given information of stacked pairs and make necessary changes in the IP model in part (b) that finds the minimum total free energy correspond to stacked pairs.
Hint: Define a new decision variable to incorporate the stacked pairs into the model.

(e) Now consider pseudo-knots, which occur **if and only if** there are two stacks, $S(i)$ and $S(i')$, starting at positions i and i' respectively, such that every pair in $S(i)$ crosses every pair in $S(i')$. Assume that RNA folds both upwards and downwards. It means that when we consider the line representation, it is possible to have pairings both above and the below of the line (see Figure 4 (right)).

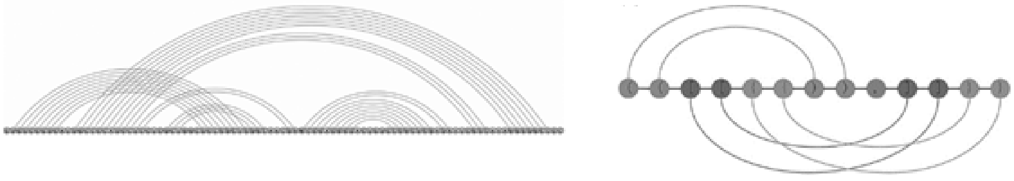


Figure 4: Line representation of pseudo-knots

However, note that cross matching is **not** still allowed, but pseudo-knots are incorporated in the IP model. Considering the information given/obtained in part (d), find a model that gives the minimum total free energy.

(f) Until now, we try to predict the secondary structure of the RNA by IP models. Now, consider a simple dynamic programming (DP) algorithm to minimize the total free energy using the information given/obtained in part (b). To construct the algorithm, begin with creating a pair with two specific points, let say (i, j). Consider the cases until two pairs form the stacked pairs. If two different pairs (*stacked pairs*) are matched, they lead to the energy level, which we try to keep it at minimum. Also note that there is a case that the structure of RNA can bifurcates in such a way that the sum of energies of two substructures is minimized.

Instructions

Please read the following instructions carefully:

(1) Formulate the models in each part separately.

(2) Solve the models using Gurobi or any other solver (CPLEX, Xpress, GAMS etc.)

(3) Prepare a written document including your precise mathematical models. Explain your objective values, constraints, decision variables and parameters explicitly.

(4) Submit your report (including members full names and ID's) as well as your Gurobi model (or your choice of solver) and all of your codes as a .zip file. The name of the .zip file should be your group number (Do not add names, ID's etc. to the file name).

(5) There will be a presentation session where you will be asked questions about your models and the project.