ISyE 3133 Project (First Report)

Sydney Bice, Johanna Lumbantobing

March 25, 2024

1 Introduction

This project involves formulating multiple integer linear programming models to address the Nucleic Acid Folding Problem. The Nucleic Acid Folding Problem involves predicting the secondary structure of a nucleic acid molecule, represented as a circular string of nucleotides, by identifying the most likely pairings of these nucleotides. Additionally, each nucleotide can be in at most one pair. The objective of this project is to model various scenarios to predict the pairing of a circular string of nucleotides given different assumptions and constraints for each model.

2 Models

2.1 Model 1: The First Crude Model

Model 1: The First Crude Model predicts the most likely pairing of a nucleic acid sequence under nested pairings. Nested pairings, also called non-crossing pairings, are pairings for which each pair is connected by a line inside the circular string of nucleotides and none of the lines cross.

2.1.1 Key Assumptions

- Only nested pairings are allowed.
- Only complementary nucleotides are allowed: (A,U) and (G,C).
- The most stable pairing is the most likely.
- The most stable pairing is the one with the most matched pairs.

2.1.2 Variables

- Let x_{ij} be a binary variable that is equal to 1 if the nucleotide at position i pairs with the nucleotide at position j, and 0 otherwise. $x_{ij} \in \{0, 1\}$

The indices i and j are integers with $0 \leq i < j < n$ where n is the length of the nucleotide sequence.

2.1.3 Objective

The objective of the model is to maximize the number of stable pairings (z):

$$\text{Maximize z} = \sum_{i<j}^{n-1} x_{ij} \tag{1}$$

2.1.4 Constraints

1. **Non-Crossing Pairing Constraint:** For any four indices $i < j < k < l$, the model ensures that not both x_{ij} and x_{kl} are 1 simultaneously:

$$x_{ij} + x_{kl} \leq 1 \quad \text{for } 0 \leq i < j < k < l \leq n-1 \tag{2}$$

2. **Single Pairing Constraint:** Ensures that each nucleotide is part of at most one pair:

$$\sum_{j=i+1}^{n-1} x_{ij} \leq 1 \quad \forall i = 0, ..., n-1 \tag{3}$$

2.1.5 Size of Formulation

- Variables: There are $\frac{n(n-1)}{2}$ binary variables x_{ij}.
- Non-Crossing Pairing Constraint: Approximately $O(n^4)$ constraints.
- Single Pairing Constraint: n constraints.
- Complementary Pairing Constraint: Approximately $\frac{n(n-1)}{2}$ constraints.

2.2 Model 2: Simple Biological Enhancements

Model 2: Simple Biological Enhancements predicts the most likely pairings given weight of different pairs.

2.2.1 Key Assumptions

- Only nested pairings are allowed.
- Any pair of nucleotides must be at least distance three away from each other.
- Complementary matched pairs are allowed: (A,U) and (G,C).
- Certain non-complementary matched pairs are allowed: (G,U) and (A,C)
- The most stable pairing is the most likely.
- The stability of a pairing is a weighted function of the number of each type of pairs, shown in Table 1.

Pair	Weight
(G, C)	3
(A, U)	2
(G, U)	0.1
(A, C)	0.05

Table 1: Weights of Pair Types for Model 2

2.2.2 Variables

- Let x_{ij} be a binary variable that is equal to 1 if the nucleotide at position i pairs with the nucleotide at position j, and 0 otherwise. $x_{ij} \in \{0, 1\}$

The indices i and j are integers with $0 \leq i < j < n$ where n is the length of the nucleotide sequence.

2.2.3 Objective

The objective of the model is to maximize the number of stable pairings (z):

$$\text{Maximize z} = \quad 3* \sum_{i<j}^{n-1} x_{ij}^{GC} + \quad 2* \sum_{i<j}^{n-1} x_{ij}^{AU} + \quad 0.1* \sum_{i<j}^{n-1} x_{ij}^{GU} + \quad 0.05* \sum_{i<j}^{n-1} x_{ij}^{AC} \tag{4}$$

Where x_{ij}^{GC} represents a pairing of nucleotides G and C at (i,j)

2.2.4 Constraints

1. Constraints 1 and 2 from Model 1
2. **Distance of 3:** Ensures that nucleotide pairs are at least a distance of 3 from each other:

$$x_{ij} + x_{ji} \leq 1 \quad \forall i = 0, 1, ..., n-5 \; and \; \forall j = i+3, i+4, ..., n-1 \tag{5}$$

2.2.5 Size of Formulation

- Variables: same as model 1
- Constraints: Approximately $\frac{n(n-1)}{2}$ constraints.

2.3 Model 3: More Complex Biological Enhancements

Model 3: More Complex Biological Enhancements predicts the most likely pairings of a nucleic acid sequence under various assumptions including stacked quartets. Stacked quartets consist of two matched pairs, (i,j) and (i+1, j-1).

2.3.1 Key Assumptions

- Any pair of nucleotides must be at least distance three away from each other.
- Complementary matched pairs are allowed: (A,U) and (G,C).
- Certain non-complementary matched pairs are allowed: (G,U) and (A,C).
- The most stable pairing is the most likely.
- The stability of a pairing is a weighted function of the number of each type of pairs, shown in Table 2, and the number of stacked quartets.

Item	Weight
(G, C) pair	3
(A, U) pair	2
(G, U) pair	0.1
(A, C) pair	0.05
Stacked quartet	1

Table 2: Weights of Pair Types for Model 3

2.3.2 Variables

- Let x_{ij} be a binary variable that is equal to 1 if the nucleotide at position i pairs with the nucleotide at position j, and 0 otherwise. $x_{ij} \in \{0, 1\}$
- Let y_{ij} be a binary variable that is equal to 1 if the nucleotide pair at position (i,j) is an inner pair of a stacked quartet, and 0 otherwise. $y_{ij} \in \{0, 1\}$

The indices i and j are integers with $0 \leq i < j < n$ where n is the length of the nucleotide sequence.

2.3.3 Objective

The objective of the model is to maximize the summation of weighted pairs and number of stacked quartets (z):

$$\text{Maximize z} = \quad 3* \sum_{i<j}^{n-1} x_{ij}^{GC} + \quad 2* \sum_{i<j}^{n-1} x_{ij}^{AU} + \quad 0.1* \sum_{i<j}^{n-1} x_{ij}^{GU} + \quad 0.05* \sum_{i<j}^{n-1} x_{ij}^{AC} + \sum_{i<j}^{n-1} y_{ij} \tag{6}$$

2.3.4 Constraints

1. All constrains from Model 2
2. **Stacked Quartet Exists:** If x_{ij} exists and $x_{i+1,j-1}$ exist, then a stacked quartet exists:

$$x_{ij} + x_{i+1,j-1} - y_{ij} \leq 1 \quad \forall \quad i, j = 0, ..., n-1 \tag{7}$$

3. **Pairing Exists:** If stacked quartet exists both X_{ij} and $X_{i+1,j-1}$ pairing exists:

$$2 * y_{ij} - x_{ij} - X_{i+1,j-1} \leq 0 \forall \quad i, j = 0, ..., n-1 \tag{8}$$

2.3.5 Size of Formulation

- Variables: Approximately $n(n-1)$ variables.
- Constraints: Approximately $n(n-1)$ additional constraints to Model 2.

2.4 Model 4: A Model with Crossing Pairs

Model 4: A Model with Crossing Pairs predicts the most likely pairings of a nucleic acid sequence under various assumptions including stacked quartets and the allowance of up to ten crossing pairs.

2.4.1 Key Assumptions

- Up to ten crossing pairs are allowed.
- Any pair of nucleotides must be at least distance three away from each other.
- Both complementary and certain non-complementary matched pairs are allowed:
  - Complementary pairs: (A, U) and (G, C).
  - Non-complementary pairs: (G, U) and (A, C).
- The most stable pairing, determined by a weighted function, is considered the most likely.
- The stability of a pairing is a weighted function of the number of each type of pair, shown in Table 3, and the number of stacked quartets.

Item	Weight
(G, C) pair	3
(A, U) pair	2
(G, U) pair	0.1
(A, C) pair	0.05
Stacked quartet	1

Table 3: Weights of Pair Types for Model 4

2.4.2 Variables

- Let x_{ij} be a binary variable that is equal to 1 if the nucleotide at position i pairs with the nucleotide at position j, and 0 otherwise. $x_{ij} \in \{0, 1\}$
- Let y_{ij} be a binary variable that is equal to 1 if the nucleotide pair at position (i,j) is an inner pair of a stacked quartet, and 0 otherwise. $y_{ij} \in \{0, 1\}$
- Let C_{ijkl} be a binary variable that is equal to 1 if pairings (i,j) and (k,l) cross, and 0 otherwise. $C_{ijkl} \in \{0, 1\}$

The indices i and j are integers with $0 \leq i < j < n$ where n is the length of the nucleotide sequence.

2.4.3 Objective

The objective of the model is to maximize the summation of weighted pairs and number of stacked quartets (z):

$$\text{Maximize z} = \quad 3* \sum_{i<j}^{n-1} x_{ij}^{GC} + \quad 2* \sum_{i<j}^{n-1} x_{ij}^{AU} + \quad 0.1* \sum_{i<j}^{n-1} x_{ij}^{GU} + \quad 0.05* \sum_{i<j}^{n-1} x_{ij}^{AC} + \sum_{i<j}^{n-1} y_{ij} \tag{9}$$

2.4.4 Constraints

1. All constraints from Model 3.
2. **Limit 10 Crossing Pairs:** Ensures there are ≤ 10 crossing pairs:

$$\sum_{i<k<j<l}^{n-1} C_{ijkl} \leq 10 \quad \forall\, 0 \leq i < k < j < l \leq n-1 \tag{10}$$

3. C_{ijkl} **Exists:** If pairs (i,j) and (k,l) are crossing, then C_{ijkl} exists:

$$X_{ij} + X_{kl} - C_{ijkl} \leq 1 \quad \forall\, 0 \leq i < k < j < l \leq n-1 \tag{11}$$

2.4.5 Size of Formulation

- Variables and Constraints: approximately $O(n^4)$.