ISYE 3133 Nucleic Acid Folding Project Formulation Report

Name: Bijie Liu, Xilei Zhu

1. Introduction

   In Biology, there are four kinds of Nucleic Acid alphabet, A, C, G, U and they form the Nucleic Acid. By pairing two characters, we enhance the bond of Nucleic Acid, and each character can only be in one pair. A pair is called Complemtary if the pair is (A, U) or (C, G), otherwise, it is called Non-Complemtary. A pairing is the set of disjoint pairs of characters in S, which denotes a string of n characters made up from $\{A, C, G, U\}$. The entire pairing is called Nested if all pairs are complemtary and when we draw a line to represent the connection, no lines are corssing. Mathematically speaking, we need to have a complemtary planar graph. Complemtary pair has higher weight, Non-Complemtary pair has little weight or none. Higher overall weight represents more stable Nucleic Acid Folding.

   In this assignment, we are givne a string S of length n. For the character of the first position of the string, denote it with $S[1]$, and the last position's character, denote it with $S[n]$.

   For our first model of Simple Biological Enhancements, we are maximizing the total weight of our Nucleic Acid. With the constrains stated in the introduction alongside with three others.

   (a) Distance of any pair must be 3 away.

   (b) Weight of different pairing is different to encourage to pair with a higher weight to gain a more stable Nucleic Acid Folidng.

   (c) We assign little weight to Non-Complemtary pairs to encourage not giving up on the non-paired characters.

   Our second model of More Complex Biological Enhancements, we are introducing several more definitions and constrains. For this model, we are maximizing the total weight plus the number of stacked quartets.

   (a) Base Stacking: A matched pair (i, j) in a nested pairing is called a Stacked Pair if either $(i + 1, j - 1)$ or $(i - 1, j + 1)$ is also a matched pair in the nested pairing.

   (b) Stakced Quartet: If (i, j) and $(i + 1, j - 1)$ are stacked pairs. $(i, i + 1, j - 1, j)$ is called stacked quartet is $i < i + 1 < j - 1 < j$.

   Stacking contributes a lot to the stability to the Nucleic Acid Folding.

   (c) Weight of stacked quartet in nested pairing: Since stacking will enhance the stability of the folding, so we are awarding more weight to those stacked quartet according to the table given.

   (d) We are also awarding more weight to a stacked quartet if it is the first or last quartet in a stack.

   (e) We are allowing crossing pairs, and the number of maxmium crossing pairs is at most 10.

2. Model for The First Crude Model

   Variables:

   (a) Binary Integer lienar variable: $P(i,j)$ for position $i,j \in \{1,\dots,n\}$ and $i<j$ such that: if i pairs with j, then $P(i,j)=1$.

   (b) $a_i = 1$ if $S[i] = A$, otherwise $a_i = 0$.

   (c) $u_i = 1$ if $S[i] = U$, otherwise $u_i = 0$.

   (d) $g_i = 1$ if $S[i] = G$, otherwise $g_i = 0$.

   (e) $c_i = 1$ if $S[i] = C$, otherwise $c_i = 0$.

   Objective Function:
   Max $\sum_{i<j} P(i,j)$, $\forall i,j \in \{1,\dots,n\}$

   Constrains:

   (a) $P(i,j) \in \{0,1\}$, $\forall i,j \in \{1,\dots,n\}$ and $i<j$.

   (b) $a_i, u_i, g_i, c_i \in \{0,1\}, \forall i \in \{1,\dots,n\}$.

   (c) For each $j \in \{1,\dots,n\}$, $\sum_{k>j} P(j,k) + \sum_{k<j} P(k,j) \leq 1$ for all $k \in \{1,\dots,n\}$. Each character can be in at most one matched pair.

   (d) For every four positions, say $i, i', j, j' \in \{1,\dots,n\}$ where $i < i' < j < j'$ such that $P(i,j) + P(i',j') \leq 1, \forall i,j,i',j' \in \{1,\dots,n\}$. Non-crossing or nested condition.

   Complemtary Pairs Constrains:

   (e) $P(i,j) \leq 2 - (u_i + u_j + g_i + g_j)$, $\forall i,j \in \{1,\dots,n\}$ and $i<j$.

   (f) $P(i,j) \leq 2 - (u_i + u_j + c_i + c_j)$, $\forall i,j \in \{1,\dots,n\}$ and $i<j$.

   (g) $P(i,j) \leq 2 - (a_i + a_j + c_i + c_j)$, $\forall i,j \in \{1,\dots,n\}$ and $i<j$.

   (h) $P(i,j) \leq 2 - (a_i + a_j + g_i + g_j)$, $\forall i,j \in \{1,\dots,n\}$ and $i<j$.

3. Model for Simple Biological Enhancements

   Variables:

   (a) Binary Integer linear variable: $P(i,j)$ for position $i,j \in \{1,\dots,n\}$ and $i<j$ such that: if i pairs with j, then $P(i,j)=1$.

   Data:

   (a) $\forall i,j \in \{1,\dots,n\}$, $i<j$, $S[i] \neq S[j]$, such that:
   $S[i], S[j] \in \{C,G\}$, then $W(i,j) = 3$.
   $S[i], S[j] \in \{A,U\}$, then $W(i,j) = 2$.
   $S[i], S[j] \in \{G,U\}$, then $W(i,j) = 0.1$.
   $S[i], S[j] \in \{A,C\}$, then $W(i,j) = 0.05$.
   Otherwise, $W(i,j) = 0$.

Objective Function:
Max $\sum_{i<j} W(i,j) * P(i,j)$, $\forall i,j \in \{1,\ldots,n\}$. Maximizing the total weight of Nucleic Acid Folding.

Constrains:

(a) $P(i,j) \in \{0,1\}$, $\forall i,j \in \{1,\ldots,n\}$, and $i < j$.

(b) $P(i,j) = 0$ if $|i-j| \leq 3, \forall i,j \in \{1,\ldots,n\}$. Any pair must be at least distance 3 away.

(c) For each $j \in \{1,\ldots,n\}$, $\sum_{k>j} P(j,k) + \sum_{k<j} P(k,j) \leq 1$ for all $k \in \{1,\ldots,n\}$. Each character can be in at most one matched pair.

(d) For every four positions, say $i,i',j,j' \in \{1,\ldots,n\}$ where $i < i' < j < j'$ such that $P(i,j) + P(i',j') \leq 1, \forall i,j,i',j' \in \{1,\ldots,n\}$. Non-crossing or nested condition.

4. Model for More Complex Biological Enhancements
Variables:

(a) Binary Integer lienar variable: $P(i,j)$ for position $i,j \in \{1,\ldots,n\}$ and $i < j$ such that: if i pairs with j, then $P(i,j) = 1$.

(b) Integer lieanr variable: $C(i,j,i',j')$ for position $i,j,i',j' \in \{1,\ldots,n\}$ such that: if (i,j) and (i',j') are matched pairs that cross, then $C(i,j,i',j') = 1$, $\forall i,j,i',j' \in \{1,\ldots,n\}$.

(c) Binary Integer linear variable $Q(i,j)$ for position $i,j \in \{1,\ldots,n\}$ and $i < j$. $Q(i,j) = 1$ if the matched pair (i,j) is the first pair in a stacked quartet, otherwise, $Q(i,j) = 0$.

(d) Binary Integer linear variable $F(i,j)$ for position $i,j \in \{1,\ldots,n\}$ and $i < j$. $F(i,j) = 1$ if $(i,i+1,j-1,j)$ is the first stacked quartet in a stack.

(e) Binary Integer linear variable $L(i,j)$ for position $i,j \in \{1,\ldots,n\}$ and $i < j$. $L(i,j) = 1$ if $(i,i+1,j-1,j)$ is the last stacked quartet in a stack.

Data:

(a) $\forall i,j \in \{1,\ldots,n\}$ and $i < j-2$, such that:
If $S[i], S[j], S[i+1], S[j-1] = A, U, A, U$ respectively, then $V(i,j) = 9$.
If $S[i], S[j], S[i+1], S[j-1] = A, U, C, G$ respectively, then $V(i,j) = 21$.
If $S[i], S[j], S[i+1], S[j-1] = A, U, G, C$ respectively, then $V(i,j) = 24$.
If $S[i], S[j], S[i+1], S[j-1] = A, U, U, A$ respectively, then $V(i,j) = 13$.
If $S[i], S[j], S[i+1], S[j-1] = C, G, A, U$ respectively, then $V(i,j) = 22$.
If $S[i], S[j], S[i+1], S[j-1] = C, G, C, G$ respectively, then $V(i,j) = 33$.
If $S[i], S[j], S[i+1], S[j-1] = C, G, G, C$ respectively, then $V(i,j) = 34$.
If $S[i], S[j], S[i+1], S[j-1] = C, G, U, A$ respectively, then $V(i,j) = 24$.

If $S[i], S[j], S[i+1], S[j-1] = G, C, A, U$ respectively, then $V(i,j) = 21$.
If $S[i], S[j], S[i+1], S[j-1] = G, C, C, G$ respectively, then $V(i,j) = 24$.
If $S[i], S[j], S[i+1], S[j-1] = G, C, G, C$ respectively, then $V(i,j) = 33$.
If $S[i], S[j], S[i+1], S[j-1] = G, C, U, A$ respectively, then $V(i,j) = 16$.
Otherwise, $W(i,j) = 0$.

(b) $\forall i, j \in \{1, \ldots, n\}$, $i < j$, $S[i] \neq S[j]$, such that:
$S[i], S[j] \in \{C, G\}$, then $W(i,j) = 3$.
$S[i], S[j] \in \{A, U\}$, then $W(i,j) = 2$.
$S[i], S[j] \in \{G, U\}$, then $W(i,j) = 0.1$.
$S[i], S[j] \in \{A, C\}$, then $W(i,j) = 0.05$.
Otherwise, $W(i,j) = 0$.

Objective Function:
Max $\sum_{i=1}^{n} V(i,j) * Q(i,j) + \sum_{i<j} W(i,j) * P(i,j) + \sum_{i<j} Q(i,j)$, $\forall i, j \in \{1, \ldots, n\}$ and $i < j$.

Constrains:

(a) $P(i,j) \in \{0,1\}$, $\forall i, j \in \{1, \ldots, n\}$, and $i < j$.

(b) $C(i,j,i',j') \in \{0,1\}$, $\forall i, j \in \{1, \ldots, n\}$, and $i < j$.

(c) $Q(i,j) \in \{0,1\}$, $\forall i, j \in \{1, \ldots, n\}$, and $i < j$.

(d) $F(i,j) \in \{0,1\}$, $\forall i, j \in \{1, \ldots, n\}$, and $i < j$.

(e) $L(i,j) \in \{0,1\}$, $\forall i, j \in \{1, \ldots, n\}$, and $i < j$.

(f) $P(i,j) = 0$ if $|i - j| \leq 3$. Any pair must be at least distance 3 away, $\forall i, j \in \{1, \ldots, n\}$.

(g) For each $j \in \{1, \ldots, n\}$, $\sum_{k>j} P(j,k) + \sum_{k<j} P(k,j) \leq 1$ for all $k \in \{1, \ldots, n\}$. Each character can be in at most one matched pair.

Allowing Crossing Constrains:

(h) For any choice of $i, j, i', j' \in \{1, \ldots, n\}$, such that $i < i' < j < j'$, $P(i,j) + P(i',j') - C(i,i'j,j') \leq 1$. To allow crossing pairs.

(i) $\sum_{i,i',j,j'} C(i,j,i',j') \leq 10$. Allowing up to at most $C = 10$ crossing pairs.

Counting Stacked Quartets:

(j) $P(i,j) + P(i+1,j-1) - Q(i,j) \leq 1$, $\forall i, j \in \{1, \ldots, n\}$ where $j > i$ and $i < i+1 < j-1 < j$. Enforcing if (i,j) and $(i+1,j-1)$ are in the nested pairing, then $Q(i,j)$ must be 1.

(k) $2Q(i,j) - P(i,j) - P(i+1,j-1) \leq 0$, $\forall i, j \in \{1, \ldots, n\}$. If $Q(i,j)$ is set to 1. then $P(i,j)$ and $P(i+1,j-1)$ must be set to 1.

Determining first/last stacked quartet in a stack

(l) $Q(i,j) - Q(i-1,j+1) - F(i,j) \leq 0$, $\forall i, j \in \{1, \ldots, n\}$. If (i,j) pair is the first pair in its quartet, and $(i+1,j-1)$ is the first pair in its quartet, then the stacked quartet of $(i, i+1, j, j-1)$ is the first quartet in its stacked quartet.

(m) $2F(i,j) - Q(i,j) + Q(i-1,j+1) \leq 1$, $\forall i,j \in \{1,\ldots,n\}$. If $F(i,j)$ is set to 1, then both $Q(i,j)$ and $Q(i-1,j+1)$ need to be set to 1.

(n) $P(i,j) - P(i-1,j+1) - L(i,j) \leq 0$, $\forall i,j \in \{1,\ldots,n\}$. If (i,j) pair is the last pair in its quartet, and $(i+1,j-1)$ is the last pair in its quartet accordingly, then the stacked quartet of $(i,i+1,j,j-1)$ is the last quartet in its stacked quartet.

(o) $2L(i,j) - P(i,j) + P(i-1,j+1) \leq 1$, $\forall i,j \in \{1,\ldots,n\}$. If $L(i,j)$ is set to 1, then both $Q(i,j)$ and $Q(i-1,j+1)$ need to be set to 1.

(p) $\forall i,j \in \{1,\ldots,n\}$, $W(i,j) = F(i,j) * W(i,j) + W(i,j)$. Doubled the weight if its first quartet.

(q) $\forall i,j \in \{1,\ldots,n\}$, $W(i,j) = L(i,j) * W(i,j) + W(i,j)$. Doubled the weight if its last quartet.

5. Size of formulation:
We are given a length of n string.

(a) 1.1 The first crude model
Number of Variables:

i. Number of $P(i,j) = (n-1)!$
ii. Number of $a_i = n$
iii. Number of $u_i = n$
iv. Number of $g_i = n$
v. Number of $c_i = n$

Number of Constrains:

i. (a): $(n-1)!$
ii. (b): $4n$
iii. (c): n
iv. (d): $(n-3)(n-2)(n-1)n$
v. (e-h): $4(n-1)!$

Let $f(n)$ denotes the mapping from $|S| = n$ to number of variables and constrains, and $f(n) = 2((n-1)!) + 5n + (n-3)(n-2)(n-1)n + 4(n-1)!$.

(b) 1.2 Simple Biological Enhancements
Number of Variables:

i. Number of $P(i,j) = (n - 1!)$

Number of Constrains:

i. (a): $(n-1)!$
ii. (b): $3n$
iii. (c): n
iv. (d): $(n-3)(n-2)(n-1)n$

Let $g(n)$ denotes the mapping from $|S| = n$ to number of variables and constrains, and $f(n) = 2((n-1)!) + 4n + (n-3)(n-2)(n-1)n + 4(n-1)!$.

(c) 1.3 More Complex Biological Enhancements
Number of Variables:

i. Number of $P(i,j) = (n-1)!$
ii. Number of $P(i,j) = \binom{n}{4}$
iii. Number of $Q(i,j) = (n-1)!$
iv. Number of $F(i,j) = (n-1)!$
v. Number of $L(i,j) = (n-1)!$

Number of Constrains:

i. (a): $(n-1)!$
ii. (b): $\binom{n}{4}$
iii. (c,d,e): $3((n-1)!)$
iv. (f): $3n$
v. (g): n
vi. (h): $\binom{n}{4}$
vii. (i): 1
viii. (j,k, ... , q): $8((n-1)!)$

Let $p(n)$ denotes the mapping from $|S| = n$ to number of variables and constrains, and $p(n) = 16(n-1)! + 3 * \binom{n}{4} + 4n + 1$