ACADEMY OF ATHENS

HELLENIC REPUBLIC
National and Kapodistrian University of Athens

Postgraduate Program: «Data Science and Information Technologies»

Algorithms in Structural Biology

Assignment #1

Postgraduate Student: Gavrielatos Marios
Registration Number: 7115152100023

The following packages must be installed by `install.packages()`:

1. "wordspace"
2. "hash"

Problem 1: RNA folding

The table was initialized with $j + 5 > i \Rightarrow E(i,j) = 100$ for $i > j$:

	A1	A2	U3	A4	C5	U6	C7	C8	G9	U10	U11	G12	C13	A14	G15	C16	A17	U18
A1	100	100	100	100	100	0	0	0	0	0	0	0	0	0	0	0	0	0
A2	100	100	100	100	100	100	0	0	0	0	0	0	0	0	0	0	0	0
U3	100	100	100	100	100	100	100	0	0	0	0	0	0	0	0	0	0	0
A4	100	100	100	100	100	100	100	100	0	0	0	0	0	0	0	0	0	0
C5	100	100	100	100	100	100	100	100	100	0	0	0	0	0	0	0	0	0
U6	100	100	100	100	100	100	100	100	100	100	0	0	0	0	0	0	0	0
C7	100	100	100	100	100	100	100	100	100	100	100	0	0	0	0	0	0	0
C8	100	100	100	100	100	100	100	100	100	100	100	100	0	0	0	0	0	0
G9	100	100	100	100	100	100	100	100	100	100	100	100	100	0	0	0	0	0
U10	100	100	100	100	100	100	100	100	100	100	100	100	100	100	0	0	0	0
U11	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	0	0	0
G12	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	0	0
C13	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	0
A14	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
G15	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
C16	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
A17	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
U18	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100

and bond energy $b(i,j) = -4, 0, 4$, for Watson-Crick bonds, GU, and all other possible pairs respectively.

The filled in E-table is the following:

	A1	A2	U3	A4	C5	U6	C7	C8	G9	U10	U11	G12	C13	A14	G15	C16	A17	U18
A1	100	100	100	100	100	96	96	96	96	96	92	92	92	92	88	88	84	80
A2	100	100	100	100	100	100	100	100	100	96	92	92	92	92	88	88	84	80
U3	100	100	100	100	100	100	100	100	100	96	96	96	96	92	88	88	84	84
A4	100	100	100	100	100	100	100	100	100	96	96	96	96	92	88	88	88	84
C5	100	100	100	100	100	100	100	100	100	100	100	96	96	92	88	88	88	88
U6	100	100	100	100	100	100	100	100	100	100	100	96	96	92	92	92	92	92
C7	100	100	100	100	100	100	100	100	100	100	100	96	96	96	96	96	96	96
C8	100	100	100	100	100	100	100	100	100	100	100	100	100	100	96	96	96	96
G9	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	96	96	96
U10	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	96	96
U11	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	96	96
G12	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
C13	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
A14	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
G15	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
C16	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
A17	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
U18	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100

The traceback matrix we created is the following:

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17	V18
1	100	100	100	100	100	3	1	1	1	1, 2, 3	2, 3	1, 2	1, 2	1, 2	2	1, 2	2	2, 3
2	100	100	100	100	100	100	1, 2	1, 2	1, 2	2, 3	3	1	1	1, 2	2	1, 2	2	3
3	100	100	100	100	100	100	100	1, 2	1, 2, 3	2	1, 2	1, 2, 3	1, 2	2, 3	2	1, 2	3	1, 2
4	100	100	100	100	100	100	100	100	1, 2	3	1, 3	1, 2	1, 2	2	2	1, 2	1, 2	3
5	100	100	100	100	100	100	100	100	100	1, 2	1, 2	2, 3	1, 2	2	3	1	1	1
6	100	100	100	100	100	100	100	100	100	100	1, 2	2	1, 2	3	1	1	1, 3	1
7	100	100	100	100	100	100	100	100	100	100	100	3	1	1	1, 2, 3	1, 2	1, 2	1, 2
8	100	100	100	100	100	100	100	100	100	100	100	100	1, 2	1, 2	3	1, 2	1, 2	1, 2
9	100	100	100	100	100	100	100	100	100	100	100	100	100	1, 2	1, 2	3	1, 2	1, 2, 3
10	100	100	100	100	100	100	100	100	100	100	100	100	100	100	1, 2, 3	1, 2	2, 3	1, 2
11	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	1, 2	3	1
12	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	1, 2	1, 2, 3
13	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	1, 2
14	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
15	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
16	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
17	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
18	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100

In every cell of the matrix, we store all the possible paths this cell was created from:

- Left: 1
- Down: 2
- Down left: 3
- Forth case (folds of r_j, ..., r_i comprised of 2 folds): 4

From the traceback we get 4 paths with the following bonds. Only two different bond combinations exist (bold):

1. **(A2 - U18), (U3 - A17), (C5 - G15), (U6 - A14), (C7 - G12)**

   Backtrack path (row, col): (1 18), (2 18), (3 17), (4 16), (4 15), (5 15), (6 14), (7 13), (7 12)

2. **(A2 - U18), (U3 - A17), (C5 - G15), (U6 - A14), (C7 - G12)**

   Backtrack path (row, col): (1 18), (2 18), (3 17), (4 16), (5 16), (5 15), (6 14), (7 13), (7 12)

3. **(A1 - U18), (U3 - A17), (C5 - G15), (U6 - A14), (C7 - G12)**

   Backtrack path (row, col): (1 18), (2 17), (3 17), (4 16), (4 15), (5 15), (6 14), (7 13), (7 12)

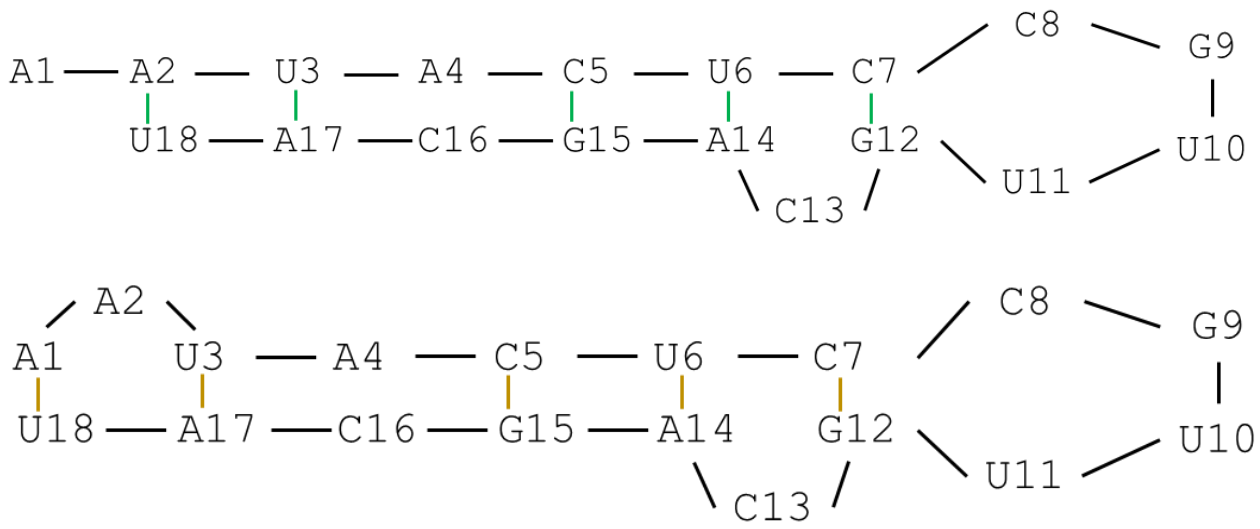
4. **(A1 - U18), (U3 - A17), (C5 - G15), (U6 - A14), (C7 - G12)**

   Backtrack path (row, col): (1 18), (2 17), (3 17), (4 16), (5 16), (5 15), (6 14), (7 13), (7 12)

The 4 paths can be seen in the following table. The green arrow corresponds to the first and second case and the yellow arrow corresponds to the third and fourth case:

	A1	A2	U3	A4	C5	U6	C7	C8	G9	U10	U11	G12	C13	A14	G15	C16	A17	U18
A1	100	100	100	100	100	96	96	96	96	96	92	92	92	92	88	88	84	80
A2	100	100	100	100	100	100	100	100	100	96	92	92	92	92	88	88	84	80
U3	100	100	100	100	100	100	100	100	100	96	96	96	96	92	88	88	84	84
A4	100	100	100	100	100	100	100	100	100	96	96	96	96	92	88	88	88	84
C5	100	100	100	100	100	100	100	100	100	100	100	96	96	92	88	88	88	88
U6	100	100	100	100	100	100	100	100	100	100	100	96	96	92	92	92	92	92
C7	100	100	100	100	100	100	100	100	100	100	100	96	96	96	96	96	96	96
C8	100	100	100	100	100	100	100	100	100	100	100	100	100	100	96	96	96	96
G9	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	96	96	96
U10	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	96	96
U11	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	96	96
G12	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
C13	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
A14	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
G15	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
C16	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
A17	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
U18	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100

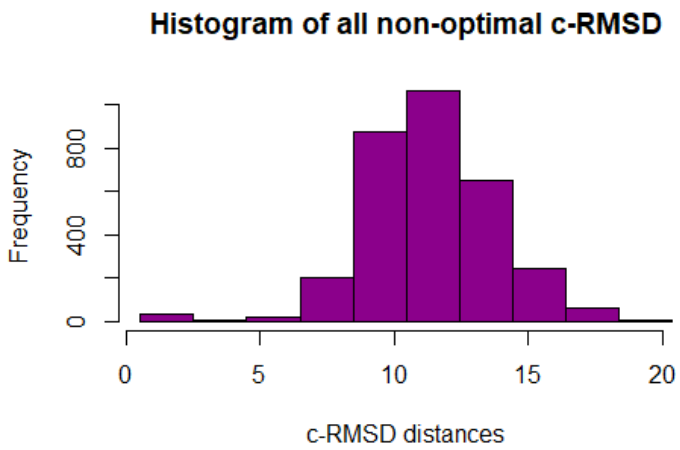
The two structures are the following:



Problem 2: c-RMSD and d-RMSD

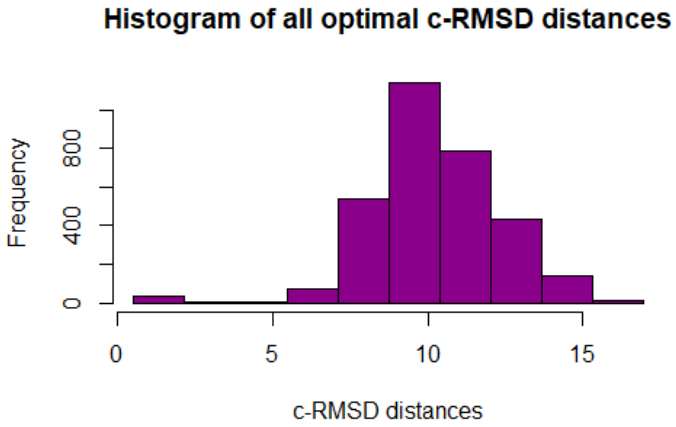
For the first problem we computed both the general c-RMSD values and the optimal minimum c-RMSD values. The results for the **non-optimal c-RMSD** are the following:

- Mean c-RMSD: 11.3047
- Median c-RMSD: 11.1957



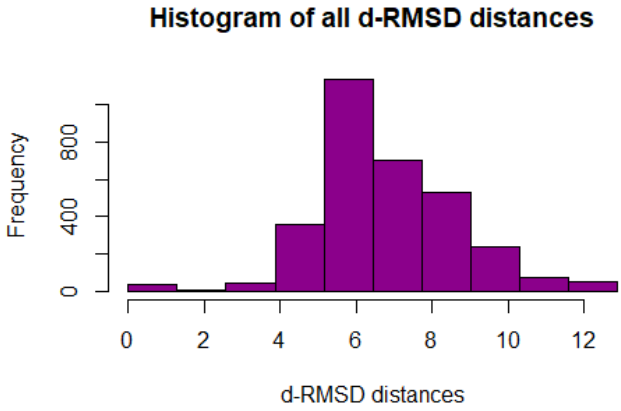
The results for the **optimal minimum c-RMSD**:

- Mean c-RMSD: 10.2168
- Median c-RMSD: 10.0910



We repeated the same experiment using all k = $\binom{n}{2}$ distances in order to calculate the d-RMSD value. The results are the following:

- Mean d-RMSD: 6.7929
- Median d-RMSD: 6.4686



Furthermore, it is true that $\frac{c-RMSD}{\sqrt{n}} \leq d - RMSD \leq 2 \cdot c - RMSD$

Problem 3: Distances

We created a TXT file (covid.txt) with the coordinates of 50 Ca atoms indexed A102 to A151 of the main protease of SARS-COV-2 (PDB id: 6LU7). We construct the distance matrix M with dimensions 50x50. We then create a 51x51 Cayley-Menger matrix (B) by appending a 0th row and a 0th column to distance matrix M.

$$\begin{bmatrix} 0 & 1 \cdots 1 \\ 1 & \\ \vdots & M \\ 1 & \end{bmatrix}$$

1. The Cayley-Menger matrix has **rank(B) = 5** which means that the distance matrix M expresses a 3D conformation which is true because embeddable matrices in $\mathbb{R}^3$ correspond to 3D conformations (theorem: Cayley:1841, Menger'28).

2. We perturb entries of the Cayley-Menger matrix B by 5% and 10%. We *randomly* chose if we are going to add or subtract the 5% or 10% of every value from itself. The perturbated matrices have both rank = 51.
3. We created the Gram matrix G for every perturbated matrix and we applied SVD: $G = V\Sigma V^T$. We chose the 3 largest singular values of Σ in order to force rank(G)=3 by defining the diagonal matrix Σ'.
4. We calculated the output coordinates for every perturbation: $P = \sqrt{\Sigma'}\, V^T$ which is a 50x3 matrix.
5. We calculated the c-RMSD of the 2 perturbated structures against the original structure:
    a. **c-RMSD of 5% perturbation vs original structure: 63.051**
    b. **c-RMSD of 10% perturbation vs original structure: 63.1165**