BioMed Data Mining Term Project 2024 Fall
Due Date 11:59PM Dec 28

Objective:

Implement a program to predict the secondary structure **commonly shared** in an RNA family.

Descriptions:

1. You can implement the prediction program based on the methods introduced in lectures, e.g., dot matrix sequence comparison, sequence covariation analysis, GPRM, or any other methods you can find in literature.
2. Your program must be written in C/C++ or Python. No other programming languages are permitted.

Evaluation criteria:

The performance of your program will be evaluated according to the number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) of your predicted basepairing.

> Note. Two canonical basepairing rules (i.e., A-U, G-C) and one non-canonical basepairing (i.e., G-U) should be considered in your program.

Truth **Prediction**	Basepair	Non-basepair
Basepair	TP	FP
Non-basepair	FN	TN

Input to your program:

A set of RNA sequences in FASTA format.

E.g.,

```
>seq1
ACCUUGGGGAAGGAAGGU
>seq2
UGCAUGGGGAAGGAUGCA
>seq3
UGUCAGGGGGUGGCA
```

Output from your program:

An alignment that shows the prediction of basepairing.

```
>seq1
ACCUUGGGGAAGGAAGGU
>seq2
UGCAUGGGGAAGGAUGCA
>seq3
UGUCAGGGG---GUGGCA
(((((.........)))))
```

Datasets:

Three sets of RNA sequences in FASTA format.

What to submit:

1. Your source program. Make sure it can be compiled and executed.
2. A description file that explains the method your program is based on.
3. The output from your program.

Demo:

Some students will be selected at random by TAs to demonstrate their programs in Lab.