

Assignment 4

Sangeet M
M.Tech AI
19322

How to run:

- past the folder navigation to where the files are stored in the variable 'path'
- if the variables are in the folder, sharing with the program, leave it as how it was
- run the code Assignment4.py

Program explained:

- The program is split into 3 parts
 - Initialisation
 - Body
 - Results
- Initialization:
 - The data is fetched from the files and stored in respective variables
 - A, C, G, and T are variables where the number of "A," "C," "G," and "T" occurrences are stored cumulatively.
 - The index ranges for exons are stored in RX_ind and GX_ind.
- Body:
 - The body mainly has two functions
 - Step
 - SS (Search Seq)
 - each input is sliced of the '\n,' inverted and complemented(A and T and interchanged so are C and G)
 - Then the function sub_step is called for the input, and the complemented and inverted input
 - In substep calls the SS function, which executes the BWT algorithm and returns the index or range along with a shift.
 - With this value, the exon mappings are marked

- These exon mapping are then appended into a global exon variable for probability calculation
- This is executed for all the inputs in the file reads
- Results:
 - Here the exon values are collected.
 - The probabilities for the sequence to be in which of the four slots mentioned in the class are calculated.
 - The probabilities for the sequence to be in each slot, the maximum of the probabilities and, at last, the slot number, are printed in that order.
- Outputs:

Exon : [90.5 295. 173. 241. 518.5 222.]

Probabilities : [0.12063643 0.00111204 0.55631756 0.32193397]

Max Probability : 0.5563175575580325

Max Prob Index : 3

P.S.

The configurations are as given below:

- 1) 50%,50%,50%,50%
- 2) 100%,100%,0%,0%
- 3) 33%,33%,100%,100%
- 4) 33%,33%,33%,100%