BM4321 Genomic Signal Processing
Semester 7 – 2017 Batch – 2021/2022 Academic Year

Assignment 1 – Individual - 50% of Final Grade
Promoter Discovery in Bacteria

For the specific GenBank accession of the bacterium *Salmonella enterica* assigned to the student, download the genome (in FASTA format) and protein table. Obtain 50 bases upstream and 3 bases downstream for all possible genes of the sense and antisense strands. Initially for a few known genes use the presence of codons ATG and GTG at the beginning of each coding sequence to verify your code. ***A gene can be neglected if the previous gene is less than 50 bases upstream of it.***

For the obtained sequence perform the following operations.

1. Obtain the distribution of bases between genes and the proportion of genes that can be neglected for being less than 50 bases downstream of the next one.

2. Perform a W matching local search (for an intact query) to locate a Pribnow box promoter within upstream positions 5 to 30 of each sequence. Using the first 1000 sequences, obtain a position probability matrix (PPM) with 10 positions for the Pribnow box.

3. Using a suitable entropy measure, eliminate the redundant positions of (2). Plot the distribution of the entropy vs. number of positions and hence, select a suitable threshold.

4. Perform a statistical alignment for the remaining sequences of (2) using the initial PPM of (2) and the reduced PPM of (3). Compare the two results and hence for the aligned sequences determine the proportion of genes that do not have promoters. For the statistical alignment you may use the thresholds of -1 to -5 (in decrements of 1) after normalizing with the consensus score. Give your result in terms of the proportion of genes used for testing that have detectable promoters.

5. Repeat (2) to (4) by randomly selecting (i) 10 and (ii) 100 samples from the initial 1000 samples used for the PPM. Compare and comment on the three results. ***You have to use the same testing set used in (2) to (4).***

6. Using the reduced Pribnow box PPM of (3) perform a statistical alignment for the given thresholds for the remaining 14 genomes of *S. enterica*. Find the proportion of genes used for testing that have detectable promoters for each genome.

Deliverables:

1. Brief report (.pdf via LMS) with results and discussion of the questions along with the number of base pairs and genes in the sense and antisense stands of the organism. Maximum 8 pages. Code, protein tables, FASTA files, table of contents, *acknowledgements are not required to be included in the report.*

2. Soft copies of code, the FASTA file of genome, protein table and any derivative sequences (.txt or FASTA) have to accompany the LMS submission. These have to be compressed into a single file.

Due on: 2021.10.11

No.	Student ID	Sequence
1	160014D	NZ_CP066047.1
2	170004G	NZ_CP028172.1
3	170015P	NZ_CP030194.1
4	170017A	NZ_CP030231.1
5	170131R	NZ_CP030238.1
6	170146R	NZ_CP037891.1
7	170147V	NZ_CP040380.1
8	170173V	NZ_CP046277.1
9	170208K	NZ_CP046279.1
10	170213V	NZ_CP046280.1
11	170221T	NZ_CP046291.1
12	170259P	NZ_CP053581.1
13	170348M	NZ_CP060508.1
14	170401V	NZ_CP069518.1
15	170418E	NZ_CP075108.1

Upeka Premaratne (upeka@uom.lk 0719538433)
2021.09.17