Algoritmos Bioinformática /Bioinformática 2022/2023

Group Assignment 1

Saccharomyces cerevisiae is a type of yeast that is used in baking, brewing, and winemaking. It is unicellular and one of the simplest eukaryotic organism that is widely studied in genetics and molecular biology. It has a relatively simple genome of approximately 12 million base pairs. It has 16 chromosomes and contains approximately 6,000 genes.
Alternative splicing is relatively rare in *S. cerevisiae* and the majority of the genes have one single exon.

In this assignment, the goal is to identify the set of open reading frames (ORFs) that are located in a segment of the yeast genome. This segment will consist in the first 30kb of the genomic sequence of chromosome I that has already been selected for posterior analysis. For that you will only consider the first 30,000 bp of the provided sequence (sequence_chr1.fasta), ignoring the remaining sequence.

The elements of the genomic sequence can be visualized in different platforms, including the UCSC Genome Browser. You can go to this tool and select the yeast genome for visualization:
https://genome.ucsc.edu/cgi-bin/hgGateway?hgsid=1569500417_h2K0XvpPLwGvUWUoEqBslBjakSwP

In the multiple tracks available, you should focus on the "*Refseq Curated*" track that shows in blue segments the coordinates of different ORFs along this genome.

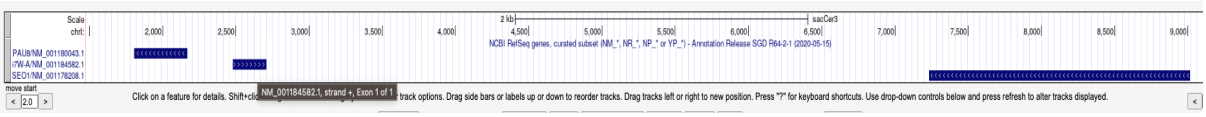

Figure 1: a screenshot of the selected region (1.5KB to 9KB) from chromosome 1. The blue segments indicate the start and end of the ORF. Each ORF represents a gene.

Additionally, a table is provided with an annotation that contains the coordinates of these ORFs along the genome. The file, *genes_chr1.gtf*, has an extension called *GTF,* standing for *Gene Transfer Format*.

For instance, the second gene in this figure is represented by the following line in the file:

chrI sacCer3_ncbiRefSeq exon 2480 2707 0.000000 + . gene_id "NM_001184582.1"; transcript_id "NM_001184582.1";

Field 1 represents the chromosome, field 3 the type of event (exon), fields 4 and 5 are the coordinates. Field 7 is the genomic strand or direction (+) and the gene id of the ORF is given by "NM_001184582.1";

A. Get statistics

Note that the genome has two strands running from 5'->3', also called positive strand or + strand. The strand running from 3'->5' is called the negative or – strand. You should search for ORFs in both directions. The positive (+) strand is given by the direct analysis of the provided sequence. The negative strand is obtained by searching the reverse complement of the positive strand.

The genomic sequence should be read from both strands (+ and - strands) to obtain the following statistics:

1. Length of the sequence.
2. Frequency (in %) of A, C, G, T.
3. GC content.
4. Number of Start (AUG) codons found.
5. Number of Stop Codons (UAA, UAG, UGA).
6. Most and least frequent codons.

B. Get ORFs

Using the genome sequence as input, identify all the potential ORFs in the positive and negative strands.

- An ORF is defined as the region that begins with the start codon (AUG) and ends with the stop codon (UAA, UAG, UGA). Note that we are working directly on the DNA sequence so U should be replaced by T.
- For a given region, if alternative start codons are found, select the longest ORF.
- Select all ORFs with a minimum length of 150 nucleotides (50 amino acids).

In this step, you should output the following information:

7. A **file** with all the protein sequences named ***all_potential_proteins.txt***, with a sequence per line.
8. A **file** with the genomic coordinates of all the ORFs, named ***orf_coordinates.txt***. The genomic coordinates correspond the start and end position in the genome in the format:

    Start1, End1, ORF1

    Start2, End2, ORF2

    …..

    StartN, EndN, ORFN

C. Overlap with annotation

Compare the results you obtained with those from the annotation in file *genes_chr1.gtf*. For the overlap comparison focus on the entries of the type Exon (3rd field).

The overlap between sequences A and B can be calculated as:

Overlap(A, B) = $\frac{|A \cap B|}{\min(A,B)}$, where the intersection region between A and B is defined by the length of the genomic region in common between A and B. For this exercise, we will define a slightly different version of the overlap, where A is one of the ORFs in the annotation and B is an ORF from your list. Thus, the overlap is defined taking in account the length of the ORF in the annotation.

Overlap (A, B) = $\frac{|A \cap B|}{|A|}$

9. For each of the ORFs in the annotation file, you should output the longest overlap obtained with an ORF in your list. Use the identifiers in the field "gene_id" to identify the ORF, e.g.

    NM_001180043.1 72%

    NM_001184582.1 93%

    ….

The above example is totally hypothetical. However, in that case, NM_001180043.1 has an overlap of 72% with one of your ORFs, being 72 the percentage corresponding to the longest overlap. **Hint:** Define a function that compares the overlap between two sequences as defined above and test for each ORF in annotation the overlap with each of the ORFs in your list.

10. Output

For this exercise, you should submit a python script named **yeast_orfs_chr1.py**.

Note that all submissions in a different format (including Jupyter notebooks) will not be considered!

In the script create a header with the name, number and course of each element of the group. Also, add any notes that you find relevant regarding the contribution of each element to the work. For instance, if an element only participated in a specific part of the work, explicitly mention that in the header as a comment.

The file should be run in the terminal as:

python yeast_orfs_chr1.py sequence_chr1.fasta genes_chr1.gtf

- Points **1 to 6 and 9**, should be written to **output**, with **the one item per line.**
- Points **7 and 8** should save the results in **files** with the names as indicated above.