# ΥΠΟΛΟΓΙΣΤΙΚΗ ΒΙΟΛΟΓΙΑ
# 2η εργασία

## Οδηγίες

Να παραδώσετε μέσω eclass, έναν συμπιεσμένο φάκελο zip με όνομα τον αριθμό μητρώου σας, που να περιέχει:

i. ένα **pdf ή doc** με τις απαντήσεις σας για το Μέρος Α, το ονοματεπώνυμό σας και τον αριθμό μητρώου σας
ii. ένα Perl script «**first_exon_from_bed.pl**» με τον κώδικά σας για το Μέρος Β
iii. το BED αρχείο «**first_exons_coordinates.bed**» που δημιουργήσατε στο Μέρος Β
iv. το BED αρχείο «**exonic_length_per_transcript.txt**» που δημιουργήσατε στο Μέρος Β
v. το BED αρχείο «**human_exons_prCoding_exercise_set.bed**», όπως αυτό σας δίνεται στο eClass.

Για απορίες μπορείτε να χρησιμοποιήσετε την περιοχή συζητήσεων στο eClass του μαθήματος.

## Μέρος Α (4 μονάδες)

Ύστερα από αναζήτηση στο internet και το υλικό του μαθήματος να απαντήσετε στις εξής ερωτήσεις:

1. Τα BED αρχεία χρησιμοποιούνται για την αποθήκευση γονιδιωματικών περιοχών. Να αναφέρετε τι πληροφορίες αποθηκεύουμε σε κάθε μία από τις στήλες του αρχείου. Ποιες στήλες είναι υποχρεωτικές και ποιες προαιρετικές;

2. Ο προσδιορισμός της αρχής και ο προσδιορισμός του τέλους μιας περιοχής σε ένα BED αρχείο ακολουθούν αρίθμηση που ξεκινάει από το μηδέν (zero-based) ή από το 1 (one-based);

3. Σε ένα BED αρχείο υπάρχει η εξής γραμμή που προσδιορίζει μια γονιδιωματική περιοχή:

| | | | | | |
|---|---|---|---|---|---|
| chrM | 8366 | 8572 | ATPase8 | 0 | + |

**α)** Τι σημαίνει η συντομογραφία «chrM»;
**β)** Σε ποια θέση του χρωμοσώματος ξεκινάει η γονιδιωματική περιοχή;
**γ)** Σε ποια θέση του χρωμοσώματος τελειώνει η γονιδιωματική περιοχή;
**δ)** Ποιο είναι το μήκος αυτής της γονιδιωματικής περιοχής;

## Μέρος Β (6 μονάδες)

Το BED αρχείο «**human_exons_prCoding_exercise_set.bed**» που σας δίνεται στο eClass περιέχει πληροφορίες για τις θέσεις στο γονιδίωμα των εξωνίων οχτώ ανθρώπινων γονιδίων. Οι τιμές της 4ης στήλης είναι οι κωδικοί του γονιδίου (Ensembl Gene id) και του μεταγράφου (Ensembl Transcript id) χωρισμένοι με το σύμβολο "@".

Σας ζητείται να δημιουργήσετε κώδικα στη γλώσσα Perl (να τον αποθηκεύσετε ως «**first_exon_from_bed.pl**»), που θα διαβάζει το αρχείο «**human_exons_prCoding_exercise_set.bed**» και θα δημιουργεί:

- Ένα νέο BED αρχείο με ονομασία «**first_exons_coordinates.bed»** το οποίο θα περιέχει μόνο τις γραμμές που περιγράφουν το πρώτο εξώνιο κάθε μεταγράφου.
- Ένα νέο text αρχείο με ονομασία «**exonic_length_per_transcript.txt»** που θα έχει 2 στήλες, η 1η στήλη θα είναι το Ensembl Transcript id και η 2η θα είναι το μήκος της εξωνικής περιοχής του μεταγράφου (δηλαδή, το άθροισμα των μηκών των επιμέρους εξωνίων του).

## Σημειώσεις

1. Όταν το γονίδιο βρίσκεται στον '+' κλώνο του γονιδιώματος το πρώτο εξώνιό του είναι το πρώτο από αριστερά, ενώ όταν το γονίδιο βρίσκεται στο '-' κλώνο του γονιδιώματος το πρώτο εξώνιο του είναι το πρώτο από δεξιά.
2. Για τον προσδιορισμό του μήκους πρέπει να λάβετε υπόψη αν η αρίθμηση στα BED αρχεία είναι zero-based ή one-based (Μέρος Α, ερώτηση 2).
3. Το αρχείο «**first_exons_coordinates.bed**» θα έχει την παρακάτω μορφή:

```
chr1    92414928    92415239    ENSG00000137948@ENST00000394530 0    +
chr9    136320770   136320921   ENSG00000160323@ENST00000371910 0    +
chr8    41522323    41522779    ENSG00000029534@ENST00000522231 0    -
chr16   72088522    72088556    ENSG00000257017@ENST00000569639 0    +
chr10   75415576    75415830    ENSG00000166317@ENST00000394810 0    -
```

4. Το αρχείο «**exonic_length_per_transcript.txt**» θα έχει την παρακάτω μορφή:

```
ENST00000511269 1758
ENST00000394530 3189
ENST00000394810 4917
ENST00000369829 906
```

5. Χρήσιμη είναι η εντολή «`exists`» για τον έλεγχο ύπαρξης ενός ζεύγος key:value σε ένα hash (Lecture 5, slide 13).
6. Να συμβουλευτείτε την 2η άσκηση της 5ης διάλεξης, με τη λύση της (Lecture 5, Exercise 2).