ΥΠΟΛΟΓΙΣΤΙΚΗ ΒΙΟΛΟΓΙΑ
1η εργασία

Οδηγίες

Να ανεβάσετε στο eclass, ένα 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** (αρχείο με κατάληξη .pl) με τον κώδικά σας για το Μέρος Β.
iii. το **fasta** αρχείο «yersinia_genes.fasta» που δημιουργήσατε στο Μέρος Β.
iv. το **fasta** αρχείο «yersinia_genome.fasta», 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. Οι DNA/RNA αλληλουχίες σε ένα fasta αρχείο είναι γραμμένες έτσι ώστε:
    a) το 5’ άκρο να βρίσκεται στην αρχή .
    b) το 3’ άκρο να βρίσκεται στην αρχή.
    c) είτε το 5’ άκρο είτε το 3’ άκρο μπορεί να βρίσκεται στην αρχή.
2. Τι συμβολίζουν τα γράμματα ‘N’ και ‘Y’ σε μία αλληλουχία ενός fasta αρχείου; (Εξετάστε την περίπτωση το fasta αρχείο να περιέχει DNA αλληλουχίες και την περίπτωση να περιέχει αλληλουχίες πρωτεϊνών.)
3. Κατά τη μεταγραφή ενός βακτηριακού γονιδίου μεταγράφεται και η αλληλουχία Shine–Dalgarno που διαθέτει ή όχι; Ποιος είναι ο ρόλος αυτής της αλληλουχίας;
4. Στη γλώσσα Perl ποια εντολή και με ποιες προϋποθέσεις μπορούμε να χρησιμοποιήσουμε, αντί για την εντολή ‘print’, ώστε αυτόματα να τυπώνεται στο τέλος η αλλαγή γραμμής (δηλαδή ο ειδικός χαρακτήρας “/n”).

Μέρος Β (6 μονάδες) *(Lecture 4, Exercise 3)*

Δημιουργήστε ένα αρχείο perl που θα διαβάζει το γονιδίωμα του βακτηρίου *Yersinia pestis* από το αρχείο «**yersinia_genome.fasta**» (ελέγξτε το υλικό στο eClass) και θα βρίσκει τις πιθανές θέσεις έναρξης και τέλους των γονιδίων του.

Η **αρχή** κάθε γονιδίου ξεκινάει με το ακόλουθο μοτίβο στην μία από τις δύο αλυσίδες του δίκλωνου DNA:

Υπάρχει μια συναινετική αλληλουχία 8 βάσεων γνωστή ως αλληλουχία Shine-Dalgarno (TAAGGAGG) ακολουθούμενη από 4-10 βάσεις πριν από το κωδικόνιο έναρξης (ATG). Ωστόσο, υπάρχουν παραλλαγές της ακολουθίας Shine-Dalgarno οι οποίες ακολουθούν το μοτίβο: [TA][AC]AGGA[GA][GA].

Διαβάζοντας ανά τριπλέτες, το **τέλος** του γονιδίου προσδιορίζεται από το κωδικόνιο λήξης TAA, TAG ή TGA.

Εκτυπώστε σε ένα αρχείο fasta εξόδου ("**yersinia_genes.fasta**") τις αλληλουχίες γονιδίων που έχετε βρει. Στην κεφαλίδα (header line) κάθε γονιδίου αποθηκεύστε τις εξής πληροφορίες διαχωρισμένες με "|":

- Αύξων αριθμός (π.χ. 1, 2, 3, κ.λπ.)
- "+" ή "-", εάν το γονίδιο βρίσκεται στον αρχικό DNA κλώνο (αυτός που αποτυπώνεται στο αρχείο yersinia_genome.fasta) ή στον αντιστρόφως συμπληρωματικό, αντίστοιχα
- θέση έναρξης (όπου ξεκινά η ακολουθία Shine-Dalgarno)
- θέση τέλους (θέση της τρίτης βάσης του κωδικονίου λήξης)
- μήκος γονιδίου

Η θέση έναρξης και τέλους του γονιδίου θα πρέπει να αναφέρονται στον αρχικό κλώνο (“+” strand) και όχι στον αντιστρόφως συμπληρωματικό (“-“ strand).

Σημειώσεις

Το αρχείο **yersinia_genes.fasta** που θα φτιάξετε θα έχει την παρακάτω μορφή:

```
>185|+|4483927|4483971|45
TCAGGAAAATGGAATCTGATGGATTCTATCTACTGCCATAAGTAA
>186|+|4484478|4485443|966
TCAGGAGAGAAAATGCGATTACGTTTAGCTTTATTTTCGTCATTACTGGCGGCAACCTTT
>187|+|4574172|4574255|84
TCAGGAGATACTGGCTGAATGCGATGTCAACATCGACCACACCACGATTTATCGTTAGGT
>188|-|4595007|4594457|551
TAAGGAGGGAGGGGCTGATGTCTGAATTTGTAACTGTAGCTCGCCCCTACGCCAAAGCAC
>189|-|4493425|4492034|1392
TCAGGAAAAATTATGAGCGCATCTAAGCAAGATGTACAAGATTTTGAGTTTATAATTTCA
>190|-|4468709|4467675|1035
TCAGGAGAGTTTATGGTCACTTTTGAGACAGTTATGGAAATTAAAATCCTGCACAAGCAC
>191|-|4407983|4407864|120
TAAGGAAGCCCCATGGCCACCCGAGTTCCTGCAAGCAATGGGGGGCCATCCCCTGCCCATA
>192|-|4405338|4405139|200
TCAGGAAAACGCTCGGTATGCCCCAAGCCGGCATACAACATCACCGGTGCATTAGCCAGA
>193|-|4303406|4303362|45
TCAGGAAGTAAGTTAATGCTCACATACTCACCAGCCTGCCAATAG
>194|-|4251648|4251620|29
AAAGGAGATATAATATGTTAGAAGAATAA
>195|-|4202795|4202247|549
TAAGGAGAAAACATGCCGATTCCTGAGCGTTTAACACCAGGCAAAGCCACTAAAAATCGC
```

Επεξήγηση της 7^{ης} γραμμής :

- “>”: η γραμμή αυτή είναι κεφαλίδα (header line)
- “188”: αυτό είναι το 188^{ο} γονίδιο που βρέθηκε.
- “-“: η κωδική αλυσίδα του γονιδίου, στην οποία βρέθηκε η αλληλουχία Shine-Dalgarno και το ανοιχτό πλαίσιο ανάγνωσης με τον σωστό προσανατολισμό, είναι ο αντιστρόφως συμπληρωματικός κλώνος.
- “4595007”: σε αυτή τη θέση του γονιδιώματος ξεκινάει η αλληλουχία Shine-Dalgarno
- “4594457”: αυτή είναι η θέση του τρίτου νουκλεοτιδίου του κωδικονίου λήξης
- "551": αυτό είναι το μήκος του γονιδίου (από τη θέση 4595007 έως 4594457). Προσοχή, επειδή η κωδική αλυσίδα αυτού του γονιδίου είναι ο αντιστρόφως συμπληρωματικός κλώνος του γονιδιώματος, η αρχή είναι στα δεξιά και το τέλος του γονιδίου είναι στα αριστερά. Επίσης, το μήκος είναι 551 βάσεις και όχι 550.