

Η Εργασία είναι προαιρετική. Όμως θα μετρήσει σαν μπόνους **στον τελικό βαθμό** παίρνοντας 1 βαθμό σαν μπόνους. Ομάδες μέχρι 3 άτομα (1,2 ή 3) επιτρέπονται.

Εργασία

Στο μάθημα της Πέμπτης 18/3 είδαμε μία προσέγγιση για την εύρεση de-novo* TFBS. Η διαδικασία (περιγράφεται και στο link <https://rpubs.com/idaios/740848>) περιλαμβάνει την χρήση ενός μοτίβου πυρήνα, δηλαδή ενός μοτίβου που υπάρχει αρκετές φορές αυτούσιο, και στην συνέχεια βρίσκουμε και άλλα substrings που απέχουν το πολύ κάποια συγκεκριμένη απόσταση από τον πυρήνα.

Η εργασία ουσιαστικά ζητάει να αναπτύξετε μια μεθοδολογία ευρεσης de novo, TFBS χωρίς την χρήση του πυρήνα. Συγκεκριμένα

1. Αναπτύξτε μια μεθοδολογία για την de-novo ευρεση TFBS χωρίς την χρήση μοτίβου-πυρήνα
2. Συγκρίνετε τα αποτελέσματα μεταξύ της δικής σας μεθοδολογίας και της μεθοδολογίας που αναπτύξαμε στην τάξη με την χρήση πυρήνα
3. Υπάρχουν περιπτώσεις όπου η δική σας μεθοδολογία υπερτερεί;
4. Χρησιμοποιήστε και τα 2 dataset που αναφέρονται παρακάτω
5. Βρείτε μοτίβα μήκος 7 βάσεων

*de-novo: εξ αρχής, χωρίς να πούμε ποιο μοτίβο ψάχνουμε.

Το αρχείο με αλληλουχίες DNA στην τάξη (ευκολο Dataset μαλλον, βρίσκεται εδώ) http://139.91.162.101/teaching/hy390_2021/5utr.fa

Επίσης χρησιμοποιήστε το παρακάτω dataset. Είναι πιο ρεαλιστικό και μάλλον λίγο δυσκολότερο.

http://139.91.162.101/teaching/hy390_2021/proteasome_online.fa

Επίσης όσοι ενδιαφέρεστε για motif discovery μπορείτε να δείτε το PhD Thesis του Jaak Vilo, εδώ

http://www.uretec.com/u/vilo/Publications/Jaak_Vilo_PhD.pdf

Σημείωση: Στα δεδομένα του proteasome δεν έχω τρέξει την διαδικασία ευρεσης μοτίβων. Όμως σε μια παλιότερη μελέτη είχα βρει ότι υπάρχουν μοτίβα (μήκους 7 βάσεων). Οπότε μπορείτε να πειραματιστείτε, αλλά ίσως τα αποτελέσματα να μην είναι και πολύ καλά.