Εργασία Βιοπληροφορικής, εαρινό εξάμηνο ακαδημαϊκού έτους 2023-2024

Ημερομηνία παράδοσης: Ημερομηνία εξέτασης μαθήματος, ώρα 23:59

Η εργασία είναι απαλλακτική και εκπονείται σε ομάδες 1-2 φοιτητών. Μπορεί επίσης να παραδοθεί στην εξεταστική Σεπτεμβρίου.

Παραδοτέα μέσω της ενότητας εργασιών στο e-class:

α) η τεκμηρίωση της εργασίας σε ένα αρχείο pdf, στην πρώτη σελίδα της οποίας αναγράφονται τα ονοματεπώνυμα των φοιτητών και οι ΑΜ. Δεν θα βαθμολογηθούν εργασίες που δεν περιέχουν τεκμηρίωση ή που δεν αναφέρουν τα ονόματα των μελών της ομάδας στην τεκμηρίωση.

β) τα αρχεία source code σε ένα συμπιεσμένο αρχείο με όνομα source2024.zip (ή .rar ή άλλη σχετική κατάληξη).

γ) οποιαδήποτε άλλα συνοδευτικά αρχεία η ομάδα κρίνει απαραίτητα σε ένα συμπιεσμένο αρχείο με το όνομα auxiliary2024.zip (ή .rar ή άλλη σχετική κατάληξη).

Θέμα

i) Δίνονται τα παρακάτω patterns από το αλφάβητο A, C, G, T: pattern1=AATTGA, pattern2=CGCTTAT, pattern3=GGACTCAT και pattern4=TTATTCGTA. Σχεδιάστε και υλοποιήστε μία μέθοδο σύνθεσης συμβολοσειράς, η οποία λειτουργεί ως εξής: α) επιλέγει ένα έως τρία σύμβολα με τυχαίο τρόπο και τα τοποθετεί στην αρχή της συμβολοσειράς. β) Επιλέγει κάθε ένα από τα παραπάνω patterns μία φορά, με τη σειρά που αναγράφονται και αντικαθιστά το πολύ δύο σύμβολα σε τυχαίες θέσεις, είτε με ένα άλλο τυχαία επιλεγμένο σύμβολο (για κάθε θέση ξεχωριστά) είτε με την κενή συμβολοσειρά (διαγραφή συμβόλου). Ότι προκύπτει συνενώνεται με την υφιστάμενη έως εκείνη τη στιγμή συμβολοσειρά. γ) Προσθέτει ένα έως δύο τυχαία σύμβολα στο τέλος της συμβολοσειράς. Δημιουργήστε συνολικά 50 συμβολοσειρές με τον αλγόριθμο σύνθεσης. Διαλέξτε με τυχαίο τρόπο 15 συμβολοσειρές και τοποθετήστε τις σε ένα σύνολο datasetA και τις υπόλοιπες σε ένα σύνολο datasetB.

ii) Σχεδιάστε και υλοποιήστε αλγόριθμο πολλαπλής στοίχισης για τις συμβολοσειρές του συνόλου datasetA. Για τη σύγκριση δύο συμβολοσειρών, υιοθετήστε αλγόριθμο καθολικής στοίχισης ο οποίος ορίζει ότι: η οριζόντια και η κάθετη μετάβαση προσθέτουν gap penalty -α, η τοπική ομοιότητα προσθέτει score +1 και η τοπική ανομοιότητα προσθέτει penalty -α/2. Επίσης, δυνατοί πρόγονοι του κόμβου (i,j) είναι οι (i-1,j-1), (i,j-1), (i-1,j). Εκτυπώστε το αποτέλεσμα της πολλαπλής στοίχισης. Αν όλα τα ΑΜ των μελών της ομάδας καταλήγουν σε περιττό ψηφίο τότε α=1. Σε κάθε άλλη περίπτωση α=2.

iii) Με βάση το αποτέλεσμα της πολλαπλής στοίχισης κατασκευάστε HMM profile και ρυθμίστε τους σχετικούς πίνακες πιθανοτήτων. Υπολογίστε τα alignment scores και alignment paths για τις ακολουθίες του datasetB.

Αποδεκτές γλώσσες υλοποίησης είναι οι Python και Matlab. Κάθε ερώτημα πρέπει να συνοδεύεται από τεκμηρίωση της λύσης.

ΚΑΛΗ ΕΠΙΤΥΧΙΑ