Πρώτο Σύνολο Ασκήσεων 2020-2021

(Βασική βιβλιογραφική πηγή: Dan Gusfield, Algorithms on String Trees and Sequences, κεφ. 1-15.)

Ερώτημα 1

(α) Με δεδομένο μία συμβολοσειρά κειμένου T μήκους n δώστε αλγόριθμο με χρήση δέντρου επιθεμάτων (suffix tree) που μετά από O(n) προεπεξεργασία μπορεί για κάθε δοθείσα συμβολοσειρά P μήκους *m* και μία τιμή k μεταξύ 1 έως n, να τσεκάρει σε O(m) χρόνο αν υπάρχει εμφάνιση του P στο T πριν από τη θέση k.

(β) Σχεδιάστε ένα αποδοτικό αλγόριθμο που βρίσκει τη μικρότερη μη επαναλαμβανόμενη συμβολοσειρά σε ένα κείμενο, δηλαδή τη μικρότερη συμβολοσειρά που εμφανίζεται μόνο μία φορά στο κείμενο.

Υπόδειξη: χρησιμοποιήστε δέντρο επιθεμάτων (suffix tree).

Ερώτημα 2

(α) Σας δίνεται μία συλλογή από k (k>2) ακολουθίες. Επιδείξτε αλγόριθμο ο οποίος εντοπίζει επαναλήψεις (δηλαδή την εμφάνιση της ίδιας συμβολοσειράς δύο φορές) σε κάθε ακολουθία, όπου η συμβολοσειρά που επαναλαμβάνεται είναι η ίδια σε όλες τις ακολουθίες. Εξετάστε τα προβλήματα που ανακύπτουν αν προσπαθήσουμε να βάλουμε περιορισμούς στα κενά ανάμεσα στις δύο εμφανίσεις της συμβολοσειράς σε κάθε ακολουθία.

(β) Έστω σύνολο ***k*** συμβολοσειρών μήκους ***n*** χαρακτήρων η κάθε μία. Προτείνετε έναν αλγόριθμο που να βρίσκει το μέγιστο κοινό πρόθεμα κάθε ζεύγους συμβολοσειρών του συνόλου. Η χρονική πολυπλοκότητα του αλγορίθμου να είναι *O(k*n+α)*, όπου ***α,*** η απάντηση δηλ. το συνολικό πλήθος των μέγιστων επιθεμάτων που υπάρχουν.

Υπόδειξη: χρησιμοποιήστε γενικευμένο δέντρο επιθεμάτων (suffix tree).

Ερώτημα 3

(α) Δείξτε πως μπορούμε να μετρήσουμε τον αριθμό των διακριτών υποσυμβολοσειρών μίας συμβολοσειράς *T* σε O(*n*) χρόνο, όπου το μήκος της συμβολοσειράς *T* είναι *n*. Δείξτε επίσης πως μπορούμε να αναφέρουμε ένα αντίγραφο από κάθε διακριτή υποσυμβολοσειρά σε χρόνο ανάλογο του μήκους όλων των διακριτών συμβολοσειρών.

(β) Έστω ότι έχουμε ένα γενικευμένο δένδρο επιθεμάτων για *k* συμβολοσειρές με τους δείκτες επιθεμάτων (suffix links). Δείξτε πώς μπορεί να μετατραπεί η δομή αυτή στο αντίστοιχο Aho-Corasick αυτόματο των *k* συμβολοσειρών σε γραμμικό χρόνο.

(γ) Θεωρήστε μία συμβολοσειρά *S n* χαρακτήρων και έστω S^r η αναστροφή της. Μελετήστε τη σχέση που υπάρχει ανάμεσα στο δέντρο επιθεμάτων της *S* και στο δέντρο επιθεμάτων της S^r

Υπόδειξη: χρησιμοποιήστε γενικευμένο δέντρο επιθεμάτων (suffix tree).

Ερώτημα 4

Δίνονται οι ακολουθίες v= AAGTACCGGA και w= CCTCGTGAATT. Υποθέστε ότι το κόστος στοίχισης είναι +1 και ότι το κόστος ασυμφωνίας καθώς και το κόστος στοίχισης με κενό είναι -1.

Ι.Υπολογίστε τις τιμές κελιών του πίνακα δυναμικού προγραμματισμού για τον υπολογισμό της ολικής στοίχισης ανάμεσα στις ακολουθίες v και w. Ποια είναι η τιμή της βέλτιστης

ολικής στοίχισης και σε ποια στοίχιση αυτή η τιμή αντιστοιχίζεται; (Για την εύρεση της στοίχισης απαιτείται η σχεδίαση πληροφορίας οπισθοδρόμησης).

II. Υπολογίστε τις τιμές κελιών του πίνακα δυναμικού προγραμματισμού για τον υπολογισμό της τοπικής στοίχισης ανάμεσα στις ακολουθίες v και w. Ποια είναι η τιμή της βέλτιστης τοπικής στοίχισης και σε ποια στοίχιση αυτή η τιμή αντιστοιχίζεται; (Για την εύρεση της στοίχισης απαιτείται η σχεδίαση πληροφορίας οπισθοδρόμησης).
Υπόδειξη: απλή εφαρμογή της θεωρίας.

Ερώτημα 5

Έστω δύο συμβολοσειρές n χαρακτήρων και μία παράμετρος k. Σε κάθε συμβολοσειρά υπάρχουν n-k+1 υποσυμβολοσειρές μήκους k και άρα $\Theta(n^2)$ ζευγάρια τέτοιων συμβολοσειρών, όπου κάθε στοιχείο του ζευγαριού ανήκει σε διαφορετική συμβολοσειρά. Για κάθε τέτοιο ζεύγος σαν σκορ *προσεγγιστικού* ταιριάσματος ορίζεται πόσοι αντίστοιχοι χαρακτήρες ταιριάζουν όταν συγκριθούν οι υποσυμβολοσειρές μήκους k. Θέλουμε να υπολογίζουμε όλες τις τιμές αυτές. Η απλοϊκή προσέγγιση μπορεί να λύσει το πρόβλημα για κάθε δυνατό ζεύγος σε χρόνο $\Theta(kn^2)$. Δώστε αλγόριθμο που να το λύνει σε χρόνο $\Theta(n^2)$.
Υπόδειξη: χρησιμοποιήστε δυναμικό προγραμματισμό.

Ερώτημα 6

Έστω ότι θέλουμε να βρούμε τη βέλτιστη καθολική στοίχιση χρησιμοποιώντας μια μέθοδο βαθμολόγησης με συγγενική ποινή ασυμφωνίας. Δηλαδή το μπόνους για ταίριασμα είναι +1, η ποινή για προσθαφαίρεση είναι -ρ και η ποινή για x συνεχόμενες ασυμφωνίες είναι –(ρ+σx). Διατυπώστε ένα αλγόριθμο O(nm) που στοιχίζει δύο αλληλουχίες μήκους n και m αντίστοιχα με συγγενικη ποινή ασυμφωνίας.
Υπόδειξη: χρησιμοποιήστε δυναμικό προγραμματισμό.

Ερώτημα 7

Υποθέστε πως σε ένα εργαστήριο βιοπληροφορικής σαν ζητούν να εντοπίσετε ομοιότητες ανάμεσα στα γονίδια αντισωμάτων του ανθρώπινου οργανισμού (*homo sapiens*) και σε αυτά του ψαριού-ζέβρα *Danio rerio (zebrafish)* με απώτερο σκοπό την αποτελεσματικότερη αντιμετώπιση αντιγόνων που προσβάλουν τον άνθρωπο. Αφού κατεβάσετε τα αντίστοιχα γονίδια (IG-Heavy chain) από την Βάση Δεδομένων Ανοσολογίας IMGT/LIGM-DB (http://www.imgt.org/ligmdb/), αναπτύξτε πρόγραμμα το οποίο θα δέχεται ως είσοδο τα δύο αυτά γονίδια, και θα εντοπίζει την μέγιστη κοινή υποσυμβολοσειρά τους. Θα πρέπει να αναφέρετε την υποσυμβολοσειρά καθώς επίσης και όλες τις εμφανίσεις της. Δώστε κώδικα ή περιγράψτε (στο χαρτί) με ένα παράδειγμα από τα γονίδια των βάσεων το τελικό αποτέλεσμα.
Υπόδειξη: χρησιμοποιήστε δέντρο επιθεμάτων (suffix tree).

Ερώτημα 8

(α) Έστω ότι λαμβάνετε στην κατοχή σας την ακόλουθη ακολουθία :
> unknown sequence
ggggtcaaagcaatcccagtgagcatccacgtcaatgtcattctcttctctgccatccttat
tgtgttaaccatggtggggacagccttcttcatgtacaatgcttttggaaaaccttttga
aactctgcatggtcccctagggctgtacctttgagcttcatttcaggctcctgtggctg
tcttgtcatgatattgtttgcctctgaagtgaaaatccatcacctctcagaaaaaattga
aattataaagaagggacttatgtctacaaaacgcaaagtgaaaaatataccacctcattc

Χρησιμοποιήστε την βάση της NCBI για να βρείτε λεπτομέρειες για αυτή την ακολουθία. Χρησιμοποιήστε κάποιο εργαλείο Nucleotide BLAST για να βρείτε ακολουθίες που περιέχουν το τμήμα. (https://blast.ncbi.nlm.nih.gov/Blast.cgi) Προσδιορίστε τον οργανισμό και το όνομα του γονιδίου που ***πιθανόν*** η ακολουθία προέρχεται (το πρώτο στη λίστα απάντησης). Ποιο εργαλείο χρησιμοποιήσατε για την απάντηση; Ποιο είναι το accession number της GenBank για την ακολουθία που ταιριάζει καλύτερα στην ακολουθία ερώτημα.

Με το accession number της GenBank ανασύρετε την εγγραφή για την ακολουθία αυτή και όσες πληροφορίες μπορείτε

- Ποια είναι παλαιότερη αναφορά στο γονίδιο αυτό
- Σε ποια θέση (locus) βρίσκεται το γονίδιο αυτό
- Πόσες παράλληλες εμφανίσεις του γονιδίου αυτού έχουν αναγνωριστεί και σε ποια χρωμοσώματα
- Ποιο το όνομα της πρωτεΐνης που παράγεται από αυτό το γονίδιο. Ποιος ο κωδικός της Genbank για την πρωτεΐνη αυτή;
- Σε ποια θέση της δεδομένης ακολουθίας αρχίσει η μετάφραση ;
- Πόσα αμινοξέα κωδικοποιούνται από το γονίδιο;
- Ποια είναι η λειτουργία της πρωτεΐνης που κωδικοποιείται από το γονίδιο

Στην NCBI κάντε μία νέα αναζήτηση (translated blast search - blastx) χρησιμοποιώντας την άγνωστη ακολουθία. Ποια συμβολοσειρά ανασύρεται αυτή τη φορά (accession number) Ταυτίζεται με την πρώτη αναζήτηση όσον αφορά την πρωτεΐνη της πρώτης αναζήτηση; Αν όχι, μελετήστε τη διαφορά των δύο πρωτεινών

(β) Επισκεφθείτε την σελίδα www.ncbi.nlm.nih.gov. Επιλέξτε **'TaxBrowser'.** Η ταξονομία του NCBI περιέχει οργανισμούς που έχουν κατατεθεί οι ακολουθίες τους. Δείτε την επιλογή **'Taxonomy Statistics'**.

- Πόσα Είδη Βακτηρίων υπάρχουν στην Βάση δεδομένων;
- Για το έτος 1999, πόσα νέα είδη βακτηρίων προστέθηκαν;
  - Επιλέξτε **'Taxonomy Statistics' ->' Extinct organisms'**
  - Στα Έντομα επιλέξτε **'*Libanorhinus succinus*'.**
- Ποιοι άλλοι οργανισμού ανήκουν σε αυτή τη κατηγορία;
- Πόσες ακολουθίες έχουν εισαχθεί από αυτόν τον οργανισμό;
- Ποιο το όνομα του γονιδίου που έχει αλληλουχηθεί από αυτόν τον οργανισμό;
- Πόσα ζευγάρια βάσεων έχει αυτή η DNA εγγραφή;

Επισκεφθείτε το **PUBMED '8505967'** για να ανασύρετε πληροφορίες για την δημοσίευση