ΠΡΩΤΗ ΕΡΓΑΣΙΑ 2024-205

1. Βασική βιβλιογραφική πηγή:

Dan Gusfield, Algorithms on String Trees and Sequences, Cambridge University Press, 1997

Βιοπληροφορική και Λειτουργική Γονιδιωματική, Jonathan Pevsner, ΑΚΑΔΗΜΑΪΚΕΣ ΕΚΔΟΣΕΙΣ Ι. ΜΠΑΣΔΡΑ & ΣΙΑ Ο.Ε., 1η/2019

Εισαγωγή στους Αλγοριθμους Βιοπληροφορικής, Neil C. Jones, Pavel Pevzner, Εκδόσεις Κλειδάριθμος, 2008

Biological Modeling: A Short Tour, January 25, 2023, by Phillip Compeau (Author, Editor), Mert Inan (Author), Noah Lee (Author), Shuanger Li (Author), Chris Lee (Author). https://biologicalmodeling.org/

https://rosalind.info/problems/locations/

2. Για την υλοποίηση (όπου απαιτείται) μπορείτε να χρησιμοποιήσετε όποια γλώσσα προτιμάτε. Για την τελευταία άσκηση είναι υποχρεωτική η επιλογή της Biopython (https://biopython.org/, https://en.wikipedia.org/wiki/Biopython)

3. Τα ερωτήματα χωρίς βαθμολογική συνεισφορά (υποερώτημα iii και iv στην άσκηση 3) δεν προσμετρούνται στην αξιολόγηση είναι απλά για ενασχόληση.

4. Στο ερώτημα 4 αρκεί μια διερεύνηση των σχετικών papers και τεχνικών πλήθους 1-2 σελίδες

Ερώτημα 1

(i) Επισκεφτείτε τις ακόλουθες σελίδες σύγχρονων εργαλείων πρόβλεψης πρωτεϊνικών δομών: **Alphafold3** (https://alphafoldserver.com, https://github.com/sokrypton/ColabFold, https://www.nature.com/articles/s41586-024-07487-w, και **Esmfold2** (https://github.com/facebookresearch/esm, https://www.science.org/doi/10.1126/science.ade2574, https://www.biorxiv.org/content/10.1101/2022.07.20.500902v1.full.pdf) και περιγράψτε με λίγα λόγια την βασική τους λειτουργικότητα.

Υπόδειξη: Αρκεί σαν αποτέλεσμα Η ΑΝΑΦΟΡΑ (1-3 σελίδες) ΑΠΛΗΣ ΧΡΗΣΗΣ ΤΩΝ ΔΙΑΦΟΡΩΝ ΕΡΓΑΛΕΙΩΝ, ΟΧΙ Η ΕΠΙΛΥΣΗ ΚΑΘΕ ΠΡΟΒΛΗΜΑΤΟΣ. Δηλαδή σκοπός της άσκησης είναι να έλθετε σε επαφή με κάποια έτοιμα εργαλεία ΟΧΙ Η ΕΜΠΕΙΡΗ ΧΡΗΣΗ ΑΥΤΩΝ.

(ii)[1]Πραγματοποιήστε μία αναζήτηση BLASTP (protein-to-protein) από την βάση δεδομένων NCBI (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins) χρησιμοποιώντας για παράδειγμα την ελαφριά αλυσίδα της φεριττίνης (NP_000137). Από τον κατάλογο αποτελεσμάτων επιλέξετε 10 πρωτείνες κάνοντας κλικ στο πλαίσιο δίπλα από την καθεμία και στη συνέχεια αποθηκεύστε τις με μορφή FASTA.

(iii)[1] Χρησιμοποιώντας τις αλληλουχίες σε μορφή FASTA που ανακτήσατε, κάντε τις μεταξύ τους στοιχίσεις χρησιμοποιώντας προγράμματα **πολλαπλών στοιχίσεων** από το EBI (https://www.ebi.ac.uk/Tools/msa/) όπως τα MAFFT, MUSCLE και T-COFFEE. Αποθηκεύστε και συγκρίνετε τα αποτελέσματά τους. Υπάρχουν διαφορές; Πως μπορείτε να αξιολογήσετε ποια στοίχιση είναι η πλέον ακριβής; Δοκιμάστε διαφορετικές επιλογές στις παραμέτρους που δίνονται από τα εργαλεία.

Ερώτημα 2[2]

Το πρόβλημα της μέγιστης κοινής υποακολουθίας δύο συμβολοσειρών S_1, S_2, μπορεί να θεωρηθεί σαν μία ειδική περίπτωση του προβλήματος της εύρεσης της απόστασης μετασχηματισμού με βάρη, των δύο συμβολοσειρών. Πιο συγκεκριμένα έστω u το μήκος της μεγαλύτερης κοινής υποακολουθίας δύο συμβολοσειρών με μήκος n και m. Χρησιμοποιώντας τα βάρη πράξεων $d=1$, $r=2$ και $e=0$ (d είναι το βάρος ένθεσης/διαγραφής, r είναι το βάρος αντικατάστασης, και e είναι το βάρος ταιριάσματος) υποστηρίζουμε ότι $D(n,m)=m+n-2u$ ή $u=(m+n-D(n,m))/2$. Συνεπώς η ποσότητα $D(n,m)$ ελαχιστοποιείται όταν μεγιστοποιείται το u. Αποδείξτε την ορθότητα ή όχι της προαναφερόμενης υπόθεσης και εξηγήστε με λεπτομέρεια πως μπορείτε να εντοπίσετε τη μεγαλύτερη κοινή υποακολουθία μεταξύ δύο ακολουθιών S1, και S2. Υλοποιήστε τον αλγόριθμο σε γλώσσα προγραμματισμού της επιλογής σας (κατά προτίμηση python).

Ερώτημα 3

(i) Προσπελάστε τη βάση δεδομένων NCBI για να μελετήσετε τον κορονοϊό SARS-CoV-2 στο σύνδεσμο https://www.ncbi.nlm.nih.gov/sars-cov-2/. Χρησιμοποιήστε την εγγραφή με δεδομένα ακολουθίας για τον SARS-CoV-2

---

[1] Βιοπληροφορική και Λειτουργική Γονιδιωματική, Jonathan Pevsner, ΑΚΑΔΗΜΑΪΚΕΣ ΕΚΔΟΣΕΙΣ Ι. ΜΠΑΣΔΡΑ & ΣΙΑ Ο.Ε., 1η/2019

[2] Dan Gusfield, Algorithms on Strings, Trees and Sequences, Cambridge University Press

https://www.ncbi.nlm.nih.gov/nuccore/NC_045512 για να κατεβάσετε την ακολουθία της spike (ακίδα) πρωτείνης του κορονοϊού. Στη συνέχεια από το σύνδεσμο https://www.uniprot.org/uniprotkb/A0A023SFE5/entry κατεβάστε την ακολουθία της spike (ακίδα) πρωτεϊνης για τον κορονοϊό **Middle East respiratory syndrome-related coronavirus (MERS-CoV)** (https://www.who.int/news-room/fact-sheets/detail/middle-east-respiratory-syndrome-coronavirus-(mers-cov)) **και υλοποιήστε τον αλγόριθμο του ερωτήματος 2 με κατάλληλα βάρη ώστε να εντοπίσετε την μέγιστη κοινή υποακολουθία τους.** Αναφέρετε το τελικό αποτέλεσμα.

(ii) Δείτε τη δομή των δύο πρωτεϊνών του προηγούμενου ερωτήματος χρησιμοποιώντας το ab-initio εργαλείο swiss-modeller (https://swissmodel.expasy.org/interactive) και κατεβάστε τα αρχεία .pdb[3]. Στη συνέχεια συγκρίνετε τις δομές των δύο πρωτεινών χρησιμοποιώντας το εργαλείο Dali στην ηλεκτρονική διεύθυνση http://ekhidna.biocenter.helsinki.fi/dali/. Κάντε τις παρατηρήσεις σας σχετικά με την συσχέτιση ακολουθιών και δομών.

(iii) **(υποερώτημα χωρίς βαθμολογική συνεισφορά)**: προσπαθήστε να επιλύσετε το πρόβλημα πρόβλεψης πρωτεϊνικής δομής αξιοποιώντας τους αλγορίθμους **Alphafold3** και **Esmfold2** (υποερώτημα 1.i)

(iv) (υποερώτημα χωρίς βαθμολογική συνεισφορά): αν κάποιος θέλει να εμβαθύνει περισσότερο, μπορεί να επισκεφτεί το διαδικτυακό τόπο https://biologicalmodeling.org/coronavirus/home με παραπλήσια (όχι όμως ίδια) ερωτήματα και να μελετήσεις τις εκεί επιλύσεις

Ερώτημα 4 (ΕΠΙΛΕΓΕΤΕ ΜΟΝΟ ΜΙΑ ΑΠΟ ΔΥΟ ΕΠΙΛΟΓΕΣ για να φτάσετε στο άριστα, αν επιλέξετε και τις δύο θα έχετε επιπλέον βαθμολογικό bonus)

Πρώτη επιλογή (ερευνητική-προβληματισμός -διερεύνηση λύσεων στη βιβλιογραφία)

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.

Υπόδειξη: χρήση γενικευμένου δένδρου επιθεμάτων και επέκταση αντίστοιχης άσκησης που θα κάνουμε στο μάθημα θεωρίας για μία συμβολοσειρά. Διερευνήστε σχετικές λύσεις στο διαδίκτυο , την χρήση τεχνικών με k-mers, και την βιβλιογραφία που ακολουθεί. **Επί της ουσίας αρκεί μια διερεύνηση των σχετικών papers και παρουσίαση τεχνικών σε 1-2 σελίδες.**

Σχετική Βιβλιογραφία:

- ✓ Gerth Stølting Brodal, Rune B. Lyngsø, Christian N. S. Pedersen, Jens Stoye: Finding Maximal Pairs with Bounded Gap. CPM 1999: 134-149
- ✓ Gerth Stølting Brodal, Christian N. S. Pedersen: Finding Maximal Quasiperiodicities in Strings. CPM 2000: 397-411
- ✓ A. Bakalis, Costas S. Iliopoulos, Christos Makris, Spyros Sioutas, Evangelos

[3] Το **.pdb (Protein Data Bank) format** είναι ένα πρότυπη μορφή αρχείου που χρησιμοποιείται για την αποθήκευση **τρισδιάστατων δομών βιολογικών μακρομορίων**, όπως είναι οι πρωτεΐνες. Το αρχείο αυτό περιέχει πληροφορίες σχετικά με τη χωρική διάταξη των ατόμων μέσα στο μόριο, καθώς και πρόσθετες λεπτομέρειες όπως δεσμούς, αλληλεπιδράσεις και σχολιασμούς.

Theodoridis, Athanasios K. Tsakalidis, Kostas Tsichlas: Locating Maximal Multirepeats in Multiple Strings Under Various Constraints. Comput. J. 50(2): 178-185 (2007).

Δεύτερη επιλογή [4]

(α) 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.

(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 $O(k*n+\alpha)$, όπου ***α,*** η απάντηση δηλ. το συνολικό πλήθος των μέγιστων επιθεμάτων που υπάρχουν.

Υπόδειξη: χρησιμοποιήστε γενικευμένο δέντρο επιθεμάτων (suffix tree).

Ερώτημα 5 (απλή εφαρμογή θεωρίας, χωρίς κώδικα)

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

I. 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 6. Ασκήσεις με χρήση BioPython

Οι παρακάτω δύο ασκήσεις απαιτούν απλή χρήση της **Biopython** (θα εξηγηθεί αναλυτικά στο φροντιστήριο πως θα χρησιμοποιηθεί). Πιο συγκεκριμένα ως γλώσσα υλοποίησης προτείνεται η python με έμφαση στην βιβλιοθήκη της BioPython και τα εργαλεία BLASTP και Clustal Omega.

Άσκηση α. Υπάρχουν περίπου 20000 γονίδια του ανθρώπου τα οποία εκφράζονται για να συντεθούν οι πρωτείνες. Να επιλέξετε μια πρωτεΐνη από αυτές (λ.χ. αξιοποιώντας την βάση δεδομένων UniProt (https://www.uniprot.org/uniprotkb?query=human) ή χρησιμοποιώντας τις spike πρωτείνες του ερωτήματος 2) και χρησιμοποιώντας το εργαλείο BLASTP (αναπτύσσοντας κώδικα σε Biopython) να βρείτε τουλάχιστον 8 ομόλογες ακολουθίες από άλλους οργανισμούς. Να δημιουργήσετε στοιχίσεις μεταξύ των ακολουθιών αυτών τόσο σε ζεύγη όσο και μία πολλαπλή. Τέλος, αποθηκεύσετε τις στοιχίσεις σε αρχεία με κατάληξή .aln[5].

Άσκηση β. Σκοπός αυτής της άσκησης είναι να χρησιμοποιήσετε το **BLASTP**, για να αναγνωρίσετε μια άγνωστη πρωτεΐνη, αξιοποιώντας την **Biopython**. Θα αναζητήσετε ομόλογες

[4] Neil Jones, Pavel Pevzner, An Introduction to Bioinformatics Algorithms, The MIT Press

[5] https://web.mit.edu/meme_v4.11.4/share/doc/clustalw-format.html

αλληλουχίες στη βάση δεδομένων **NR** και θα βρείτε την πηγή της άγνωστής πρωτεΐνης. Επιπλέον, θα αναζητήσετε τη σχετική επιστημονική δημοσίευση που περιγράφει αυτήν την πρωτεΐνη και θα συνοψίσετε τα βασικά ευρήματά της.

ΑΓΝΩΣΤΗ ΠΡΩΤΕΙΝΗ:

>mystery_sequence
GATGAPGIAGAPGFPGARGAPGPQGPSGAPGPKXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXGVQGPPGPQGPR
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXGSAGPPGATGFP
GAAGRXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXGVVGLPGQR

Το γράμμα **"X"** αντιστοιχεί σε άγνωστα αμινοξέα, όμως γνωρίζετε ότι το όνομα του οργανισμού που ανήκει η άγνωση πρωτείνη ξεκινά με **“Tyr”.**