ΠΡΩΤΗ ΕΡΓΑΣΙΑ 2025-2026

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. Συνολικά προτείνεται η επιλογή της Biopython (https://biopython.org/, https://en.wikipedia.org/wiki/Biopython)

3. Τα ερωτήματα χωρίς βαθμολογική συνεισφορά (υποερωτήματα (iii) και (iv) στην 4) δεν προσμετρούνται στην αξιολόγηση είναι απλά για ενασχόληση.

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 **Esmfold** (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 **Hemoglobin subunit beta Accession (RefSeq): NP_000509** *(ή άλλη πρωτεΐνη της επιλογής σας)*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

Στόχος είναι η μελέτη της διατήρησης μιας πρωτεΐνης σε διαφορετικά είδη θηλαστικών. Επιλέξτε την ανθρώπινη πρωτεΐνη Cytochrome C αφού εντοπίσετε το Accession Number της στη βάση Uniprot και παραθέστε το link και κάποιο screenshot. (Απ: Accession Number = P99999). Πραγματοποιήστε αναζήτηση για ομόλογες ακολουθίες εντός της κλάσης των θηλαστικών (Mammalia) χρησιμοποιώντας το εργαλείο BLASTP. Πραγματοποιήστε στοίχιση, ΟΧΙ κατ ανάγκην πολλαπλή, των ακολουθιών που βρήκατε μαζί με την ανθρώπινη. Τέλος, να υπολογίστε το ποσοστό ταυτότητας (identity percentage) των ομόλογων ακολουθιών σε σχέση με την ανθρώπινη.

Υποδείξεις

- Βήμα 1 (Fetching): Χρησιμοποιήστε τη βιβλιοθήκη Entrez της Biopython ώστε να σταλθεί ένα αίτημα στο NCBI που θα επιστρέψει την ακολουθία σε μορφή FASTA, χρησιμοποιώντας το μοναδικό ID της πρωτεΐνης.
- Βήμα 2 (BLAST Query): Χρησιμοποιήστε τη συνάρτηση qblast της Biopython. Πρέπει να οριστεί ο σωστός αλγόριθμος (blastp), η βάση (nr) και –πολύ σημαντικό– να το φίλτρο entrez_query (βλ. https://biopython.org/docs/latest/Tutorial/chapter_entrez.html) ώστε να περιοριστεί η αναζήτηση στα θηλαστικά.

---

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

- Βήμα 3 (Data Collection): Το BLAST θα επιστρέψει ένα αντικείμενο με πολλά "hits". Επαναληπτικά, για κάθε hit, θα πρέπει να διατηρηθεί η κατάλληλή πληροφορία (Accession ID) και να ξαναγίνει ένα αίτημα στο Entrez ώστε να ληφθεί η κατάλληλη ακολουθία/ιες.
- Βήμα 4 (Alignment):Το εργαλέιο Clustal Omega μπορεί να χρησιμοποιηθεί για πολλαπλή στοίχιση. ΟΜΩΣ, αυτή τη στιγμή δεν είναι διαθέσιμο στο site τους. Άρα για το λόγο αυτό στα πλαίσια της άσκησης αρκεί να γίνει στοίχιση μεταξύ της ακολουθίας ενδιαφέροντος και αυτών που επέστρεψε το BLAST. Η biopython υποστηρίζει τέτοια pairwise στοίχιση με τη χρήση του PairwiseAligner. Σε περίπτωση που θέλετε να χρησιμοποιήσετε πολλαπλή στοίχιση μπορείτε να χρησιμοποιήσετε το MUSCLE σε συνδυασμό με biopython (βλ. https://biopython.org/docs/dev/Tutorial/chapter_msa.html)
- Βήμα 5 (Identity Calculation): Για κάθε ακολουθία, να δημιουργηθεί μια συνάρτηση που συγκρίνει το κάθε γράμμα της με το αντίστοιχο γράμμα της ανθρώπινης ακολουθίας (στη θέση i) και μετράει πόσες φορές ταυτίζονται.

Ερωτήσεις

- Στο βήμα 2 για ποιους λόγους χρησιμοποιήθηκε ο αλγόριθμος blastp η βάση nr και για πιο λόγο θέλουμε να περιορίσουμε τη σύγκριση στα θηλαστικά; Τι αναμένετε να συμβεί αν δεν περιοριστεί η αναζήτηση;
- Αν κατά την αναζήτηση BLAST θέταμε πολύ "χαλαρά" κριτήρια (π.χ. πολύ υψηλό E-value), ποιος θα ήταν ο κίνδυνος για τα αποτελέσματα της στοίχισης;
- Υπάρχουν περιοχές στην πρωτεΐνη που είναι απόλυτα συντηρημένες (100% ταυτότητα) σε όλα τα αποτελέσματα της αναζήτησης ;
- ποια είδη παρουσιάζουν μεγαλύτερη απόκλιση από τον άνθρωπο;

Ερώτημα 3

Σας δίνεται μια σύντομη αλληλουχία αμινοξέων που αντιστοιχεί σε ένα τμήμα μιας πολύ γνωστής ανθρώπινης πρωτεΐνης.

1. Χρησιμοποιήστε το BLASTP για να ταυτοποιήσετε ποια είναι η πρωτεΐνη και σε ποιον οργανισμό ανήκει.
2. Επιβεβαιώστε τη λειτουργία της πρωτεΐνης μέσω μιας σύντομης βιβλιογραφικής αναζήτησης (PubMed/NCBI).

Αλληλουχία (Input): GIVEQCCTSICSLYQLENYCN

Υποδείξεις

- Βήμα 1 (Input): Ορίστε την παραπάνω αλληλουχία σε μια μεταβλητή Python.
- Βήμα 2 (BLAST): Χρησιμοποιήστε τη συνάρτηση qblast. Επειδή η αλληλουχία είναι μικρή, το BLAST μπορεί να σου επιστρέψει πολλά αποτελέσματα. Δώσε βάση στο πρώτο hit (αυτό με το χαμηλότερο E-value).
- Βήμα 3 (Ανάγνωση αποτελεσμάτων): Διαβάστε το alignment.title. Θα δεις το όνομα της πρωτεΐνης (π.χ. Insulin) και τον οργανισμό (π.χ. Homo sapiens).
- Βήμα 4 (Επιβεβαίωση): Μόλις πάρετε το όνομα, ψάξετε στο Google ή στο NCBI "Function of [Όνομα Πρωτεΐνης]" για να επιβεβαιώσετέ ότι αυτό που βρήκε το BLAST ταιριάζει με τη βιολογική γνώση.

Ερωτήσεις

- Ποιο είναι το όνομα της πρωτεΐνης που βρήκε το BLAST και σε ποιον οργανισμό ανήκει; (Συμφωνεί με την υπόθεση ότι είναι "ανθρώπινη";)
- Αν το E-value του πρώτου αποτελέσματος είναι εξαιρετικά χαμηλό (κοντά στο μηδέν). Τι σημαίνει αυτό για την αξιοπιστία της απάντησης; Ποιο ήταν το E-Value της αναζήτησης που κάνατε;
- Με βάση μια απλή αναζήτηση, ποιος είναι ο κύριος ρόλος αυτής της πρωτεΐνης στο ανθρώπινο σώμα;
- Αν τρέχατε το ίδιο BLAST με την αλληλουχία αυτή, αλλά περιόριζες την αναζήτηση μόνο σε "Bacteria" (μέσω entrez_query), τι αναμένετε ως αποτέλεσμα; Θα βρεθεί η ανθρώπινη ινσουλίνη ή κάτι άλλο; Μπορείτε να δοκιμάσετε να κάνετε την αναζήτηση αυτή

Ερώτημα 4

(i) Προσπελάστε τη βάση δεδομένων NCBI για να μελετήσετε τον κορονοϊό **Severe acute respiratory syndrome coronavirus 2 (**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 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/P59594/entry κατεβάστε την ακολουθία της spike (ακίδα) πρωτεΐνης για τον σχετιζόμενο κορονοϊό **Severe acute respiratory syndrome coronavirus (SARS-CoV) και** χρησιμοποιήστε το εργαλείο **EMBOSS Needle** του EBI (https://www.ebi.ac.uk/jdispatcher/psa/emboss_needle) για να συγκρίνετε τις δύο ακολουθίες μεταξύ τους.

(ii) Δείτε τη δομή των δύο πρωτεινών του προηγούμενου ερωτήματος χρησιμοποιώντας το ab-initio εργαλείο swiss-modeller (https://swissmodel.expasy.org/interactive) και κατεβάστε τα αρχεία .pdb[2]. Στη συνέχεια συγκρίνετε τις δομές των δύο πρωτεινών χρησιμοποιώντας το εργαλείο RCSB στην ηλεκτρονική διεύθυνση https://www.rcsb.org/alignment ακολουθιών και δομών.

(iii) **(υποερώτημα χωρίς βαθμολογική συνεισφορά)**: προσπαθήστε να επιλύσετε το πρόβλημα πρόβλεψης πρωτεϊνικής δομής αξιοποιώντας τους αλγορίθμους **Alphafold3** και **Esmfold** (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 μελετήσει τις εκεί επιλύσεις.

Ερώτημα 5 (απλή εφαρμογή θεωρίας, χωρίς κώδικα)

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

---

[2] 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.

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

(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) Διερευνήστε **(σε το πολύ μία σελίδα)** τεχνικές συνδυασμού Bloom Filters, Suffix Arrays και Burrows Wheeler για δεικτοδότηση βιολογικών δεδομένων με σημείο εστίασης το άρθρο:
Ondřej Sladký, Pavel Veselý, Karel Břinda, From Superstring to Indexing: a space-efficient index for unconstrained k-mer sets using the Masked Burrows-Wheeler Transform (MBWT), Bioinform Adv. 2025 Nov 12;6(1):vbaf290. doi: 10.1093/bioadv/vbaf290