Πρώτο Σύνολο Ασκήσεων 2023-2024

1. Βασική βιβλιογραφική πηγή:

Dan Gusfield, Algorithms on String Trees and Sequences, κεφ. 1-15.

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. Μελετήστε και την επιλογή της Biopython (https://biopython.org/, https://en.wikipedia.org/wiki/Biopython)

3. Τα ερωτήματα χωρίς βαθμολογική συνεισφορά (υποερώτημα iii και iv στην άσκηση 2 και μη υποχρεωτικό ερώτημα στην άσκηση 7) δεν προσμετρούνται στην αξιολόγηση είναι απλά για ενασχόληση.

4. Στα ερωτήματα 3,4 αρκεί μια διερεύνηση των σχετικών papers και τεχνικών μήκους από 1-2 σελίδες, ενώ στα ερωτήματα 1,2 διερεύνηση των σχετικών εργαλείων από 1-3 σελίδες.

Ερώτημα 1

(i) Αντικείμενο της συγκεκριμένης άσκησης είναι η διερεύνηση έτοιμων εργαλείων λογισμικού σχετικά με χειρισμό προβλημάτων βιοπληροφορικής. Πιο συγκεκριμένα θα πρέπει να δοκιμάσετε **(όχι επίλυση)** τα παραδείγματα χρήσης εργαλείων λογισμικού που καταγράφονται στη σελίδα https://rosalind.info/problems/list-view/?location=bioinformatics-armory της Rosalind (https://rosalind.info/problems/locations/) και για καθένα από αυτά να κάνετε μία μικρή αναφορά χρήσης και εμπειρίας με διάφορα δεδομένα.

(ii) Υπάρχουν πολλά ελεύθερα προσβάσιμα εργαλεία για την πολλαπλή στοίχιση αλληλουχιών. Στο ερώτημα αυτό θα κάνετε μία (αρκεί και η θεωρητική) σύγκριση των εργαλείων στις Βάσεις Δεδομένων NCBI και EBI. Επισκεφθείτε τον ιστότοπο του NCBI και του EBI και αναφέρετε τα βασικά χαρακτηριστικά των εργαλείων πολλαπλή στοίχισης που προσφέρουν. Για το NCBI τα βασικά εργαλεία είναι στα links: https://www.ncbi.nlm.nih.gov/projects/msaviewer/, https://www.ncbi.nlm.nih.gov/tools/cobalt/re_cobalt.cgi και για το EBI Sequence Manipulation Suite στην ιστοσελίδα https://www.ebi.ac.uk/jdispatcher/msa/ο οποίος εξασφαλίζει πρόσβαση σε ένα μεγάλο αριθμό εργαλείων.

Υπόδειξη: Αρκεί σαν αποτέλεσμα Η ΑΝΑΦΟΡΑ 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.

Ερώτημα 2[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 την ελαφριά αλυσίδα της φεριττίνης (NP_000137). Από τον κατάλογο αποτελεσμάτων επιλέξετε 3 πρωτείνες κάνοντας κλικ στο πλαίσιο δίπλα από την καθεμία και στη συνέχεια αποθηκεύστε τις με μορφή FASTA.

(i) Στοιχίστε τις ακολουθίες χρησιμοποιώντας το πρόγραμμα T-COFFEE ή οποιοδήποτε πρόγραμμα από το ερώτημα 1, και αξιολογήστε την στοίχιση με το πρόγραμμα TCS (http://www.tcoffee.org). Σημειώστε την βαθμολογία.

(ii) Δείτε τη δομή των τριών πρωτεϊνών του προηγούμενου ερωτήματος χρησιμοποιώντας το εργαλείο swiss-modeller (https://swissmodel.expasy.org/interactive) και κατεβάστε τα αρχεία .pdb (η μορφή αρχείου που χρησιμοποιείται από την Protein Data Bank). Στη συνέχεια συγκρίνετε τις δομές των 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)

Υποερωτημα (iii) (υποερώτημα χωρίς βαθμολογική συνεισφορά): προσπαθήστε να επιλύσετε το πρόβλημα πρόβλεψης πρωτεϊνικής δομής με διάφορους νέους αλγόριθμους μηχανικής μάθησης (https://www.nature.com/articles/s41592-023-01790-6) όπως **AlphaFold** (https://alphafold.ebi.ac.uk, https://www.ebi.ac.uk/Tools/sss/fasta/, https://colab.research.google.com/github/deepmind/alphafold/blob/main/notebooks/AlphaFold.ipynb,

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

https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb#scrollTo=G4yBrceuFbf3) και **ESMFold** (https://www.science.org/doi/10.1126/science.ade2574, https://esmatlas.com/resources?action=fold, https://github.com/facebookresearch/esm).
Υποερώτημα (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.

Ερώτημα 3 (ερευνητική-προβληματισμός)
Έστω γενικευμένο δένδρο επιθεμάτων σε ένα σύνολο κ συμβολοσειρών. Συμβολοσειρές είναι δυνατό να προστίθενται ή να αφαιρούνται από το σύνολο αυτό. Περιγράψτε τα προβλήματα που ανακύπτουν για την δυναμική διατήρηση αυτής της δομής. Δώστε αλγόριθμους που να υποστηρίζουν τις πράξεις αυτές.
Υπόδειξη: ψάξτε στο παγκόσμιο ιστό (με μηχανές αναζήτησης ή με εργαλεία ψηφιακών βιβλιοθηκών) με το ερώτημα dynamic suffix trees/dynamic string matching και συνθέστε τα αναφερόμενα στα εκεί αποτελέσματα.

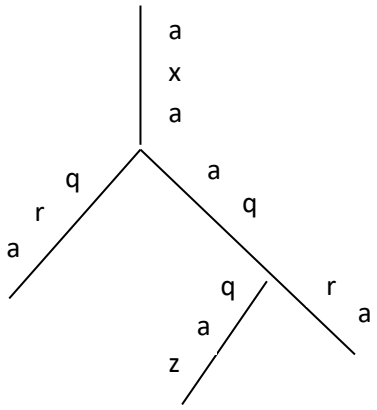
Ερώτημα 4 (ερευνητική-προβληματισμός)
Σας δίνεται μία συλλογή από k (k>2) ακολουθίες. Επιδείξτε αλγόριθμο ο οποίος εντοπίζει επαναλήψεις (δηλαδή την εμφάνιση της ίδιας συμβολοσειράς δύο φορές) σε κάθε ακολουθία, όπου η συμβολοσειρά που επαναλαμβάνεται είναι η *ίδια* σε όλες τις ακολουθίες . Εξετάστε 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 επίσης σχετικές λύσεις στο διαδίκτυο και την χρήση τεχνικών με k-mers.

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

Ερώτημα 5[2]
Έστω δένδρο όπου κάθε πλευρά έχει ετικέτα με έναν ή περισσότερους χαρακτήρες και ένα πρότυπο P. Δώστε αλγόριθμο που να εντοπίζει όλα τα υπομονοπάτια που ξεκινούν από την ρίζα και περιέχουν το πρότυπο. Παρατηρήστε ότι αν και το υπομονοπάτι πρέπει να είναι τμήμα μονοπατιού που ξεκινά από τη ρίζα, το ίδιο το υπομονοπάτι δεν χρειάζεται να ξεκινά από τη ρίζα (κοιτάξτε και σχολιασμό σχήματος). Δώστε αλγόριθμο για το συγκεκριμένο πρόβλημα που τρέχει σε χρόνο αναλογο με το συνολικό αριθμό των χαρακτήρων στις ακμές του δέντρου συν το μήκος του πρότυπου P.

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



Σχήμα*: το πρότυπο P=aqra, υπάρχει σε δύο υπομονοπάτια, που αρχίζουν από τη ρίζα. Αυτά τα μονοπάτια αρχίζουν από τη ρίζα, αλλά τα υπομονοπάτια που περιέχουν τη συμβολοσειρά aqra όχι (υπάρχει επίσης ένα άλλο υπομονοπάτι στο δέντρο που έχει ετικέτα aqra, και ξεκινά πάνω από το χαρακτήρα z, αλλά παραβιάζεται η απαίτηση ότι είναι υπομονοπάτι μονοπατιού που ξεκινά από τη ρίζα).*

Ερώτημα 6

Δίνονται οι ακολουθίες v= GUGTTGTGG και w= TCGTGAATT. 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 7. Άσκηση με Python στην βιοπληροφορική.

Να γραφτεί ένα πρόγραμμα με τη χρήση της γλώσσας προγραμματισμού Python που θα ελέγχει για την ύπαρξη περιοχών δέσμευσης των παρακάτω μεταγραφικών παραγόντων (Transcription Factor - TF). Στη συνέχεια θα εμφανίζει τις θέσεις αυτών των περιοχών.

Transcription Factor	**Consensus Sequence**
RUNX1	BHTGTGGTYW
TGIF1	WGACAGB
IKZF1	BTGGGARD

Η ακολουθία στην οποία θα γίνει ο έλεγχος φαίνεται παρακάτω. Μπορείτε να την τοποθετήσετε σε ένα .fasta αρχείο και να την διαχειριστείτε με την βοήθεια της Biopython.

- *Εμφάνιση θέσεων περιοχών για κάθε παράγοντα και αποθήκευση τους στο κατάλληλο αρχείο.*

- *Να παρουσιάσετε στατιστικά στοιχεία της ακολουθίας.*
- *Να αποθηκεύσετε σε ένα αρχείο τα στατιστικά στοιχεία που εντοπίσατε (αριθμός βάσεων, ποσοστό CG σε σχέση με την ακολουθία).*
- *Να δημιουργήσετε ένα αρχείο που θα περιλαμβάνει την συμπληρωματική ακολουθία και την ακολουθία μετά από μεταγραφή.*

Προτείνετε η χρήση των βιβλιοθηκών string, Biopython ή re (που επιτρέπει την χρήση κανονικών εκφράσεων)

```
>Sequence

GACACCTCAGTACTAGGATGTATCAGCCTGAACTAGCAGGCCTGGTTCCAAATTTTTTTATCAACACTCG
TAGGGGGATTATCCTAGAGGGGGTCTGGGATTTCTTTGACATCAGAGTATTTTTGCCTTGCTCCTTCACA
ATTTGGGAACAAATAATTTAGTGGTTATTAACCCTGGCTACGCACTGGAAACTTTAAAAATAATGCTGGT
ATGAAATTTACACAGAGTATCGTGAAAATTTTCACTGAGTACCATGTGGTTATACATTGGATAAGGCTCC
AGGAAGCAGCTACTGGAAGACAGCCATGCCAAGAGTGGTTAGTGGTTGGAATTTTGGCAAGTCAGTTTTA
GTCTGCCTTATCAAATACATGGGCATACAGATAAATCCTTAGATGGCTCTCCTACTTACTGAAACATTTT
CTATCTATCTATCTATCTATCTATCTATTTGGGAAGCTATCTATCTATCTATCATTTATTTAAGGTAGTC
TCTATCTGCCTCTGTCTCTGTCTGTCTCTGTGTCTCTGTGTCTGTCTGCTCTCTCTCTCTCTCTGTGGGA
ATCTCTCTCTGTGTGTGTGTGTGTATGTGTGTGTGTGTGTGTGTGGTGTGCATGAACATGAGTAAAATCC
ATAAGGAAACTTTCAGAGTTGGTCCTCTCCTTATATCAAATGGATCCAGGAATTAAACTCAGGTTCAATT
CTTGGTGCCTTTACTAGTTGAGCCATCTCACTGGCTCTTCATCATCTTTAGAATAAACTCACTTTATTAC
ACACACACACACACACACAACCTGGGAGTACACACACACACACAACCAAAGCCCCAACGGAAAACTACAA
TATTATAATGAATACACAGGTTCTCAACATAGTCTCTGCCACGCTTGCAGACAAAGATGAGTAGAAGTAG
AAAGAACCAGGGAAACGTGGAGCAAGTCAGAAGGAATAACAGTCAGAAGGAATAACAGTCAGAAGGAATA
ACAGTCAGAAGGAGTAACAGTCAGAAGGAATAGCAGTCAGAAGGAATAACAGTCAGAAGACAGCACAGTC
AGAAGGAATAACAGTCAGAAGGAATAACAGTCAGAAGGAATAACAGTCAGAAGGAATAACAGTCAGAAGG
AATAGCAGTCAGAAGGAATAACAGTCAGAAGGAATAACAGTCAGAAGGAATAACAGTCAAAGAAATAGCA
GTCAGAAGGAATAGCAGTCAGAAGGAATAACAGTCAAAGGAGCAGTCAGAAGGAGTAACAGTCAGAAGGA
ATAACAGTCAGAAGGAATAACAGTCAAAGGAATAGCAGTCAGAAGGAGTAACAGTCAGAGCAAACACAGA
GATGACAAAGGCAATGGGGTCAGAGACTTCACCACTCTCCAAGATCTACTATATACTCTCTCTGTGT
```

Θα παρατηρήσατε ότι οι ακολουθίες προς εξέταση δεν περιλαμβάνουν μόνο τις βάσεις A,T,G,C αλλά και άλλους χαρακτήρες. Αυτοί ονομάζονται κώδικες ασάφειας (ambiguity codes) και φαίνονται στον παρακάτω πίνακα[3].

Code	Represents
A	Adenine
G	Guanine
C	Cytosine
T	Thymine
Y	Pyrimidine (C or T)
R	Purine (A or G)
W	weak (A or T)
S	strong (G or C)
K	keto (T or G)
M	amino (C or A)
D	A, G, T (not C)
V	A, C, G (not T)
H	A, C, T (not G)
B	C, G, T (not A)

[3] https://www.dnabaser.com/articles/IUPAC%20ambiguity%20codes.html

<u>ΜΗ ΥΠΟΧΡΕΩΤΙΚΟ ΚΟΜΜΑΤΙ ΤΗΣ ΑΣΚΗΣΗΣ[4]</u>

- Software as a service → pipelines με τη χρήση της βιβλιοθήκης argparse
- Εφαρμογή Suffix Tree για αναζήτηση και χρονική σύγκριση με naïve αλγόριθμο.
- Μπορούν να εφαρμοστούν οι KMP, Boyer More;
- Χρονική σύγκριση κανονικών εκφράσεων με αλγόριθμους ακριβούς ταιριάσματος.

[4] δεν προσμετράται στην αξιολόγηση είναι απλά για ενασχόληση