ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Τμήμα Πληροφορικής

A blue and red logo

AI-generated content may be incorrect.

Εργασία Μαθήματος «Βιοπληροφορική»

|  |  |
| --- | --- |
| Τίτλος | ***Υπολογιστική Εξαγωγή Μέτρων Κεντρικότητας και Πρόγνωση Ακμών*** |
| Όνομα φοιτητή – Αρ. Μητρώου | Λαζαρίδης Χρήστος-Λάζαρος - Π22083 |
| Καλογερόπουλος Αθανάσιος – Π22223 |
| Ημερομηνία παράδοσης |  |

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

[1. Θέμα 1 3](#_Toc196500559)

[2. Θέμα 2 4](#_Toc196500560)

[3. Θέμα 3 6](#_Toc196500561)

[4. Θέμα 4 6](#_Toc196500562)

[5. Σχόλια υλοποίησης 6](#_Toc196500563)

[6. Βιβλιογραφία 6](#_Toc196500564)

ΠΙΝΑΚΑΣ ΠΙΝΑΚΩΝ

[Table 1: Παράδειγμα πίνακα συχνοτήτων 5](#_Toc196500582)

[Table 2: Αρχικό lookup table 6](#_Toc196500583)

# Θέμα 1

Στο θέμα 1 μας ζητείτε να δημιοουργήσουμε 3 dataset με συγκεκριμένο τρόπο.

Μας δώθηκαν 4 συγκεκριμένα patterns του αλφάβητου

* Pattern 1 = *“ΑΤΤΑGA“*
* Pattern 2 = *“ΑCGCΑTΤT“*
* Pattern 3 =*”ΑCGCΑTΤT“*
* Pattern 4 =*”ATTTCAGT“*

Από τα οποία πρέπει να δημιουργήσουμε 100 συμβολοσειρές , τις οποίες τις δημιουργούμε μεταλλάσοντας και επαυξάνοντας τα παραπάνω patterns με τον εξής τρόπο

* Κάθε sequence αρχίζει με 1 έως 3 τυχαία σύμβολα του αλφάβητου
* Για κάθε καινούργιο sequence που θα δημιουργηθεί επιλέγουμε τα 4 patterns με τη σειρά από το πρώτο στο τέταρτο, μια φορά το καθένα και αλλάζουμε από κανένα μέχρι δύο σύμβολα σε τυχαίες θέσεις, είτε με ένα τυχαίο επιλεγμένα σύμβολο του αλφάβητου, είτε με το κενό, διαγράφοντας το σύμβολο, και τελικά προσθέτουμε το αποτέλεσμα της μετάλλαξης στη μέχρι τώρα συμβολοσειρά, μέχρι να έχουμε προσθέσει το αποτέλεσμα της τυχαίας μετάλλαξης και των 4 patterns.
* Τέλος κάθε sequence θα τελειώνει με 1 έως 2 τυχαία σύμβολα του αλφάβητου.

Η διαδικασία επαναλαμβάνεται 100 φορές και τα 100 sequences που δημιουργούνται θα εισαχθούν με τυχαίο τρόπο σε 3 datasets

* Dataset A: 10 sequences
* Dataset B: 70 sequences
* Dataset C: 20 sequences

Για τις ανάγκες του θέματος κάνουμε χρήση των συναρτήσεων

* **mutate\_pattern**: Εφαρμόζει έως *max\_mutations* τυχαίες αλλαγές (substitution ή deletion) στο δοθέν pattern.
  + Arguments: *pattern* (σε μορφή string) και *max\_mutations,* δηλαδή ένας ακέραιος με τον αριθμό των μεταλλάξεων (default 2)
  + Output: Το μεταλλαγμένο pattern σε μορφή string
* **create\_random\_sequence:** δημιουργεί τυχαία ακολουθία από *min\_len* μέχρι *max\_len*
  + Arguments: 2 ακέραιοι *min\_len* και *max\_len* που δηλώνουν το ελάχιστο και το μέγιστο επιτρεπτό μήκος της συμβολοσειράς
  + Output: Η τυχαία παραγώμενη συμβολοσειρά σε μορφή string

Ακόμη χρησιμοποιούνται και 2 μέθοδοι **save\_fasta** και **save\_txt** που σώζουν τα παραγώμενο datasets σε αρχεία τύπου fasta και txt αντίστοιχα. Για το θέμα δημιουργήθηκε το script **thema\_1.py**

# Θέμα 2

Και τα 2 ΑΜ της ομάδας καταλήγουν σε περριτό αριθμό, επομένως για τις ανάγκες του θέματος α =2.

Θα πρέπει να δημιουργηθεί αλγόριθμος πολλαπλής στοίχισης βασισμένος σε έναν αλγόριθμο global alignment (καθολικής στοίχησης) ο οποίος σύμφωνα με την εκφώνηση θα οριστεί ως εξής

* Η οριζόντια και η κάθετη μετάβαση στο πλέγμα δυναμικού προγραμματισμού θα τιμωρείτε με gap penalty -2
* Η τοπική ομοιότητα θα προσθέτει score +1 (αρχικά, θα μιλήσουμε για το lookup table αργότερα)
* Η τοπική ανομοιότητα θα επιβάλει penalty -1
* Δυνατοί πρόγονοι του κόμβου (i,j) θα είναι οι κόμβοι (i-1, j-1), (i, j-1) και (i-1, j) δηλαδή, ο από κάτω του, ο αριστερά του και ο κάτω και αριστέρα διαγώνιος
* Προφανώς ο αλγόριθμος αποτελεί αλγόριθμο μεγιστοποίησης με βάση την ομοιότητα 2 συμβολοσειρών, οπότε επιλέγουμε πάντα τη μετάβαση που μεγιστοποιεί το σκορ
* Η στοίχηση των 2 προφίλ γίνεται μέσο backtracking με τον εξής τρόπο:
  + Διαγραφή για κάθετη κίνηση (προσθήκη “\_” στη δεύτερη ακολουθία)
  + Εισαγωγή για οριζόντια κίνηση (προσθήκη “\_” στη πρώτη ακολουθία)
  + Όταν εφαρμώσαμε διαγώνια κίνηση τότε μιλάμε για αντιστοίχηση (δεν μας ενδιαφέρει αν τα σύμβολα είναι ίδια)

Για τη διαδικασία της πολλαπλής στοίχισης θα υιοθετηθεί η διαδικασία hierarchical clustering η οποία παραδόθηκε στην 7η διάλεξη του μαθήματος και:

* Για κάθε εποχή (μέχρι να μείνουν με 2 sequences)
  + Δημιουργούμε ένα πίνακα ανομοιοτήτων
  + Επιλέγουμε το το λιγότερο ανόμοιο ζευγάρι
  + Εφαρμόζουμε τον αλγόριθμο καθολικής στοίχισης στο ζευγάρι
  + Φτιάχνουμε την ακολουθία συνένεσης (όχι στη τελευταία εποχή)
  + Αντικαθιστούμε το ζευγάρι με την ακολουθία συνένεσης (όχι στη τελευταία εποχή)
* Κάνουμε unfold τα αποτελέσματα

Για τη δημιουργία του πίνακα ανομοιοτήτων θα χρησιμοποιήσουμε τον αλγόριθμο k-mer ο οποίος διδάχθηκε στην 6η διάλεξη του μαθήματος και ορίζεται ως εξής

* Ορίζουμε κάποιο k
* Με τεχνική συρρώμενου παραθύρου βρίσκουμε όλες τις πιθανές k-αδες και μετράμε τη συχνότητα τους, φτιάχνοντας ένα πίνακα συχνοτήτων ο οποίος θα μπορούσε να μοιάζει ως εξής:

Παράδειγμα: 2 συμβολοσειρές

* + Seq1: “ATGCTAGC”
  + Seq2: “ATGCGC”

Ο πίνακας συχνοτήτων θα ήταν

|  |  |  |
| --- | --- | --- |
|  | Seq1 | Seq2 |
| ATG | 1 | 1 |
| TGC | 1 | 1 |
| GCT | 1 | 0 |
| CTA | 1 | 0 |
| TAG | 1 | 0 |
| AGC | 1 | 0 |
| GCG | 0 | 1 |
| CGC | 0 | 1 |

Table 1: Παράδειγμα πίνακα συχνοτήτων

Από εκεί και πέρα μπορούμε να διανυσματοποιήσουμε το παραπάνω πίνακα για κάθε sequence και να υπολογίσουμε της απόστασης συνημιτόνου

Ορίζουμε την απόσταση συνημιτόνου (cosine distance) ως εξής

όπου:

το εσωτερικό γινόμενο των a και b

το ευκλείδιο μήκος ενός διανύσματος

Τελικά θα καταλήξω με ένα άνω τριγωνικό πίνακα, του οποίου το κελί (i , j) αποτελεί το

* Για την ακολουθία συνένεσης θα χρησιμοποιήσουμε την εξής διαδικασία στις συμβολοσειρές μετά το alignment (το αποτέλεσμα δηλαδή του backtracking στο πλέγμα δυναμικού προγραμματισμού)
  + Αν τότε
  + Αν τότε (και αντίστοιχα)
  + Αν τότε ειςάγω ένα καινούργιο σύμβολο (πχ X) το οποίο ορίζεται ως:

Τα scores αυτά ανανεώνουν το lookup table,έτσω ώστε το X θα προσθέσει score = 0.5 αν βρεθεί με τα ή στον αλγόριθμο στοίχισης. Προφανώς στο lookup table οι βαθμολογίες αρχικά είναι 1 για όμοια σύμβολα και 0 για ανόμοια σύμβολα, δηλαδή

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | A | C | G | T |
| A | 1 | 0 | 0 | 0 |
| C | 0 | 1 | 0 | 0 |
| G | 0 | 0 | 1 | 0 |
| T | 0 | 0 | 0 | 1 |

Table 2: Αρχικό lookup table

Τα scores για κάθε σύμβολο προφανώς και μπορεί να αλλάξουν, αν ένα σύμβολο χρειάζεται να αντιπροσωπεύει περισσότερα σύμβολα, με τα scores να αλλάζουν με βάση το frequency

* Τέλος απαιτείτε να κάνουμε unfold το αποτέλεσμα της πολλαπλής στοίχισης. Ουσιαστικά αυτό είναι το αποτέλεσμα όλων των στοιχισμένων συμβολοσειρών, δηλαδή κάθε στοιχισμένη ακολουθία από την οποία φτιάχνουμε κάποια ακολουθία συνένεσης, θεωρείτε στοιχισμένη, οπότε αν πχ στη πρώτη εποχή στοιχίσω 2 ακολουθίες  *“ACCGTTACCATAC” και “”AGTTTACATC”* και λάβω τη στοίχιση *“ACCGTT\_ACCATAC” και “A\_\_GTTTAC\_AT\_C”* , προφανώς και θα φτιάξουμε την ακολουθία συνένεσης *“ACCGTTTACCATAC”* και οι 2 στοιχισμένες ακολουθίες που θα αφαιρεθούν από τη διαδικασία θα προστεθούν στο αποτέλεσμα της πολλαπλής στοίχισης και θα θεωρηθούν πλέον πλήρως στοιχισμένες. Όποια από τις αρχικές ακολουθίες βγαίνει από τη διαδικασία θεωρείτε στοιχισμένη. Το unfold είναι στη περίπτωση μας η εκτύπωση της στοίχησης
* **Clustering:** Κατά τη δημιουργία του αλγόριθμου σε python βρήκαμε ένα θεμελιώδες πρόβλημα της παραπάνω στρατηγικής το οποίο οφείλεται στο γεγονός ότι σε κάποια επανάληψη μπορεί να επιλεχτούν προς στοίχηση 2 ακολουθίες συνένεσης. Αν σε αυτές εισαχθεί κάποιο κενό, τότε προστίθεται πληροφορία σε ακολουθία συνένεσης που δεν προήλθε από το dataset και αυτό οδηγεί σε ένα αποσυγχρονισμό των στηλών των ακολουθιών που συνέβαλαν στη δημιουργία της αντίστοιχης ακολουθίας συνένεσης. Τελικά κάποιες ακολουθίες είχαν μικρότερο μήκος από κάποιες άλλες. Αυτό λύνετε με 2 τρόπους: Ο ένας θα ήταν να χρησιμοποιούμε πάντα έστω μια ακολουθία από τις αρχικές, και ο δεύτερος (και αυτός που επιλέχθηκε) θα ήταν να διατηρούμε πληροφορία, ομαδοποιόντας μαζί (clustering) όλες τις ακολουθίες που συνέβαλαν στη δημιουργία μιας ακολουθίας συνένεσης. Αν σε αυτή εισαχθεί κάποιο κενό το εισάγουμε μέσο μιας διαδικασίας propagation σε όλες τις ακολουθίες του αντίστοιχου cluster

# Θέμα 3

# Θέμα 4

# Σχόλια υλοποίησης

# Βιβλιογραφία