1ο project στο μάθημα Εισαγωγή στη Βιοπληροφορική

Άγκο Μπεσιάνα 1059662
Ζεκυριά Αθανασία 1059660

Διδάσκοντες: Μακρής Χρήστος, Μεγαλοοικονόμου Βασίλειος

Τμήμα Μηχανικών Ηλεκτρονικών Υπολογιστών και Πληροφορικής
Πανεπιστήμιο Πατρών

Ελλάδα
2021

Περιεχόμενα

Ερώτημα 1

1.α) Στο συγκεκριμένο ερώτημα μας ζητείται η σύνταξη αλγορίθμου με χρήση suffix tree που ελέγχει, έπειτα από προεπεξεργασία O(n) χρόνου, εάν μία συμβολοσειρά P μήκους m εμφανίζεται σε μία συμβολοσειρά T μήκους n, πριν από την τιμή k (με $1 \leq k \leq n$) σε χρόνο O(m).

Ο αλγόριθμος που προτείνουμε παίρνει ως εισόδους τις συμβολοσειρές T και P και τον αριθμό k που συμβολίζει τη θέση που θέλουμε να σταματήσει η σύγκριση των δύο συμβολοσειρών. Ακολουθεί ο αλγόριθμος:

1. Εισαγωγή T, P και k.
2. Δημιούργησε suffix tree ST για την συμβολοσειρά T.
3. Ξεκίνα από τον πρώτο χαρακτήρα του P και τη ρίζα του ST και σύγκρινε κάθε χαρακτήρα των συμβολοσειρών 1-1.
    α. Για τον τωρινό χαρακτήρα του P, αν υπάρχει ακμή από τον τωρινό κόμβο του ST, τότε ακολούθα την ακμή.
    β. Αν δεν υπάρχει ακμή, δεν εμφανίζεται η συμβολοσειρά P στην συμβολοσειρά T.
4. Αν έχουν ελεγχθεί όλοι οι χαρακτήρες του P στο ST και ο αριθμός του φύλλου που εμφανίζεται η συμβολοσειρά P είναι μικρότερο του k, τότε έχουμε πλήρες ταίριασμα πριν την θέση k.

Ο αλγόριθμος για το ταίριασμα των δύο συμβολοσειρών χρειάζεται χρόνο O(m) όπου m είναι το μήκος της συμβολοσειράς P. Επίσης η δημιουργία του suffix tree χρειάζεται χρόνο O(n) όπου n το μήκος της συμβολοσειράς T.

1.β) Στο ερώτημα αυτό, μας ζητείται ο σχεδιασμός αλγορίθμου που ανακαλύπτει την ελάχιστη συμβολοσειρά που εμφανίζεται μόνο μία φορά σε ένα κείμενο. Θα χρησιμοποιήσουμε ένα suffix tree καθώς χρησιμοποιούνται εκτενώς σε προβλήματα συμβολοσειρών.

Ο αλγόριθμος [1] ξεκινάει με είσοδο κείμενο έστω T. Ακολουθεί ο αλγόριθμος:

1. Εισαγωγή T.
2. Δημιούργησε suffix tree ST για το κείμενο T.
3. Για κάθε θέση p του δέντρου, ακολουθεί η διαδικασία εύρεσης του LSUS(p) (locational shortest unique substring).

    α. Βρες το φύλλο του κόμβου που αντιστοιχεί στο επίθεμα ST [p,n].

    β. Αν η ετικέτα της ακμής του φύλλου είναι $, τότε το LSUS(p) δεν υπάρχει και επίστρεψε null. Αλλιώς, συνέχισε τη διαδικασία εύρεσης.

    γ. Έστω L το μήκος της ετικέτας της ακμής του φύλλου χωρίς το $.

    δ. Τότε το LSUS(p) είναι ST[p, n-L+1].

    ε. Αποθήκευσε το LSUS(p).

4. Σύγκρινε όλα τα LSUS και επίλεξε το μικρότερο.

Ο παραπάνω αλγόριθμος χρειάζεται O(n) χρόνο για την εύρεση του LSUS για μία συγκεκριμένη θέση, άρα για m θέσεις απαιτεί χρόνο m*O(n).

Ερώτημα 2

2.α) Στο συγκεκριμένο ερώτημα, μας ζητείται αλγόριθμος για την εύρεση των κοινών υποσυμβολοσειρών που επαναλαμβάνονται και στις k ακολουθίες που μας δίνονται.

Μέχρι τώρα έχουμε χρησιμοποιήσει μόνο suffix trees, αλλά αυτά δεν είναι επαρκή για την σύνταξη του εν λόγω αλγορίθμου. Για τον λόγο αυτό θα χρησιμοποιήσουμε τα Γενικευμένα Δέντρα Επιθεμάτων (Generalized Suffix Tree). Ένα GST αποτελεί ένα suffix tree αλλά για πολλές συμβολοσειρές ταυτόχρονα. Αυτή είναι η ουσιαστική διαφορά του σε σχέση με το απλό suffix tree.

Άρα ο συγκεκριμένος αλγόριθμος θα ξεκινήσει κατασκευάζοντας ένα GST για όλες τις συμβολοσειρές μας. Η κατασκευή του δέντρου αυτού απαιτεί O(T1+T2+..+TN), όπου T1..TN αποτελούν τις συμβολοσειρές.

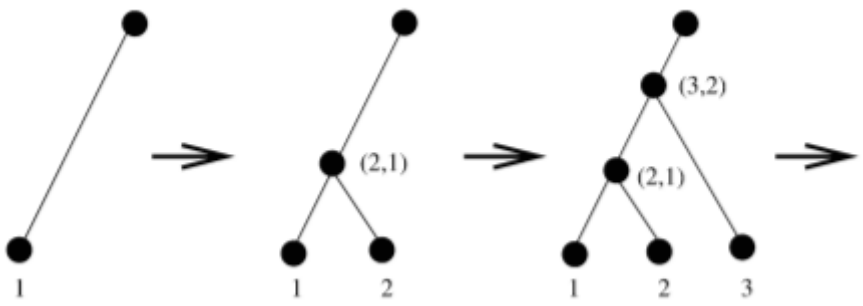
Η αρχική μας σκέψη είναι να σχεδιάσουμε αλγόριθμο ο οποίος εκτελεί έλεγχο για την εύρεση επαναλαμβανόμενων υποσυμβολοσειρών σε μία εκ των δεδομένων συμβολοσειρών και στην συνέχεια, με χρήση GST για k συμβολοσειρές να εξετάζεται αν οι υποσυμβολοσειρές που βρέθηκαν στην πρώτη αναζήτηση είναι κοινά. Ωστόσο, βρήκαμε έναν έτοιμο αλγόριθμο που βασίζεται στην σκέψη αυτή, στο άρθρο [2], ο οποίος βρίσκει λύση σε γραμμικό χρόνο.

Ο αλγόριθμος που προτείνουμε είναι ο ακόλουθος:

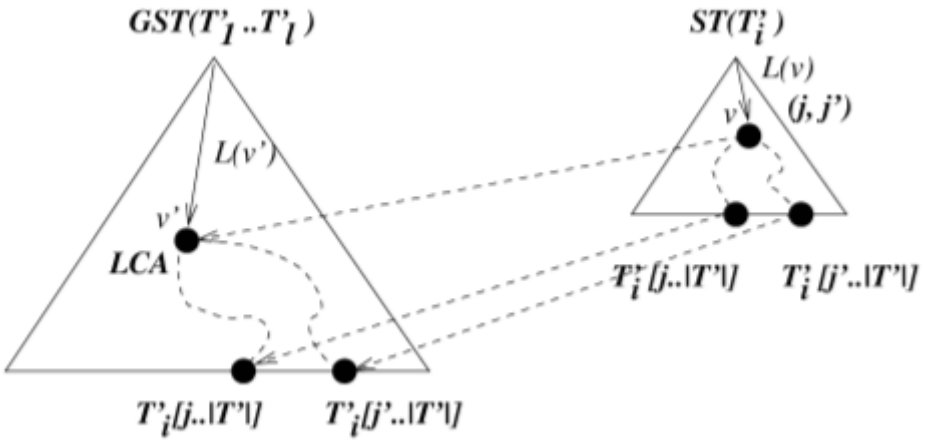
1. Δημιούργησε μια νέα συμβολοσειρά T_i' για κάθε $1 \leq i \leq l$ ώστε να λαμβάνονται υπόψη οι αντίστροφες και οι αντίστροφα συμπληρωμένες επαναλήψεις.
2. Φτιάξε suffix tree ST(T_i') για κάθε $1 \leq i \leq l$ και GST για όλες αυτές τις συμβολοσειρές.
3. Βρες τις supermaximal επαναλήψεις (επαναλήψεις οι οποίες δεν βρίσκονται μέσα σε μία άλλη μεγαλύτερη επανάληψη) του T_i' για κάθε i στο GST.

Στο βήμα 2, κατά τη κατασκευή του ST, διατηρείται πληροφορία για τον χαμηλότερο κοινό πρόγονο (LCA) δύο φύλλων που αντιπροσωπεύουν $T_i'[j \dots |T_i'|]$ και $T_i'[j'' \dots |T_i'|]$.

Στο βήμα 3, ουσιαστικά, βρίσκουμε τις supermaximal επαναλήψεις για κάθε ξεχωριστή συμβολοσειρά χρησιμοποιώντας το suffix tree της. Στη συνέχεια, χρησιμοποιώντας αυτή τη πληροφορία και τη πληροφορία του βήματος 2, βρίσκουμε το LCA για δύο φύλλα στο GST και καταλήγουμε στο επιθυμητό αποτέλεσμα.



Εικόνα 1: Βήμα 1



Εικόνα 2: Βήμα 3

Όσον αφορά το πρόβλημα σχετικά με τους περιορισμούς στα κενά ανάμεσα στις δύο εμφανίσεις της συμβολοσειράς σε κάθε ακολουθία, ο αλγόριθμος αυτός το προσπερνάει.

2.β) Στο ερώτημα αυτό, μας ζητείται ο σχεδιασμός αλγορίθμου για την εύρεση του μέγιστου κοινού προθέματος κάθε ζεύγους συμβολοσειρών n χαρακτήρων ενός συνόλου k συμβολοσειρών. Το πρόβλημα αυτό ταυτίζεται στην βιβλιογραφία με το πρόβλημα εύρεσης της μέγιστης κοινής επέκτασης μεταξύ συμβολοσειρών, άρα θα αναλύσουμε αυτό το πρόβλημα.

Ακολουθεί ο αλγόριθμος:

1.Δημιούργησε Generalized Suffix Tree ST για το σύνολο των συμβολοσειρών T1..TN.

2.Προεπεξεργάσου το ST δέντρο ώστε ο χαμηλότερος κοινός πρόγονος των φύλλων του ST για κάθε ζεύγος συμβολοσειρών και ένα ζεύγος δεικτών i, j να μπορεί να βρεθεί σε σταθερό χρόνο.

3.Βρες τον χαμηλότερο κοινό πρόγονο των φύλλων του δέντρου που αντιστοιχούν στα αντίστοιχα επιθέματα των συμβολοσειρών που εξετάζεις. Έστω ότι αυτός ο κόμβος είναι ο v.

4.Η συμβολοσειρά που αποτελεί την ετικέτα του μονοπατιού προς το v αποτελεί την μέγιστη υποσυμβολοσειρά έστω της ακολουθίας T1 ξεκινώντας από τον δείκτη i, η οποία ταιριάζει με την υποσυμβολοσειρά της ακολουθίας T2 ξεκινώντας από το σημείο με δείκτη j.

5.Όποτε βρίσκεις την μέγιστη κοινή επέκταση, άρα και το μέγιστο κοινό πρόθεμα, μεταξύ ζεύγους συμβολοσειρών, αύξησε τον μετρητή a κατά 1 και μετά την τελευταία σύγκριση ζεύγους συμβολοσειρών, εμφάνισε το a, που πλέον αποτελεί το πλήθος των μέγιστων επιθεμάτων.

Σημείωση: Ο αλγόριθμος αυτός δουλεύει για ζεύγη συμβολοσειρών κάθε φορά. Με αυτόν τον τρόπο προκύπτει και η χρονική πολυπλοκότητα που ζητείται, O(k*n +a), καθώς ο αλγόριθμος θα εκτελεστεί k φορές για τα k διαφορετικά ζεύγη συμβολοσειρών των n χαρακτήρων η καθεμία.

Ερώτημα 3

3.α) Στο ερώτημα αυτό μας ζητείται η εύρεση των διακριτών υποσυμβολοσειρών μιας συμβολοσειράς Τ σε O(n) χρόνο, όπου n είναι το μήκος της Τ, και η αναφορά κάθε αντιγράφου τους σε χρόνο ανάλογο του μήκους όλων των διακριτών συμβολοσειρών.

Η εύρεση των διακριτών υποσυμβολοσειρών είναι εύκολη διαδικασία. Ουσιαστικά, κάθε διακριτή υποσυμβολοσειρά της συμβολοσειράς είναι το μονοπάτι από τη ρίζα που τελειώνει μετά από διαφορετικό χαρακτήρα. Π.χ. έχοντας την συμβολοσειρά abb, οι διακριτές συμβολοσειρές είναι abb, ab, a, bb, b. Επομένως, ο αριθμός των υποσυμβολοσειρών, αποτελεί τον αριθμό των χαρακτήρων στο δέντρο.

Φτιάχνουμε ένα ukkonen suffix tree για την συμβολοσειρά χρόνου O(n). Ισχύει ότι κάθε υποσυμβολοσειρά αποτελεί πρόθεμα κάποιου επιθέματος της συμβολοσειράς κι ότι όλα τα επιθέματα υπάρχουν στο αντίστοιχο trie. Άρα υπάρχει 1-1 αντιστοιχία μεταξύ υποσυμβολοσειρών και μονοπατιών στο trie και 1-1 αντιστοιχία μεταξύ των χαρακτήρων του δέντρου και των όχι άδειων μονοπατιών. Αυτό συμβαίνει γιατί, κάθε ξεχωριστό όχι άδειο μονοπάτι τελειώνει σε ξεχωριστή θέση μετά τον τελευταίο του χαρακτήρα, και το μονοπάτι μέχρι τη συγκεκριμένη θέση, ακολουθώντας κάθε χαρακτήρα είναι μοναδικό.

Κρατώντας την θέση που τελειώνει κάθε διακριτή υποσυμβολοσειρά, είναι εύκολο μετά με μια απλή αναζήτηση να εντοπίσουμε την αντίστοιχη υποσυμβολοσειρά και να εμφανίσουμε το αντίγραφό της. Αυτό ενισχύεται και από το skip/count trick που έχουν τα suffix trees που κατασκευάζονται με τον τρόπο που έχει προτείνει ο Ukkonen.

3.β) Στο δεύτερο ερώτημα της άσκησης μας ζητείται η μετατροπή ενός GST k συμβολοσειρών (στο οποίο έχουν ήδη τοποθετηθεί suffix links) σε ένα αντίστοιχο αυτόματο Aho-Corasick των k συμβολοσειρών σε γραμμικό χρόνο.

Γνωρίζουμε ότι για την μετατροπή των suffix links σε αυτόματο AC είναι απαραίτητη η δημιουργία τριών συναρτήσεων: GoTo, Failure, Output.

GoTo: Δημιουργούμε trie. Και για όλους τους χαρακτήρες που δεν έχουν άκρο στη ρίζα, προσθέτουμε ένα άκρο πίσω στη ρίζα.

Failure: Για μια κατάσταση s, βρίσκουμε το μεγαλύτερο κανονικό επίθεμα το οποίο είναι ένα κανονικό πρόθεμα κάποιου pattern. Αυτό γίνεται με Breadth First Traversal of Trie.

Output: Για μια κατάσταση s, αποθηκεύονται δείκτες όλων των λέξεων που τελειώνουν σε s. Οι δείκτες αυτοί αποθηκεύονται ως αντιστοιχία bit (με τη χρήση τιμών OR bit). Χρησιμοποιείται, επίσης Breadth First Traversal με αποτυχίες (Failure).

Για τη διαδικασία αυτή, εκμεταλλευόμαστε την ιδιότητα των υποσυμβολοσειρών του δέντρου που λέγεται κλειστότητα (substring-closed). Έστω P⊆ Σ* με Σ* το σύνολο του αλφαβήτου που χρησιμοποιείται στις συμβολοσειρές του δέντρου, ισχύει ότι $p = \{s \in \Sigma^* | t \in P, s\ υποσυμβολοσειρά\ του\ t\ \}$. Στην συγκεκριμένη περίπτωση, οι ακμές και τα suffix links του GST του P είναι ισόμορφα των GoTo και των Failure functions ενός AC αυτόματου του P για κάθε διακλαδωμένο κόμβο. Παρακάτω αναπαρίσταται η παραπάνω διαδικασία:

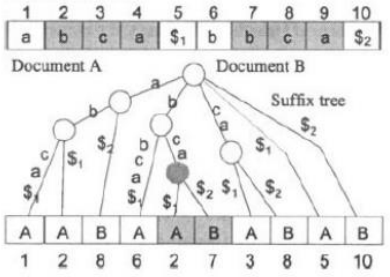


Fig. 2. The generalized suffix tree (GST) ST for $S = \{abca\$_1, bbca\$_2\}$. The shadowed node corresponds to the substring bca and two shadowed leaves correspond to the two occurrences of bca. Suffix links are omitted. (Sec. 4.2)

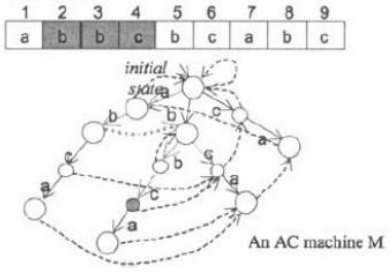


Fig. 3. The AC-pattern matching machine M for $P = \{abca, bbca, bca, ca, a\}$, which corresponds to all internal nodes of the GST in Fig. 2. The solid and the dotted lines indicate goto and failure functions, respectively. (Sec. 4.3)

[3]

Συμπεραίνουμε πως η χρονική πολυπλοκότητα μετατροπής του GST με suffix links σε AC είναι O(N), με N το μέγιστο μήκος των υποσυμβολοσειρών.

3.γ) Θεωρώντας συμβολοσειρά S n χαρακτήρων και αναστροφή αυτής S^r, μας ζητείται να μελετήσουμε τη σχέση των δύο suffix trees τους.

- Παρατηρούμε ότι για κάθε κόμβο v στο ST(S) υπάρχει αντίστοιχος μοναδικός κόμβος στο ST(S^r) για το αντίστροφο label.
- Έστω κόμβος v1 με ancestor v2 στο ST(S^r), τότε το label του v2 θα είναι πρόθεμα του label του v1.
- Αν ένας κόμβος στο ST(S^r) είναι $, τότε το κανονικό label είναι πρόθεμα.
- Μόνο οι κόμβοι $ έχουν αντίστοιχο κανονικό label στο ST(S).

Ερώτημα 4

4.α) Στο ερώτημα αυτό εφαρμόζουμε ολική στοίχιση των συμβολοσειρών v = AAGTACCGGA και w = CCTCGTGAATT. Αυτό επιτυγχάνεται με τον πιο γνωστό αλγόριθμο δυναμικού προγραμματισμού ολικής στοίχισης, τον Needleman-Wunsch.

Στην πρώτη γραμμή και στήλη έχουμε τους δείκτες για κάθε αζωτούχα βάση της ακολουθίας βάσει του βιβλίου του Gusfield.

Για κάθε κελί του πίνακα υπολογίζουμε την τιμή του:

1. Το μονοπάτι του πάνω ή αριστερού κελιού αναπαριστά ένα ταίριασμα με κενό, επομένως, στις τιμές των κελιών αυτών προστίθενται το κόστος στοίχισης με κενό.

2. Το μονοπάτι του διαγώνιου κελιού αναπαριστά ένα ταίριασμα ή μία ασυμφωνία, επομένως, στη τιμή του κελιού αυτού προστίθεται το κόστος ταιριάσματος ή ασυμφωνίας αντίστοιχα.

Στο κελί τελικά εισάγεται η μέγιστη αυτών των τιμών. Η δεύτερη γραμμή και η δεύτερη στήλη αρχικοποιούνται με το κόστος της στοίχισης με κενό. Κάθε φορά που συμπληρώνεται η τιμή ενός κελιού, θα σημειώνεται και από ποιο κελί προήλθε για να γίνεται μετά το backtracking. Το backtracking αρχίζει από το τελευταίο κελί του

πίνακα, η τιμή του οποίου είναι και η τιμή της ολικής στοίχισης. Η στοίχιση που θα προκύψει από το backtracking θα είναι και η βέλτιστη.

		C	C	T	C	G	T	G	A	A	T	T
	0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10	-11
A	-1	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10	-11
A	-2	-2	-2	-3	-4	-5	-6	-7	-8	-9	-10	-11
G	-3	-3	-3	-3	-4	-3	-4	-5	-6	-7	-8	-9
T	-4	-4	-4	-2	-3	-4	-2	-3	-4	-5	-6	-7
A	-5	-5	-5	-3	-3	-4	-3	-3	-2	-3	-4	-5
C	-6	-4	-4	-4	-2	-3	-4	-4	-3	-3	-4	-5
C	-7	-5	-3	-4	-3	-3	-4	-5	-4	-4	-4	-5
G	-8	-6	-4	-4	-4	-2	-3	-3	-4	-5	-5	-5
G	-9	-7	-5	-5	-5	-3	-3	-2	-3	-4	-5	-6
A	-10	-8	-6	-6	-6	-4	-4	-3	-1	-2	-3	-4

Ο παραπάνω πίνακας παράχθηκε με την χρήση του [global alignment app].

Τα βαμμένα κελιά απεικονίζουν την επιλεγμένη στοίχιση με τελικό score: -4:

C	C	-	T	-	C	-	G	T	G	A	A	T	T
A	A	G	T	A	C	C	G	-	G	A	-	-	-

4.β) Ο αλγόριθμος δυναμικού προγραμματισμού τοπικής στοίχισης είναι ο Smith-Waterman. Οι δύο κύριες διαφορές μεταξύ αυτού και του αλγορίθμου ολικής στοίχισης είναι ότι στην τοπική στοίχιση δεν συναντώνται αρνητικές τιμές στα κελιά του πίνακα και το ενδιαφέρον μας πλέον στρέφεται στις περιοχές με την μεγαλύτερη ομοιότητα. Οι αρνητικές τιμές μετατρέπονται αυτόματα σε 0 και το backtracking δεν ξεκινάει από το τελευταίο κελί του πίνακα, αλλά από το κελί με τη μέγιστη ομοιότητα και καταλήγει στο πρώτο κελί που συναντάει με τιμή 0.

D(i,j)		C	C	T	C	G	T	G	A	A	T	T
	0	0	0	0	0	0	0	0	0	0	0	0
A	0	0	0	0	0	0	0	0	1	1	0	0
A	0	0	0	0	0	0	0	0	1	2	1	0
G	0	0	0	0	0	1	0	1	0	1	1	0
T	0	0	0	1	0	0	2	1	0	0	2	2
A	0	0	0	0	0	0	1	1	2	1	1	1
C	0	1	1	0	1	0	0	0	1	1	0	0
C	0	1	2	1	1	0	0	0	0	0	0	0
G	0	0	1	1	0	2	1	1	0	0	0	0
G	0	0	0	0	0	1	1	2	1	0	0	0
A	0	0	0	0	0	0	0	1	3	2	1	0

Τα βαμμένα κελιά απεικονίζουν την επιλεγμένη στοίχιση με τελικό score: 3:

C	G	T	G	A
C	G	_	G	A

Ο αλγόριθμος αγνοεί την μικρή διαφορά του ενός χαρακτήρα (T) μεταξύ των δύο ακολουθιών, ώστε να καταλήξει στην βέλτιστη τοπική στοίχιση.

Ερώτημα 5

Διαθέτουμε 2 συμβολοσειρές n χαρακτήρων όπου η καθεμία αποτελείται από n-k+1 υποσυμβολοσειρές μήκους k και μία παράμετρο k. Επομένως, έχουμε $\Theta(n^2)$ ζευγάρια τέτοιων συμβολοσειρών όπου το κάθε στοιχείο ανήκει σε διαφορετική συμβολοσειρά.

Για κάθε ζεύγος ως σκορ προσεγγιστικού ταιριάσματος ορίζεται πόσοι χαρακτήρες ταιριάζουν όταν συγκριθούν οι υποσυμβολοσειρές μήκους k και στόχος μας είναι να υπολογίσουμε όλες αυτές τις τιμές, δίνοντας έναν αλγόριθμο που το λύνει σε χρόνο $\Theta(n^2)$.
Θεωρούμε ως t τα ζευγάρια των δύο συμβολοσειρών και S είναι οι θέσεις έναρξης.

Αλγόριθμος:

1. Αρχικά, δημιουργούμε έναν δισδιάστατο πίνακα το οποίο περιέχει σε κάθε κελί ένα ζεύγος που τα οποία πρέπει να ελεγχθούν.
2. Θέτουμε το σκορ προσεγγιστικού ταιριάσματος μηδέν.
3. Για κάθε S (Si=[1,…,n-k+1]) εάν το σκορ της θέσης έναρξης είναι μεγαλύτερο από το best score τότε θεωρούμε αυτό ως best score, και θεωρούμε ότι το καλύτερο μοτίβο είναι (S1,S2,…,St).
4. Τέλος, επιστρέφουμε (return) το καλύτερο μοτίβο.

Ερώτημα 6

Οι μεταλλάξεις στο DNA συνήθως προκαλούνται από σφάλματα κυρίως κατά τη διαδικασία του αναδιπλασιασμού. Στο συγκεκριμένο ερώτημα μας ζητείται να βρούμε μια βέλτιστη στοίχιση χρησιμοποιώντας μια μέθοδο βαθμολόγησης με συγγενική ποινή ασυμφωνίας. Επομένως, οι ποινές για συνεχόμενες x ασυμφωνίες είναι –(-ρ+σx) όπου ρ > 0 , το μπόνους για κάθε ταίριασμα είναι +1 και η ποινή για προσθήκη και αφαίρεση συμβολοσειρών -ρ.

Ο σκοπός μιας στοίχισης, είναι να εντοπίσει τη βέλτιστη, όταν η διαδρομή είναι γνωστή, η παράθεση των ζευγών των συμβόλων που αντιστοιχούν στα κελιά του πίνακα με την «καλύτερη» διαδρομή, αντιστοιχεί στην τελική στοίχιση.

Επίσης, για δύο ακολουθίες μήκους m το κόστος εύρεσης της βέλτιστης στοίχισης είναι (n*m).

Αλγόριθμος :

Έστω ότι έχουμε ως είσοδο 2 ακολουθίες S1 και S2. Χρησιμοποιώντας δυναμικό προγραμματισμό έχουμε:

- Συμβολίζουμε ως (Si,Sj) την στοίχιση του στοιχείου i με το στοιχείο j της δεύτερης ακολουθίας.
- Τοποθετούμε τις δύο αλληλουχίες σε έναν πίνακα n*m διαστάσεων όπου κάθε στοιχείο του πίνακα είναι η τιμή του σκορ για την καλύτερη στοίχιση.
- Η γραμμή 0 και στήλη 0 αναπαριστούν το κόστος αν προσθέταμε διαδοχικά σφάλματα (κενά) και στις δυο ακολουθίες
- Όσο τα στοιχεία i και j είναι διαφορετικά του μηδέν για οριζόντια στοίχιση έχουμε $S_{i,j-1}$ -ρ , για κάθετη $S_{i-1,j}$ -ρ και για διαγώνια στοίχιση $S_{i-1,j}$ +1 αν S1=S2 ή $S_{i-1,j}$ –(-ρ+σx) αν S1≠S2 .
- Έπειτα ως σκορ παίρνουμε το μέγιστο ανάμεσα στις οριζόντιες, κάθετες και διαγώνιες στοιχίσεις.
- Αν το σκορ ταυτίζεται με την οριζόντια τότε i παραμένει ίδιο ενώ j γίνεται j-1.
- Αν το σκορ ταυτίζεται με την κάθετη τότε το i γίνεται i-1 και το j παραμένει ίδιο.
- Αν όμως ταυτίζεται με τη διαγώνια τότε και τα δυο γίνονται: i-1 και j-1 όπου έχουμε ταίριασμα.

Τέλος, παρατηρούμε ότι για δύο ακολουθίες μήκους m το κόστος εύρεσης της βέλτιστης στοίχισης είναι O(n*m).

Ερώτημα 7

Στο συγκεκριμένο ερώτημα μας ζητήθηκε να εντοπίσουμε ομοιότητες ανάμεσα στο γονίδια αντισωμάτων του ανθρώπινου οργανισμού (homo sapiens) και των γονιδίων του ψαριού-ζέβρα Danio rerio (zebrafish) με σκοπό την αντιμετώπιση αντιγόνων που προσβάλλουν τον άνθρωπο.

Κατεβάσαμε τα αντίστοιχα γονίδια (IG-Heavy chain) από τη Βάση Δεδομένων Ανοσολογίας IMGT/LIGM-DB (http://www.imgt.org/ligmdb/).

Danio rerio (zebrafish):

Accession number	**AF273876**
Sequence version	1
Secondary accession numbers	
IMGT annotation level	by annotators
Definition	Danio rerio clone VH101 immunoglobulin heavy chain variable region mRNA, partial cds.
Species	*Danio rerio* (zebrafish)
Taxonomy	*Gnathostomata; Teleostomi; Euteleostomi; Actinopterygii; Actinopteri; Neopterygii; Teleostei; Osteoglossocephalai; Clupeocephala; Otomorpha; Ostariophysi; Otophysi; Cypriniphysae; Cypriniformes; Cyprinoidei; Cyprinidae; Danio; Danio rerio*
Sequence length	426
IMGT/LIGM-DB dates	29-MAR-2001
	22-FEB-2013 (v. 9)

```
Length : 426 BP
Composition : 129 A; 83 C; 97 G; 117 T; 0 other

     cgtatgatta ttatttcata ctctgattta catcagcagc ttcaagatga agaatgctct        60
     ctgcttactg ctgctctcat tctgtctaca gcgtataaaa tgtcaaagta tggagtcgat       120
     tgaaagctca gtgcagagaa agcctggaga aactctgact ctgtcctgca gaggatccgg       180
     gttcagcttt agctcctatt acatgcactg gatcaggcaa caagctggaa aacctcttgt       240
     gtggattgga ggcactggct atggctatgt tgaatccttt aaaggaagag gtgaaatcac       300
     cagagataat tcaaagagta tgacatatct gaaactatca ggtctgacag tagaagattc       360
     agccgtgtat tactgtgcaa gacaatataa caactacaat gctgcctttg actactgggg       420
     aaaagg                                                                 426
```

Homo Sapiens:

Accession number	**A03907**
Sequence version	1
Secondary accession numbers	
IMGT annotation level	keyword level
Definition	H.sapiens antibody D1.3 variable region protein
Species	*Homo sapiens* (human)
Taxonomy	*cellular organisms; Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Deuterostomia; Chordata; Craniata; Vertebrata; Gnathostomata; Teleostomi; Euteleostomi; Sarcopterygii; Dipnotetrapodomorpha; Tetrapoda; Amniota; Mammalia; Theria; Eutheria; Boreoeutheria; Euarchontoglires; Primates; Haplorrhini; Simiiformes; Catarrhini; Hominoidea; Hominidae; Homininae; Homo; Homo sapiens*
Sequence length	412
IMGT/LIGM-DB dates	11-MAR-1998
	14-JAN-2013 (v. 5)

```
Length : 412 BP
Composition : 105 A; 109 C; 104 G; 94 T; 0 other

     tcagagcatg gctgtcctgg cattactctt ctgcctggta acattcccaa gctgtatcct        60
     ttcccaggtg cagctgaagg agtcaggacc tggcctggtg gcgccctcac agagcctgtc       120
     catcacatgc accgtctcag ggttctcatt aaccggctat ggtgtaaact gggttcgcca       180
     gcctccagga aagggtctgg agtggctggg aatgatttgg ggtgatggaa acacagacta       240
     taattcagct ctcaaatcca gactgagcat cagcaaggac aactccaaga gccaagtttt       300
     cttaaaaatg aacagtctgc acactgatga cacagccagg tactactgtg ccagagagag       360
     agattatagg cttgactact ggggccaagg caccactctc acagtctcct ca               412
```

Για την εύρεση του μήκους της μέγιστης υποσυμβολοσειράς θα πάρουμε από κάθε ακολουθία γονιδίων ένα μικρό τμήμα λόγω του μεγάλου μήκους τους.

Έστω X= ctgtcct από το γονίδιο του ανθρώπινου οργανισμού και Y= ctctgat από το γονίδιο Danio rerio (zebrafish).

Το Suffix tree μίας συμβολοσειράς S[1…n] είναι ένα compressed trie που περιέχει ως κλειδιά όλα τα επιθέματα S[i…n], $1 \leq i \leq n$.

Χρησιμοποιώντας το δέντρο επιθεμάτων (suffix tree) έχουμε:

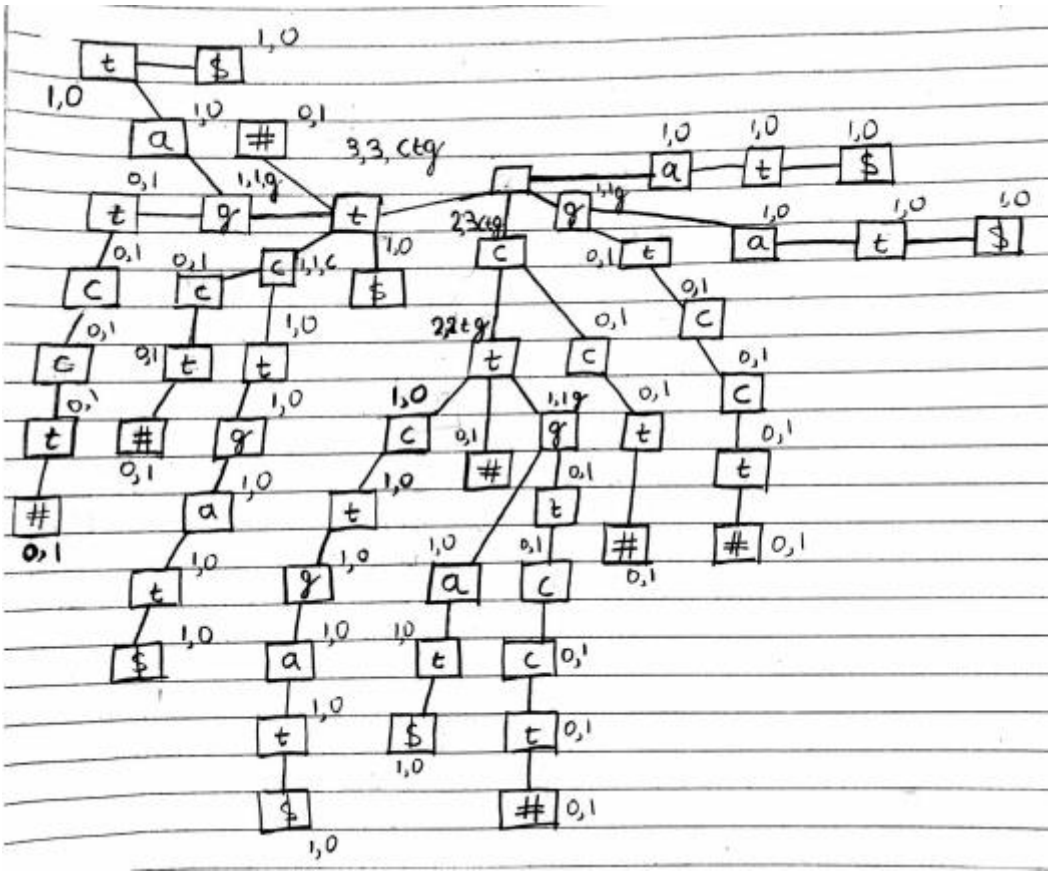
- X = ctgtcct

X = t #, ct #, cct #, tcct #, gtcct #,
tgtcct #, ctgtcct #

- Y = ctctgat

Y = t $, at $, gat $, tgat $, ctgat $
tctgat $, ctctgat $

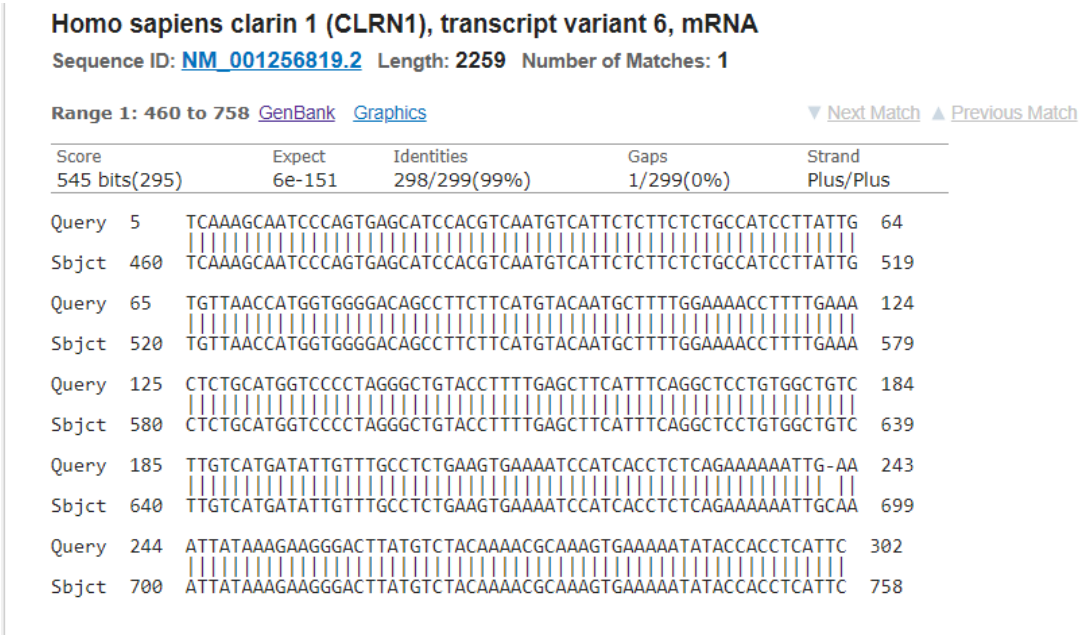
Στη συνέχεια βλέπουμε ότι η μέγιστη υποσυμβολοσειρά είναι η ctg και εμφανίζεται φορά. Τα επιθέματα της ακολουθίας X σημειώνονται με 0,1 , ενώ της ακολουθίας Y με 1,0 .



Ερώτημα 8

8.α)

Το εργαλείο BLAST (Basic Local Alignment Search Tool) είναι βασικό για την αναζήτηση τοπικών στοιχίσεων. Συγκεκριμένα, πρόκειται για ένα υπολογιστικό εργαλείο που μας δίνει τη δυνατότητα να συγκρίνουμε δεδομένες ακολουθίες με τις καταχωρήσεις στις τρέχουσες βιολογικές βάσεις δεδομένων. Επίσης, επιτρέπει την πλήρη αντιστοίχιση μεταξύ δύο ακολουθιών, πρωτεϊνικών ή νουκλεοτιδικών, εντοπίζοντας κοινές υπο-ακολουθίες ίδιου μήκους από τη δοσμένη ακολουθία και από το σύνολο ακολουθιών μιας βάσης δεδομένων κάνοντας χρήση του αλγορίθμου ς Needleman-Wunsch.



Homo sapiens clarin 1 (CLRN1), transcript variant 6, mRNA

Sequence ID: NM_001256819.2 Length: 2259 Number of Matches: 1

Range 1: 460 to 758 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
545 bits(295)	6e-151	298/299(99%)	1/299(0%)	Plus/Plus

```
Query  5    TCAAAGCAATCCCAGTGAGCATCCACGTCAATGTCATTCTCTTCTCTGCCATCCTTATTG  64
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  460  TCAAAGCAATCCCAGTGAGCATCCACGTCAATGTCATTCTCTTCTCTGCCATCCTTATTG  519

Query  65   TGTTAACCATGGTGGGGACAGCCTTCTTCATGTACAATGCTTTTGGAAAACCTTTTGAAA  124
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  520  TGTTAACCATGGTGGGGACAGCCTTCTTCATGTACAATGCTTTTGGAAAACCTTTTGAAA  579

Query  125  CTCTGCATGGTCCCCTAGGGCTGTACCTTTTGAGCTTCATTTCAGGCTCCTGTGGCTGTC  184
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  580  CTCTGCATGGTCCCCTAGGGCTGTACCTTTTGAGCTTCATTTCAGGCTCCTGTGGCTGTC  639

Query  185  TTGTCATGATATTGTTTGCCTCTGAAGTGAAAATCCATCACCTCTCAGAAAAAATTG-AA  243
            |||||||||||||||||||||||||||||||||||||||||||||||||||||||| ||
Sbjct  640  TTGTCATGATATTGTTTGCCTCTGAAGTGAAAATCCATCACCTCTCAGAAAAAATTGCAA  699

Query  244  ATTATAAAGAAGGGACTTATGTCTACAAAACGCAAAGTGAAAAATATACCACCTCATTC  302
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  700  ATTATAAAGAAGGGACTTATGTCTACAAAACGCAAAGTGAAAAATATACCACCTCATTC  758
```

Η συγκεκριμένη ακολουθία προέρχεται από το γονίδιο του οργανισμού Homo sapiens και ονομάζεται clarin 1 (CLRN1) , transcript variant 6 , mRNA με accession number της GenBank: NM_001256819 και version: NM_001256819.2 .

```
LOCUS       NM_001256819            2259 bp    mRNA    linear   PRI 06-APR-2021
DEFINITION  Homo sapiens clarin 1 (CLRN1), transcript variant 6, mRNA.
ACCESSION   NM_001256819
VERSION     NM_001256819.2
```

Χρησιμοποιώντας το accession number της GenBank οι πληροφορίες που βρήκαμε για το γονίδιο είναι οι εξής:

- Η παλαιότερη αναφορά στο γονίδιο αυτό ήταν το 1993.
- Το γονίδιο βρίσκεται στη θέση (locus) : NM_001256819 2259 bp mRNA linear PRI 06-APR-2021 .
- Έχουν αναγνωριστεί 2 παράλληλες εμφανίσεις του γονιδίου στο χρωμόσωμα 3.
- Το όνομα της πρωτεϊνης που παράγεται είναι clarin-1 isoform e και ο κωδικός της GenBank είναι NP_001243748.1 .

  ```
  /product="clarin-1 isoform e"
  /protein_id="NP_001243748.1"
  ```

- Η θέση της δεδομένης ακολουθίας στην οποία ξεκινάει η μετάφραση είναι στη θέση 20και τελειώνει στη θέση 577 (20…577).

  ```
  20..577
  /gene="CLRN1"
  ```

- Από το γονίδιο κωδικοποιούνται 185 αμινοξέα.
- Λειτουργία:
  Η clarin 1 είναι μια πρωτεΐνη που στους ανθρώπους κωδικοποιείται από το γονίδιο CLRN1. Η κωδικοποιημένη πρωτεΐνη έχει βρεθεί σε αρκετές περιοχές του σώματος συμπεριλαμβάνοντας αισθητήρια κύτταρα στο εσωτερικό του αυτιού. Μπορεί, επομένως, να είναι σημαντική για την ανάπτυξη και την ομοιόσταση στο εσωτερικό του αυτιού και του αμφιβληστροειδούς, ο οποίος είναι ο ιστός που ανιχνεύει το φως που ευθυγραμμίζει το πίσω μέρος του ματιού. Επίσης, οι μεταλλάξεις σε αυτό το γονίδιο έχουν συσχετιστεί με τον τύπο του συνδρόμου Usher IIIa.

Αναζήτηση (translated blast search – blastx)) χρησιμοποιώντας την άγνωστη ακολουθία:

clarin-1 isoform c [Homo sapiens]

Sequence ID: NP_443721.1 **Length:** 120 **Number of Matches:** 2

See 5 more title(s) See all Identical Proteins(IPG)

Range 1: 14 to 105 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
160 bits(404)	2e-48	Compositional matrix adjust.	81/92(88%)	84/92(91%)	0/92(0%)	+1

```
Query  4    VKAIPVSIHVNVILFSAILIVLTMVGTAFFMYNAFGKPFETLHGPLGLYLLSFISGSCGC  183
            +KAIPVSIHVNVILFSAILIVLTMVGTAFFMYNAFGKPFETLHGPLGLYLLSFISGSCGC
Sbjct  14   LKAIPVSIHVNVILFSAILIVLTMVGTAFFMYNAFGKPFETLHGPLGLYLLSFISGSCGC  73

Query  184  LVMILFASEVKIHHLSEKIEIIKKGLMSTKRK  279
             LVMILFASEVKIHHLSEKI   K+G    K +
Sbjct  74   LVMILFASEVKIHHLSEKIANYKEGTYVYKTQ  105
```

Range 2: 94 to 113 GenPept Graphics ▼ Next Match ▲ Previous Match ▲ First Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
43.1 bits(100)	0.018	Compositional matrix adjust.	20/20(100%)	20/20(100%)	0/20(0%)	+3

```
Query  243  NYKEGTYVYKTQSEKYTTSF  302
            NYKEGTYVYKTQSEKYTTSF
Sbjct  94   NYKEGTYVYKTQSEKYTTSF  113
```

```
LOCUS       NP_443721                120 aa            linear   PRI 06-APR-2021
DEFINITION  clarin-1 isoform c [Homo sapiens].
ACCESSION   NP_443721
VERSION     NP_443721.1
```

- Το όνομα της πρωτεΐνης είναι clarin-1 isoform c.
- Accession number: NP_443721 .

Στις δύο αναζητήσεις έχουμε το ίδιο γονίδιο clarin 1 (CLRN1) παρόλα αυτά οι δύο αναζητήσεις δεν ταυτίζονται μεταξύ τους γιατί στην δεύτερη περίπτωση προκύπτει η clarin-1 isoform c.

8.β)

- Υπάρχουν συνολικά 32.453 είδη βακτηρίων.

Ranks:	higher taxa	genus	species	lower taxa	total
Archaea	470	215	745	0	1,430
Bacteria	5,007	4,442	22,097	907	32,453
Eukaryota	62,270	91,970	470,302	33,297	657,839
Fungi	5,481	6,978	50,282	1,516	64,257
Metazoa	45,152	64,989	244,712	16,522	371,375
Viridiplantae	7,948	16,272	161,776	14,898	200,894
Viruses	1,590	1,489	4,633	10	7,722
All taxa	69,366	98,117	497,766	34,214	699,463

- Για το έτος 1999 προστέθηκαν επιπλέον 667 νέα βακτήρια.

Ranks:	higher taxa	genus	species	lower taxa	total
Archaea	7	5	17	0	29
Bacteria	52	89	488	38	667
Eukaryota	900	2,867	9,111	498	13,376
Fungi	61	276	1,232	43	1,612
Metazoa	670	1,501	3,997	261	6,429
Viridiplantae	100	909	3,455	187	4,651
Viruses	5	6	67	0	78
All taxa	964	2,967	9,683	536	14,150

1999: 01/01 02/01 03/01 04/01 05/01 06/01 07/01 08/01 09/01 10/01 11/01 12/01 1999/01/01
01/31 02/28 03/31 04/30 05/31 06/30 07/31 08/31 09/30 10/31 11/30 12/31 1999/12/31

Στη συνέχεια επιλέγουμε 'Taxonomy Statistics' ->' Extinct organisms' και στην κατηγορία έντομα 'Libanorhinus succinus'.

- Σε αυτήν την κατηγορία ανήκουν οι οργανισμοί :

Lineage *(full)*

cellular organisms; Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Protostomia; Ecdysozoa; Panarthropoda; Arthropoda; Mandibulata; Pancrustacea; Hexapoda; Insecta; Dicondylia; Pterygota; Neoptera; Endopterygota; Coleoptera; Polyphaga; Cucujiformia; Curculionoidea; Nemonychidae; Libanorhinus

Genetic code: Translation table 1 (Standard)
Mitochondrial genetic code: Translation table 5 (Invertebrate Mitochondrial)

- Από αυτόν τον οργανισμό έχουν εισαχθεί 2 ακολουθίες, που αποτελούν τον γενετικό του κώδικα (νουκλεοτιδική ακολουθία) και τον μιτοχονδριακό γενετικό του κώδικα (πρωτεϊνική ακολουθία).

```
LOCUS       LBNRR18S                315 bp    DNA     linear   INV 03-JAN-1997
DEFINITION  Lebanorhinus succinus 18S ribosomal RNA gene, partial sequence.
ACCESSION   L08072
VERSION     L08072.1
KEYWORDS    18S ribosomal RNA.
SOURCE      Libanorhinus succinus
  ORGANISM  Libanorhinus succinus
            Eukaryota; Metazoa; Ecdysozoa; Arthropoda; Hexapoda; Insecta;
            Pterygota; Neoptera; Endopterygota; Coleoptera; Polyphaga;
            Cucujiformia; Nemonychidae; Libanorhinus.
REFERENCE   1  (bases 1 to 315)
  AUTHORS   Cano,R.J., Poinar,H.N., Pieniazek,N.J., Acra,A. and Poinar,G.O. Jr.
  TITLE     Amplification and sequencing of DNA from a 120-135-million-year-old
            weevil
  JOURNAL   Nature 363 (6429), 536-538 (1993)
   PUBMED   8505978
COMMENT     On Jan 3, 1997 this sequence version replaced gi:293030.
FEATURES             Location/Qualifiers
     source          1..315
                     /organism="Libanorhinus succinus"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:27445"
     gene            1..315
                     /gene="18S rRNA"
     rRNA            <1..>315
                     /gene="18S rRNA"
                     /product="18S ribosomal RNA"
                     /note="120-135 million year old nemonychid weevil entombed
                     in Lebanese amber and flanked by the primers NS1 and NS2"
```

- Το όνομα του γονιδίου που έχει αλληλουχηθεί από αυτόν τον οργανισμό είναι Lebanorhinus succinus 18S ribosomal RNA.
- Η συγκεκριμένη εγγραφή έχει 315 ζευγάρια βάσεων.

8.γ) Μέσω του αριθμού PMID της PUBMED που μας δόθηκε βρέθηκε το άρθρο με τον εξής τίτλο:
Internal fight threatens future of Strangeways laboratory in Britain
(Εσωτερικές διαμάχες στο εργοστάσιο Strangeways στη Βρετανία)
Από την σελίδα του άρθρου στην PUBMED μπορούμε να ανασύρουμε διάφορες πληροφορίες, όπως το όνομα του συγγραφέα του άρθρου, **D Dickson**, το έτος που δημοσιεύτηκε το άρθρο, **10 Ιουνίου 1993**, αλλά και το DOI link που μας οδηγεί στην ιστοσελίδα Nature.com, όπου μπορούμε να κατεβάσουμε και να διαβάσουμε το άρθρο.

Επειδή το άρθρο μας φάνηκε αρκετά ενδιαφέρον, ακολουθεί η περίληψη του:

Το μέλλον του βρετανικού εργοστασίου βρίσκεται σε κίνδυνο έπειτα από την διαδοχή του διευθυντή John Dingle από τον Peter Lachmann (καθηγητή ανοσολογίας στο πανεπιστήμιο). Προέκυψε διαφωνία ανάμεσα στο ανώτατο επιστημονικό προσωπικό και τους διαχειριστές όσο αναφορά την αλλαγή επιστημονικών προτεραιοτήτων. Οι διαχειριστές θέλησαν να ασχοληθούν με την έρευνα για τη φλεγμονή, ενισχύοντας τις σχέσεις με το πανεπιστήμιο, ενώ οι άλλοι θεώρησαν πως πρόκειται για εξαγορά από την Ιατρική σχολή.
Το διοικητικό συμβούλιο, το οποίο είναι σημαντικός υποστηρικτής του εργοστασίου επισημαίνει ότι το εργαστήριο πρόκειται να ζητήσει από τη MRC υποστήριξη για την έρευνα της φλεγμονής, τονίζοντας ότι πρέπει να αποδειχθεί βελτίωση στην προοπτική επίτευξης εκτεταμένων εννοιολογικών προόδων στη βιοϊατρική. Ως αποτέλεσμα μεγάλο μέρος του ανώτατου προσωπικού που διαφωνεί με το επιστημονικό περιεχόμενο των σχεδίων του Lachmann μετακόμισε στο Ινστιτούτο Ζωικής Φυσιολογίας και Γενετικής, ενώ η MRC δεν θα ληφθεί καμία απόφαση για το μέλλον του Strangeways μέχρι να λάβει λεπτομερή ερευνητικά σχέδια από τον Lachmann. Παρόλο που ο Lachmann υποστηρίζει ότι δεν θα υποχωρήσει από τις προτεινόμενες αλλαγές πολλοί υποστηρικτές θεωρούν ότι η ζημία στο εργοστάσιο έχει ήδη γίνει.

Swiss loosen restrictions on foreign workers
(Ελάφρυνση περιορισμών για τους αλλοδαπούς εργαζόμενους)

Η Ελβετία χαλαρώνει τους περιορισμούς για αλλοδαπούς εργαζόμενους διευκολύνοντας τους ερευνητές να εγκαταλείψουν τη χώρα τους προσωρινά. Συμφώνα με έναν κυβερνητικό εκπρόσωπο ο περιορισμός στην ανταλλαγή ανθρώπων-εργαζομένων ενδέχεται να παρεμποδίσει την ανάκαμψη της χώρας. Για το λόγο αυτό ψηφίστηκε νέος νόμος που εξαλείφει την απαίτηση να διαφημίζονται οι κενές θέσεις εργασίας και επιτρέπει σε αλλοδαπούς να εγκαταλείψουν τη χώρα για να παρακολουθήσουν μαθήματα χωρίς να χρειάζεται να υποβάλουν εκ νέου αίτηση για άδεια εργασίας. Επιπλέον πολλοί ξένοι επιστήμονες στην Ελβετία απασχολούνται από μεγάλες φαρμακευτικές εταιρείες επιτρέποντας πλέον τα συνοριακά περάσματα από τη Γερμανία και τη Γαλλία.

Αναφορές

[1] S. V. T. M. O. K. Paniz Abedin, «A Survey on Shortest Unique Substring Queries,» 6 September 2020.

[2] C. S. I. K. P. Inbok Lee, «Linear time algorithm for the longest common repeat problem,» *Journal of Discrete Algorithms,* 16 May 2006.

[3] H. A. S. A. R. Fujino, « Discovering Unordered and Ordered Phrase Association Patterns for Text Mining,» *Knowledge Discovery and Data Mining. Current Issues and New Applications,* pp. 281-293, 2000.