Αλέξανδρος Ξιάρχος
st1059619@ceid.upatras.gr
1059619

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ · ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Η/Υ ΚΑΙ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΕΙΣΑΓΩΓΗ ΣΤΗ ΒΙΟΠΛΗΡΟΦΟΡΙΚΗ

ΠΡΩΤΟ ΣΥΝΟΛΟ ΑΣΚΗΣΕΩΝ · 2023–2024

ΠΕΡΙΕΧΟΜΕΝΑ

1 ΕΡΩΤΗΜΑ 1

1.1 ΕΡΓΑΛΕΙΑ ΓΙΑ ΧΕΙΡΙΣΜΟ ΠΡΟΒΛΗΜΑΤΩΝ ΒΙΟΠΛΗΡΟΦΟΡΙΚΗΣ

Η σελίδα της Rosalind περιλαμβάνει κάποια βασικά προβλήματα, με σκοπό μια πρώτη εξοικείωση στο τομέα της Βιοπληροφορικής.

1.1.1 "Introduction to the Bioinformatics Armory": (SMS 2)

Το πρώτο πρόβλημα αφορά την εύρεση των αριθμών των νουκλεοτιδίων από μια ακολουθία DNA. Ένα εργαλείο για την ανάλυση της ακολουθίας είναι το Sequence Manipulation Suite (SMS) 2. Πρόκειται για μια συλλογή Javascript προγραμμάτων για τη δημιουργία, στοίχιση και ανάλυση μικρών DNA και πρωτεϊνικών ακολουθιών. [15] Χρησιμοποιώντας το DNA Stats, εισάγουμε το Sample Dataset και εισάγεται το πλήθος των νουκλεοτιδίων, και το ποσοστό εμφάνισής τους:

Pattern	Times found:	Percentage
g	17	24.29
a	20	28.57
c	21	30
3	12	17.14

1.1.2 "GenBank Introduction": Αναζήτηση

Αφορά τη βάση δεδομένων GenBank. [10] Μπορούμε να αναζητήσουμε ακολουθίες νουκλεοτιδίων και πρωτεϊνών, όπως επίσης και βιβλιογραφικές δημοσιεύσεις.

1.1.3 "Data Formats": Formats της GenBank

Στην GenBank για να υπάρχει μια συνέπεια στην αναπαράσταση των νουκλεοτιδικών ακολουθιών ακολουθείται ένα συγκεκριμένο format που περιλαμβάνει το header, τα χαρακτηριστικά της ακολουθίας και την ίδια την ακολουθία. Το εργαλείο GenBank to Fasta του SMS 2 [16] επιτρέπει την αντιγραφή κάποιου entry από το GenBank και τη μετατροπή του σε FASTA, τη πιο δημοφιλέστερη μορφή αναπαράστασης ακολουθιών. [1] Για παράδειγμα:

```
GenBank to FASTA results
>Strongylocentrotus purpuratus fascin (FSCN1) mRNA, complete cds.
acttgaaagtggataaaatcgactgataccaaaacaacattgttttacagaagtggtcgt
ttgaggacatcaacatatttcacaatgcctgctatgaatttaaaatacaaatttggcctg
```

1.1.4 "New Motif Discovery": Αναζήτηση Motifs σε ακολουθίες

Με το εργαλείο MEME (Multiple Em for Motif Elicitation) [13], εισάγοντας ακολουθίες που περιλαμβάνει motif [2] , εξάγεται η κανονική έκφραση (regular expression) του συγκεκριμένου motif.

[1] Η μορφή FASTA αποτελείται από ένα header (που ξεκινάει με το σύμβολο ">" και ένα αγνωριστικό της ακολουθίας), και τα δεδομένα της αλληλουχίας. Κάθε εγγραφή της GenBank είναι πολύ πιο λεπτομερής σε σχέση με τη FASTA: περιλαμβάνει πληροφορίες όπως το μήκος της ακολουθίας, ημερομηνία τροποποίησης, περιγραφή της ακολουθίας, χαρακτηριστικά της ακολουθίας όπως επίσης και δημοσιεύσεις που σχετίζονται με την ακολουθία.

[2] Πρόκειται για ένα μοτίβο συγκεκριμένης λειτουργικής σημασίας που επαναλαμβάνεται σε μια ακολουθία. Ο εντοπισμός τους έχει σημασία για την κατανόηση της λειτουργίας των αλληλουχιών.

```
Motif TISWYQ MEME-1 regular expression
--------------------------------------------------------
TISWYQ

Motif YQPARIKEFAK MEME-2 regular expression
--------------------------------------------------------
[YQ][ALQ][PC][AGV]R[IR][KV][ERS]F[AMN][KC]
```

Η κανονική έκφραση του μοτίβου που εξήχθη από τη δεύτερη ακολουθία για παράδειγμα, έχει ως πρώτη θέση πάντα το Y ή το Q, στη δεύτερη θέση πάντα τη A, L ή Q,

1.1.5 "Pairwise Global Alignment": Στοίχιση ακολουθιών

Στο εργαλείο Needle [8] μπορούμε να εισάγουμε τα ID από δύο GenBank entries, με σκοπό την ολική στοίχισή τους.

Κομμάτι του αποτελέσματος που εξάγεται:

```
%# Length: 142
%# Identity:     122/142 (85.9%)
%# Similarity:   131/142 (92.3%)
%# Gaps:           0/142 ( 0.0%)
%# Score: 648.0
```

1.1.6 "FASTQ format introduction": Μετατροπή FASTQ σε FASTA

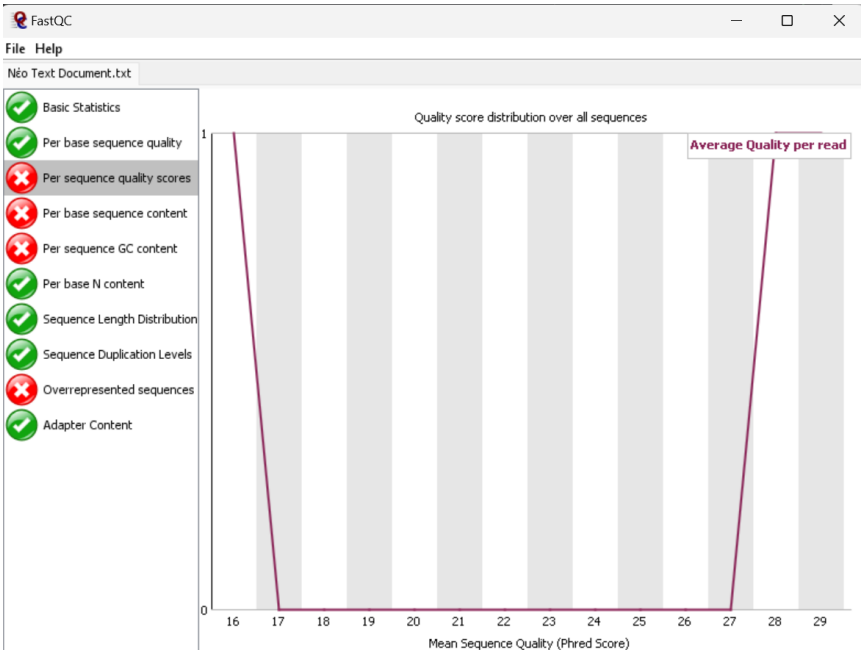
Η διαφορά του FASTQ αρχείου σε σχέση με το FASTA είναι πως περιλαμβάνει quality scores (πληροφορίες ποιότητας) για κάθε νουκλεοτίδιο στην ακολουθία. Αναπαρίσταται ως μια "γραμμή ποιότητας", κάτω από την ακολουθία, συνδεδμένη με ένα "+".

Υπάρχουν διαφορετικοί online convertors που μπορούν να το μετατρέψουν σε FASTA, όπως ο Sequence Conversion της Bugaco, [2] στον οποίον ανεβάζουμε ένα FASTQ αρχείο και το μετατρέπουμε σε αρχείο `.fasta`.

1.1.7 "Read Quality Distribution": Per sequence quality analysis

Το FastQC [**FastQC**] είναι λογισμικό ανάγνωσης ακολουθιακών δεδομένων, το οποίο μπορεί να εξάγει γραφικά και πίνακες ελέγχου ποιότητας των ακολουθιών.

```
INPUT:
    @Rosalind_0041
    GGCCGGTCTATTTACGTTCTCACCCGACGTGACGTACGGTCC
    +
    6.3536354;.151<211/0?::6/-2051)-*"40/.,+%)
    @Rosalind_0041
    TCGTATGCGTAGCACTTGGTACAGGAAGTGAACATCCAGGAT
    +
    AH@FGGGJ<GB<<9:GD=D@GG9=?A@DC=;:?>839/4856
    @Rosalind_0041
    ATTCGGTAATTGGCGTGAATCTGTTCTGACTGATAGAGACAA
    +
    @DJEJEA?JHJ@8?F?IA3=;8@C95=;=?;>D/:;74792
```



Σχήμα 1.1: Γραφικό περιβάλλον FastQC.

1.1.8 "Protein Translation": SMS 2 Translate

Μέσω του εργαλείου Translate του SMS 2 [17], μπορούμε να μεταφράσουμε την αλληλουχία των νουκλεοτιδίων σε αμινοξέα. Για παράδειγμα:

```
INPUT:
    >test
    ATGGCCATGGCGCCCAGAACTGAGATCAATAGTACCCGTATTAACGGGTGA
OUTPUT:
    >rf 1 test
    MAMAPRTEINSTRING*
```

1.1.9 "Read Filtration by Quality": FASTQ Quality Filter

Μπορούμε να "καθαρίσουμε" αδιάφορα κομμάτια από τις αλληλουχίες, με στόχο τη βελτίωση της ποιότητας των δεδομένων μας, χρησιμοποιώντας το FASTQ Quality Filter της Galaxy. [9]

Από το εργαλείο εξάγεται το αρχείο `Galaxy2-[Filter_by_quality_on_data_1].fastqsanger` το οποίο περιλαμβάνει μόνο τα φιλταρισμένα entries.

1.1.10 "Complementing a Strand of DNA": SMS 2 Reverse Complement

Το Reverse Complement του SMS 2 επιστρέφει τα συμπληρωματικά νουκλεοτίδια. Για παράδειγμα:

```
INPUT:
    >Rosalind_12
    GACTCCTTTGTTTGCCTTAAATAGATACATATTTACTCTTGACTCTTTT...
    ...GTTGGCCTTAAATAGATACATATTTGTGCGACTCCACGAGTGATTCGTA
    >Rosalind_37
    ATGGACTCCTTTGTTTGCCTTAAATAGATACATATTCAACAAGTGTGCA...
    ...CTTAGCCTTGCCGACTCCTTTGTTTGCCTTAAATAGATACATATTTG
OUTPUT:
    The best non-identical alignments are:     ls-w bits E(1) %_id  %_sim  alen
    Rosalind_37                    (  96) [f]  465 35.8 1.6e-07 0.763 0.774   93
    +-                                         308 19.1   0.017 0.549 0.593   91
    +-                                         252 13.1    0.65 0.476 0.563  103
    +-                                         244 12.3    0.85 0.489 0.564   94
    +-                                         235 11.3    0.98 1.000 1.000   34
    Rosalind_37                    (  96) [r]  229 10.7       1 0.442 0.526   95
```

1.1.11 "Suboptimal Local Alignment": Lalign

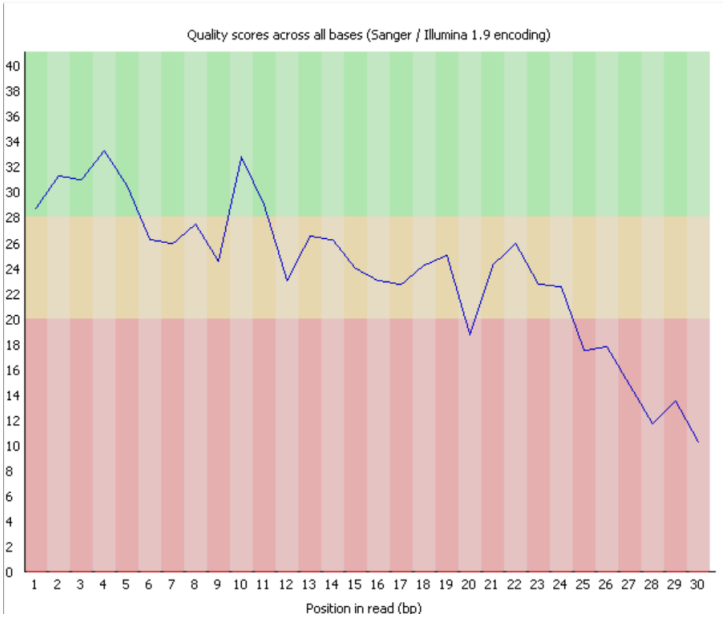
Το εργαλείο Lalign [11] βρίσκει επαναλαμβανόμενες εσωτερικές ακολουθίες νουκλεοτιδίων ή πρωτεϊνών, στοιχίζοντας ξένες υπακολουθίες, ψάχνοντας ομοιότητες. Για παράδειγμα:

```
INPUT:
    >Rosalind_48
    GCATA
OUTPUT:
    >Rosalind_48 reverse complement
    TATGC
```

1.1.12 "Base Quality Distribution": Per Base Sequence Quality

Το FastQC [**FastQC**] εμφανίζει διάγραμμα με τη μετρική Base Call Quality. Για παράδειγμα:

```
INPUT:
    @Rosalind_0029
    GCCCCAGGGAACCCTCCGACCGAGGATCGT
    +
    >?F?@6<C<HF?<85486B;85:8488/2/
    @Rosalind_0029
    TGTGATGGCTCTCTGAATGGTTCAGGCAGT
    +
    @J@H@>B9:B;<D==:<;:,<::?463-,,
    @Rosalind_0029
    CACTCTTACTCCCTAGCCGAACTCCTTTTT
    +
    =88;99637@5,4664-65)/?4-2+)$)$
    @Rosalind_0029
    GATTATGATATCAGTTGGCTCCGAGAGCGT
    +
    <@BGE@8C9=B9:B<>>>7?B>7:02+33.
```



Σχήμα 1.2: Διάγραμμα μετρικής Base Call Quality.

1.1.13 "Global Multiple Alignment": Clustal

Το πρόγραμμα Clustal πραγματοποιεί ολική στοίχιση ακολουθιών.

```
OUTPUT:
    Rosalind_7       -----CACGTCTGTTCGCCTAAAACTTTGATTGCCGGCCTACGCTAGTTAGTTA 49
    Rosalind_28      GGGGTCATGGCTGTTTGCCTTAAACCCTTGGCGGCCTAGCCGTAATGTTT---- 50
    Rosalind_51      --TCCTATGTTTGTTTGCCTCAAACTCTTGGCGGCCTAGCCGTAAGGTAAG--- 49
    Rosalind_18      ---GACATGTTTGTTTGCCTTAAACTCGTGGCGGCCTAGCCGTAAGTTAAG--- 48
    Rosalind_23      --ACTCATGTTTGTTTGCCTTAAACTCTTGGCGGCTTAGCCGTAACTTAAG--- 49
                           * *  **** **** ****       * *           *
```

1.1.14 "Finding Genes with ORFs":

Το OTF Finder του SMS 2 ξεχωρίζει το κωδικόνιο έναρξης και λήξης και επιστρέφει τη μεγαλύτερη πρωτεϊνική ακολουθία. Για παράδειγμα:

```
INPUT:
    >sample sequence
    gcggcggcggcggcggcggcggcggcggcggcggcggcggcggcggcggcggcggcggcggcggcggcggcggcg...
OUTPUT:
>Translation of ORF number 1 in reading frame 1 on the direct strand.
AAAAAAAAAAAAAAAAAAAAAAAAAAAAA*
```

1.1.15 "Base Filtration by Quality":

Το FASTQ Quality Filter του Galaxy μπορεί να χρησιμοποιηθεί για να καθαρίσουμε τα entries που δεν ικανοποιούν κάποιο threshold. Για παράδειγμα για window size 20:

```
INPUT:
    @Rosalind_0049
    GCAGAGACCAGTAGATGTGTTTGCGGACGGTCGGGCTCCATGTGACACAG
    +
    FD@@;C<AI?4BA:=>C<G=:AE=><A??>764A8B797@A:58:527+,
OUTPUT:
    @Rosalind_0049
    GCAGAGACCAGTAGATGTGTTTGCGGACGGTCGGGCTCCATGTGACAC
    +
    FD@@;C<AI?4BA:=>C<G=:AE=><A??>764A8B797@A:58:527
```

1.2 ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ NCBI & EBI

Η βάση δεδομένων NCBI (National Center for Biotechnology Information) [14] χρησιμοποιεί το COBALT [5] ως εργαλείο πολλαπλής στοίχισης. ΤΟ COBALT (Constraint-Based Multiple Alignment Tool) λαμβάνει motif και ομοιότητες από υπάρχουσες βάσεις δεδομένων, τα οποία μετά αξιοποιεί για τη στοίχιση των ακολουθιών. Είναι πιο αποτελεσματικό σε συγκεκριμένα είδη πρωτεϊνών, είναι πιο υπολογιστικά απαιτητικό γενικά.

Αντίθετα, η βάση δεδομένων EBI [7] (European Bioinformatics Institute) χρησιμοποιεί το Cluster Omega [4]. Το Cluster Omega είναι εξαιρετικά γρήγορο και ευέλικτο καθώς χρησιμοποιεί προοδευτική σύγκριση (progressive alignment) για την κατασκευή πολλαπλών στοιχίσεων. Αυτό επιτυγχάνεται με τη δημιουργία ιεραρχικών δομών (guide trees) που αναπαριστούν τις ομοιότητες μέσα στις ακολουθίες, οι οποίες στη συνέχεια στοιχίζονται. Μπορεί να στοιχίσει ταυτόχρονα πολλαπλές ακολουθίες, προσφέροντας υψηλή ακρίβεια και κλιμακωτή απόδοση.

2 ΕΡΩΤΗΜΑ 2

Μετά από αναζήτηση, χρησιμοποιούμε τις εξής αλυσίδες φεριττίνης:

```
>AAH13928.1 Ferritin, light polypeptide [Homo sapiens]
    MSSQIRQNYSTDVEAAVNSLVNLYLQASYTYLSLGFYFDRDDVALEGVSHFFRELAEEKREGYERLLKMQNQRGGRALFQ
    DIKKPAEDEWGKTPDAMKAAMALEKKLNQALLDLHALGSARTDPRLCDFLETHFLDEEVKLIKKMGDHLTNLHRLGGPEA
    GLGEYLFERLTLKHD
>NP_001126850.1 ferritin light chain [Pongo abelii]
    MSSQIRQNYSTDVEAAVNSLVNMYLQASYTYLSLGFYFDRDDVALEGVSHFFRELAEEKREGYERLLKMQNQRGGRALFQ
    DIKKPAEDEWGKTPDAMKAAMALEKKLNQALLDLHALGSAHTDPHLCDFLETHFLDEEVKLIKKMGDHLTNLHRLGGPEA
    GLGEYLFERLTLKHD
>XP_063672238.1 ferritin light chain-like [Pan troglodytes]
    MFWQFGGPAGLSLASTVFGRNRSGDSLPASDRPPISSPLATSGTIFSAISCFWDLPAPFLWLAPSCQPTMSSQIRQNYST
    DVEAAVNSLVNLYLQASYTYLSLGFYFDRDDVALEGVSHFFRELAEEKREGYERLLKMQNQRGGRALFQDIKKPAEDEWG
    KTPDAMKAAMALEKKLNQALLDLHALGSAHTDPHLCDFLETHFLDEEVKLIKKMGDHLTNLHRLGGPEAGLGEYLFERLT
    LKHD
```

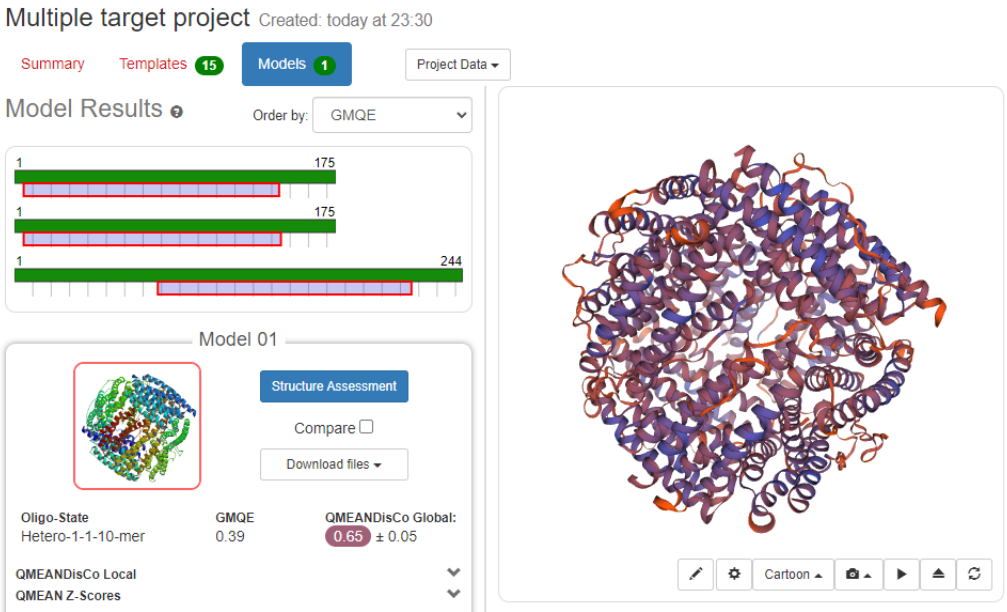
2.1 ΣΤΟΙΧΙΣΗ ΑΚΟΛΟΥΘΙΩΝ

Για τη στοίχιση των ακολουθιών χρησιμοποιούμε το εργαλείο T-COFFEE [18].

Σχήμα 2.1: Αποτελέσματα εργαλείου T-COFFEE.

Βλέπουμε πως υπάρχει μια πολύ καλή βαθμολογία (99), το οποίο σημαίνει πως υπάρχει μεγάλη ομοιότητα ανάμεσα στις ακολουθίες.

2.2 ΣΥΓΚΡΙΣΗ ΔΟΜΩΝ



Σχήμα 2.2: Αποτελέσματα εργαλείου swiss-modeller.

Αφού εξάγουμε το αρχείο `.pdb` μέσω του swiss-modeller, συγκρίνουμε τις δομές χρησιμοποιώντας το Dali.

```
1287:  8jb0-F 15.3  2.1  128   161   13   MOLECULE: BACTERIOFERRITIN;
1288:  8jb0-T 15.3  2.0  128   162   14   MOLECULE: BACTERIOFERRITIN;
1289:  8jb0-S 15.3  2.1  129   161   13   MOLECULE: BACTERIOFERRITIN;
1290:  2pyb-A 15.3  1.9  131   151    7   MOLECULE: NEUTROPHIL ACTIVATING PROTEIN;
1291:  6zlq-Z 15.2  2.3  136   174   75   MOLECULE: FERRITIN;
1292:  6zlq-G 15.2  2.3  136   174   75   MOLECULE: FERRITIN;
1293:  6job-B 15.2  2.4  137   172   50   MOLECULE: FERRITIN HEAVY CHAIN;
1294:  6job-A 15.2  2.4  137   172   50   MOLECULE: FERRITIN HEAVY CHAIN;
1295:  5obb-F 15.2  2.4  136   174   49   MOLECULE: FERRITIN HEAVY CHAIN;
1296:  1xz1-A 15.2  2.4  137   168   82   MOLECULE: FERRITIN LIGHT CHAIN;
```

3 ΕΡΩΤΗΜΑ 3

Τα γενικευμένα δέντρα επιθεμάτων (generalized suffix trees) επιτρέπουν την αποθήκευση και την αναζήτηση πολλαπλών συμβολοσειρών, εν αντιθέσει με τα δέντρα επιθεμάτων (suffix trees) που αφορούν μια συγκεκριμένη συμβολοσειρά. Πρόκειται για μια **στατική δομή δεδομένων**, μιας και κατασκευάζεται για κάποιες συγκεκριμένες συμβολοσειρές που ορίζονται εξ' αρχής. Ως αποτέλεσμα, η δομή δεν έχει σχεδιαστεί για να δέχεται εύκολα τροποποιήσεις, όπως είναι η εισαγωγή νέων συμβολοσειρών ή η διαγραφή υπάρχοντων. Συγκεκριμένα:

3.1 ΠΡΟΒΛΗΜΑΤΑ ΣΤΑΤΙΚΩΝ ΔΕΝΤΡΩΝ

Για να μπορέσει να εισαχθεί μια νέα συμβολοσειρά στο γενικευμένο δέντρο επιθεμάτων, είναι απαραίτητη η ανακατασκευή ολόκληρου του δέντρου μιας και άλλαξε η είσοδος.

Έτσι όλα τα υπάρχοντα μονοπάτια –που αναπαριστούν τα υπάρχοντα επιθέματα– χρειάζεται να ανανεωθούν για να συμβαδίζουν με τις αλλαγές της εισόδου. Παρόμοια ανανέωση απαιτείται και με διαγραφή κάποιας συμβολοσειράς, αφού τροποποιείται και πάλι η είσοδος του δέντρου.

Η αναδιάρθρωση των μονοπατιών από την αρχή μετά από κάθε εισαγωγή και διαγραφή δεν είναι αποδοτική καθώς κάθε φορά είναι απαραίτητο να επαναυπολογιστεί ολόκληρο το δέντρο.

3.2 ΔΥΝΑΜΙΚΟΙ ΑΛΓΟΡΙΘΜΟΙ

Είναι σαφές ότι είναι απαραίτητος ένας δυναμικός τρόπος διαχείρισης της δομής, ώστε να μη χρειάζεται η ανακατασκευή όλων των μονοπατιών κάθε φορά που αλλάζει η είσοδος του δέντρου, αλλά παρά μόνο των μονοπατιών που επηρεάζονται.

3.2.1 Δυναμικό δέντρο επιθεμάτων του McCreight

Ο McCreight προτείνει έναν νέο αλγόριθμο [12], ο οποίος κατασκευάζει το δέντρο επιθεμάτων σταδιακά (Algorithm M): Ξεκινάει από τη ρίζα και το πρώτο χαρακτήρα της συμβολοσειράς. Για κάθε επόμενο χαρακτήρα επεκτείνει το δέντρο προσθέτοντας τα επιθέματα που ξεκινούν από αυτόν τον χαρακτήρα, με τελική πολυπλοκότητα $O(n)$. Κατ' αυτόν τον τρόπο, δεν είναι απαραίτητη η πρότερη γνώση όλων των συμβολοσειρών, συντελώντας σε μια *κάπως* δυναμική μορφή δέντρου, από την άποψη ότι δε χρειάζεται η πρωτύτερη γνώση ολόκληρης της εισόδου για να ξεκινήσει η εισαγωγή των επιθεμάτων.

3.2.2 Δυναμικό δέντρο επιθεμάτων των Choi - Lam

Οι Choi - Lam προτείνουν μια νέα δυναμική υλοποίηση για το δέντρο επιθεμάτων. [3] Κατά την εισαγωγή μιας νέας συμβολοσειράς, το δέντρο ψάχνει το μεγαλύτερο επίθεμα και μετά προσθέτει τα νέα επιθέματα κάνοντας τις απαραίτητες μεταβολές στις ακμές και στους κόμβους. Στην εισαγωγή χρησιμοποιείται μια ιδέα

που παρουσίασε και ο McCreight, τα suffix links.

Αντίστοιχα κατά τη διαγραφή μιας συμβολοσειράς s, αναγνωρίζονται και διαγράφονται οι ακμές / επιθέματα που σχετίζονται με το s και ο αλγόριθμος ανανεώνει τις ετικέτες των φύλλων τους. Το αποτέλεσμα είναι σε κάθε περίπτωση να επανακατασκευάζεται το δέντρο ακέραια, επιτρέποντας τη δυναμική προσθήκη και διαγραφή συμβολοσειρών σε $O(nlogA)$ χρόνο.

3.2.3 Online αλγόριθμος του Ukkonen

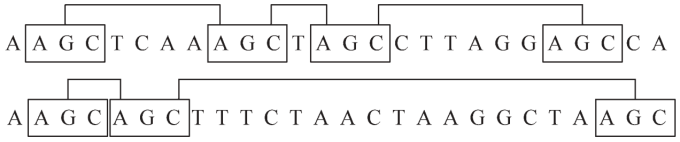
Ο Ukkonen προτείνει έναν νέο αλγόριθμο κατασκευής δέντρων επιθεμάτων σε γραμμικό χρόνο. [19] Ο αλγόριθμος διαχειρίζεται την εισαγωγή και διαγραφή των συμβολοσειρών με έναν τρόπο που επιτρέπει την ανανέωση του δέντρου χωρίς να χρειάζεται η ανακατασκευή όλων των μονοπατιών.

Για κάθε νέο χαρακτήρα που εισάγεται, ο αλγόριθμος ανανεώνει το δέντρο επιθεμάτων επεκτείνοντας τα υπάρχοντα επιθέματα με τον νέο χαρακτήρα. Όταν διαγράφεται μια συμβολοσειρά, ο αλγόριθμος διασχίζει το δέντρο για να εντοπίσει και να διαγράψει τους κόμβους που αντιστοιχούν στα επιθέματα του διαγραμμένου χαρακτήρα. Αυτά επιτυγχάνονται σε γραμμικό χρόνο.

3.2.4 LCP αλγόριθμος των Cole - Hariharan

Τέλος οι Cole - Hariharan προτείνουν έναν LCP (Longest Common Prefix) αλγόριθμο για το δέντρο επιθεμάτων, με $O(\log n)$ χειρότερο χρόνο για εισαγωγές και διεγραφές. [6]

4 ΕΡΩΤΗΜΑ 4

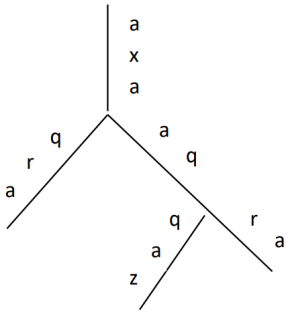


Ένα **Γενικευμένο Δέντρο Επιθεμάτων** περιέχει όλα τα suffixes από πολλαπλές συμβολοσειρές. Κάθε κόμβος του δέντρου αναπαριστά ένα κοινό suffix από αυτές τις συμβολοσειρές, και τα φύλλα πληροφορίες για τις συμβολοσειρές και τη θέση του επιθέματος μέσα στην συμβολοσειρά.

Επομένως, για την εύρεση επαναλήψεων, διασχίζουμε το δέντρο και συλλέγουμε τους κόμβους/φύλλα. Αν ένας κόμβος περιλαμβάνει k φορές το ίδιο επίθεμα, τότε αυτή η συμβολοσειρά επαναλαμβάνεται σε διαφορετικές ακολουθίες. Σε περίπτωση που υπάρχουν περιορισμοί στα κενά, τότε θα λάβουμε υπόψη μας και τις θέσεις των επιθεμάτων, όπως είναι αποθηκευμένες στα φύλλα.

Με χρήση του **K-mers** αλγορίθμο μπορούμε να σπάσουμε τη συμβολοσειρά σε μικρότερα κομμάτια μήκους k, όπου k το μέγεθος της συμβολοσειράς που επαναλαμβάνεται, και να συγκρίνουμε τα σπασμένα κομμάτια της με την επαναλαμβανόμενη. Για πιο αποδοτική σύγκριση μπορούμε να χρησιμοποιήσουμε κάποιο hashtable. Στην περίπτωση που υπάρχουν περιορισμοί στα κενά μπορούμε να προσπεράσουμε κάποιες συγκρίσεις συμβολοσειρών όταν έχουμε βρει match.

5 ΕΡΩΤΗΜΑ 5



Έστω πρότυπο P και κείμενο T. Ο μόνος αλγόριθμος που μπορεί να τρέξει realtime, δηλαδή σε $O(n+m)$ χρόνο, όπου n ο συνολικός αριθμός χαρακτήρων του T και m το μήκος του προτύπου P, είναι ο **Knuth-Morris-Pratt** αλγόριθμος.

Εφαρμόζουμε τον αλγόριθμο αφού πρώτα διασχίσουμε το δέντρο, όπου για κάθε χαρακτήρα $\in P$ ακολουθούμε τις ακμές που αντιστοιχούν σε αυτό. Μπορεί να χρησιμοποιηθεί κάποιος DFS αλγόριθμος για την εύρεση όλων των εμφανίσεων.

6 ΕΡΩΤΗΜΑ 6

6.1 ΟΛΙΚΗ ΣΤΟΙΧΙΣΗ

Οι τιμές του $D(i,j)$ ορίζονται ως

$$D(i,j) = \min\{D(i-1,j)+1, D(i,j-1)+1, D(i-1,j-1)+t(i+j))\}$$

όπου $t(i,j)=1$ αν υπάρχει συμφωνία και $t(i,j)=0$ αν υπάρχει ασυμφωνία.

Ο πίνακας δυναμικού προγραμματισμού είναι ο εξής:

D		G	U	G	T	T	G	T	G	G
	0	←1	←2	←3	←4	←5	←6	←7	←8	←9
T	↑1	↖1	↖←2	↖←3	↖3	↖←4	←5	↖←6	←7	←8
C	↑2	↖↑2	↖2	↖←3	↖←↑4	↖4	↖←5	↖←6	↖←7	↖←8
G	↑3	↖2	↖←↑3	↖2	←3	←4	↖4	←5	←6	←7
T	↑4	↑3	↖3	↑3	↖2	↖←3	←4	↖4	←5	←6
G	↑5	↖4	↖↑4	↖3	↑3	↖3	↖3	←4	↖4	↖←5
A	↑6	↑5	↖↑5	↑4	↖↑4	↖↑4	↖↑4	↖4	↖←↑5	↖5
A	↑7	↑6	↖↑6	↑5	↖↑5	↖↑5	↖↑5	↖↑5	↖5	↖←↑6
T	↑8	↑7	↖↑7	↑6	↖5	↖5	↖←↑6	↖5	↖←↑6	↖6
T	↑9	↑8	↖↑8	↑7	↖↑6	↖5	↖←6	↖↑6	↖6	↖←↑7

Η τιμή της βέλτιστης ολικής στοίχισης είναι 7. Μπορούμε να ακολουθήσουμε πολλά διαφορετικά μονοπάτια οπισθοδρόμησης. Για το σκιασμένο προκύπτει η εξής στοίχιση:

```
G U G T T G T G G
    | |
T C G T G A A T T
```

6.2 ΤΟΠΙΚΗ ΣΤΟΙΧΙΣΗ

Οι τιμές του $v(i,j)$ ορίζονται ως

$$v(i,j) = \max\{v(i-1,j) + s(S_1(i), \_), v(i,j-1) + s(\_, S_2(j)), v(i-1,j-1) + s(S_1(i), S_2(j)), 0\}$$

v		G	U	G	T	T	G	T	G	G
	0	0	0	0	0	0	0	0	0	0
T	0	0	0	0	↘1	↘1	0	↘1	0	0
C	0	0	0	0	0	0	0	0	0	0
G	0	↘1	0	↘1	0	0	↘1	0	↘1	↘1
T	0	0	0	0	↘1	↘1	0	↘2	→1	0
G	0	↘1	0	↘1	0	0	↘2	→1	↘3	↘→2
A	0	0	0	0	0	0	↘1	↘1	↓2	↘2
A	0	0	0	0	0	0	0	0	↓1	↘↓1
T	0	0	0	0	0	↘1	↘1	0	↘1	0
T	0	0	0	0	0	↘1	↘2	→1	↘1	0

Η τιμή της βέλτιστης τοπικής στοίχισης είναι 3.

```
G U G T T G T G G
      |   | | |
      T C G T G A A T T
```