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. [14] 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. [9] 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 με ορισμένο header, τα χαρακτηριστικά της ακολουθίας και την ίδια την ακολουθία. Μέσω του εργαλείου GenBank to Fasta του SMS 2 [15] μπορούμε να αντιγράψουμε κάποιο entry από το GenBank και να το μετατρέψουμε σε FASTA, την κλασική αναπαράσταση νουκλεοτιδίων. 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) [12], εισάγοντας ακολουθίες που περιλαμβάνει motif (δηλαδή ένα επαναλαμβανόμενο μοτίβο), εξάγεται η κανονική έκφραση του συγκεκριμένου motif.

1.1.5 "Pairwise Global Alignment": Στοίχιση ακολουθιών

Στο εργαλείο Needle [7] μπορούμε να εισάγουμε τα ID από δύο GenBank entries. 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
```

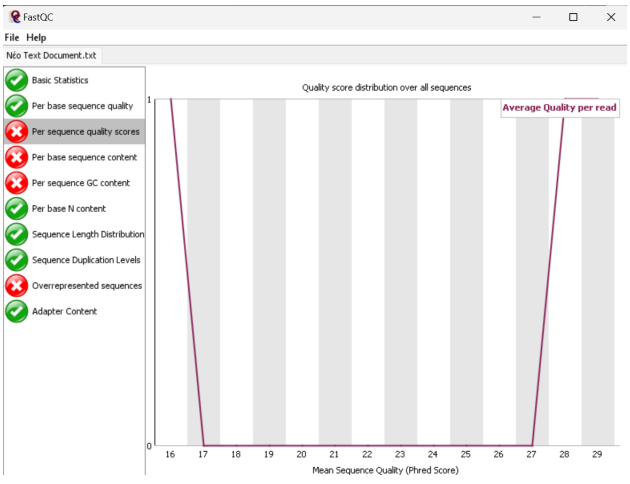
1.1.6 "FASTQ format introduction": Μετατροπή FASTQ σε FASTA

Ένα FASTQ αρχείο είναι μια μορφή αρχείου που αποθηκεύει μια ακολουθία και επιπλέον πληροφορία για αυτή (quality scores). Υπάρχουν διαφορετικοί online convertors που μπορούν να το μετατρέψουν σε FASTA, όπως ο Sequence Conversion της Bugaco, [1] 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
```



1.1.8 "Protein Translation": SMS 2 Translate

Μέσω του εργαλείου Translate του SMS 2 [16], 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” entries χρησιμοποιώντας συγκεκριμένο threshold (quality cut-off value και ποσοστό entries που να ικανοποιούνται από αυτό) χρησιμοποιώντας το FASTQ Quality Filter της Galaxy. [8] 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

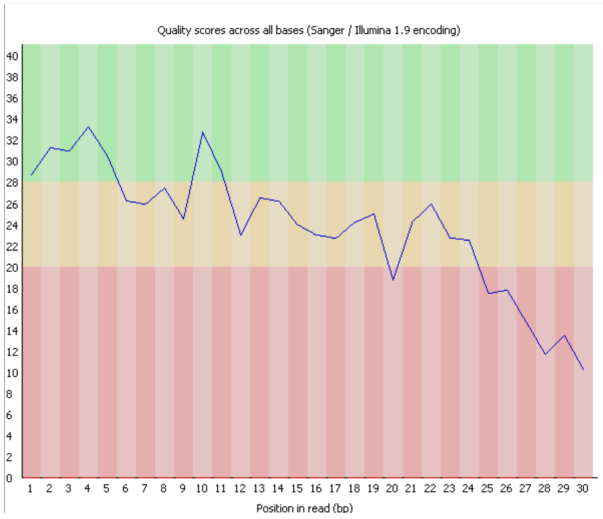
Το εργαλείο Lalign [10] βρίσκει επαναλαμβανόμενες εσωτερικές ακολουθίες νουκλεοτιδίων ή πρωτεϊνών, στοιχίζοντας ξένες υπακολουθίες ψάχνοντας ομοιότητες. Για παράδειγμα:

```
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.
```



1.1.13 "Global Multiple Alignment": Clustal

Το πρόγραμμα Clustal πραγματοποιεί 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) [13] χρησιμοποιεί το COBALT [4] ως εργαλείο πολλαπλής στοίχισης. ΤΟ COBALT (Constraint-Based Multiple Alignment Tool) χρησιμοποιεί motif μοτίβα και ομοιότητες από υπάρχουσες βάσεις δεδομένων, τα οποία μετά αξιοποιεί για τη στοίχιση των ακολουθιών. Είναι πιο αποτελεσματικό σε συγκεριμένα είδη πρωτεϊνών.

Αντίθετα, η βάση δεδομένων EBI [6] (European Bioinformatics Institute) χρησιμοποιεί το Cluster Omega [3]. Το Cluster Omega είναι εξαιρετικά γρήγορο και ευέλικτο καθώς χρησιμοποιεί ιεραρχικές δομές (guide trees) που αναπαριστούν τις συσχετίσεις μέσα στην ακολουθία. Μπορεί να στοιχίσει ταυτόχρονα πολλαπλές ακολουθίες, έχοντας ως αποτέλεσμα τον εντοπισμό διατηρημένων περιοχών σε διαφορετικές αλληλουχίες, προσφέροντας υψηλή ακρίβεια και κλιμακωτή απόδοση.

2 ΕΡΩΤΗΜΑ 2

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 [17].

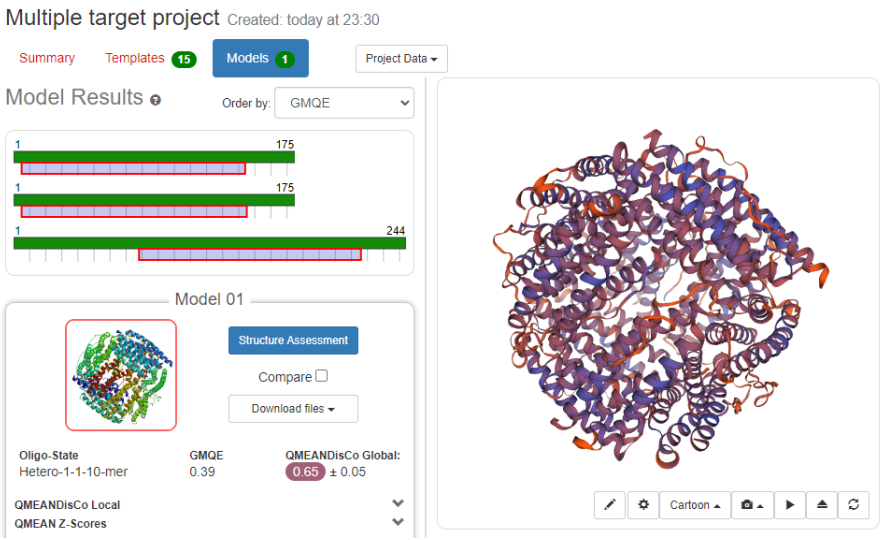
MSA

The multiple sequence alignment result as produced by T-coffee.

```
T-COFFEE, Version_11.00 (Version_11.00)
Cedric Notredame
SCORE=999
*
 BAD AVG GOOD
*
AAH13928.1          :  99
NP_001126850.1      :  99
XP_063672238.1      :  99
cons                :  99

AAH13928.1          ------------------------------------------------------------
NP_001126850.1      ------------------------------------------------------------
XP_063672238.1      MFWQFGGPAGLSLASTVFGRNRSGDSLPASDRPPISSPLATSGTIFSAISCFWDLPAPFLWLAPSC

cons                ************************************************************


AAH13928.1          ---MSSQIRQNYSTDVEAAVNSLVNLYLQASYTYLSLGFYFDRDDVALEGVSHFFRELAEEKREGY
NP_001126850.1      ---MSSQIRQNYSTDVEAAVNSLVNMYLQASYTYLSLGFYFDRDDVALEGVSHFFRELAEEKREGY
XP_063672238.1      QPTMSSQIRQNYSTDVEAAVNSLVNLYLQASYTYLSLGFYFDRDDVALEGVSHFFRELAEEKREGY

cons                   ********************:***************************************


AAH13928.1          ERLLKMQNQRGGRALFQDIKKPAEDEWGKTPDAMKAAMALEKKLNQALLDLHALGSARTDPRLCDF
NP_001126850.1      ERLLKMQNQRGGRALFQDIKKPAEDEWGKTPDAMKAAMALEKKLNQALLDLHALGSAHTDPHLCDF
XP_063672238.1      ERLLKMQNQRGGRALFQDIKKPAEDEWGKTPDAMKAAMALEKKLNQALLDLHALGSAHTDPHLCDF

cons                ************************************************:***:****


AAH13928.1          LETHFLDEEVKLIKKMGDHLTNLHRLGGPEAGLGEYLFERLTLKHD
NP_001126850.1      LETHFLDEEVKLIKKMGDHLTNLHRLGGPEAGLGEYLFERLTLKHD
XP_063672238.1      LETHFLDEEVKLIKKMGDHLTNLHRLGGPEAGLGEYLFERLTLKHD

cons                ************************************************
```

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

2.2 ΣΥΓΚΡΙΣΗ ΔΟΜΩΝ



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). 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 [11], ο οποίος κατασκευάζει το δέντρο επιθεμάτων σταδιακά (Algorithm M), προσθέτοντας ένα επίθεμα τη φορά. 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.

Προφανώς, η πρόταση του McCreight δεν αποτελεί μια καθαρόαιμη δυναμική δομή δεδομένων.

3.2.2 Δυναμικό δέντρο επιθεμάτων των Choi - Lam

Οι Choi - Lam προτείνουν μια νέα δυναμική υλοποίηση για το δέντρο επιθεμάτων. [2] Κατά την εισαγωγή μιας νέας συμβολοσειράς, το δέντρο ψάχνει το μεγαλύτερο επίθεμα με παρόμοιο τρόπο όπως περιγράφει ο McCreight, 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 online αλγόριθμο κατασκευής δέντρων επιθεμάτων σε γραμμικό χρόνο. [18] 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. [5]

Βιβλιογραφία

[1] *Bugaco - RNA/RNA Sequence Converter*. URL: https://meme-suite.org/meme/tools/meme.

[2] Y. Choi και T.W. Lam. "Dynamic suffix tree and two-dimensional texts management". Στο: *Information Processing Letters* 61.4 (Φεβ. 1997), σσ. 213–220. DOI: 10.1016/s0020-0190(97)00018-5.

[3] *Cluster Omega*. URL: https://www.ebi.ac.uk/jdispatcher/msa/clustalo.

[4] *COBALT*. URL: https://www.ncbi.nlm.nih.gov/tools/cobalt/re_cobalt.cgi.

[5] Richard Cole και Ramesh Hariharan. "Dynamic LCA queries on trees". Στο: *SIAM Journal on Computing* 34.4 (Ιαν. 2005), σσ. 894–923. DOI: 10.1137/s0097539700370539.

[6] *EBI*. URL: https://www.ebi.ac.uk/.

[7] *EMBOSS Needle*. URL: https://meme-suite.org/meme/tools/meme.

[8] *Galaxy FastQ Quality Filter*. URL: https://usegalaxy.org/root?tool_id=cshl_fastq_quality_filter.

[9] *GenBank*. URL: https://www.ncbi.nlm.nih.gov/genbank/.

[10] *Lalign*. URL: https://www.ebi.ac.uk/jdispatcher/psa/lalign/.

[11] Edward M. McCreight. "A space-economical suffix tree construction algorithm". Στο: *Journal of the ACM* 23.2 (Απρ. 1976), σσ. 262–272. DOI: 10.1145/321941.321946.

[12] *MEME - Multiple Em for Motif Elicitation*. URL: https://meme-suite.org/meme/tools/meme.

[13] *NCBI*. URL: https://www.ncbi.nlm.nih.gov/.

[14] *Sequence Manipulation Suite, Version 2*. URL: https://www.bioinformatics.org/sms2/.

[15] *SMS - GenBank to FASTA*. URL: https://www.bioinformatics.org/sms2/genbank_fasta.html.

[16] *SMS - Translate*. URL: https://www.bioinformatics.org/sms2/translate.html.

[17] *T-Coffee*. URL: https://tcoffee.crg.eu/apps/tcoffee/do:regular.

[18] E. Ukkonen. "On-line construction of suffix trees". Στο: *Algorithmica* 14.3 (Σεπτ. 1995), σσ. 249–260. DOI: 10.1007/bf01206331.