

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ · ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Η/Υ ΚΑΙ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΕΙΣΑΓΩΓΗ ΣΤΗ ΒΙΟΠΛΗΡΟΦΟΡΙΚΗ

ΠΡΩΤΟ ΣΥΝΟΛΟ ΑΣΚΗΣΕΩΝ · 2023–2024

ΠΕΡΙΕΧΟΜΕΝΑ

1	ΕΡΩΤΗΜΑ 1	2
1.1	ΕΡΓΑΛΕΙΑ ΓΙΑ ΧΕΙΡΙΣΜΟ ΠΡΟΒΛΗΜΑΤΩΝ ΒΙΟΠΛΗΡΟΦΟΡΙΚΗΣ	2
1.1.1	"Introduction to the Bioinformatics Armory": {SMS 2}	2
1.1.2	"GenBank Introduction": Αναζήτηση	2
1.1.3	"Data Formats": Formats της GenBank	2
1.1.4	"New Motif Discovery": Αναζήτηση Motifs σε ακολουθίες	2
1.1.5	"Pairwise Global Alignment": Στοιχίση ακολουθιών	2
1.1.6	"FASTQ format introduction": Μετατροπή FASTQ σε FASTA	3
1.1.7	"Read Quality Distribution": Per sequence quality analysis	3
1.1.8	"Protein Translation": SMS 2 Translate	3
1.1.9	"Read Filtration by Quality": FASTQ Quality Filter	4
1.1.10	"Complementing a Strand of DNA": SMS 2 Reverse Complement	4
1.1.11	"Suboptimal Local Alignment": Lalign	4
1.1.12	"Base Quality Distribution": Per Base Sequence Quality	4
1.1.13	"Global Multiple Alignment": Clustal	5
1.1.14	"Finding Genes with ORFs":	5
1.1.15	"Base Filtration by Quality":	5
1.2	ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ NCBI & EBI	6
2	ΕΡΩΤΗΜΑ 2	6
2.1	ΣΤΟΙΧΙΣΗ ΑΚΟΛΟΥΘΙΩΝ	6
2.2	ΣΥΓΚΡΙΣΗ ΔΟΜΩΝ	7
3	ΕΡΩΤΗΜΑ 3	8
3.1	ΠΡΟΒΛΗΜΑΤΑ ΣΤΑΤΙΚΩΝ ΔΕΝΤΡΩΝ	8
3.1.1	Πομπυλοκότητα στην τροποποίηση εισόδου	8
3.1.2	Χρονική πομπυλοκότητα	8
3.2	ΔΥΝΑΜΙΚΟΙ ΑΛΓΟΡΙΘΜΟΙ	8
3.2.1	Δυναμικό δέντρο επιθεμάτων του McCreight	8
3.2.2	Δυναμικό δέντρο επιθεμάτων των Choi - Lam	8

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. [13] 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. [8] 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 [14] 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 fascic (FSCN1) mRNA, complete cds.
acttgaaagtggataaaatcgactgataccaaaacaacattgttttacagaagtggctcgt
ttgaggacatcaacatatatttcacaatgcctgctatgaatttaaaatacaaatgtggcctg
```

1.1.4 "New Motif Discovery": Αναζήτηση Motifs σε ακολουθίες

Με το εργαλείο MEME (Multiple Em for Motif Elicitation) [11], 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 [6] 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 converters 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.

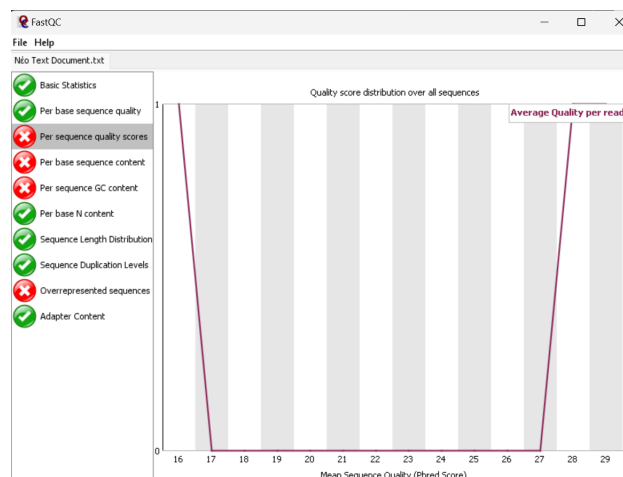
1.1.7 "Read Quality Distribution": Per sequence quality analysis

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

```

INPUT:
@Rosalind_0041
GGCCGGTCTATTTACGTTCTACCCGACGTGACGTACGGTCC
+
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 [15], μπορούμε να μεταφράσουμε την αλληλουχία των νουκλεοτιδίων σε αμινοξέα. Για παράδειγμα:

```

INPUT:
>test
ATGGCCATGGCGCCAGAACTGAGATCAATAGTACCCGTATTAACGGGTGA
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. [7] Εξάγεται το αρχείο Galaxy2-[Filter_by_quality_on_data_1].fastqsanger το οποίο περιλαμβάνει μόνο τα φιλτραρισμένα 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
ATGGACTCCTTTGTTTGCCTTAAATAGATACATATTCACAAGTGTGCA...
...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 [9] βρίσκει επαναλαμβανόμενες εσωτερικές ακολουθίες νουκλεοτιδίων ή πρωτεϊνών, στοιχίζοντας ξένες υποακολουθίες 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_0049
  GCAGAGACCAGTAGATGTGTTTGC GGACGGTCGGGCTCCATGTGACACAG
  +
  FD@@;C<AI?4BA:=>C<G=:AE=><A??>764A8B797@A:58:527+,
OUTPUT:
  @Rosalind_0049
  GCAGAGACCAGTAGATGTGTTTGC GGACGGTCGGGCTCCATGTGACAC
  +
  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) [12] 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 συγκεκριμένα είδη πρωτεϊνών.

Αντίθετα, η βάση δεδομένων EBI [5] (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 φερριτίνης:

```

>AAH13928.1 Ferritin, light polypeptide [Homo sapiens]
  MSSQIRQNYSTDVEAAVNSLVNLYLQASYTYLSLGFYFDRDDVALEGVSHFFRELAEEKREGYERLLKMQNQRGGRALFQ
  DIKKPAEDEWGTKTPDAMKAAMALEKKLNQALLDLHALGSARTDPRLCDFLETHFLDEEVKLIKMGDHLTNLHRLGGPEA
  GLGEYLFERLTLKHD
>NP_001126850.1 ferritin light chain [Pongo abelii]
  MSSQIRQNYSTDVEAAVNSLVNMYLQASYTYLSLGFYFDRDDVALEGVSHFFRELAEEKREGYERLLKMQNQRGGRALFQ
  DIKKPAEDEWGTKTPDAMKAAMALEKKLNQALLDLHALGSAHTDPHLCDFLETHFLDEEVKLIKMGDHLTNLHRLGGPEA
  GLGEYLFERLTLKHD
>XP_063672238.1 ferritin light chain-like [Pan troglodytes]
  MFWQFGGPAGLSLASTVFGNRSGDSLPAASDRPPISSPLATSGTIFSAISCFWDLPAFLWLAPSCQPTMSSQIRQNYST
  DVEAAVNSLVNLYLQASYTYLSLGFYFDRDDVALEGVSHFFRELAEEKREGYERLLKMQNQRGGRALFQDIKKPAEDEWG
  KTPDAMKAAMALEKKLNQALLDLHALGSAHTDPHLCDFLETHFLDEEVKLIKMGDHLTNLHRLGGPEAGLGEYLFERLT
  LKHD

```

2.1 ΣΤΟΙΧΙΣΗ ΑΚΟΛΟΥΘΙΩΝ

Για τη στοίχιση των ακολουθιών χρησιμοποιούμε το εργαλείο T-COFFEE [16].



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:	6z1q-Z	15.2	2.3	136	174	75	MOLECULE: FERRITIN;
1292:	6z1q-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 πολυηχητών 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

3.1.1 Πολυπλοκότητα στην τροποποίηση εισόδου

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

3.1.2 Χρονική πολυπλοκότητα

Η αναδιάρθρωση των μονοπατιών από την αρχή μετά από κάθε εισαγωγή και διαγραφή δεν είναι αποδοτική καθώς κοστίζει $O(n)$ χρόνο ανά τροποποίηση (n : συνολικός αριθμός χαρακτήρων των συμβολοσειρών)

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 αλληλαζει η είσοδος του δέντρου, αλλιώς παρά μόνο των μονοπατιών που επηρεάζονται.

3.2.1 Δυναμικό δέντρο επιθεμάτων των Choi - Lam

Οι Choi - Lam προτείνουν μια νέα υλοποίηση για το δέντρο επιθεμάτων, με την ίδια πολυπλοκότητα με τα στατικά δέντρα, αλλιώς με τη διαφορά ότι το δέντρο πλέον δε διατηρεί αποθηκευμένες συμβολοσειρές που διαγράφονται. [2]

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] *EBI*. URL: <https://www.ebi.ac.uk/>.
- [6] *EMBOSS Needle*. URL: <https://meme-suite.org/meme/tools/meme>.
- [7] *Galaxy FastQ Quality Filter*. URL: https://usegalaxy.org/root?tool_id=cshl_fastq_quality_filter.
- [8] *GenBank*. URL: <https://www.ncbi.nlm.nih.gov/genbank/>.
- [9] *Lalign*. URL: <https://www.ebi.ac.uk/jdispatcher/psa/lalign/>.
- [10] 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.
- [11] *MEME - Multiple Em for Motif Elicitation*. URL: <https://meme-suite.org/meme/tools/meme>.
- [12] *NCBI*. URL: <https://www.ncbi.nlm.nih.gov/>.
- [13] *Sequence Manipulation Suite, Version 2*. URL: <https://www.bioinformatics.org/sms2/>.
- [14] *SMS - GenBank to FASTA*. URL: https://www.bioinformatics.org/sms2/genbank_fasta.html.
- [15] *SMS - Translate*. URL: <https://www.bioinformatics.org/sms2/translate.html>.
- [16] *T-Coffee*. URL: <https://tcoffee.crg.eu/apps/tcoffee/do:regular>.