

Εισαγωγή στην Βιοπληροφορική

Πρώτο Σύνολο Ασκήσεων

Λουδάρος Ιωάννης (1067400) - Χριστίνα Κρατημένου (1067495)



Μπορείτε να δείτε την τελευταία έκδοση του Project εδώ ή σκανάροντας τον κωδικό QR που βρίσκεται στην επικεφαλίδα.

Περιγραφή Αναφοράς

Παρακάτω παραθέτω τις απαντήσεις μου στο “Πρώτο Σύνολο Ασκήσεων” του μαθήματος “Εισαγωγή στην Βιοπληροφορική” καθώς και σχόλια τα οποία προέκυψαν κατά την εκπόνηση του.

Περιεχόμενα

1. Ερώτημα 1	3
Introduction to the Bioinformatics Armory	3
GenBank Introduction	4
Data Formats	5
New Motif Discovery	6
Pairwise Global Alignment	7
FASTQ format introduction	8
Read Quality Distribution	9
Protein Translation	10
Read Filtration by Quality	11
Complementing a Strand of DNA	12
Suboptimal Local Alignment	13
Base Quality Distribution	14
Global Multiple Alignment	15
Finding Genes with ORFs	16
Base Filtration by Quality	17
2. Ερώτημα 2	18
Παρατηρήσεις για την χρήση του εργαλείου	18
Παρατηρήσεις σε σχέση με τα αποτελέσματα	18
3. Ερώτημα 3	19

Υποερώτημα (α)	19
<i>Ο Αλγόριθμος</i>	19
<i>Παρατηρήσεις</i>	19
Υποερώτημα (β)	19
<i>Ο Αλγόριθμος</i>	19
<i>Παρατηρήσεις</i>	20
4. Ερώτημα 4	20
Τα Θέματα που Εντοπίστηκαν	20
Διαθέσιμες Λύσεις	20
5. Ερώτημα 5	22
<i>Ο Αλγόριθμος</i>	22
<i>Παρατηρήσεις</i>	22
5. Ερώτημα 6	23
Υποερώτημα (α)	23
<i>Δημιουργία Πίνακα Βαθμολογίας Στοίχισης</i>	23
<i>Traceback Πίνακα και Παραγωγή στοίχισης</i>	23
<i>Παρατηρήσεις</i>	24
Υποερώτημα (β)	24
<i>Ο Αλγόριθμος</i>	24
<i>Παρατηρήσεις</i>	24
7. Ερώτημα 7	26
Ελέγχουμε τον ισχυρισμό της εκφώνησης	26
Διατύπωση Αλγορίθμου	27
Υλοποίηση	27
8. Ερώτημα 8	28
SARS-CoV-2	28
Bat-RaTG13	28
Μέγιστη κοινή Υποακολουθία	28
9. Ερώτημα 9	29

Απαντήσεις

1. Ερώτημα 1

Introduction to the Bioinformatics Armory

Στο ερώτημα αυτό γίνεται αναφορά του εργαλείου DNA STATS του Sequence Manipulation Suite.

The screenshot shows the Sequence Manipulation Suite DNA Stats interface. On the left, there's a sidebar with various bioinformatics tools like Format Conversion, Combine FASTA, EMBL to FASTA, EMBL Feature Extractor, EMBL Sequence Extractor, Filter DNA, Filter Protein, Filter FASTA, GenBank Feature Extractor, GenBank Sequence Extractor, Sort by Three, Range Extractor DNA, Range Extractor Protein, Reverse Complement, Split Codons, Sort by TAA, Three to One, Window Extractor DNA, Window Extractor Protein, Sequence Analysis, Codon Plot, Codon Usage, CpG Islands, DNA Molecular Weight, DNA Nucleotide Find, DNA Stats, Fuzzy Search DNA, Fuzzy Search Protein, Ident and Sim, Align Two, Alignments, Aligns for Digest, ORF Finder, pairwise Align Codons, pairwise Align DNA, pairwise Align Protein, PCR Products, PCR Products, protein GRAVY, protein Motif, protein Motif Point, protein Molecular Weight, protein Statistics, and protein Statistics End.

The main content area has a title "Sequence Manipulation Suite: DNA Stats". It contains a text area for pasting sequences, with a sample sequence provided:

```
AACCGCTCTCCGACGCCAACCCGAGCGACCGGTGCAATGGATCACACCGAGGGCA  
TTTATGCCATAACCCATCGGATGTCGAAGTGTCTTGGCGGGGAACTGGGGCAO  
CCGATTGGCGTAGTGTGGGCTTGGGATAATTATACCTGATTCTCGGATTCTCCCTTTTA  
GTATCGAAAAAGACTGTCTAGCCAGGGCGAAGACCTTTTGTTGAATAGCTTACCCC  
GATATGGGACTCCTGTATCTCTCTTGGCATATTAGCTCCGCTTGCATA
```

Below the text area are buttons for "Submit", "Clear", and "Reset". A note says: "This page requires JavaScript. See browser compatibility." and "You can mirror this page or use it off-line." At the bottom, it shows "Sun 14 Jun 00:36:59 2020" and "Valid XHTML 1.0; Valid CSS".

On the right, there are links for "new window | home | citation".

Below this, under "DNA Stats results", is a table titled "Results for 875 residue sequence "Untitled" starting "GGCGCGAACAC"". The table lists all possible dinucleotide patterns along with their counts and percentages. The most frequent pattern is "gg" at 5.72%.

Pattern:	Times found:	Percentage:
g	216	24.69
a	220	25.14
t	215	24.57
c	224	25.60
n	0	0.00
u	0	0.00
r	0	0.00
y	0	0.00
s	0	0.00
w	0	0.00
k	0	0.00
m	0	0.00
b	0	0.00
d	0	0.00
h	0	0.00
v	0	0.00
gg	50	5.72
ga	50	5.72
gt	53	6.06
gc	63	7.21
gn	0	0.00
gg	55	6.29
aa	56	6.41
at	56	6.41
ac	52	5.95
an	0	0.00
tg	50	5.72
ta	56	6.41
tt	58	6.64
tc	51	5.84
tn	0	0.00
cg	60	6.86
ca	58	6.64
ct	48	5.49
cc	58	6.64
cn	0	0.00
ng	0	0.00
na	0	0.00
nt	0	0.00
nc	0	0.00
nn	0	0.00
g,c	440	50.29
a,t	435	49.71
r,y,s,w,k	0	0.00
b,h,d,v,n	0	0.00
r,y,s,w,k,m,b,d,h,v,n	0	0.00

Όπως φαίνεται στις εικόνες αριστερά, η χρήση του εργαλείου είναι ιδιαίτερα απλή. Για να το δοκιμάσουμε αρκεί να κατεβάσουμε το Dataset του προβλήματος. Υστερα απλά αντιγράφουμε τα περιεχόμενα του Dataset μέσα στο πλαίσιο κειμένου, πατάμε “submit” και μας επιστρέφεται η ανάλυση που φαίνεται παρακάτω.

Όπως φαίνεται αριστερά, στο συγκεκριμένο Dataset, υπάρχουν:

Βάση	Πλήθος
Αδενίνη	220
Θυμίνη	215
Κυτοσίνη	224
Τουανίνη	216

GenBank Introduction

Σε αυτό το ερώτημα μαθαίνουμε πως να κάνουμε απλά ερωτήματα με περιορισμούς στην GenBank.

Μπορούμε να χρησιμοποιήσουμε την ιστοσελίδα της GeoBank και να κάνουμε την αναζήτηση μας μέσω γραφικού περιβάλλοντος. Μπορούμε επίσης να χρησιμοποιήσουμε το Entrez, μια μηχανή αναζήτησης που μπορεί να χρησιμοποιηθεί μέσω της Biopython.

Μπορούμε να διακρίνουμε στις αντίστοιχες εικόνες την διαδικασία της αναζήτησης. Συγκεκριμένα, το παράδειγμα που διατίθεται από τη ROSALIND επιστρέφει 54 αποτελέσματα.

```
1 from Bio import Entrez
2 Entrez.email = "your_name@your_mail_server.com"
3 handle = Entrez.esearch(db="nucleotide", term='''Zea
mays''[Organism] AND rbcL[Gene] ')
4 record = Entrez.read(handle)
5 record["Count"]
```

Data Formats

Σε αυτό το πρόβλημα εμβαθύνουμε στο [Format](#) που χρησιμοποιούν οι απαντήσεις των ερωτημάτων της GenBank. Επίσης γνωρίζουμε το εργαλείο [GenBank to FASTA](#).

Sequence Manipulation Suite: GenBank to FASTA

GenBank to FASTA accepts a GenBank file as input and returns the entire DNA sequence in FASTA format. Use this program when you wish to quickly remove all of the non-DNA sequence information from a GenBank file.

Paste the contents of one or more GenBank files into the text area below. Input limit is 200,000,000 characters.

```
LOCUS  SUSTFASCIN 2320 bp mRNA INV 14-MAR-2000
DEFINITION Strongylocentrotus purpuratus fascin (FSNC1) mRNA, complete cds.
ACCESSION L12047
VERSION L12047.1 GI:161470
KEYWORDS actin bundling protein; fascin;
SOURCE Strongylocentrotus purpuratus
ORGANISM Strongylocentrotus purpuratus
Eukaryota; Metazoa; Echinodermata; Echinozoa; Echinoidea; Echinacea; Echinoids; Strongylocentrotidae; Strongylocentrotus.
REFERENCE 1 (bases 1 to 2320)
AUTHORS Kane,R.E.
TITLE Actin polymerization and interaction with other proteins in temperature-induced gelation of sea urchin egg extracts
JOURNAL Cell Biol. 71 (3), 704-714 (1976)
MEDLINE 7709152B
REFERENCE 2 (bases 1 to 2320)
AUTHORS Bryan,J. and Kane,R.E.
TITLE Separation and interaction of the major components of sea urchin actin
JOURNAL J. Mol. Biol. 125 (2), 207-224 (1978)
MEDLINE 7909118A
REFERENCE 3 (bases 1 to 2320)
AUTHORS Bryan,J., Edwards,R., Matsudaira,P., Otto,J. and Wulfkuhle,J.
TITLE Fasnc, an echinoid actin-bundling protein, is a homolog of the Drosophila singe gene product
JOURNAL Proc. Natl. Acad. Sci. U.S.A. 90 (19), 9115-9119 (1993)
MEDLINE 94022326
FEATURES
source 1..2320
organism="Strongylocentrotus purpuratus"
/db_xref="taxon:7668"
/dev_stage="larva"
gene "FSNC1"
/genes="FSNC1"
CDS B5..1575
/genomic_start="5"
function="actin bundling protein"
/note="putative"
/codon_start=1
```

*This page requires JavaScript. See browser compatibility.
You can mirror this page or use it off-line.

[new window](#) | [home](#) | [citation](#)

Sun 14 Jun 00:36:20 2020
Valid XHTML 1.0; Valid CSS

Formal Conversion

- Combine FASTA
- EMBL to FASTA
- EMBL File Extractor
- EMBL to FASTA Extractor
- Filter DNA
- Filter RNA
- GenBank to FASTA
- GenBank Feature Extractor
- GenBank Trans Extractor
- One to Three
- Range Extractor DNA
- Range Extractor Protein
- Reverse Complement
- Split Codons
- Split RNA
- Three to One
- Window Extractor DNA
- Window Extractor Protein

Sequences Analysis

- Codon Plot
- Codon Usage
- Copy Number
- DNA Molecular Weight
- DNA Pattern Find
- DNA Statistics
- Fuzzy Search DNA
- Fuzzy Search Protein
- Identify
- Multi Rev Trans
- Mutate for Digest
- ORF Finder
- Parwises Align Codons
- Parwises Align Proteins
- Protein Coding Region
- Protein Coding Region Protein
- PCR Primer Stats
- Protein Domains
- Protein GRAVY
- Protein Isoelectric Point
- Protein Molecular Weight
- Protein Pattern Find
- Protein States
- Restriction Digest
- Restriction Summary
- Reverse Translate
- Translators

Sequence Figures

- Color Align Conservation
- Color Align Properties
- Group Align
- Group Protein
- Primer Map
- Restriction Map
- Translation Map

Random Sequences

- Mutate
- Mutate Protein
- Random Coding DNA
- Random DNA Sequence
- Random DNA Regions
- Random Protein Sequence
- Random Protein Regions
- Sample DNA
- Sample Protein
- Shuffle DNA
- Shuffle Protein

Miscellaneous

- IUPAC codes
- GenBank

Το Format που χρησιμοποιεί η GenBank μας δίνει πολλές πληροφορίες που πιθανόν να μη χρειαζόμαστε. Το GenBank to FASTA μας επιστρέφει μια επιστρέφει την ακολουθία των βάσεων ώστε να την εκμεταλλευτούμε όπως θέλουμε.

New Motif Discovery

Γνωρίζουμε το εργαλείο MEME (Multiple Em for Motif Elicitation).

The screenshot shows the MEME Suite 5.4.1 interface. The main title is "MEME" with the subtitle "Multiple Em for Motif Elicitation". The version is 5.4.1. On the left, there's a sidebar with various links: Motif Discovery, Motif Enrichment, Motif Scanning, Motif Comparison, Gene Regulation, Manual, Guides & Tutorials, Sample Outputs, File Format Reference, Databases, Download & Install, Help, Alternate Servers, Authors & Citing, Recent Jobs, and a link to Previous version 5.3.3.

Data Submission Form

Perform motif discovery on DNA, RNA, protein or custom alphabet datasets.

Select the motif discovery mode (Classic mode selected)

Select the sequence alphabet (DNA, RNA or Protein selected)

Input the primary sequences

Enter sequences in which you want to find motifs. (Protein)

Select the site distribution

How do you expect motif sites to be distributed in sequences? (Zero or One Occurrence Per Sequence (zoops))

Select the number of motifs

How many motifs should MEME find? (3)

Input job details

(Optional) Enter your email address.

(Optional) Enter a job description.

Advanced options

What should be used as the background model? (0-order model of sequences)

How wide can motifs be? Minimum width: 6 Maximum width: 50

How many sites must each motif have? Minimum sites: 1 Maximum sites: 800

Can motif sites be on both strands? (DNA/RNA only) Search given strand only

Should MEME restrict the search to palindromes? (DNA only) Look for palindromes only

Should MEME shuffle the sequences? Shuffle the sequences

Note: if the combined form inputs exceed 80MB the job will be rejected.

Start Search | Clear Input

Version 5.4.1 | Please send comments and questions to: meme-suite@uw.edu | Powered by Open

Home | Documentation | Downloads | Authors | Citing

Το εργαλείο MEME είναι ένα εργαλείο για να βρίσκουμε μοτίβα μέσα σε συμβολοσειρές. Αφού δώσουμε το input (είτε σε μορφή αρχείου, είτε χρησιμοποιώντας το πλαίσιο κειμένου), το MEME εκτελεί για εμάς ένα “job”. Τα αποτελέσματα αυτού μπορούμε να τα δούμε σε διαφορετικές μορφές, όπως φαίνεται και στην δεύτερη εικόνα. Ενδεικτικά μπορείτε να δείτε πως φαίνονται οι παρακάτω μορφές κάνοντας κλικ στα αντίστοιχα κουμπιά:

MIME HTML Output

MAST HTML Output

Your MEME job is complete. The results should be displayed below.

Job Details ...

Results

- MEME HTML output
- MEME XML output
- MEME text output
- MAST HTML output
- MAST XML output
- MAST text output
- (Primary) Sequences

Status Messages

- Parsing arguments
- Arguments ok
- Starting memo
- memo sequences.fa -protein -oc . -nostatus -time 14400 -mod zoops -nmotifs 3 -minw 6 -maxw 50 -objfun classic -markov_order 0
- memo ran successfully in 0.87 seconds
- Starting mast
- mast memo sequences.fa -oc . -nostatus
- mast ran successfully in 0.15 seconds
- Done

Version 5.4.1 | Please send comments and questions to: meme-suite@uw.edu | Powered by Open

Home | Documentation | Downloads | Authors | Citing

Previous version 5.3.3

Pairwise Global Alignment

Σε αυτό το σημείο, μαθαίνουμε για ένα εργαλείο ολικής στοίχισης ακολουθιών RNA και DNA, το **Needle**.

The screenshot shows the EMBL-EBI Bioinformatics Tools website with the URL <https://www.ebi.ac.uk/bioinformatics/tools/pairwise-sequence-alignment>. The page title is "Pairwise Sequence Alignment". It includes a brief description of what pairwise sequence alignment is used for, a note about multiple sequence alignment (MSA), and sections for "Global Alignment" and "Local Alignment". Each section lists tools like Needle (EMBOSS), Stretcher (EMBOSS), and GSEARCH2SEQ, along with their descriptions and links to launch them. A banner at the bottom asks for cookie consent.

Μπορείτε να δείτε παρακάτω τα αποτελέσματα του παραδείγματος που προσφέρει το εργαλείο.

Example job

The screenshot shows the EMBOSS Needle web interface. The title is "EMBOSS Needle". The "Input form" tab is selected. The "Pairwise Sequence Alignment" section contains two text areas for entering protein sequences. Below each area is a "Choose File" button and a link to "Use a example sequence". An "AND" section follows, containing another text area for the second sequence. The "STEP 2 - Set your pairwise alignment options" section includes an "OUTPUT FORMAT" dropdown set to "pair" and a note about default settings. At the bottom, there's a checkbox for "Be notified by email" and a "Submit" button.

FASTQ format introduction

To FASTQ είναι άλλο ένα format για αναπαράσταση βιολογικών ακολουθιών. Εδώ βλέπουμε εργαλεία για την μετατροπή του FASTQ στο γνωστό μας FASTA.

The screenshot shows the "Online sequence conversion tool" interface. It has a sidebar with "Automatic Sequence Annotation" and "Easy Drag-n-drop Operation". The main area shows "Convert from: abi" to "to: clustal". Below this is a table of formats:

Format	About format
abi	Reads the ABI "Sanger" capillary sequence traces files, including the PHRED quality scores for the base calls. This allows ABI to FASTQ conversion. Note each ABI file contains one and only one sequence (so there is no point in indexing the file).
abi-trim	Same as "abi" but with quality trimming with Mott's algorithm.
ace	Reads the contig sequences from an ACE assembly file. Uses Bio.Sequencing.Ace internally clustal. The alignment format of Clustal X and Clustal W. See also the Bio.Clustalw module.
cif-atom	Uses Bio.PDB.MMCIFParser to determine the (partial) protein sequence as it appears in the structure based on the atomic coordinates.
cif-seqres	Reads a macromolecular Crystallographic Information File (mmCIF) file to determine the complete protein sequence as defined by the .pdbx_poly_seq_scheme records.
clustal	The alignment format of Clustal X and Clustal W.
embl	The EMBL flat file format. Uses Bio.GenBank internally.
fasta	This refers to the input FASTA file format introduced for Bill Pearson's FASTA tool, where each record starts with a '>' line. Resulting sequences have a generic alphabet by default.

Τέτοια εργαλεία είναι για παράδειγμα το sequenceconversion.bugaco.com ή το FASTQ to FASTA του usegalaxy.org.

Αριστερά μπορείτε να δείτε τα δύο αυτά εργαλεία. Μπορείτε επίσης να δείτε ένα παράδειγμα εκτέλεσης κάνοντας κλικ στο παρακάτω κουμπί. Εκεί θα βρείτε το αρχείο εισόδου ως "fastq_example.fq" και το αποτέλεσμα ως "fasta_result.fasta".

The screenshot shows the Galaxy web interface with the "Tools" menu open. Under "GENERAL TEXT TOOLS", "FASTA/FASTQ", and "GENOMICS ANALYSIS", the "FASTQ to FASTA" tool is selected. The tool configuration includes:

- Input FASTQ file:** A batch mode input field with "No fastqsanger.fastqsanger.gz" selected.
- Discard sequences with unknown (N) bases:** Set to "yes".
- Rename sequence names in output file (reduces file size):** Set to "yes".
- Execute:** A button to run the tool.
- What it does:** A description of the tool's function.
- Example:** An example of Solexa-FASTQ data and its conversion to FASTA.

FASTQ to FASTA

Read Quality Distribution

Σειρά έχει ο έλεγχος ποιότητας. Υπάρχουν μετρικές οι οποίες μας δείχνουν την ποιότητα της ακολουθίας μας, ώστε να δούμε αν θα εξάγουμε από αυτή σωστά δεδομένα. Παρακάτω βλέπουμε εικόνες από το εργαλείο FASTQC.

The screenshot shows the Babraham Bioinformatics website. Under 'Download Babraham Bioinformatics Projects', it lists:

- ASAP**: Performing allele-specific alignments in Next-Gen Sequencing samples with mixed genetic background. Includes links to Release Notes, User Guide v0.1.2 (pdf), v0.1.2.tar.gz, and a test dataset.
- Bareback**: A tool to shuffle low complexity sequence to the end of Illumina sequencing runs. Includes links to README and Bareback_v1.0.tar.gz.
- Bismark**: A bisulfite read mapper and methylation caller. Includes links to Release Notes, User Guide v0.21.0, a test dataset, v0.22.3.tar.gz (which includes bismark2summary, deduplicate_bismark, bismark2bedGraph, coverage2cytosine and other source code), and an RRBS Guide (PDF, last updated 25 Jan 2017).

Μπορείτε να δείτε τα αποτελέσματα του Sample Dataset που δίνεται από τη Rosalind, πατώντας παρακάτω:

FASTQC Results

The screenshot shows the FastQC High Throughput Sequence QC Report interface. It features a logo with a stylized 'Q' and text indicating the report version is 0.11.9. Below this, copyright information is provided:

FastQC High Throughput Sequence QC Report
Version: 0.11.9
www.bioinformatics.babraham.ac.uk/projects/
© Simon Andrews, Pierre Lindenbaum, Brian Howard, Phil Ewels 2011–17,
Picard BAM/SAM reader ©The Broad Institute, 2013
BZip decompression ©Matthew J. Francis, 2011
Base64 encoding ©Robert Harder, 2012
Java HDF5 reader ©ETH, CISD and SIS, 2007–14

At the bottom, there is a note: "Use File > Open to select the sequence file you want to check".

Protein Translation

Το πακέτο SMS 2 που είχαμε χρησιμοποιήσει από το πρώτο κιόλας ερώτημα, έχει εργαλείο μετάφρασης που μπορεί να μετατρέψει μια αλληλουχία βάσεων, σε αλληλουχία αμινοξέων. Χρησιμοποιούμε λοιπόν το εργαλείο **Translate**.

SMS

Format Conversion

- Combine FASTA
- EMBL to FASTA
- EMBL Feature Extractor
- EMBL Trans Extractor
- Filter FASTA
- Filter Protein
- GenBank to FASTA
- GenBank Feature Extractor
- GenBank Trans Extractor
- One to Three
- Range Extractor DNA
- Range Extractor Protein
- Reverse Complement
- Split FASTA
- Split FASTA
- Three to One
- Window Extractor DNA
- Window Extractor Protein

Sequence Analysis

- Codon Plot
- Codon Usage
- CpG Islands
- DNA Molecular Weight
- DNA Pattern Find
- DNA Stats
- Fuzzy Search DNA
- Fuzzy Search Protein
- Ident and Sim
- Multi Rev Trans
- Mutate for Digest
- ORF Finder
- Pairwise Align Codons
- Pairwise Align DNA
- Pairwise Align Protein
- PCR Primer Stats
- PCR Products
- Protein GRAVY
- Protein Isoelectric Point
- Protein Molecular Weight
- Protein Pattern Find
- Primer Stats
- Restriction Digest
- Restriction Summary
- Reverse Translates
- Translate

Sequence Figures

- Color Align Conservation
- Color Align Properties
- Group DNA
- Group Protein
- Primer Map
- Restriction Map
- Translation Map
- Translators Sequence

Sequence Manipulation Suite:

Translate

Translate accepts a DNA sequence and converts it into a protein in the reading frame you specify. Translate supports the entire IUPAC alphabet and several genetic codes.

Paste a raw sequence or one or more FASTA sequences into the text area below. Input limit is 200,000,000 characters.

```
>sample sequence  
gckugcgaygatty
```

```
>sample sequence 2  
ggwggggggatggcgcgaggaaatgcacgttgcgtatgtgcgcgcgcgtgcgcggaaatgcggaaaaaa  
taacatgataattatcacacaacttcgtgtatgtgtactatatttttcgtcatctccggcg;
```

- Translate in reading frame 1 on the direct strand.
- Use the standard (1) genetic code.

*This page requires JavaScript. See browser compatibility.
*You can mirror this page or use it off-line.

[new window](#) | [home](#) | [citation](#)

Sun 14 Jun 00:37:01 2020
Valid XHTML 1.0; Valid CSS

΄Οπως βλέπουμε αριστερά, εισάγουμε στο πλαίσιο κειμένου τις αλληλουχίες βάσεων, και μας επιστρέφεται η αλληλουχία των αμινοξέων.

```
Translate results

>rf 1 sample sequence
ACDEF

>rf 1 sample sequence 2
GGGGGEEDDVVVVAAAARRSSKKNNMIIITTTTW*CC**YLLFFSSSSRRRQQHHLLL
LPPPP
```

Read Filtration by Quality

Όταν έχουμε κακή ποιότητα αλληλουχιών, μπορούμε να φίλτράρουμε τα δεδομένα μας ώστε να κρατήσουμε μόνο τα αξιόλογα. Αυτό μπορούμε να το καταφέρουμε μέσω του εργαλείου **FASTQ Quality Filter**.

The screenshot shows the Galaxy web interface. On the left, the 'Tools' sidebar is open, displaying various genomic analysis tools categorized under 'GENERAL TEXT TOOLS', 'GENOMIC FILE MANIPULATION', and 'COMMON GENOMICS TOOLS'. In the center, the 'History' panel shows a workflow: 'sample.fq' is input into the 'Filter by quality on data 1' tool, which then produces 'sample.fq' as output. A success message indicates the filter was executed successfully. The bottom status bar shows 'Using 0%'. The top navigation bar includes links for 'Workflow', 'Visualize', 'Shared Data', 'Help', 'Login or Register', and 'galaxy.org'.

Αριστερά βλέπουμε το εργαλείο, το οποίο χρησιμοποιούμε, τα δεδομένα εισόδου, και τελικά, τα φίλτραρισμένα δεδομένα.

```
sample.fq
@Rosalind_0049_1
GCAGAGACCACTAGATGTGTTGCGGACGGTCGGCTCCATGTGACACAG
+
FD@C;C<AI?4BA:=>C<G=:AE=><A??>764A8B797@A:58:527+,
@Rosalind_0049_2
AATGGGGGGGGAGACAAAATACGGCTAACGGCAGGGTCTTGATGTCT
+
1<<65:793967<4:92568-34:.>1;2752)24')*15;1,.3*3+*!
@Rosalind_0049_3
ACCCCATACGGCGAGCGTCAGCATCTGATATCCTCTTCAATCCTAGCTA
+
B:EI>JDB5=>DA?E6B@CA?C;=@@C:6D:3=@49;@87;::;;?8+
```

```
display
@Rosalind_0049_1
GCAGAGACCACTAGATGTGTTGCGGACGGTCGGCTCCATGTGACACAG
+
FD@C;C<AI?4BA:=>C<G=:AE=><A??>764A8B797@A:58:527+,
@Rosalind_0049_3
ACCCCATACGGCGAGCGTCAGCATCTGATATCCTCTTCAATCCTAGCTA
+
B:EI>JDB5=>DA?E6B@CA?C;=@@C:6D:3=@49;@87;::;;?8+
```

Complementing a Strand of DNA

Σε μια αλυσίδα DNA, γνωρίζουμε ότι οι βάσεις σχηματίζουν συμπληρωματικά ζεύγη. Χρησιμοποιούμε το εργαλείο Reverse Complement του SMS 2 πακέτου, ώστε να πάρουμε το αντεστραμμένο συμπλήρωμα μιας αλυσίδας.

The screenshot shows the SMS web interface with the "Reverse Complement" tool selected. The input sequence is:

```
>Rosalind_64
ATAT
>Rosalind_48
GCATA
```

The output sequence is:

```
• reverse-complement
ATAT
GCATA
```

Χρησιμοποιώντας το Sample Dataset που δίνει η Rosalind, παίρνουμε το αποτέλεσμα που περιμέναμε. Ότι δηλαδή, μια από τις δύο ακολουθίες, ταυτίζεται με το αντεστραμμένο συμπλήρωμα της.

The screenshot shows the bioinformatics.org website displaying the "Reverse Complement results". The results are identical to those shown in the SMS interface:

```
>Rosalind_64 reverse complement
ATAT
>Rosalind_48 reverse complement
TATGC
```

Suboptimal Local Alignment

Μεταξύ δύο ακολουθιών, μπορούμε να βρούμε πολλαπλά εναλλακτικά τοπικά ταιριάσματα με το εργαλείο Lalign.

The screenshot shows the LALIGN tool page. At the top, there's a navigation bar with links for EMBL-EBI Services, Research, Training, Industry, About us, and a search icon. Below that is a sub-navigation bar for Tools > Pairwise Sequence Alignment > LALIGN. The main content area has a teal header "Pairwise Sequence Alignment". It starts with a brief description: "LALIGN finds internal duplications by calculating non-intersecting local alignments of protein or nucleotide sequences." Below this is a section titled "STEP 1 - Enter your protein sequences". It asks for a pair of sequences, with "PROTEIN" selected. A text input field contains the sequence for Rosalind_12: >Rosalind_12 GACTCCTTGTTCGCCTAAATAGATACATATTACTCTTGACTCTTGGCCTAAATAGATACA TTTTGTCGACTCCACCGAGTGATTGTA. There's also a file upload option. Below this is another section for "AND" where the second sequence, Rosalind_37, is entered: >Rosalind_37 ATGGACTCCTTGTTCGCCTAAATAGATACATATTCAACAAAGTGTGCACTTAGCCTGCCGACTCC TTGTTGCCTAAATAGATACATATTG. This section also includes a file upload option. The next section, "STEP 2 - Set your pairwise alignment options", contains a note about default settings and a "More options..." link. The final section, "STEP 3 - Submit your job", has a checkbox for email notifications and a "Submit" button.

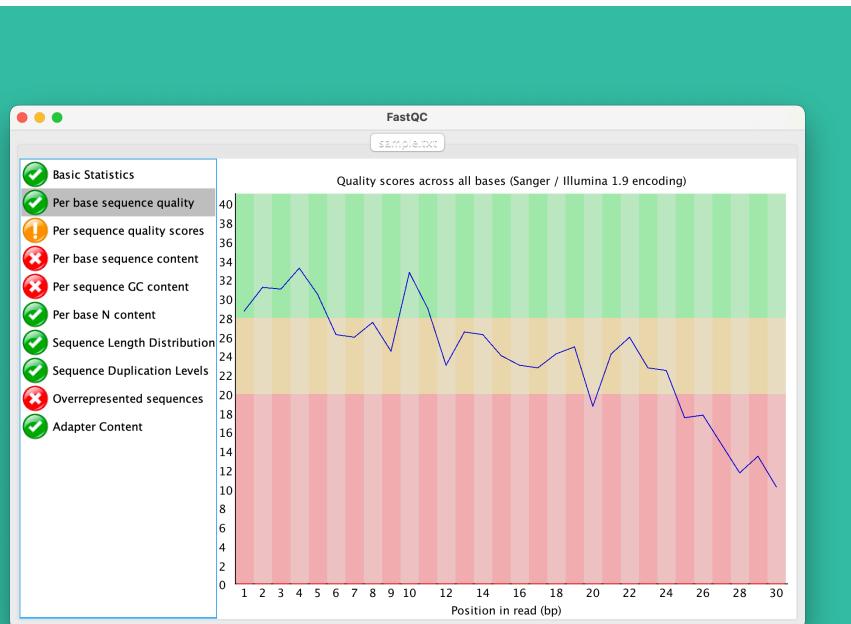
Αριστερά βλέπουμε την χρήση του εργαλείου.

Μπορείτε να δείτε τα αποτελέσματα χρήσης του εργαλείου παρακάτω:

LALIGN Results

Base Quality Distribution

Ανάλυση ποιότητας μπορούμε να κάνουμε και σε επίπεδο εμφανίσεων βάσεων. Χρησιμοποιούμε παρακάτω το εργαλείο FASTQC για άλλη μια φορά.



Μπορείτε παρακάτω να δείτε το sample Dataset της Rosalind καθώς και τα αποτελέσματα του FASTQC.

FASTQC Results

Global Multiple Alignment

Σε αυτό το παράδειγμα χρησιμοποιούμε το Clustal Omega. Ένα εργαλείο για Multiple Sequence Alignment. Είναι συνέχεια του παροχημένου πλέον εργαλείου ClustalW2.

The screenshot shows the Clustal Omega web interface. At the top, there's a navigation bar with links for EMBL-EBI Services, Research, Training, Industry, About us, and Bioinformatics Tools FAQ. Below that is a sub-navigation bar for Tools > Multiple Sequence Alignment > Clustal Omega. The main content area is titled "Multiple Sequence Alignment". It contains a note about Clustal Omega being a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments between three or more sequences. It also mentions that for two sequences, pairwise sequence alignment tools should be used instead. An important note states that the tool can align up to 4000 sequences or a maximum file size of 4 MB. The "STEP 1 - Enter your input sequences" section has a text area where sequences are pasted. The sequences shown are from Rosalind datasets 18, 23, 51, and 7. Below this is a file upload section. The "STEP 2 - Set your parameters" section includes an "OUTPUT FORMAT" dropdown set to "ClustalW with character counts". A note says "The default settings will fulfill the needs of most users." and a "More options..." link. The "STEP 3 - Submit your job" section has a checkbox for "Be notified by email" and a "Submit" button.

Όπως βλέπουμε, παίρνοντας τις ακολουθίες που δίνονται από τη Rosalind και δίνοντας τες στο Clustal Omega, μας επιστρέφεται η πολλαπλή στοίχιση που προκύπτει.

The screenshot shows the Clustal Omega web interface after a job has been submitted. The top navigation bar and sub-navigation bar are identical to the previous screenshot. The main content area is titled "Results for job clustalo-l20220501-233053-0227-74084013-p1m". It features tabs for Alignments, Result Summary, Guide Tree, Phylogenetic Tree, Results Viewers, and Submission Details. The "Alignments" tab is selected. Below the tabs is a "Download Alignment File" button and a "Hide Colors" link. The main content area displays the CLUSTAL W(1.2.4) multiple sequence alignment. The sequences are color-coded, with Rosalind_7 in red, Rosalind_51 in blue, Rosalind_23 in green, Rosalind_18 in orange, and Rosalind_28 in purple. The alignment shows the consensus sequence and the positions of each sequence. At the bottom, a note says "PLEASE NOTE: Showing colors on large alignments is slow."

Finding Genes with ORFs

Από μια αλληλουχία βάσεων RNA μπορούμε να ανιχνεύσουμε κωδικόνια ώστε να περάσουμε σε μια συμβιολοσειρά πρωτεΐνης.

Sequence Manipulation Suite:
ORF Finder

ORF Finder searches for open reading frames (ORFs) in the DNA sequence you enter. The program returns the range of each ORF, along with its protein translation. Use ORF Finder to search newly sequenced DNA for potential protein encoding segments. ORF Finder supports the entire IUPAC alphabet and several genetic codes.

Paste the text into the text area below. Input limit is 100,000,000 characters.

```
AGCCATGTAGCTAAC...  
TTTGGATAAGCCTGAATGATCCGAGTAGCATCTCGAG
```

Submit Clear Reset

• ORFs can begin with: any codon .
 • Search for ORFs in reading frame 1 on the direct strand.
 • Only return ORFs that are at least 30 codons long.
 • Use the standard (1) genetic code.

*This page requires JavaScript. See browser compatibility.
 *You can mirror this page or use it off-line.

Sun 14 Jun 00:36:59 2020
 Valid XHTML 1.0; Valid CSS

[new window](#) | [home](#) | [citation](#)

Sequence Analysis
[-Codon Plot](#)
[-Codon Usage](#)
[-CG Content](#)
[-DNA Molecular Weight](#)
[-DNA Pattern Find](#)
[-DNA Stats](#)
[-DNA Search DNA](#)
[-Fuzzy Search Protein](#)
[-Ident and Sim](#)
[-Multi Rev Trans](#)
[-Nucleic Acid Digest](#)
[-ORF Finder](#)
[-Pairwise Align Codons](#)
[-Pairwise Align DNA](#)
[-Pairwise Align Protein](#)
[-PCR Primer Stats](#)
[-PCR Products](#)
[-Protein GRAVY](#)
[-Protein Isoelectric Point](#)
[-Protein Molecular Weight](#)
[-Protein Pattern Find](#)
[-Protein Stats](#)
[-Protein Digest](#)
[-Restriction Summary](#)
[-Reverse Translate](#)
[-Translate](#)

Sequence Figures
[-Color Align Conservation](#)
[-Color Align Properties](#)
[-Group DNA](#)
[-Group Protein](#)
[-Infernal](#)
[-Restriction Map](#)
[-Translation Map](#)

Random Sequences
[-Mutate DNA](#)
[-Mutate Protein](#)
[-Random Coding DNA](#)
[-Random DNA Sequence](#)
[-Random DNA Regions](#)
[-Random Protein Sequence](#)
[-Random Protein Regions](#)
[-Sample DNA](#)
[-Sample Protein](#)
[-Shuffle DNA](#)
[-Shuffle Protein](#)

Miscellaneous
[-IUPAC codes](#)
[-Genetic codes](#)
[-Browser compatibility](#)
[-Mirror this site](#)
[-Use this site off-line](#)
[-About this site](#)
[-Acknowledgments](#)
[-Reference](#)

Αριστερά βλέπουμε το εργαλείο, το οποίο χρησιμοποιούμε, τα δεδομένα εισόδου, και τελικά, τα κωδικόνια που ανιχνεύτηκαν.

bioinformatics.org

ORF Finder results
Results for 96 residue sequence "Untitled" starting "AGCCATGTAG"

>ORF number 1 in reading frame 1 on the reverse strand extends from base 1 to base 48.
 CTGAGATGCTACTCGATCATTCAAGCTTATTCCAAAAGAGACTCTAA

>Translation of ORF number 1 in reading frame 1 on the reverse strand.
 LRCYSDHSGLFQKRL*

>ORF number 2 in reading frame 1 on the reverse strand extends from base 49 to base 81.
 TCCAAGTCGGGGTCATCCCCATGTAACCTGA

>Translation of ORF number 2 in reading frame 1 on the reverse strand.
 SKSRGHPHV*

>ORF number 3 in reading frame 1 on the reverse strand extends from base 82 to base 96.
 GTTAGCTACATGGCT

>Translation of ORF number 3 in reading frame 1 on the reverse strand.
 VSYMA

Base Filtration by Quality

Εδώ χρησιμοποιούμε το FASTQ Quality Trimmer. Ένα εργαλείο το οποίο τριμάρει την ακολουθία μας σύμφωνα με παραμέτρους που έχουν να κάνουν με την ποιότητα.

The screenshot shows the Galaxy web interface with the following details:

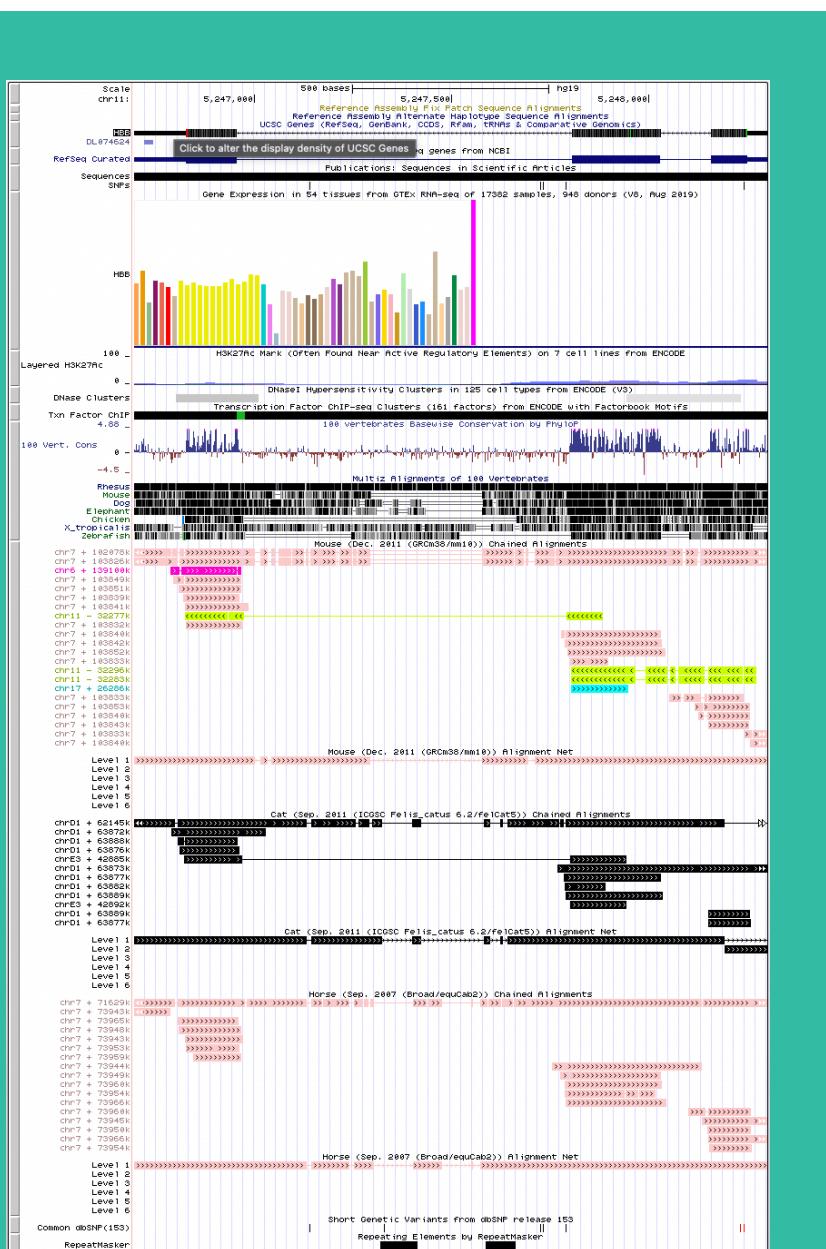
- Header:** Galaxy logo, Workflow, Visualize, Shared Data, Help, User, Notifications, and a progress bar showing "Using 0%".
- Left Sidebar (Tools):**
 - Get Data
 - Send Data
 - Collection Operations
 - GENERAL TEXT TOOLS** (selected)
 - Text Manipulation
 - Filter and Sort
 - Join, Subtract and Group
 - Datamash
 - GENOMIC FILE MANIPULATION**
 - FASTA/FASTQ
 - FASTQ Quality Control
 - SAM/BAM
 - BED
 - VCF/BCF
 - Nanopore
 - Convert Formats
 - Lift-Over
 - COMMON GENOMICS TOOLS**
 - Interactive tools
 - Operate on Genomic Intervals
- Middle Panel (Tool Configuration):** The "FASTQ Quality Trimmer" tool is selected. The configuration form includes:
 - Input: "5: FASTQ Quality Trimmer on data 3"
 - Checkboxes: "Keep reads with zero length" (unchecked), "Trim ends" (unchecked).
 - Dropdowns: "5' and 3'" (selected), "Window size" (set to 1), "Step size" (set to 1).
 - Text input: "Maximum number of bases to exclude from the window during aggregation" (set to 0).
 - Dropdown: "Aggregate action for window" (set to "min score").
 - Text input: "Trim until aggregate score is" (set to ">= 0.0").
 - Checkboxes: "Quality score" (unchecked), "Email notification" (unchecked).
- Right Panel (History):** An "Unnamed history" section shows:
 - 2 shown, 3 deleted
 - 1.64 KB
 - Items:
 - "5: FASTQ Quality Trimmer on data 3" (green background)
 - "3: sample.fq" (green background)

Μπορείτε να δείτε την είσοδο, καθώς και την έξοδο που παράγει το εργαλείο παρακάτω:

FASTQ Quality Trimmer Results

2. Ερώτημα 2

Σε αυτό το ερώτημα πραγματοποιούμε στοίχιση ακολουθιών μέσα από το εργαλείο Genome Browser που δίνεται από τον διαδικτυακό τόπο UCSC.



Παρατηρήσεις για την χρήση του εργαλείου

Παρατηρήθηκε ότι κάνοντας κλικ στον σύνδεσμο Genome Browser από το header της ιστοσελίδας οδηγούσε από προεπιλογή στο hg39. Χρειάστηκε η πλοιόγηση σε αυτό, από το body της ιστοσελίδας ώστε να δοθεί η επιολογία να αλλάξει η έκδοση του Human Assembly.

Παρατηρήσεις σε σχέση με τα αποτελέσματα

Αριστερά μπορούμε να δούμε τα αποτελέσματα της στοίχισης της ακολουθίας ζευγών βάσεων που αποτελούν το γονίδιο HBB στον άνθρωπο, με τις αντίστοιχες ακολουθίες στο άλογο, τη γάτα και το ποντίκι. Πατώντας τα παρακάτω κουμπιά φαίνονται καθαρά τα αποτελέσματα ξεχωριστά για το κάθε είδος που εξετάζοτας.

Όπως μας γίνεται φανερό, υπάρχει μεγαλύτερη αντιστοιχία με το ποντίκι. Κάτι το οποίο είναι αναμενόμενο αφού εξετάζοντας το

“UCSC SPECIES TREE AND CONNECTED ASSEMBLY HUBS”

βλέπουμε ότι ο κοινός μας πρόγωνος είναι πιο πρόσφατος από τον αντίστοιχο που μας συνδέει με το άλογο και την γάτα.

Στοίχιση Άνθρωπος - Άλογο

Στοίχιση Άνθρωπος - Γάτα

Στοίχιση Άνθρωπος - Ποντίκι

3. Ερώτημα 3

Υποερώτημα (α)

Ο Αλγόριθμος

1. Δημιουργούμε το δέντρο επιθεμάτων της συμβολοσειράς T σε $O(|T|)$ χρόνο ($|T|=n$ σύμφωνα με την εκφώνηση)
2. Ξεκινώντας από τη ρίζα, σύγκρινε έναν προς έναν τους χαρακτήρες του P , ακολουθώντας το κατάλληλο μονοπάτι.
 - 2.1. Εάν εμφανιστεί κάποιο μη-ταίριασμα, τότε το πρότυπο δεν εμφανίζεται στην ακολουθία.
 - 2.2. Διαφορετικά,
 - 2.2.1. Αν κάποιο από τα φύλλα που βρίσκονται κάτω από τον κόμβο του τελευταίου χαρακτήρα έχει τιμή $< k$, επέστρεψε **Αληθές**.
 - 2.2.2. Διαφορετικά επέστρεψε **Ψευδές**.

Παρατηρήσεις

Ο αλγόριθμος που μόλις περιγράψαμε απαντάει στο ερώτημα. Σε χρόνο $O(n)$ δημιουργεί το δέντρο επιθεμάτων (στο βήμα 1) και σε χρόνο $O(m)$ εντοπίζει τα φύλλα που περιέχουν τις ενδιαφέρουσες τιμές που θα συγκρίνουμε με το k .

Υποερώτημα (β)

Ο Αλγόριθμος

1. Δημιουργούμε το γενικευμένο δέντρο επιθεμάτων με όλες τις λέξεις του κειμένου, βάζοντας τες με την σειρά που εμφανίζονται στο κείμενο.
2. Φτιάχνουμε μια λίστα A , για τις μικρότερες συμβολοσειρές που εμφανίζονται μόνο μία φορά στο κείμενο.
3. Κατά την προεπεξεργασία σημειώνουμε:
 1. Σε κάθε φύλλο, το βάθος του.
 2. Σε κάθε ενδιάμεσο κόμβο, το βάθος του(v), καθώς και το βάθος φύλλου-απογόνου του με το ελλάχιστο βάθος, σε μια μεταβλητή d .
4. Για κάθε μονοπάτι που ξεκινάει από τη ρίζα, άρχισε να το διαπερνάς.
 1. Σε κάθε κόμβο που συναντάς, να επιλέγεις τον επόμενο που έχει το μικρότερο δυνατό βάθος (άλλα όχι το ίδιο με τον τρέχοντα κόμβο), μέχρι να φτάσεις σε ένα φύλλο.
 2. Μόλις φτάσεις στο φύλλο, αποθήκευσε τις ετικέτες που συνάντησες στο μονοπάτι που πήρες μέχρι τον πρώτο χαρακτήρα της τελευταίας ετικέτας πριν από αυτό το φύλλο, σε μια μεταβλητή a .
 1. Αν η συμβολοσειρά a έχει μήκος ίσο με την πρώτη συμβολοσειρά που ήδη υπάρχει στην λίστα A , τότε πρόσθεσε την a στην A .
 2. Αν η συμβολοσειρά a έχει μήκος μικρότερο από κάποια την πρώτη συμβολοσειρά που ήδη υπάρχει στην λίστα A
 1. Άδειασε την λίστα A
 2. Αποθήκευσε σε αυτή την συμβολοσειρά a .
 3. Άλλιώς, συνέχισε.
5. Τερμάτισε μόλις τερματίσουν όλα τα μονοπάτια που ξεκινάνε από την ρίζα με αυτόν τον τρόπο και επέστρεψε την λίστα A .

Παρατηρήσεις

Προσοχή: Το ερωτηματικό δεν συνεισφέρει στο βάθος κανενός κόμβου.

Ο παραπάνω αλγόριθμος εξυπηρετεί το ζητούμενο του ερωτήματος. Μόλις τελειώσει, θα έχουμε εντοπίσει την μικρότερη συμβολοσειρά (μέτρηση μήκους όπως αναφέρεται στο βήμα 4) που εμφανίζεται μόνο μία φορά στο κείμενο (αυτή την πληροφορία την αντλούμε από τις ετικέτες των φίλων στο βήμα 2).

Αυτό που κάνει ουσιαστικά ο παραπάνω αλγόριθμος είναι:

1. Βρες το μικρότερο μοναδικό προθέματα κάθε επιθέματος.
2. Διάλεξε τα μικρότερα μεταξύ αυτών.

4. Ερώτημα 4

Τα Θέματα που Εντοπίστηκαν

Τα suffix trees είναι εξαιρετικά σημαντικές δομές για την βιοπληροφορική. Εκ πρώτης όψεως όμως, συνειδητοποιούμε ότι συνοδεύονται από κάποια σημαντικά προβλήματα. Κάποια περιγράφονται παρακάτω:

- Ο χειρισμός τους γίνεται δύσκολος όταν χρειάζεται να αφαιρούμε και να προσθέτουμε στοιχεία συχνά. Αν για παράδειγμα αφαιρέσουμε την λέξη “άνθρωπος” (ενώ γνωρίζαμε ότι είναι η πρώτη λέξη που εισάγεται στο δέντρο η οποία περιλαμβάνει το γράμμα “α”) από ένα γενικευμένο suffix tree που περιέχει όλες τις λέξεις τις ελληνικής γλώσσας, τότε πρέπει να ξαναχτιστεί σημαντικά μεγάλο μέρος του δέντρου, αφού το α είναι εξαιρετικά συχνό γράμμα, και άρα, υπάρχει εξαιρετικά μεγάλο πλήθος επιθεμάτων που ξεκινάνε από αυτό.
- Είναι σημαντικά ογκώδεις δομές. Απαιτούν χώρο $O(n \log n)$ για να αποθηκεύσουν ένα string μεγέθους n , ενώ γνωρίζουμε ότι για να αποθηκεύσουμε ένα τέτοιο string χρειαζόμαστε μόνο $n \log n$ bits (όπου σ είναι το μέγεθος του αλφαριθμητικού). Αυτό το γεγονός, σε συνδυασμό με την ανάγκη τα Suffix Trees να βρίσκονται στην κύρια μνήμη ώστε να έχει νόημα η χρήση τους, δημιουργεί πρόβλημα στην εφαρμογή τους σε πολύ μεγάλα δεδομένα.

Διαθέσιμες Λύσεις

Φαίνεται ότι μια πρώτη προσπάθεια να επιλυθούν τα παραπάνω εστιάστηκε στην αποδοτικότερη αποθήκευση τους¹, ενώ αργότερα εμφανίστηκαν δομές οι οποίες προσομοίαζαν τα suffix trees, όπως τα suffix arrays². Η βιβλιογραφεία φαίνεται να καταλήγει σε μια δομή η οποία άρει και τα δύο προβλήματα:

Το Δυναμικό Συμπιεσμένο Δέντρο Επιθεμάτων

Δεν προλάβαμε να αναλύσουμε τις πηγές που βρήκαμε. Παρόλα αυτά, παραθέτουμε την πιο περιεκτική από αυτές.

Dynamic Fully-Compressed
Suffix Trees

¹ Giegerich, R., Kurtz, S., Stoye, J.: Efficient implementation of lazy suffix trees. Softw. Pract. Exper. 33(11), 1035–1049 (2003)

² Manber, U., Myers, E.W.: Suffix arrays: A new method for on-line string searches. SIAM J. Comput. 22(5), 935–948 (1993)

5. Ερώτημα 5

Θα αναπτύξουμε έναν αλγόριθμο που ουσιαστικά εκτελεί DFS με τον επιπλέον περιορισμό ότι αναζητεί τη συμβολοσειρά που ζητείται.

Ο Αλγόριθμος

1. Σε ένα δέντρο επιθεμάτων ξεκίνα από τη ρίζα και εκτέλεσε DFS.
2. Όσο εκτελείται ο DFS συνέκρινε τους χαρακτήρες που σαρώνονται με τους αντίστοιχους ζητούμενους.
3. Κάθε φορά που εντοπίζεται η ζητούμενη υποσυμβολοσειρά, επέστρεφε την ταμπέλα του τελευταίου κόμβου που επισκέφθηκες πριν βρεθεί.

4. Τερμάτισε μόλις ο αλγόριθμος έχει επισκεφθεί όλα τα φύλλα.

Action	Score
Ταίριασμα	1
Προσθαφαίρεση	-ρ
✗ Συνεχόμενες ασυμφωνίες	-(ρ+σχ)

Παρατηρήσεις

Μόλις ολοκληρωθεί ο αλγόριθμος θα έχουμε μια λίστα με τις λέξεις μέσα στο κείμενο όπου εμφανίζουν την ζητούμενη υποσυμβολοσειρά.

Ο αλγόριθμος τηρεί την προϋπόθεση ότι πρέπει να τρέχει σε χρόνο ανάλογο με το συνολικό αριθμό των χαρακτήρων στις ακμές του δέντρου.

5. Ερώτημα 6

Υποερώτημα (a)

Το πρόβλημα που καλούμαστε να αντιμετωπίσουμε είναι μια ειδική περίπτωση Weighted Edit Distance, οπότε μας γίνεται ξεκάθαρο ότι πρέπει να χρησιμοποιήσουμε δυναμικό προγραμματισμό για τον υπολογισμό της βέλτιστης καθολικής στοίχισης. Έστω λοιπόν 2 συμβολοσειρές S_1 και S_2 μεγέθους n , m αντίστοιχα και έστω $V[i, j]$ η τιμή της βέλτιστης στοίχισης των προθεμάτων $S_1[1, i]$ και $S_2[1, j]$.

Ορίζουμε τις παρακάτω συναρτήσεις υπολογισμού του score της στοίχισης, λαμβάνοντας υπόψην τα κόστη/κέρδη στον πίνακα αριστερά.

- $V(i, 0) = -\rho \cdot i$
- $V(0, j) = -\rho \cdot j$
- $V(i, j) = \max [V(i - 1, j) - \rho, \\ V(i, j - 1) - \rho, \\ V(i - 1, j - 1) + s(S_1(i), S_2(j))]$

$$\text{όπου } s(S_1(i), S_2(j)) = \begin{cases} 1, & \text{αν } S_1(i) = S_2(j) \\ -(\rho + \sigma x), & \text{αλλιώς} \end{cases}$$

Τώρα που έχουμε τις συναρτήσεις που θα υπολογίζουν το Score της στοίχισης μας ας σχεδιάσουμε τον αλγόριθμο.

Ο αλγόριθμος μας χωρίζεται σε 2 τμήματα:

1. Δημιουργία Πίνακα Βαθμολογίας Στοίχισης
2. Traceback Πίνακα και Παραγωγή στοίχισης

Δημιουργία Πίνακα Βαθμολογίας Στοίχισης

1. Δημιουργούμε τον Πίνακα Βαθμολογίας Στοίχισης μεγέθους n^m υπολογίζοντας ανά γραμμή όλα τα scores και προσθέτοντας δείκτες, σύμφωνα με το κελί που χρησιμοποιήθηκε για τον υπολογισμό του κάθε score.
2. Καθώς υπολογίζουμε τον πίνακα, σε κάθε κελί, προτού υπολογιστεί το $V(i, j)$, γίνονται τα παρακάτω:

1. Αν για τον υπολογισμό του score χρειάζεται το $V(i-1, j-1)$:
 1. Αν $S_1[i] = S_2[j]$, τότε αποθήκευσε στο κελί $x=0$
 2. Άλλιώς $x = \text{το } x \text{ του κελιού στο οποίο είναι αποθηκευμένο το } V(i-1, j-1) + 1$

Traceback Πίνακα και Παραγωγή στοίχισης

1. Ξεκινώντας από την θέση (n, m) ακολουθούμε τους δείκτες μέχρι την θέση $(0, 0)$
2. Εκτελώντας το Traceback ταυτόχρονα δημιουργούμε τις στοιχίσεις από δεξιά προς τα αριστερά ως εξής:
 - 2.1. Όταν βρισκόμαστε σε κελί $[i, j]$ στο οποίο καταλήγει διαγώνιος δείκτης τότε τοποθετούμε τους χαρακτήρες $S_1[i]$ και $S_2[j]$ τον ένα κάτω από τον άλλο.
 - 2.2. Όταν βρισκόμαστε σε κελί $[i, j]$ στο οποίο καταλήγει οριζόντιος δείκτης τότε τοποθετούμε τους χαρακτήρες _ και $S_2[j]$ τον ένα κάτω από τον άλλο.

2.3.Όταν βρισκόμαστε σε κελί (i,j) στο οποίο καταλήγει κάθετος δείκτης τότε τοποθετούμε τους χαρακτήρες $S_1[i]$ και $_$ τον ένα κάτω από τον άλλο.

Παρατηρήσεις

Action	Weight
Ένθεση / Διαγραφή - d	1
Αντικατάσταση - r	2
Ταίριασμα - e	0

Στον παραπάνω αλγόριθμο η δημιουργία του πίνακα γίνεται σε χρόνο $O(nm)$ ενώ το Traceback σε χρόνο $O(n+m)$. Άρα ο αλγόριθμος εκτελείται πράγματι σε χρόνο $O(nm)$ όπως ζητείται.

Υποερώτημα (β)

Το πρόβλημα που καλούμαστε να λύσουμε είναι συγγενικό του Longest Common Extension, όπου $i=j=1$. Έτσι θα χρησιμοποιήσουμε ένα γενικευμένο suffix tree ώστε να το επιλύσουμε.

Ο Αλγόριθμος

1. Δημιούργησε μια δομή λεξικού που για κλειδιά έχει όλα τα πιθανά ζευγάρια των κ συμβολοσειρών. Σε κάθε τιμή του λεξικού να υπάρχει μια θέση για μια συμβολοσειρά και μια θέση για έναν ακέραιο l. Αρχικοποίησε όλα τα l στο 0.
2. Δημιούργησε ένα γενικευμένο δέντρο επιθεμάτων με δυνατότητα απάντησης ερωτημάτων lcs. Προσοχή, σε αυτή την προεπεξεργασία, στους ενδιάμεσους κόμβους να σημειώνεις τις συμβολοσειρές μόνο αν υπάρχει φύλλο-απόγονος που να δηλώνει ότι σε αυτό τερματίζει το πρώτο επίθεμα της αντίστοιχης συμβολοσειράς. Επίσης να αποθηκεύεις και το βάθος του κόμβου, d.

	-	C	A	T	G	C
-	0	1	2	3	4	5
T	1	2	3	2	3	4
A	2	3	2	3	4	5
T	3	4	3	2	3	4
C	4	3	4	3	4	3

3. Για κάθε συμβολοσειρά:

$$D(4,5) = 5 + 4 - 2 * 2$$

$$3 = 9 - 4$$

$$3 = 5$$

Ψευδές

3.1.Ξεκίνα από το φύλλο του πρώτου επιθέματος της α (του επιθέματος δηλαδή που ξεκινάει από την αρχή της λέξης, της λέξης ολόκληρης), και ταξίδευε προς τη ρίζα.

3.1.1.Σε κάθε ενδιάμεσο κόμβο που θα βρίσκεις που είναι κοινός πρόγονος με άλλη συμβολοσειρά (έστω β), εξέτασε αν το d του κόμβου είναι μεγαλύτερο του l του λεξικού στο κλειδί (α,β).

3.1.1.1.Αν είναι , τότε με κλειδί (α,β) αποθήκευσε το d στο l και το $\alpha[1,d]$

3.2.Πήγαινε στην επόμενη συμβολοσειρά

Παρατηρήσεις

- Για να μην χρειαστεί να χτίσουμε ένα λεξικό με όλα τα πιθανά ζευγάρια, μπορούμε να χρησιμοποιήσουμε μια λίστα, στην οποία θα δημιουργούμε ένα ζευγάρι (l, ζ ευγάρι συμβολοσειρών) μόνο αν όντως εμφανιστεί κοινός πρόγονος μεταξύ τους.
- Ο παραπάνω αλγόριθμος όντως βρίσκει ποιο είναι το κενό πρόθεμα όλων των πιθανών ζευγαριών των κ συμβολοσειρών που δίνονται σαν είσοδο σε χρόνο $O(k^*n+a)$

Python Script

7. Ερώτημα 7

Ας ξεκινήσουμε γράφοντας τις εξισώσεις που θα μας βοηθήσουν στην παραγωγή του πίνακα δυναμικού προγραμματισμού.

$$D(i,0) = i * 1$$

$$D(0,j) = j * 1$$

$$\begin{aligned} D(i,j) = \min[D(i-1,j) + d, \\ D(i,j-1) + d, \\ D(i-1,j-1) + t(i,j)] \end{aligned}$$

Περισσότερες Πληροφορίες

Ελέγχουμε τον ισχυρισμό της εκφώνησης

$$D(n,m) = m + n - 2u$$

$$u = \frac{m + n - D(n,m)}{2}$$

Δοκιμάζουμε να τρέξουμε μερικά παραδείγματα. Και καταλήγουμε σε αντιπαράδειγμα. Όπως βλέπουμε, η μέγιστη κοινή υποακολουθία του παραδείγματος είναι η “AT” και είναι μήκους 2.

Διατύπωση Αλγορίθμου

Παρόλα αυτά, όπως είναι λογικό, μέσα από τον πίνακα μας μπορούμε να δημιουργήσουμε μια στοίχιση η οποία θα περιλαμβάνει την μεγαλύτερη κοινή υποακολουθία των δύο συμβολοσειρών με αυτόν τον τρόπο:

1. Δημιουργησε τον πίνακα δυναμικού προγραμματισμού χρησιμοποιώντας τις εξισώσεις που εμφανίζονται παραπάνω.

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

ion=0, max_position

ελί που πληρεί την προϋπόθεση:

j=0

length=0

G	A	T	_	C	G	T	G	A	A	T	T
G	G	T	T	C	G	T	G	G	A	_	_

πορείτε να δείτε παρακάτω screenshots από

S1 : kalimera
S2 : oli mera kai oli nixta
LCS: mera

S1 : abalamaba
S2 : alamabada
LCS: alamaba

D	G	T	C	A	T	A	G	A
0	0	S1 : ACGTAAAT						
G	0	1	S2 : TACAGTAT					
G	0	1	LCS: GTA					
T	0	0	0	1	0	0	2	1
T	0	0	0	1	0	0	1	2
C	0	0	0	0	2	1	0	0
G	0	1	0	0	1	3	2	1
T	0	0	0	1	0	2	4	3
G	0	1	0	0	0	1	3	5
G	0	1	0	0	0	1	2	4
A	0	0	2	1	0	0	1	3

T C G T G A
* * * * * * * *
T C G T _ G A

T C G T G G A
* * * * * * * *
T C G T G _ A

T C G T G G A
* * * * * * * *
T C G T G A A

Local Alignment Table

Results

Local Alignment App

8. Ερώτημα 8

Παρακάτω εναποθέτουμε τις διαθέσιμες πληροφορίες των δύο ιών, Καθώς επίσης και την μέγιστη κοινή υποακολουθία μεταξύ τους.

SARS-CoV-2

Απομονωμένη η ακίδα

Περισσότερες Πληροφορίες

Bat-RaTG13

Απομονωμένη η ακίδα

Μέγιστη κοινή Υποακολουθία

```
RSVASQSIAYTMSLGAENSVAYSNNIAIPTNFTISVTTEILPVSMKTSVDCTMYICGDSTECSNLLQYGSFCT
QLNRALTGIAVEQDKNTQEVTFAQVKQIYKTPPIKDFGGFNFSQILPDPSKRSFIEDLLFNKVTLADAGFIKQY
GDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVT
QNVLYENQKLIANQFNSAIGKIQDSSLSSSTASALGKLQDVVNQNAQALNTLVKQLSSNFGAISSVLNDILSRLDKV
EAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVV
FLHVTYVPAQEKNFTTAPAIHDGKAHFREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSG
```

9. Ερώτημα 9

Για την καλύτερη οπτικοποίηση των αποτελεσμάτων έγινε χρήση εργαλείων που βρέθηκαν στο διαδύκτιο.
Για να έχετε πρόσβαση σε αυτά τα εργαλεία μπορείτε να χρησιμοποιήσετε τα αντίστοιχα κουμπιά.

Αριστερά φαίνεται ο πίνακας δυναμικού προγραμματισμού που προέκυψε. Η δεύτερη γραμμή και η δεύτερη στήλη αρχικοποιούνται με το κόστος στοίχισης με το κενό. Κάθε φορά που συμπληρώνεται η τιμή ενός κελιού σημειώνεται από ποιο κελί προήλθε, για την εκτέλεση του Traceback.

Το Traceback 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 Traceback είναι η βέλτιστη.

Στην τοπική στοίχιση δεν συναντώνται αρνητικές τιμές στα κελιά του πίνακα, ενώ δίνεται έμφαση στις περιοχές με την μεγαλύτερη ομοιότητα. Οι αρνητικές τιμές μετατρέπονται σε 0 και το Traceback δεν αρχίζει από το τελευταίο κελί, αλλά από τα κελιά με τη μέγιστη ομοιότητα, και καταλήγει στο πρώτο κελί με τιμή 0.

Η τοπική στοίχιση μπορεί να οδηγήσει σε πολλές βέλτιστες εναλλακτικές τις οποίες μπορείτε να δείτε στον πίνακα “Results” που υπάρχει αριστερά.