

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
Υποερώτημα (β)	19
4. Ερώτημα 4	20
Τα θέματα που εντοπίστηκαν.	20
Διαθέσιμες λύσεις	20
5. Ερώτημα 5	21
6. Ερώτημα 6	22
7. Ερώτημα 7	23
8. Ερώτημα 8	24
9. Ερώτημα 9	25

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 SMS DNA Stats interface. The main content area displays a sequence of DNA bases: AACGCTTCTCCAGGCCAACCCAGCGACCGGTGCAATAGATCACACCGAGGGCA ATTAGCCATAACCCATCGGATOTCGGAAGTOATCCTTGGCGGGGAAGTCGGCA CGGATTGGGTATGTGGGCTTGGGATAATTAACTGATTCTCGGATTCTCCCTTTA GTATCGAAAAAGACTGTCAATGCCAGGGCGAAGACCTTTTGTTGAATAGCTTACCC GATATGGGACTCCCTGATCTCTCTTGCATATTAGCTCCGCTTGCATA

Sequence Manipulation Suite: DNA Stats

Paste the raw sequence or one or more FASTA sequences into the text area below. Input limit is 500,000,000 characters.

Submit Clear Reset

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

new window | home | citation

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

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
ag	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
hc	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.

SMS

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  SUSDASCIN 2320 bp mRNA    INV 14-MAR-2000
DEFINITION Strongylocentrotus purpuratus fascin (FSCN1) 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; Echinoidea; 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 Dev Biol 71 (3), 704-714 (1976)
MEDLINE 7051438
REFERENCE 2 (bases 1 to 2320)
AUTHORS Bryan,J. and Kane,R.E.
TITLE Separation and interaction of the major components of sea urchin egg gel
JOURNAL J. Mol. Biol. 125 (2), 207-224 (1978)
MEDLINE 79091184
REFERENCE 3 (bases 1 to 2320)
AUTHORS Bryan,J., Edwards,M., Matsudaira,P., Otto,J. and Wulfkuhle,J.
TITLE Fascin, 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"
/deазвание="larva"
gene 1..2320
/genes="FSCN1"
CDS B5..1575
/genomic_start=1
/functions="actin bundling protein"
/note="putative"
/codon_start=1
```

Submit Clear Reset

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

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

Run 14 Jun 00:36:2020
Valid XHTML 1.0; Valid CSS

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).

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:

MEME HTML Output

MAST HTML Output

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/align/pairwise>. 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". Under "Global Alignment", it lists "Needle (EMBOSS)", "Stretcher (EMBOSS)", and "GGSEARCH2SEQ". Under "Local Alignment", it lists "Water (EMBOSS)" and "Matcher (EMBOSS)". A cookie consent banner at the bottom states: "This website requires cookies, and the limited processing of your personal data in order to function. By using the site you are agreeing to this as outlined in our Privacy Notice and Terms of Use." with a "I agree, dismiss this banner" link.

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

Example job

The screenshot shows the EMBOSS Needle web interface with the URL <https://www.ebi.ac.uk/emboss/needle>. The page title is "EMBOSS Needle". It has tabs for "Input form", "Web services", "Help & Documentation", and "Bioinformatics Tools FAQ". The main content area is titled "Pairwise Sequence Alignment" and says "EMBOSS Needle reads two input sequences and writes their optimal global sequence alignment to file." It has two main input sections: "STEP 1 - Enter your protein sequences" and "STEP 2 - Set your pairwise alignment options". The "STEP 1" section has a dropdown menu "Enter a pair of" set to "PROTEIN" and a text area for pasting sequences. The "STEP 2" section includes "OUTPUT FORMAT" set to "pair" and a note about default settings. At the bottom is 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" search bar set to "FASTQ to FASTA". The left sidebar lists various genomic tools under categories like GENERAL TEXT TOOLS, GENOMIC FILE MANIPULATION, and COMMON GENOMICS TOOLS. The main panel shows the "FASTQ to FASTA" tool configuration with fields for "Input FASTQ file" (set to "No fastqsanger.fastqsanger.gz"), "Discard sequences with unknown (N) bases" (set to "yes"), and "Rename sequence names in output file (reduces file size)" (set to "yes"). At the bottom are "Execute" and "What it does" sections.

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.

Babraham Bioinformatics

About | People | Services | Projects | Training | Publications

Download Babraham Bioinformatics Projects

All Babraham Bioinformatics code is released under the GNU General public license. You should be aware that some of the downloads on this page include code from other projects which is available under different license terms.

All Babraham Bioinformatics projects are distributed WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.

If you have any problems with any of the software you downloaded from here then [please report them to us](#) and we'll do our best to get them sorted out.

If you want to stay up to date with changes to our existing projects, or with new software we have released then you can [follow us on twitter](#) You can also see demonstrations of our software on our [YouTube channel](#).

ASAP Performing allele-specific alignments in Next-Gen Sequencing samples with mixed genetic background

- [Release Notes](#)
- [ASAP User Guide v0.1.2 \(pdf\)](#)
- [ASAP_v0.1.2.tar.gz](#)
- [ASAP test dataset](#)

Bareback A tool to shuffle low complexity sequence to the end of Illumina sequencing runs

- [README](#)
- [Bareback_v1.0.tar.gz](#) (includes back-shuffling script)

Bismark A bisulfite read mapper and methylation caller

- [Release Notes](#)
- [Bismark User Guide v0.21.0](#)
- [Bismark test dataset](#)
- [Bismark_v0.22.3.tar.gz](#) (includes bismark2summary, deduplicate_bismark, bismark2bedGraph, coverage2cytosine and other source code)
- [RRBS Guide](#) (PDF, last updated 25 Jan 2017)

ChIPMonk ChIP-on-Chip analysis tool

- [README](#)
- [Release Notes](#) Please read these if you are upgrading from an older version!!
- [ChIPMonk v1.2.3 \(Win/Linux zip file\)](#)
- [ChIPMonk v1.2.3 \(Mac DMG image\)](#)
- [Source Code for ChIPMonk v1.2.3 \(zip file\)](#)
- [Example data file](#) [85MB] (unzip before use) [Original paper](#)

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

FASTQC Results

FastQC

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

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**.

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
gckucgcgagarty

>sample sequence 2
ggwgggggggggtgcggaggaaatggacgttgcggcagctgcggcaggagaatgcggaaaaaa
taacatgtataatcatcgacaaactaccctggtgatgtctgtatattacttgttttcgticatctccggcg
```

Submit Clear Reset

- 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

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

Not Secure — bioinformatics.org

Translate results

```
>rf 1 sample sequence
ACDEF

>rf 1 sample sequence 2
GGGGGEEDVVVVAARRSSKKNNMIIITTTTW*CC**YYLLFFSSSSRRRQQHHLLL
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 under categories like 'GENERAL TEXT TOOLS', 'GENOMIC FILE MANIPULATION', and 'COMMON GENOMICS TOOLS'. In the center, a tool configuration panel for 'Filter by quality' is shown. It indicates that the tool has been successfully executed and added one job to the queue. The input file is listed as '1: (unavailable) sample.fq'. The output is '2: Filter by quality on data 1'. A note says you can check the status of queued jobs and view the resulting data by refreshing the History panel. The History panel on the right shows an 'Unnamed history' with two entries: '2: Filter by quality on data 1' and '1: sample.fq'. The total size of the history is 605 b.

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

A terminal window titled 'sample.fq' displays the contents of a FASTQ file. The sequence starts with '@Rosalind_0049_1' followed by the sequence data and quality scores. The sequence ends with 'B:EI>JDB5=>DA?E6B@CA?C;=@@C:6D:3=@49;@87;::;;?8+'.

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

A terminal window titled 'display' shows the same FASTQ sequence data as the previous terminal window, indicating the filtered output. The sequence starts with '@Rosalind_0049_1' and ends with 'B:EI>JDB5=>DA?E6B@CA?C;=@@C:6D:3=@49;@87;::;;?8+'.

```
@Rosalind_0049_1
GCAGAGACCGTAGATGTGTTGCGGACGGTCGGCTCCATGTGACACAG
+
FD@@;C<AI?4BA:=>C<G=:AE=><A??>764A8B797@A:58:527+,
@Rosalind_0049_3
ACCCCATAACGGCGAGCGTCAGCATCTGATATCCTCTTCAATCCTAGCTA
+
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.

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.

Suboptimal Local Alignment

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

The screenshot shows the LALIGN 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 main header "LALIGN" and a sub-header "Pairwise Sequence Alignment". A sub-sub-header "LALIGN finds internal duplications by calculating non-intersecting local alignments of protein or nucleotide sequences." follows. The main content area is divided into sections:

- STEP 1 - Enter your protein sequences:** A dropdown menu set to "PROTEIN". Below it is a text area containing a protein sequence: >Rosalind_12 GACTCCTTGTTCGCCTAAATAGATACATATTACTCTTGACTCTTGTGGCCTAAATAGATACA TTTTGCGACTCCACGAGTGATTGTA
- Or, upload a file:** A "Choose File" button.
- AND**: A section for entering the second protein sequence. It includes a "Choose File" button and a text area containing: >Rosalind_37 ATGGACTCCTTGTTCGCCTAAATAGATACATATTCAACAAAGTGTGCACTTAGCCTGCCGACTCC TTTGTTGCCTAAATAGATACATATTG
- STEP 2 - Set your pairwise alignment options:** A note stating "The default settings will fulfill the needs of most users." and a "More options..." link.
- STEP 3 - Submit your job:** A checkbox for "Be notified by email (Tick this box if you want to be notified by email when the results are available)" 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: Rosalind_18, Rosalind_23, Rosalind_51, Rosalind_51, and Rosalind_7. Below this is a file upload section with a placeholder "Or, upload a file: Choose File... no file selected". There are buttons for "Use a example sequence", "Clear sequence", and "See more example inputs". 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 (Tick this box if you want to be notified by email when the results are available)" 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". Below this is a navigation bar with tabs for Alignments, Result Summary, Guide Tree, Phylogenetic Tree, Results Viewers, and Submission Details. The "Alignments" tab is active. There are buttons for "Download Alignment File" and "Hide Colors". The main content area displays the CLUSTAL W(1.2.4) multiple sequence alignment. The sequences shown are Rosalind_7, Rosalind_51, Rosalind_23, Rosalind_18, and Rosalind_28. The alignment shows the sequences with their respective lengths (49, 49, 49, 48, 50) and a color-coded representation of the sequence identities. A note at the bottom 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.

```
AGCCATGTAGCTA...TCAGTAGCTCT
```

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](#)

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 1 in reading frame 1 on the reverse strand extends from base 49 to base 81.
FCCAAGTCCGGGTCATCCCCATGAAACCTGA

>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

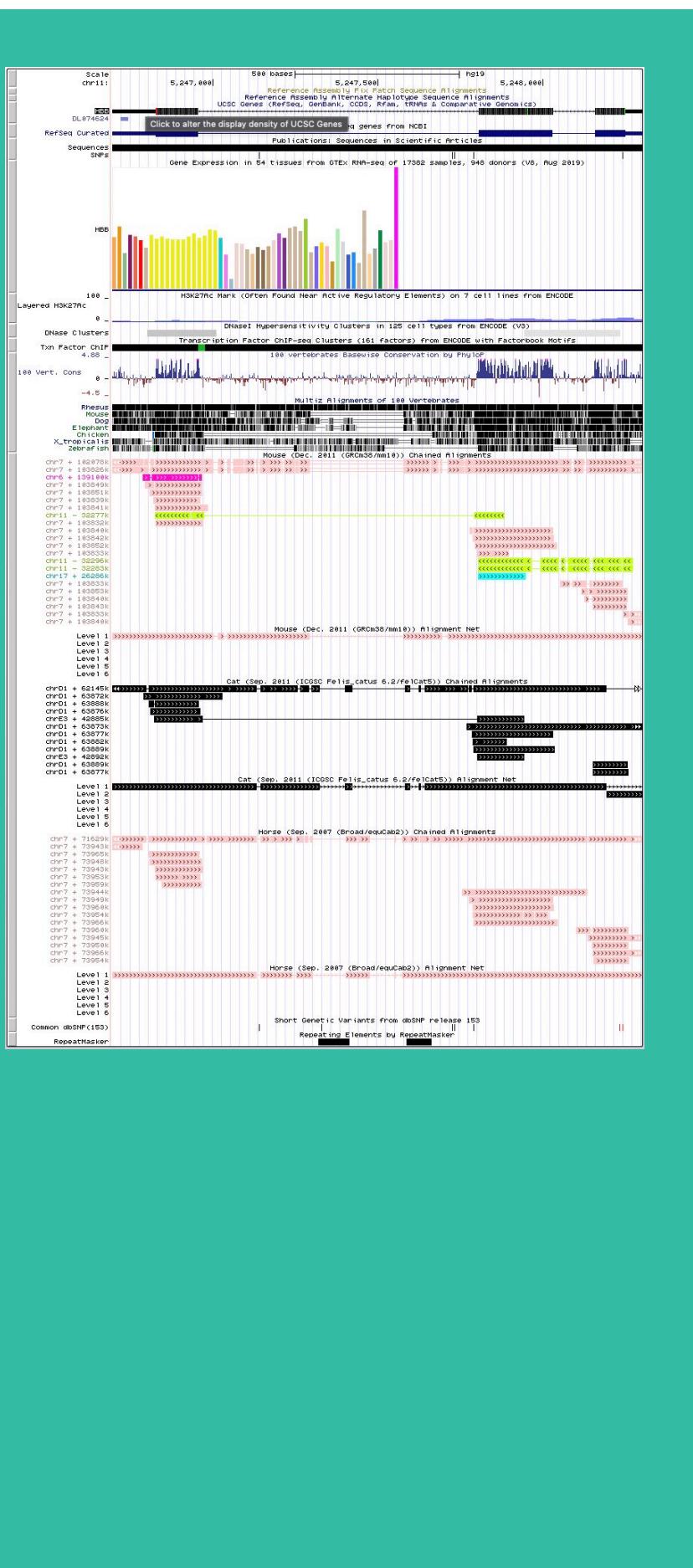
Εδώ χρησιμοποιούμε το FASTQ Quality Trimmer. Ένα εργαλείο το οποίο τριμάρει την ακολουθία μας σύμφωνα με παραμέτρους που έχουν να κάνουν με την ποιότητα.

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 επιλογή 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 εξετάστηκε.

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)

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**.

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 παραπάνω αλγόριθμος εξυπηρετεί το ζητούμενο του ερωτήματος. Μόλις τελειώσει, θα έχουμε εντοπίσει

1. Δημιουργούμε το γενικευμένο δέντρο επιθεμάτων του κειμένου.
2. Εξετάζουμε αν υπάρχουν φύλλα τα οποία να έχουν αποθηκευμένη μόνο μια τιμή.
3. Για κάθε φύλλο, ταξιδεύουμε προς τη ρίζα και σημειώνουμε το μήκος του μονοπατιού.
4. Επιστρέφουμε την λέξη στην οποία αντιστοιχεί το μονοπάτι με το μικρότερο μήκος.

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).

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.

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.

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.

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.

6. Ερώτημα 6

Θα χρησιμοποιήσουμε δυναμικό προγραμματισμό για τον υπολογισμό της βέλτιστης καθολικής στοίχισης. Εφαρμόζουμε λοιπόν τα κόστη/κέρδη τα οποία αναγράφονται αριστερά.

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

7. Ερώτημα 7

8. Ερώτημα 8

9. Ερώτημα 9