

Popular bioinformatics tools in Galaxy: II

Yanbin Yin

Spring 2013

Homework assignment 5

Watch Galaxy 101 (<http://screencast.g2.bx.psu.edu/galaxy101/>) and other videos I put on course page 19-Feb

Use PSU Galaxy main server or glu galaxy server:

1. Get data from UCSC: phenotype and disease group -> cosmic track -> cosmicRaw table
2. Get data: phenotype and disease -> cosmic -> cosmic (chr22)
3. Google cosmic to learn about it
4. Join the two tables according to cosmic id, design operation flow to:
 - Find what chr22 genes have the most cancer mutations (e.g. show top 10)
 - Find what organs have the most cancer mutations in chr22
5. Create workflow and apply to four other chromosomes

Write a report (in **word or ppt**) to include all the operations and screen shots.

Office hour:

Due on March 5 (send by email)

Tue, Thu and Fri 2-4pm, MO325A

Or email: yyin@niu.edu

Outline

- Hands on practice
 - Introduce tools menu
 - How to upload/download
 - How to copy datasets from one history to another
 - EMBOSS

<https://main.g2.bx.psu.edu>

The screenshot shows the Galaxy web interface at <https://main.g2.bx.psu.edu>. The browser address bar at the top displays the URL. The main content area features a central banner for "Galaxy 101: Store, Manage, and Share Start small data with Libraries. The very first tutorial you need An in-depth tutorial". Below the banner is a section titled "Live Quickies" with four cards: "ed fastQ" (Galaxy quickie # 14), "454 Mapping: Single End" (Galaxy quickie # 15), "Uploading Data using FTP" (Galaxy quickie # 17), and "Managing account histories" (Galaxy quickie # 19). To the left, a "Tools" sidebar lists various genomic analysis tools. On the right, a "History" panel is shown, with the "Unnamed history" entry highlighted by a red box. A message in the history panel states: "Your history is empty. Click 'Get Data' on the left pane to start". The word "galaxy-2" is overlaid in red text on the right side of the interface.

Tools

[Get Data](#)

[Send Data](#)

[ENCODE Tools](#)

[Lift-Over](#)

[Text Manipulation](#)

[Convert Formats](#)

[FASTA manipulation](#)

[Filter and Sort](#)

[Join, Subtract and Group](#)

[Extract Features](#)

[Fetch Sequences](#)

[Fetch Alignments](#)

[Get Genomic Scores](#)

[Operate on Genomic Intervals](#)

[Statistics](#)

[Graph/Display Data](#)

[Regional Variation](#)

[Multiple regression](#)

[Multivariate Analysis](#)

[Evolution](#)

[Motif Tools](#)

[Multiple Alignments](#)

[Metagenomic analyses](#)

[Genome Diversity](#)

[Phenotype Association](#)

[EMBOSS](#)

[NGS TOOLBOX BETA](#)

[NGS: QC and manipulation](#)

[NGS: Mapping](#)

[NGS: SAM Tools](#)

[NGS: GATK Tools \(beta\)](#)

[NGS: Variant Detection](#)

[NGS: Indel Analysis](#)

[NGS: Peak Calling](#)

Hands on practice: getting data into Galaxy

Tools

search tools

Get Data

- [Upload File from your computer](#)
- [UCSC Main table browser](#)
- [UCSC Archaea table browser](#)
- [BX table browser](#)
- [EBI SRA ENA SRA](#)
- [BioMart Central server](#)
- [GrameneMart Central server](#)
- [Flymine server](#)
- [modENCODE fly server](#)
- [modENCODE modMine server](#)
- [Ratmine server](#)
- [YeastMine server](#)
- [modENCODE worm server](#)
- [WormBase server](#)
- [EuPathDB server](#)
- [EncodeDB at NHGRI](#)

Upload File (version 1.1.3)

File Format:

Auto-detect
Which format? See help below

File:

No file chosen
TIP: Due to browser limitations, uploading files larger than 2GB is guaranteed to fail. To method (below) or FTP (if enabled by the site administrator).

URL/Text:

<http://cys.bios.niu.edu/yyin/teach/PBB/cesa-pr.fa>

Here you may specify a list of URLs (one per line) or paste the contents of a file.

Files uploaded via FTP:

File	Size	Date
Your FTP upload directory contains no files.		

This Galaxy server allows you to upload files via FTP. To upload some files, log in to the **main.g2.bx.psu.edu** using your Galaxy credentials (email address and password).

Convert spaces to tabs:

Yes
Use this option if you are entering intervals by hand.

Genome:

yyin@niu.edu@main.g2.bx.psu.edu – FileZilla

Host: main.g2.bx.psu.edu Username: yyin@niu.edu Password: Port: Quickconnect

Response: 200 Type set to Command: PASV Response: 227 Entering Passive Mode (128,118,250,4,151,192). Command: LIST Response: 150 Opening BINARY mode data connection for file list Response: 226 Transfer complete Status: Directory listing successful

main.g2.bx.psu.edu

FTP option (for data size > 2GB)

Local site: /Users/yanbinyin/Downloads/

.cups
.filezilla
► .novell
.ssh
Desktop
► Documents
► Downloads
► Google Drive

Filename	Filesize	Filetype	Last n
..			
.DS_Store	24,580	File	02/
FileZilla_3.6.0.2_i686-apple-dar...	5,553,726	bz2-file	02/
seminar announcement 022213....	78,372	Portable Doc ...	02/
download (3)	0	File	02/
download (2)	0	File	02/
Skype_6.2.0.11117.dmg	39,230,620	dmg-file	02/
Fulbright Flyer 4.2.13.pdf	231,636	Portable Doc ...	02/
NIHMS1674-supplement-Supple...	506,372	Portable Doc ...	02/
Evolution and diversity of bacter ...	155,136	Word Docum...	02/

Selected 1 file. Total size: 78,372 bytes

Remote site: /

Empty directory.

1st step: download and unzip filezilla
<http://filezilla-project.org/>

2nd step:
Open filezilla and type in address, username and password

8

FASTA manipulation

Goal: find how many exons on chr22
are shorter than 10nt

Get the chr22 exon position file

Tools

Get Data

- UCSC Main table browser
- UCSC Archaea table browser
- BX table browser
- EBI SRA ENA SRA
- BioMart Central server
- GrameneMart Central server
- Flymine server
- modENCODE fly server
- modENCODE modMine server
- Ratmine server
- YeastMine server
- modENCODE worm server
- WormBase server
- EuPathDB server
- EncodeDB at NHGRI
- EpiGRAPH server
- GenomeSpace import from file

Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks and to retrieve DNA sequence covered by a track. For help in using this application see [Using the Table Browser](#), a description of the controls in this form, the [User's Guide](#) for general information and sample queries, and the OpenHelix Table Browser [tutorial](#) for a narrated presentation of the software features and usage. For more complex queries, you may want to use [Galaxy](#) or our [public MySQL server](#). To examine the biological function of your set through annotation enrichments, send the data to [GREAT](#). Refer to the [Credits](#) page for the list of contributors and usage restrictions associated with these data. All tables can be downloaded in their entirety from the [Sequence annotation Downloads](#) page.

clade: Mammal genome: Human assembly: Feb. 2009 (GRCh37/hg19)

group: Genes and Gene Prediction Tracks track: RefSeq Genes

table: refGene describe table schema

region: genome ENCODE Pilot regions position chr22:1-5130

identifiers (names/accessions): paste list upload list

filter: create

intersection: create

correlation: create

output format: BED – browser extensible data Send output

output file: (leave blank to keep output in browser)

file type returned: plain text gzip compressed

Create one BED record per gene

Whole Gene
 Upstream by 200
 Exons plus 0
 Introns plus 0
 5' UTR Exons
 Coding Exons
 3' UTR Exons
 Downstream by 200

get output summary/statistics

To reset all user cart settings (including custom tracks), [click here](#).

Note: if a feature is close to the end of a chromosome it may be truncated in order to fit the display.

Send query to Galaxy

Cancel

Get the exon sequences

Tools

search tools

[Get Data](#)

[Send Data](#)

[ENCODE Tools](#)

[Lift-Over](#)

[Text Manipulation](#)

[Convert Formats](#)

[FASTA manipulation](#)

[Filter and Sort](#)

[Join, Subtract and Group](#)

[Extract Features](#)

[Fetch Sequences](#)

- [Extract Genomic DNA using coordinates from assembled/unassembled genomes](#)

[Fetch Alignments](#)

[Get Genomic Scores](#)

[Operate on Genomic Intervals](#)

[Statistics](#)

chr22	16258185	16258303	NM_001136213_cds_1_0_chr22_16
chr22	16266928	16267095	NM_001136213_cds_2_0_chr22_16
chr22	16268136	16268181	NM_001136213_cds_3_0_chr22_16
chr22	16269872	16269943	NM_001136213_cds_4_0_chr22_16
chr22	16275206	16275277	NM_001136213_cds_5_0_chr22_16
chr22	16277747	16277885	NM_001136213_cds_6_0_chr22_16
chr22	16279194	16279301	NM_001136213_cds_7_0_chr22_16
Extract Genomic DNA (versi			18
			92
Fetch sequences for intervals in:			35
6: chr22 exon seq			04
Interpret features when possible:			40
Yes			99
Only meaningful for GFF, GTF data			14
Source for Genomic Data:			53
Locally cached			56
Output data type:			19
Interval			52
Execute			58
chr22	17450832	17451083	NM_001037814_cds_6_0_chr22_17

Convert to fasta sequence format

Galaxy Using 0%

Tools

search tools

[Get Data](#)

[Send Data](#)

[ENCODE Tools](#)

[Lift-Over](#)

[Text Manipulation](#)

[Convert Formats](#)

[FASTA manipulation](#)

[Filter and Sort](#)

[Join, Subtract and Group](#)

[Extract Features](#)

[Fetch Sequences](#)

Chr	Start	End	Gene ID	Description	Strand	Sequence
chr22	16258185	16258303	NM_001136213_cds_1_0	chr22_16258186_r	0	TGATGAACAAAATGATACTCAGAACGAACTTTC
chr22	16266928	16267095	NM_001136213_cds_2_0	chr22_16266929_r	0	GTTGAAGAAGAAAATGAGAACGACCGAAGTACT
chr22	16268136	16268181	NM_001136213_cds_3_0	chr22_16268137_r	0	GAAATGTCCTCAAGAACCGAACATAAAATAAGGGT
chr22	16269872	16269943	NM_001136213_cds_4_0	chr22_16269873_r	0	AACAAGACTTAAAGCTGACATCAGAGGAAGAGT
chr22	16275206	16275277	NM_001136213_cds_5_0	chr22_16275207_r	0	AATTGCCAGTTACTTCTGACTACAAAGAAAAA
chr22	16277747	16277885	NM_001136213_cds_6_0	chr22_16277748_r	0	AACTGCTCTCATACTTGCTGTATGTTGTGGATCC
chr22	16279194	16279301	NM_001136213_cds_7_0	chr22_16279195_r	0	CATGGCCTCACACCACTGTTACTTGGTGTACAT
chr22	16282144	16282318	NM_001136213_cds_8_0	chr22_16282145_r	0	GGCGTACAATGCCAGGAAGATGAATGTCGTTA
chr22	16282477	16282592	NM_001136213_cds_9_0	chr22_16282478_r	0	GACTGCTCTACATCTGGCCTCTGCCAATGGAAA
chr22	16287253	16287885	NM_001136213_cds_10_0	chr22_16287254_r	0	ATGGTGGCTGAGGCTGGTCAATGCCGGCTGCC
chr22	16448823	16449804	NM_001005239_cds_0_0	chr22_16448824_r	0	ATGTGTCCTTGACCTTGAGGTCACTGCCCTA
chr22	17071766	17073440	NM_014406_cds_0_0	chr22_17071767_r	0	ATGGACACGACAGTCCCTCAGCCCTGGAGCTC
chr22	17264508	17265299	NM_175878_cds_0_0	chr22_17264509_r	0	CATTGCTGATGACATTCCCTGTTATCAGTTA
chr22	17280660	17280914	NM_175878_cds_1_0	chr22_17280661_r	0	GTGTTGCACACCATTAGAAATTACCAACAAATG
chr22	17288628	17288963	NM_175878_cds_2			ATGAAAGA

FASTA manipulation

- [Compute sequence length](#)
- [Filter sequences by length](#)
- [Concatenate FASTA alignment by species](#)
- [FASTA-to-Tabular converter](#)
- [Tabular-to-FASTA converter](#)
- [FASTA Width formatter](#)
- [RNA/DNA converter](#)
- [Collapse sequences](#)

Tabular-to-FASTA (version 1.1.0)

Tab-delimited file:
6: chr22 exon seq

Title column(s):
**c1
c2
c3
c4**

Multi-select list – hold the appropriate key while clicking

Sequence column:
c7

Execute

Format sequence with fixed width

[Get Data](#)

[Send Data](#)

[ENCODE Tools](#)

[Lift-Over](#)

[Text Manipulation](#)

[Convert Formats](#)

[FASTA manipulation](#)

▪ [Compute sequence length](#)

▪ [Filter sequences by length](#)

▪ [Concatenate FASTA alignment by species](#)

[FASTA manipulation](#)

▪ [Compute sequence length](#)

▪ [Filter sequences by length](#)

▪ [Concatenate FASTA alignment by species](#)

▪ [FASTA-to-Tabular converter](#)

▪ [Tabular-to-FASTA converts tabular file to FASTA format](#)

▪ [FASTA Width formatter](#)

▪ [RNA/DNA converter](#)

▪ [Collapse sequences](#)

```
>NM_001136213_cds_1_0_chr22_16258186_r  
TGATGAACAAAATGATACTCAGAACACTTTCTGAAGAACAGAACACTGGAAATTACAAGATGAGATTCTGATTCAAGAAAAGCAGATAGAAGTGGCTGAAAATGAATT  
>NM_001136213_cds_2_0_chr22_16266929_r  
GTTGAAGAAGAAATGAAGAACCGAACGGAGTACTCATATGGGATTCCCAGAAACCTGACTAACGGTGCCACTGCTGACAATGGTATGGATTAAATTCCACCAAGGAAAGCA  
>NM_001136213_cds_3_0_chr22_16268137_r  
GAAATGTCTCAAGAACCCAGAAATAAGGGTGGTATAGAAAG  
>NM_001136213_cds_4_0_chr22_16269873_r  
AACAAAGACTTAAAGCTGACATCAGAGAACGGCTAAAGGAAGTGAAAATGCCAGGCCAG  
>NM_001136213_cds_5_0_chr22_16275207_r  
AATTTGCCAGTACTTCTGACTACAAAGAAAAACAGATACTAAAAGTCTCTCTGAAAACAGCAATCCAG  
>NM_001136213_cds_6_0_chr22_16277748_r  
AACTGCTCTCATACTTGTGTATGGATCGGAAAGTATACTGCAGCCTCTACTTGAGCAAAACATTGATGTATCTCTCAAGATCTATCTGGACAGACGCCAGAGAGTAT  
>NM_001136213_cds_7_0_chr22_16279195_r  
CATGGCCTCACACCACGTGTACTTGTTACATGAGCAAAACAGCAAGTGGTGAATTAAATCAAGAAAAAGCAAATTAAATGCACTGGATAGATATGGAAG
```

FASTA Width (version 1.0.0)

Library to re-format:

7: chr22 exon fasta

New width for nucleotides strings:

50

Use 0 for single line out.

Execute

```
>NM_001136213_cds_1_0_chr22_16258186_r  
TGATGAACAAAATGATACTCAGAACACTTTCTGAAGAACAGAACACTGGAAATTACAAGATGAGATTCTGATTCAAGAAAAGCAGATAGAAGTGGCTGAAAATGAATT  
GAATATTACAAGATGAGATTCTGATTCAAGAAAAGCAGATAGAAGTGGCTGAAAATTAATCAAGAAAAAGCAAATTAAATGCACTGGATAGATATGGAAG  
>NM_001136213_cds_2_0_chr22_16266929_r  
GTTGAAGAAGAAATGAAGAACCGAACGGAGTACTCATATGGGATTCCCAGAAACCTGACTAACGGTGCCACTGCTGACAATGGTATGGATTAAATTCCACCAAGGAAAGCA  
AAACCTGACTAACGGTGCCACTGCTGACAATGGTATGGATTAAATTCCACCAAGGAAAGCAGAACACCTGAAAGCCAGCAATTCTGACACTGAG  
AATGAACAGTATCACAG  
>NM_001136213_cds_3_0_chr22_16268137_r  
GAAATGTCTCAAGAACCCAGAAATAAGGGTGGTATAGAAAG  
>NM_001136213_cds_4_0_chr22_16269873_r  
AACAAAGACTTAAAGCTGACATCAGAGAACGGCTAAAGGAAAGAGTCACAAAGGCTAAAGGA  
AGTGAAGAACGCCAGAG
```

FASTA manipulation

- [Compute sequence length](#)
- **[Filter sequences by length](#)**
- [Concatenate FASTA alignment by species](#)
- [FASTA-to-Tabular converter](#)
- [Tabular-to-FASTA](#) converts tabular file to FASTA format
- [FASTA Width formatter](#)
- [RNA/DNA converter](#)
- [Collapse sequences](#)

```
>NM_031413_cds_17_0_chr22_18032535_f
AGCTAG
>NM_018943_cds_0_0_chr22_18593632_f
ATG
>NM_001035247_cds_0_0_chr22_19438263_r
GATGA
>NM_001035247_cds_11_0_chr22_19466606_r
ATG
>NM_005659_cds_11_0_chr22_19466606_r
ATG
>NM_000407_cds_0_0_chr22_19711093_f
ATGGGCTCCG
>NM_016449_cds_0_0_chr22_23950951_r
GGTGA
```

Filter sequences by length (v)

Fasta file:

8: fixed length

Minimal length:

0

Maximum length:

10

Setting to '0' will return all sequences

Execute

8: FASTA Width on data 7

7,459 sequences

format: fasta, database: hg19



9: Filter sequences by length on data 8

44 sequences

format: fasta, database: hg19



Galaxy

Analyze Data Workflow Shared Data Visualization Cloud Help User

Using 0%

Tools	chr22	16258185	16258303	NM_001136213_cds_1_0_chr22_16258186_r	0	-	TGATGAACAAAATGATACTCAGAAGCAACTTC
search tools	chr22	16266928	16267095	NM_001136213_cds_2_0_chr22_16266929_r	0	-	GTTGAAGAAGAAATGAGAAGCACCGAACGTACT
Get Data	chr22	16268136	16268181	NM_001136213_cds_3_0_chr22_16268137_r	0	-	GAAATGTCTCAAGAACCCAGAAATAATAAGGGT
Send Data	chr22	16269872	16269943	NM_001136213_cds_4_0_chr22_16269873_r	0	-	AACAAGACTTAAAGCTGACATCAGAGGAAGAGT
ENCODE Tools	chr22	16275206	16275277	NM_001136213_cds_5_0_chr22_16275207_r	0	-	AATTGCCAGTTACTTCTGACTACAAAGAAAA
Lift-Over	chr22	16277747	16277885	NM_001136213_cds_6_0_chr22_16277748_r	0	-	AACTGCTCTCATACTTGCTGTATTTGTGGATCC
Text Manipulation	chr22	16279194	16279301	NM_001136213_cds_7_0_chr22_16279195_r	0	-	CATGCCCTCACACCACTGTTACTTGGTGTACAT
Convert Formats	chr22	16282144	16282318	NM_001136213_cds_8_0_chr22_16282145_r	0	-	GCCGTACAATGCCAGGAAGATGAATGTGCGTTA
FASTA manipulation	chr22	16282477	16282592	NM_001136213_cds_9_0_chr22_16282478_r	0	-	GACTGCTCTACATCTGGCCTCTGCCAATGGAAA
Filter and Sort	chr22	16287253	16287885	NM_001136213_cds_10_0_chr22_16287254_r	0	-	ATGGTGGCTGAGGCTGGTTCAATGCCGGCTGCC
Join, Subtract and Group	chr22	16448823	16449804	NM_001005239_cds_0_0_chr22_16448824_r	0	-	ATGTGCCCCTTGACCTTGCAGCTCACTGCCCTA
Extract Features	chr22	17071766	17073440	NM_014406_cds_0_0_chr22_17071767_r	0	-	ATGGACAGCACAGTCCTTCAGCCCTGGAGCTC
Fetch Sequences	chr22	17264508	17265299	NM_175878_cds_0_0_chr22_17264509_r	0	-	CATTGCTGATGACATTTCCTGTTACAGTTAC
	chr22	17280660	17280914	NM_175878_cds_1_0_chr22_17280661_r	0	-	GTGTTGACACCAATTAGAAATTACCAACAAATG
	chr22	17288628	17288963	NM_175878_cds_2_0_chr22_17288629_r	0	-	ATGGAGACAGTGTGTTGAAGAGATGGATGAAGAA

Filter and Sort

- Filter data on any column using simple expressions
- Sort data in ascending or descending order
- Select lines that match an expression
- Filter on ambiguities in polymorphism datasets

Filter (version 1.1.0)

Filter:

6: chr22 exon seq

Dataset missing? See TIP below.

With following condition:

len(c7)<= 10

Double equal signs, ==, must be used
Select tool.

Number of header lines to skip:

0

Execute

The other way to find out: using filter option

Hands on practice: Text manipulation

Goal: find what genes on chr22 has the largest number of exons

Add one more column

chr22	16258185	16258303	NM_001136213_cds_1_0_chr22_16258186_r	0	-
chr22	16266928	16267095	NM_001136213_cds_2_0_chr22_16266929_r	0	-
chr22	16268136	16268181	NM_001136213_cds_3_0_chr22_16268137_r	0	-
chr22	16269872	16269943	NM_001136213_cds_4_0_chr22_16269873_r	0	-
chr22	16275206	16275277	NM_001136213_cds_5_0_chr22_16275207_r	0	-
chr22	16277747	16277885	NM_001136213_cds_6_0_chr22_16277748_r	0	-
chr22	16279194	16279301	NM_001136213_cds_7_0_chr22_16279195_r	0	-
chr22	16282144	16282318	NM_001136213_cds_8_0_chr22_16282145_r	0	-

Text Manipulation

- Add column to an existing dataset
- Compute an expression on every row
- Concatenate datasets tail-to-head
- Condense consecutive characters
- Convert delimiters to TAB
- Merge Columns together

Compute (version 1.1.0)

Add expression:

c3-c2

as a new column to:

5: chr22 exon

Dataset missing? See TIP below

Round result?:

NO

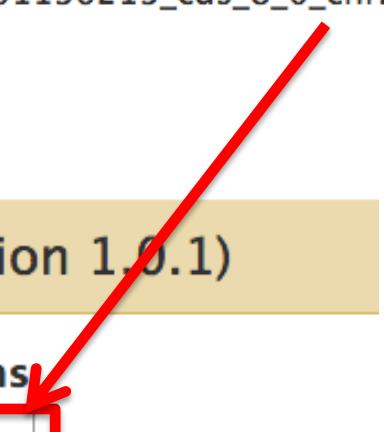
Execute

chr22	16258185	16258303	NM_001136213_cds_1_0_chr22_16258186_r	0	-
chr22	16266928	16267095	NM_001136213_cds_2_0_chr22_16266929_r	0	-
chr22	16268136	16268181	NM_001136213_cds_3_0_chr22_16268137_r	0	-
chr22	16269872	16269943	NM_001136213_cds_4_0_chr22_16269873_r	0	-
chr22	16275206	16275277	NM_001136213_cds_5_0_chr22_16275207_r	0	-
chr22	16277747	16277885	NM_001136213_cds_6_0_chr22_16277748_r	0	-
		16279301	NM_001136213_cds_7_0_chr22_16279195_r	0	-
		16282318	NM_001136213_cds_8_0_chr22_16282145_r	0	-

Text Manipulation

- Add column to an existing dataset
- Compute an expression on every row
- Concatenate datasets tail-to-head
- Condense consecutive characters
- Convert delimiters to TAB
- Merge Columns together
- Create single interval as a new dataset
- Cut columns from a table
- Change Case of selected columns

Cut (version 1.0.1)

Cut columns: 

Delimited by:

From:

Execute

This column has gene ID and exon info

We can cut out this gene ID part and count how many times each ID appear

Further cut the gene ID part
The delimiter is NOT tab!!!

Cut (version 1.0.1)

Cut columns:

c1,c2

Delimited by:

Underscore

From:

15: exon-id Cut on data 5

Execute

NM_001136213_cds_1_0_chr22_16258186_r
NM_001136213_cds_2_0_chr22_16266929_r
NM_001136213_cds_3_0_chr22_16268137_r
NM_001136213_cds_4_0_chr22_16269873_r
NM_001136213_cds_5_0_chr22_16275207_r
NM_001136213_cds_6_0_chr22_16277748_r
NM_001136213_cds_7_0_chr22_16279195_r

NM	001136213

Continue to cut the two cols to save into two files and then join them using new delimiter

Cut (version 1.0.1)

Cut columns: c1

Delimited by: Tab

From: 19: Cut c1, c2 on data 15

Execute

NM	Cut (version 1.0.1)	001136213
NM	Cut columns: <input type="text" value="c2"/> c2	001136213
NM	Delimited by: Tab	001136213
NM	From: 19: Cut c1, c2 on data 15	001136213
NM	Execute	001136213
NM		001136213

Tools

paste paste

Text Manipulation

- Paste two files side by side

EMBOSS

- pasteseq Insert one sequence into another

Workflows

- All workflows

Paste (version 1.0.0)

Paste: 22: Cut c1 on data 19

and: 23: Cut c2 on data 19

Delimit by: Underscore

Execute

NM_001136213
NM_001136213
NM_001136213
NM_001136213
NM_001136213
NM_001136213
NM_001136213
NM_001136213
NM_001136213
NM_001136213

Now we can count the occurrence !!!

Tools

count

Text Manipulation

- Line/Word/Character count of a dataset

Filter and Sort

GFF

- Filter GFF data by feature count using simple expressions

Statistics

- Count occurrences of each record

EMBOSS

Count (version 1.0.0)

from dataset:

24: Paste on data 23 and data 22 ▾

Dataset missing? See TIP below

Count occurrences of values in column(s):

c1

Multi-select list - hold the appropriate key while selecting items.

Delimited by:

Tab ▾

Execute

13	NM_000026
9	NM_000106
4	NM_000185
9	NM_000262
16	NM_000268
15	NM_000343
9	NM_000355
5	NM_000362
13	NM_000395
9	NM_000398

Tools

- sort**
- [Filter and Sort](#)
- [Sort data in ascending or descending order](#)
- [BEDTools](#)
- [Intersect multiple sorted BED files](#)
- [Workflows](#)
- [All workflows](#)

Sort (version 1.0.1)

Sort Dataset: 25: Count on data 24

on column: c1

with flavor: Numerical sort

everything in: Descending order

Column selections

Add new Column selection

Execute

55	NM_058004
42	NM_001136029
42	NM_001242896
42	NM_032608
41	NM_002972
41	NM_014662
40	NM_002473
39	NM_001242897
37	NM_021096

User ▾ Using 0%

3A	NM_0580
DC5	NM_0011
DC5	NM_0012
18B	NM_0326
1	NM_0029
DC5	NM_0146
9	NM_0024
DC5	NM_0012
NA1I	NM_0210
NA1I	NM_0010
IN1	NM_0011
IN1	NM_0122
IN1	NM_0012
NB2	NM_0124
5R1	NM_0142
.	NM_0010
DL1	NM_0070
DO	NM_0014
DL1	NM_0018
.	NM_0147
AL3	NM_0152
1	NM_1525

History

- [HISTORY LISTS](#)
- [Saved Histories](#)
- [Histories Shared with Me](#)
- [CURRENT HISTORY](#)
- [Create New](#)
- [Copy History](#)
- [Copy Datasets](#)
- [Share or Publish](#)
- [Extract Workflow](#)
- [Dataset Security](#)
- [Resume Paused Jobs](#)
- [Collapse Expanded Datasets](#)
- [Include Deleted Datasets](#)
- [Include Hidden Data](#)
- [Unhide Hidden Data](#)
- [Purge Deleted Data](#)
- [Show Structure](#)
- [Export to File](#)
- [Delete](#)
- [Delete Permanent](#)

OTHER ACTIONS

The following list contains each tool that was run to create the dataset; you wish to include in the workflow.

Tools which cannot be run interactively and thus cannot be incorporated into a workflow:

Workflow name

Create Workflow

Tool

Upload File

History

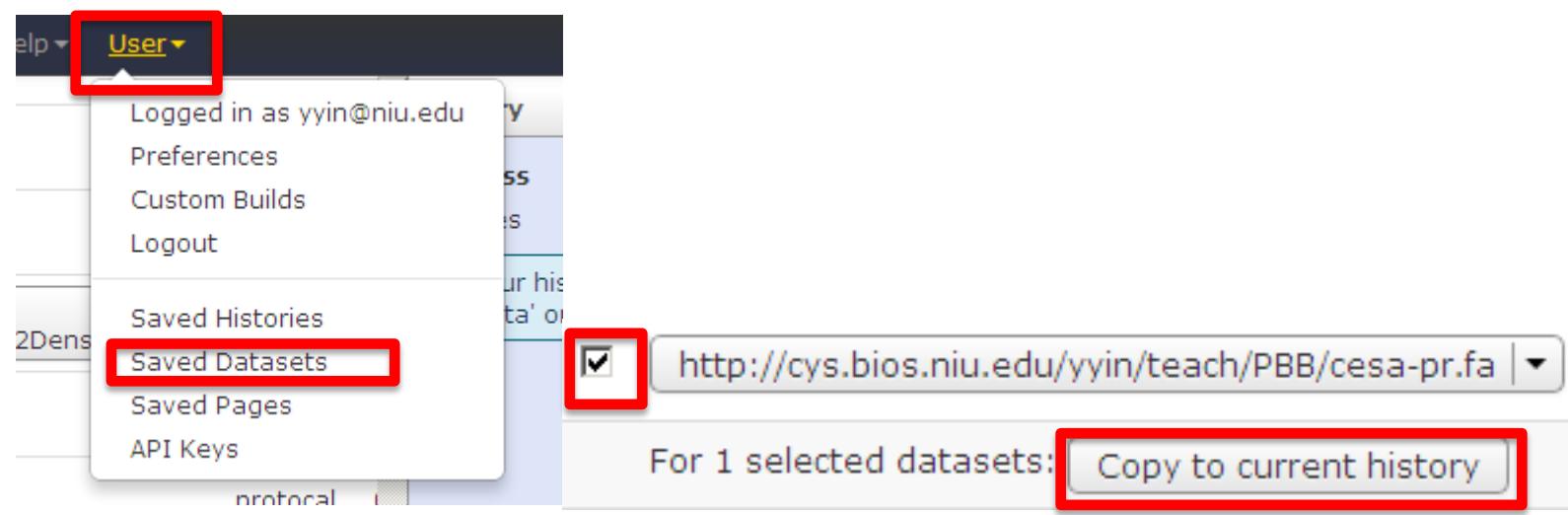
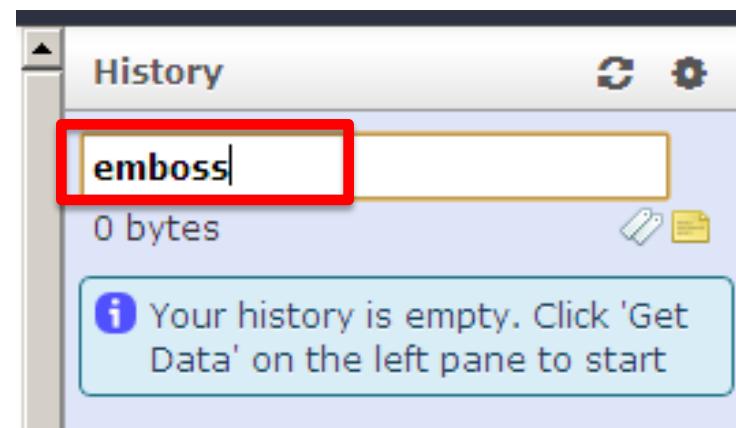
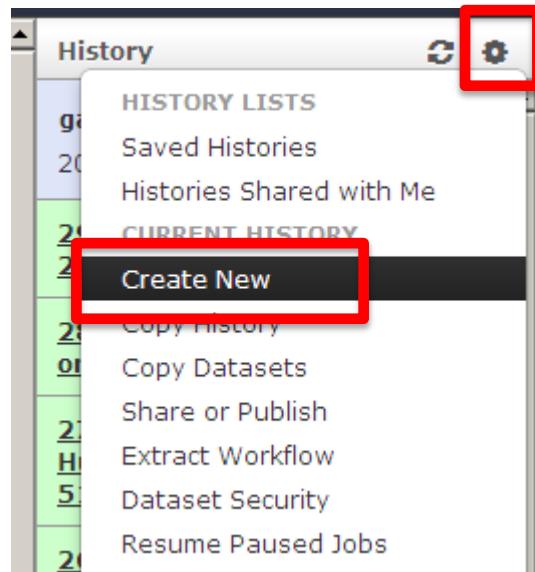
3: ces

Try to run the work flow on Chr20 exons

Hands on practice: EMBOSS

[EMBOSS: European Molecular Biology
Open Software Suite](#)

Create a new history



Upload File (version 1.1.3)

File Format:

Which format? See help below

File: No file chosen

TIP: Due to browser limitations, uploading files larger
URL method (below) or FTP (if enabled by the site adi

URL/Text:

Here you may specify a list of URLs (one per line) or p

Upload File (version 1.1.3)

File Format:

Which format? See help below

File: No file chosen

TIP: Due to browser limitations, uploading files larger than 2MB must be done via URL method (below) or FTP (if enabled by the site administrator).

URL/Text:

Here you may specify a list of URLs (one per line) or p

Copy & past only the first protein seq

Upload File (version 1.1.3)

File Format:
Auto-detect
Which format? See help below

File:
 No file chosen
TIP: Due to browser limitations, uploading files larger than 2GiB
URL method (below) or FTP (if enabled by the site administrator)

URL/Text:

```
>AT2G21770.1|AT2G21770.1|cesA
MINTGGRLIAGSHNRNEFVLINADDTARIRS
AEELSGQTCKICRDEIELTDNGEPFIACNE
CAFPTCRPCYERREGNQACPQCGTRYK
RIKGSPRVEGDEEDDDIDDLEHEFYGMDPE
```

Here you may specify a list of URLs (one per line) or paste the

History	
emboss	36.0 KB
7: AtCesA	
6: CesA alignment	
2: nucleotide seq	
1: CesA proteins	

Google Microphone Search

Web Images Maps Shopping Videos More Search tools

About 1,420,000 results (0.25 seconds)

EMBOSS Homepage
emboss.sourceforge.net/
The European Molecular Biology Open Software Suite. An open source project started by the EMBnet community in order to replace proprietary systems like ...

→ C emboss.sourceforge.net
recarb - Google Sea... George Mason Univers... Customize Links Free Hotmail RealPlayer Windows Marketplace Windows Media

emboss

About • **Applications** • GUIs • Servers • Downloads • Licence • User docs • Development involved • Support • Meetings • News • Credits

EMBOSS was most recently funded from May 2009 to Dec 2011 by BBSRC grant BB/I00425X/1

Funded from May 2006 to April 2009 by BBSRC grant BB/D018358/1

About EMBOSS

Overview • Uses • FAQ Citing EMBOSS

A high-quality package of free, Open Source software for molecular biology ... *more >*

Applications

EMBOSS • EMBASSY • Groups Proposed

EMBOSS Applications

Contents

- Introduction
- Application groups (CVS & stable releases)
 - CVS (developers) release
 - Stable release 6.4.0
 - Stable release 6.3.0
 - Stable release 6.2.0

Hundreds of useful commands

Group	Description
Acd	Acd file utilities
Alignment	Sequence comparison and alignment
Alignment consensus	Merging sequences to make a consensus
Alignment differences	Finding differences between sequences
Alignment dot plots	Dot plot sequence comparisons
Alignment global	Global sequence alignment
Alignment local	Local sequence alignment
Alignment multiple	Multiple sequence alignment
Assembly fragment assembly	DNA sequence assembly
Data resources	Data resources
Data retrieval	Data retrieval
Data retrieval chemistry data	Chemistry data retrieval
Data retrieval feature data	Sequence feature data retrieval
Data retrieval ontology data	Ontology data retrieval
Data retrieval resource data	Resource data retrieval
Data retrieval sequence data	Sequence data retrieval

- banana Bending and curvature plot in B-DNA

banana (version 5.0.0)

On query:



Execute

banana predicts bending of a normal (B) DNA double helix, using the method of Goodsell & Dickerson, NAR 1994 11;22(24):5497-5503. The program calculates the magnitude of local bending and macroscopic curvature at each point along an arbitrary B-DNA sequence

<http://emboss.sourceforge.net/apps/release/6.4/emboss/apps/banana.html>

Base	Bend	Curve
t	0.0	0.0
A	11.7	0.0
A	13.9	0.0
G	14.1	0.0
A	13.3	0.0
T	10.6	0.0
A	14.9	0.0
C	17.7	0.0
C	17.7	0.0
T	22.3	0.0
C	26.9	0.0
G	18.5	0.0
A	5.0	0.0
A	1.3	0.0
A	5.9	0.0
T	9.2	0.0
A	5.9	0.0
T	1.3	0.0
T	0.0	0.0
T	3.4	0.0
T	7.8	4.2
A	5.9	5.4
T	1.3	6.8
T	5.7	7.8
T	15.3	8.6
G	19.7	9.4
C	20.7	10.3
A	12.4	11.4

- geecee Calculates fractional GC content of nucleic acid sequences

geecee (version 5.0.0)

Sequences:

2: nucleotide seq ▾

Execute

```
#Sequence    GC content
contig00008    0.46
```

<http://emboss.sourceforge.net/apps/cvs/emboss/apps/geecee.html>

- dan Calculates DNA RNA/DNA melting temperature

On query:

Window Size:

Step size (shift increment):

Create a graph:
 Yes

DNA Concentration (nM):

Salt concentration (mM):

Execute

Output the DeltaG, DeltaH and DeltaS values:
 Yes

Temperature at which to calculate the DeltaG, DeltaH and DeltaS values:

Dan calculates the **melting temperature (Tm)** and the percentage of G+C nucleotides for windows over a nucleic acid sequence, optionally plotting them.

<http://emboss.sourceforge.net/apps/release/6.4/emboss/apps/dan.html>

- [fuzznuc](#) Nucleic acid pattern search

fuzznuc searches for a specified PROSITE-style pattern in nucleotide sequences. They can specify a search for an exact sequence or they can allow various ambiguities, matches to variable lengths of sequence and repeated subsections of the sequence. One or more nucleotide sequences are read from file. The output is a standard EMBOSS report file that includes data such as location and score of any matches

fuzznuc (version 5.0.1)

Sequences:
2: nucleotide seq

Search pattern:
aaaaat

Number of mismatches:
0

Search complementary strand:
No

Output Report File Format:
SeqTable

Execute

```
#####
# Program: fuzznuc
# Rundate: Tue 19 Feb 2013 00:37:18
# Commandline: fuzznuc
#   -sequence /galaxy/main_pool/pool6/files/00
#   -outfile
/galaxy/main_pool/pool4/tmp/job_working_directory
#   -pattern aaaaat
#   -pmismatch 0
#   -complement no
#   -rformat2 seqtable
#   -auto
#   Report_format: seqtable
#   Report_file:
/galaxy/main_pool/pool4/tmp/job_working_directory
#####

=====
#
# Sequence: contig00008      from: 1      to: 862
# HitCount: 1
#
# Pattern_name Mismatch Pattern
# pattern1                  0 aaaaat
#
# Complement: No
#
=====

Start      End  Pattern_name Mismatch Sequence
95        100  pattern1          . AAAAAAT
```

<http://emboss.sourceforge.net/apps/release/6.4/emboss/apps/fuzznuc.html>

- plotorf Plot potential open reading frames

plotorf plots **potential open reading frames** (ORFs) for an input nucleotide sequence

plotorf (version 5.0.0)

Sequence:

2: nucleotide seq ▾

Start codons:

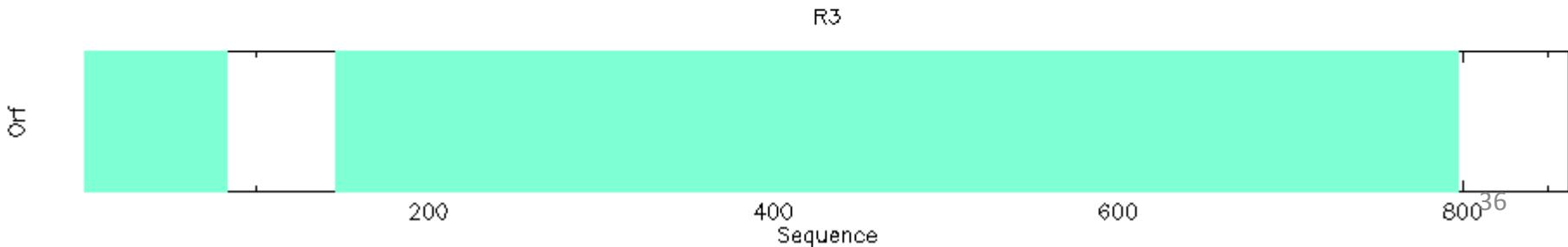
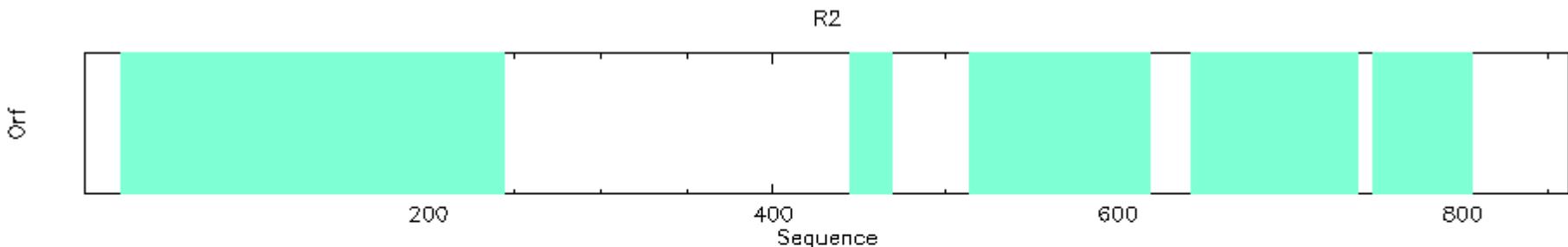
ATG

Stop codons:

TAA

Execute

<http://emboss.sourceforge.net/apps/release/6.4/emboss/apps/plotorf.html>



prettyseq reads a nucleotide sequence and writes an output file containing in a **clean** format the sequence with the translation

- prettyseq Output sequence with translated ranges

PRETTYSEQ of contig00008 from 1 to 862

prettyseq (version 5.0.0)

Sequence:

2: nucleotide seq ▾

Add a ruler:

Yes 

Number translations:

Yes

Number DNA sequence:

Yes 

Width of screen:

60

Execute

<http://emboss.sourceforge.net/apps/release/6.4/emboss/apps/prettyseq.html>

garnier is an implementation of the original Garnier Osguthorpe Robson algorithm (GOR I) for predicting protein secondary structure

- garnier Predicts protein secondary structure

garnier (version 5.0.0)

Sequences:

7: AtCesA

In their paper, GOR mention that if you are analyzing, you can do better in which provide 'decision constants' (dsheet (extend) terms. So, $idc=0$ says various combinations of $dch.dcs$ offset.

idc 0

Output Report File Format:

TagSeq

Execute

<http://emboss.sourceforge.net/apps/release/6.4/emboss/apps/garnier.html>

pepinfo (version 5.0.0)

- [pepinfo](#) Plots simple amino acid properties in parallel

Sequence:

7: AtCesA

Window size for hydropathy averaging:

9

Choose a plot type:

Histogram of general properties

Execute

pepinfo plots various amino acid properties in parallel for an input protein sequence

<http://emboss.sourceforge.net/apps/release/6.4/emboss/apps/pepinfo.html>

Charged residues in cesA from position 1 to 1088



Positive residues in cesA from position 1 to 1088



Negative residues in cesA from position 1 to 1088



- pepstats Protein statistics

pepstats (version 5.0.0)

Sequence:

Include charge at N and C terminus:

Execute

PEPSTATS of cesA from 1 to 108

Molecular weight = 123446.86	Residues = 1088
Average Residue Weight = 113.462	Charge = 5.5
Isoelectric point = 6.6610	
A280 Molar Extinction Coefficient = 211800	
A280 Extinction Coefficient 1mg/ml = 1.72	
Improbability of expression in inclusion bodies = 0.695	

PEPSTATS of cslA from 1 to 534

Molecular weight = 61558.14	Residues = 534
Average Residue Weight = 115.277	Charge = 20.0
Isoelectric Point = 9.4005	
A280 Molar Extinction Coefficient = 109670	
A280 Extinction Coefficient 1mg/ml = 1.78	
Improbability of expression in inclusion bodies = 0.790	

- [plotcon Plot quality of conservation of a sequence alignment](#)

Sequence:

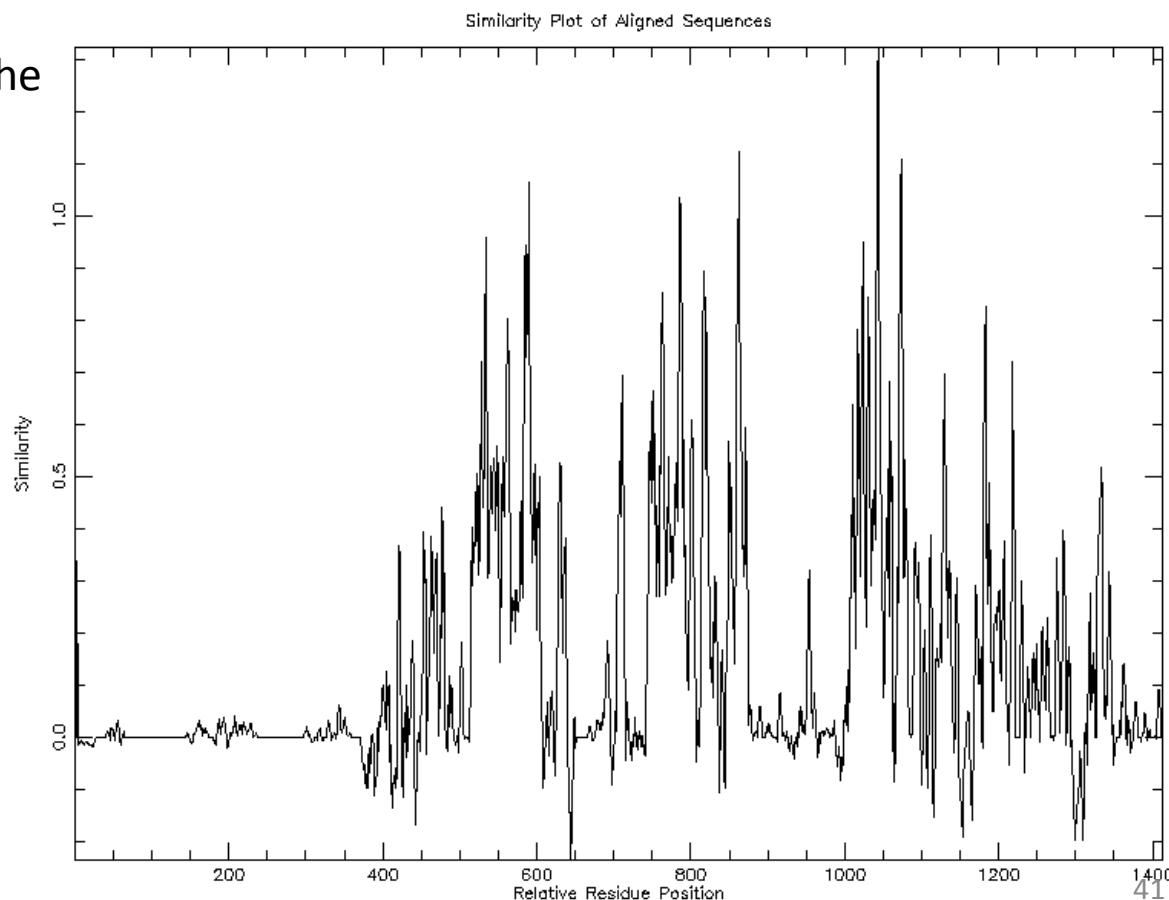
6: CesA alignment ▾

Number of columns to average alignment quality over:

4

Execute

plotcon reads a sequence alignment and draws a plot of the **sequence conservation** within windows over the alignment



Next class: Galaxy III - basic NGS
analysis