

Ćwiczenia nr 5.

Wykorzystanie baz danych i narzędzi analitycznych dostępnych online

I. Zasoby NCBI

Strona:

<http://www.ncbi.nlm.nih.gov/>

stanowi punkt startowy dla eksploracji powiązanych ze sobą zasobów Narodowego Centrum Informacji Biotechnologicznej USA.

Zgromadzone tam informacje umożliwiają odpowiedź na liczne pytania jakie nasuwają się ekologom stosującym w swoich badaniach dane molekularne.

Przykłady takich pytań:

1. Co wiadomo „od strony molekularnej” o moim organizmie badawczym?
2. Czy genom jakiegoś organizmu blisko spokrewnionego z moim gatunkiem jest znany? A może projekt sekwencjonowania takiego genomu jest w toku?
3. Ile i jakich sekwencji DNA lub białek dla mojego organizmu zostało już opisanych?
4. Jeżeli sekwencja interesującego mnie genu nie jest znana u tego organizmu, to czy może sekwencje tego genu znane są u jego bliskich krewnych?
5. Czy zidentyfikowano mikrosatelity dla mojego gatunku? A jeżeli nie to może dla jakichś blisko spokrewnionych gatunków?
6. Uzyskała(e)m sekwencję DNA. Jak sprawdzić czy to gen i organizm o który mi chodziło?
7. Co wiadomo o danym genie? Kto go już badał i u jakich organizmów?
8. Czy istnieją dane o zmienności interesującego mnie genu w populacjach jakiegoś gatunku z danej grupy taksonomicznej?

Z możliwościami i zasobami NCBI można zapoznać się samodzielnie eksplorując strony NCBI.

Pomocny w zrozumieniu ich działania może być podręcznik *The NCBI Handbook*:

<http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=handbook>

oraz instrukcje na stronie:

<http://www.ncbi.nlm.nih.gov/guide/training-tutorials/>

Genomy ilu gatunków zsekwencjonowano dotychczas?

<http://www.ncbi.nlm.nih.gov/genomes/static/gpstat.html>

Genomy jakich ssaków są „skończone”? A jakich roślin?

Skąd bierze się ogromna dysproporcja w liczbie skończonych genomów między bakteriami a eukariotami?

Sekwencje całych genomów są dostępne na serwerach NCBI, możemy je sobie ściągnąć i analizować lokalnie, na własnym komputerze, np.:

ftp://ftp.ncbi.nlm.nih.gov/genomes/H_sapiens/Assembled_chromosomes/

GenBank

<http://www.ncbi.nlm.nih.gov/>

DNA&RNA -> GenBank

Dane w GenBanku mają standardowy format zawierający znacznie więcej niż tylko samą sekwencję kwasu nukleinowego czy białka.

Na stronie na której teraz jesteśmy znajduje się przykładowy rekord oraz wyjaśnienie co oznaczają poszczególne pola:

GenBank Flat File Format

Click on any link in this sample record to see a detailed description of that data element or field. All of the descriptions are included on this page, so it can be printed as a single document. You can also return to the [Alphabetical Quicklinks Table](#) or [Resource Guide](#)

```

LOCUS       SCU49845       5028 bp    DNA             PLN             21-JUN-1999
DEFINITION   Saccharomyces cerevisiae TCP1-beta gene, partial cds, and Axl2p
              (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION    U49845
VERSION      U49845.1   GI:1293613
KEYWORDS     .
SOURCE       Saccharomyces cerevisiae (baker's yeast)
ORGANISM     Saccharomyces cerevisiae
              Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;
              Saccharomycetales; Saccharomycetaceae; Saccharomyces.
REFERENCE    1 (bases 1 to 5028)
AUTHORS      Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W.
TITLE        Cloning and sequence of REV7, a gene whose function is required for
              DNA damage-induced mutagenesis in Saccharomyces cerevisiae
JOURNAL      Yeast 10 (11), 1503-1509 (1994)
PUBMED       7871890
REFERENCE    2 (bases 1 to 5028)
AUTHORS      Roemer,T., Madden,K., Chang,J. and Snyder,M.
TITLE        Selection of axial growth sites in yeast requires Axl2p, a novel
              plasma membrane glycoprotein
JOURNAL      Genes Dev. 10 (7), 777-793 (1996)
PUBMED       8846915
REFERENCE    3 (bases 1 to 5028)
AUTHORS      Roemer,T.
TITLE        Direct Submission
JOURNAL      Submitted (22-FEB-1996) Terry Roemer, Biology, Yale University, New
              Haven, CT, USA
FEATURES             Location/Qualifiers
     source           1..5028
                     /organism="Saccharomyces cerevisiae"
                     /db_xref="taxon:4932"
                     /chromosome="IX"
                     /map="9"
     CDS               1..206
                     /codon_start=3
                     /product="TCP1-beta"
                     /protein_id="AAA98665.1"
                     /db_xref="GI:1293614"
                     /translation="SSLYNGISTSGLDLNNGTIADMRQLGIVESTYKLRVAVSSASEA
                     AEVLLRVNDNIIRARPTANRQHN"
     gene              687..3158
                     /gene="AXL2"
     CDS               687..3158
                     /gene="AXL2"
                     /note="plasma membrane glycoprotein"
                     /codon_start=1
  
```

Zakończono

Rekordy mogą być złożone i zawierać dużo informacji, co zależy od fragmentu genomu i organizmu, a przede wszystkim od zainteresowania badaczy. Poniższy link pokazuje taki złożony rekord:

http://www.ncbi.nlm.nih.gov/nuccore/NM_133378

Format w jakim wyświetlana jest sekwencja można zmieniać, „natywnym” formatem sekwencji w GenBanku jest ASN.1, który możemy znaleźć pod Display Settings. Jest on trudno czytelny dla człowieka, lecz wygodny dla programów przeszukujących bazy danych.

Gdy interesuje nas określony takson

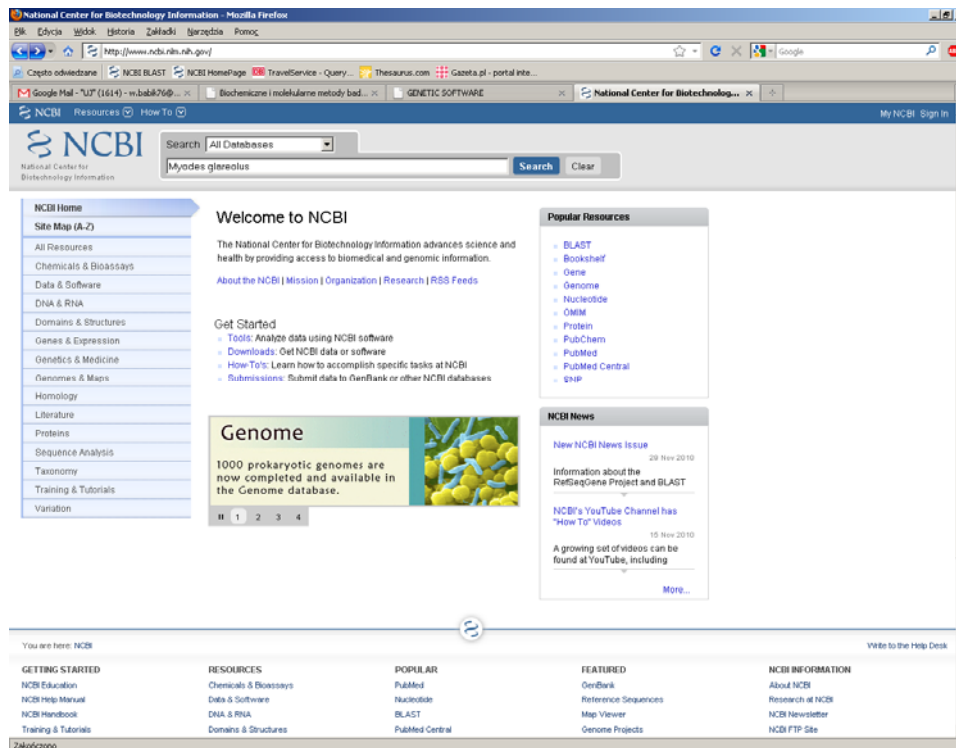
Będziemy chcieli zobaczyć jakie informacje na temat normicy rudej (*Myodes glareolus*) znajdują się w zasobach NCBI.

Na stronie:

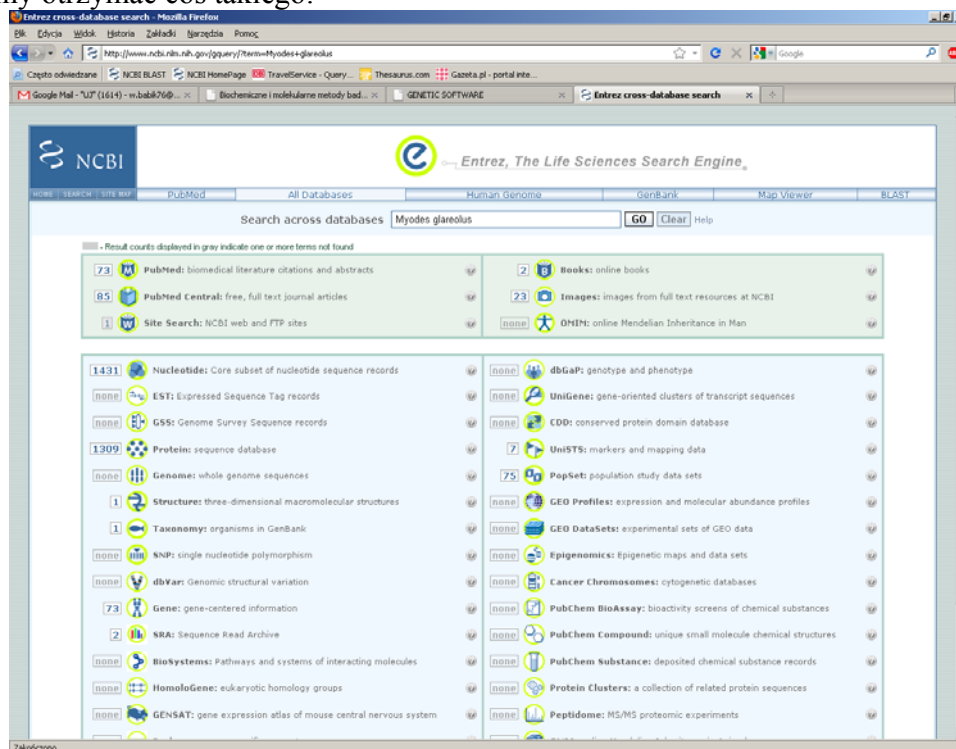
<http://www.ncbi.nlm.nih.gov/>

Domyślnie przeszukiwane są wszystkie bazy danych.

Wpisujemy w polu wyszukiwania: *Myodes glareolus*



Powinniśmy otrzymać coś takiego:



Ile sekwencji nukleotydów dla normicy jest w banku genów?

A ile sekwencji białek?

Czy wszystkie bazy zawierają informację o normicy?

Kliknijmy Taxonomy, a gdy ukaże się nowe okno kliknijmy *Myodes glareolus*. Powinniśmy dostać stronę:

Myodes glareolus

Taxonomy ID: 447135
 Genbank common name: Bank vole
 Inherited blast name: rodents
 Rank: species
 Genetic code: Translation table 1 (Standard)
 Mitochondrial genetic code: Translation table 2 (Vertebrate Mitochondrial)
 Other names:
 synonym: *Clethrionomys glareolus*
 common name: bank vole

[\(Linkage/ full\)](#)

[cellular organisms](#); [Eukaryota](#); [Fungi/Metazoa group](#); [Metazoa](#); [Eumetazoa](#); [Platenia](#); [Coclonata](#); [Deuterostomia](#); [Chordata](#); [Gnathostomata](#); [Teleostomi](#); [Euteleostomi](#); [Sarcopterygii](#); [Tetrapoda](#); [Amniota](#); [Mammalia](#); [Theria](#); [Euthera](#); [Euarchontoglires](#); [Glires](#); [Rodentia](#); [Sciurognathi](#); [Muridae](#); [Cricetidae](#); [Arvicolinae](#); [Myodes](#)

Database name	Direct links
Nucleotide	1,117
Protein	945
Structure	1
Popset	50
UniSTS	14
PubMed Central	246
SEA Experiments	2
Taxonomy	1

External Information Resources (NCBI LinkOut)

LinkOut	Subject	LinkOut Provider
Clethrionomys glareolus taxonomy	taxonomy/phylogenetic	Arctos System Database
Myodes glareolus taxonomy	taxonomy/phylogenetic	
Myodes glareolus with GenBank sequence accessions	taxonomy/phylogenetic	
DNA barcoding: Myodes glareolus	taxonomy/phylogenetic	
Clethrionomys glareolus	taxonomy/phylogenetic	Barcode of Life
Clethrionomys glareolus (Schreber, 1780)	taxonomy/phylogenetic	Catalog of Life
Myodes glareolus Schreber 1780	taxonomy/phylogenetic	Integrated Taxonomic Information System
Myodes glareolus Schreber 1780	taxonomy/phylogenetic	Mammal Species of the World
Wikispecies	taxonomy/phylogenetic	iPhylo

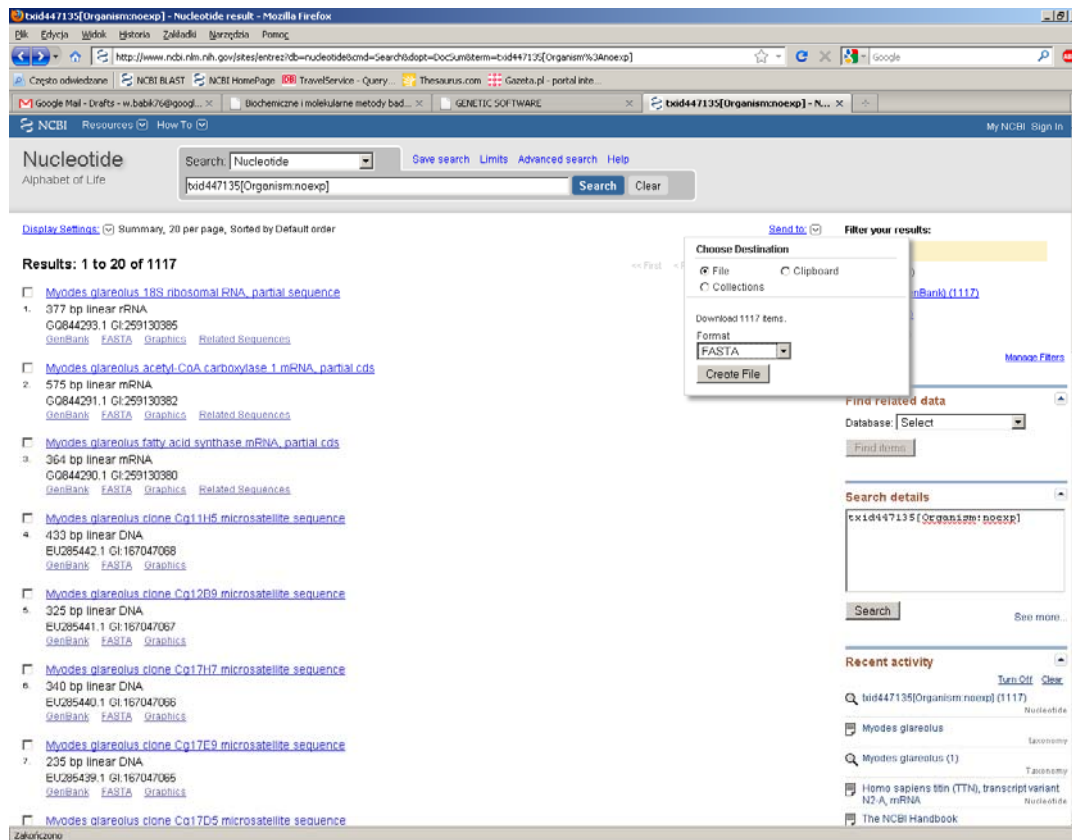
Notes:
 Groups interested in participating in the LinkOut program should visit the [LinkOut home page](#).
 A list of our newest non-bibliographic LinkOut providers can be found [here](#).
 To see LinkOut links in this lineage click [here](#).

Zakończono

która zawiera informacje o normicy, włączając w to jej pełną hierarchiczną systematykę według systemu przyjętego w NCBI.

Chcielibyśmy ściągnąć całą bazę sekwencji normicowych żeby np. wykorzystać ją do szybkiego sprawdzania czy sekwencje uzyskane przez nas z biblioteki genomowej normicy są już w bazie danych.

Gdy klikniemy elementy w kolumnie Direct links, uzyskamy dostęp do rekordów dotyczących normicy we odpowiednich bazach danych. Kliknijmy liczbę obok Nucleotide. Dostajemy spis sekwencji, domyślnie w skróconej formie (Summary), ale możemy sobie je wszystkie wyeksportować np. w formacie FASTA:



sekwencje zapisane w formacie FASTA do pliku będą odczytywane przez wiele programów, możemy w ten sposób stworzyć lokalną bazę normicznych sekwencji i przeszukiwać ją na własnym komputerze.

Czy w banku genów są jakieś mikrosatellity dla nornicy?

Na stronie <http://www.ncbi.nlm.nih.gov/>

w oknie wyszukiwania ograniczamy wyszukiwanie do bazy Nucleotide i wpisujemy `Myodes glareolus [organism] AND microsatellite`.

Gdybyśmy chcieli sprawdzić czy są jakieś mikrosatellity dla innych gatunków z tego samego rodzaju wpisujemy:

`Myodes [organism] NOT glareolus AND microsatellite`

II. BLAST

BLAST (Basic Local Alignment Search Tool) to rodzina programów pozwalających na szybkie przeszukiwanie baz sekwencji i znajdowanie sekwencji podobnych do sekwencji użytej jako zapytanie (*Query*).

Gdy zapytaniem jest sekwencja nukleotydów można przeszukiwać:

- bazy danych nukleotydów (algorytmy: blastn, megablast, discontinuous megablast)
- bazy białek (algorytm: blastx) - sześć możliwych ramek translacji

Gdy zapytaniem jest sekwencja białek można przeszukiwać:

- bazy białek (algorytmy: pblast, psi-blast, phi-blast)
- bazy nukleotydów (algorytm: tblastn)

BLAST można wykorzystywać również do bardziej wyspecjalizowanych celów, które wymieniono na stronie:

<http://blast.ncbi.nlm.nih.gov/Blast.cgi>

w sekcji Specialized BLAST.

Przeszukanie baz sekwencją nukleotydów

Na stronie technik:

http://www.eko.uj.edu.pl/molecol/index.php?option=com_content&view=article&id=101:wbnz-839&catid=7&Itemid=59

otwórz plik Cw_5_sekw1.txt

skopiuj znajdującą się w nim sekwencję wraz z nagłówkiem do schowka, przejdź do strony:

<http://blast.ncbi.nlm.nih.gov/Blast.cgi>

wybierz nucleotide blast

Wklej sekwencję w odpowiednim polu

Wybierz Database... Others (nr etc.)

kliknij BLAST

po kilku sekundach pojawi się wynik

Sekwencja jakiego genu była w pliku? Jakiego organizmu?

Najważniejsze parametry opisujące wynik to bit score i E-value. Bit score opisuje jakość dopasowania. E-value jest statystyczną miarą istotności wyniku. Jest to oczekiwana liczba dopasowań o takim podobieństwie do sekwencji-zapytania jakie otrzymano w analizie, w bazie losowych sekwencji o wielkości równej wielkość bazy przeszukiwanej.

Wyniki otrzymane na stronie przedstawione są w formie łatwej dla odczytania przez człowieka, gdy wyniki analizowane są przez programy, często wyświetla się je w innym formacie, który można wybrać klikając Download w oknie z wynikami BLASTa.

Poniżej przycisku BLAST jest rozwijalna sekcja Algorithm parameters, która pozwala na precyzyjne dostosowanie działania programu do naszych potrzeb.

Czy nasza sekwencja ma homologii wśród bezkręgowców?

Na stronie gdzie ustawia się parametry BLASTa w polu Organism wpisz Metazoa NOT vertebrata?

Następnie spróbuj wyszukiwania z opcją Optimize for Somewhat similar sequences (blastn)

O czym świadczą wyniki?

Przeszukiwanie baz białek

Przeszukaj analizowaną sekwencją nukleotydów bazy sekwencji białek. Dlaczego analiza trwa dłużej?

Otwórz plik Cw_5_sekw2.txt

Sekwencją białka z tego pliku przeszukaj bazę białek dla bezkręgowców

Analiza wielu sekwencji

Otwórz plik Cw_5_sekw3.txt

Skopiuj wszystkie sekwencje do schowka.

Wejdź na stronę :

<http://blast.ncbi.nlm.nih.gov/Blast.cgi>

wybierz Nucleotide BLAST

wklej sekwencje w odpowiednie pole

i ustaw parametry zgodnie ze wzorem:

Nucleotide BLAST: Search nucleotide databases using a nucleotide query - Mozilla Firefox

BLAST Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

NCBI BLAST/blastn suite

blastn blast blastx tblastn tblastx

BLASTn programs search nucleotide databases using a nucleotide query. [more...](#) [Reset page](#) [Bookmark](#)

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [Clear](#) [Query subrange](#)

From

To

Or, upload file [Przełóż...](#)

Job Title

Enter a descriptive title for your BLAST search [?](#)

☐ Align two or more sequences [?](#)

Choose Search Set

Database ☐ Human genomic + transcript ☐ Mouse genomic + transcript ☒ Others (nr etc.):

☒ Nucleotide collection (nr/nt) [?](#)

Organism [Optional](#) ☐ Exclude [+](#)

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown. [?](#)

Exclude [Optional](#) ☐ Models (XM/XP) ☐ Uncultured/environmental sample sequences

Entrez Query [Optional](#) [?](#)

Program Selection

Optimize for ☒ Highly similar sequences (megablast) ☐ More dissimilar sequences (discontiguous megablast) ☐ Somewhat similar sequences (blastn)

Choose a BLAST algorithm [?](#)

BLAST Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly similar sequences)

☐ Show results in a new window

[Algorithm parameters](#) **Note: Parameter values that differ from the default are highlighted in yellow and marked with + sign**

Zakończono

kliknij BLAST

O czym świadczą otrzymane wyniki?

Jak zmieni się wynik gdy w sekcji Program selection wybierzesz Somewhat similar sequences (blastn)?

Przeszukaj tymi samymi sekwencjami bazę sekwencji białkowych

O czym świadczą otrzymane wyniki?

III. Galaxy

Galaxy to zbiór narzędzi pozwalających na manipulację dużymi zbiorami danych genomowych, meta genomowych, kodów kreskowych DNA. Daje łatwy dostęp do zasobów wielu baz danych i pozwala na integrację tych zasobów z naszymi danymi oraz na analizę naszych danych w kontekście informacji dostępnych w bazach.

Analizy można prowadzić online, lub też zainstalować Galaxy lokalnie, co przydaje się gdy analizowane zbiory danych są bardzo duże.

My spróbujemy określić różnorodność bakterii w niewielkiej próbce 200 sekwencji rRNA otrzymanych metodą sekwencjonowania nowej generacji 454.

Plik z danymi ściągamy ze strony kursu:

http://www.eko.uj.edu.pl/molecol/index.php?option=com_content&view=article&id=101:wbnz-839&catid=7&Itemid=59

Sekwencje do Galaxy

W przeglądarce internetowej wchodzimy na stronę Galaxy

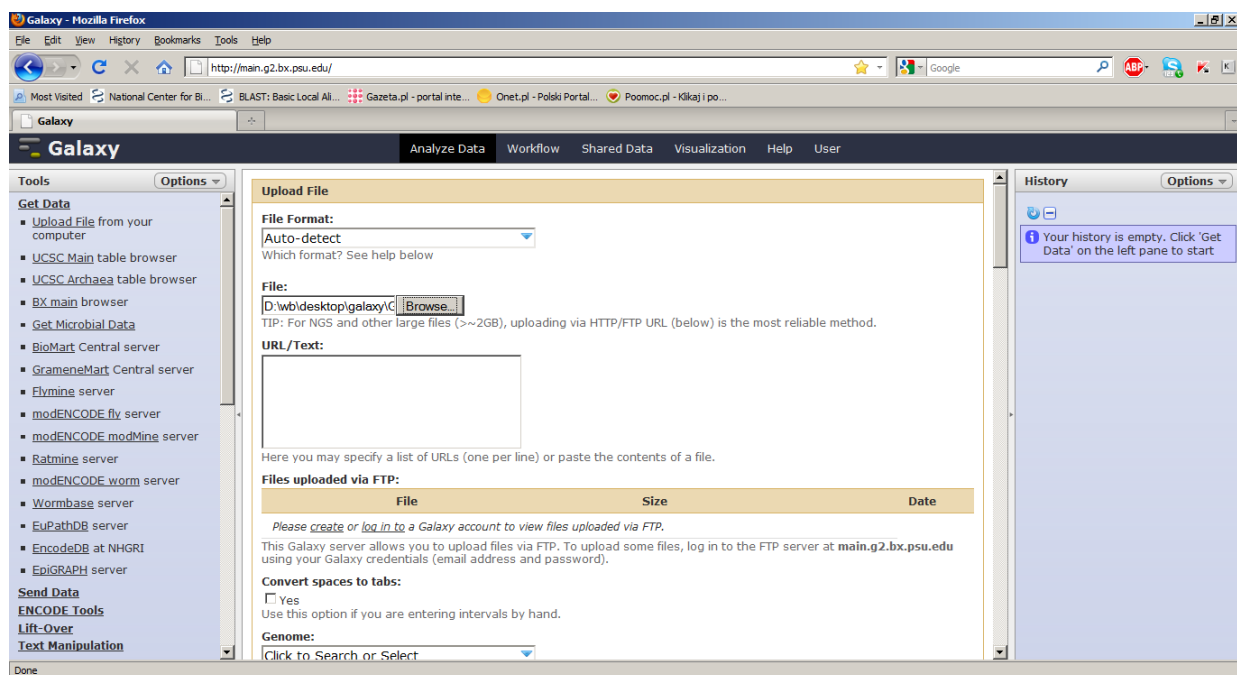
<http://leskowiec.eko.uj.edu.pl:8080/>

Aby używać serwisu musicie się zarejestrować.

Ładujemy nasze dane do Galaxy:

Get Data-> Upload File

Wskaż plik i naciśnij Execute



Przeszukujemy bazę nukleotydów GenBank programem Megablast, który (stosunkowo) szybko wyszukuje sekwencje o wysokim podobieństwie

NGS Mapping->MEGABLAST

Oraz wybieramy bazę nt i opcje jak na obrazu poniżej



Procedura może chwilę potrwać

Filtrowanie danych

Chcemy uzyskać tabelę z długościami sekwencji

Fasta manipulation -> compute sequence length

Execute

Galaxy - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://main.g2.bx.psu.edu/

Most Visited National Center for Bi... BLAST: Basic Local Ali... Gazeta.pl - portal inte... Onet.pl - Polski Portal... Poomoc.pl - Kikaj i po...

Galaxy

Analyze Data Workflow Shared Data Visualization Help User

Tools Options

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-to-Tabular converts FASTA file to tabular format
Tabular-to-FASTA converts tabular file to FASTA format
FASTA Width formatter
RNA/DNA converter
Collapse sequences
Filter and Sort
Join, Subtract and Group
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Statistics
Done

Compute sequence length

Compute length for these sequences:
1: GF5_A_200.fasta

How many title characters to keep?:
0
'0' = keep the whole thing

Execute

What it does

This tool counts the length of each fasta sequence in the file. The output file has two columns per line (separated by tab): fasta titles and lengths of the sequences. The option *How many characters to keep?* allows to select a specified number of letters from the beginning of each FASTA entry.

Example

Suppose you have the following FASTA formatted sequences from a Roche (454) FLX sequencing run:

```
>EYKX4VC02DGL05 length=108 xy=1826_0455 region=2 run=B_2007_11_07_16_15_57_
TCCGCGCCGACAGCAGCCATCTGGATTCGCGCCGACATGACCATGCGCCGCTCCGACAG
TCCGCGCCGACAGCAGCCATCTGGATTCGCGCCGACATGACCATGCGCCGCTCCGACAG
>EYKX4VC02D4882 length=60 xy=1873_8872 region=2 run=B_2007_11_07_16_15_57_
AATAAATCTAAATCCGACAGCTGCGCAATACTGCACTGCACTGCACTGCACTGCACTG
```

Running this tool while setting *How many characters to keep?* to 14 will produce this:

```
EYKX4VC02DGL05 108
EYKX4VC02D4882 60
```

History Options

3: Compute sequence length on data 1
2: Megablast on data 1
1: GF5_A_200.fasta

łączenie tabeli wynikowej Megablasta oraz tabeli z długościami sekwencji:
Join, Subtract and Group->Join Two Queries

Galaxy - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://main.g2.bx.psu.edu/

Most Visited National Center for Bi... BLAST: Basic Local Ali... Gazeta.pl - portal inte... Onet.pl - Polski Portal... Poomoc.pl - Kikaj i po...

Galaxy

Analyze Data Workflow Shared Data Visualization Help User

Tools Options

Send Data
ENCODE Tools
Lift-Over
Text Manipulation
Convert Formats
FASTA manipulation
Filter and Sort
Join, Subtract and Group
Join two Queries side by side on a specified field
Compare two Queries to find common or distinct rows
Subtract Whole Query from another query
Group data by a column and perform aggregate operation on other columns.
Column Join
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Statistics
Graph/Display Data
Regional Variation
Done

Join two Queries

Join:
2: Megablast on data 1

using column:
c1

with:
3: Compute sequence ..h on data 1

and column:
c1

Keep lines of first input that do not join with second input:
No

Keep lines of first input that are incomplete:
No

Fill empty columns:
No

Execute

History Options

3: Compute sequence length on data 1
2: Megablast on data 1
1: GF5_A_200.fasta

filtrowanie:

Filter and Sort -> Filter

kolumna c5 – długość wyrównania (alignment)

kolumna c15 – długość sekwencji

chcemy żeby dopasowanie obejmowało przynajmniej połowę długości sekwencji: ≥ 0.5

Galaxy - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://main.g2.bx.psu.edu/

Most Visited National Center for Bi... BLAST: Basic Local Ali... Gazeta.pl - portal inte... Onet.pl - Polski Portal... Poomoc.pl - Kikaj i po...

Galaxy

Analyze Data Workflow Shared Data Visualization Help User

Tools Options

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
Done

Filter

Filter:
4: Join two Queries ... and data 2
Query missing? See TIP below.

With following condition:
c5/c15 >= 0.5
Double equal signs, ==, must be used as shown above. To filter for an arbitrary string, use the Select tool.

Execute

Double equal signs, ==, must be used as "equal to" (e.g., c1 == 'chr22')

TIP: Attempting to apply a filtering condition may throw exceptions if the data type (e.g., string, integer) in every line of the columns being filtered is not appropriate for the condition (e.g., attempting certain numerical calculations on strings). If an exception is thrown when applying the condition to a line, that line is skipped as invalid for the filter condition. The number of invalid skipped lines is documented in the resulting history item as a "Condition/data issue".

TIP: If your data is not TAB delimited, use Text Manipulation->Convert

Syntax

The filter tool allows you to restrict the dataset using simple conditional statements.

- Columns are referenced with **c** and a **number**. For example, **c1** refers to the first column of a tab-delimited file
- Make sure that multi-character operators contain no white space (e.g., **<=** is valid while **< =** is not valid)
- When using 'equal-to' operator **double equal sign '==' must be used** (e.g., **c1=='chr1'**)
- Non-numerical values must be included in single or double quotes (e.g., **c6=='+'**)
- Filtering condition can include logical operators, but **make sure operators are all lower case** (e.g., **(c1!='chrX' and c1!='chrY')** or **not c6=='+'**)

Example

History Options

Unnamed history

5: Filter on data 4

4: Join two Queries on data 3 and data 2

3: Compute sequence length on data 1

2: Megablast on data 1

1: GF5_A_200.fasta

Taksonomia DNA

Metagenomic analyses -> Fetch taxonomic representation

Przypisanie reprezentacji taksonomicznej

Galaxy - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://main.g2.bx.psu.edu/

Most Visited National Center for Bi... BLAST: Basic Local Ali... Gazeta.pl - portal inte... Onet.pl - Polski Portal... Poomoc.pl - Kikaj i po...

Galaxy

Analyze Data Workflow Shared Data Visualization Help User

Tools Options

Multivariate Analysis
Evolution
Metagenomic analyses
Fetch taxonomic representation
Summarize taxonomy
Draw phylogeny
Find diagnostic hits
Find lowest diagnostic rank
Poisson two-sample test
Human Genome Variation
EMBOSS
NGS TOOLBOX BETA
NGS: QC and manipulation
NGS: Mapping
NGS: SAM Tools
NGS: Indel Analysis
NGS: Peak Calling
NGS: RNA Analysis
RGENETICS
SNP/WGA: Data: Filters
SNP/WGA: QC: LD: Plots
SNP/WGA: Statistical Models

Fetch taxonomic representation

Show taxonomic representation for:
5: Filter on data 4

GIs column:
c2
select column containing GI numbers

Name column:
c1
select column containing identifiers you want to include into output

Execute

Use Filter and Sort->Filter to restrict output of this tool to desired taxonomic ranks. You can also use Text Manipulation->Cut to remove unwanted columns from the output.

What it does

Fetches taxonomic information for a list of GI numbers (sequences identifiers used by the National Center for Biotechnology Information <http://www.ncbi.nlm.nih.gov>).

Example

Suppose you have BLAST output that looks like this:

queryId	targetGI	identity	alignmentLength	mismatches	gaps	score
11_FYRX4VC01B00K1_265	1430919	90.09	212	15	6	252.00

and you want to obtain full taxonomic representation for GIs listed in *targetGI* column. If you set parameters as shown here:

History Options

5: Filter on data 4

4: Join two Queries on data 3 and data 2

3: Compute sequence length on data 1

2: Megablast on data 1

1: GF5_A_200.fasta

Filtrowanie sekwencji unikatowych na określonym poziomie taksonomicznym

Metagenomic analyses -> Find lowest diagnostics rank

The screenshot shows the Galaxy web interface in a Mozilla Firefox browser. The main panel displays the 'Find lowest diagnostic rank' tool configuration. The tool is set to 'for taxonomy dataset:' with a dropdown menu showing '6: Fetch taxonomic r..n on data 5'. Below this, there is a section 'require the lowest rank to be at least:' with a dropdown menu showing 'Class'. An 'Execute' button is visible. The 'What it does' section explains that the tool identifies the lowest taxonomic rank for which a metagenomic sequencing read is diagnostic. The 'Example' section provides a detailed explanation of the tool's output, showing taxonomic profiles for two reads, `read_1` and `read_2`, and their corresponding taxonomic labels. The 'History' panel on the right shows a list of tools executed, including '6: Fetch taxonomic representation on data 5', '5: Filter on data 4', '4: Join two Queries on data 3 and data 2', '3: Compute sequence length on data 1', '2: Megablast on data 1', and '1: GFS_A_200.fasta'.

Podsumowanie danych

Metagenomic analyses → Summarize taxonomy

Wszystkie kroki naszej analizy są zapisywane, możemy je odtworzyć, zapisać w pliku, wykonać ponownie, przesłać komuś jako Workflow:

History → Options → Extract Workflow

The screenshot shows the Galaxy web interface in a Mozilla Firefox browser. The main panel displays the 'Extract Workflow' process. The 'Workflow name' is 'Workflow constructed from history "Unnan"'. There are buttons for 'Create Workflow', 'Check all', and 'Uncheck all'. The 'Tool' section lists tools that can be included in the workflow, including 'Upload File', 'Megablast', 'Compute sequence length', 'Join two Queries', and 'Filter'. The 'History items created' section shows a list of tools executed, including '1: GFS_A_200.fasta', '2: Megablast on data 1', '3: Compute sequence length on data 1', and '4: Join two Queries on data 3 and data 2'. The 'History' panel on the right shows a list of tools executed, including '8: Summarize taxonomy on data 7', '7: Find lowest diagnostic rank on data 6', and '6: Fetch taxonomic representation on data 5'.