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:

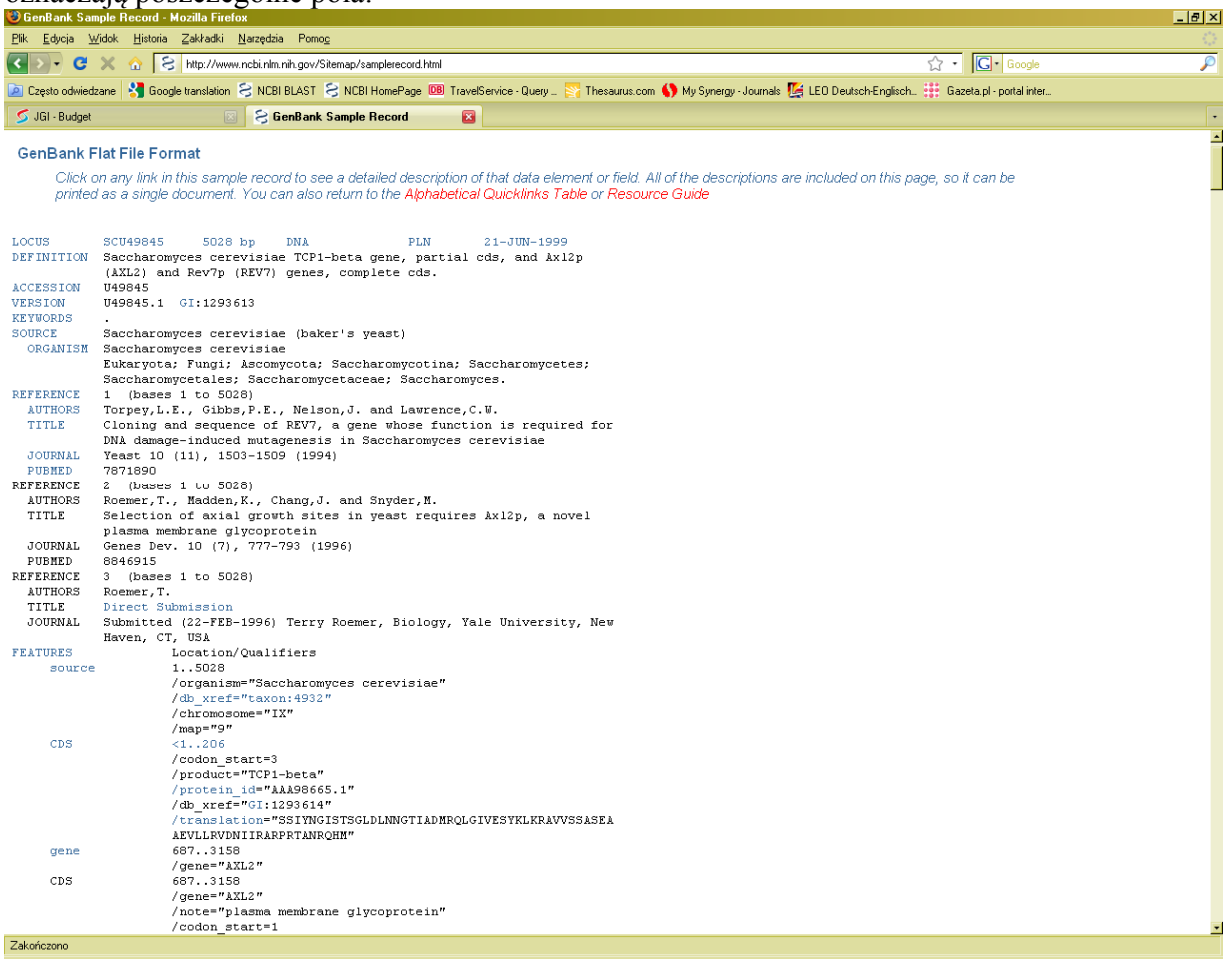
1. Co wiadomo „od strony molekularnej" o moim organizmie badawczym?
2. Czy genom jakiegoś organizmu blisko spokrewnionego z moim gatunkiem jest znany? A może projekt sekwencjonowania takiego genomu jest w toku?
3. Ile i jakich sekwencji DNA lub białek dla mojego organizmu zostało już opisanych?
4. Jeżeli 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/

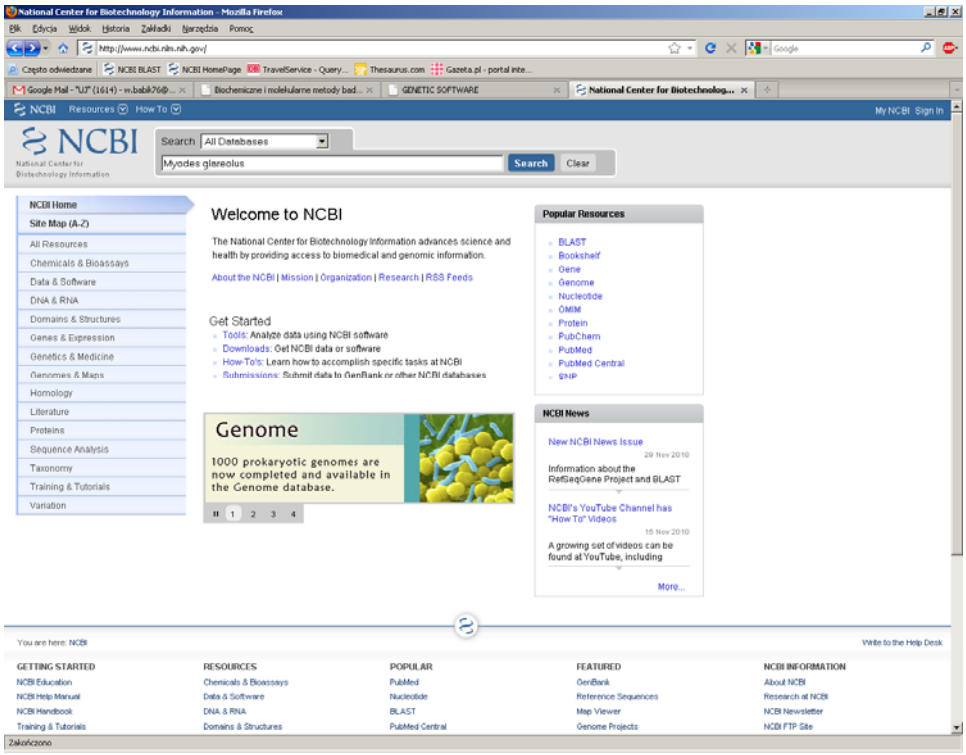
GenBank
http://www.ncbi.nlm.nih.gov/
`DNA&RNA->GenBank`

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:

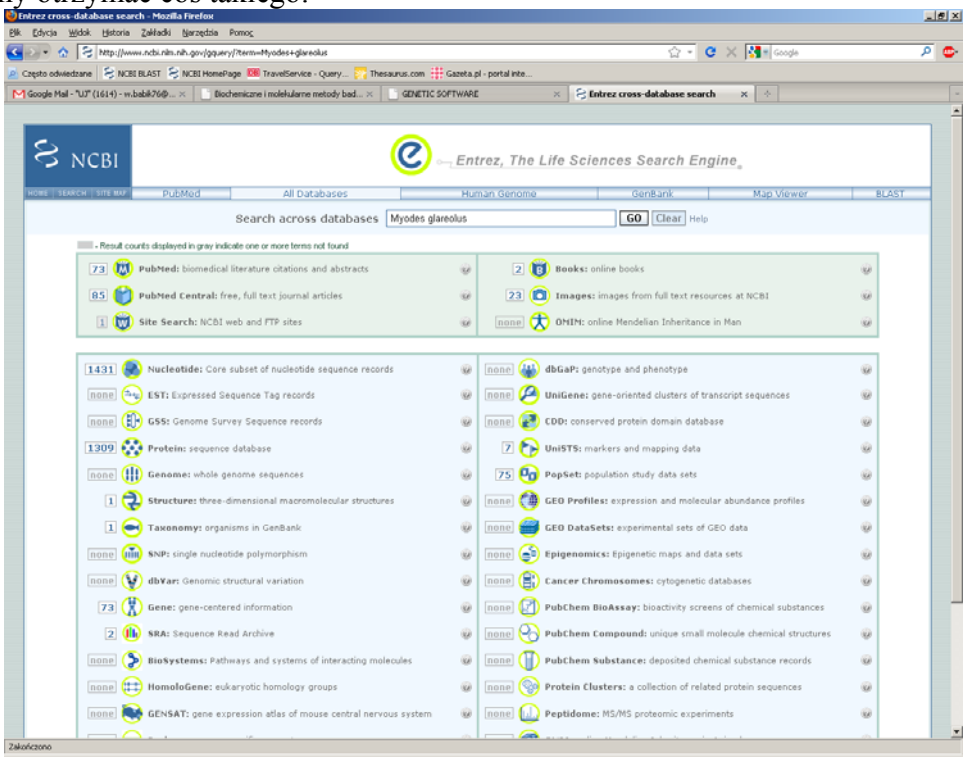


Rekordy mogą być złożone i zawierać dużo informacji, co zależy od fragmentu genomu i organizmu, a przede wszystkim od zainteresowania badaczy. Poniższy link pokazuje taki złożony rekord:
http://www.ncbi.nlm.nih.gov/nuccore/NM_133378
Format w jakim wyświetlana jest sekwencja można zmieniać, „natywnym" 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 nornicy 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`

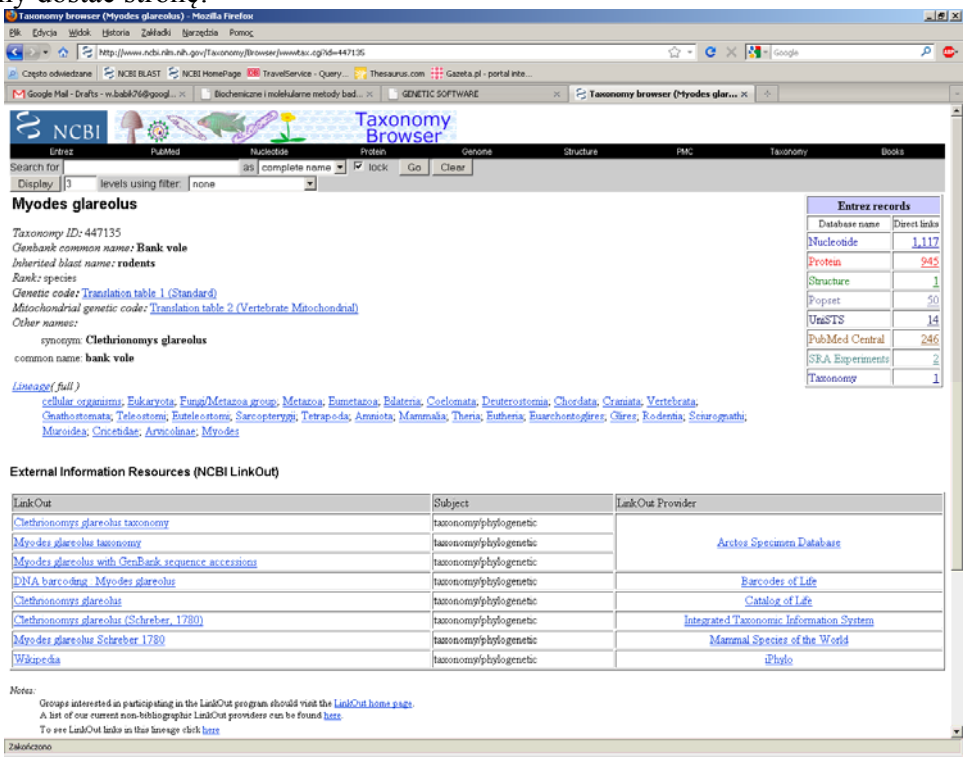


Powinniśmy otrzymać coś takiego:



Ile sekwencji nukleotydów dla nornicy jest w banku genów?
A ile sekwencji białek?
Czy wszystkie bazy zawierają informację o nornicy?

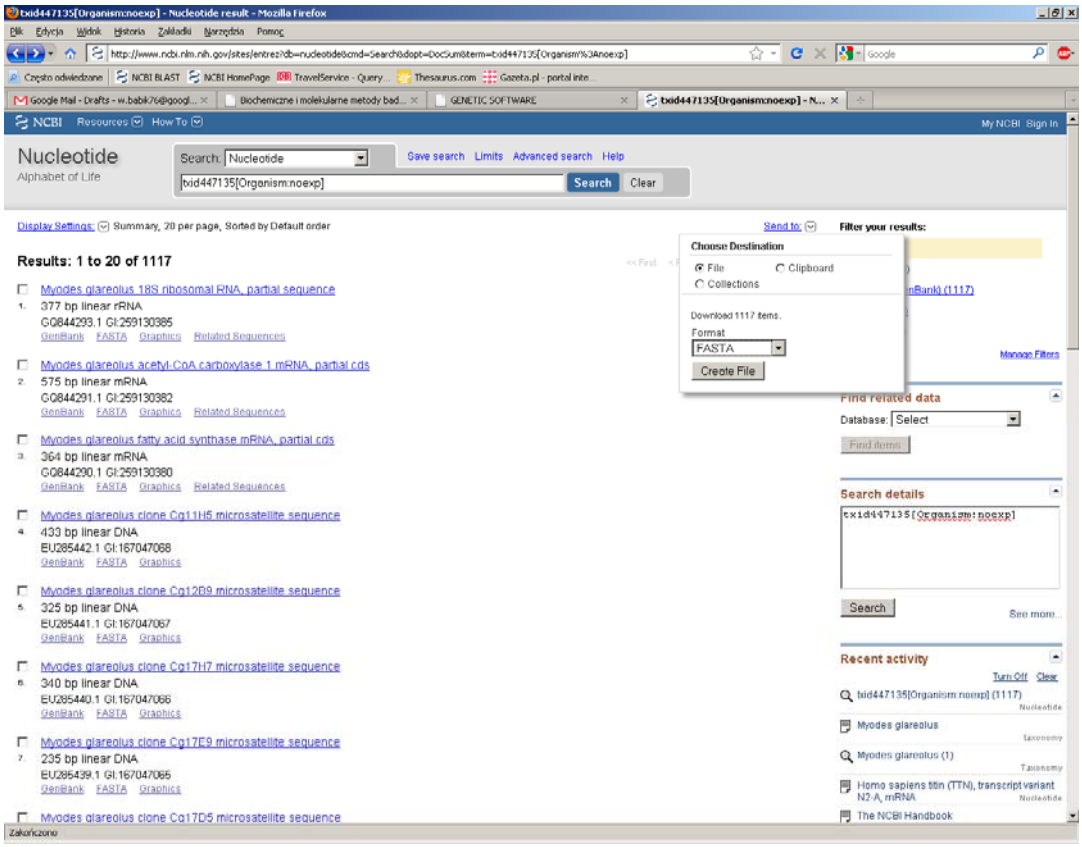
Kliknijmy `Taxonomy`, a gdy ukaże się nowe okno kliknijmy `Myodes glareolus`. Powinniśmy dostać stronę:



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

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

Gdy klikniemy elementy w kolumnie `Direct links`, uzyskamy dostęp do rekordów dotyczących nornicy 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 nornicowych sekwencji i przeszukiwać ją na własnym komputerze.

Czy w banku genów są jakieś mikrosatelity 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 mikrosatelity dla innych gatunków z tego samego rodzaju wpiszemy:
`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.

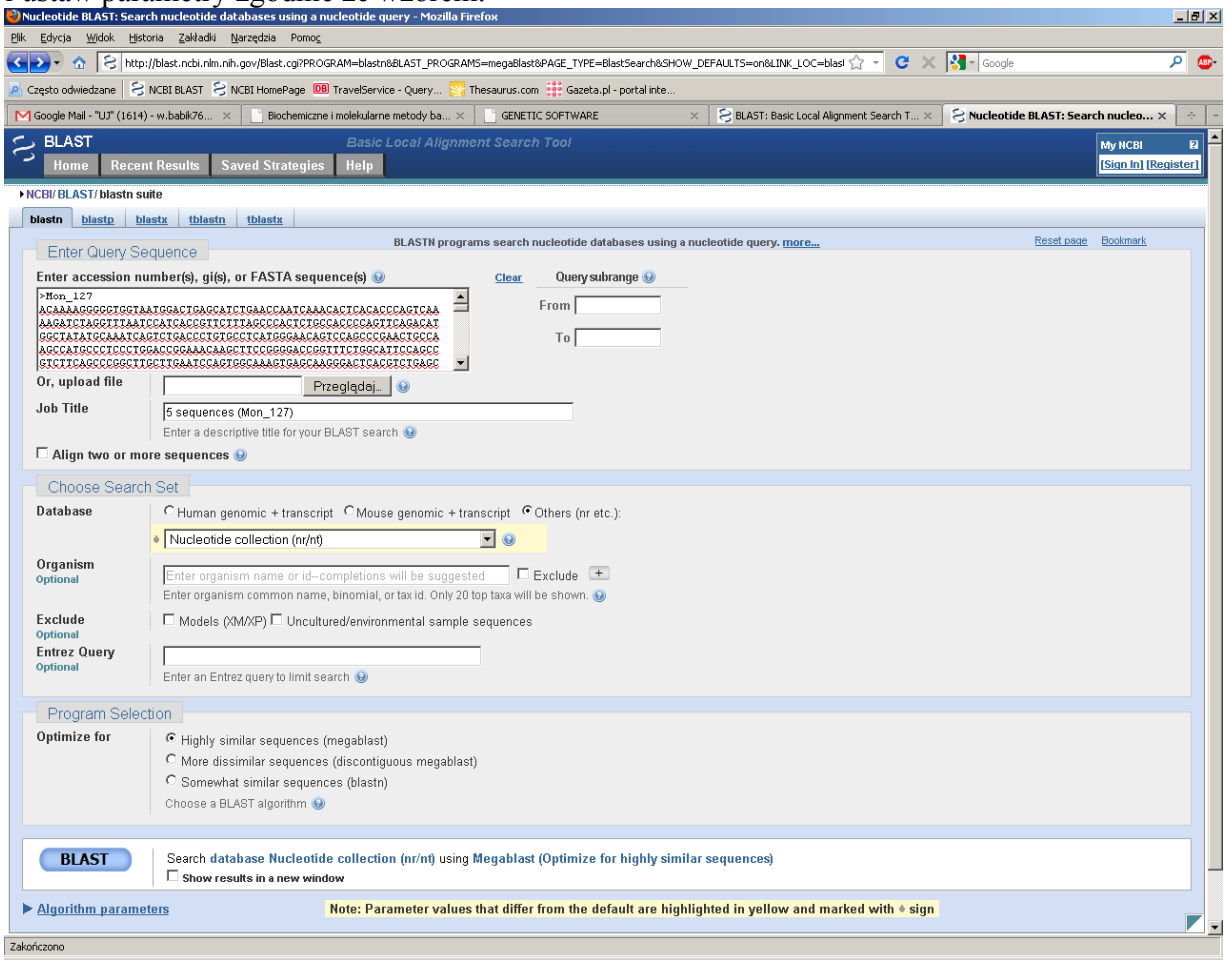
Czy nasza sekwencja ma homologi wśród bezkręgowców?
Na stronie gdzie ustawia się parametry BLASTa w polu `Organism` wpisz `Metazoa NOT vertebrata?`
Następnie spróbuj wyszukiwania z opcją `Optimize for Somewhat similar sequences (blastn)`
O czym świadczą wyniki?

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:



kliknij `BLAST`
O czym świadczą otrzymane wyniki?
Jak zmieni się wynik gdy w sekcji `Program selection` wybierzesz `Somewheat 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 próbie 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.

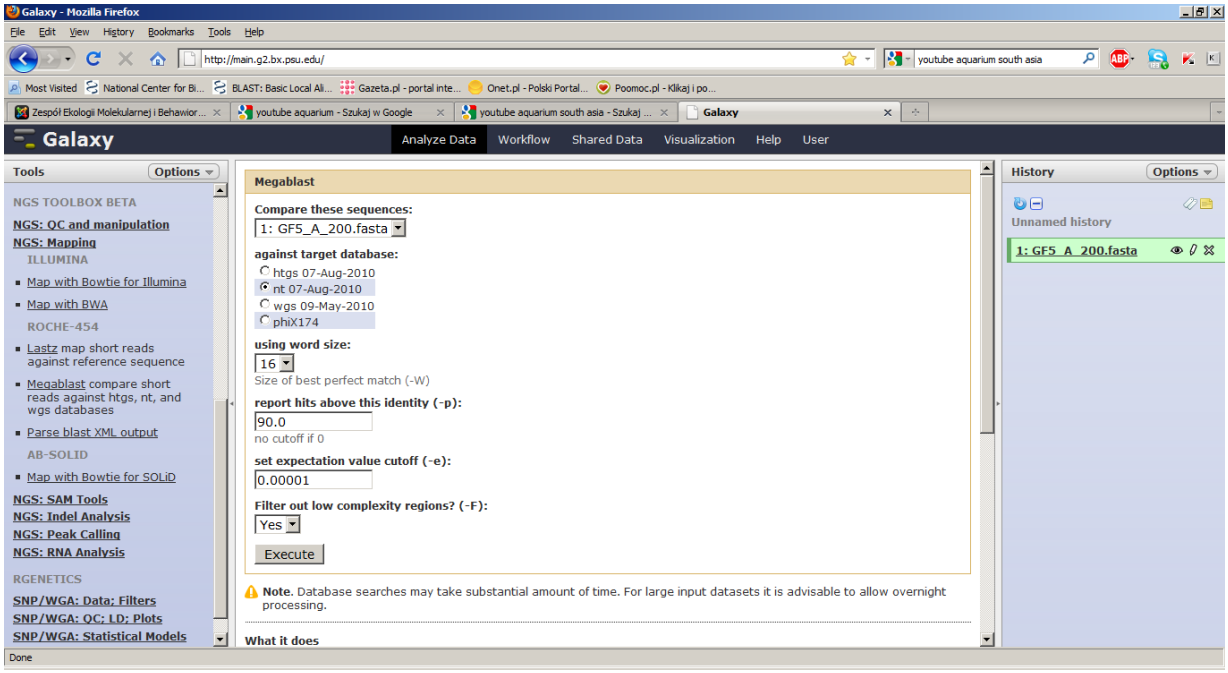
Ładujemy nasze dane do Galaxy:

`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



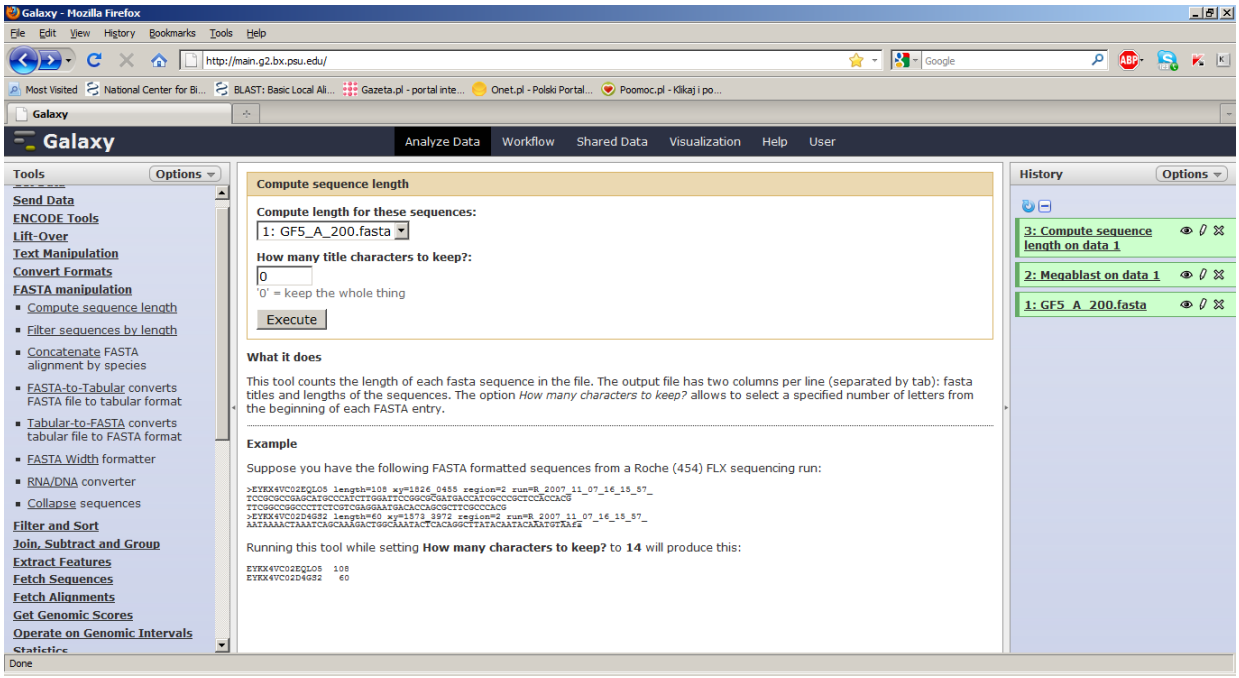
Procedura może chwilę potrwać

Filtrowanie danych

Chcemy uzyskać tabelę z długościami sekwencji

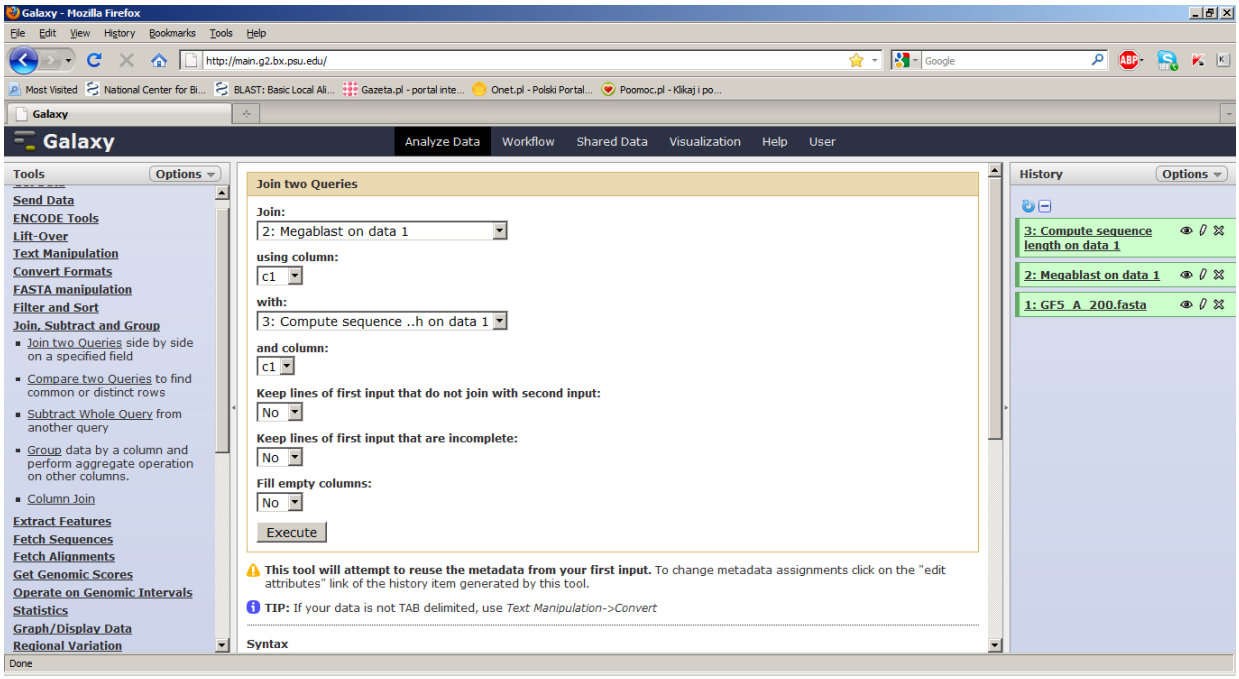
`Fasta manipulation -> compute sequence length`

`Execute`



łączenie tabeli wynikowej Megablasta oraz tabeli z długościami sekwencji:

`Join, Subtract and Group->Join Two Queries`



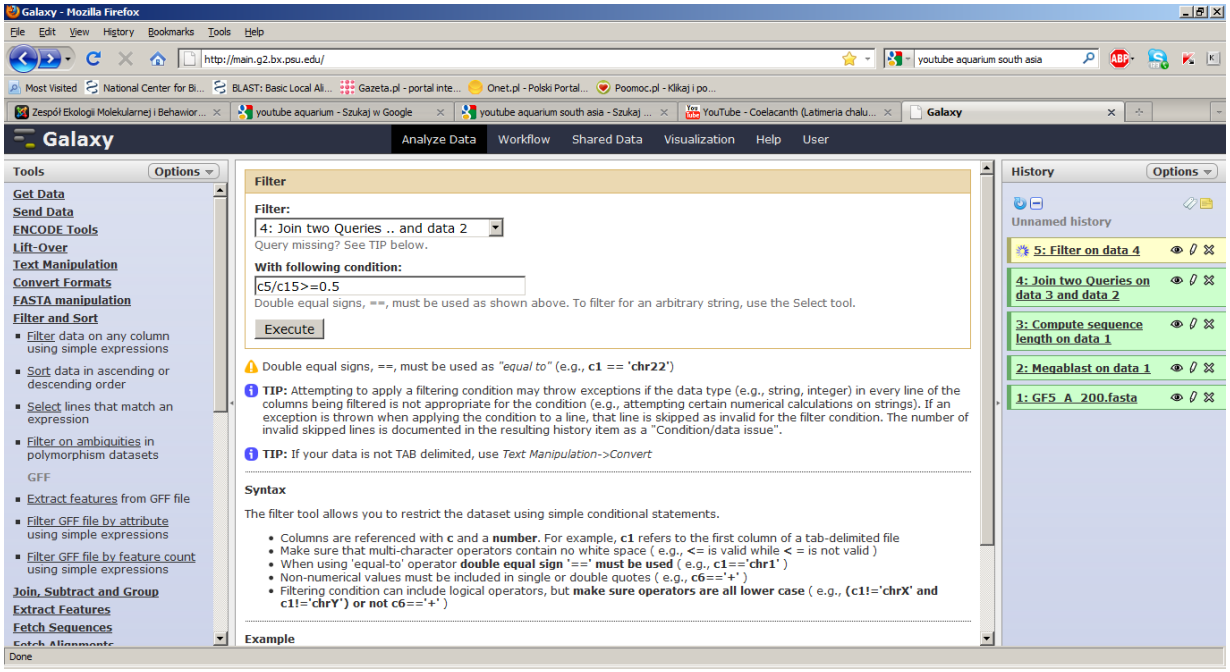
filtrowanie:

`Filter and Sort -> Filter`

kolumna c5 – długość wyrównania (alignment)

kolumna c15 – długość sekwencji

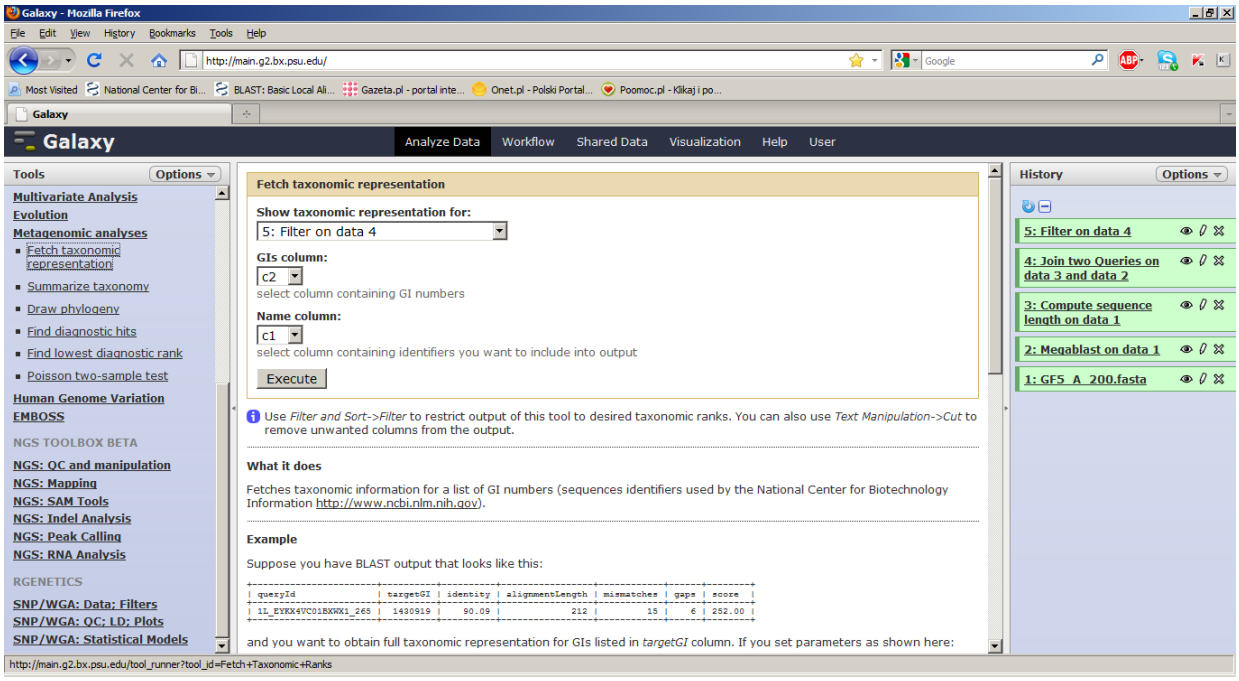
chcemy żeby dopasowanie obejmowało przynajmniej połowę długości sekwencji: >= 0.5



Taksonomia DNA

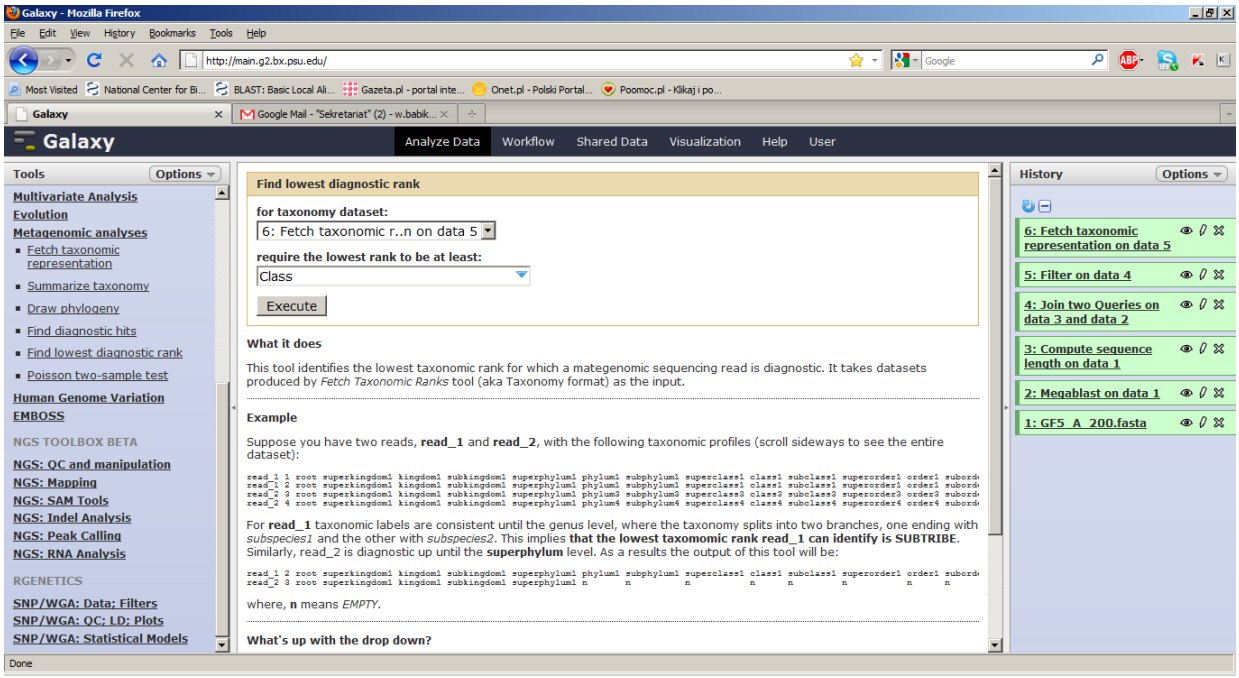
`Metagenomic analyses -> Fetch taxonomic representation`

Przypisanie reprezentacji taksonomicznej



Filtrowanie sekwencji unikatowych na określonym poziomie taksonomicznym

`Metagenomic analyses ->Find lowest diagnostics rank`



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`