

# Podstawy tworzenia drzew filogenetycznych

---

Grzegorz Góralski

Zakład Cytologii i Embriologii Roślin  
Instytut Botaniki  
Uniwersytet Jagielloński

1. Drzewa filogenetyczne
2. Etapy tworzenia drzew filogenetycznych - teoria
  - Wybór i zbieranie sekwencji
  - Dopasowanie sekwencji
  - Modele ewolucji molekularnej
  - Metody konstrukcji i szacowania wiarygodności drzew filogenetycznych
3. Część praktyczna
  - Wyszukiwanie danych w bazie GenBank
  - Wybór oprogramowania
  - Dobór modelu ewolucji molekularnej
  - Generowanie drzewa
  - Poprawki w wygenerowanym drzewie
  - Samodzielna praca
4. Horyzontalny Transfer Genów (HGT)

Dane do pobrania: [ggoralski.pl](http://ggoralski.pl)

zakładka: Metody laboratoryjne w badaniach genetycznych I

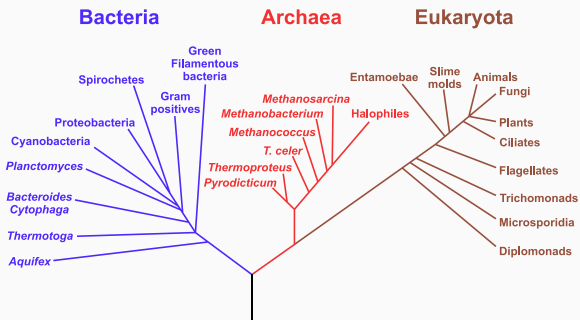
## Drzewa filogenetyczne

---

# Czym są drzewa filogenetyczne? I

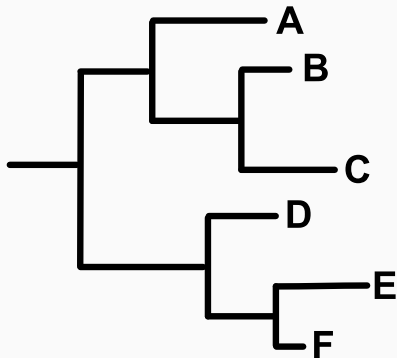
- Drzewa filogene tyczne w sposób graficzny starają się oddać pokrewieństwo organizmów

## Phylogenetic Tree of Life



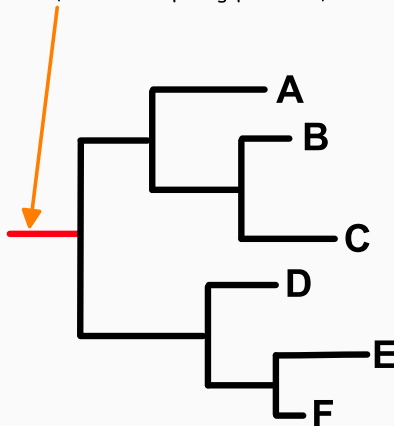
Drzewo filogenetyczne (Wikipedia)

- Tworzenie drzew filogenetycznych opiera się na badaniu podobieństw i różnic pomiędzy organizmami
- W tego typu badaniach bierze się pod uwagę cechy morfologiczne, anatomiczne itp. i/lub genetyczne
- W drzewach opartych na fragmentach DNA bierze się pod uwagę różnice w badanej sekwencji pomiędzy organizmami
- Zasadniczo im bliżej spokrewnione ze sobą organizmy, tym mniejsze powinniśmy obserwować różnice w DNA między nimi



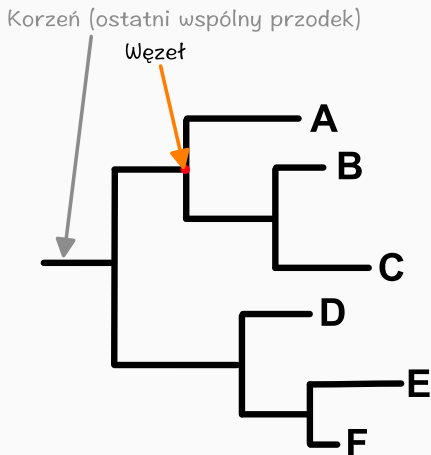
Struktura drzewa

Korzeń (ostatni wspólny przodek)



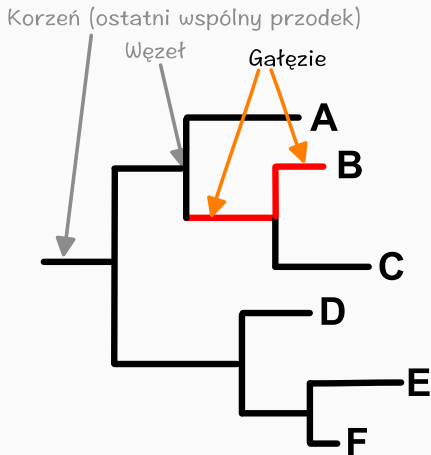
Struktura drzewa



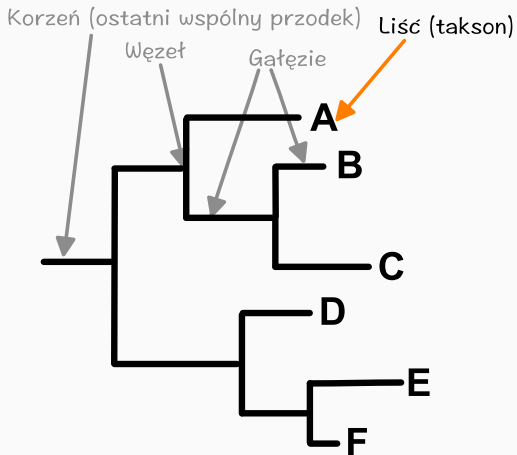


Struktura drzewa

# Struktura drzewa

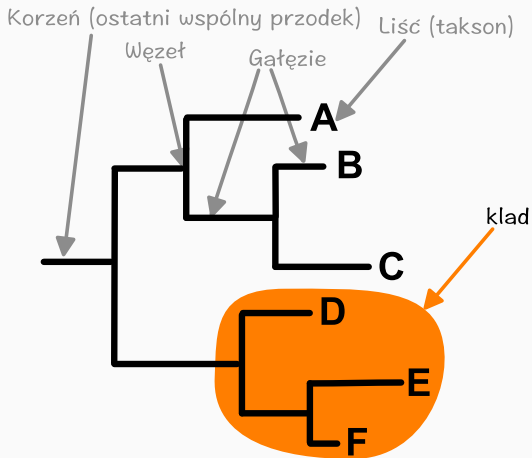


Struktura drzewa



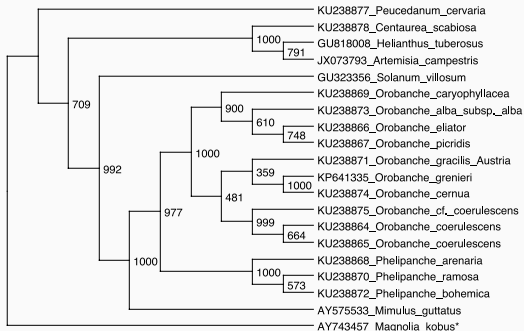
Struktura drzewa

# Struktura drzewa



Struktura drzewa

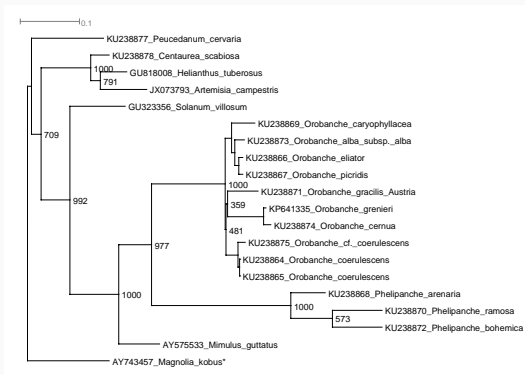
- Kladogram: pokazuje pokrewieństwa ale długość gałęzi nie pokazuje liczby mutacji



Kladogram

# Podstawowe typy drzew

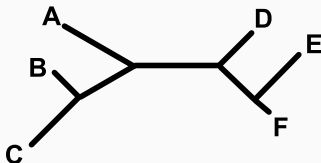
- Filogram: pokazuje pokrewieństwa, długość gałęzi odpowiada liczbie zmian



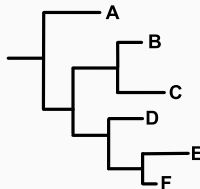
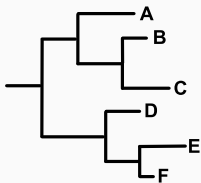
Filogram

## Nieukorzenione drzewa

- Po wygenerowaniu drzewa otrzymujemy informację o podobieństwie sekwencji ale nie o **kolejności** rozdzielania się taksonów
- Mamy więc więc **nieukorzenione drzewo**, które można przedstawić tak:



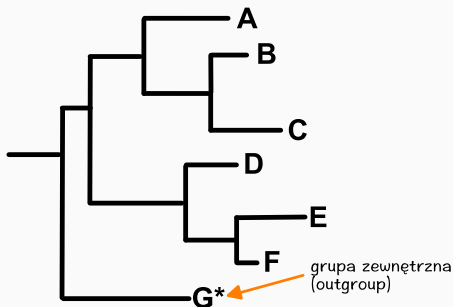
- Można je zinterpretować na wiele sposobów, np:



- itd...

# Ukorzenianie

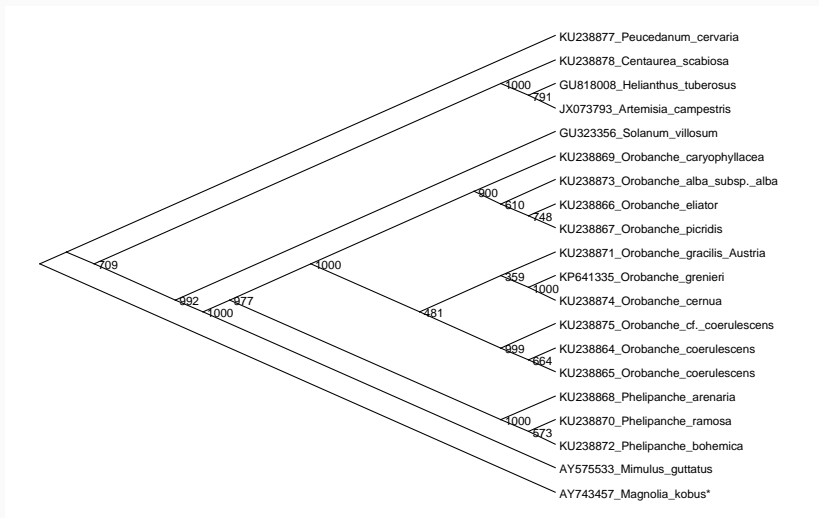
- Rozwiązaniem tego problemu jest dodanie grupy zewnętrznej czyli „outgrupy” (ang. *outgroup*)
- Grupą zewnętrzną powinien być organizm, który jest dalej spokrewniony od pozostałych, niż one między sobą. Czyli taki, który najwcześniej oddzielił się od pozostałych taksonów w grupie.
- Na przykład dla badanych gatunków człowieka (*Homo habilis*, *H. erectus*, *H. sapiens* itp) mógłby być to szympans
- Dodanie outgrupy pozwala właściwie zorientować (ukorzenić) drzewo tak, aby poszczególne rozgałęzienia odpowiadały ich kolejności w czasie.





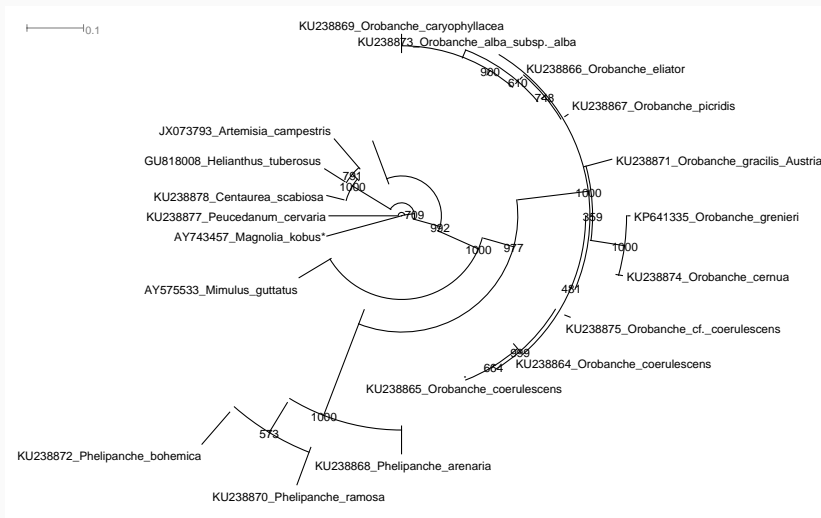
- Drzewa mogą przybierać różne formy, poza już pokazanymi na przykład:

- Drzewa mogą przybierać różne formy, poza już pokazanymi na przykład:



# Typy drzew

- Drzewa mogą przybierać różne formy, poza już pokazanymi na przykład:



Konstruowanie drzewa filogenetycznego składa się z wielu etapów:

- Wybór rodzaju sekwencji odpowiedniej dla zestawu badanych taksonów (zmiennosc, dostepnosc sekwencji etc.)
- Zebranie sekwencji (sekwencje własne, bazy danych)
- Wybór algorytmów/oprogramowania do dopasowania sekwencji, budowy drzewek oraz ich wizualizacji
- Wstępne automatyczne dopasowanie sekwencji
- Ręczne poprawki: dokładniejsze dopasowanie sekwencji, przycięcie
- Wybranie modelu ewolucji molekularnej
- Budowanie drzewa
- Tworzenie fologramu/kladogramu
- Poprawki: wskazanie outgrupy, obracanie gałęzi, wybór typu drzewa itp.

## Etapy tworzenia drzew filogenetycznych - teoria

---

Wybór sekwencji - kluczowe dla skonstruowania drzewa filogenetycznego, jest wybranie odpowiedniego rodzaju sekwencji. Należy przy tym wziąć pod uwagę kilka aspektów.

- **Homologie, analogie, ortologi i paralogi**

- Przy tworzeniu drzew filogenetycznych porównuje się sekwencje **homologiczne** (pochodzące od wspólnej sekwencji ancestralnej)
- Dwa rodzaje sekwencji homologicznych:
  - **ortologi**: sekwencje, które miały wspólnego przodka zaraz przed procesem specjacji
  - **paralogi**: sekwencje, które powstały w skutek duplikacji, czyli miały wspólnego przodka przed zduplikowaniem
- Do konstruowania drzew filogenetycznych powinno się używać ortologów.
- Podobne sekwencje mogą powstać z niespokrewnionych sekwencji w wyniku dostosowania genów do pełnienia tych samych funkcji. Takie podobieństwo nazywamy **homoplazją** a geny **analogicznymi**
- Tworzenie drzew filogenetycznych opartych o geny analogiczne nie ma sensu.

### • Tempo ewolucji

- Różne sekwencje zmieniają się w różnym tempie
- Generalnie geny ewoluują dużo wolniej niż sekwencje niekodujące (np. geny białek histonowych są bardzo konserwatywne, białek kolagenowych są bardzo zmienne)
- Geny także różnią się tempem ewolucji
- Jeśli badamy blisko spokrewnione organizmy należy wybrać szybko ewoluujące sekwencje
- Mniej zmienne odcinki DNA będą się lepiej nadawać do badań mniej spokrewnionych taksonów
- Sekwencję należy dobrać tak aby jej tempo ewolucji pozwalało odróżnić poszczególne taksony (powinny być widoczne różnice pomiędzy bliskimi taksonami)
- Jednocześnie nie może być zbyt wysokie, ponieważ wtedy trudno dopasować sekwencje a podobieństwa mogą mieć przyczynę przypadkową (pomiędzy dwoma losowymi sekwencjami powinno być ok. 1/4 zgodnych nukleotydów).

- **Aspekty praktyczne:**
  - Łatwość zaczepienia starterów itp.
- **Dostępność w bazach**
  - Jeśli nie dysponujemy wszystkimi potrzebnymi sekwencjami będzie trzeba je uzupełnić z baz danych, dlatego powinniśmy wybrać taką sekwencję, która występuje w bazach danych (cieszy się zainteresowaniem innych badaczy)



- Sekwencje mogą pochodzić z badań własnych
- Często stosuje się również sekwencje pobrane z baz danych, z których najbardziej znany jest GenBank

The screenshot shows the NCBI Nucleotide search interface. At the top, there's a navigation bar with 'NCBI', 'Resources', 'How To', and 'Sign in to NCBI'. Below that, the search bar contains 'Nucleotide' and 'orobanche ITS' with a 'Search' button. The main content area displays search results for 'orobanche ITS', showing 'Items: 41 to 60 of 2164'. The results are listed in a table with columns for item number, title, and details. The first two items are highlighted. On the right side, there are filters for 'Results by taxon' and 'Find related data'.

NCBI Resources How To Sign in to NCBI

Nucleotide Nucleotide orobanche ITS Search Help

Summary 20 per page Sort by Default order

Species: Plants (2,111), Fungi (14), Protists (6), Bacteria (14), Viruses (1), Customize ...

Molecule types: genomic DNA/RNA (2,109), mRNA (54), rRNA (1), Customize ...

Source databases: INSDC (GenBank) (2,159), RefSeq (5), Customize ...

Items: 41 to 60 of 2164

<< First < Prev Page 3 of 109 Next > Last >>

41. [Orobanche transcaucasica ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit \(rbcL\) pseudogene, partial sequence; plastid](#)  
1,244 bp linear DNA  
Accession: AY582272.1 GI: 46410832  
[GenBank](#) [FASTA](#) [Graphics](#) [PopSet](#)

42. [Orobanche lutea ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit \(rbcL\) pseudogene, partial sequence; plastid](#)  
1,212 bp linear DNA  
Accession: AY582206.1 GI: 46410766  
[GenBank](#) [FASTA](#) [Graphics](#) [PopSet](#)

Filters: [Manage Filters](#)

Send to: ▾

Results by taxon

Top Organisms [Tree](#)

- Orobanche crenata (393)
- Orobanche cernua (255)
- Orobanche gracilis (191)
- Phelipanche aegyptiaca (190)
- Phelipanche ramosa (135)
- All other taxa (1000)

More...

Find related data

Database: Select

Find items

- Sekwencje można zbierać w plikach o różnym formacie.
- Do najbardziej znanych należy format **FASTA**
- Plik FASTA jest zwykłym plikiem tekstowym w którym dane są sformatowane w następujący sposób:

```
>KC879635_Magnolia_stellata ← informacje  
CTGCTAACTCTCAGTTTGGTCCTACTTCTGGTTCATTTTGTTACTAAAACGG ← sekwencja  
AACTCAGTACCAAATGCTTGGCAATCCTTGGTAGAGCTTATTCATGATTTTCGT  
CCGGTAAACGAACAAATAGGTGGTCTTTCCGGAAATGTTCAACAAAAGTTTTTC
```

```
>AF095276_Solanum_tuberosum  
CTACTAACTCTCAGTTTGGTCCTACTTTTGGTTTATTTTGTTACTAAAAGGG  
AACTCAGTACCAAATGCTTGGCAATCCTTGGTAGAGCTTATTTATGATTTTCGT  
CCGGTAAACGAACAAATAGGTGGTCTTTCCGGAAATGTTAAACAAAAGTTTTTC
```

- Na etapie zbierania sekwencji w pliku FASTA najwygodniej jest używać do tego celu edytora tekstu.

- **Pliki tekstowe:**
  - Są uniwersalnym sposobem zapisu tekstu. Można je otworzyć i edytować w dowolnym edytorze tekstu.
  - Można je otwierać w trybie tekstowym (w terminalu)
  - Nie zawierają wizualnego formatowania tekstu (kursywa, tekst pogrubiony, kolor itp.) ale można stworzyć pliki tekstowe umożliwiające osiągnięcie tego celu, np: pliki html, tex, markdown itp.
- Do edycji plików tekstowych - używamy **edytora tekstu**, nie mylić z **procesorem tekstu** (np. Word)
- Edytor tekstu to program służący do edycji tekstu ale nie jego formatowania. Ewentualne kolorowanie, pogrubianie itp. widoczne w oknie edycji (kolorowanie składni) służy ułatwieniu pracy z tekstem ale nie jest zapisywane w samym pliku. Jest to przydatne zwłaszcza w programowaniu.

- Przykłady edytorów tekstu:
  - Pracujące w trybie tekstowym, mogą mieć także interfejs graficzny, zwykle wieloplatformowe:  
Vim, Emacs, Pico, Nano, JED.
  - Pracujące w trybie graficznym, często dla jednego systemu operacyjnego:  
TextMate (Mac OS X), Notepad++ (Windows), Gedit (Linux, Mac OS X, Windows), Kate (Linux, Mac OS X, Windows), TextWrangler (Mac OS X)
- Do dalszej pracy na naszych ćwiczeniach sugeruję użyć w zależności od systemu operacyjnego:
  - Linux: pluma lub gedit - powinny być domyślnie zainstalowane
  - Windows: Notepad++ - <https://notepad-plus-plus.org>
  - Mac OS X: TextMate 2 - <http://macromates.com>

## Symbole w sekwencjach nukleotydów

- W zapisie sekwencji nukleotydów używa się standardowo symboli standardu IUPAC:

Symbol IUPAC	znaczenie
A	Adenina
C	Cytozyna
G	Guanina
T (lub U)	Tymina (lub Uracyl)
R	A lub G
Y	C lub T
S	G lub C
W	A lub T
K	G lub T
M	A lub C
B	C lub G lub T
D	A lub G lub T
H	A lub C lub T
V	A or C or G
N	nieznany nukleotyd
- lub .	brak nukleotydu

## Na czym polega dopasowanie sekwencji („Alignment“)?

- Jeśli ustawimy pod sobą sekwencje w kolejnych liniach, zwykle uzyskamy podobny rezultat:

```
GAAGTAGCGGTATGCAATTA  
AACTAGCCGCATACAATTA  
CAAGAACTGCGCTAAACTTAGCCA  
AACTGCGCTAAACTTAGC
```

- Jak widać sekwencje „nie pasują” do siebie.

## Na czym polega dopasowanie sekwencji („Alignment“)?

- Dopasowanie sekwencji polega na takim ich ustawieniu, żeby w kolumnach kolejne nukleotydy były możliwie dobrze do siebie dopasowane. Na przykład tak:

```
---GAAGTAGCGGTATGCAATTA----  
----AACTAGCCGCATACAATTA----  
CAAGAACT-GCGCTAAACACTTAGCCA  
----AACT-GCGCTA-ACACTTAGC--
```

- Dopasowanie uzyskano wstawiając w odpowiednie miejsca znaki - odpowiadające miejscom wystąpienia indeli (insercji lub delecji)
- Sens dopasowania polega na tym, aby w kolumnach znajdowały się nukleotydy (miejsca) względem siebie homologiczne.

## Dopasowanie sekwencji - różne możliwości

- Dopasowanie sekwencji nie zawsze jest jednoznaczne. Rozważmy taki przykład:

```
TAAACGCTT
TAACCTT
```

- Sekwencje można dopasować na wiele sposobów, np:

```
TAAACGCTT  TAAACGCTT  TAAACGCTT  TAAACGCTT
TAA-C-CTT  T-AAC-CTT  TA-AC-CTT  TAA--CCTT
```

- W praktyce trzeba przy wyrównaniu brać pod uwagę różne czynniki. m. in:
  - kontekst (dopasowanie pozostałych sekwencji)
  - prawdopodobieństwo danej mutacji (np. indel mniej prawdopodobny niż substytucja, różne prawdopodobieństwa poszczególnych rodzajów substytucji)



- Ręczne dopasowanie sekwencji jest zwykle bardzo czasochłonne. Zwłaszcza jeśli są długie, są bardzo zmienne, jest ich wiele, gdy posiadają indeli.
- W praktyce stosuje się programy komputerowe, które wykorzystują odpowiednie algorytmy.
- Obecnie jest dostępna duża liczba tego typu programów, powstają także nowe.
- Różnią się one szybkością i dokładnością dopasowania.
- Trzeba jednak pamiętać, że efektywność ich działania zależy także od ustawienia odpowiednich parametrów, np. wartości „kar” za otwarcie i przedłużenie indeli.
- Wynik otrzymany po automatycznym dopasowaniu wymaga często dalszych ręcznych korekt. Zwłaszcza gdy sekwencje są bardzo zmienne i zawierają wiele indeli.

## Odległość między sekwencjami

- Z czasem, dzięki gromadzeniu mutacji, sekwencje ewoluują niezależnie od siebie różniąc się coraz bardziej.
- Zatem, im więcej różnic między sekwencjami tym organizmy z których pochodzą są mniej ze sobą spokrewnione - czyli dawniej miały wspólnego przodka. Im sekwencje są bardziej podobne, tym pokrewieństwo bliższe.
- Kluczowym zagadnieniem dla rekonstrukcji ewolucji na podstawie porównania sekwencji DNA jest określenie „odległości” między nimi, czyli określenie jak bardzo się od siebie różnią.
- Najprościej można to zrobić poprzez obliczenie udziału różnic pomiędzy dopasowanymi sekwencjami.

$$O = \frac{R}{L}$$

Gdzie: O - odległość, R - liczba miejsc z różnicami, L - liczba wszystkich miejsc

## Odległości między sekwencjami

- Można by z powyższych rozważań wyciągnąć wniosek, że tak liczona odległość powinna wzrastać liniowo wraz z upływem czasu.
- W rzeczywistości problem jest bardziej złożony.
- Wraz z upływem czasu wzrasta prawdopodobieństwo wielokrotnych mutacji w tym samym miejscu. Jeśli np. w tym samym miejscu wydarzą się mutacje:



to te dwie zmiany będą widoczne jako jedna zmiana, jedna z nich będzie „ukryta”.

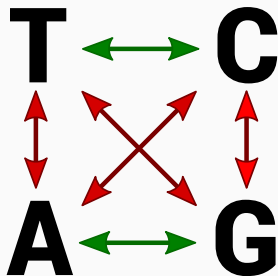
- Co więcej, może dojść to mutacji:



W takim przypadku, choć doszło do dwu zmian, nie będą one widoczne

- Tak więc odległość rośnie liniowo kiedy sekwencje różnią się nieznacznie, z czasem tempo jej wzrost stopniowo maleje.

- Aby oszacować liczbę substytucji (indele w tych rozważaniach ignorujemy), stosuje się odpowiednie modele.
- Poszczególne nukleotydy mogą się zmieniać „każdy w każdy” zgodnie z poniższym schematem (czerwony - transwersje, zielony - tranzycje):



- Każda z rodzajów substytucji przebiega z określonym prawdopodobieństwem
- Można je przedstawić w postaci maczy.
- W najprostszym modelu, zwanym modelem **Jukes-Cantora (JC69)** każda z substytucji ma takie samo prawdopodobieństwo:

	T	C	A	G
T	$-3\alpha$	$\alpha$	$\alpha$	$\alpha$
C	$\alpha$	$-3\alpha$	$\alpha$	$\alpha$
A	$\alpha$	$\alpha$	$-3\alpha$	$\alpha$
G	$\alpha$	$\alpha$	$\alpha$	$-3\alpha$

Gdzie:  $\alpha$  - prawdopodobieństwo mutacji.

Wartość  $-3\alpha$  wynika z tego, że p-stwo w każdym rzędzie = 0.

- Bardziej złożone modele uwzględniają różne prawdopodobieństwo poszczególnych substytucji - np. tranzycje (puryna na purynę lub pirymidyna na pirymidynę) zdarzają się częściej niż transwersje (puryna na pirymidynę lub odwrotnie).
- Do najbardziej znanych modeli należą m. in.: JC69, K80, TN93, GTR, HKY86
- Modele ewolucji molekularnej wyznacza się dla danego zestawu sekwencji w celu obliczania odległości ewolucyjnych między nimi oraz wiarygodności drzew filogenetycznych.
- Dobór modelu wykonuje się przy pomocy odpowiedniego oprogramowania, które pozwala także ewentualnie dopasować dodatkowe parametry takie jak wartości  $G$  (*gamma distribution*) oraz  $I$  (*proportion of invariable sites*), których nie będziemy tu omawiać.

- Istnieje szereg metod rekonstrukcji drzew filogenetycznych, np.:
  - UPGMA (Unweighted Pair-Group Method using arithmetic Averages)
  - Metoda minimalnej ewolucji (ME - Minimum Evolution)
  - Metoda najbliższego sąsiada (NJ - Neighbor-Joining)
  - Metoda największej oszczędności (MP - Maximum Parsimony)
  - Metoda największej wiarygodności (ML - Maximum Likelihood)
  - Metody Bayesowskie (Bayesian Methods)
- Tutaj nie będziemy ich szerzej omawiać. W części praktycznej użyjemy metody ML.

## Metody szacowania wiarygodności drzew filogenetycznych

- Z tworzeniem drzew filogenetycznych związane jest szacowanie ich wiarygodności.
- Do najczęściej stosowanych należy metoda **bootstrap**
- Zasada działania metody bootstrap:
  - z zestawu dopasowanych sekwencji wielokrotnie losuje się zestawy kolumn
  - niektóre kolumny mogą zostać wylosowane wielokrotnie inne w ogóle
  - w ten sposób tworzy się zestawy (fragmentów) sekwencji o długości odpowiadającej dopasowaniu początkowemu
  - dla tych zestawów konstruuje się drzewa filogenetyczne i porównuje ich topologie
  - na tej podstawie oblicza się wiarygodność poszczególnych rozgałęzień: im większe wsparcie w sekwencjach dla danego rozgałęzienia tym częściej pojawia się ono w wygenerowanych drzewach, tym większa zatem jest wiarygodność węzła
  - wartość bootstrap określa odsetek drzew w których występował dany węzeł.



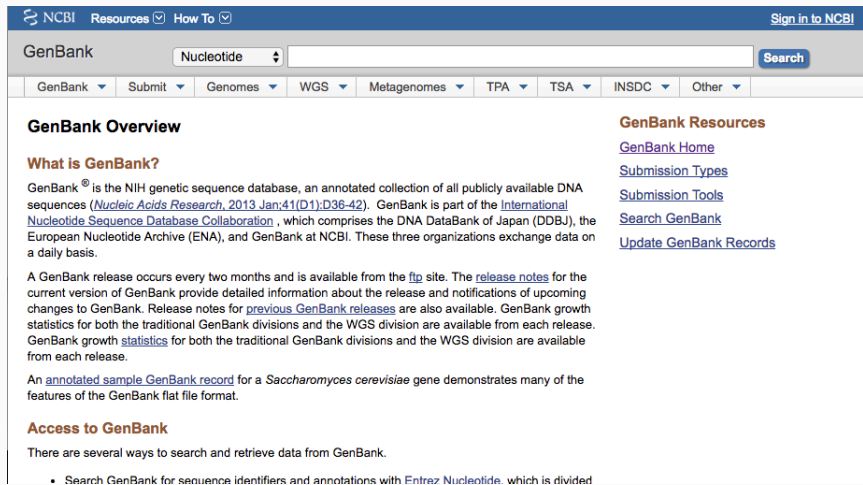
## Część praktyczna

---

- **GenBank** to ogólnie dostępna przez Internet baza zawierająca m. in. sekwencje nukleotydów dostępna pod adresem:  
<https://www.ncbi.nlm.nih.gov/genbank/>
- Inne podobne bazy to: ENA (European Nucleotide Archive) oraz DDBJ (DNA Data Bank of Japan).
- Wszystkie trzy bazy synchronizują dane.

Po wejściu na stronę GenBank-u

<https://www.ncbi.nlm.nih.gov/genbank/> widać informacje na temat bazy oraz pole wyszukiwania:



The screenshot shows the top navigation bar of the GenBank website. It includes the NCBI logo, links for 'Resources' and 'How To', and a 'Sign in to NCBI' button. Below this is the 'GenBank' header with a search dropdown menu set to 'Nucleotide' and a search button. A secondary navigation bar contains dropdown menus for 'GenBank', 'Submit', 'Genomes', 'WGS', 'Metagenomes', 'TPA', 'TSA', 'INSDC', and 'Other'. The main content area is titled 'GenBank Overview' and contains several sections: 'What is GenBank?' with a paragraph describing the database, 'GenBank Resources' with a list of links, and 'Access to GenBank' with introductory text and a bullet point.

**GenBank Overview**

**What is GenBank?**

GenBank<sup>®</sup> is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences (*Nucleic Acids Research*, 2013 Jan;41(D1):D36-42). GenBank is part of the [International Nucleotide Sequence Database Collaboration](#), which comprises the DNA DataBank of Japan (DDBJ), the European Nucleotide Archive (ENA), and GenBank at NCBI. These three organizations exchange data on a daily basis.

A GenBank release occurs every two months and is available from the [ftp](#) site. The [release notes](#) for the current version of GenBank provide detailed information about the release and notifications of upcoming changes to GenBank. Release notes for [previous GenBank releases](#) are also available. GenBank growth statistics for both the traditional GenBank divisions and the WGS division are available from each release. GenBank growth [statistics](#) for both the traditional GenBank divisions and the WGS division are available from each release.

An [annotated sample GenBank record](#) for a *Saccharomyces cerevisiae* gene demonstrates many of the features of the GenBank flat file format.

**Access to GenBank**

There are several ways to search and retrieve data from GenBank.

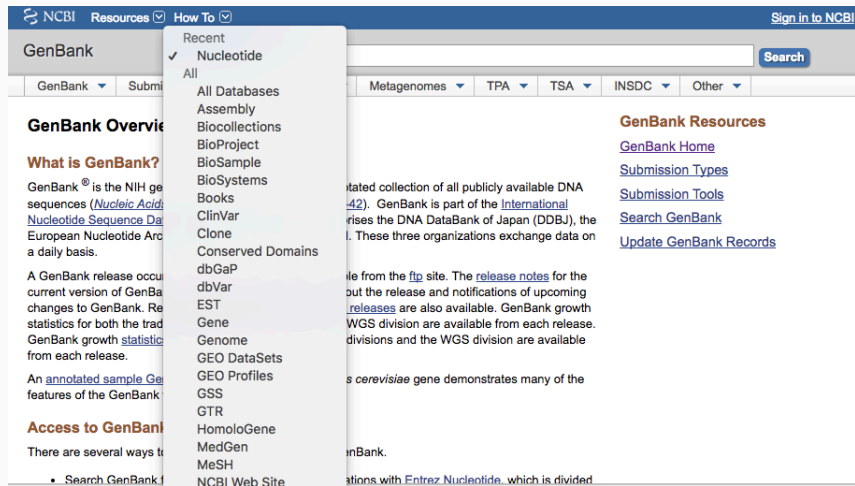
- Search GenBank for sequence identifiers and annotations with [Entrez Nucleotide](#), which is divided

**GenBank Resources**

- [GenBank Home](#)
- [Submission Types](#)
- [Submission Tools](#)
- [Search GenBank](#)
- [Update GenBank Records](#)

# GenBank - menu wyszukiwania

W zależności od tego czego szukamy , można wybrać odpowiednią opcję z rozwijanego menu. Jednak do dalszej pracy użyjemy domyślnego wyszukiwania nukleotydów („Nucleotide”)



The screenshot shows the NCBI GenBank search page. At the top, there is a navigation bar with "NCBI Resources" and "How To" menus, and a "Sign in to NCBI" link. Below this is the "GenBank" search area, which includes a search input field and a "Search" button. A dropdown menu is open, showing a list of search categories. The "Recent" section contains "Nucleotide" (checked) and "All". The "All" section lists various databases and resources such as "All Databases", "Assembly", "Biocollections", "BioProject", "BioSample", "BioSystems", "Books", "ClinVar", "Clone", "Conserved Domains", "dbGaP", "dbVar", "EST", "Gene", "Genome", "GEO DataSets", "GEO Profiles", "GSS", "GTR", "HomoloGene", "MedGen", "MeSH", and "NCBI Web Site". To the right of the search area, there are tabs for "Metagenomes", "TPA", "TSA", "INSDC", and "Other". Below the search area, there is a "GenBank Overview" section with a "What is GenBank?" heading, followed by a "GenBank Resources" section with links to "GenBank Home", "Submission Types", "Submission Tools", "Search GenBank", and "Update GenBank Records".

# GenBank - wyszukiwanie sekwencji

Sekwencje można wyszukiwać np. wg. taksonu, rodzaju sekwencji i innych słów kluczowych. Np. wyszukiwanie wszystkich sekwencji dla kukurydzy (*Zea mays*) zwróci taki wynik.

The screenshot shows the NCBI Nucleotide search interface. At the top, there are navigation links for 'NCBI Resources' and 'How To', and a 'Sign in to NCBI' button. The search bar contains 'Zea mays' and has a 'Search' button. Below the search bar, there are options for 'Create alert' and 'Advanced'. The main content area is divided into several sections:

- Species:** A list of taxonomic groups including Animals (349), Plants (927,175), Fungi (9,094), Protists (11,148), Bacteria (3,432), Archaea (6), and Viruses (442). There is also a 'Customize ...' link.
- Molecule types:** A list of molecule types including DNA/RNA (578,844), mRNA (230,668), and rRNA (24). There is also a 'Customize ...' link.
- Source databases:** A list of source databases including INSDC (GenBank) (973,000) and RefSeq (72,069). There is also a 'Customize ...' link.
- Summary:** A dropdown menu set to '20 per page' and a 'Sort by Default order' dropdown.
- Send:** A dropdown menu.
- Filters:** A link to 'Manage Filters'.
- Items: 1 to 20 of 1045335:** A list of search results. The first result is 'Found 5172751 nucleotide sequences. Nucleotide (1045335) EST (2023274) GSS (2104142)'. Below this are three numbered results, each with a checkbox, a title, and a description with accession and GI numbers, and links to 'GenBank', 'FASTA', and 'Graphics':
  - [Zea mays 18S rRNA, complete sequence](#)  
1. 1,576 bp linear rRNA  
Accession: AH001709.2 GI: 1059791817  
[GenBank](#) [FASTA](#) [Graphics](#)
  - [Zea mays cultivar BSS53 chromosome 4 clone BAC 072, complete sequence](#)  
2. 106,246 bp linear DNA  
Accession: AF528565.1 GI: 23928433  
[GenBank](#) [FASTA](#) [Graphics](#)
  - [Zea mays 22-kDa alpha zein gene cluster, complete sequence](#)  
3. 78,101 bp linear DNA  
Accession: AF031569.1 GI: 2832242
- Results by taxon:** A section with a dropdown arrow, containing 'Top Organisms [Tree]' and a list of organisms: *Zea mays* (702681), *Triticum turgidum* (137485), *Triticum urartu* (84574), uncultured organism (25645), *Pythium arrhenomanes* (10980), and All other taxa (83970). There is a 'More...' link.
- Find related data:** A section with a dropdown arrow, containing a 'Database:' label and a 'Select' dropdown menu, and a 'Find items' button.
- Search details:** A section with a dropdown arrow, containing a search query: '"Zea mays" (Organism) OR Zea

Można zawęzić wynik dla konkretnego rodzaju sekwencji, np. *trnL-trnF*:

The screenshot shows the NCBI Nucleotide search interface. The search query is "Zea mays trnL-trnF". The results are displayed in a list format, showing two items. The first item is a 492 bp linear DNA sequence, and the second is a 749 bp linear DNA sequence. Both items are described as "Zea mays trnL-trnF intergenic spacer, complete sequence; and tRNA-Phe (trnF) gene, partial sequence; chloroplast". The search results are filtered by taxon, showing "Zea mays (84)". The search details section shows the search criteria: ("Zea mays"[Organism] OR Zea mays[All Fields]) AND trnL-trnF[All Fields].

NCBI Resources How To Sign in to NCBI

Nucleotide Nucleotide Zea mays trnL-trnF Search

Create alert Advanced Help

Species Summary 20 per page Sort by Default order Send: Filters: Manage Filters

Plants (84)

Customize ...

Molecule types Items: 1 to 20 of 84

genomic DNA/RNA (84) << First < Prev Page 1 of 5 Next > Last >>

Customize ...

Source databases 492 bp linear DNA

INSDC (GenBank) (84) Accession: KM385497.1 GI: 726966080

Customize ... GenBank FASTA Graphics PopSet

Genetic compartments 749 bp linear DNA

Chloroplast (84) Accession: GQ870012.1 GI: 300089518

Plastid (84) GenBank FASTA Graphics PopSet

Sequence length

Custom range...

Release data

https://www.ncbi.nlm.nih.gov/nuccore/GQ870012.1 chloroplast

Results by taxon

Top Organisms [Tree]

Zea mays (84)

Find related data

Database: Select

Find items

Search details

("Zea mays"[Organism] OR Zea mays[All Fields]) AND trnL-trnF[All Fields]

Search

UWAGA: sekwencje są różnie opisane np. opisy mogą nie zawierać skrótowych nazw sekwencji - wtedy sekwencja może nie zostać znaleziona.

Po kliknięciu w jeden z wyników pokazują się szczegółowe informacje, zawierające m. in. takie dane jak: nazwa sekwencji, organizm, listę autorów, publikację źródłową itp. Do najważniejszych należy unikalny numer identyfikacyjny (*accession number*)

GenBank Send to: ▾

**Zea mays tRNA-Leu (trnL) gene, partial sequence; trnL-trnF intergenic spacer, complete sequence; and tRNA-Phe (trnF) gene, partial sequence; chloroplast**

GenBank **QQ870012.1**

[FASTA](#) [Graphics](#) [PopSet](#)

---

[Go to:](#)

LOCUS QQ870012 749 bp DNA linear PLN 30-AUG-2011

DEFINITION Zea mays tRNA-Leu (trnL) gene, partial sequence; trnL-trnF intergenic spacer, complete sequence; and tRNA-Phe (trnF) gene, partial sequence; chloroplast.

ACCESSION QQ870012

VERSION QQ870012.1

KEYWORDS .

SOURCE chloroplast Zea mays

ORGANISM [Zea mays](#)  
Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta; Spermatophyta; Magnoliophyta; Liliopsida; Poales; Poaceae; PACMAD clade; Panicoideae; Andropogonodae; Andropogoneae; Tripsacinae; Zea.

REFERENCE 1 (bases 1 to 749)

AUTHORS Teerawatananon,A., Jacobs,S.W.L. and Hodkinson,T.R.

TITLE Molecular evolution of the grass subfamily Panicoideae (Poaceae): based on chloroplast and nuclear DNA sequences

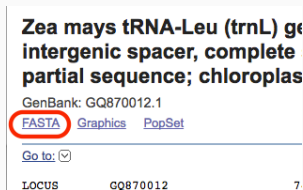
JOURNAL Unpublished

Dalej znajdziemy m. in. samą sekwencję nukleotydów:

```
JOURNAL      Unpublished
REFERENCE    2 (bases 1 to 749)
AUTHORS      Teerawatananon,A., Jacobs,S.W.L. and Hodkinson,T.R.
TITLE        Direct Submission
JOURNAL      Submitted (03-SEP-2009) Botany, Trinity College, Dublin D2, Ireland
FEATURES     Location/Qualifiers
             source                1..749
                                     /organism="Zea mays"
                                     /organelle="plastid:chloroplast"
                                     /mol_type="genomic DNA"
                                     /specimen_voucher="WK 105/THNHM"
                                     /db_xref="taxon:4577"
             misc_feature        <1..>749
                                     /note="contains tRNA-Leu (trnL), trnL-trnF intergenic
                                     spacer, and tRNA-Phe (trnF)"
ORIGIN
1 tgggcaatcc tgagccaat cccctttttg aaaaacaagt ggttctcaaa ctagaacca
61 aaggaaaagg ataggtgcag agactcaatg gaagctgttc taacgaatcg aagtaataac
121 gattaatcac agaacccata ttataatata ggttctttat tttattttta gaatgaaatt
181 aggaatgatt atgaaataga aaattcataa ttttttttta gaattattgt gaatctattc
241 caatcaaata ttgagtaatc aaatccttca attcattggt ttcgagatct ttaatttttt
301 aaaagtggat taatcggacg aggataaaga gagagtcocca ttctacatgt caatactgac
361 aacaatgaaa tttctagtaa aaggaaaatc cgtcgcacttt ataagtcgtg agggttcaag
421 tccctctatc cccaaacctt cttttattcc ctaaccatag ttgttatcct tttttctttt
481 tatcaatggg ttaagattc actagcttcc tcattctact ctttcacaaa ggagtgcgac
541 aagaactcaa tgaactctat gctattcatt aaatagatga tttctttttt attctttttt
601 tatttattag agtagagtat cggcaaggaa tctcgcattat taattcgttt ttttaagta
661 ttattaagta agccatccac aatgcatagg actaccctcc cccatttcct aattttgaat
721 ggaatacttt attgattttt tagtccctt
//
```



Jeśli chcemy pobrać sekwencję w formacie FASTA, można to zrobić na różne sposoby. Można kliknąć link „FASTA”:



**Zea mays tRNA-Leu (trnL) gene, complete intergenic spacer, complete partial sequence; chloroplast**  
GenBank: GQ870012.1  
[FASTA](#) [Graphics](#) [PopSet](#)

---

[Go to:](#)

LOCUS	GQ870012	7
-------	----------	---

# GenBank - pobieranie sekwencji w formacie FASTA

Wtedy pokaże się strona na której możemy zaznaczyć potrzebny fragment, skopiować go i wkleić ją w edytorze tekstu do pliku FASTA w której zbieramy sekwencje.

FASTA -

Send to: -

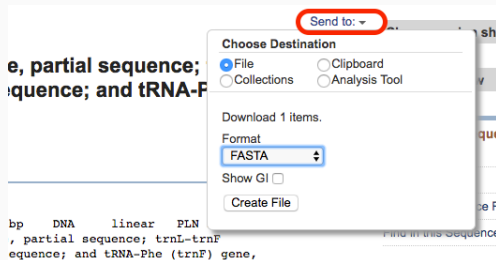
## **Zea mays tRNA-Leu (trnL) gene, partial sequence; trnL-trnF intergenic spacer, complete sequence; and tRNA-Phe (trnF) gene, partial sequence; chloroplast**

GenBank: GQ870012.1

[GenBank](#) [Graphics](#) [PopSet](#)

```
>GQ870012.1 Zea mays tRNA-Leu (trnL) gene, partial sequence; trnL-trnF intergenic spacer, complete sequence; and tRNA-Phe (trnF) gene, partial sequence; chloroplast
TGGGCAATCCTGACCCAATCCCTTTTTGAAAAACAAGTGGTTCTCAAACCTAGAACCACAAAGGAAAAGG
ATAGGTGCAGAGACTCAATGGAAGCTGTTCTAACGAATCGAACTAATAACGATTAATCACAGAACCATA
TTATAATATAGGTTCTTTATTTTATTTTGAATGAAATAGGAATGATTATGAAATAGAAAATTCATAA
TTTTTTTTTGAATTTATGTGAATCTATTCCAATCAAATATTGAGTAATCAAATCCTTCAATTCATGTT
TTCGAGATCTTTAATTTTAAAAGTGGATTAATCGGACGAGGATAAAGAGAGAGTCCCATCTACATGT
CAACTACTGACACAATGAAATTTCTAGTAAAAGGAAAATCCGTCGACTTTATAAGTCGTGAGGGTTCAAG
TCCCCTATCCCCAAACCCTCTTTTATCCCTAACCATAGTTGTTATCCTTTTTTCTTTTATCAATGGG
TTAAGATTCACTAGCTTTCCTACTTACTCTTTCACAAAGGAGTCCGACAAGAAGTCAATGAATCTTAT
GCTATTCATTAATAGATGATTTCTTTTATTTCTTTTATTTATTTAGATAGATATCGGCAAGGAA
TCTCGATTATTAATTCGTTTTTTTAAAGTATTATTAAGTAAGCCATCCACAATGCATAGGACTACCCCT
CCCATTTCCTAATTTTGAATGGAATACTTTATGATTTTTAGTCCCTT
```

Można również kliknąć „Send to” a następnie wybrać „File”, format „FASTA” i kliknąć „Create File”. Zostanie pobrany plik z sekwencją w formacie FASTA.



Wyszukaj w bazie GenBank i pobierz w formacie FASTA sekwencje:

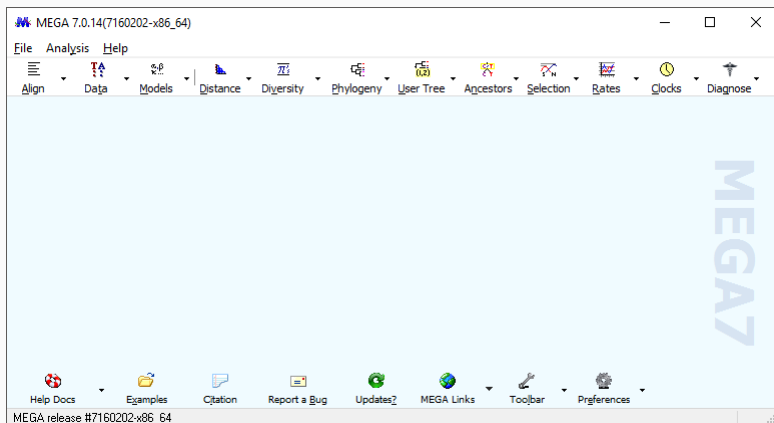
- atp6 *Magnolia stellata* nr. KC879635
- trnL-trnF *Magnolia kobus* nr. AY743457

## Wybór oprogramowania do dalszej pracy

- Dostępnych jest bardzo wiele algorytmów i implementujących je programów a także programów, które wykorzystują narzędzia dostępne przez internet
- Programy działają w formie aplikacji okienkowych lub z linii komend, niektóre na oba sposoby
- Narzędzia mogą być płatne lub darmowe
- Niektóre programy do dopasowania sekwencji: ClustalW, Mutt, Mafft, Muscle, T-Coffee
- Niektóre metody do tworzenia drzew: najbliższego sąsiada (*Neighbor-Joining*), największej wiarygodności (*Maximum Likelihood*), największej oszczędności (*Maximum Parsimony*), analiza Bayesowska (*Bayesian Interference*)
- Niektóre programy do tworzenia drzew: Phylip, IQTree, MrBayes, PhyML, RAxML, Fasttree.
- Niektóre programy do wizualizacji drzewek/ogólnego zastosowania: **MEGA**, Dendroscope, Mesquite, Jalview, PAUP\*

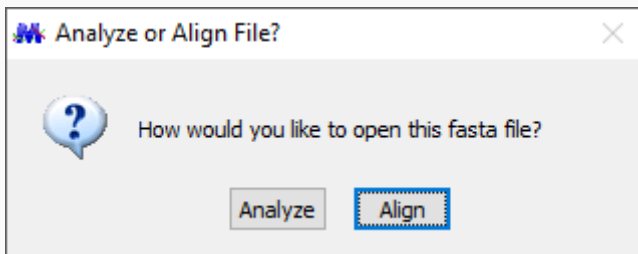
- Generalnie uważa się, że do pracy bioinformatycznej lepiej nadają się systemy UNIX-owe (Linux, Mac OS X itp.), na nie też jest dostępnych większość narzędzi bioinformatycznych ale część narzędzi działa także pod Windows.
- Ponadto systemy UNIX-owe domyślnie udostępniają wiele narzędzi niekoniecznie dedykowanych do badań bioinformatycznych ale przydatnych (np. do pracy z plikami tekstowymi)
- Na naszych warsztatach użyjemy programu MEGA pod Windows.
- Jest to darmowy program, który można pobrać ze strony:  
**<http://www.megasoftware.net>**
- Można go używać w formie okienkowej pod Windows i Mac OS X
- Jest dostępny także zestaw narzędzi do pracy w linii komend, również pod system Linux

- Po otwarciu programu pokazuje się taki interfejs:



## Otwieramy plik do wyrównania

- Otwórz w edytorze tekstu plik „atp6.fasta” z katalogu „sekwencje”, dodaj do niego pobraną sekwencję *Magnolia stellata*
- Otwórz plik w programie MEGA
- Pokaże się takie okienko, w którym wybieramy „Align ”





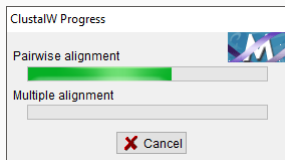
# Dopasowanie sekwencji

- Pokazuje się taki widok:

The screenshot displays the M7: Alignment Explorer (atp6.fasta) window. The interface includes a menu bar (Data, Edit, Search, Alignment, Web, Sequencer, Display, Help) and a toolbar with icons for file operations and alignment functions. Below the toolbar, there are tabs for "DNA Sequences" and "Translated Protein Sequences". The main area shows a multiple sequence alignment of DNA sequences from 20 different species. The sequences are listed in a table with columns for each position in the alignment. The sequences are color-coded by nucleotide: G (green), C (cyan), A (orange), and T (red). The alignment shows a high degree of conservation across most positions, with some gaps (indicated by dashes) and mismatches. The species listed are: 2. G8\_Orobanche\_grenieri, 3. 50\_Orobanche\_alba\_subsp.\_alba, 4. 17\_Orobanche\_picridis, 5. x12\_O.\_elatior, 6. 28\_Orobanche\_gracilis\_Austria, 7. 1\_Orobanche\_caryophyllacea, 8. JN098455\_Mimulus\_guttatus, 9. 19\_Phelipanche\_arenaria, 10. 25\_Phelipanche\_ramosa, 11. 38\_Phelipanche\_bohemica, 12. 13\_Orobanche\_coerulescens, 13. G18\_Orobanche\_cf.\_coerulescens, 14. 12\_Orobanche\_coerulescens, 15. X82388\_Helianthus\_annuus\_ssp.\_texanus, 16. z7\_Centaurea\_scabiosa, 17. ATP6f4\_Artemisia\_campestris, 18. AF095276\_Solanum\_tuberosum, 19. z3\_Peucedanum\_cervaria, 20. KC879635\_Magnolia\_stellata. At the bottom, there is a "Site #" field set to 1 and radio buttons for "with" and "w/o Gaps".

- Jak widać, część sekwencji jest „na dzień dobry” dopasowana, ale inne wymagają poprawek

- Z belki narzędziowej na górze wybieramy ikonę z literą „W” - jest to wywołanie programu **ClustalW**
- Pokazuje się okienko z postępem procesu wyrównania:



# Dopasowana automatycznie sekwencja

- W końcu widać dopasowane sekwencje:

M7: Alignment Explorer (atp6.fasta)

Data Edit Search Alignment Web Sequencer Display Help

DNA Sequences Translated Protein Sequences

Species/Abbrv	Group Name	
2. G8_Orobanche_grenieri		*** ** * * * * * * * * * * * * * * * * *
3. 50_Orobanche_alba_subsp._alba		GCCTACGTCAGCTAGGTTTGGTCCACTT
4. 17_Orobanche_picridis		-----
5. x12_O_elatior		-----
6. 28_Orobanche_gracilis_Austria		GCCTACGTCAGCTAGGTTTGGTCCACTT
7. 1_Orobanche_caryophyllacea		GCCTACGTCAGCTAGGTTTGGTCCACTT
8. JN098455_Mimulus_guttatus		GCCTACGTCAGCTAGGTTTGGTCCACTT
9. 19_Phelipanche_arenaria		GCCTACGTCAGCTAGGTTTGGTCCACTT
10. 25_Phelipanche_ramosa		GCCTACGTCAGCTAGGTTTGGTCCACTT
11. 38_Phelipanche_bohemica		GCCTACGTCAGCTAGGTTTGGTCCACTT
12. 13_Orobanche_coerulescens		GCCTACGTCAGCTAGGTTTGGTCCACTT
13. G18_Orobanche_cf._coerulescens		GCCTACGTCAGCTAGGTTTGGTCCACTT
14. 12_Orobanche_coerulescens		GCCTACGTCAGCTAGGTTTGGTCCACTT
15. X82388_Helianthus_annuus_ssp._texanus		GCCTACGTCAGCTAGGTTTGGTCCACTT
16. z7_Centaurea_scabiosa		-----
17. ATP6f4_Artemisia_campestris		-----
18. AF095276_Solanum_tuberosum		GCCTACGTCAGCTAGGTTTGGTCCACTT
19. z3_Peucedanum_cervaria		GCCTACGTCAGCTAGGTTTGGTCCACTT
20. KC879635_Magnolia_stellata		GCCTACGTCAGCTAGGTTTGGTCCACTT

Site # 22 with w/o Gaps

- Sekwencje wyglądają na (z grubsza) dopasowane, ale należy je przyciąć tak, aby miały taką samą długość.

- W tym celu zaznaczamy blok sekwencji po lewej, tak by objął część z brakującymi nukleotydami:

M7: Alignment Explorer (atp6.fasta)

Data Edit Search Alignment Web Sequencer Display Help

DNA Sequences Translated Protein Sequences

Species/Abbrv	Group Name	
2. G8_Orobanche_grenieri		*** ** *****
3. 50_Orobanche_alba_subsp._alba		*** ** *****
4. 17_Orobanche_picridis		*** ** *****
5. x12_O._elatior		*** ** *****
6. 28_Orobanche_gracilis_Austria		*** ** *****
7. 1_Orobanche_caryophyllacea		*** ** *****
8. JN098455_Mimulus_guttatus		*** ** *****
9. 19_Phelipanche_arenaria		*** ** *****
10. 25_Phelipanche_ramosa		*** ** *****
11. 38_Phelipanche_bohemica		*** ** *****
12. 13_Orobanche_coerulescens		*** ** *****
13. G18_Orobanche_cf._coerulescens		*** ** *****
14. 12_Orobanche_coerulescens		*** ** *****
15. X82388_Helianthus_annuus_ssp._texanus		*** ** *****
16. z7_Centaurea_scabiosa		*** ** *****
17. ATP6f4_Artemisia_campestris		*** ** *****
18. AF095276_Solanum_tuberosum		*** ** *****
19. z3_Peucedanum_cervaria		*** ** *****
20. KC879635_Magnolia_stellata		*** ** *****

Site # 29 with w/o Gaps

- a następnie usuwamy go (Ctrl+X lub z menu, które pojawi się po kliknięciu prawym klawiszem myszy)

- Teraz zwróć uwagę na zaznaczony na żółto fragment:

M7: Alignment Explorer (atp6.fasta)

Data Edit Search Alignment Web Sequencer Display Help

DNA Sequences Translated Protein Sequences

Species/Abbrv	Group Name	*** **	*****	*****	*		*****	****	*****
1. G15_Orobanche_cernua		G	A	T	T	A	G	T	T
2. G8_Orobanche_grenieri		G	A	T	T	A	G	T	T
3. 50_Orobanche_alba_subsp._alba		G	A	T	T	A	G	T	T
4. 17_Orobanche_picridis		G	A	T	T	A	G	T	T
5. x12_O_elatior		G	A	T	T	A	G	T	T
6. 28_Orobanche_gracilis_Austria		G	A	T	T	A	G	T	T
7. 1_Orobanche_caryophyllacea		G	A	T	T	A	G	T	T
8. JN098455_Mimulus_guttatus		G	A	T	T	A	G	T	T
9. 19_Phelipanche_arenaria		G	A	T	T	A	G	T	T
10. 25_Phelipanche_ramosa		G	A	T	T	A	G	T	T
11. 38_Phelipanche_bohemica		G	A	T	T	A	G	T	T
12. 13_Orobanche_coerulescens		G	A	T	T	A	G	T	T
13. G18_Orobanche_cf._coerulescens		G	A	T	-	A	G	T	T
14. 12_Orobanche_coerulescens		G	A	T	-	A	G	T	T
15. X82388_Helianthus_annuus_ssp._texanus		G	A	T	T	A	G	T	T
16. z7_Centaurea_scabiosa		G	A	T	T	A	G	T	T
17. ATP6f4_Artemisia_campestris		G	A	T	T	A	G	T	T
18. AF095276_Solanum_tuberosum		G	A	T	T	A	G	T	T
19. z3_Peucedanum_cervaria		G	A	T	T	A	G	T	T
20. W829635_Mammillia_stellata		G	A	T	T	A	G	T	T

Site # 7 with wo Gaps

- Wygląda na to, że nukleotydy „A” powinny się zamienić miejscami z indelami (-). Zamień je używając „wytnij” i „wklej”.

- Sekwencja atp6 jest kodująca, zasadniczo nie powinna mieć indelu długości różnej niż wielokrotność 3.
- Może to oznaczać, że badana sekwencja pochodzi z niekodującej kopii genu (pseudogenu)
- Ale przyczyną mogą też być błędy odczytu lub opracowania plików z sekwencjonera.
- W takim przypadku należałoby przyrzeć się ponownie surowym plikom z sekwencjonera.
- Teraz jednak pozostawimy sekwencje takie jakie są.

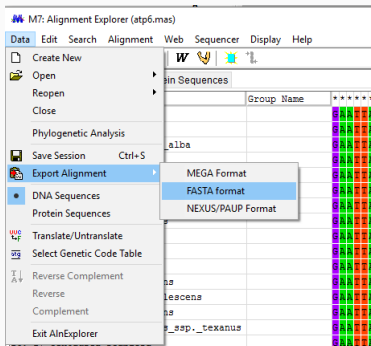




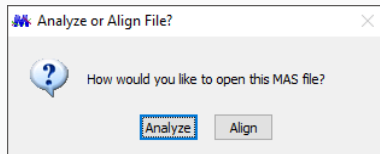


# Zapisanie pliku FASTA

- Możemy zapisać wyrównane sekwencje w pliku FASTA, najlepiej pod nową nazwą (np. atp6-aligned.fasta)

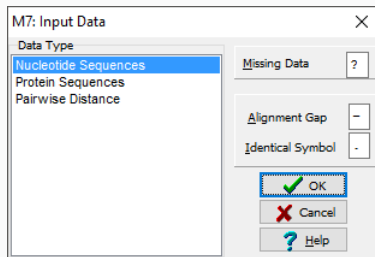


- Zapisujemy sesję w formacie MAS
- Następnie otwieramy go. Pojawia się pytanie:



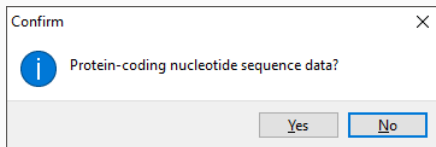
- Tym razem wybieramy „Analyze”.

- Pojawia się okienko:



- Wybieramy „Nucleotide sequence”.

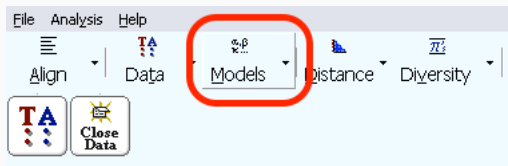
- Teraz pada pytanie czy to sekwencja kodująca:



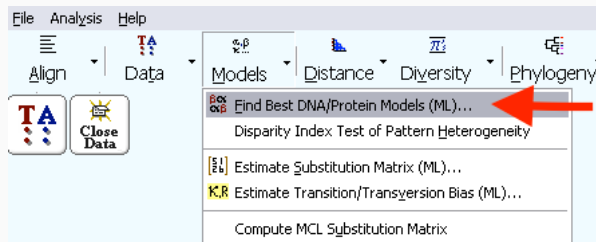
- W zasadzie jest to sekwencja kodująca, ale dla uproszczenia wybieramy opcję „No”.

# Dobór Modelu ewolucji molekularnej

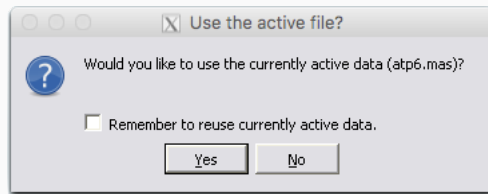
- Przed zbudowaniem drzewa, należy dobrać model ewolucji molekularnej.
- Klikamy ikonę „Models”:



- Następnie wybieramy z rozwijanego menu „Find Best DNA/Protein Models (ML)...”

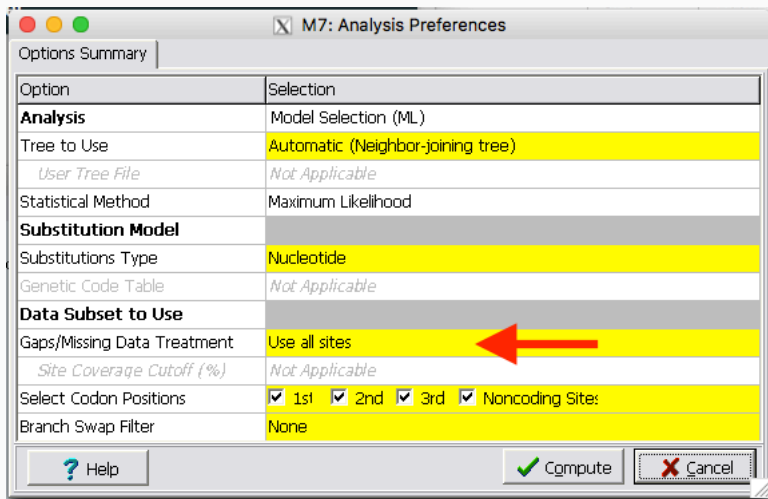


- Zatwierdzamy:



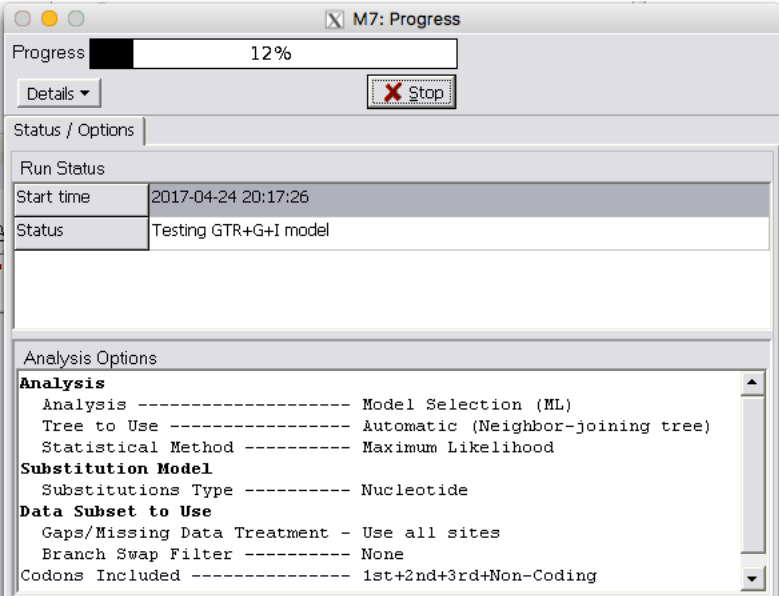
## Dobór Modelu ewolucji molekularnej

- Pokazuje się okienko z ustawieniami. Zostawiamy wartości domyślne, poza „Gaps/Missing Data Treatment” gdzie ustawiamy „Use all sites”:



# Dobór modelu ewolucji molekularnej

- Następnie trzeba chwilę poczekać aż zakończą się obliczenia:



The screenshot shows a window titled "M7: Progress" with a progress bar at 12%. Below the progress bar is a "Details" dropdown menu and a "Stop" button with a red X icon. The window is divided into sections: "Status / Options", "Run Status", and "Analysis Options".

**Run Status**

Start time	2017-04-24 20:17:26
Status	Testing GTR+G+I model

**Analysis Options**

```
Analysis
  Analysis ----- Model Selection (ML)
  Tree to Use ----- Automatic (Neighbor-joining tree)
  Statistical Method ----- Maximum Likelihood

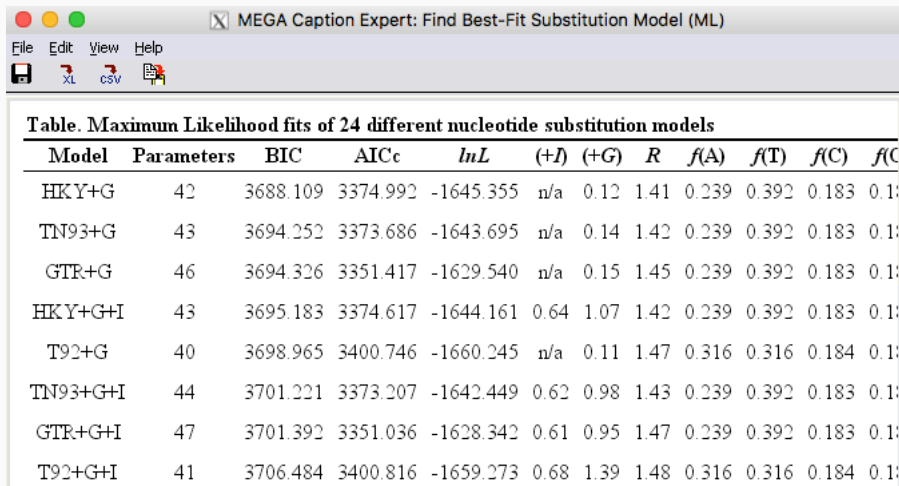
Substitution Model
  Substitutions Type ----- Nucleotide

Data Subset to Use
  Gaps/Missing Data Treatment - Use all sites
  Branch Swap Filter ----- None
  Codons Included ----- 1st+2nd+3rd+Non-Coding
```



# Wybór modelu ewolucji molekularnej

- W końcu uzyskujemy wynik:



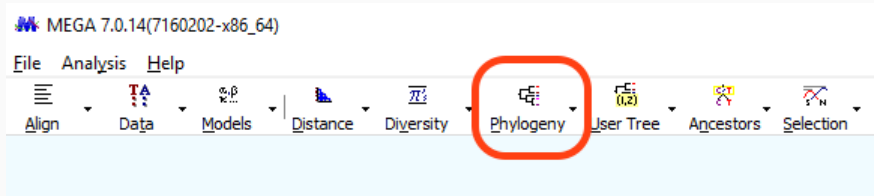
The screenshot shows a window titled "MEGA Caption Expert: Find Best-Fit Substitution Model (ML)". The window contains a table with the following data:

Table. Maximum Likelihood fits of 24 different nucleotide substitution models											
Model	Parameters	BIC	AICc	<i>lnL</i>	(+I)	(+G)	R	<i>f</i> (A)	<i>f</i> (T)	<i>f</i> (C)	<i>f</i> (G)
HKY+G	42	3688.109	3374.992	-1645.355	n/a	0.12	1.41	0.239	0.392	0.183	0.183
TN93+G	43	3694.252	3373.686	-1643.695	n/a	0.14	1.42	0.239	0.392	0.183	0.183
GTR+G	46	3694.326	3351.417	-1629.540	n/a	0.15	1.45	0.239	0.392	0.183	0.183
HKY+G+I	43	3695.183	3374.617	-1644.161	0.64	1.07	1.42	0.239	0.392	0.183	0.183
T92+G	40	3698.965	3400.746	-1660.245	n/a	0.11	1.47	0.316	0.316	0.184	0.184
TN93+G+I	44	3701.221	3373.207	-1642.449	0.62	0.98	1.43	0.239	0.392	0.183	0.183
GTR+G+I	47	3701.392	3351.036	-1628.342	0.61	0.95	1.47	0.239	0.392	0.183	0.183
T92+G+I	41	3706.484	3400.816	-1659.273	0.68	1.39	1.48	0.316	0.316	0.184	0.184

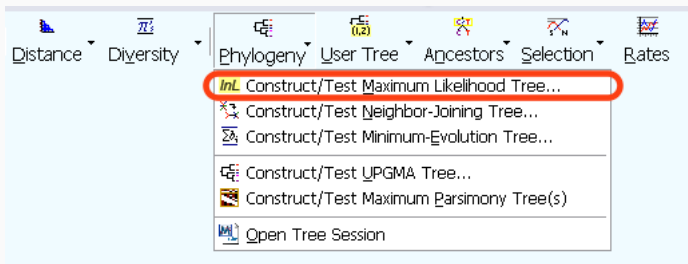
- Do budowy drzewa użyjemy model znajdujący się na pierwszym miejscu na liście, w tym wypadku: **HKY+G**

# Budujemy drzewko

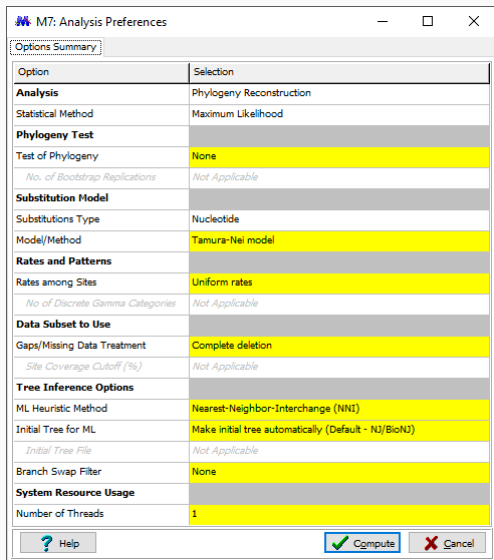
- Teraz wybieramy z menu Phylogeny



- Z menu wybierz Maximum Likelihood



- Otwiera się okno z opcjami:



Option	Selection
<b>Analysis</b>	Phylogeny Reconstruction
Statistical Method	Maximum Likelihood
<b>Phylogeny Test</b>	
Test of Phylogeny	None
<i>No. of Bootstrap Replications</i>	<i>Not Applicable</i>
<b>Substitution Model</b>	
Substitutions Type	Nucleotide
Model/Method	Tamura-Nei model
<b>Rates and Patterns</b>	
Rates among Sites	Uniform rates
<i>No of Discrete Gamma Categories</i>	<i>Not Applicable</i>
<b>Data Subset to Use</b>	
Gaps/Missing Data Treatment	Complete deletion
<i>Site Coverage Cutoff (%)</i>	<i>Not Applicable</i>
<b>Tree Inference Options</b>	
ML Heuristic Method	Nearest-Neighbor-Interchange (NNI)
Initial Tree for ML	Make initial tree automatically (Default - NJ/BioNJ)
<i>Initial Tree File</i>	<i>Not Applicable</i>
Branch Swap Filter	None
<b>System Resource Usage</b>	
Number of Threads	1

? Help      ✓ Compute      ✗ Cancel

- Ustawiamy:

M7: Analysis Preferences

Options Summary

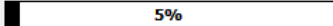
Option	Selection
<b>Analysis</b>	Phylogeny Reconstruction
Statistical Method	Maximum Likelihood
<b>Phylogeny Test</b>	
Test of Phylogeny	Bootstrap method
<i>No. of Bootstrap Replications</i>	1000
<b>Substitution Model</b>	
Substitutions Type	Nucleotide
Genetic Code Table	<i>Not Applicable</i>
Model/Method	Hasegawa-Kishino-Yano model
<b>Rates and Patterns</b>	
Rates among Sites	Gamma Distributed (G)
<i>No of Discrete Gamma Categories</i>	5
<b>Data Subset to Use</b>	
Gaps/Missing Data Treatment	Use all sites
<i>Site Coverage Cutoff (%)</i>	<i>Not Applicable</i>
Select Codon Positions	<input checked="" type="checkbox"/> 1st <input checked="" type="checkbox"/> 2nd <input checked="" type="checkbox"/> 3rd <input checked="" type="checkbox"/> Noncoding Site:
<b>Tree Inference Options</b>	
ML Heuristic Method	Nearest-Neighbor-Interchange (NNI)
Initial Tree for ML	Make initial tree automatically (Default - NJ/BioNJ)
<i>Initial Tree File</i>	<i>Not Applicable</i>
Branch Swap Filter	None
<b>System Resource Usage</b>	
Number of Threads	2


- Wartość opcji „Number of Threads” oznacza liczbę wątków używanych w obliczeniach, czyli ile równoległych obliczeń program powinien prowadzić.
- Wartość ustawiamy w zależności od liczby wykonywanych przez procesor nie kolidujących ze sobą wątków obliczeniowych.
- Na przykład procesor Pentium i7 posiadający cztery rdzenie może jednocześnie wykonywać osiem wątków.
- Ogólnie - im więcej wątków tym szybciej powinny skończyć się obliczenia.
- Jeśli nie wiemy ile rdzeni ma procesor, można pozostawić wartość 1.

# Drzewko w trakcie tworzenia

- Pojawia się okienko z postępowaniem procesu tworzenia drzewa

**M7: Progress**

Progress  **5%**

**Details** 

**Status / Options**

**Run Status**

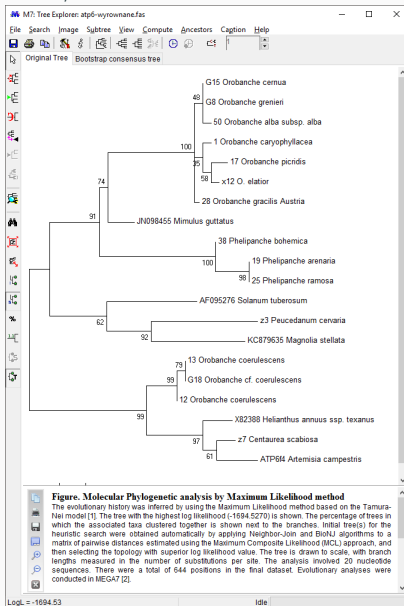
Start time	29.03.2016 17:17:46
Status	Bootstrapping ML tree
Log Likelihood	-1597.2769
Replicate No.	26 of 500

**Analysis Options**

```
Analysis
  Analysis ----- Phylogeny Reconstruction
  Statistical Method ----- Maximum Likelihood
Phylogeny Test
  Test of Phylogeny ----- Bootstrap method
  No. of Bootstrap Replications --- 500
Substitution Model
  Substitutions Type ----- Nucleotide
```

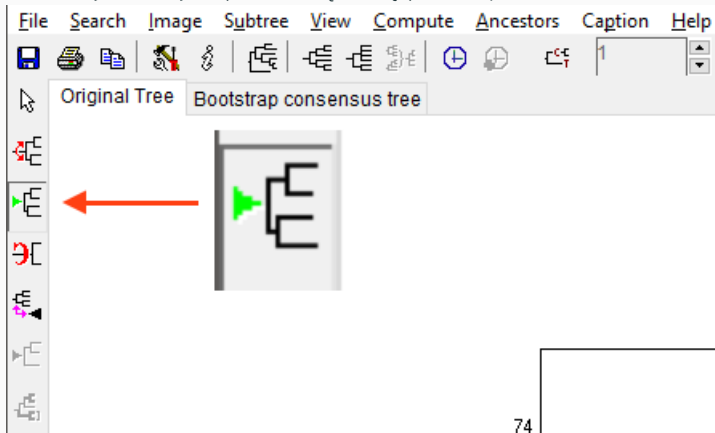
# Nieukorzenione drzewo

- W końcu widzimy drzewo, na razie nieukorzenione



# Ukorzeniecie drzewa

- Teraz należy ukorzenieć drzewo
- W tym celu wybieramy odpowiednią ikonę po lewej:

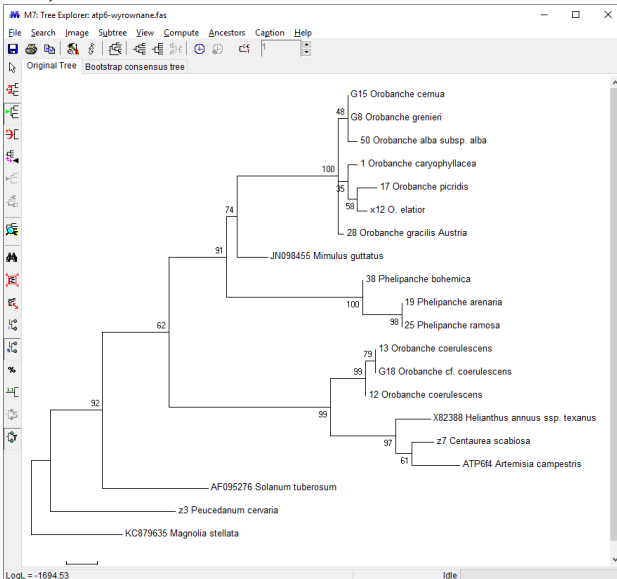


- Następnie klikamy w gałąź prowadzącą do naszej „outgrupy”, którą jest *Magnolia stellata*



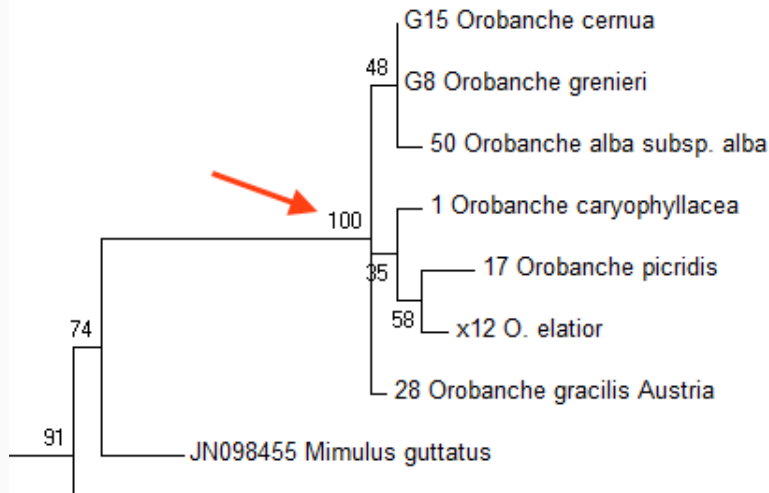
# Drzewo ukorzenie

- Teraz drzewo jest ukorzenie:



## Wartości bootstrapu

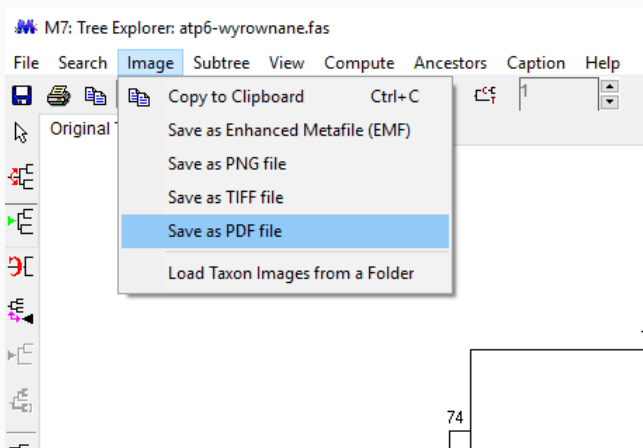
- Na drzewie widoczne są wartości bootstrapu



- Im wyższa wartość bootstrapu tym większa wiarygodność węzła.

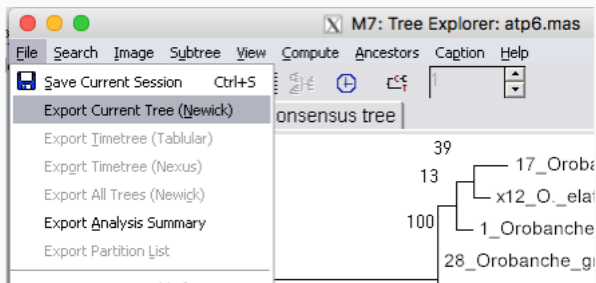
## Zapisanie drzewa jako w pliku pdf

- Plik można zapisać np. w formacie pdf:



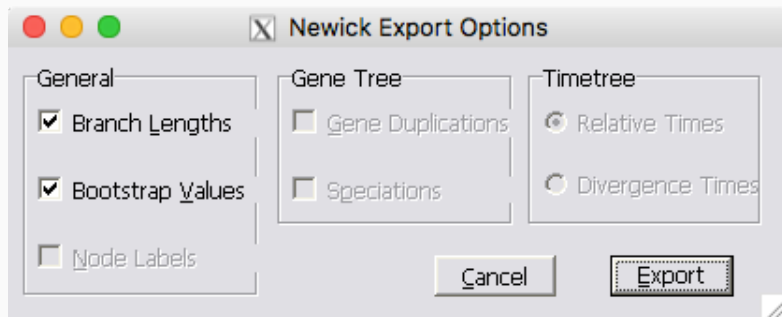
## Eksport drzewa do formatu Newick

- Mega od razu tworzy graficzną postać drzewa filogenetycznego (filogram, kladogram....) ale możemy też użyć do tego celu innego programu (iTOL, Dendroscope, Mesquite, Archeopteryx itp.).
- W taki przypadku należy zachować drzewo w formacie, który zachowa jego strukturę i parametry (długości gałęzi, wartości bootstrap) i jest odczytywany przez inne programy.
- Do takich formatów należy tekstowy format **Newick**
- Z menu okna w którym wyświetla się drzewo wybierz „Export Current Tree (Newick)“:



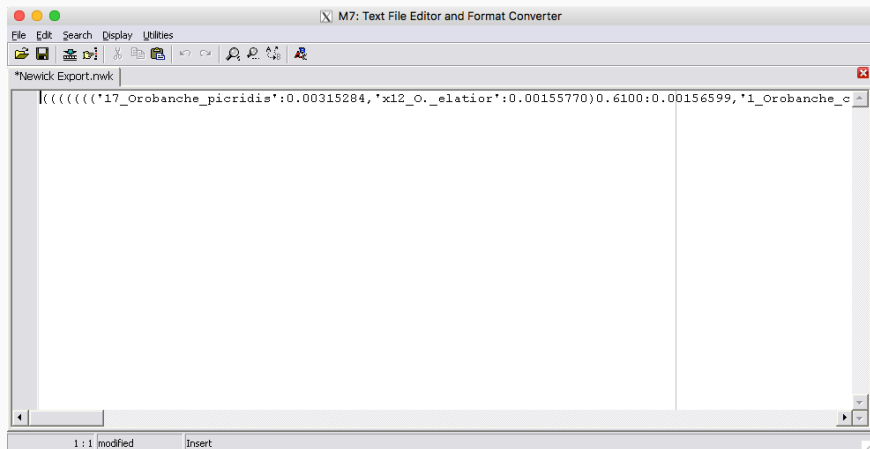
## Eksport drzewa do formatu Newick

Pokazują się opcje, jeśli chcemy wyeksportować długości gałęzi i wartości bootstrapu, odpowiednie pola powinny być zaznaczone:



# Eksport drzewa do formatu Newick

Pokazuje się wynik:

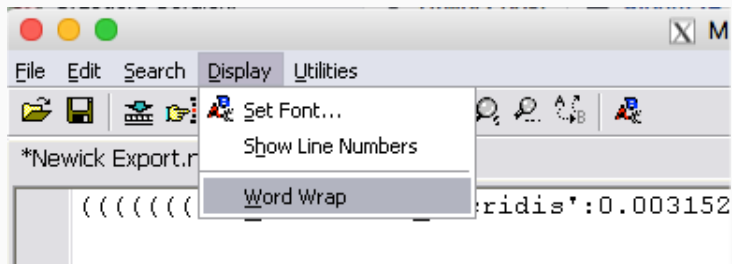


The screenshot shows a window titled "M7: Text File Editor and Format Converter" with a menu bar (File, Edit, Search, Display, Utilities) and a toolbar. The active document is named "\*Newick Export.nwk". The text content of the document is a single line of Newick format code: `|((((((( '*17_Orobanche_picridis':0.00315284, '*x12_O_elatior':0.00155770)0.6100:0.00156599, '*1_Orobanche_c`. The status bar at the bottom indicates "1 : 1" and "modified", and the cursor is in "Insert" mode.

```
|((((((( '*17_Orobanche_picridis':0.00315284, '*x12_O_elatior':0.00155770)0.6100:0.00156599, '*1_Orobanche_c
```

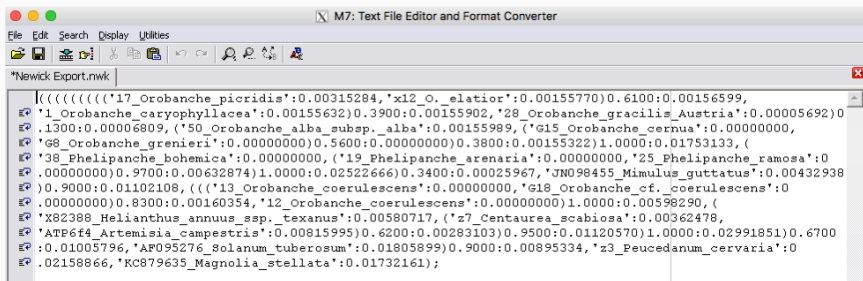
## Eksport drzewa do formatu Newick

- Tekst wyświetla się w postaci jednej długiej linii.
- Łatwiej przeanalizować jego strukturę zawijając treść. Wybieramy z menu „Display” → „Word Wrap”:



# Eksport drzewa do formatu Newick

- Teraz widać strukturę zapisu drzewa bardziej przejrzysto:

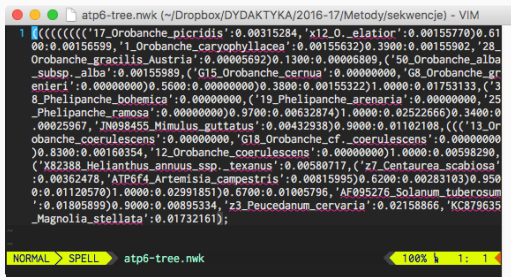


```
*Newick Export.nwk
[((( ((( ((( ('17_Orobanche_picridis':0.00315284,'x12_O_
elatior':0.00155770)0.6100:0.00156599,
'1_Orobanche_caryophyllacea':0.00155632)0.3900:0.00155902,'28_Orobanche_gracilis_Austria':0.00005692)0
.1300:0.00006809, ('50_Orobanche_alba_subsp._alba':0.00155989, ('G15_Orobanche_cernua':0.00000000,
'G8_Orobanche_grenieri':0.00000000)0.5600:0.00000000)0.3800:0.00155322)1.0000:0.01753133, (
'38_Phelipanche_bohemica':0.00000000, ('19_Phelipanche_arenaria':0.00000000,'25_Phelipanche_ramosa':0
.00000000)0.9700:0.00632874)1.0000:0.02522666)0.3400:0.00025967,'JN098455_Mimulus_guttatus':0.00432938
)0.9000:0.01102108, ((( ('13_Orobanche_coerulescens':0.00000000,'G18_Orobanche_cf._coerulescens':0
.00000000)0.8300:0.00160354,'12_Orobanche_coerulescens':0.00000000)1.0000:0.00598290, (
'X82388_Helianthus_annuus_esp._texanus':0.00580717, ('z7_Centaurea_scabiosa':0.00362478,
'ATP6f4_Artemisia_campestris':0.00815995)0.6200:0.00283103)0.9500:0.01120570)1.0000:0.02991851)0.6700
:0.01005796,'AF095276_Solanum_tuberosum':0.01805899)0.9000:0.00895334,'z3_Peucedanum_cervaria':0
.02158866,'KC879635_Magnolia_stellata':0.01732161);
```



## Eksport drzewa do formatu Newick

- Klikając na ikonę zapisu pliku lub odpowiednią pozycję menu („File” → „Save”) można zapisać zawartość do pliku tekstowego. Domyślnie pliki w formacie **Newick** mają przedłużenie **nwk**, zatem możemy plik nazwać np. **atp6-drzewo.nwk**
- Zapisany plik można otworzyć w dowolnym edytorze tekstu:



The screenshot shows a text editor window titled "atp6-tree.nwk (-~/Dropbox/DYDAKTYKA/2016-17/Metody/sekwencje) - VIM". The main content is a single line of text in Newick tree format, listing various plant species names with numerical values in parentheses, representing a phylogenetic tree. The text is as follows:

```
1 (((((((('17_Orobanche_picridis':0.00315284,'x12_0._elator':0.00155770)0.6100:0.00156599,'1_Orobanche_caryophyllacea':0.00155632)0.3900:0.00155902,'28_Orobanche_gracilis_Austria':0.00005692)0.1300:0.00006809,('50_Orobanche_alba_subsp._alba':0.00155989,('G15_Orobanche_cernua':0.00000000,'G8_Orobanche_generi':0.00000000)0.5600:0.00000000)0.3800:0.00155322)1.0000:0.01753133,('38_Phelipanche_bohemica':0.00000000,('19_Phelipanche_arenaria':0.00000000,'25_Phelipanche_ramosa':0.00000000)0.9700:0.00632874)1.0000:0.02522666)0.3400:0.00025967,'JN098455_Mimulus_guttatus':0.00432938)0.9000:0.01102108, (('13_Orobanche_coerulescens':0.00000000,'G18_Orobanche_cf._coerulescens':0.00000000)0.8300:0.00160354,'12_Orobanche_coerulescens':0.00000000)1.0000:0.00598290,('X82388_Helianthus_annuus_ssp._texanus':0.00580717,('z7_Centaurea_scabiosa':0.00362478,'ATP6f4_Artemisia_campestris':0.00815995)0.6200:0.00283103)0.9500:0.01120570)1.0000:0.02991851)0.6700:0.01005796,'AF095276_Solanum_tuberosum':0.01805899)0.9000:0.00895334,'z3_Peucedanum_cervaria':0.02158866,'KC879635_Magnolia_stellata':0.01732161);
```

At the bottom of the editor, there are status indicators: "NORMAL" and "SPELL" (with a yellow arrow pointing to "SPELL"), the filename "atp6-tree.nwk", and "100%" zoom level with a cursor icon.

- Można też w celu wizualizacji drzewa otworzyć go w jednym z programów do tego przeznaczonych.

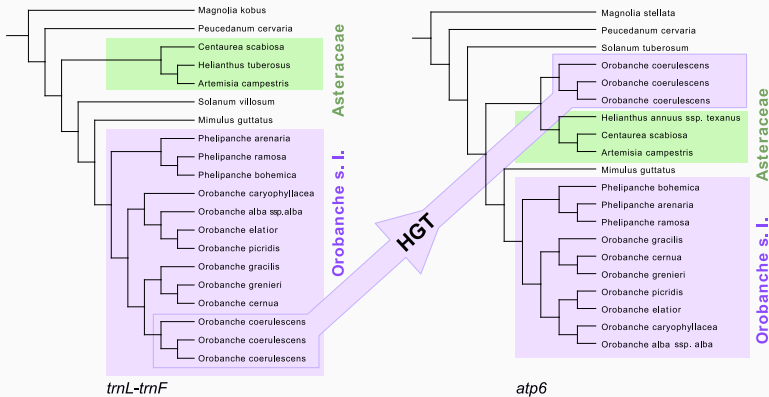
- Teraz czas na sekwencje *trnL-trnF*
- Otwórz plik **trn.fasta**, dodaj do niego sekwencję *Magnolia kobus* pobraną z GenBank
- Wstępnie wyrównaj używając ClustalW
- Zwróć uwagę na znacznie większą zmienność
- Otwórz plik z już wyrównanymi sekwencjami: **trn-aligned.fasta**
- Stwórz i ukorzeń drzewo jak w przypadku sekwencji *atp6*
- Porównaj oba drzewa
- Zwróć uwagę na *Orobanche coerulescens*
- O czym to może świadczyć?

## Horizontalny Transfer Genów (HGT)

---

# Horizontalny transfer genów

- Można zauważyć, że *Orobanche coerulescens* „przeskoczyła” z części drzewa *atp6* gdzie znajdują się jej krewniacy do części gdzie znajdują się inne gatunki
- Orobanchę* i *Phelipanchę* to rodzaje roślin pasożytniczych - pobierają one składniki odżywcze od żywicieli
- Gałąź gdzie znalazła się *Orobanche coerulescens* zawiera żywiciela tej rośliny

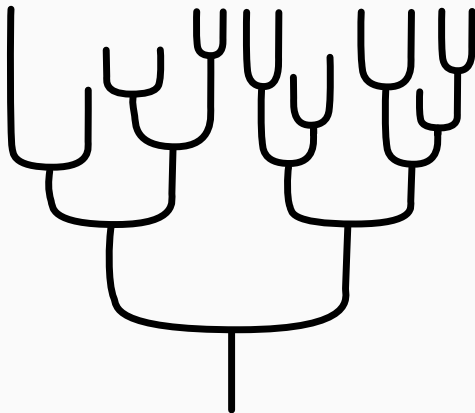


## Horizontalny transfer genów

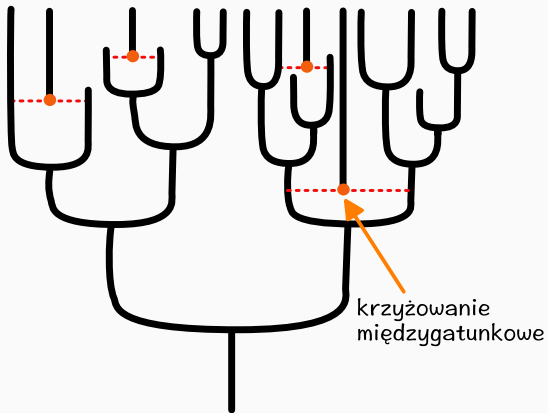
- To co widać na obrazku pokazuje efekt działania procesu zwanego **Horizontalnym Transferem Genów (*Horizontal Gene Transfer - HGT*)**
- HGT to zjawisko przenoszenia DNA pomiędzy odległymi ewolucyjnie organizmami bez udziału procesów płciowych.
- Jest to zjawisko dość częste u bakterii i pierwotniaków
- Obserwuje się go także u roślin: np. u szczepionych a także w układach pasożyt-żywiciel
- *Orobanche* (po polsku **zaraza**) właśnie są roślinami pasożytniczymi

Zaraza żółta (*Orobanche flava*)  
w Dolinie Strążyskiej



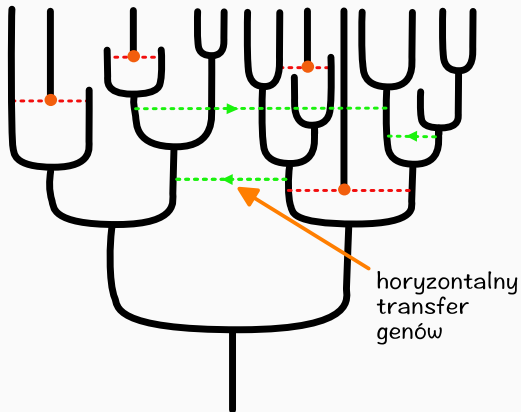


„Tradycyjne” drzewo



Drzewo uwzględniające krzyżówki międzygatunkowe





Drzewo uwzględniające krzyżówki międzygatunkowe i HGT