

# Jak zrobić drzewo filogenetyczne?

---

Grzegorz Góralski

Zakład Cytologii i Embriologii Roślin  
Instytut Botaniki  
Wydział Biologii  
Uniwersytet Jagielloński

1. Kim jestem?
2. Czym są drzewa filogenetyczne?
3. Dobór sekwencji do badań
4. Struktura drzewa filogenetycznego
5. Etapy tworzenia drzew
6. Tworzymy drzewo filogenetyczne
7. Wybór oprogramowania
8. Tworzenie drzewka w programie MEGA
9. Horyzontalny Transfer Genów (HGT)

Kim jestem?

---

dr hab. Grzegorz Góralski prof. UJ  
Zakład Cytologii i Embriologii Roślin  
Instytut Botaniki  
Wydział Biologii  
Uniwersytet Jagielloński  
ul. Gronostajowa 9 pok 2.14 (II piętro)  
e-mail: [g.goralski@uj.edu.pl](mailto:g.goralski@uj.edu.pl)  
www: [ggoralski.pl](http://ggoralski.pl)

## Czym się zajmuję?

- Liczby chromosomów i poliploidyzaacja u roślin
- Filogenetyka roślin
- Horyzontalny transfer genów
- Bioinformatyka

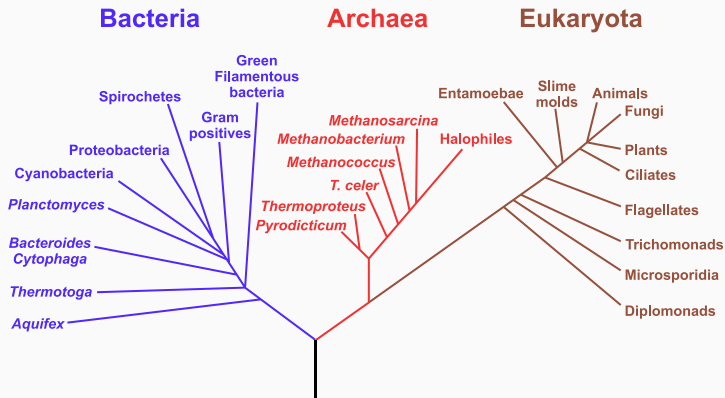
Czym są drzewa filogenetyczne?

---

## Czym są drzewa filogenetyczne? i

- Drzewa filogenetyczne w sposób graficzny starają się oddać pokrewieństwo organizmów

# Phylogenetic Tree of Life



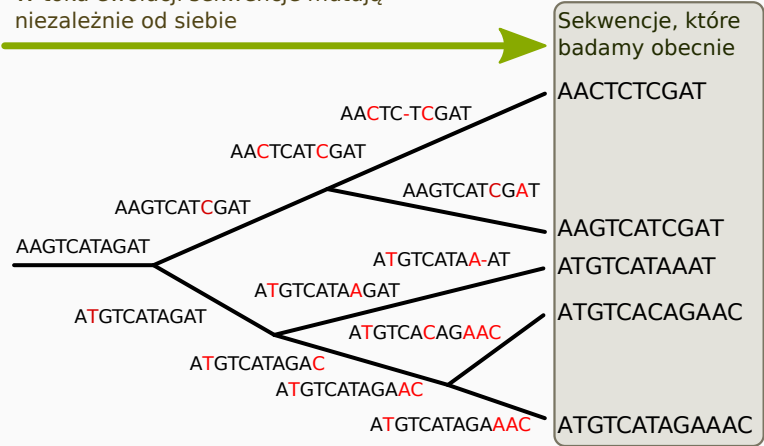
Drzewo filogenetyczne (Wikipedia)



- Ich tworzenie 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
- Można też badać sekwencje białkowe (aminokwasów) ale tu skupimy się na DNA
- Badane sekwencje otrzymane z obecnie żyjących organizmów są efektem długiej ewolucji kumulującej zmiany w DNA
- Porównując te sekwencje, staramy się odkryć ich wzajemne pokrewieństwa i ewentualnie sekwencje ich przodków

# Ewolucja sekwencji

W toku ewolucji sekwencje mutują niezależnie od siebie



## Dobór sekwencji do badań

---

## Dobór sekwencji do badań

- Pierwszym krokiem jest wybór sekwencji do badań
- Zakładamy, że jeśli porównuje się odpowiadające sobie sekwencje (na przykład konkretnego genu) to u organizmów bliżej ze sobą spokrewnionych powinny być one bardziej podobne do siebie niż w przypadku taksonów bardziej odległych ewolucyjnie.
- Wynika to z losowego gromadzenia mutacji - im więcej czasu minęło od rozdzielenia się w toku ewolucji badanych grup, tym więcej różnych mutacji powinno się skumulować w DNA.
- Takie porównania sekwencji mają oczywiście sens tylko wtedy, gdy pochodzą one od wspólnego „molekularnego” przodka, czyli są **homologiczne**
- Trzeba pamiętać, że samo podobieństwo badanych odcinków DNA jeszcze nie przesądza o ich homologiczności.
- Podobne sekwencje mogą bowiem powstać z niespokrewnionych sekwencji w wyniku dostosowania genów do pełnienia tych samych funkcji.
- Oczywiście nie nadają się one do badań filogenetycznych.

## Jakie sekwencje DNA wybrać do badań?

- Różne sekwencje zmieniają się w różnym tempie - np. niekodujące generalnie zmieniają się szybciej niż kodujące (geny) - dobór naturalny działa na nie znacznie słabiej
- Drobne zmiany fragmentów nieaktywnych DNA nie mają na ogół wpływu na organizm. Mutacje w ich obrębie mogą się więc kumulować w kolejnych pokoleniach praktycznie bez przeszkód
- W przypadku sekwencji kodujących mutacje nawet pojedynczych nukleotydów, zwłaszcza jeśli są to delecje lub insercje (zbiorczo nazywane indelami) często wpływają negatywnie na funkcjonowanie produkowanych przez gen białek lub cząsteczek RNA, zmniejszając szanse lub uniemożliwiając nosicielowi mutacji przetrwanie i przekazanie mutacji następnym pokoleniom.

## Jakie sekwencje DNA wybrać do badań?

- Mutacje genów są więc w pewnym stopniu usuwane przez dobór.
- W jakim stopniu - to zależy od rodzaju genu, miejsca i rodzaju mutacji
- Na przykład wstawienie (insercja) lub usunięcie (delecja) innej liczby nukleotydów niż trzy (długość kodonu) zmienia sposób odczytu dalszej części genu co na ogół powoduje, że kodowane przez niego białko przestaje pełnić swoje funkcje - takie mutacje są na ogół usuwane przez dobór.
- Z kolei wstawienie czy usunięcie pojedynczego aminokwasu do białka, zwłaszcza w rejonie mniej krytycznym dla jego funkcjonowania może nie wpływać na jego funkcjonowanie
- Taka mutacja może być „niewidoczna” dla doboru i przechodzić bez przeszkód do następnych pokoleń.

## Jakie sekwencje wybrać do badań?

- Geny także różnią się tempem ewolucji
- Niektóre są bardzo „wrażliwe” na zmiany, np. kodujące białka histonowe odpowiedzialne za strukturę chromatyny. Są one bardzo mało zmienne (**konserwatywne**).
- Zatem im bardziej gen jest konserwatywny tym mniej różnic zauważymy między sekwencjami pochodzącymi między badanymi organizmami.
- W innych genach, takich których struktura kodowanych przez nie białek może się zmieniać bez upośledzenia ich funkcji, zmiany kumulują się częściej. Są więc bardziej zmienne.

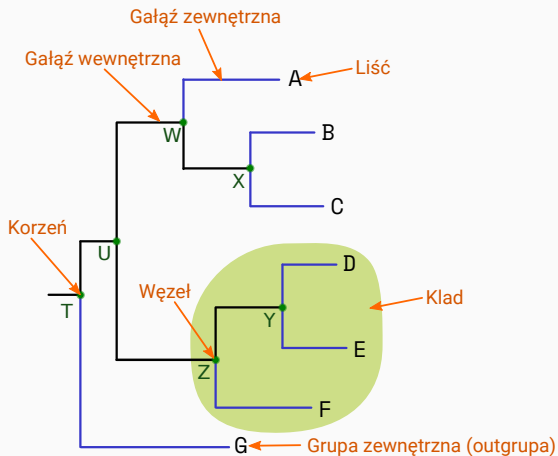
## Jakie sekwencje wybrać do badań?

- Mogłoby się wydawać, że im bardziej zmienne sekwencje tym lepiej.
- Ale zbyt dużo zmian także nie jest korzystne
- W zbyt zmiennych sekwencjach trudno znaleźć podobieństwa
- Z kolei jeśli w porównywanych sekwencjach różnic jest zbyt mało, mogą nie wystarczyć do przeprowadzenia analizy.
- Ważną konsekwencją omawianych różnic w tempie ewolucji jest to, że przy podejmowaniu decyzji którą sekwencję będzie się badać, należy wziąć pod uwagę stopień pokrewieństwa badanej grupy organizmów.
- Jeśli badamy blisko spokrewnione organizmy należy wybrać szybko ewoluujące sekwencje
- Mniej zmienne odcinki DNA będą się lepiej nadawać do badań mniej spokrewnionych taksonów



## Struktura drzewa filogenetycznego

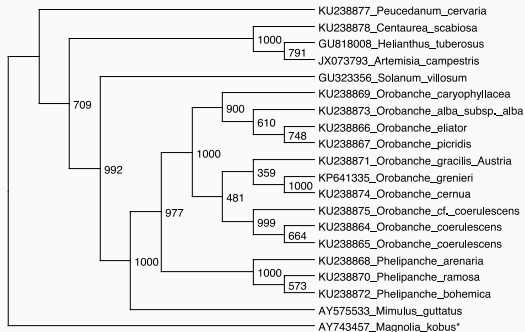
---



Struktura drzewa

- **Gałęzie** pokazują związki pomiędzy nimi. Ich długość może (w zależności od rodzaju drzewa) odpowiadać liczbie mutacji w sekwencjach nagromadzonych podczas ewolucji. Można wyróżnić gałęzie wewnętrzne prowadzące do węzłów i gałęzie zewnętrzne zakończone liśćmi.
- **Węzły** to miejsca łączenia się gałęzi - reprezentują jednostki taksonomiczne (gatunki, osobniki, odmiany itd.). Węzły wewnętrzne (nie będące liśćmi) reprezentują hipotetycznego wspólnego przodka kladu (zob. niżej)
- **Liście** są końcowymi (terminalnymi) węzłami, odpowiadają badanym sekwencjom/taksonom
- Grupa taksonów pochodzących od wspólnego przodka nazywana jest **kladem**. Niekoniecznie poszczególne klady wyróżnia się wizualnie na drzewie, ale jest to termin stosowany w opisie zależności filogenetycznych.

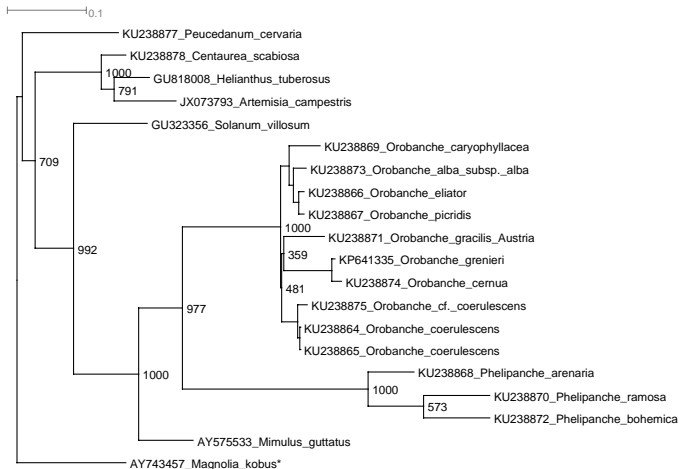
- Kladogram: pokazuje pokrewieństwa ale długość gałęzi nie pokazuje liczby mutacji



Kladogram

# Podstawowe typy drzew

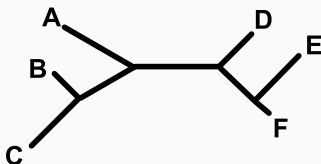
- Filogram: pokazuje pokrewieństwa, długość gałęzi odpowiada liczbie zmian



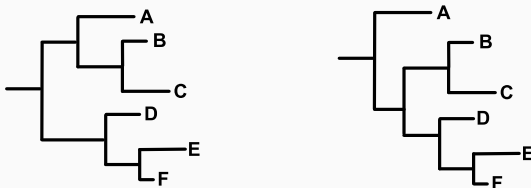
Filogram

## 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:



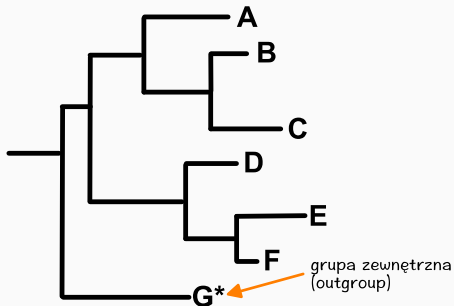
- Można je zinterpretować na wiele sposobów, np:



- itd...

# Ukorzenianie

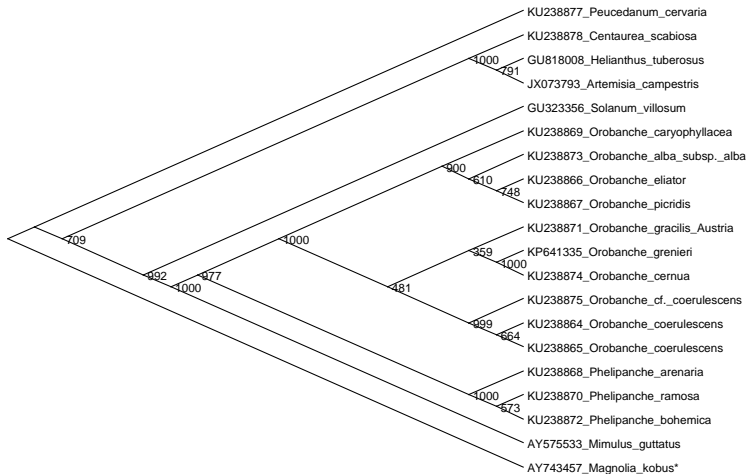
- Rozwiązaniem tego problemu jest dodanie grupy zewnętrznej czyli „outgrupy” (ang. *outgroup*)
- Grupą zewnętrzną powinien być organizm, który jest dalej spokrewniony od pozostałych, niż one między sobą. Czyli taki, który najwcześniej oddzielił się od pozostałych taksonów w grupie.
- Na przykład dla 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.



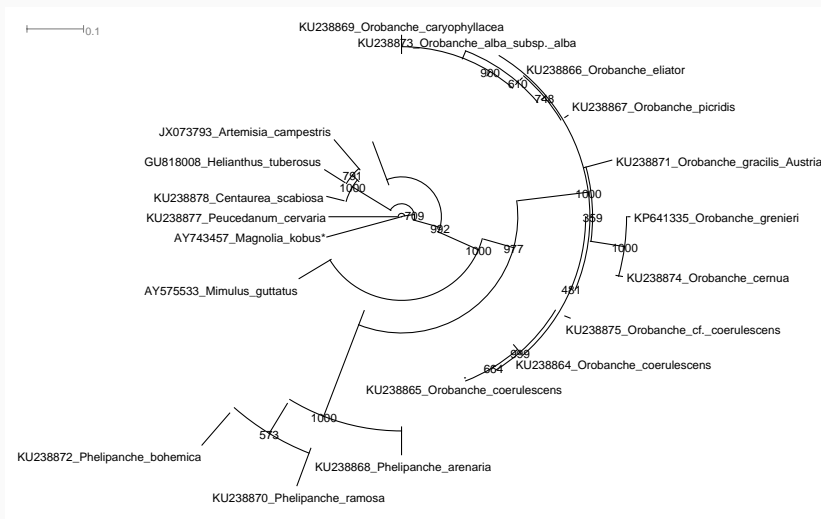
- Drzewa mogą przybierać różne formy, poza już pokazanymi na przykład:



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



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



## Etapy tworzenia drzew

---

- Wybór rodzaju sekwencji odpowiedniej dla zestawu badanych taksonów (zmienność, dostępność 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: dopasowania sekwencji, przycięcie
- Wybranie modelu ewolucji molekularnej
- Budowanie drzewa
- Poprawki drzewa: wskazanie outgrupy, obracanie gałęzi, wybór typu drzewa itp.

Tworzymy drzewo filogenetyczne

---

- Wybór sekwencji
  - **Tempo ewolucji**
    - 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).
    - Generalnie geny ewoluują dużo wolniej niż sekwencje niekodujące
    - Pomiędzy genami także występują duże różnice (np. geny białek histonowych są bardzo konserwatywne, białek kolagenowych są zmienne)
  - **Aspekty praktyczne:**
    - Łatwość badań molekularnych (replikacji DNA 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)

- W poniższym przykładzie wybrałem gen **atp6** - mitochondrialna sekwencja kodująca podjednostkę 6 syntazy ATP

- 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 is a navigation bar with 'NCBI', 'Resources', 'How To', and 'Sign in to NCBI'. Below this is a search bar with 'Nucleotide' selected and 'orobanche ITS' entered. A 'Search' button is to the right. Below the search bar, there are links for 'Create alert' and 'Advanced', and a 'Help' link.

The main content area is divided into several sections:

- Species:** A list of taxonomic groups including Plants (2,111), Fungi (14), Protists (6), Bacteria (14), Viruses (1), and a 'Customize ...' link.
- Molecule types:** A list of molecular types including genomic DNA/RNA (2,109), mRNA (54), rRNA (1), and a 'Customize ...' link.
- Source databases:** A list of databases including INSDC (GenBank) (2,159) and RefSeq (5), with a 'Customize ...' link.
- Summary:** A dropdown menu set to '20 per page' and a 'Sort by Default order' dropdown.
- Items:** A list of search results. The first two items are visible:
  - Item 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. Links: GenBank, FASTA, Graphics, PopSet.
  - Item 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. Links: GenBank, FASTA, Graphics, PopSet.
- Filters:** A link to 'Manage Filters'.
- Results by taxon:** A section showing 'Top Organisms' with a 'Tree' link. The list includes:
  - Orobanche crenata* (393)
  - Orobanche cernua* (255)
  - Orobanche gracilis* (191)
  - Phelipanche aegyptiaca* (190)
  - Phelipanche ramosa* (135)
  - All other taxa (1000)A 'More...' link is also present.
- Find related data:** A section with a 'Database:' dropdown menu set to 'Select' and a 'Find items' button.



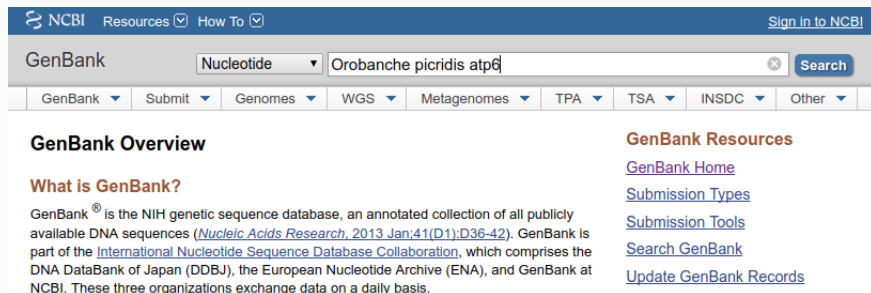
- Sekwencje można zbierać w plikach o różnym formacie.
- Do najbardziej znanych należy format **FASTA**
- Plik FASTA jest zwykłym plikiem tekstowym w którym dane są sformatowane w następujący sposób:

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

```
>AF095276_Solanum_tuberosum  
CTACTAACTCTCAGTTTGGTCCTACTTTTGGTTTATTTTGTTACTAAAAAGGG  
AACTCAGTACCAAATGCTTGGCAATCCTTGGTAGAGCTTATTTATGATTTTCGT  
CCGGTAAACGAACAAATAGGTGGTCTTTCCGGAAATGTTAAACAAAAGTTTTTC
```

# Wyszukiwanie sekwencji

- W GenBank-u można wyszukiwać sekwencję na różne sposoby. Pokażę dwa podstawowe.
- Można wyszukiwać według nazwy:



NCBI Resources How To Sign In to NCBI

GenBank Nucleotide Orobanche picridis atp6 Search

GenBank Submit Genomes WGS Metagenomes TPA TSA INSDC Other

## 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.


## GenBank Resources

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

- Można podawać nazwy taksonu, sekwencji itp.

- Jeśli wyników jest wiele, pokazuje się ich lista, wtedy klikamy na wybraną sekwencję.

## Items: 6

 Found 8 nucleotide sequences. Nucleotide (6) GSS (2)

- [Orobanche picridis isolate 17 tRNA-Leu \(trnL\) gene, partial sequence; trnL-trnF intergenic spacer, complete sequence; and tRNA-Phe \(trnF\) gene, partial sequence; chloroplast](#)

1.

865 bp linear DNA

Accession: KU238867.1 GI: 1028916351

[Taxonomy](#)

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

- [Orobanche picridis clone 17 ATPase subunit 6 \(atp6\) gene, partial cds; mitochondrial](#)

2.

642 bp linear DNA

Accession: KU180463.1 GI: 1025818453

[Protein](#) [Taxonomy](#)

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

- Jeśli wynik jest jeden, od razu otwiera się w oknie.

- Strona z wynikiem wyszukiwania zawiera szereg informacji.

NCBI Resources How To Sign in to NCBI

Nucleotide   Help

Advanced

GenBank Send to:

**Orobanche picridis clone 17 ATPase subunit 6 (atp6) gene, partial cds; mitochondrial**

GenBank: KU180463.1

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

---

Go to:

LOCUS	KU180463	642 bp	DNA	linear	PLN 10-MAY-2016
DEFINITION	Orobanche picridis clone 17 ATPase subunit 6 (atp6) gene, partial cds; mitochondrial.				
ACCESSION	KU180463				
VERSION	KU180463.1				
KEYWORDS	.				
SOURCE	mitochondrion Orobanche picridis				
ORGANISM	<a href="#">Orobanche picridis</a>				
	Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta; Spermatophyta; Magnoliophyta; eudicotyledons; Gunneridae; Pentapetalae; asterids; lamiids; Lamiales; Orobanchaceae; Orobancheae; Orobanche.				
REFERENCE	1 (bases 1 to 642)				

**Related information**

- [Protein](#)
- [Taxonomy](#)
- [PopSet](#)

- Na dole znajduje się sama sekwencja:

```

      /codon_start=1
      /product="ATPase subunit 6"
      /protein_id="ANC68065.1"
      /translation="LLTSLVLLFVHFVTKKGGGKSVPNAFQSVLELIYDFVPLVNE
QIGGLSGNVKQOFFPCISVTFTFSLFRNLQGMIPYSFTVTSHFIVTLGLSFSFIGIT
IVGFQKNGLHFLSFLPAGVPLPLAPFLVLELIPHCFRALSGLIRLFANMMAGHSLV
KILSGFAWTMLCMNDLLYFIGDPGLFIVLALTGLELGV AISQAHVSTISICIIY"
ORIGIN
   1 ctactcactc tcagtttggg cctacttttt gttcattttg ttactaaaaa gggaggagga
  61 aagtcagtac caaatgcttt tcaatccgtg ttagagctta tttatgattt tgtgccgaac
 121 ctggtaaacc acaaataggt tggtcttttc ggaaatgtga aacaacagtt tttcccttgc
 181 atctcgggta cttttacttt ttcggttatt cgtaactctc agggatgatg accttatagc
 241 ttcacagtaa caagtcattt tatcgttact ttgggtctct cattttctct tttttattggc
 301 attactatag tgggatttca aaaaaatggg cttcattttt taagcttctc attaccgcga
 361 ggagtcccac tgccggttagc acctttttta gtactccttg agctaatacc tcattgtttt
 421 cgcgatttaa gcttaggaat acgtttattt gctaatatga tggccgggtca tagtttagta
 481 aagattttaa gtgggttcgc ttggactatg ctatgtatga atgatctttt atatttcata
 541 ggggatcctg gtcctttatt tatagttcct gcattaaccg gtcttgaatt aggtgtagct
 601 atatcacaag ctcatgtttc tacgatctca atctgtattt ac
//

```

- Aby wyświetlić sekwencję w formacie FASTA, można wybrać (na górze strony z wynikami, poniżej nagłówka) „FASTA”

## **Orobanche picridis c cds; mitochondrial**

GenBank: KU180463.1

[FASTA](#)

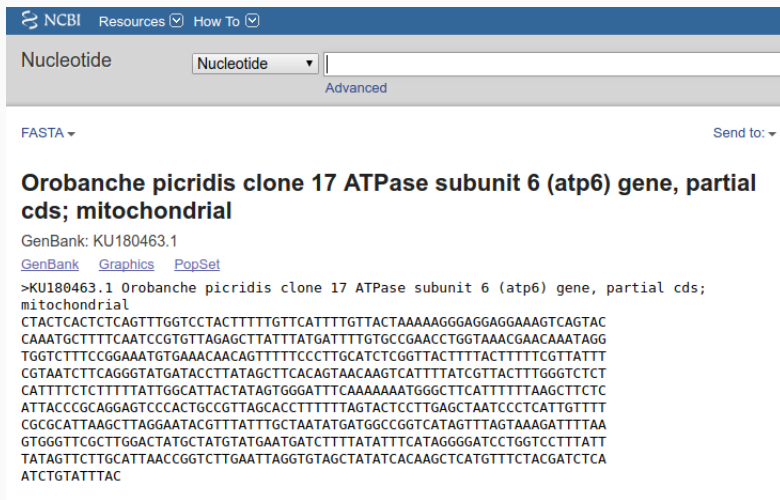
[Graphics](#)

[PopSet](#)

---

[Go to:](#)

- Wyświetla się wynik w formacie FASTA



NCBI Resources How To

Nucleotide Nucleotide Advanced

FASTA Send to:

**Orobanche picridis clone 17 ATPase subunit 6 (atp6) gene, partial cds; mitochondrial**

GenBank: KU180463.1

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

>KU180463.1 Orobanche picridis clone 17 ATPase subunit 6 (atp6) gene, partial cds; mitochondrial

```
CTACTCACTCTCAGTTGGTCCTACTTTTTGTTTCATTTTGTACTAAAAAGGGAGGAGGAAAGTCAGTAC
CAAATGCCTTTTCAATCCGTGTAGAGCTTATTTATGATTTTGTGCCGAACCTGGTAAACGAACAAATAGG
TGGTCTTTCCGGAAATGTGAAACAACAGTTTTTCCCTTGCCATCTCGGTTACTTTTACTTTTTTCGTTATTT
CGTAATCTTCAGGGTATGATACCTTATAGCTTCACAGTAACAAGTCATTTTATCGTTACTTTGGGCTCTC
CATTTTCTCTTTTTATTGGCATTACTATAGTGGGATTTCAAAAAATGGGCTTCATTTTTTAAAGCTTCTC
ATTACCCGAGGAGTCCCACTGCCGTTAGCACCTTTTTTAGTACTCCTTGAGCTAATCCCTCATTGTTTT
CGCGCATTAAAGCTTAGGAATACGTTTTATTTGCTAATATGATGGCCGGTCATAGTTTGTAAAGATTTTAA
GTGGGTTCCGCTGGACTATGCTATGATGAATGATCTTTATATTTTCATAGGGGATCCTGGTCTTTATT
TATAGTTCTTGCAATTAACGGGCTTGAATTAGGTGTAGCTATATCACAAAGCTCATGTTTCTACGATCTCA
ATCTGTATTTAC
```

- Wynik można zapisać w formacie FASTA wybierając odpowiednie opcje z menu po prawej stronie:

FASTA ▾

**Orobanche picridis clone 17 ATPase subunit 6 (atp6) cds; mitochondrial**

GenBank: KU180463.1

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

>KU180463.1 Orobanche picridis clone 17 ATPase subunit 6 (atp6) gene, mitochondrial

CTACTCACTCTCAGTTTGGTCCTACTTTTTGTTTCATTTTGTACTAAAAAGGGAGGAGGAAAGTCAGTAA  
CAAATGCTTTTCAATCCGTGTTAGAGCTTATTTATGATTTTGTGCCGAACCTGGTAAACGAACAAATAC  
TGGTCTTTCGGAAATGTGAAACAACAGTTTTCCCTTGCATCTCGTTACTTTTACTTTTTCGTTATT  
CGTAATCTTCAGGGTATGATACCTTATAGCTTCACAGTAACAAGTCATTTTATCGTTACTTTGGGTC  
CATTTTCTCTTTTTATTGGCATTACTATAGTGGGATTTCAAAAAAATGGCTTCATTTTTTAAGCTTCT  
ATTACCCGAGGAGTCCCACTGCCGTTAGCACCTTTTTTAGTACTCCTTGAGCTAATCCCTCATTGTT  
CGCGCATTAAGCTTAGGAATACGTTTATTTGCTAATATGATGGCCGGTCATAGTTTAGTAAAGATTTT  
GTGGGTTTCGCTGGACTATGCTATGATGAATGATCTTTTATATTTTCATAGGGGATCCTGGTCCTTTA  
TATAGTTCTTGCATTAACCGGTCTTGAATTAGGTGTAGCTATATCACAAAGCTCATGTTTCTACGATCTC  
ATCTCTATTAC

Send to: ▾

- Complete Record
- Coding Sequences
- Gene Features

**Choose Destination**

- File
- Clipboard
- Collections
- Analysis Tool

Download 1 items.

Format

FASTA ▾

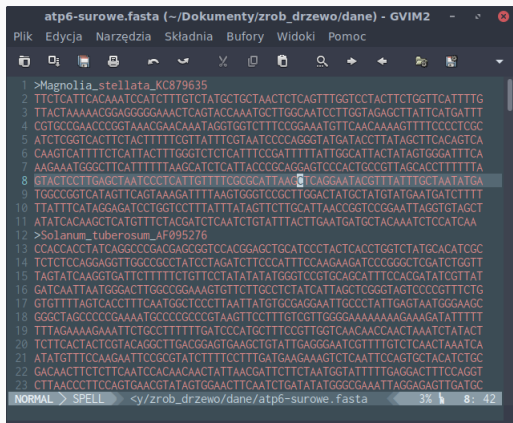
Show GI

Create File



## Zapis w pliku FASTA

- Można też zaznaczyć opis oraz sekwencję, skopiować i wkleić do ulubionego edytora tekstu. Pod Windows może to być np. Notepad++, tu jest pokazany gVim:



```
atp6-surowe.fasta (~/.Dokumenty/zrob_drzewo/dane) - GVIM2
Plik Edycja Narzędzia Składnia Bufory Widoki Pomoc

1 >Magnolia_stellata_KC879635
2 TTCTCAATCCAAATCCATCTTGTCTATGCTGCTAACTCTCAGTTTGGTCTACTCTGTGGTTCATTTTGG
3 TTAATAAAAAACGGAGGGGAACTCAGTACCAAAATGCTTGGCAATCCTTGGTAGAGCTATTTCATGATTT
4 CGTGCCGAAACCCGGTAAACGAACAAATAGTGGTCTTTCGGAAATGTTCAACAAAAGTTTTCCCTCGC
5 ATCTCGGTCACTTCTACTTTTCGTATTTTCGTAATCCCAAGGATGATACCTTATAGCTTACAGTCA
6 CAAGTCATTTTCTCATTACTTTGGGTCTCTCAITTCGATTTTTATTGGCACTACTATAGTGGGATTTCA
7 AAGAAATGGGCTTCATTTTTAAGCATCTCAATCCCGCAGGAGTCCCACTGCCGTAGCACCTTTTTTA
8 GTACTCCTTGAGCTAATCCCTCATTGTTTTCGCGCATTAAAGTACAGGAATACGTTTATTGTCAATATGA
9 TGGCCGGTCATAGTTCAGTAAAGATTTTTAAGTGGGTCCGCTGGACTATGCTATGATGAATGATCTTTT
10 TTATTTTCATAGGAGATCCTGGTCTTTTATTATAGTCTTGCATTAACCGGTCCGGAATAGGTGAGCT
11 ATATCAACAGCTCATGTTTCTACCATCTCAATCTGATTTACTTGAATGATGCTACAAATCTCCATCAA
12 >Solanum_tuberosum_AF095276
13 CCACCACCTATCAGGCCCGACGAGCGGTCCACGGAGTGCATCCCTACTCACCTGGTCTATGCACATCGC
14 TCTCTCCAGGAGGTTGGCCGCCTATCTCAGATCTTCCCAATTTCCAAGAAGTCCCGGGCTCGATCTGGTT
15 TAGTATCAAGGTGATTTCTTTCTGTTCCTATATATATGGGTCGTGCAGCATTTCCACGATATCGTAT
16 GATCAATTAATGGGACTTGGCCGGAAGTGTCTTGCCTCTATCATTAGCTCGGTAGTCCCGTTTTCTG
17 GTGTTTTAGTCACTTTCATAGGCTCCCTTAATTTATGTGCGAGGAATTTGCCCTATTGAGTAAATGGGAAGC
18 GGGCTAGCCCCCGAAAAATGCCCGCCCGTAAGTTCCTTTGTGTTGGGGAAAAAAGAAAGATATTTTTT
19 TTTAGAAAAAGAAATCTGCCTTTTGTGATCCCATGCTTTCCGTTGGTCAACAAACCAACTAAATCTATACT
20 CTTCACACTCGTACAGGCTTGACGGAGTGAAGCTGATATTGAGGGAATCGTTTTGTCTCAACTAAATCA
21 ATATGTTTCCAAGAAATCCCGCTATCTTTTCCCTTGTGATGAAGAAAGTCTCAATCCAGTGCATACATCTGC
22 GACAATCTCTTCAATCCACAACAATAAAGCATCTTCTAATGGTATTTTGAGGACTTCCAGGT
23 CTTAACCCCTCCAGTGAACGTATAGTGGAACTCAATCTGATATATGGGCGAAATAGGAGAGTTGATGC

NORMAL > SPELL <y/zrob_drzewo/dane/atp6-surowe.fasta < 3% 8: 42
```

- Może to być wygodniejsze rozwiązanie, ponieważ na końcu i tak chcemy mieć wszystkie sekwencje w jednym pliku.

## Wybór oprogramowania

---

- 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, MrBayes, PhyML, RAxML
- 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 ale część narzędzi działa także pod Windows.
- Pokażę jak zrobić proste drzewo filogenetyczne (pomijam etap wyboru modelu ewolucji) z użyciem 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

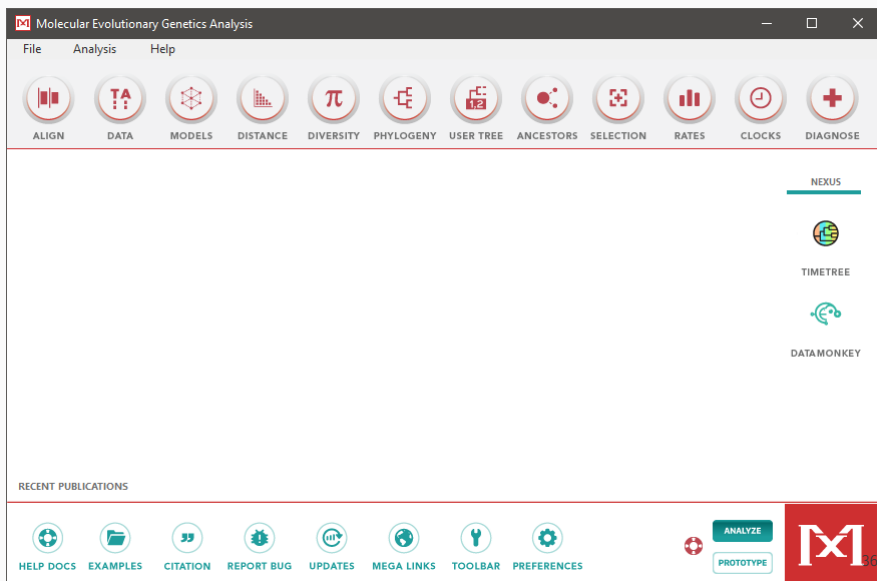
## Tworzenie drzewka w programie MEGA

---

- Pobierz plik „dane.zip”, który znajdziesz pod adresem:  
[http://ggoralski.pl/?page\\_id=3276](http://ggoralski.pl/?page_id=3276)
- Rozpakuj plik - powstanie katalog „dane”
- Pobierz i zainstaluj program Mega ze strony:  
<https://www.megasoftware.net/>

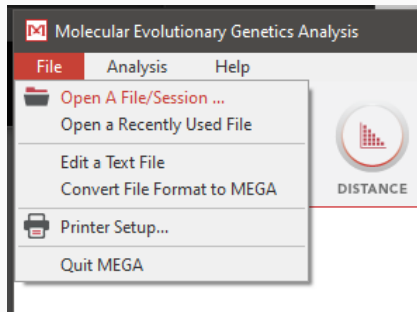
# MEGA - interfejs

- Po otwarciu programu pokazuje się taki interfejs:

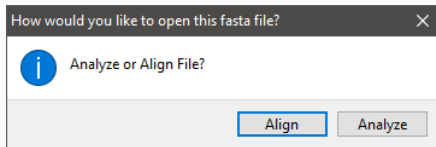


## Otwieramy plik do wyrównania

- Otwieramy plik z sekwencjami „atp6-surowe.fasta” z katalogu „dane”



- Pokaże się takie okienko, w którym wybieramy „Align”







- Teraz trzeba sekwencje dopasować.
- Dopasowanie polega na tym, żeby poszczególne nukleotydy w kolumnach odpowiadały tym samym miejscom w sekwencji DNA
- Jest to łatwiejsze jeśli są to takie same nukleotydy
- Jeśli się różnią, to oznacza, że w toku ewolucji jeden z nich przeszedł w drugi (substytucja)
- Trudniej jest dopasować sekwencje, jeśli zawierają one insercje i delecje, w takim przypadku nie zawsze jest wiadome, które miejsca odpowiadają sobie sekwencjach.

- Na przykład mamy dwie sekwencje:

AGCCTTAG

AGCTTAG

- Można je dopasować tak:

AGCCTTAG

AGC-TTAG

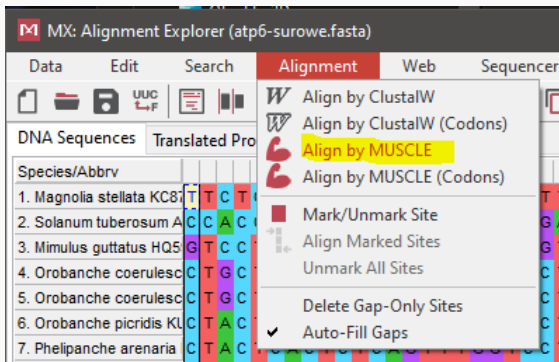
- Znak „-” oznacza brakujący nukleotyd
- Albo tak:

AGCCTTAG

AG-CTTAG

- W naszym przykładzie użyjemy sekwencji bez indeli.

- Z menu na górze wybieramy „Align” a następnie „Align By Muscle”



- Następnie wybieramy „OK”

- Pojawia się okno z opcjami, zostawiamy wartości domyślne i klikamy „OK”:

Option	Setting
<b>GAP PENALTIES</b>	
Gap Open	-400.00
Gap Extend	0.00
<b>MEMORY/ITERATIONS</b>	
Max Memory in MB	2048
Max Iterations	16
<b>ADVANCED OPTIONS</b>	
Cluster Method (Iterations 1,2)	UPGMA
Cluster Method (Other Iterations)	UPGMA
Min Diag Length (Lambda)	24

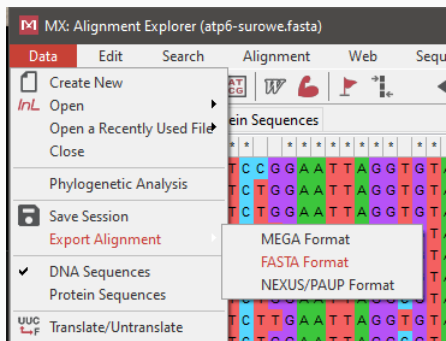
? Help    Reset    X Cancel    ✓ OK





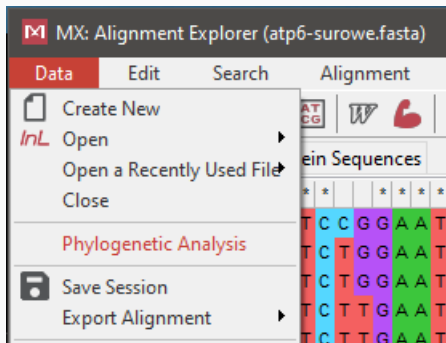
# Zapisanie pliku FASTA

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

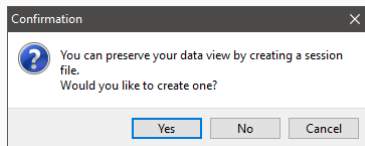
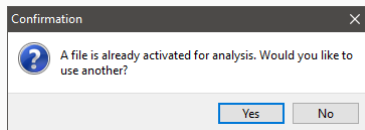
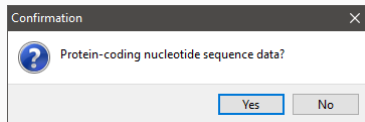




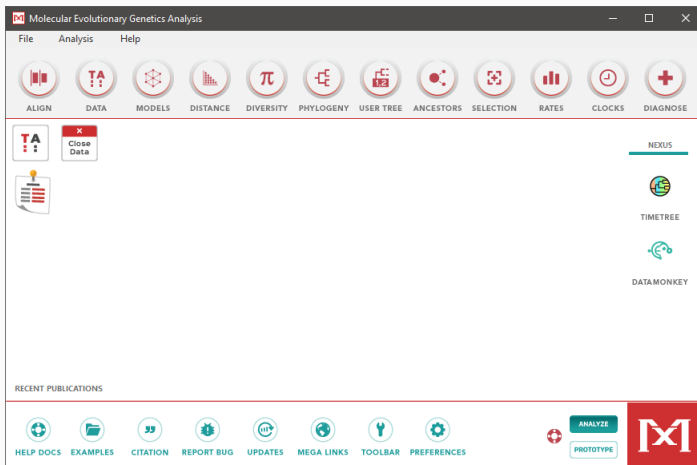
- Teraz wybieramy z menu opcję „Phylogenetic analysis”



- Następnie pojawiają się trzy okienka, z pytaniami, na wszystkie odpowiadamy "Yes"

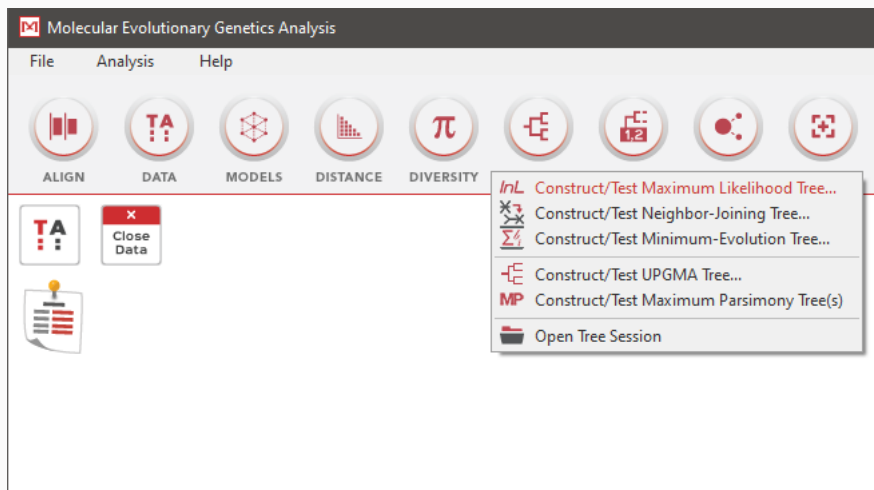


- Zmienia się okno główne, pojawiają się nowe ikony:



# Budujemy drzewko

- Teraz wybieramy z menu Phylogeny opcję Maximum Likelihood



- Zatwierdzamy w okienku pytanie o użycie aktualnych danych.

- Otwiera się okno z opcjami:

MX: Analysis Preferences

Phylogeny Reconstruction

Option	Setting
<b>ANALYSIS</b>	
Statistical Method →	Maximum Likelihood
<b>PHYLOGENY TEST</b>	
Test of Phylogeny →	None
No. of Bootstrap Replications →	Not Applicable
<b>SUBSTITUTION MODEL</b>	
Substitutions Type →	Nucleotide
Genetic Code Table →	Not Applicable
Model/Method →	Tamura-Nei model
<b>RATES AND PATTERNS</b>	
Rates among Sites →	Uniform Rates
No of Discrete Gamma Categories →	Not Applicable
<b>DATA SUBSET TO USE</b>	
Gaps/Missing Data Treatment →	Use all sites
Site Coverage Cutoff (%) →	Not Applicable
Select Codon Positions →	<input checked="" type="checkbox"/> 1st <input checked="" type="checkbox"/> 2nd <input checked="" type="checkbox"/> 3rd <input checked="" type="checkbox"/> Noncoding Sites
<b>TREE INFERENCE OPTIONS</b>	
ML Heuristic Method →	Nearest-Neighbor-Interchange (NNI)
Initial Tree for ML →	Make initial tree automatically (Default - NJ/BioNJ)
Initial Tree File →	Not Applicable
Branch Swap Filter →	None
<b>SYSTEM RESOURCE USAGE</b>	
Number of Threads →	4

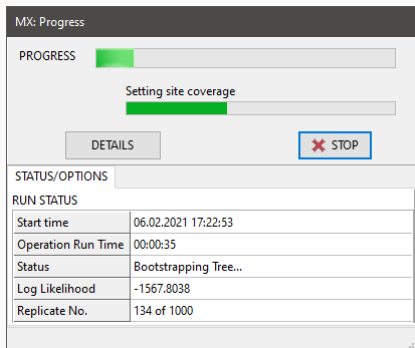
? Help    X Cancel    ✓ OK

- Ustawiamy:

MX: Analysis Preferences	
Phylogeny Reconstruction	
Option	Setting
<b>ANALYSIS</b>	
Statistical Method →	Maximum Likelihood
<b>PHYLOGENY TEST</b>	
Test of Phylogeny →	Bootstrap method
No. of Bootstrap Replications →	1000
<b>SUBSTITUTION MODEL</b>	
Substitutions Type →	Nucleotide
Genetic Code Table →	Not Applicable

- Wartość „Bootstrap” można ustawić na niższą, zwłaszcza gdy masz wolny komputer (np. na 100)

- Pojawia się okienko z postępem procesu tworzenia drzewa

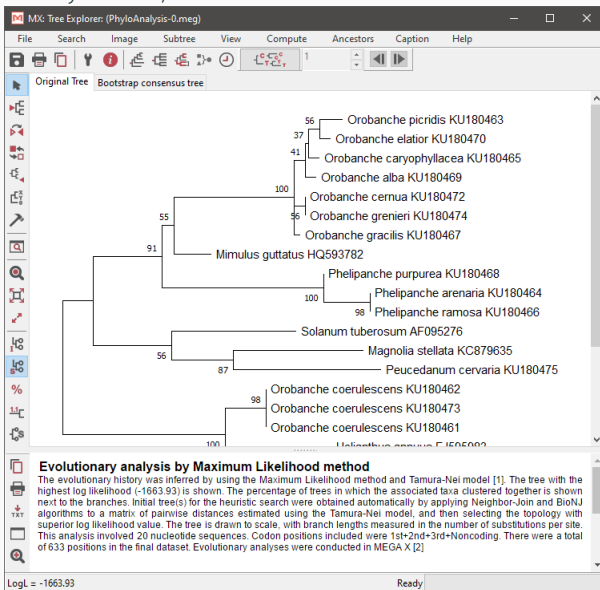


The screenshot shows a dialog box titled "MX: Progress". It features a main progress bar at the top, which is partially filled with green. Below it, a sub-progress bar is labeled "Setting site coverage" and is also partially filled with green. There are two buttons: "DETAILS" and "STOP" (with a red 'X' icon). Below the progress bars is a section titled "STATUS/OPTIONS" containing a table with the following data:

RUN STATUS	
Start time	06.02.2021 17:22:53
Operation Run Time	00:00:35
Status	Bootstrapping Tree...
Log Likelihood	-1567.8038
Replicate No.	134 of 1000

# Nieukorzenione drzewo

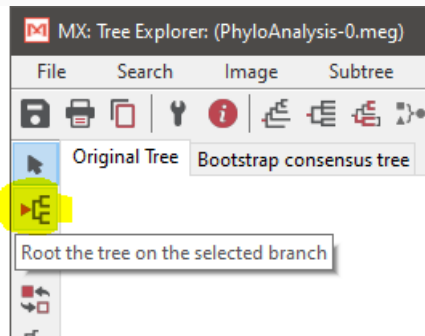
- W końcu widzimy drzewo, na razie nieukorzenione



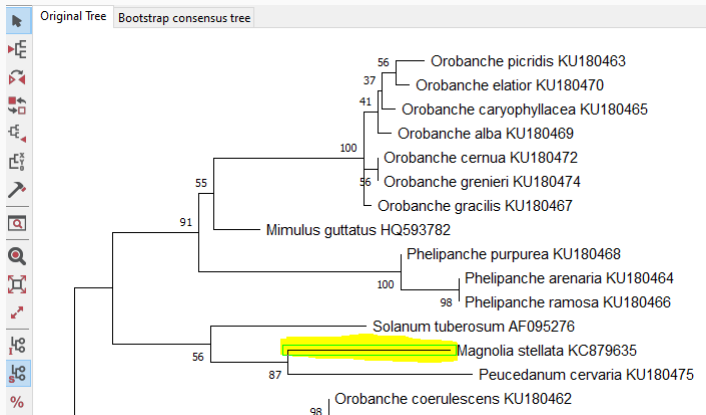


# Ukorzenianie drzewa

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

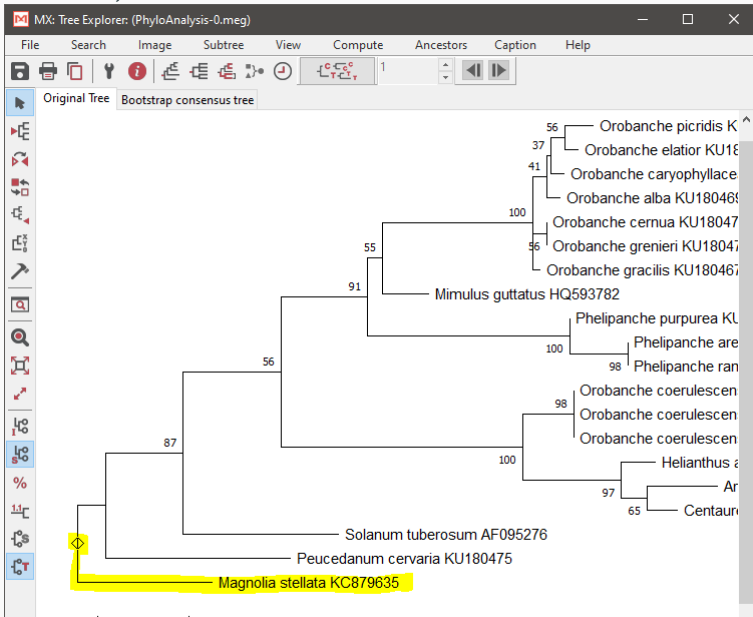


- Następnie klikamy w gałąź prowadzącą do naszej „outgrupy”, którą jest *Magnolia stellata*



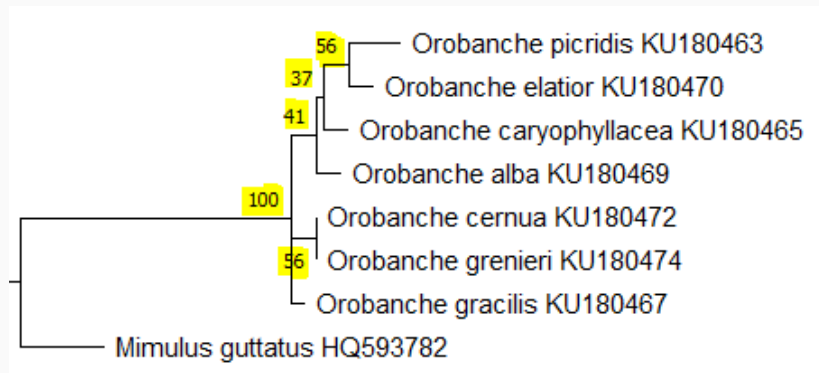
# Drzewo ukorzone

- Teraz drzewo jest ukorzone:



## Wartości bootstrapu

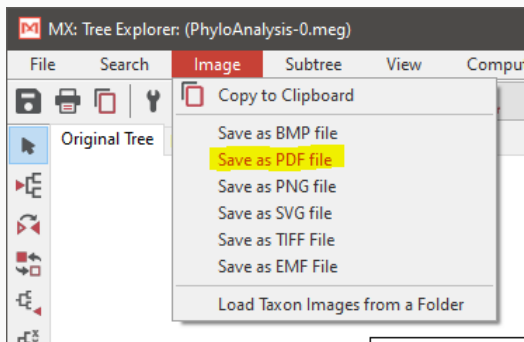
- Na drzewie widoczne są wartości bootstrapu



- Im wyższa wartość bootstrapu tym większa wiarygodność węzła.

## Zapisanie drzewa jako w pliku pdf

- Plik można zapisać np. w formacie pdf:

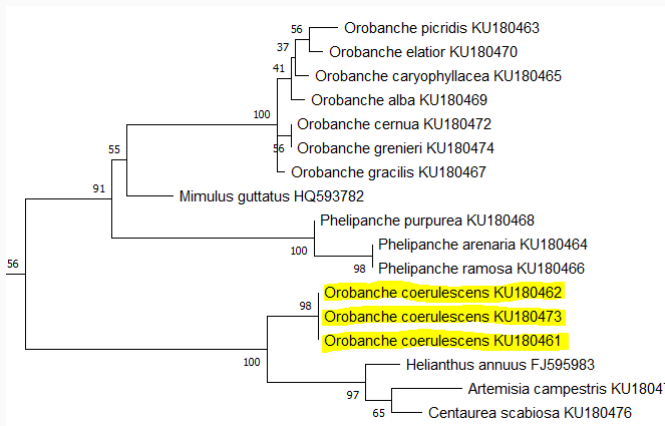


## Horizontalny Transfer Genów (HGT)

---

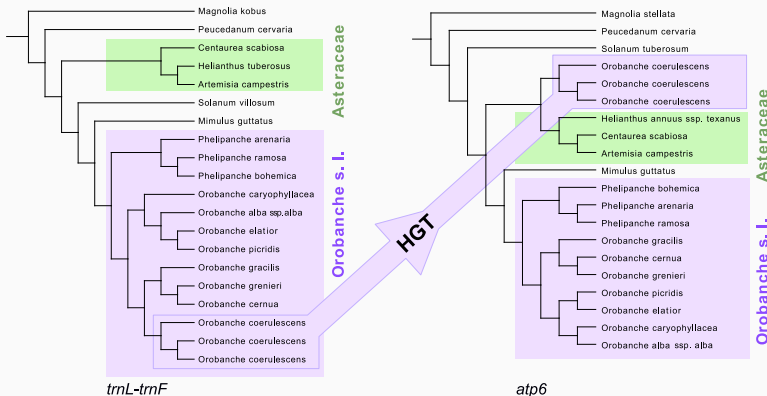
## Horizontalny transfer genów

- Można zauważyć, że *Orobanche coerulescens* „przeskoczyła” z części drzewa *atp6* gdzie znajdują się jej krewniacy do części gdzie znajdują się inne gatunki



# Horizontalny transfer genów

- *Orobanche* i *Phelipanche* to rodzaje roślin pasożytniczych - pobierają one składniki odżywcze od żywicieli
- Gałąź gdzie znalazła się *Orobancha coerulescens* zawiera żywiciela tej rośliny



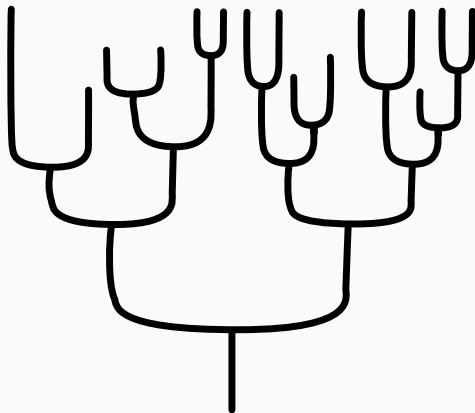


## Horizontalny transfer genów

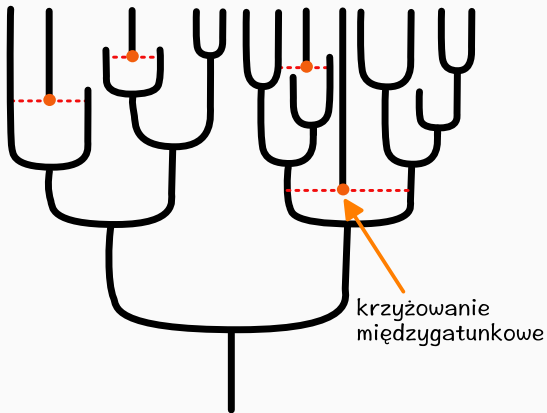
- To co widać na obrazku pokazuje efekt działania procesu zwanego **Horizontalnym Transferem Genów (Horizontal Gene Transfer - HGT)**
- HGT to zjawisko przenoszenia DNA pomiędzy odległymi ewolucyjnie organizmami bez udziału procesów płciowych.
- Jest to zjawisko dość częste u bakterii i pierwotniaków
- 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

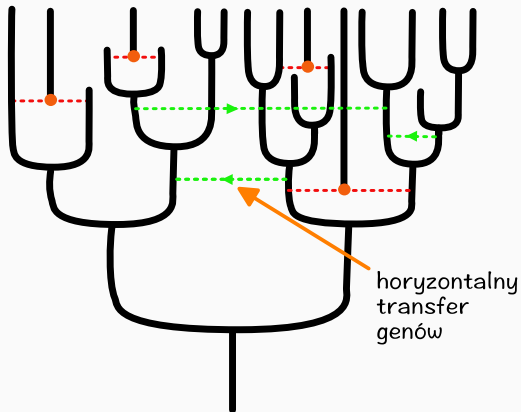




„Tradycyjne” drzewo



Drzewo uwzględniające krzyżówki międzygatunkowe



Drzewo uwzględniające krzyżówki międzygatunkowe i HGT

Dziękuję za uwagę.

Prezentacja, oraz pliki przydatne przy tworzeniu omawianego drzewa są dostępne na mojej stronie internetowej pod adresem:

[ggoralski.pl](http://ggoralski.pl)