# Jak zrobić drzewo filogenetyczne?

Grzegorz Góralski

Zakład Cytologii i Embriologii Roślin
Instytut Botaniki
Wydział Biologii
Uniwersytet Jagielloński

# Spis treści

# Kim jestem?

# 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
www: ggoralski.pl

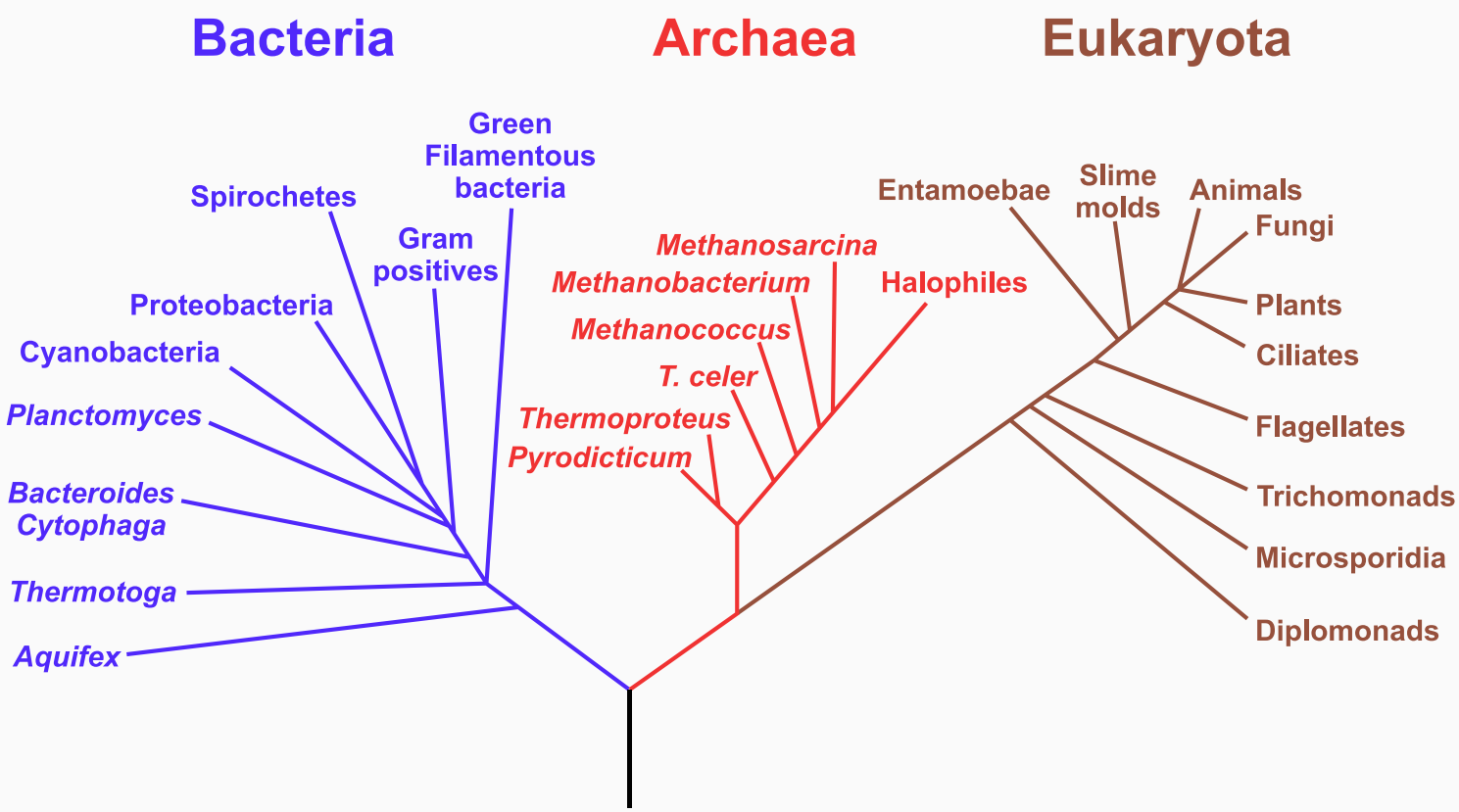
# Czym się zajmuję?

- Liczby chromosomów i poliploidyzacja 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

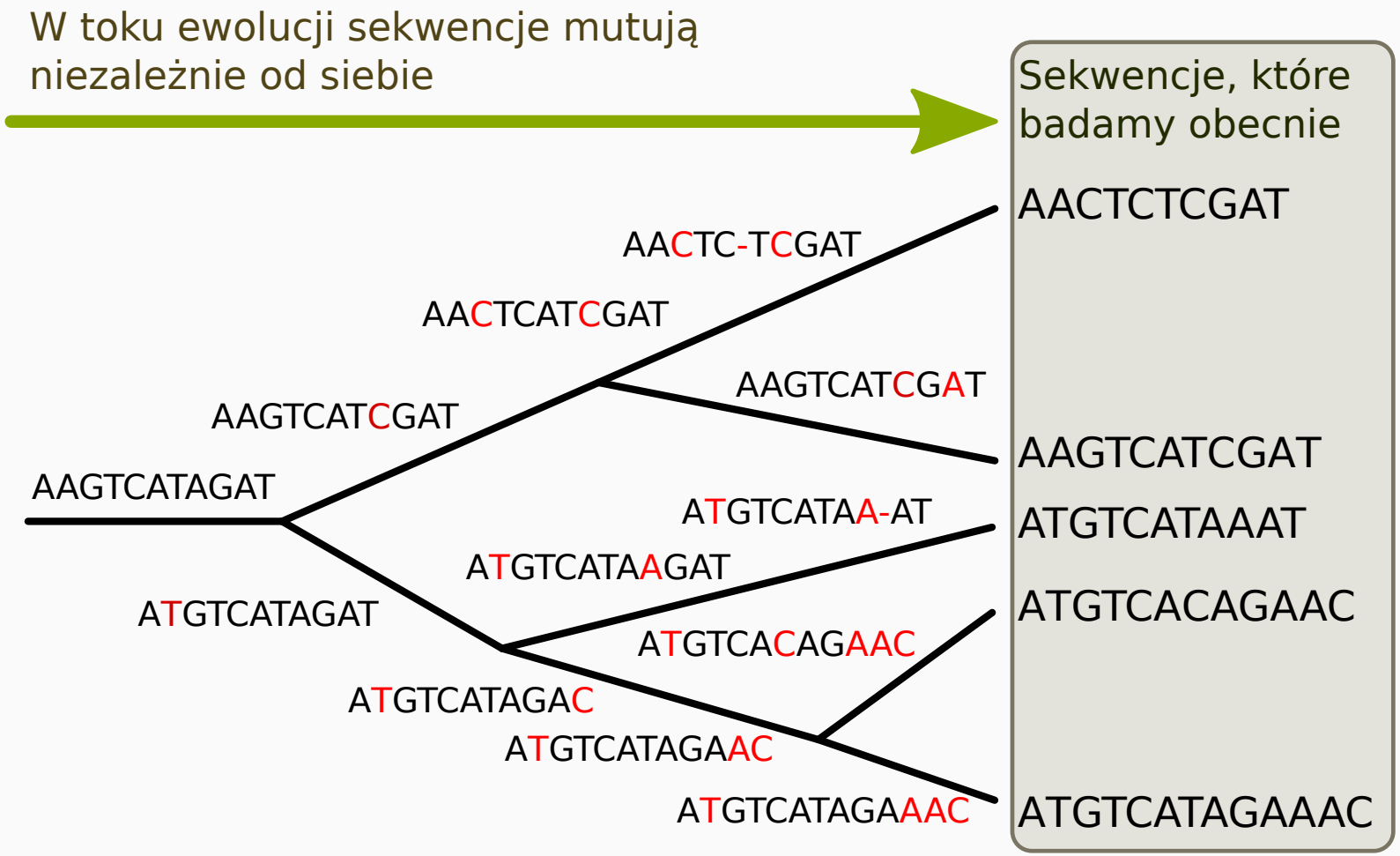



Drzewo filogenetyczne (Wikipedia)

# Drzewa filogenetyczne

- 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



# 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?

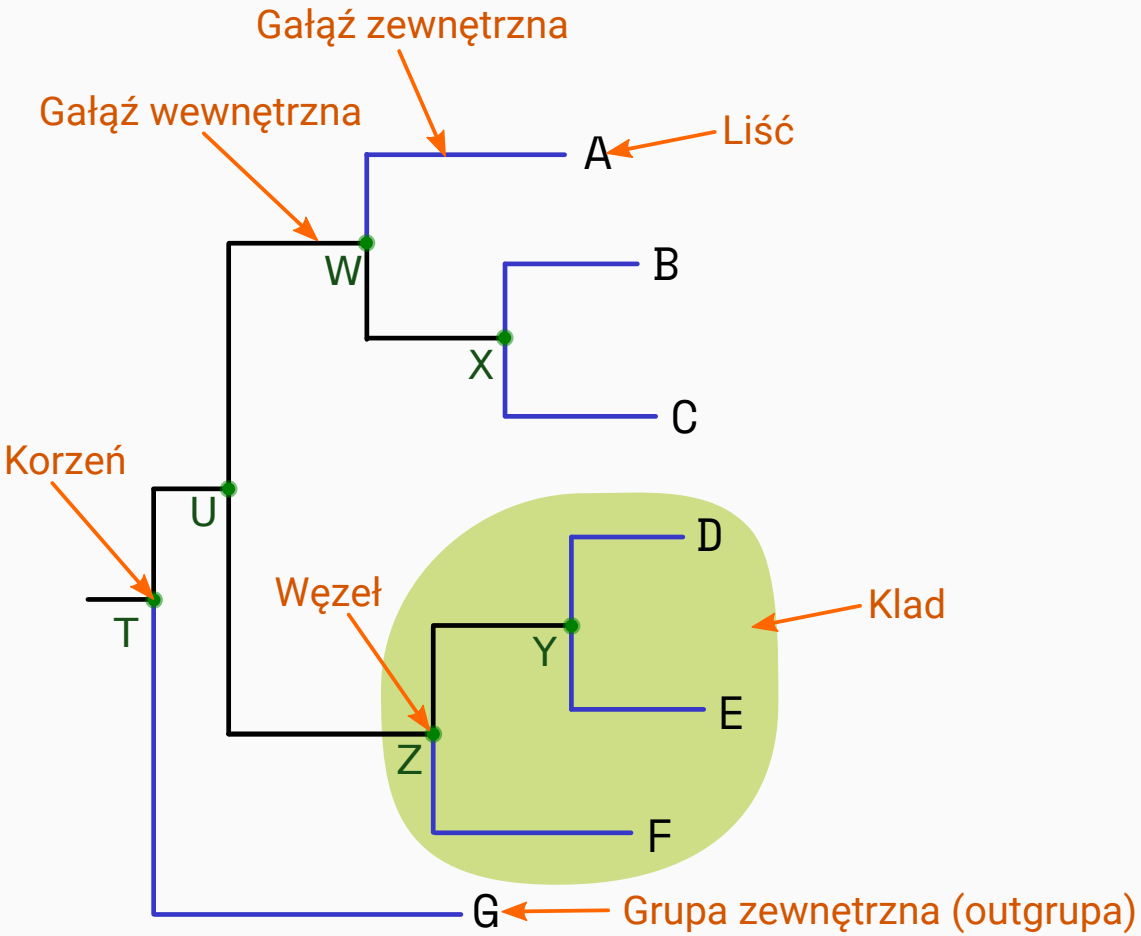
- Geny także różnią się tempem ewolucji
- Niektóre są bardzo „wrażliwe” na zmiany, np. kodujące białka histonowe odpowiedzialne za strukturę chromatyny. Są one bardzo mało zmienne (**konserwatywne**).
- Zatem im bardziej gen jest konserwatywny tym mniej różnic zauważymy między sekwencjami pochodzącymi 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



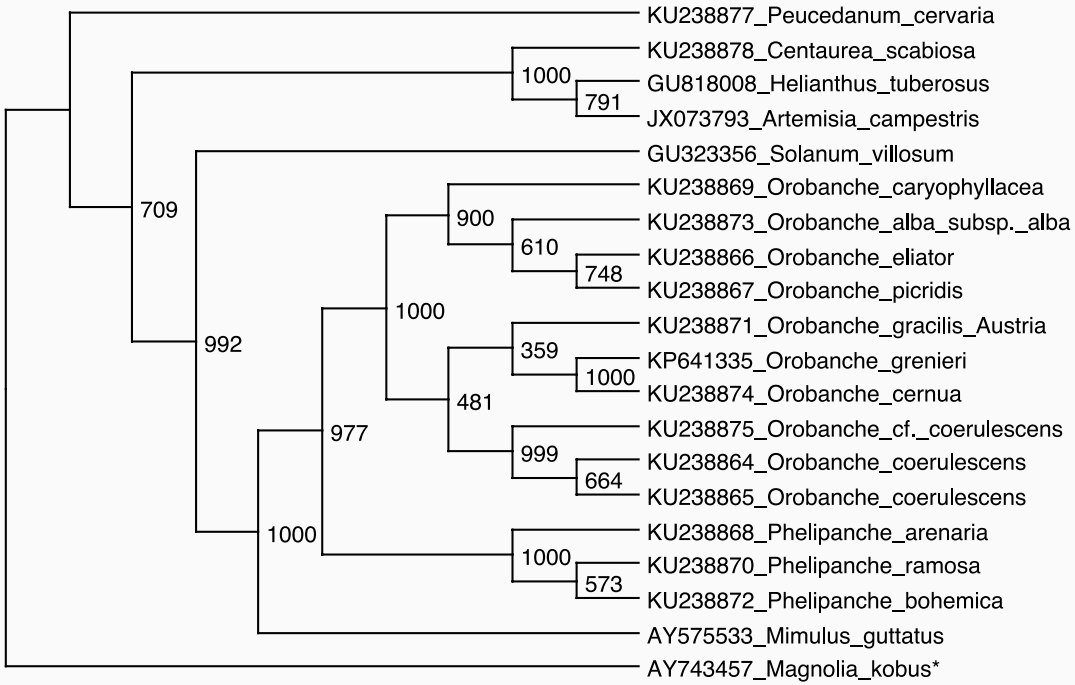

Struktura drzewa

# Elementy 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.

# Podstawowe typy drzew

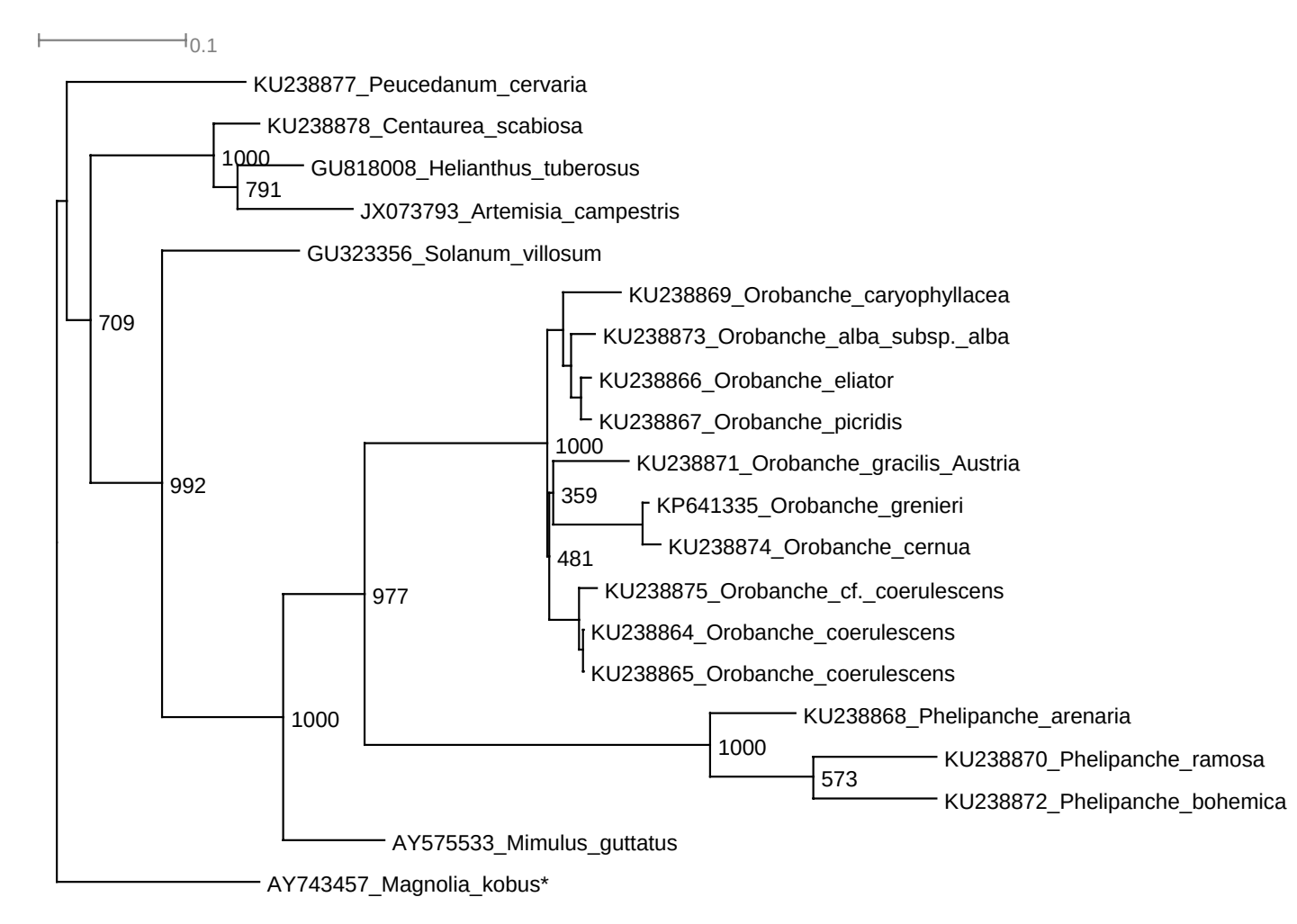
- Kladogram: pokazuje pokrewieństwa ale długość gałęzi nie pokazuje liczby mutacji




Kladogram

# Podstawowe typy drzew

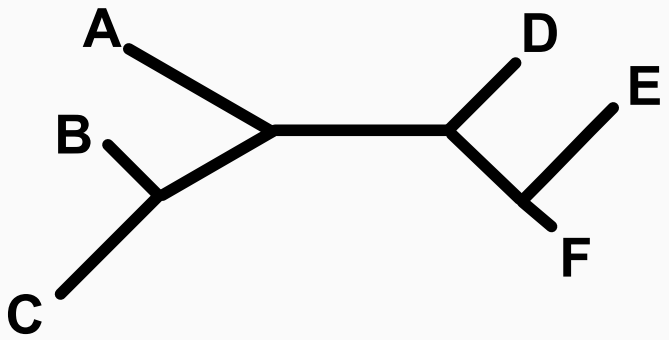
- Filogram: pokazuje pokrewieństwa, długość gałęzi odpowiada liczbie zmian



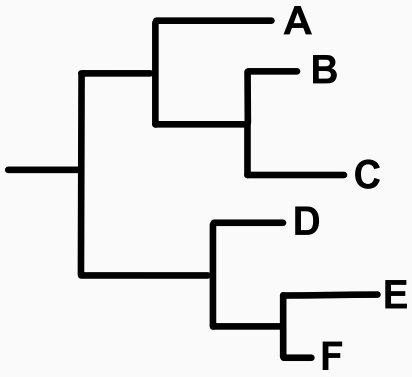

Filogram

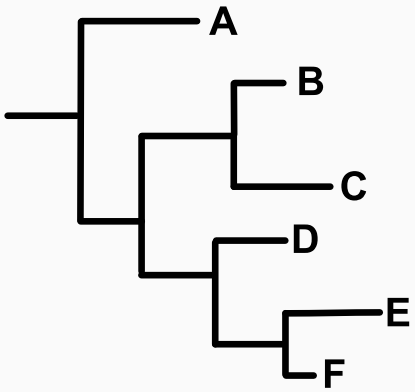
# Nieukorzenione drzewa

- 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:

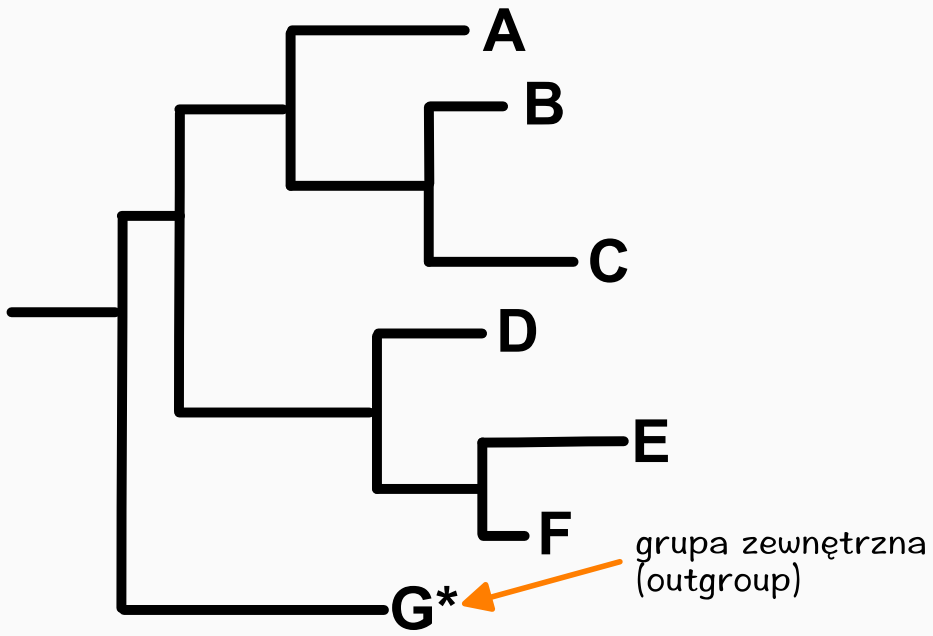



- itd...

# 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.

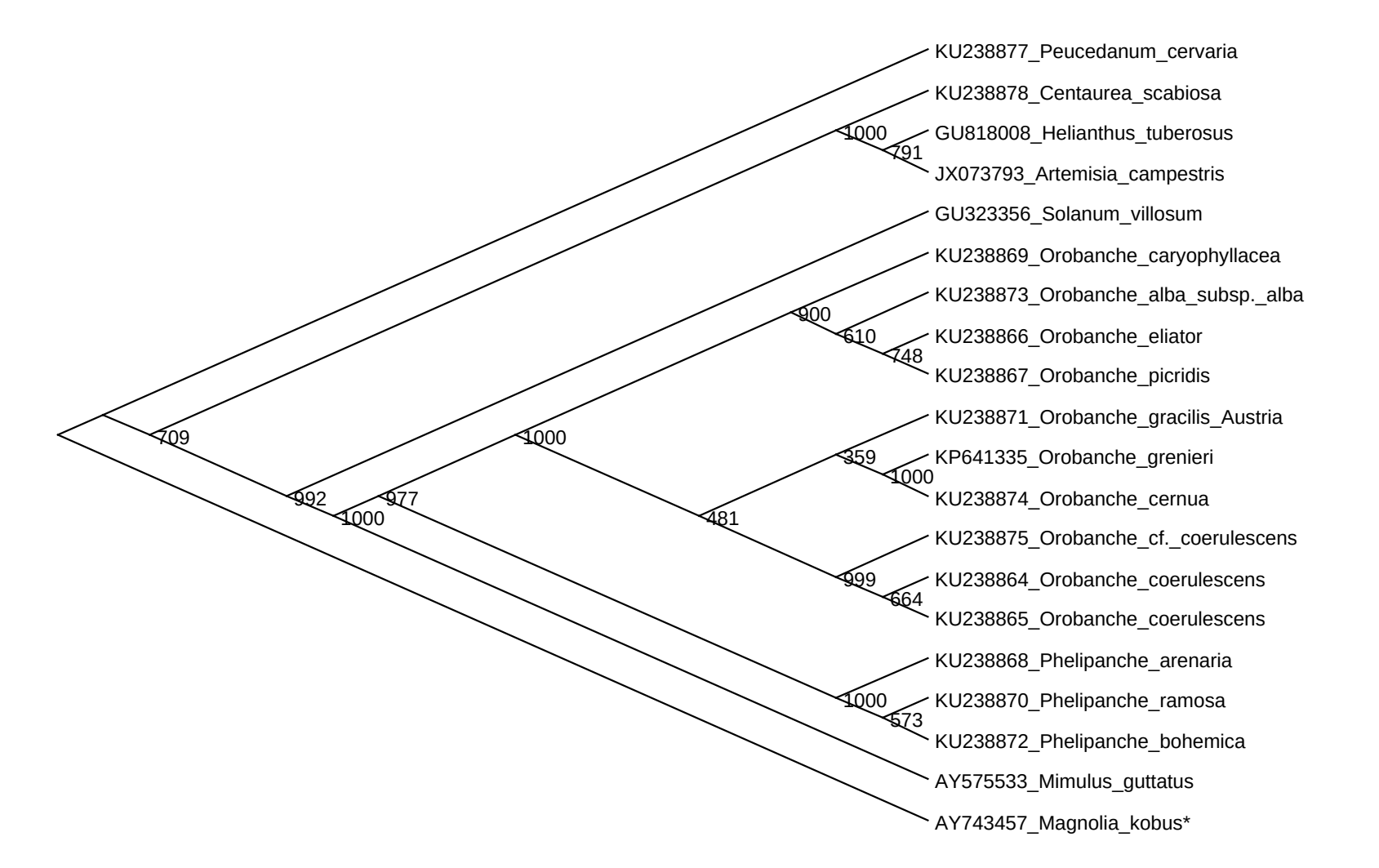


# Typy drzew

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

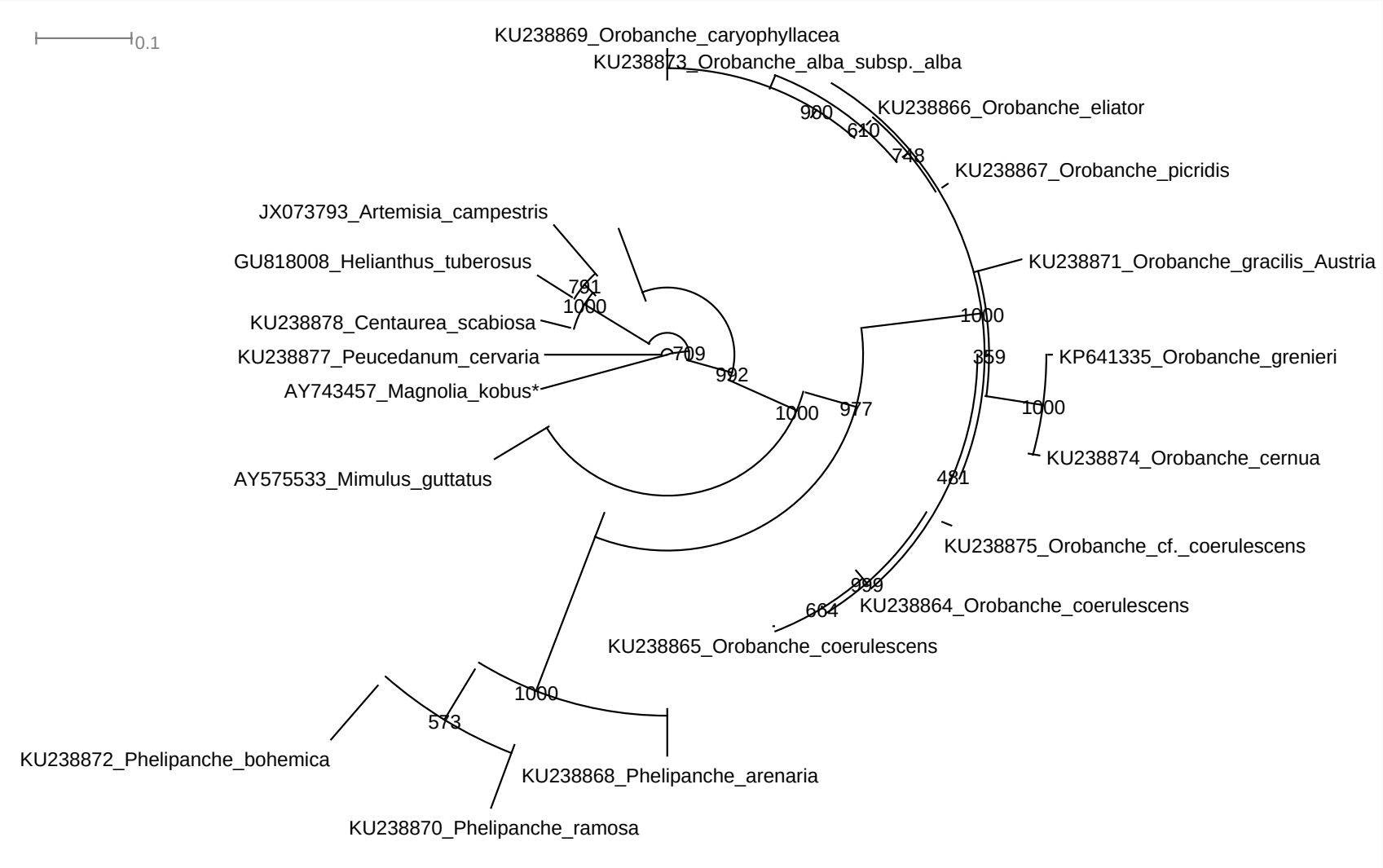
# Typy drzew

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



# Typy drzew

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



# Etapy tworzenia drzew

# 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

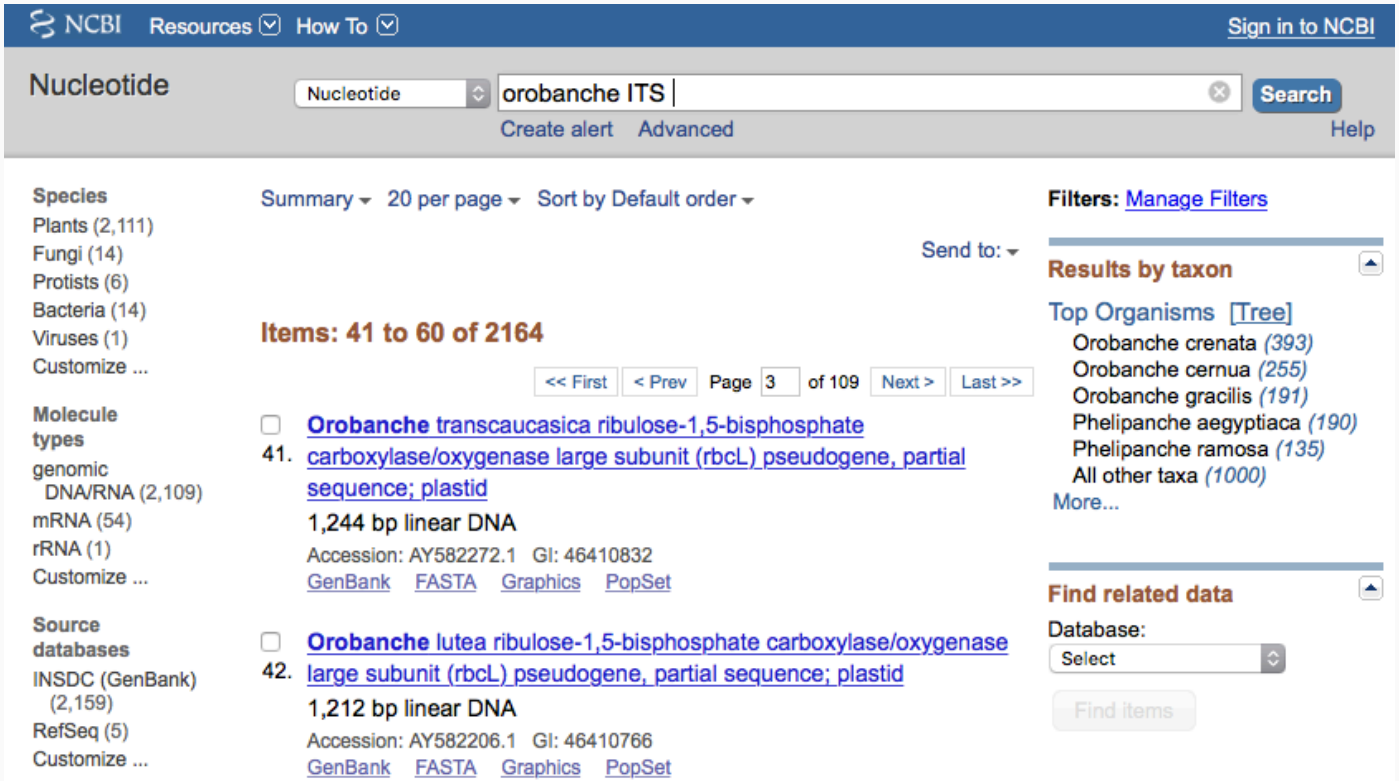
- Wybór sekwencji
  - **Tempo ewolucji**
    - Sekwencję należy dobrać tak aby jej tempo ewolucji pozwalało odróżnić poszczególne taksony (powinny być widoczne różnice pomiędzy bliskimi taksonami)
    - Jednocześnie nie może być zbyt wysokie, ponieważ wtedy trudno dopasować sekwencje a podobieństwa mogą 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)

# Wybrane sekwencje

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

# Zebranie sekwencji

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



# Format FASTA

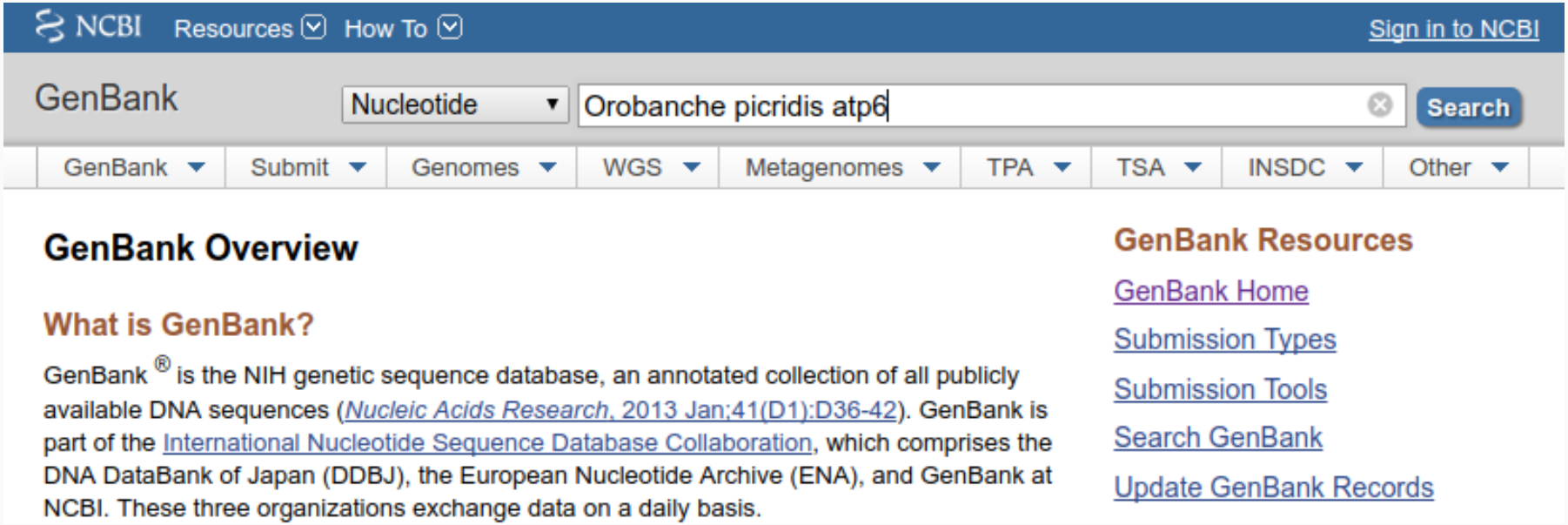
- Sekwencje można zbierać w plikach o róznym formacie.
- Do najbardziej znanych należy format **FASTA**
- Plik FASTA jest zwykłym plikiem tekstowym w którym dane są sformatowane w następujący sposób:

```
>KC879635_Magnolia_stellata ← informacje
CTGCTAACTCTCAGTTTGGTCCTACTTCTGGTTCATTTTGTTACTAAAAACGG ← sekwencja
AACTCAGTACCAAATGCTTGGCAATCCTTGGTAGAGCTTATTCATGATTTCGT
CCGGTAAACGAACAAATAGGTGGTCTTTCCGGAAATGTTCAACAAAAGTTTTC

>AF095276_Solanum_tuberosum
CTACTAACTCTCAGTTTGGTCCTACTTTTGGTTTATTTTGTTACTAAAAAGGG
AACTCAGTACCAAATGCTTGGCAATCCTTGGTAGAGCTTATTTATGATTTCGT
CCGGTAAACGAACAAATAGGTGGTCTTTCCGGAAATGTTAAACAAAAGTTTTC
```

# Wyszukiwanie sekwencji

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




- Można podawać nazwy taksonu, sekwencji itp.

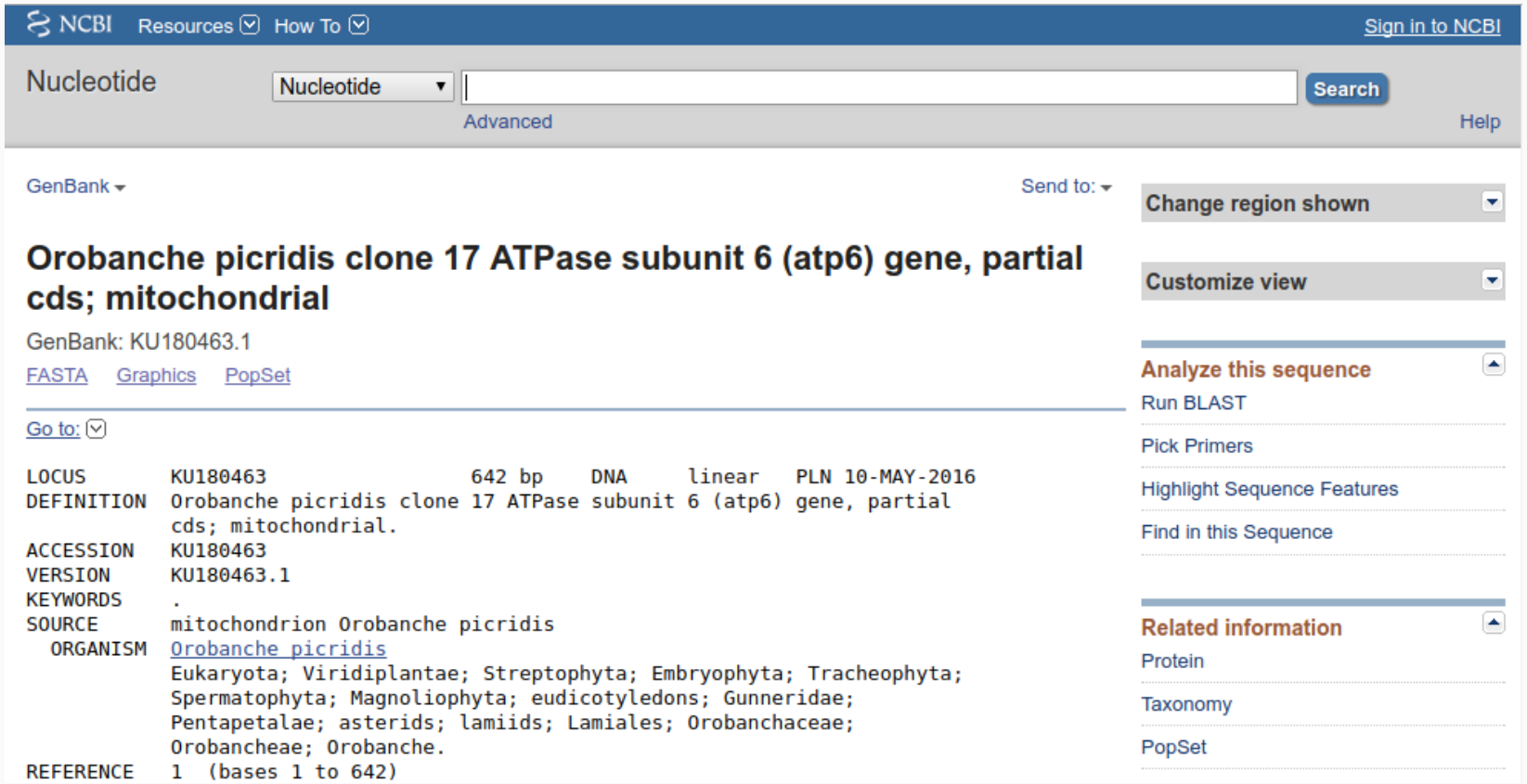
# Wyszkiwanie sekwencji

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


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

# Wynik

- Strona z wynikiem wyszukiwania zawiera szereg informacji.



# Wynik

- Na dole znajduje się sama sekwencja:

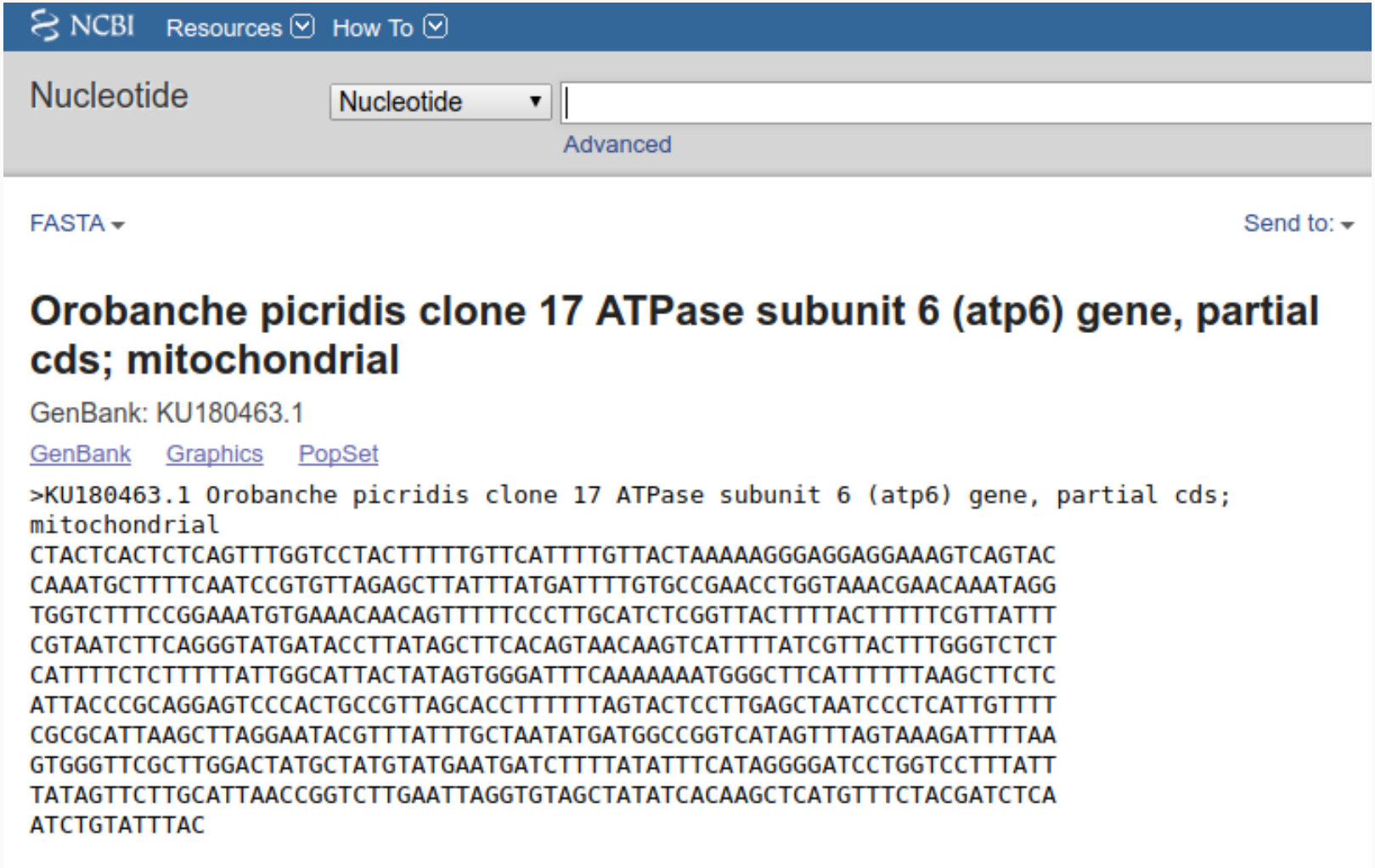
```
                     /codon_start=1
                     /product="ATPase subunit 6"
                     /protein_id="ANC68065.1"
                     /translation="LLTLSLVLLFVHFVTKKGGGKSVPNAFQSVLELIYDFVPNLVNE
                     QIGGLSGNVKQQFFPCISVTFTFSLFRNLQGMIPYSFTVTSHFIVTLGLSFSLFIGIT
                     IVGFQKNGLHFLSFSLPAGVPLPLAPFLVLLELIPHCFRALSLGIRLFANMMAGHSLV
                     KILSGFAWTMLCMNDLLYFIGDPGPLFIVLALTGLELGVAISQAHVSTISICIY"
ORIGIN
        1 ctactcactc tcagtttggt cctacttttt gttcattttg ttactaaaaa gggaggagga
       61 aagtcagtac caaatgcttt tcaatccgtg ttagagctta tttatgattt tgtgccgaac
      121 ctggtaaacg aacaaatagg tggtctttcc ggaaatgtga aacaacagtt tttcccttgc
      181 atctcggtta cttttacttt ttcgttattt cgtaatcttc agggtatgat accttatagc
      241 ttcacagtaa caagtcattt tatcgttact ttgggtctct cattttctct ttttattggc
      301 attactatag tgggatttca aaaaaatggg cttcattttt taagcttctc attacccgca
      361 ggagtcccac tgccgttagc acctttttta gtactccttg agctaatccc tcattgtttt
      421 cgcgcattaa gcttaggaat acgtttattt gctaatatga tggccggtca tagtttagta
      481 aagattttaa gtgggttcgc ttggactatg ctatgtatga atgatctttt atatttcata
      541 ggggatcctg gtcctttatt tatagttctt gcattaaccg gtcttgaatt aggtgtagct
      601 atatcacaag ctcatgtttc tacgatctca atctgtattt ac
//
```

# Wyświetlenie w formacie FASTA

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

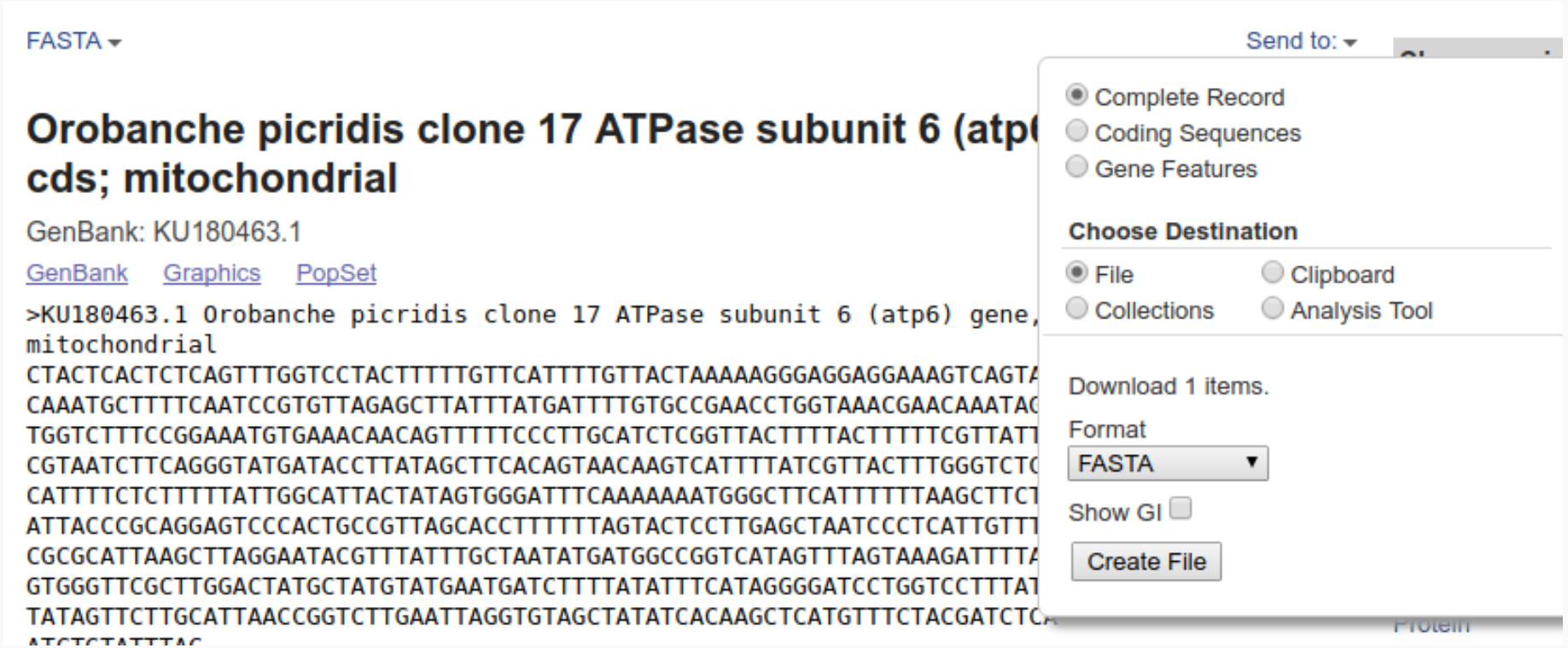
# Wyświetlanie w formacie FASTA

- Wyświetla się wynik w formacie FASTA



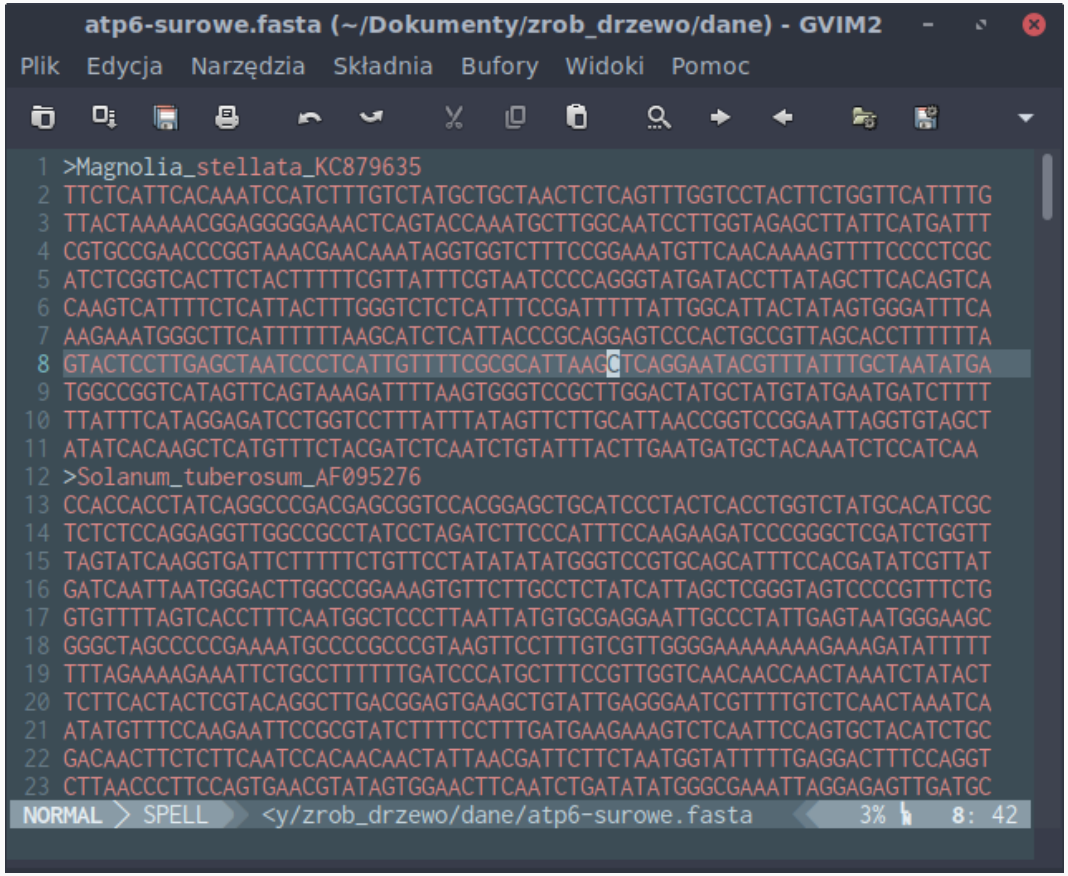
# Zapis formatu FASTA

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



# 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:




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

# Wybór oprogramowania

# 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 (*Neghbor-Joining*), największej wiarygodności (*Maximum Likehood*), 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*

# MEGA

- Generalnie uważa się, że do pracy bioinformatycznej lepiej nadają się systemy UNIX-owe (Linux, Mac OS X itp.), na nie tez jest dostępniych 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

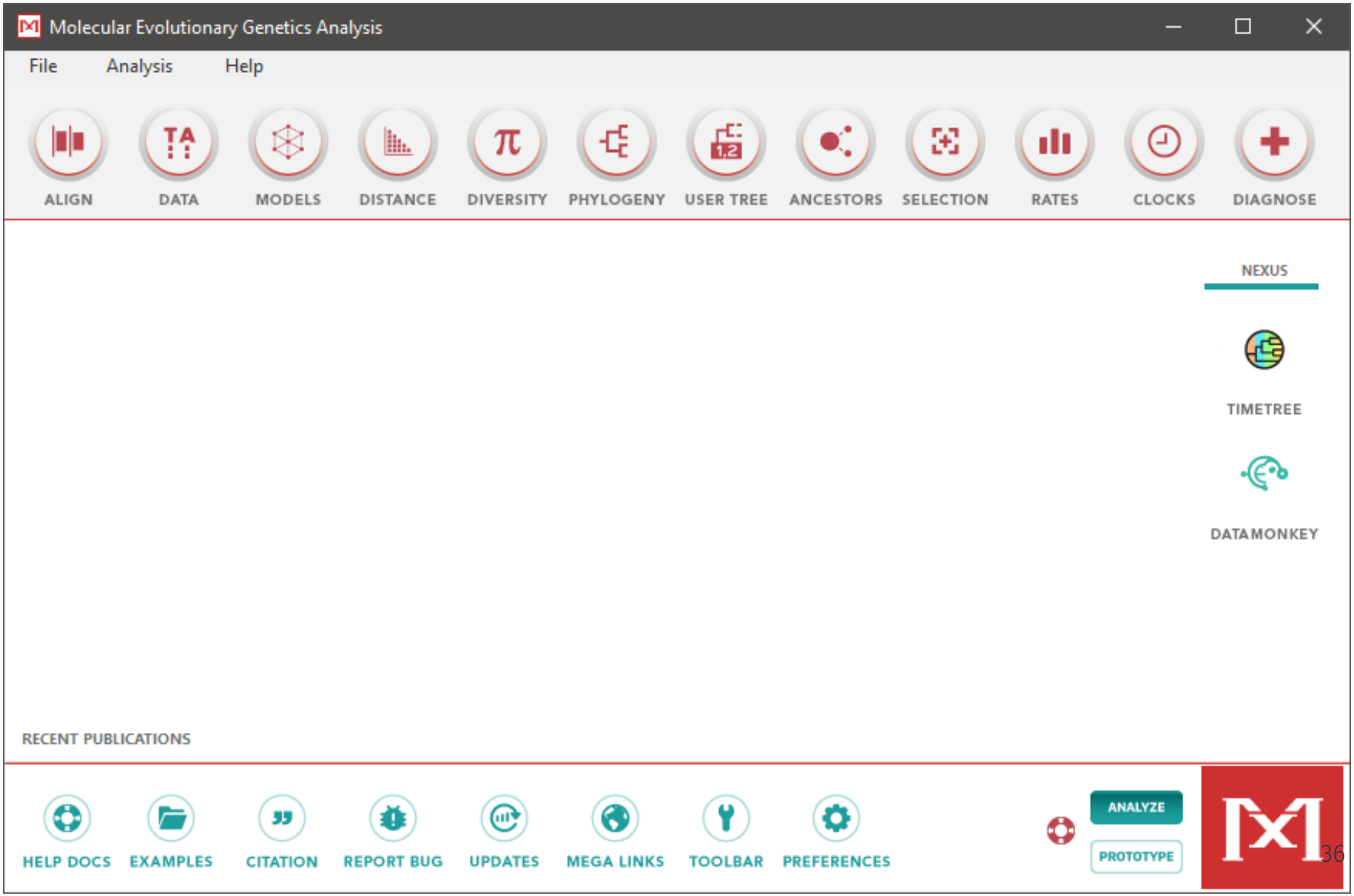
# Tworzenie drzewka w programie MEGA

# Sekwecje

- Pobierz plik „dane.zip”, który znajdziesz pod adresem: `http://ggoralski.pl/?page_id=3276`
- Rozpakuj plik - powstanie katalog „dane”
- Pobierz i zainstaluj program Mega ze strony: `https://www.megasoftware.net/`

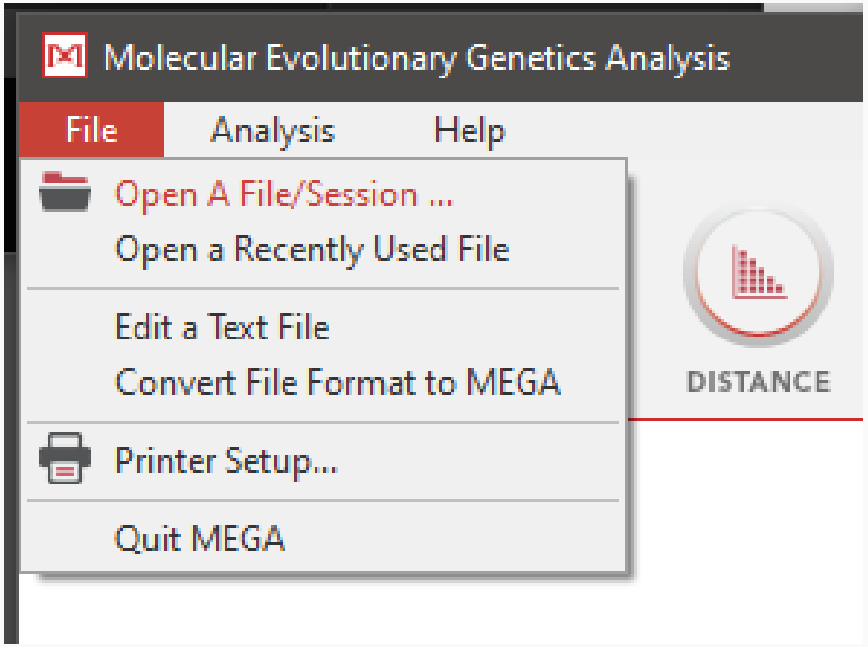
# MEGA - interfejs

- Po otwarciu programu pokazuje się taki interfejs:

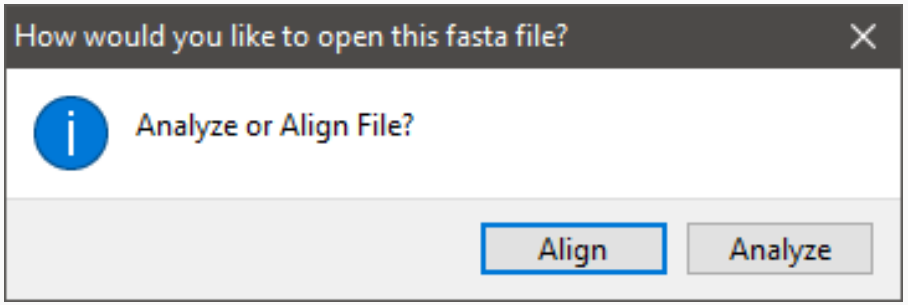


# Otwieramy plik do wyrównania

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

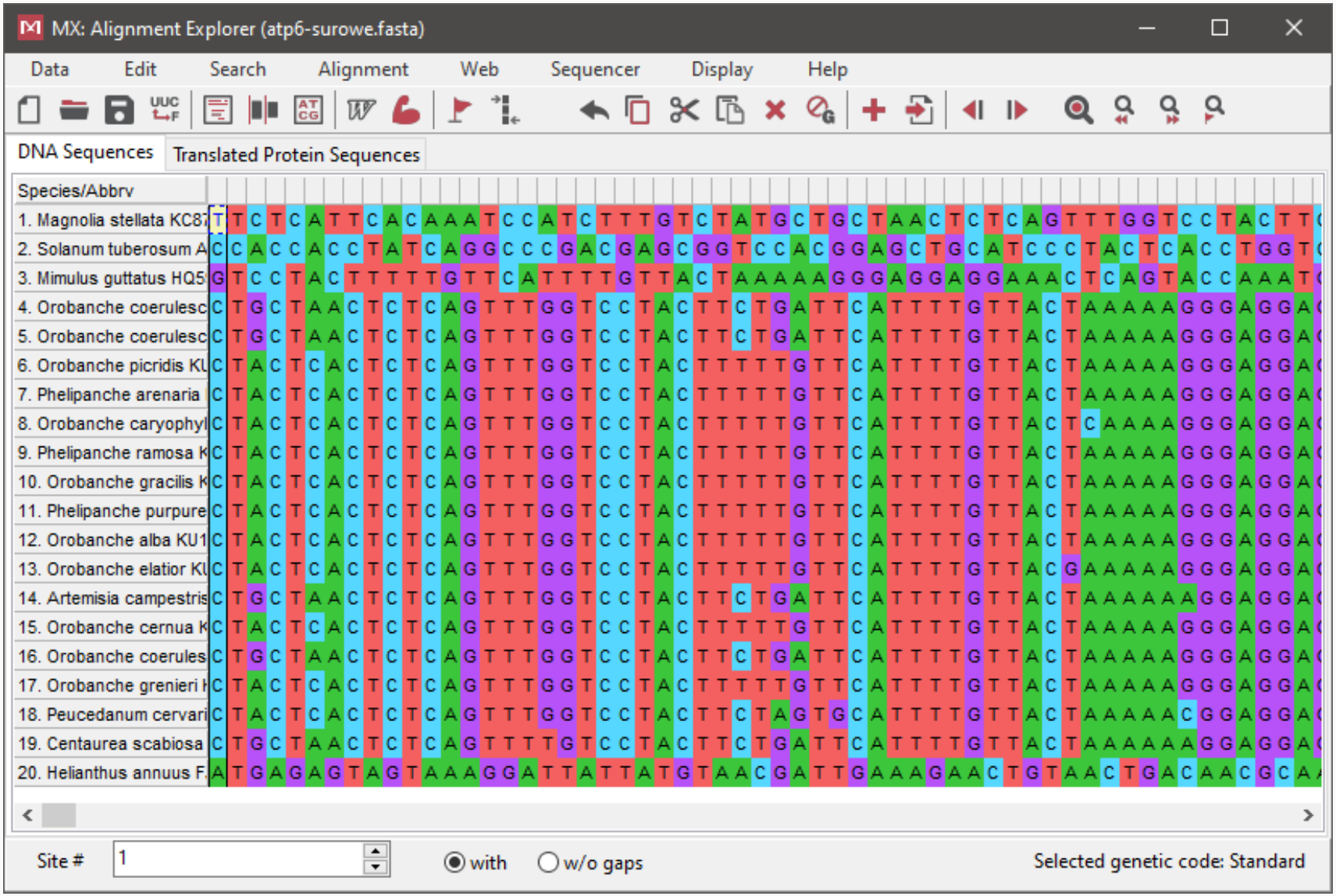



- Pokaże się takie okienko, w którym wybieramy „Align”



# Dopasowanie sekwencji

- Pokazuje się taki widok:



# Dopasowanie sekwencji

- 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
```

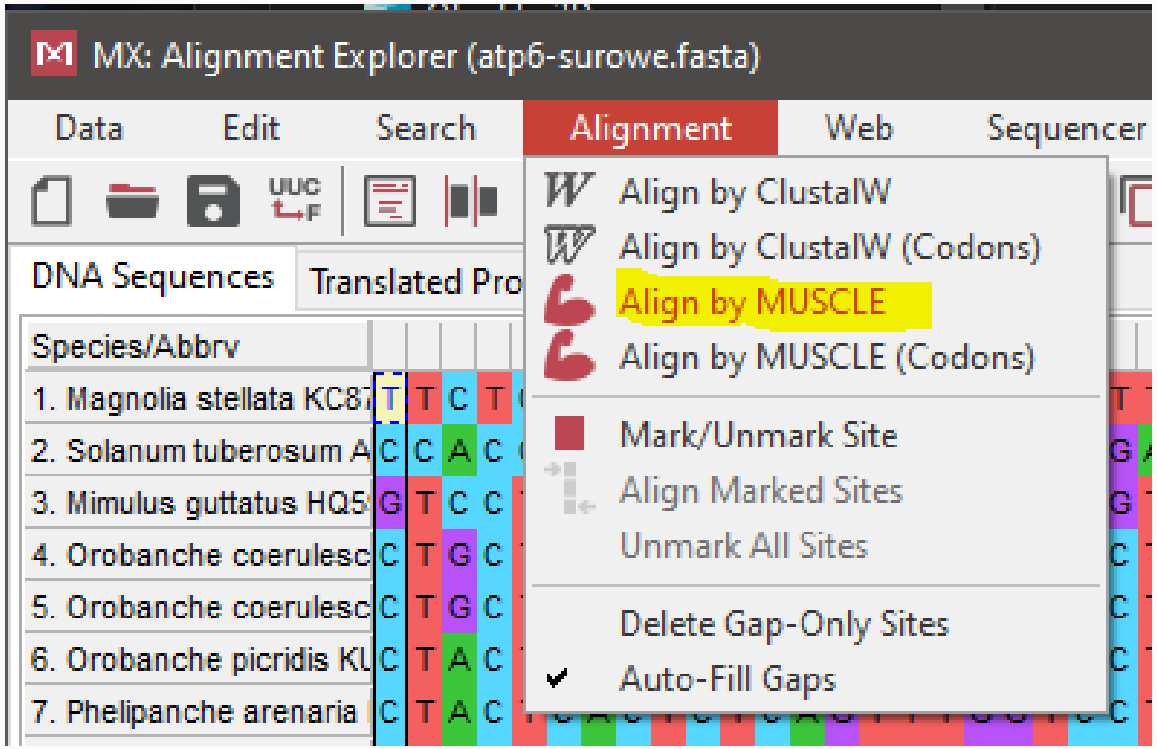
- Znak „-” oznacza brakujący nukleotyd
- Albo tak:

```
AGCCTTAG
AG-CTTAG
```

- W naszym przykładzie użyjemy sekwencji bez indeli.

# Dopasowanie c.d.

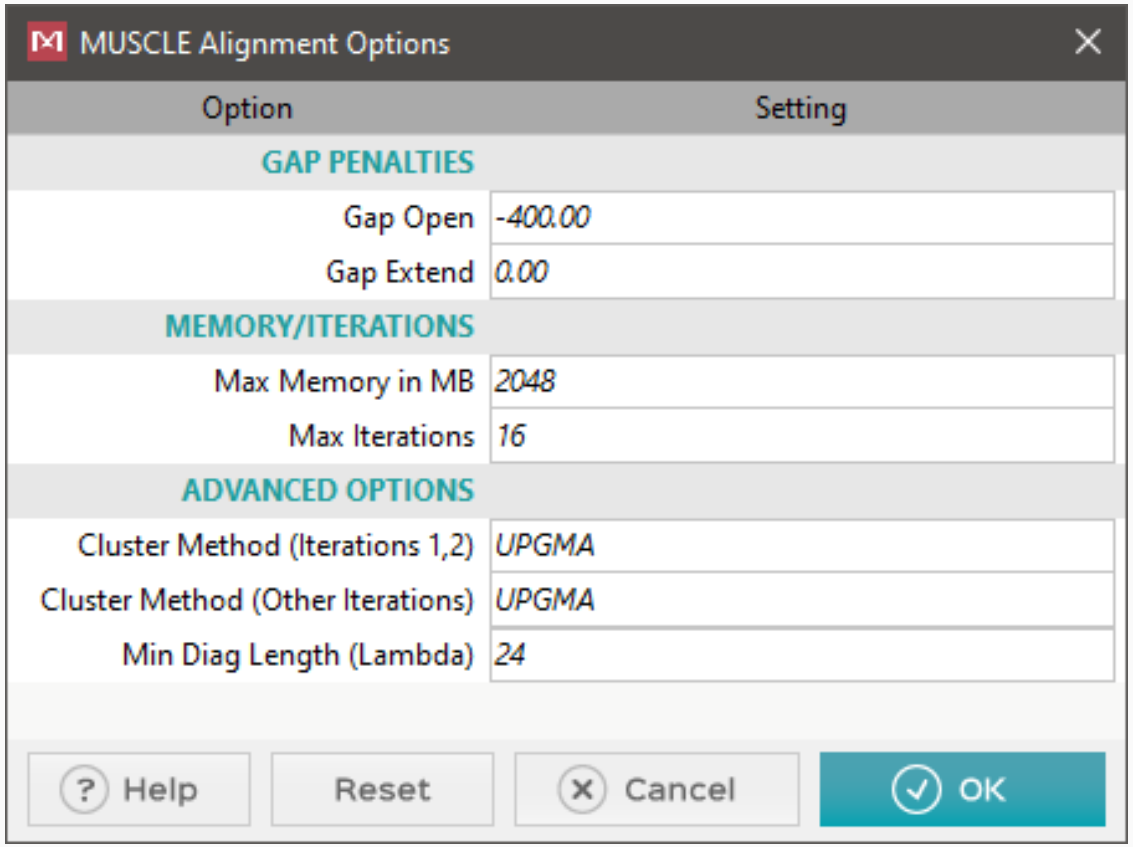
- Z menu na górze wybieramy „Align” a następnie „Align By Muscle”




- Następnie wybieramy „OK”

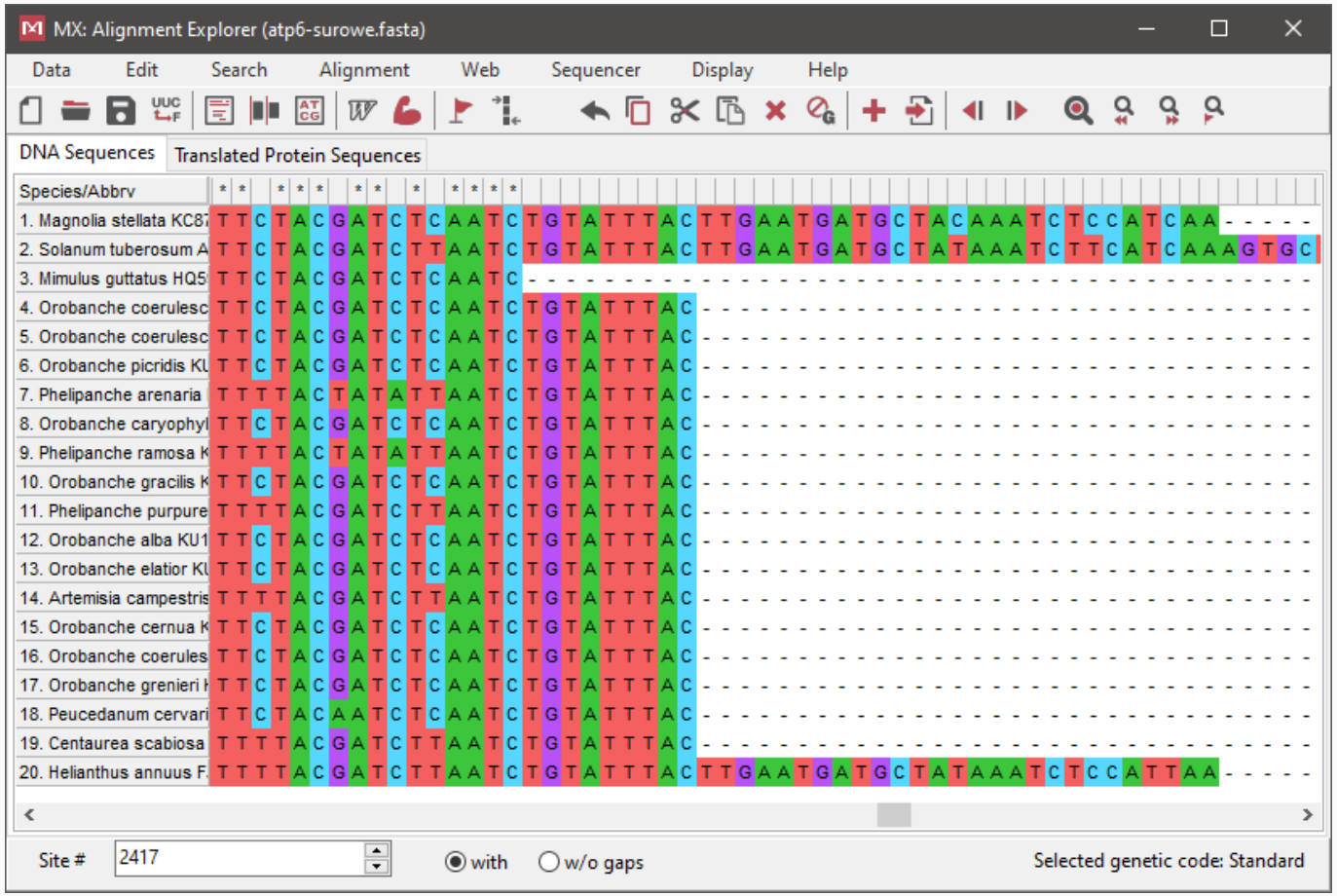
# Dopasowanie c.d.

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



# Dopasowanie c.d.

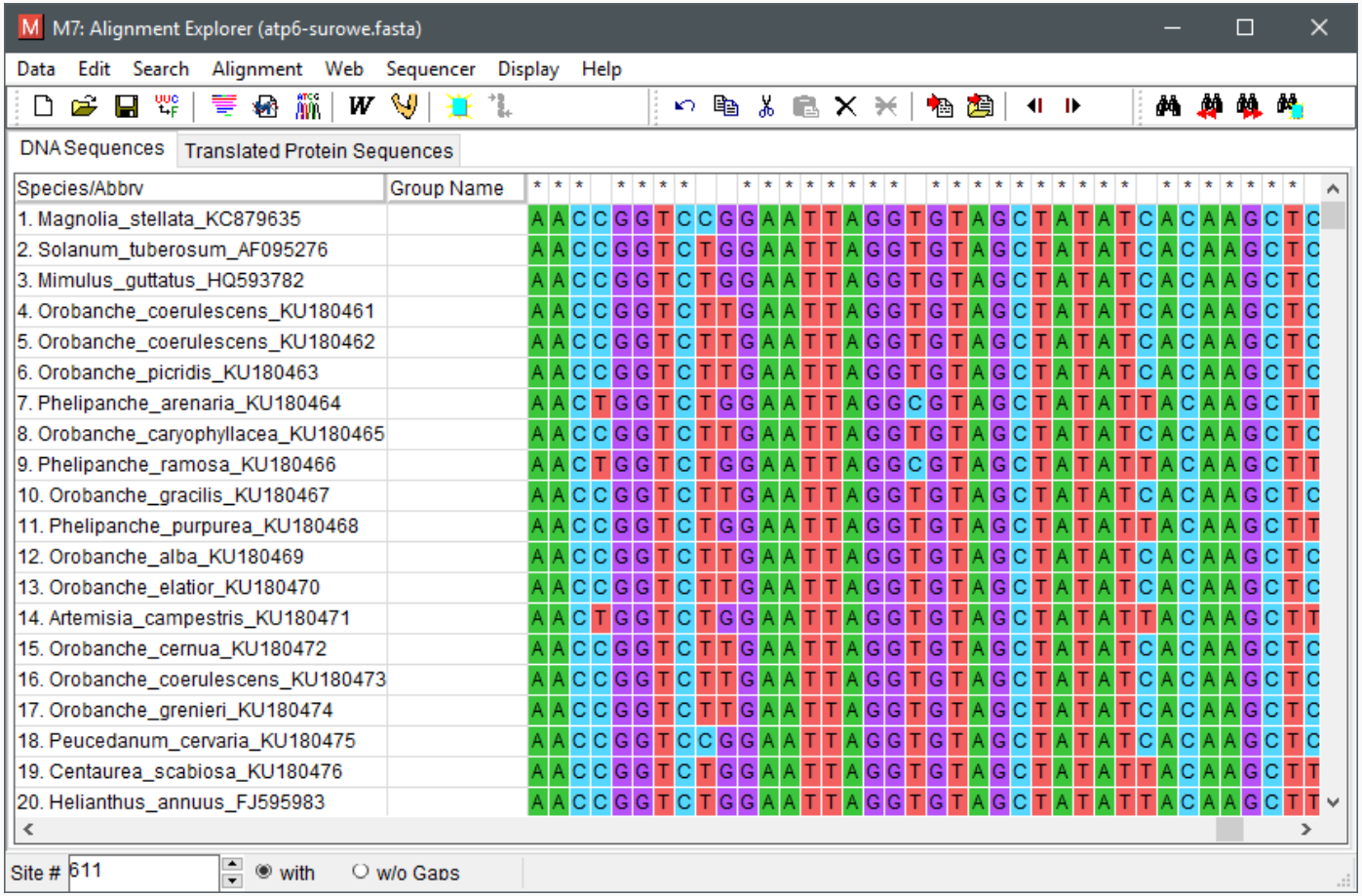
- Pojawia się pasek postępu
- Otrzymujemy wyrównanie sekwencje, ale z obu końców niektóre sekwencje są wyraźnie dłuższe:




- Zaznaczamy „ wystające” części i je usuwamy (klawisz „Delete”).

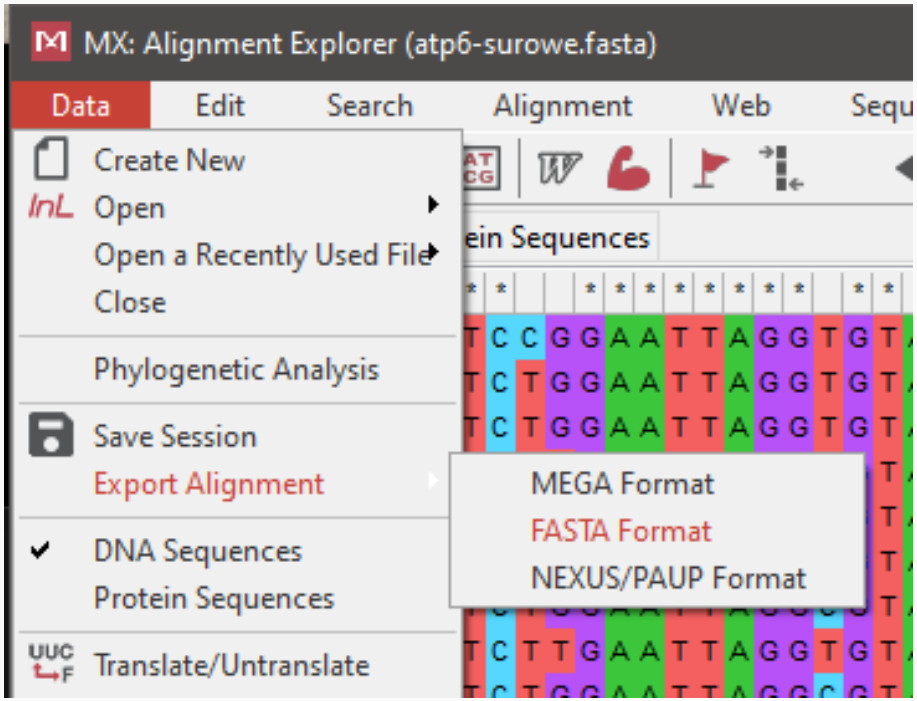
# Sekwencje dopasowane

- W końcu sekwencje zostały dopasowane.



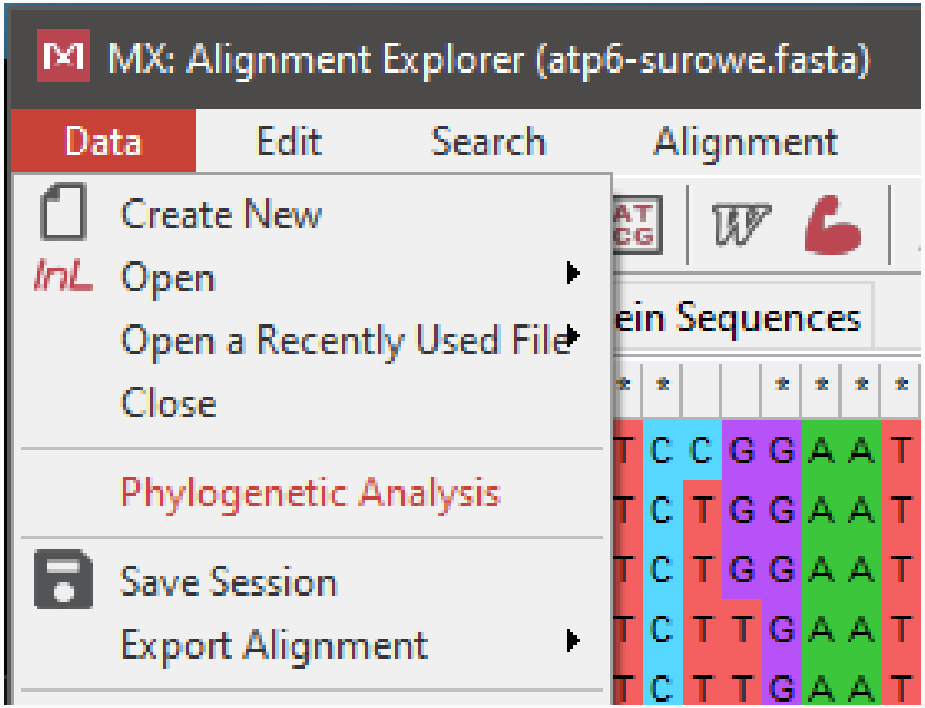
# Zapisanie pliku FASTA

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



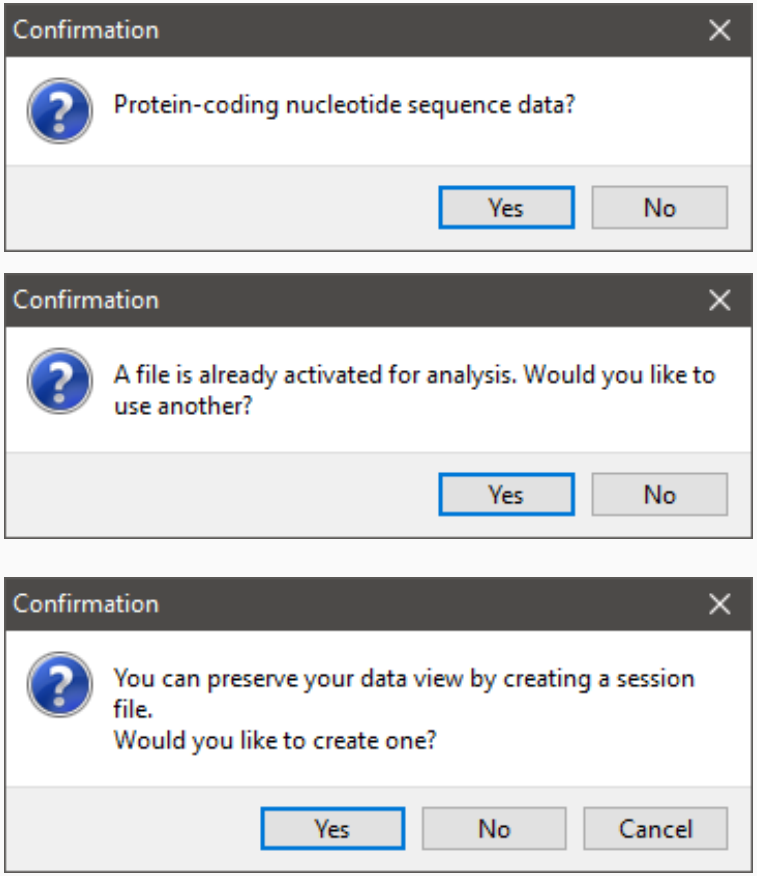
# Tworzenie drzewa

- Teraz wybieramy z menu opcję „Phylogenetic analysis”



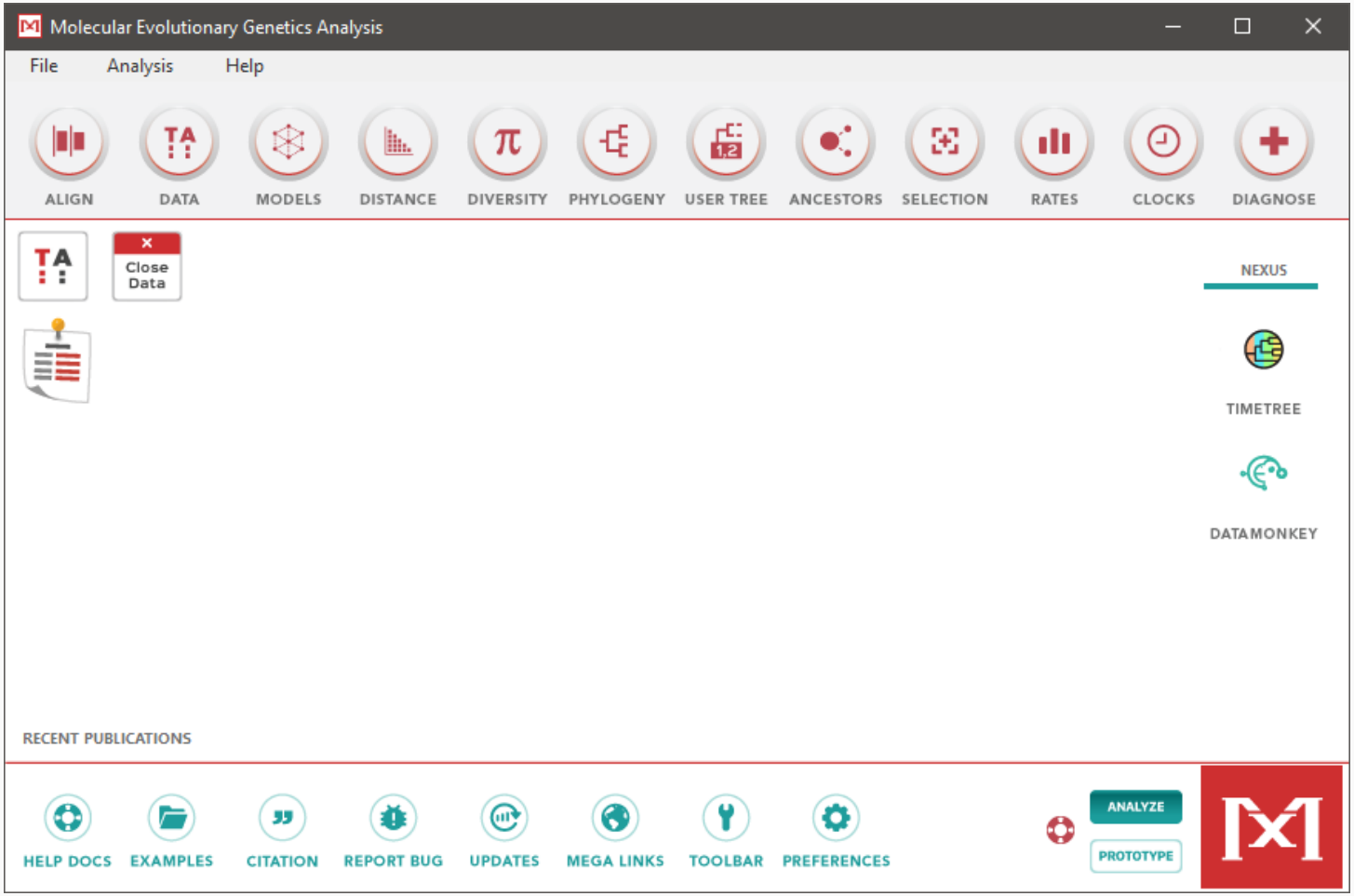
# Tworzenie drzewa

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



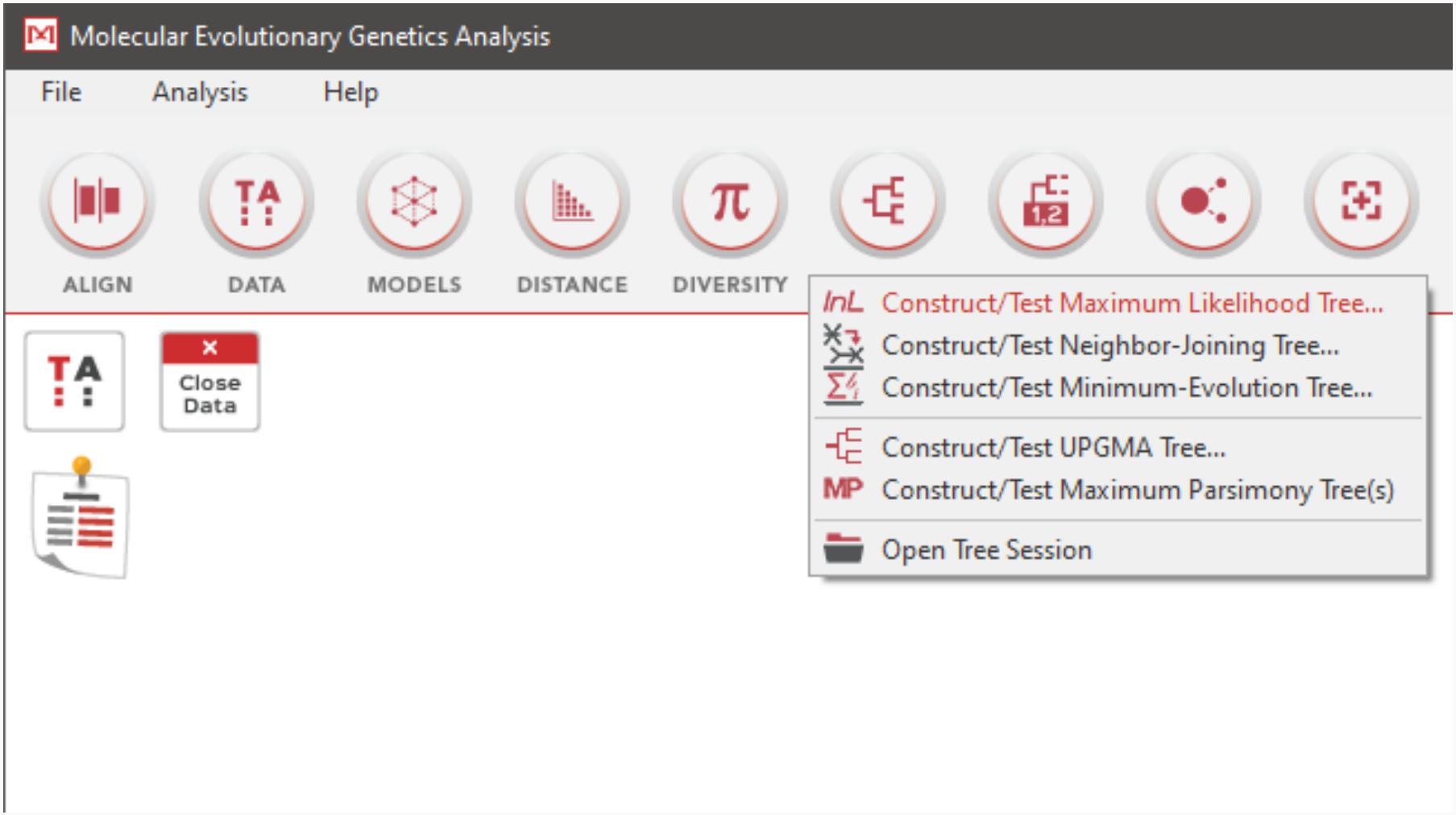
# Tworzenie drzewa

- Zmienia się okno główne, pojawiają się nowe ikony:



# Budujemy drzewko

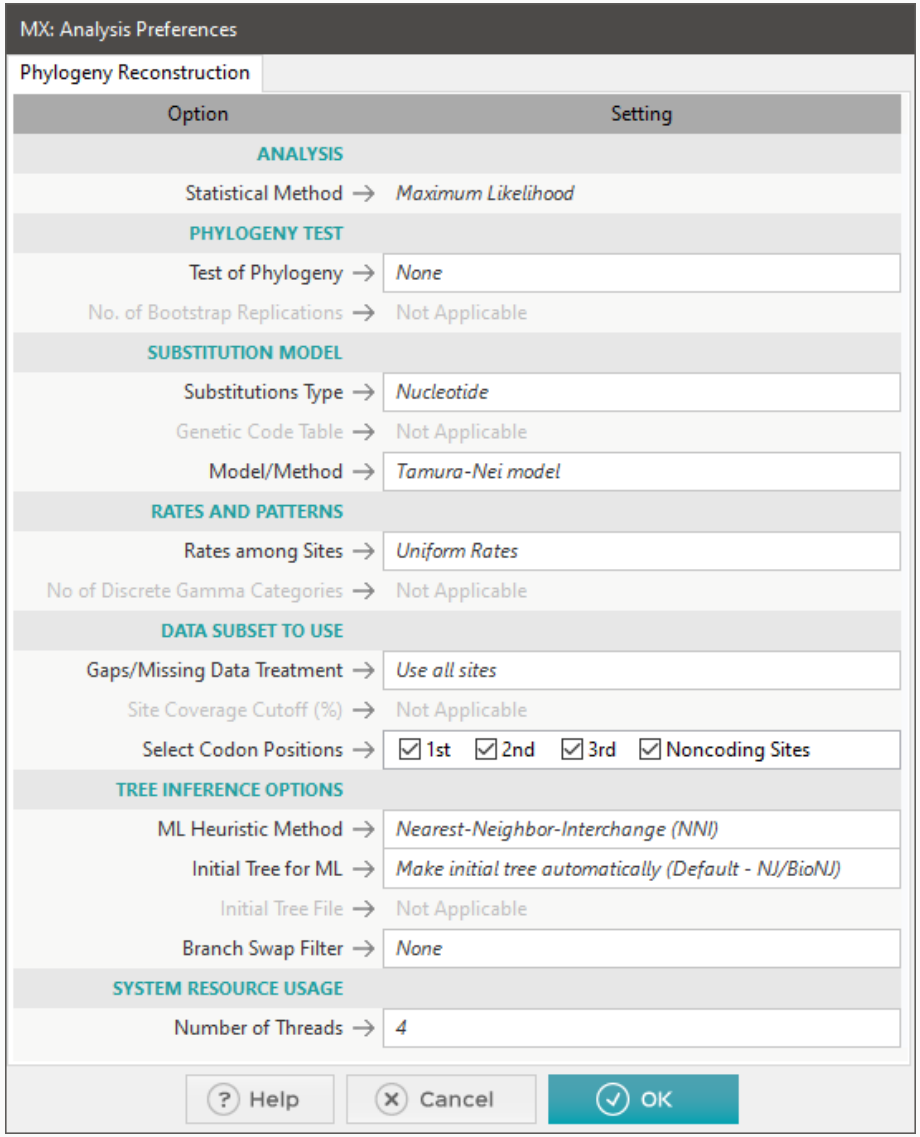
- Teraz wybieramy z menu `Phylogeny` opcję `Maximum Likelihood`




- Zatwierdzamy w okienku pytanie o użycie aktualnych danych.

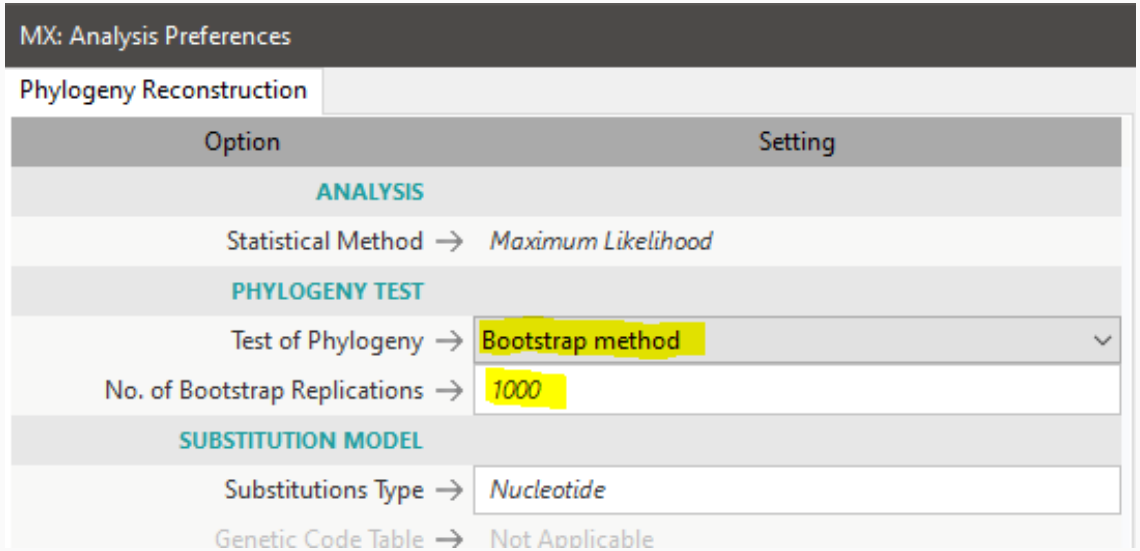
# Opcje

- Otwiera się okno z opcjami:



# Opcje

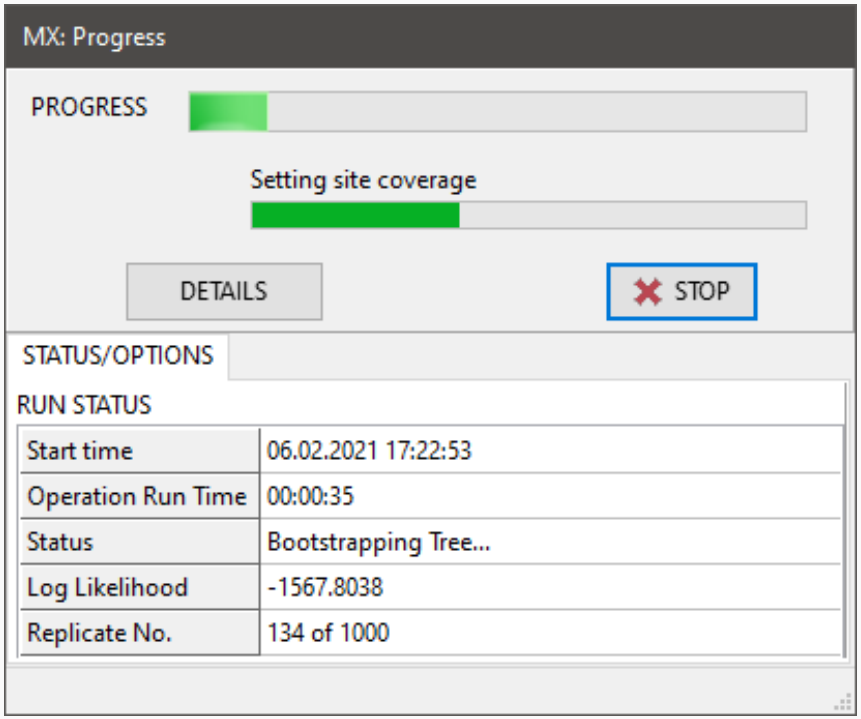
- Ustawiamy:




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

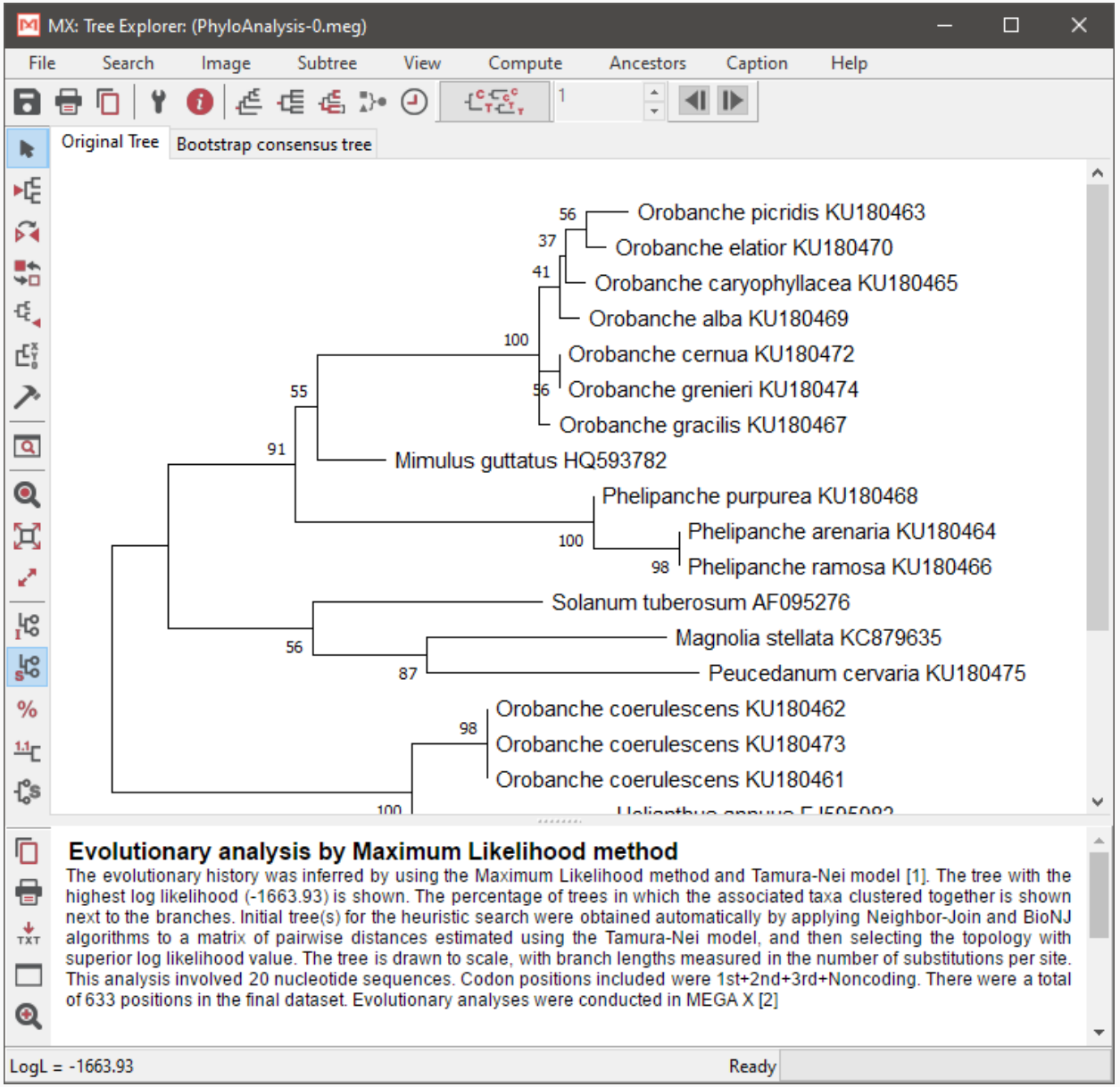
# Drzewko w trakcie tworzenia

- Pojawia się okienko z postępem procesu tworzenia drzewa



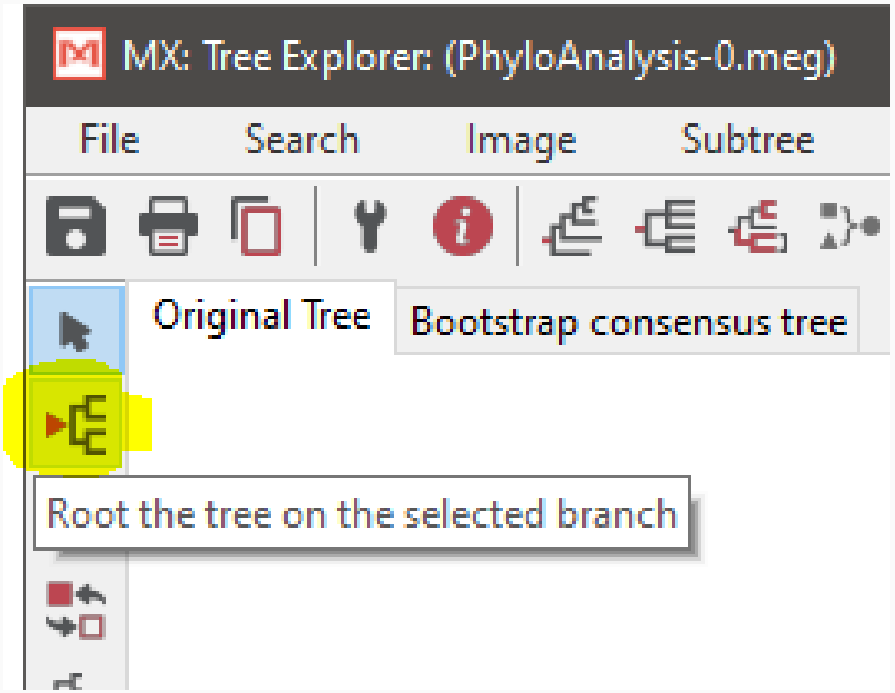
# Nieukorzenione drzewo

- W końcu widzimy drzewo, na razie nieukorzenione



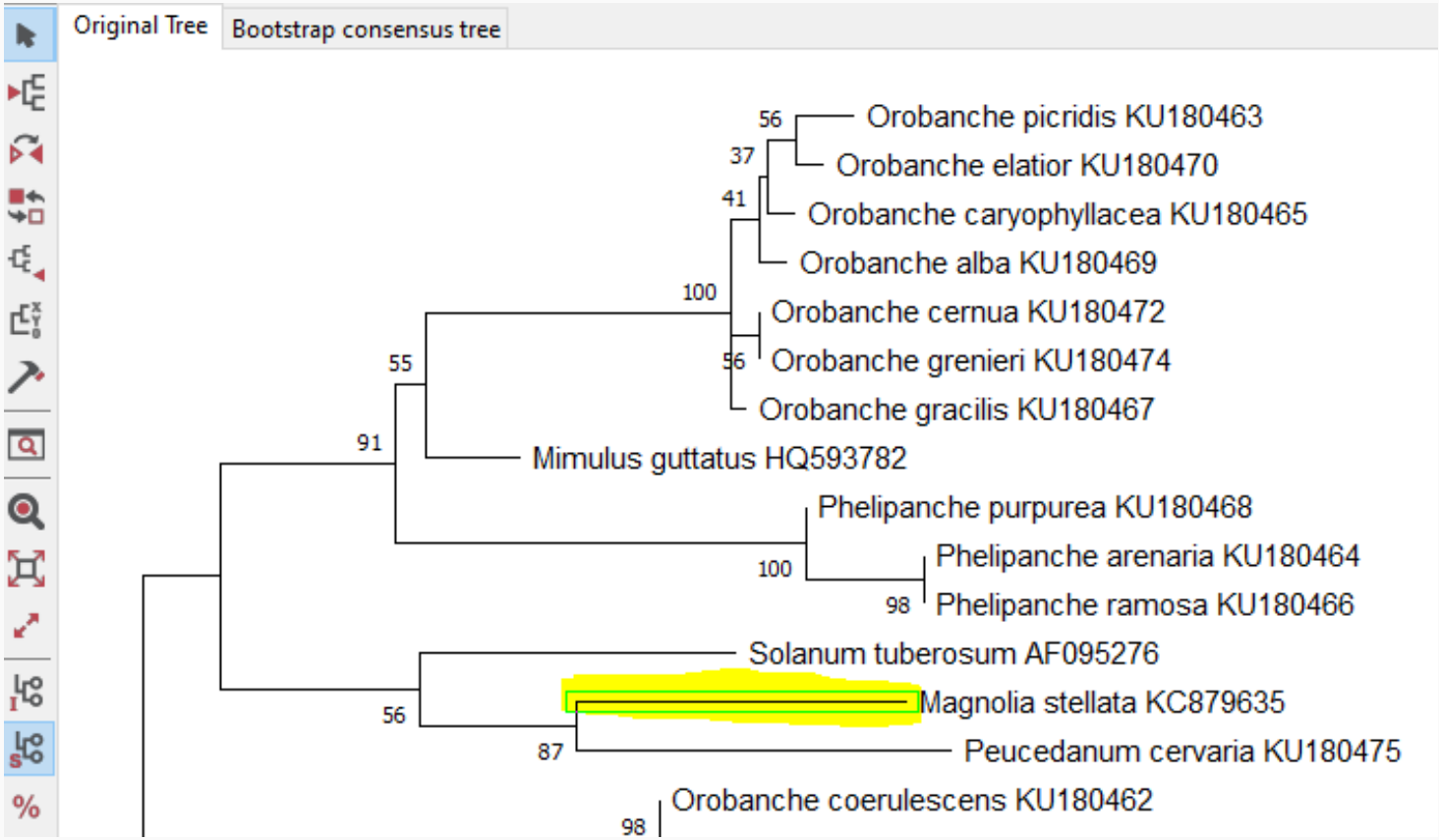
# Ukorzenianie drzewa

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



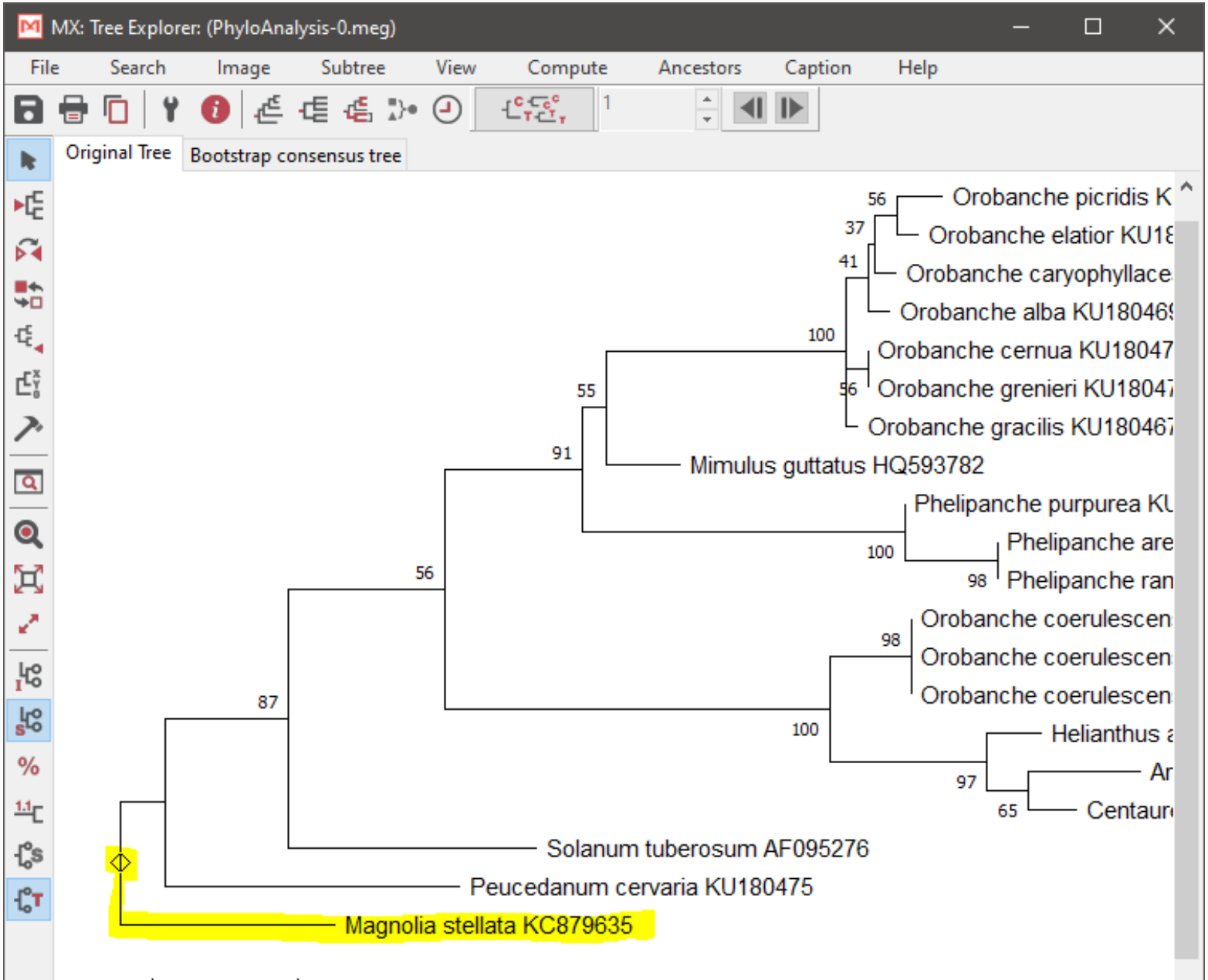
# Ukorzenianie drzewa

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



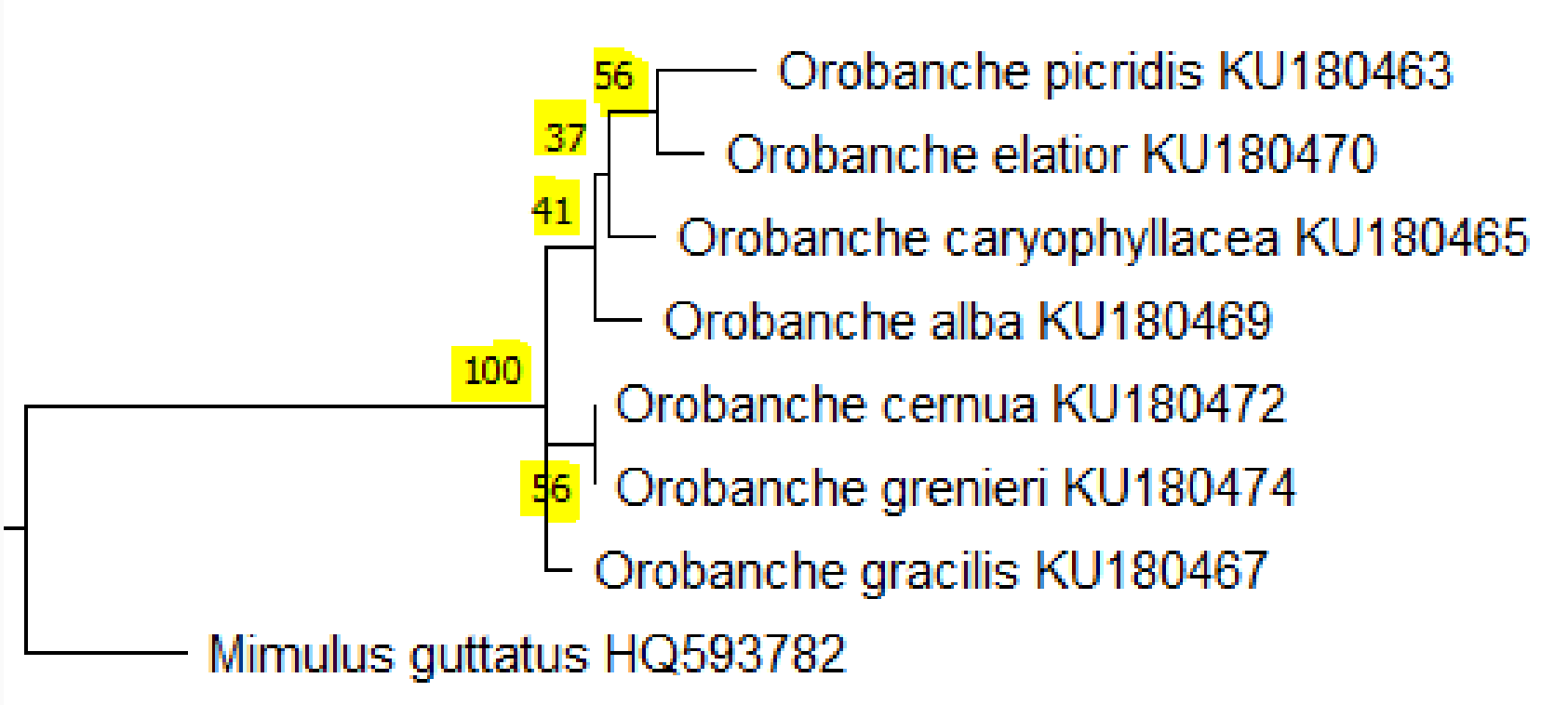
# Drzewo ukorzenione

- Teraz drzewo jest ukorzenione:



# Wartości bootstrapu

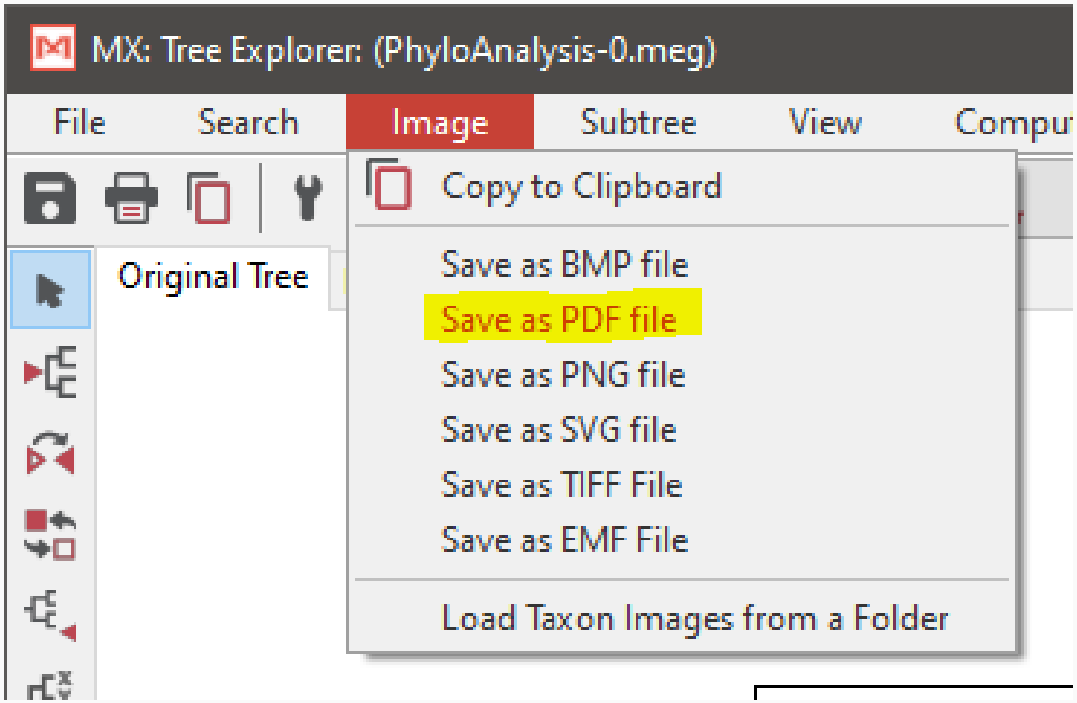
- Na drzewie widoczne są wartości bootstrapu




- Im wyższa wartość bootstrapu tym większa wiary godność węzła.

# Zapisanie drzewa jako w pliku pdf

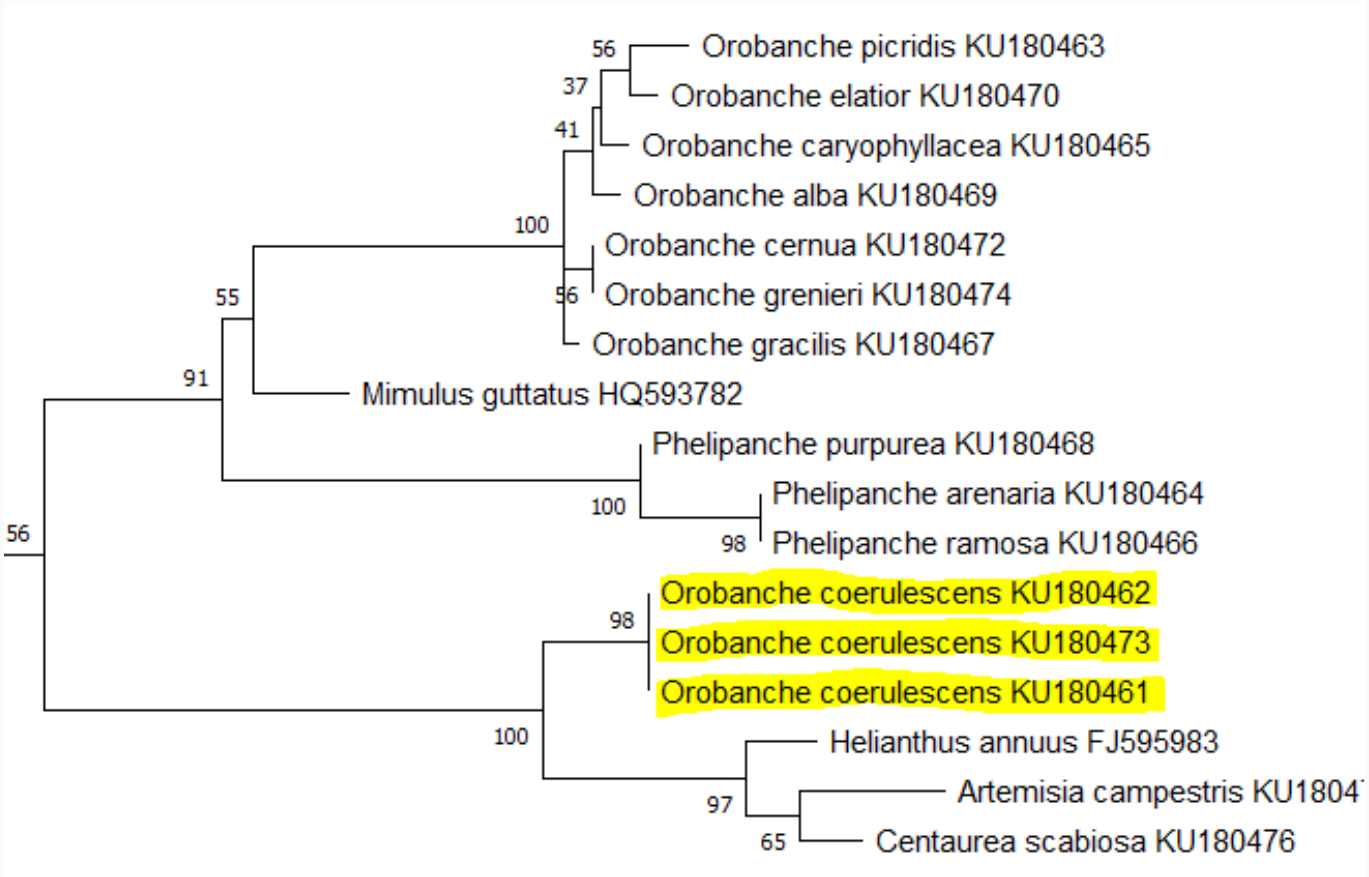
- Plik można zapisać np. w formacie pdf:



# Horyzontalny Transfer Genów (HGT)

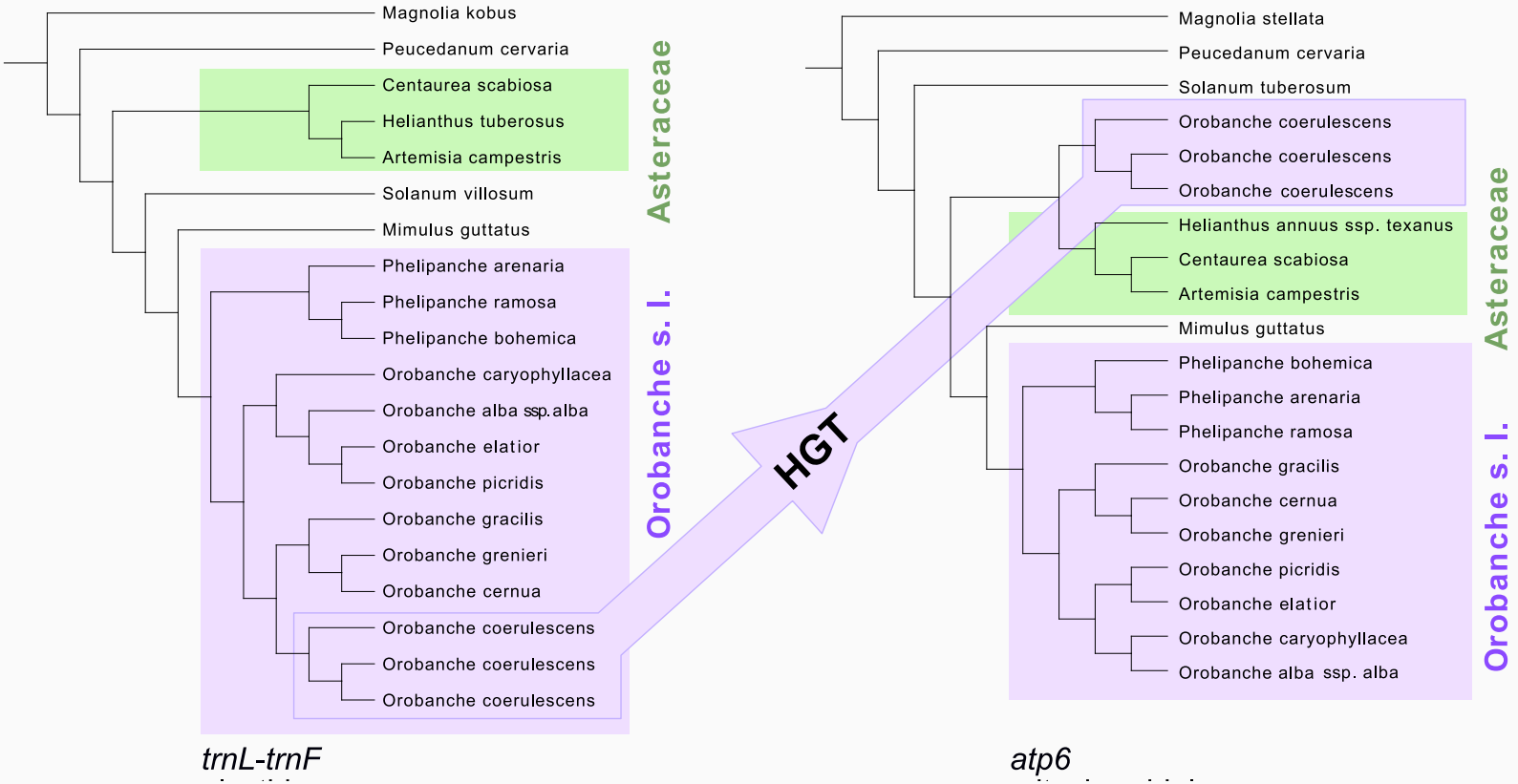
# Horyzontalny transfer genów

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



# Horyzontalny transfer genów

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



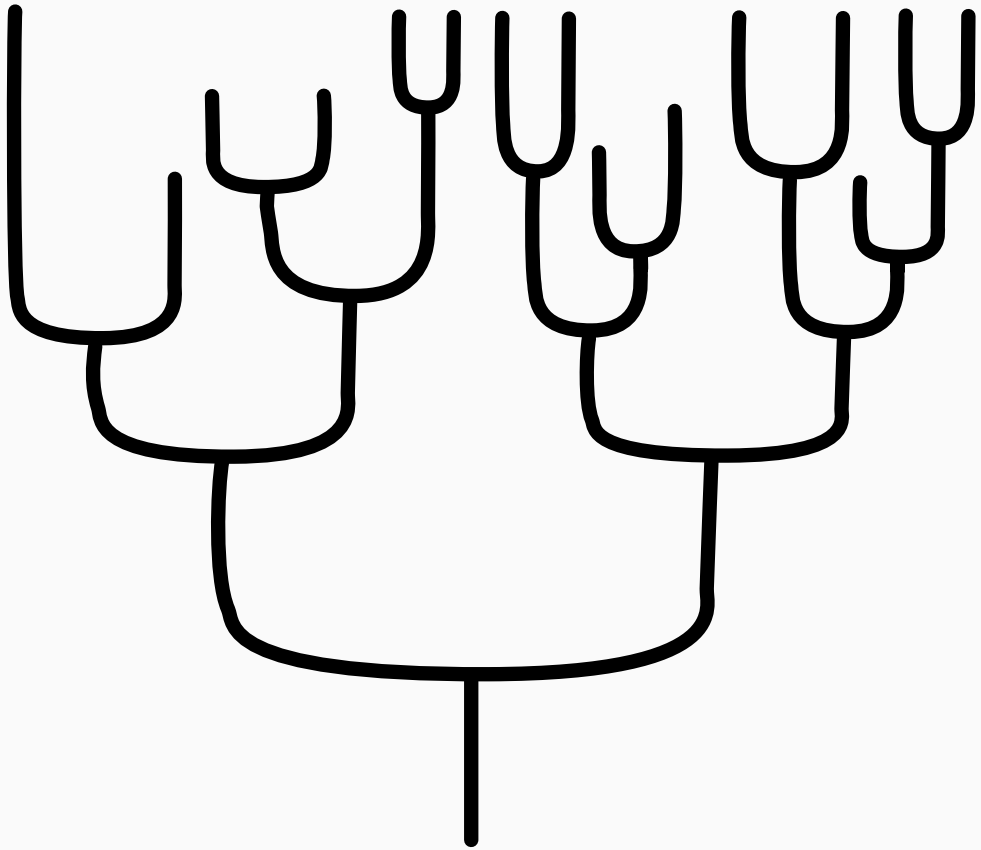
# Horyzontalny transfer genów

- To co widać na obrazku pokazuje efekt działania procesu zwanego Horyzontalnym 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

# Orobanche

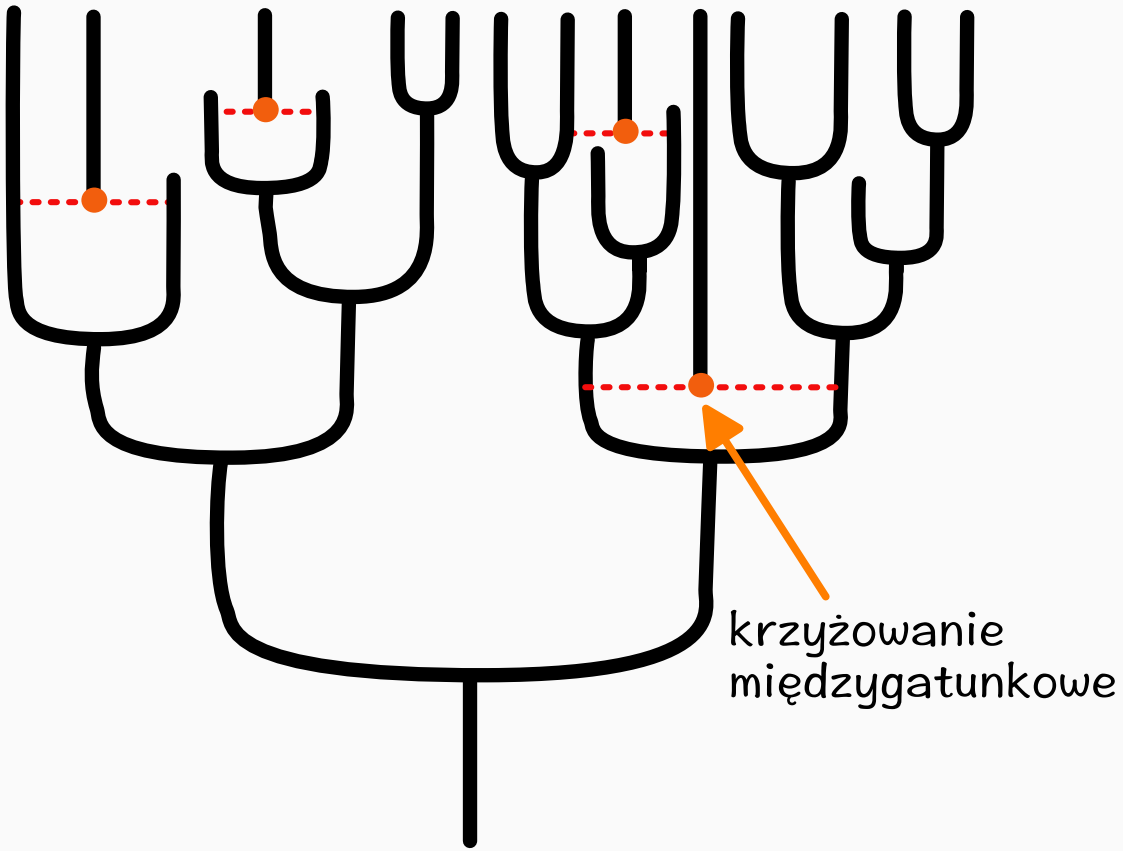
Zaraza żółta (*Orobanche flava*)
w Dolinie Strążyskiej

# Drzewo życia



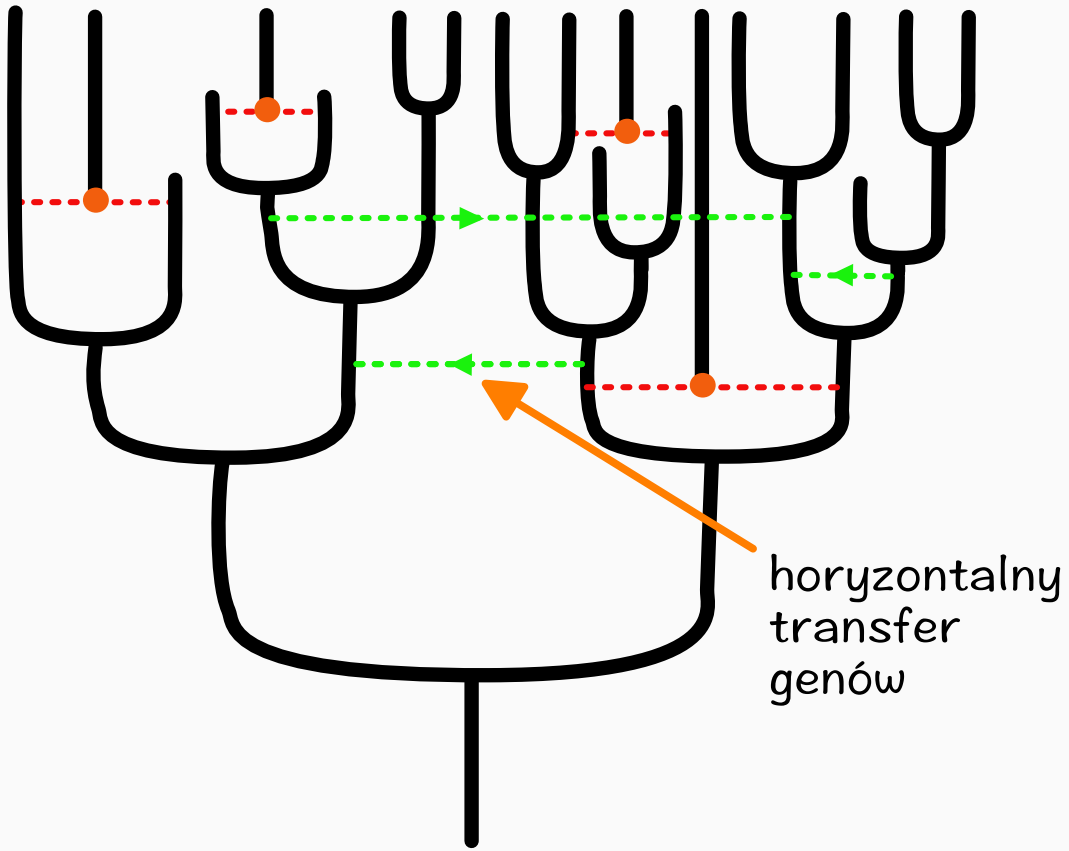
„Tradycyjne” drzewo

# Drzewo życia




Drzewo uwzględniające krzyżówki międzygatunkowe

# Drzewo życia




Drzewo uwzględniające krzyżówki międzygatunkowe i HGT

Dziękuję za uwagę.

Prezentacja, oraz pliki przydatne przy tworzeniu omawianego drzewa są dostępne ma mojej stronie internetowej pod adresem:
`ggoralski.pl`