Podstawy tworzenia drzew filogenetycznych

Grzegorz Góralski

Zakład Cytologii i Embriologii Roślin
Instytut Botaniki
Uniwersytet Jagielloński

Spis treści

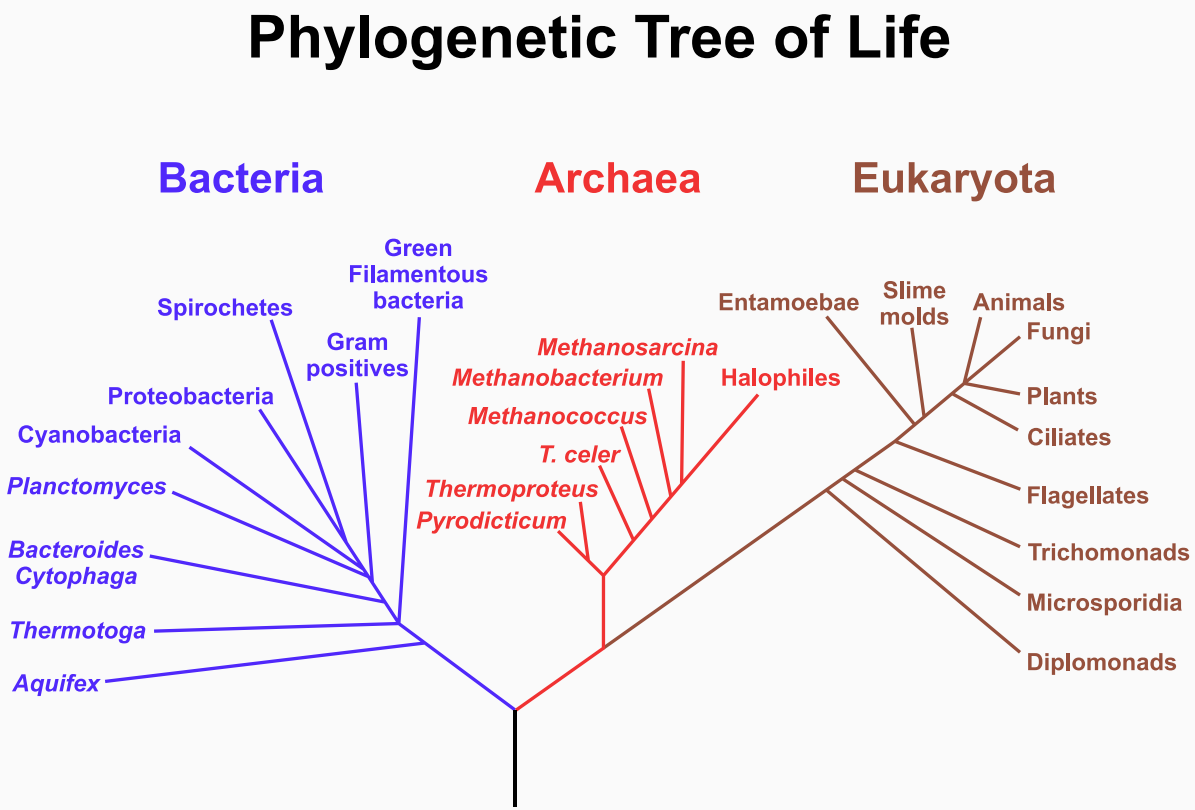
Dane

Dane do pobrania: `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

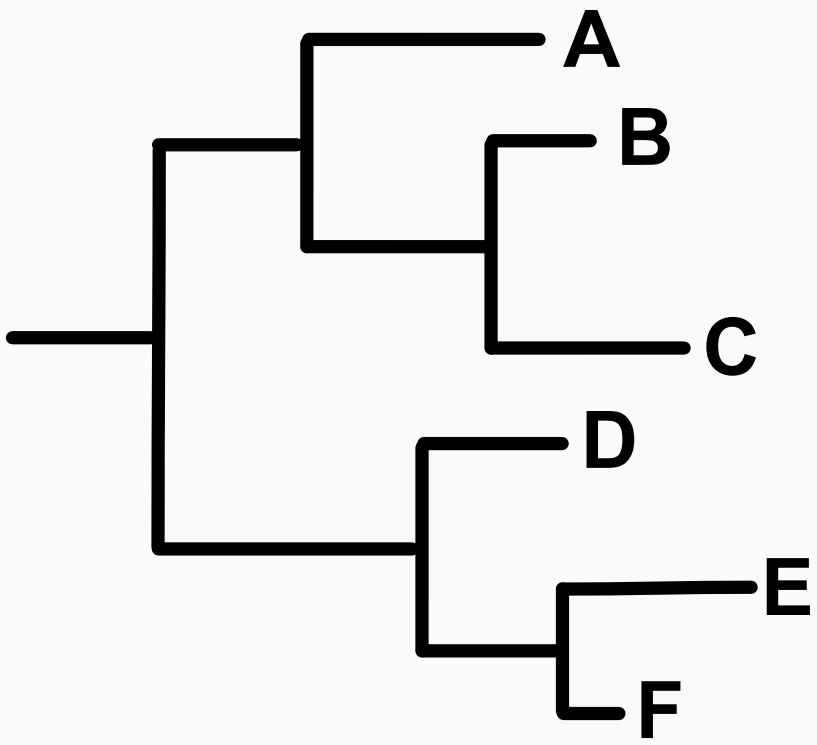


Drzewo filogenetyczne (Wikipedia)

Drzewa filogenetyczne

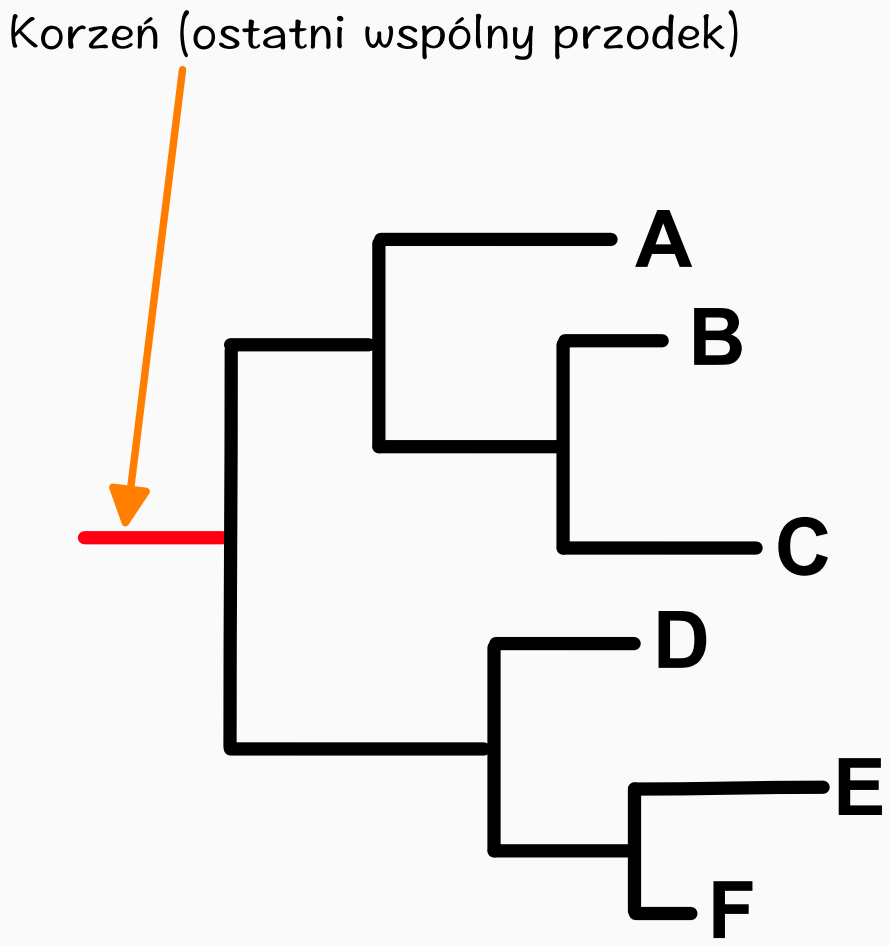
- Tworzenie drzew filogenetycznych opiera się na badaniu podobieństw i różnic pomiędzy organizmami
- W tego typu badaniach bierze się pod uwagę cechy morfologiczne, anatomiczne itp. i/lub genetyczne
- W drzewach opartych na fragmentach DNA bierze się pod uwagę różnice w badanej sekwencji pomiędzy organizmami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

Struktura drzewa



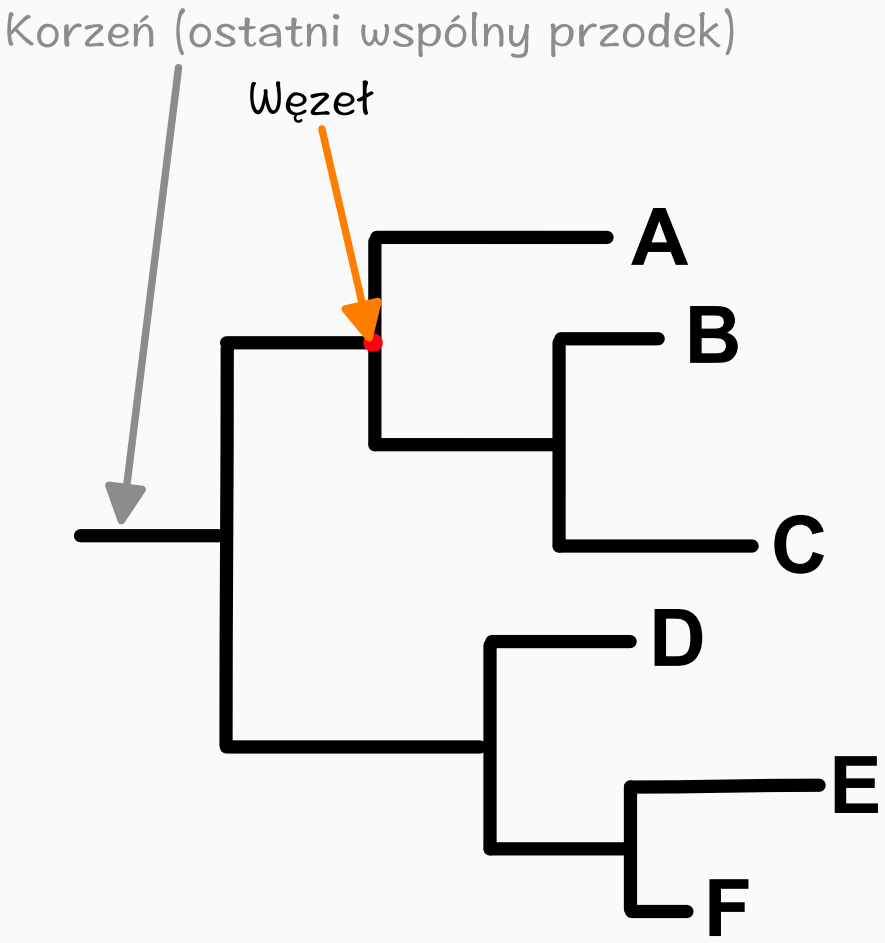
Struktura drzewa

Struktura drzewa



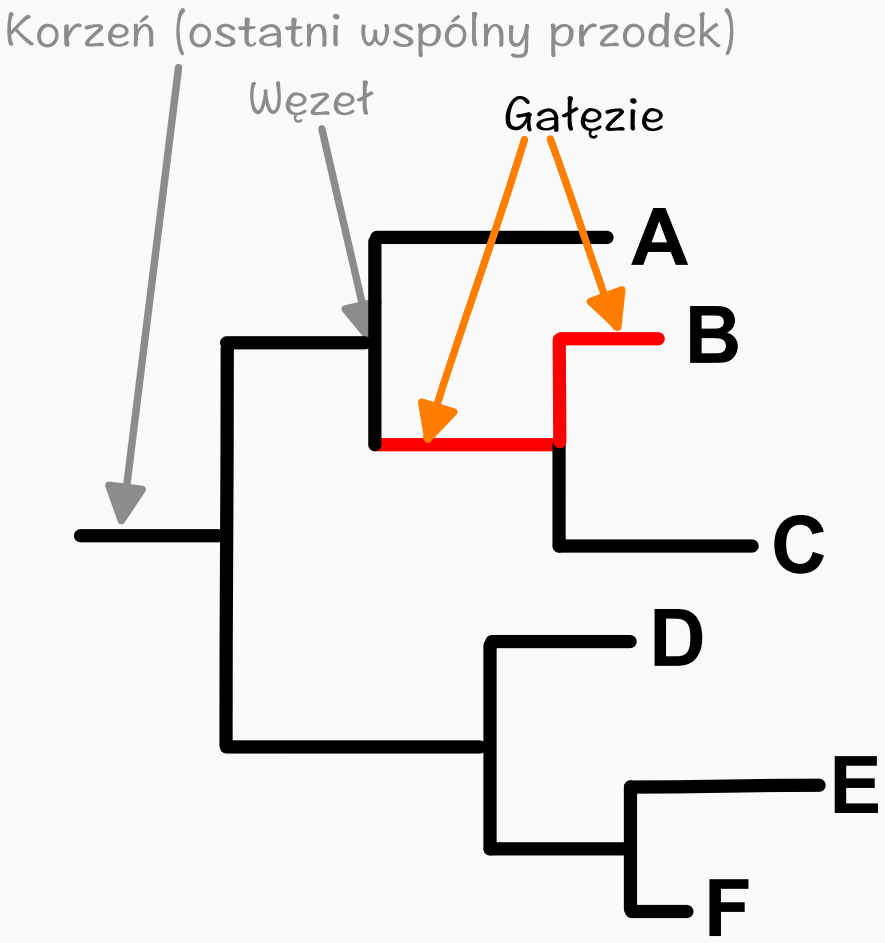
Struktura drzewa

Struktura drzewa



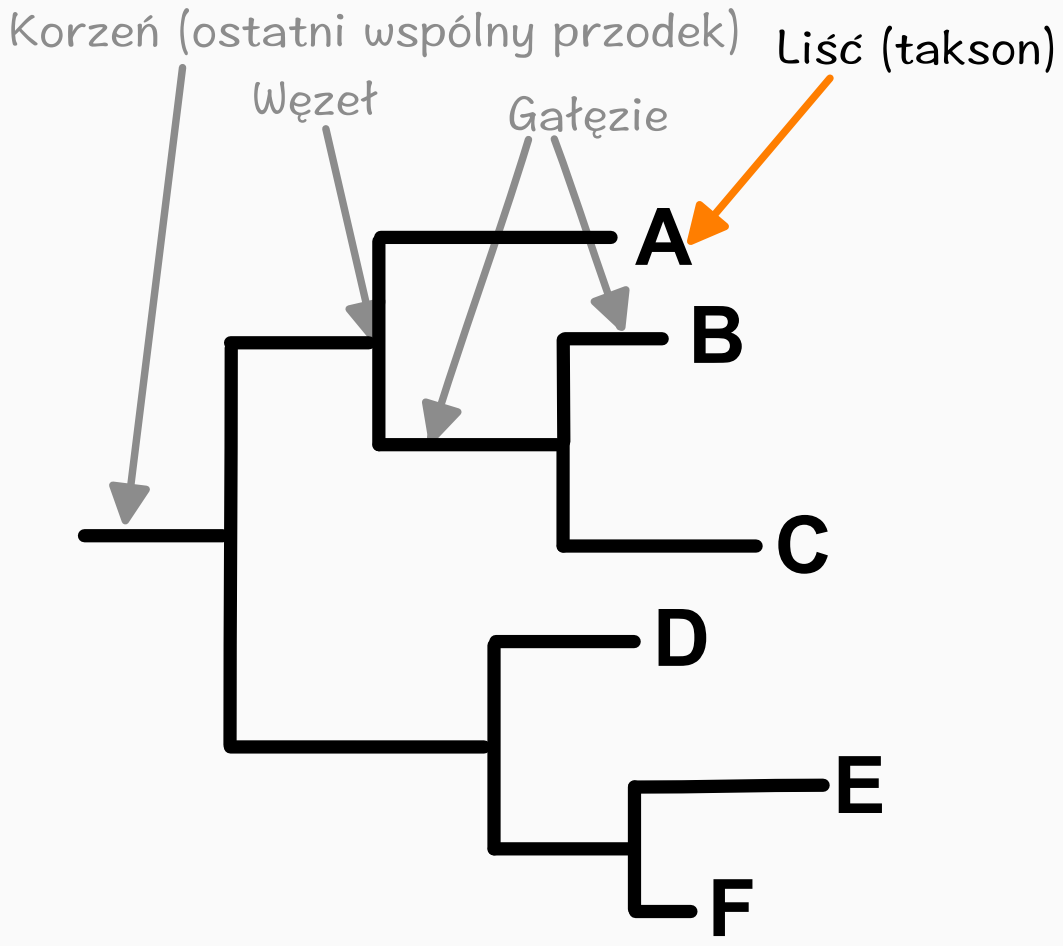
Struktura drzewa

Struktura drzewa



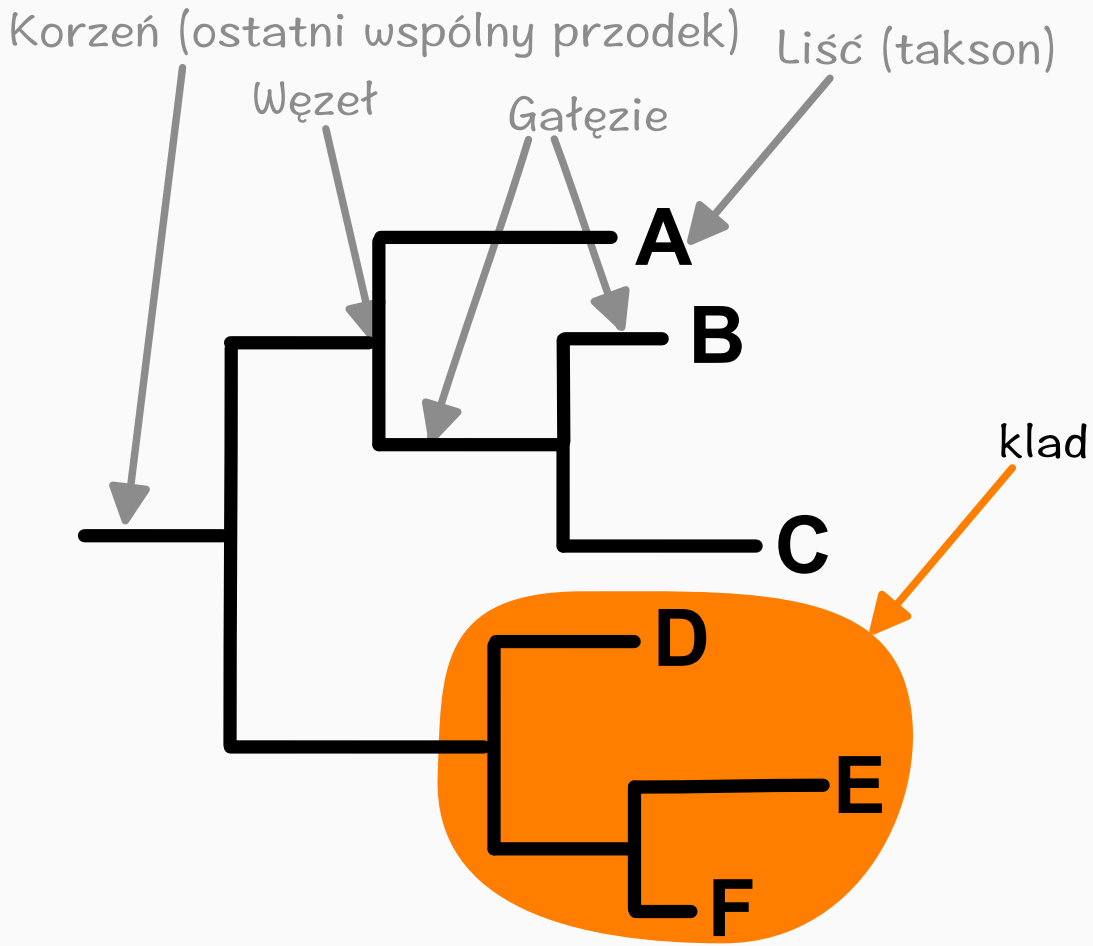
Struktura drzewa

Struktura drzewa



Struktura drzewa

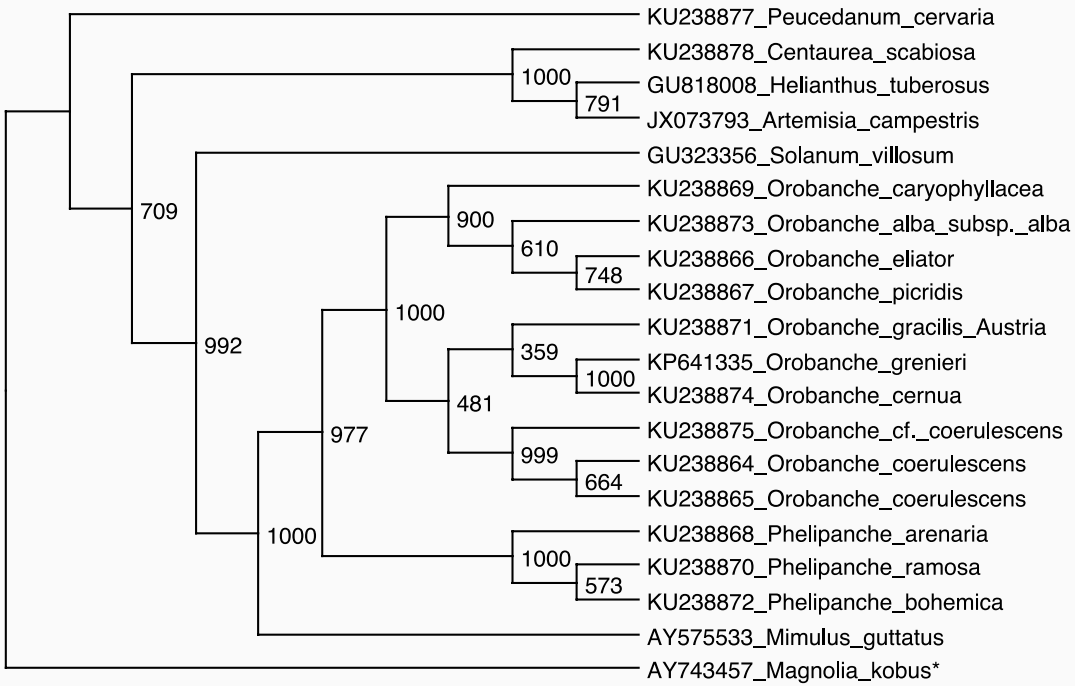
Struktura drzewa



Struktura drzewa

Podstawowe typy drzew

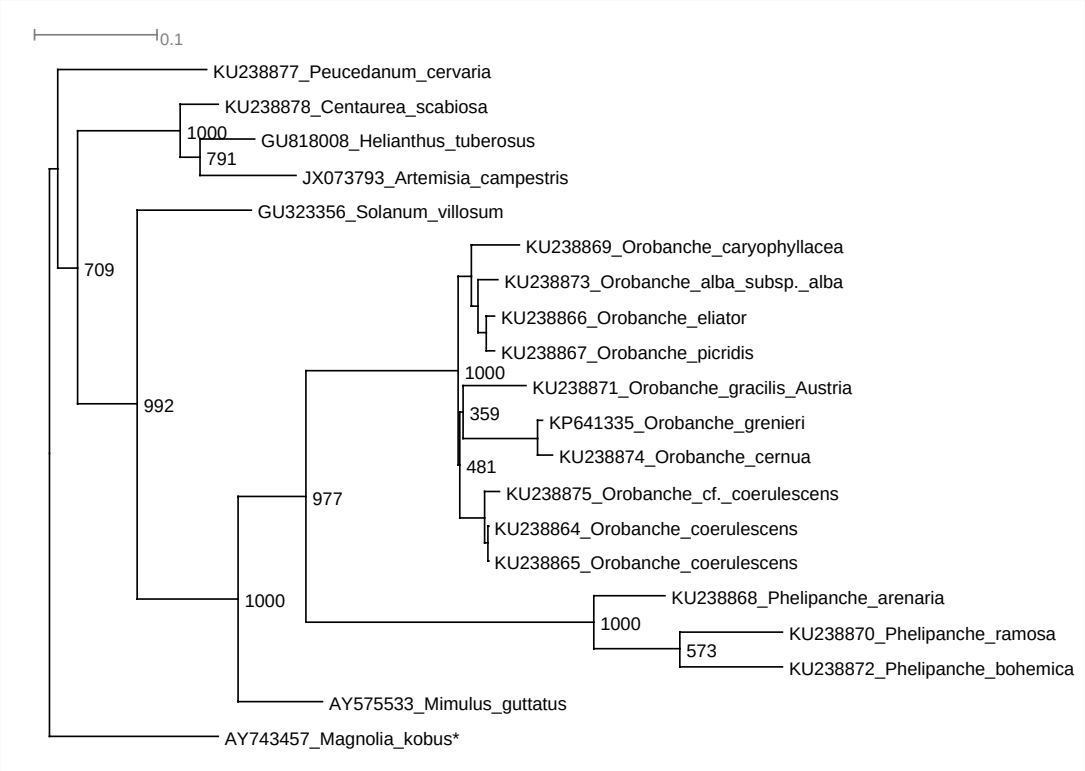
- Kladogram: pokazuje pokrewieństwa ale długość gałęzi nie pokazuje liczby mutacji



Kladogram

Podstawowe typy drzew

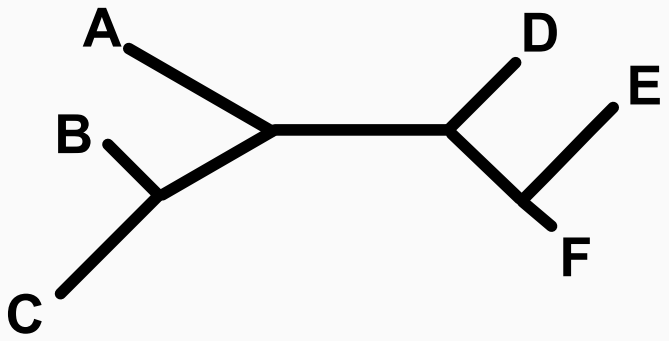
- Filogram: pokazuje pokrewieństwa, długość gałęzi odpowiada liczbie zmian



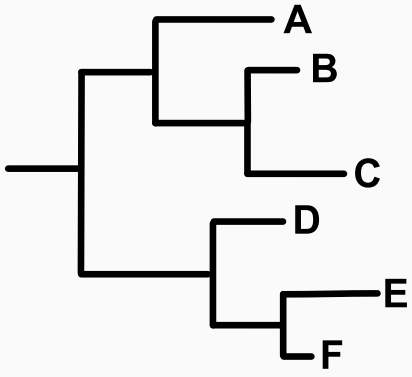
Filogram

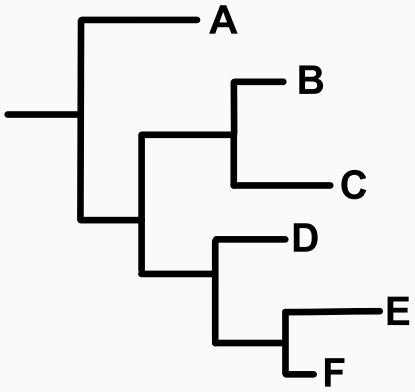
Nieukorzenione drzewa

- 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:

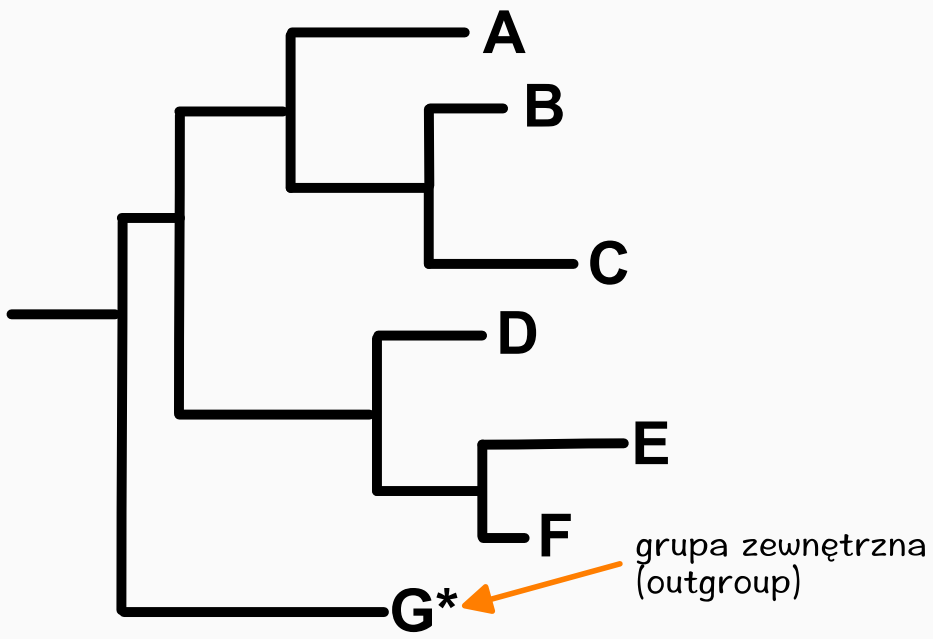




- itd...

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.

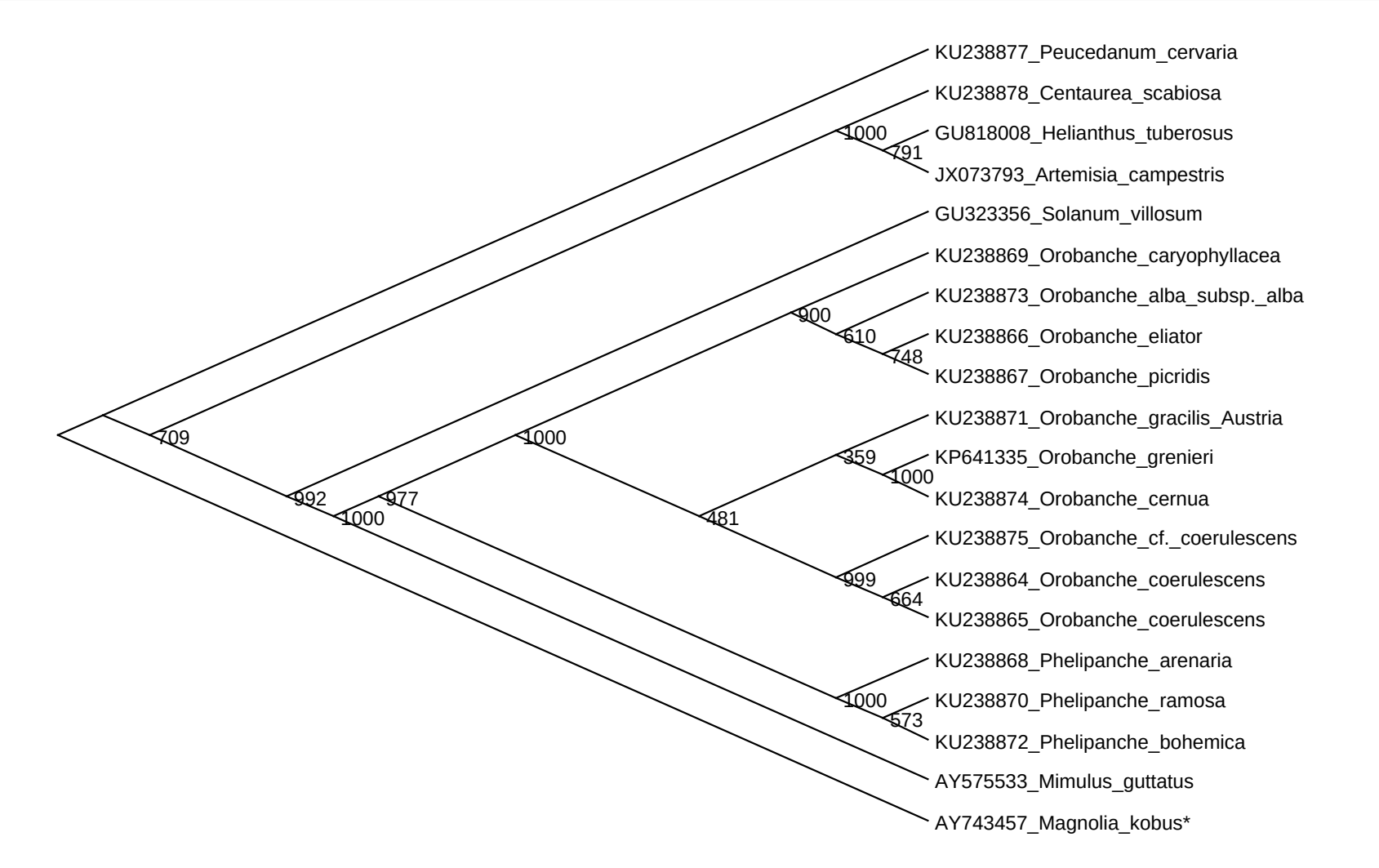


Typy drzew

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

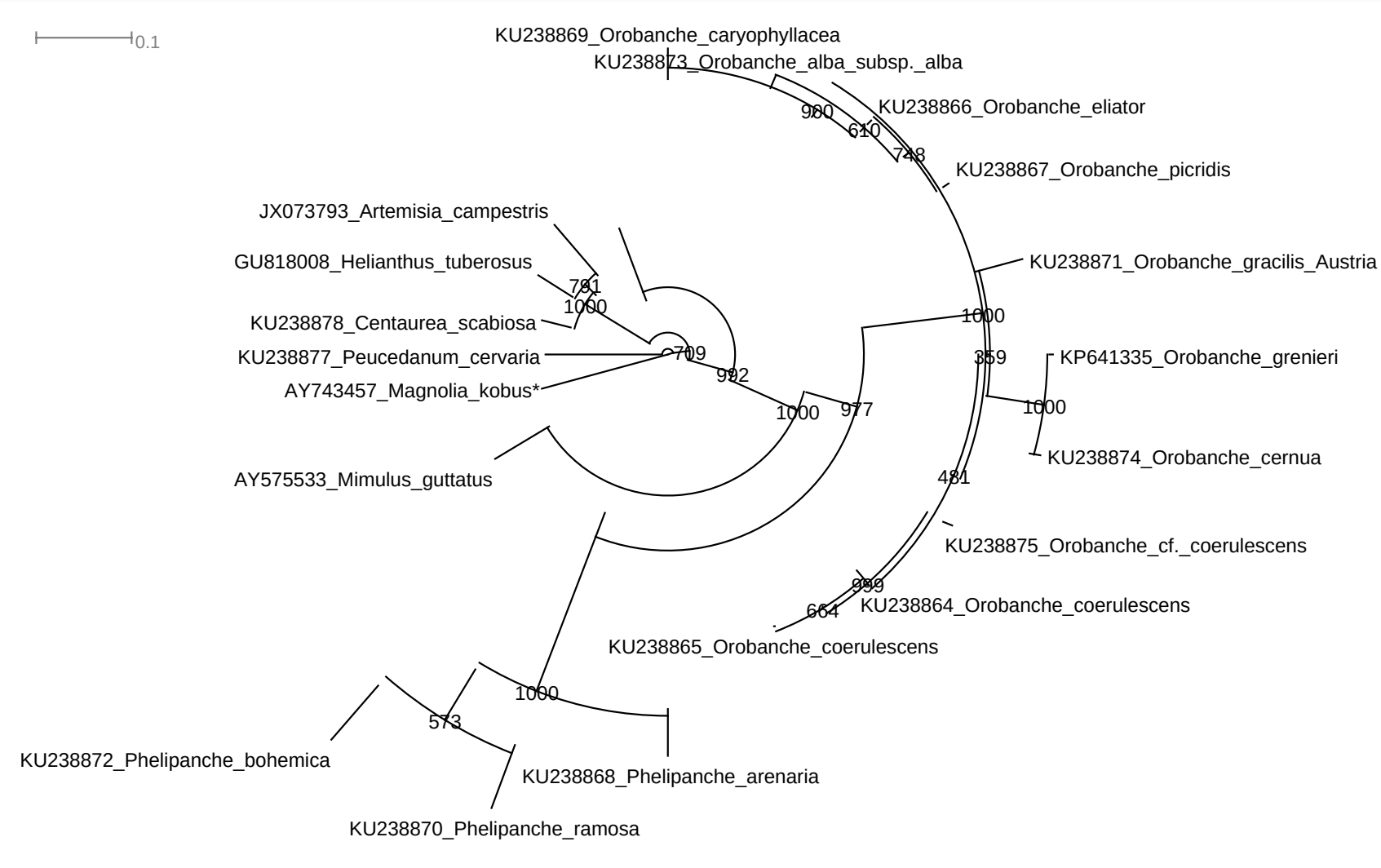
Typy drzew

- 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

Konstruowanie drzewa filogenetycznego składa się z wielu etapów:

- 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: 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 rodzaju sekwencji

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.

Wybór rodzaju sekwencji

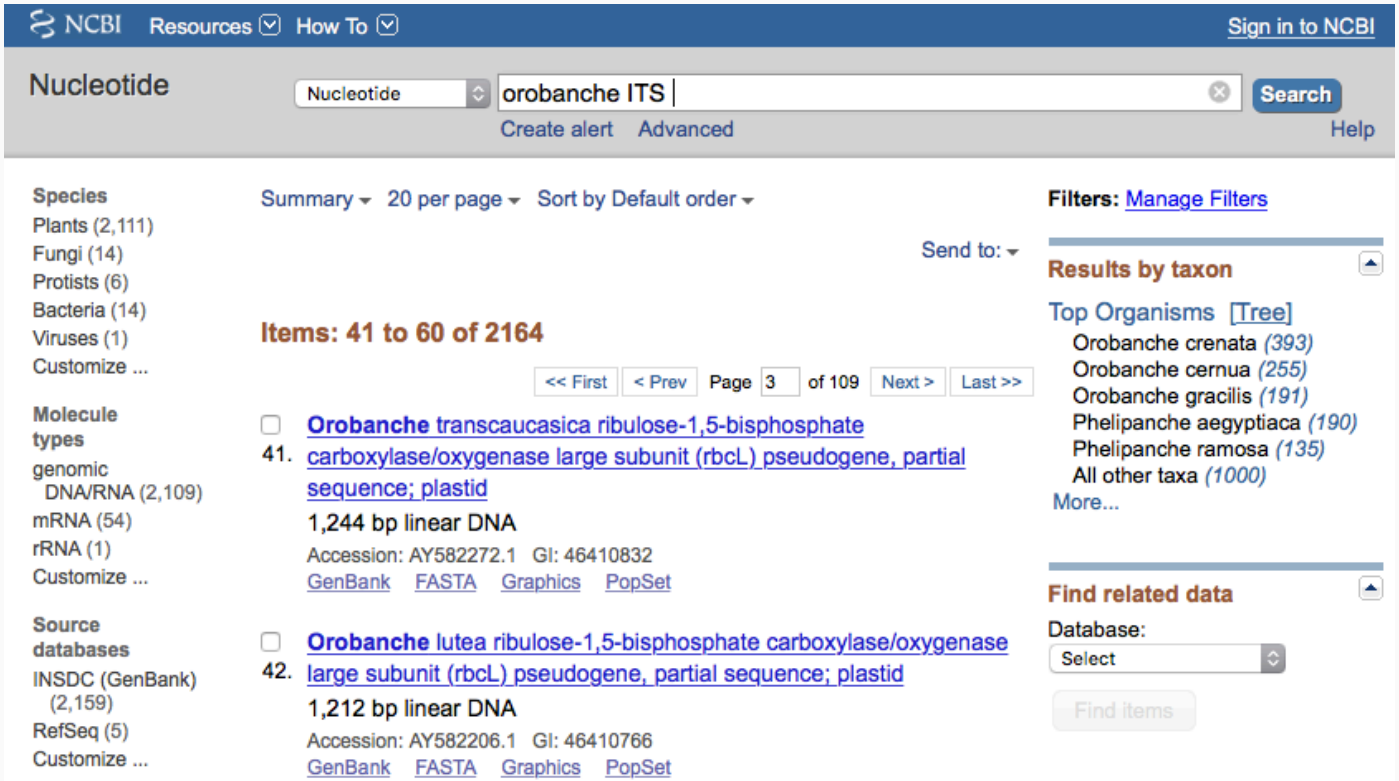
- Tempo ewolucji
  - Różne sekwencje zmieniają się w różnym tempie
  - Generalnie geny ewoluują dużo wolniej niż sekwencje niekodujące (np. geny białek histonowych są bardzo konserwatywne, białek kolagenowych są bardzo zmienne)
  - Geny także różnią się tempem ewolucji
  - Jeśli badamy blisko spokrewnione 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).

Wybór rodzaju sekwencji

- **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)

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 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
AACTCAGTACCAAATGCTTGGCAATCCTTGGTAGAGCTTATTCATGATTTCGT
CCGGTAAACGAACAAATAGGTGGTCTTTCCGGAAATGTTCAACAAAAGTTTTC

>AF095276_Solanum_tuberosum
CTACTAACTCTCAGTTTGGTCCTACTTTTGGTTTATTTTGTTACTAAAAAGGG
AACTCAGTACCAAATGCTTGGCAATCCTTGGTAGAGCTTATTTATGATTTCGT
CCGGTAAACGAACAAATAGGTGGTCTTTCCGGAAATGTTAAACAAAAGTTTTC
```

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

Edycja plików tekstowych

- **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 **edytoru 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

- 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
CAAGAACTGCGCTAAACACTTAGCCA
AACTGCGCTAACACTTAGC
```

- 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)

Automatyczne dopasowanie

- Ręczne dopasowanie sekwencji jest zwykle bardzo czasochłonne. Zwłaszcza jeśli są długie, są bardzo zmienne, jest ich wiele, gdy posiadają indele.
- 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 ewoluujące niezależnie od siebie różnią 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.
- Kluczową zagadnieniem dla rekonstrukcji ewolucji na podstawie porównania sekwencji DNA jest określenie „**odległości**” miedzy 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:

  $$A \to T, T \to G$$

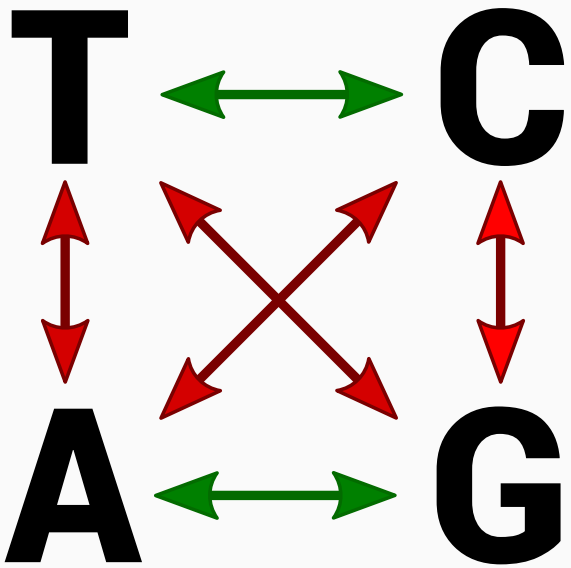
  to te dwie zmiany będą widoczne jako jedna zmiana, jedna z nich będzie „ukryta”.
- Co więcej, może dojść to mutacji:

  $$A \to T, T \to A$$

  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.

Modele ewolucji molekularnej

- 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):



Modele ewolucji molekularnej

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

	T	C	A	G
T	-3α	α	α	α
C	α	-3α	α	α
A	α	α	-3α	α
G	α	α	α	-3α

Gdzie: α - prawdopodobieństwo mutacji.
Wartość -3α wynika z tego, że p-stwo w każdym rzędzie = 0.

Modele ewolucji molekularnej

- 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.

Metody konstrukcji drzew filogenetycznych

- 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.

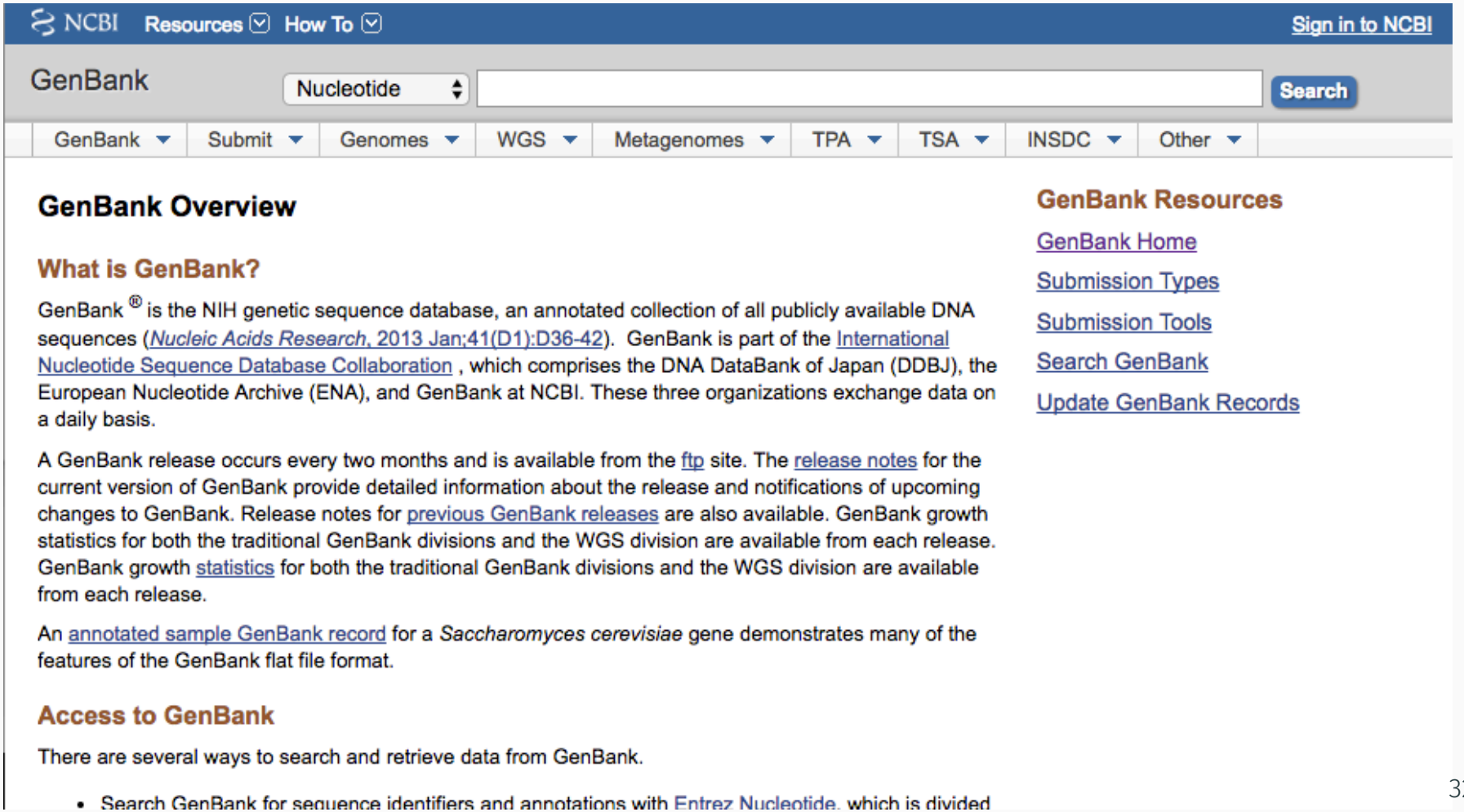
Część praktyczna

Internetowe bazy sekwencji nukleotydów

- **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.

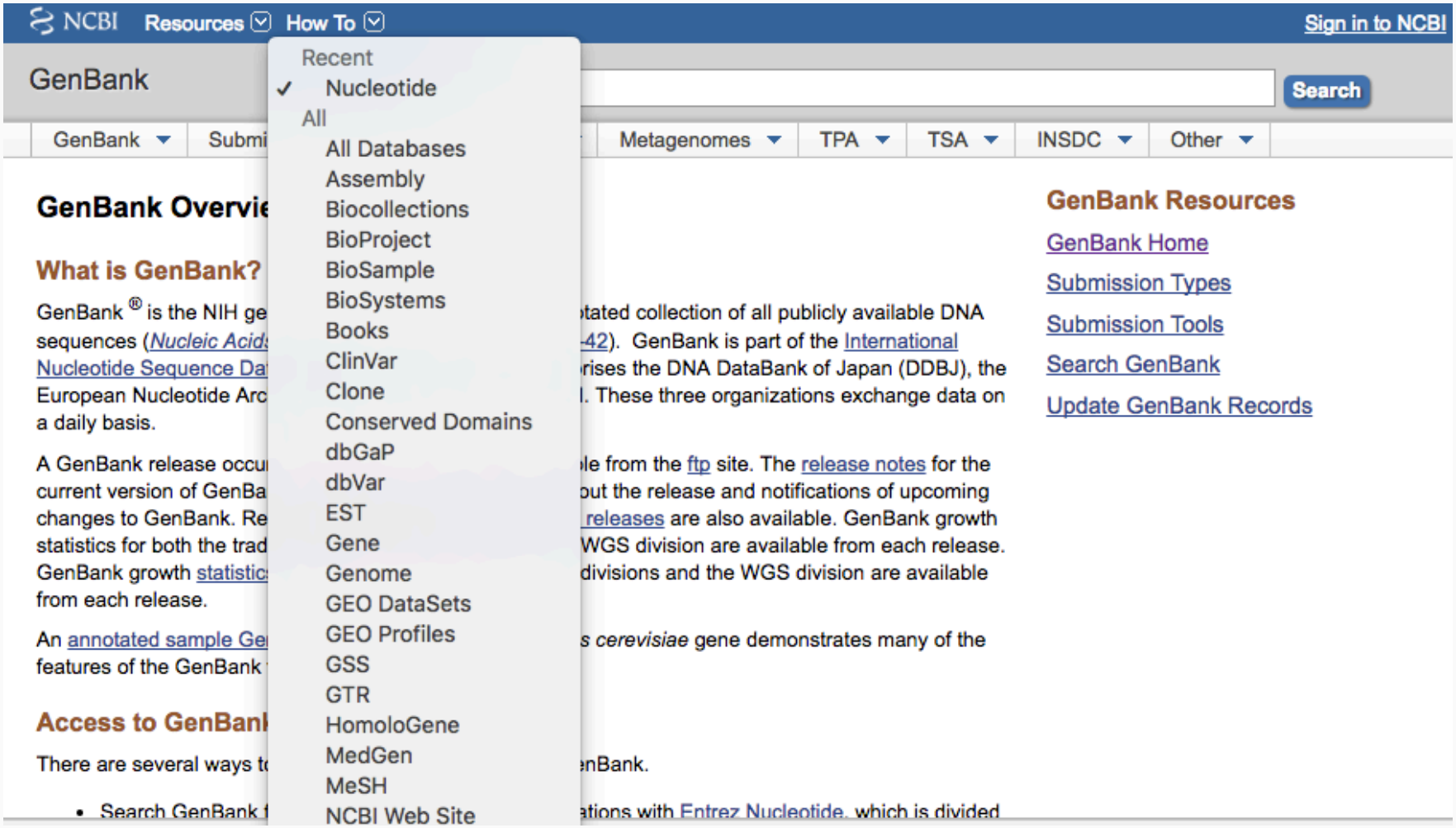
GenBank - strona główna

Po wejściu na stronę GenBank-u `https://www.ncbi.nlm.nih.gov/genbank/` widać informacje na temat bazy oraz pole wyszukiwania:



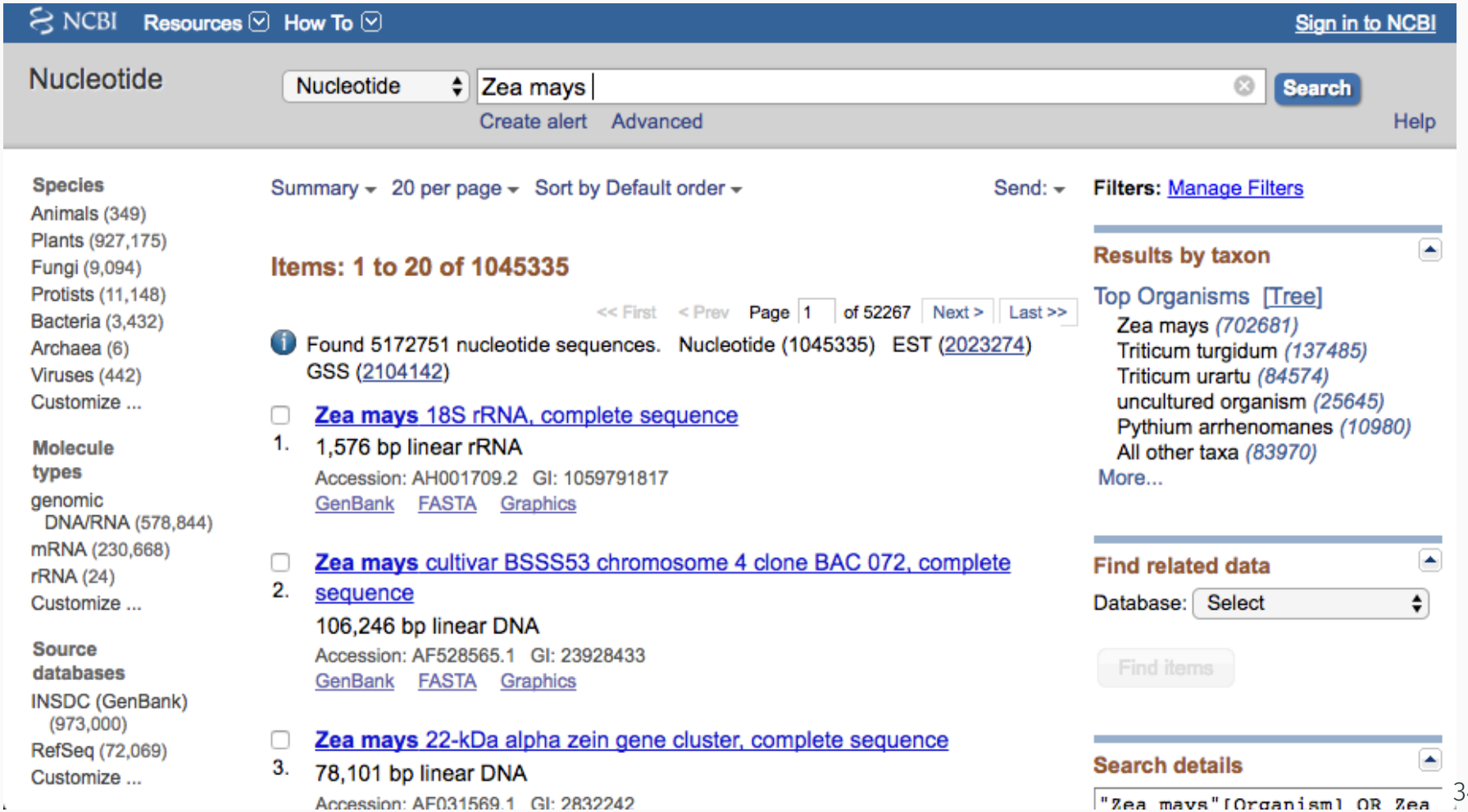
GenBank - menu wyszukiwania

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



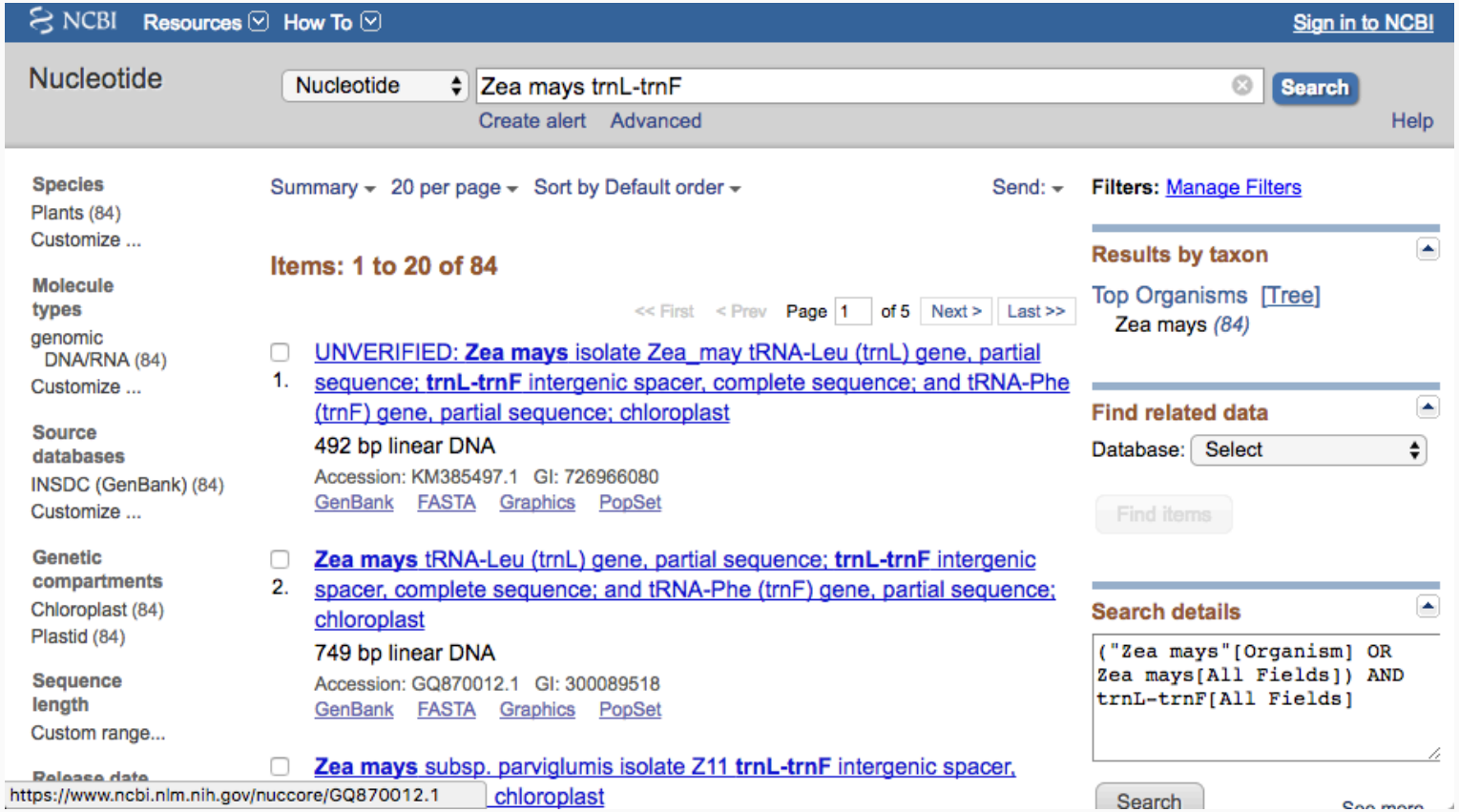
GenBank - wyszukiwanie sekwencji

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



GenBank - wyszukiwanie sekwencji

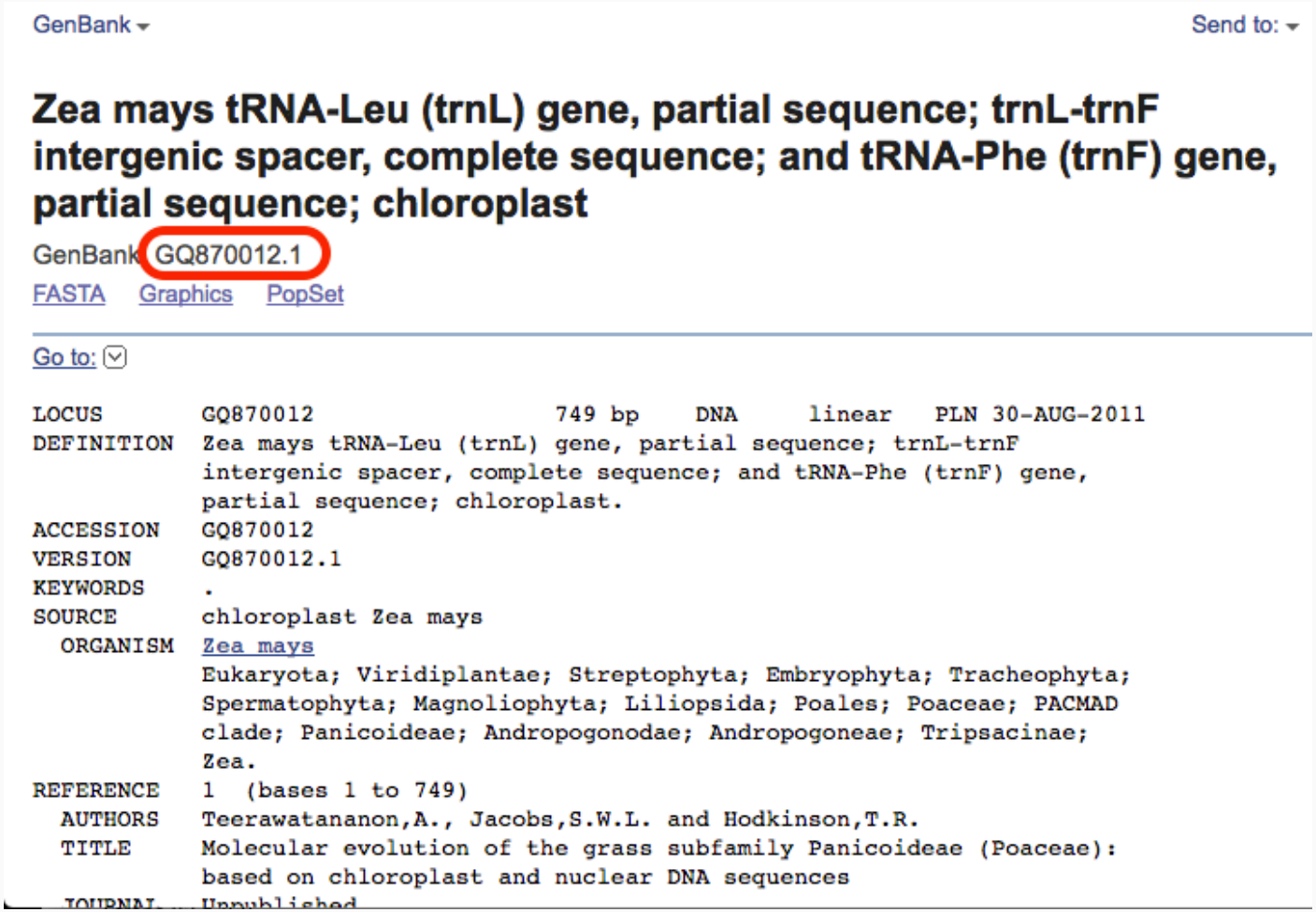
Można zawęzić wynik dla konkretnego rodzaju sekwencji, np. *trnL-trnF*:



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

GenBank - dane sekwencji

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 GQ870012.1

FASTA Graphics PopSet

Go to:

```
LOCUS       GQ870012                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   GQ870012
VERSION     GQ870012.1
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
```

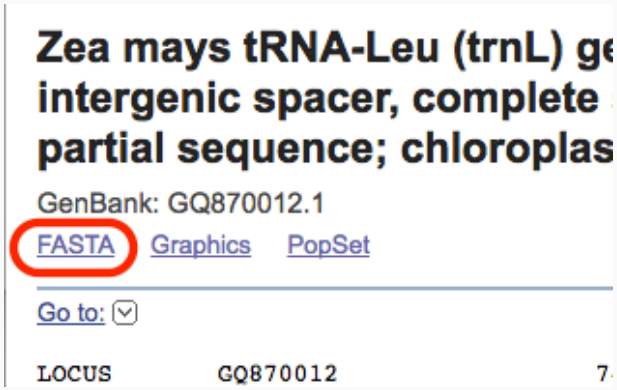
GenBank - dane sekwencji

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 tgagccaaat ccctttttg aaaaacaagt ggttctcaaa ctagaaccca
       61 aaggaaaagg ataggtgcag agactcaatg gaagctgttc taacgaatcg aagtaataac
      121 gattaatcac agaacccata ttataatata ggttctttat tttatttta gaatgaaatt
      181 aggaatgatt atgaaataga aaattcataa ttttttttta gaattattgt gaatctattc
      241 caatcaaata ttgagtaatc aaatccttca attcattgtt ttcgagatct tttaattttt
      301 aaaagtggat taatcggacg aggataaaga gagagtccca ttctacatgt caatactgac
      361 aacaatgaaa tttctagtaa aaggaaaatc cgtcgacttt ataagtcgtg agggttcaag
      421 tccctctatc cccaaaccct cttttattcc ctaaccatag ttgttatcct ttttttcttt
      481 tatcaatggg tttaagattc actagctttc tcattctact ctttcacaaa ggagtgcgac
      541 aagaactcaa tgaatcttat gctattcatt aaatagatga tttctttttt atttcttttt
      601 tatttattag agtagagtat cggcaaggaa tctcgattat taattcgttt tttttaagta
      661 ttattaagta agccatccac aatgcatagg actacccctc cccatttcct aattttgaat
      721 ggaatacttt attgattttt tagtccctt
//
```

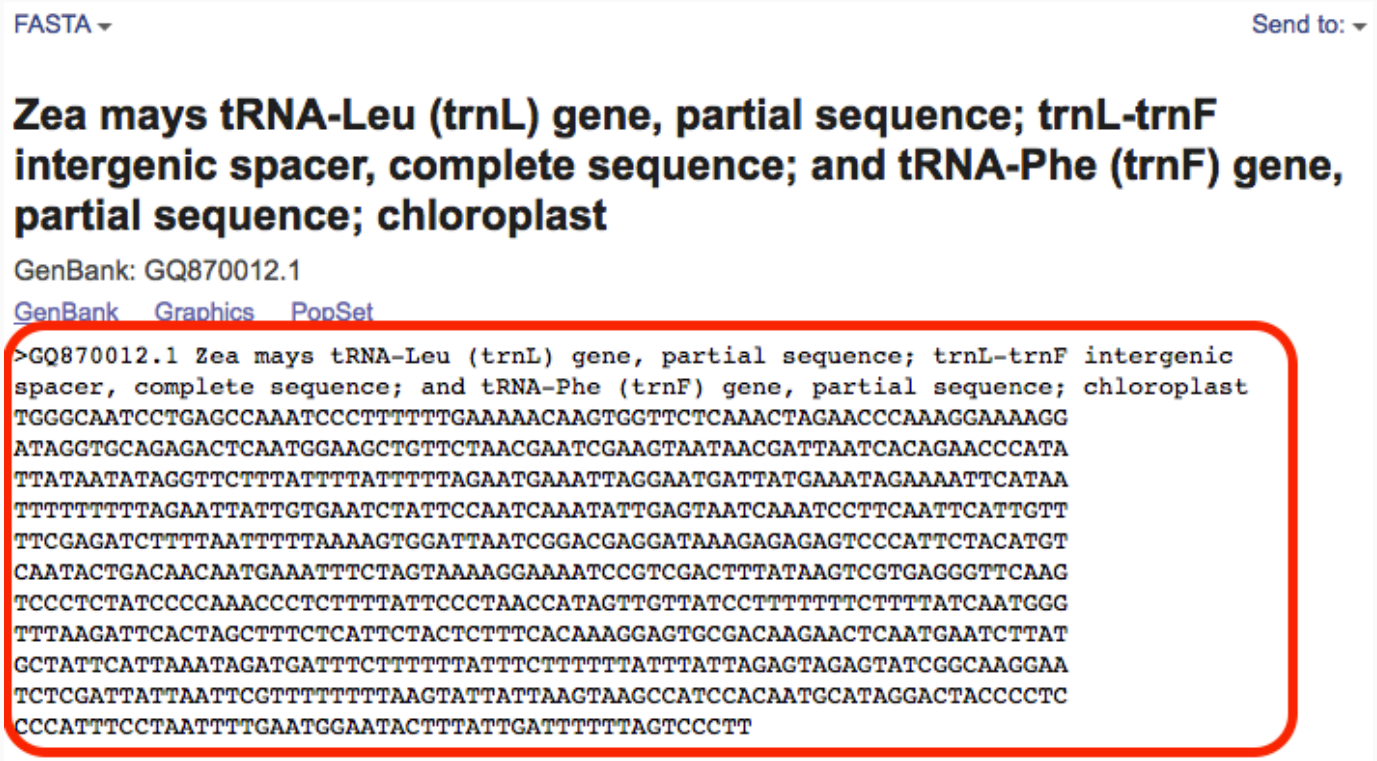
GenBank - pobieranie sekwencji w formacie FASTA

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



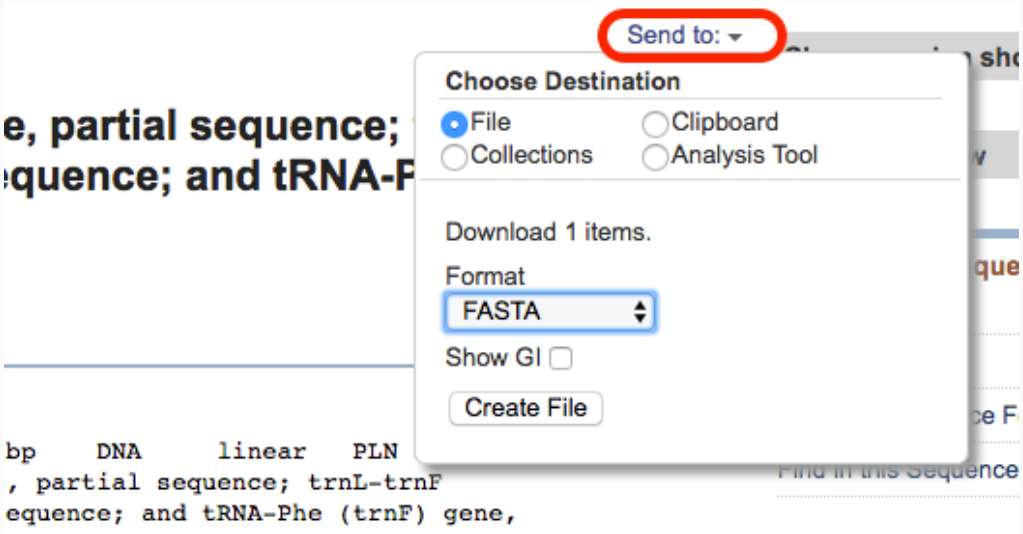
GenBank - pobieranie sekwencji w formacie FASTA

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.



GenBank - pobieranie sekwencji w formacie FASTA

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



Wyszukanie sekwencji

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

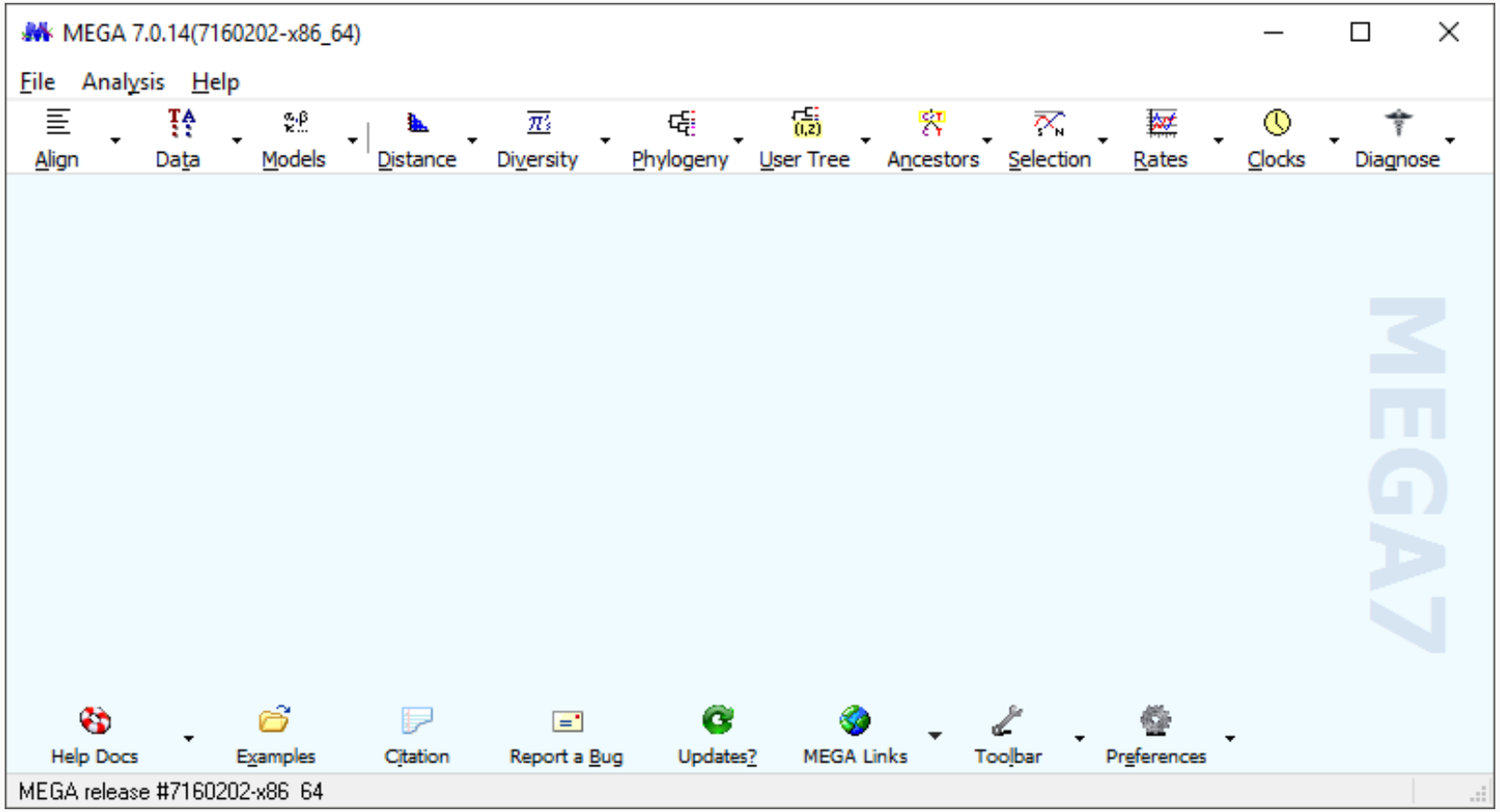
- Dostępnych jest bardzo wiele algorytmów i implementujących je programów a także programów, które wykorzystują narzędzia dostępne przez internet
- Programy działają w formie aplikacji okienkowych lub z linii komend, niektóre na oba sposoby
- Narzędzia mogą być płatne lub darmowe
- Niektóre programy do dopasowania sekwencji: ClustalW, Mutt, Mafft, Muscle, T-Coffee
- Niektóre metody do tworzenia drzew: najbliższego sąsiada (*Neghbor-Joining*), największej wiarygodności (*Maximum Likehood*), 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*

MEGA

- 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

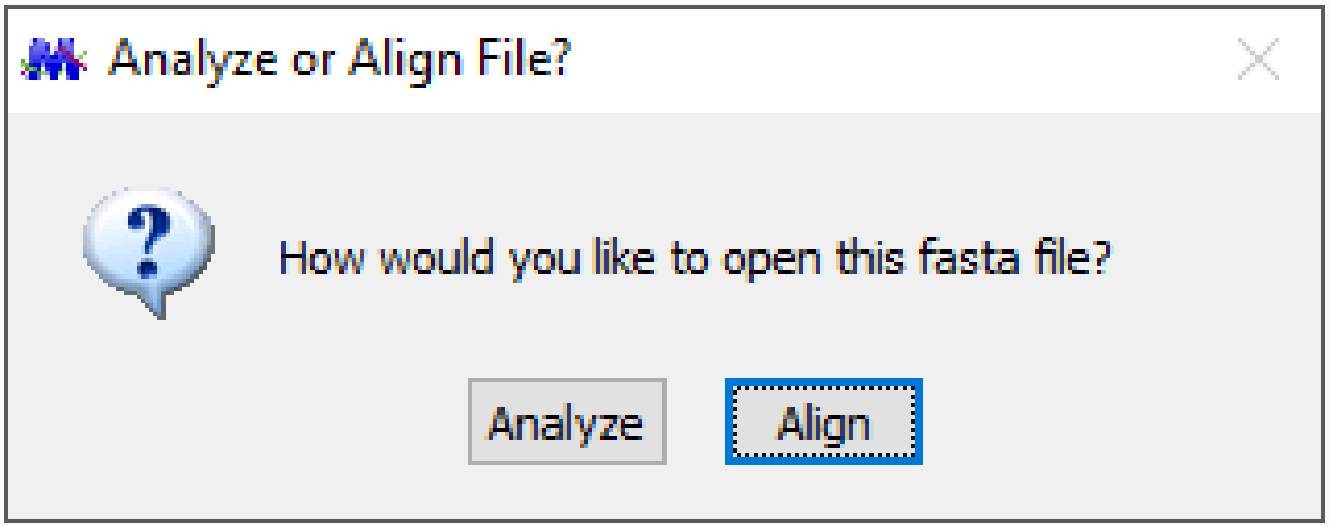
MEGA - interfejs

- Po otwarciu programu pokazuje się taki interfejs:



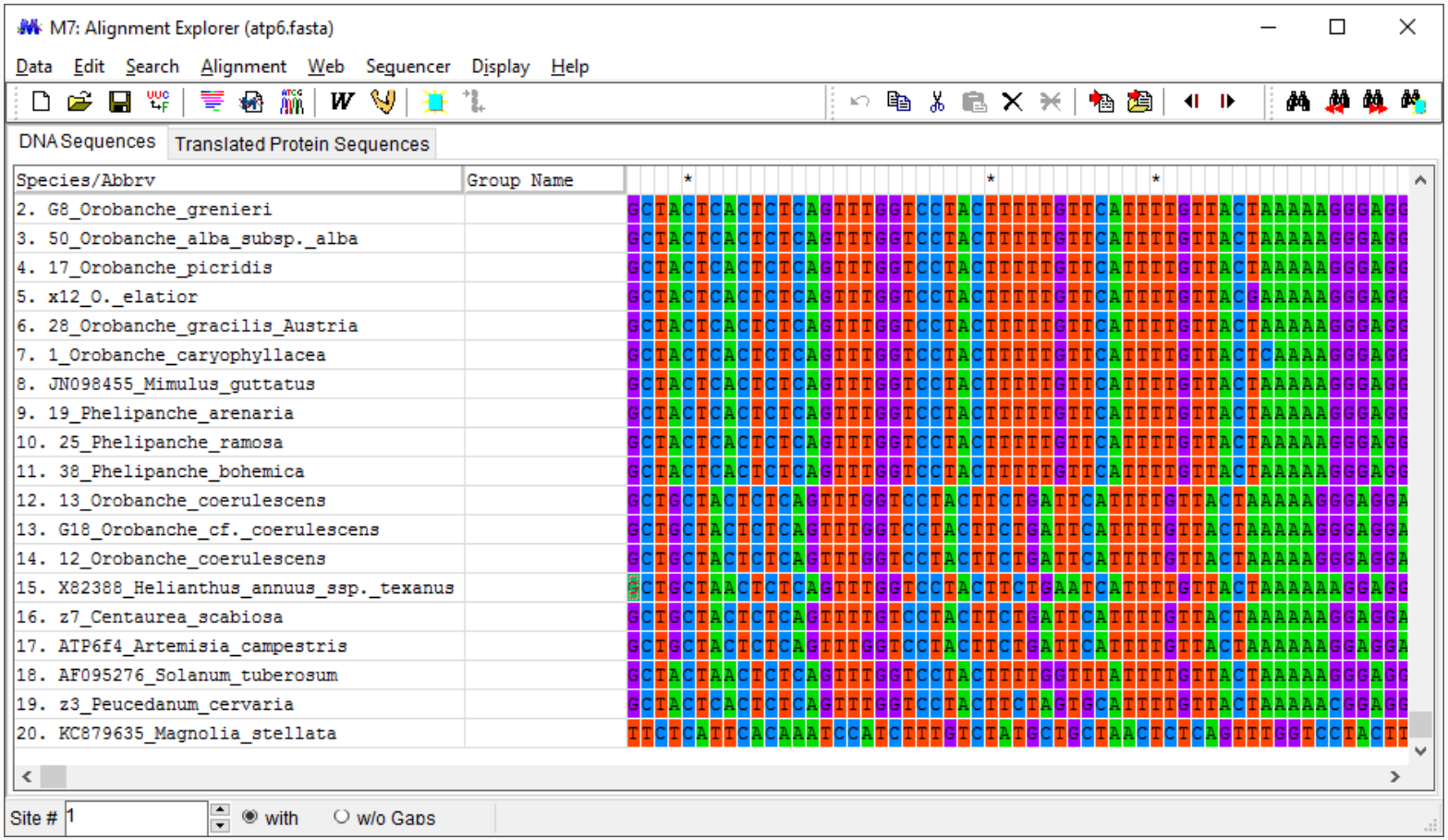
Otwieramy plik do wyrównania

- 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

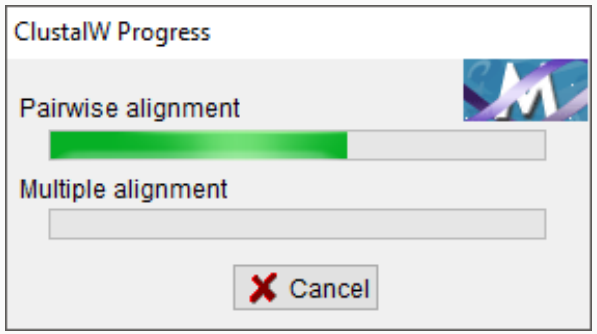
- Pokazuje się taki widok:



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

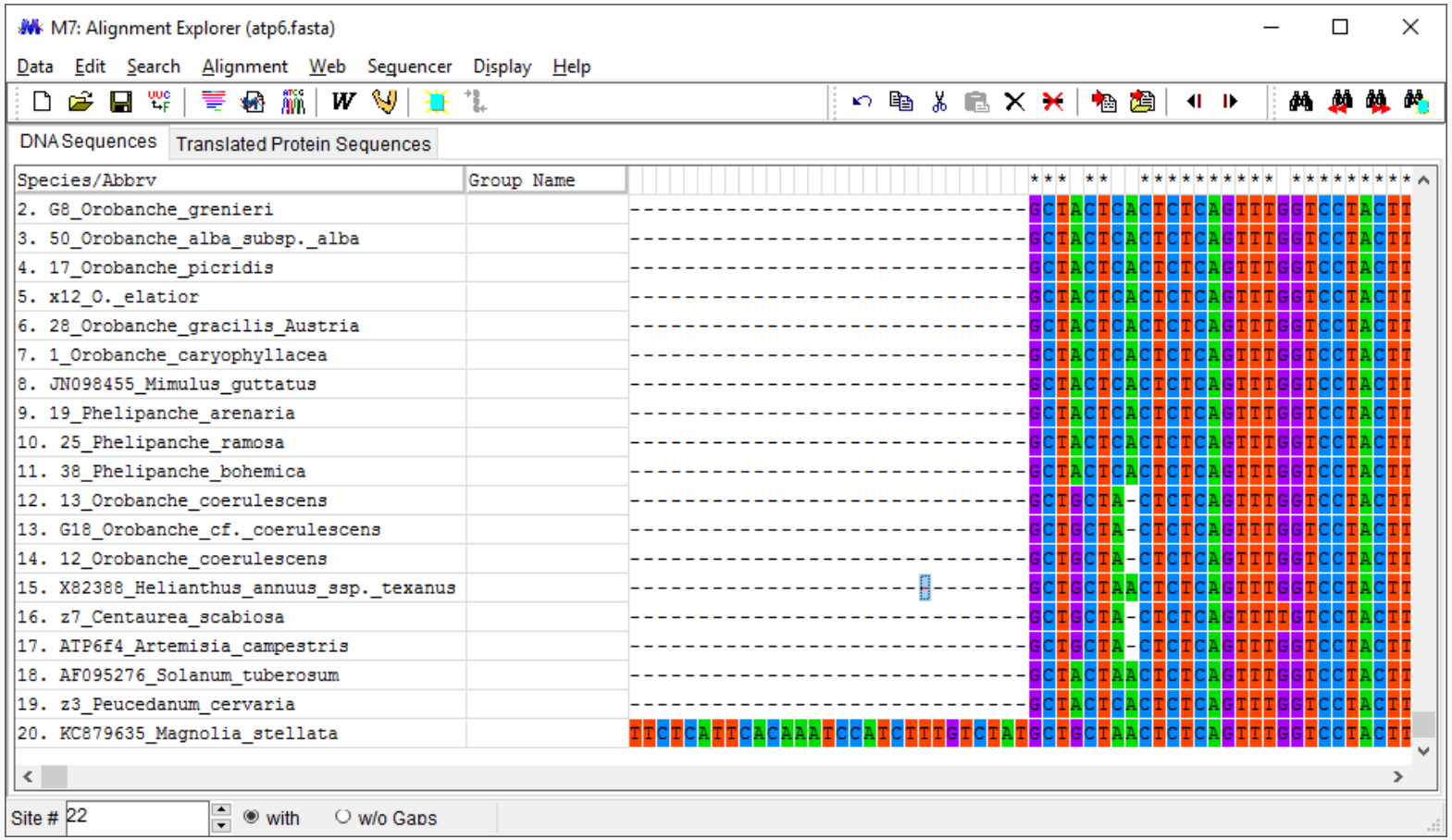
Dopasowanie c.d.

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



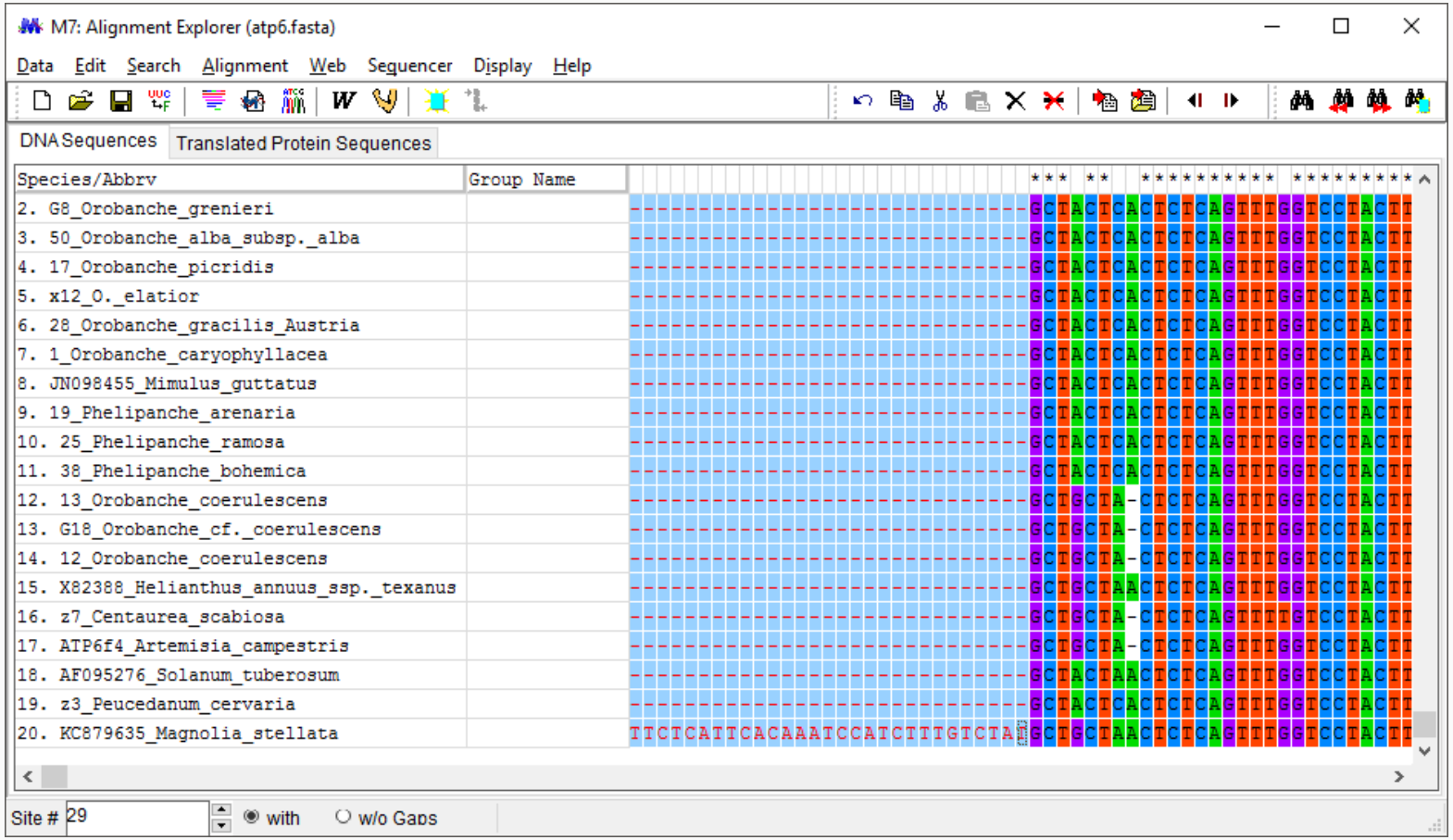
Dopasowana automatycznie sekwencja

- W końcu widać dopasowane sekwencje:



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

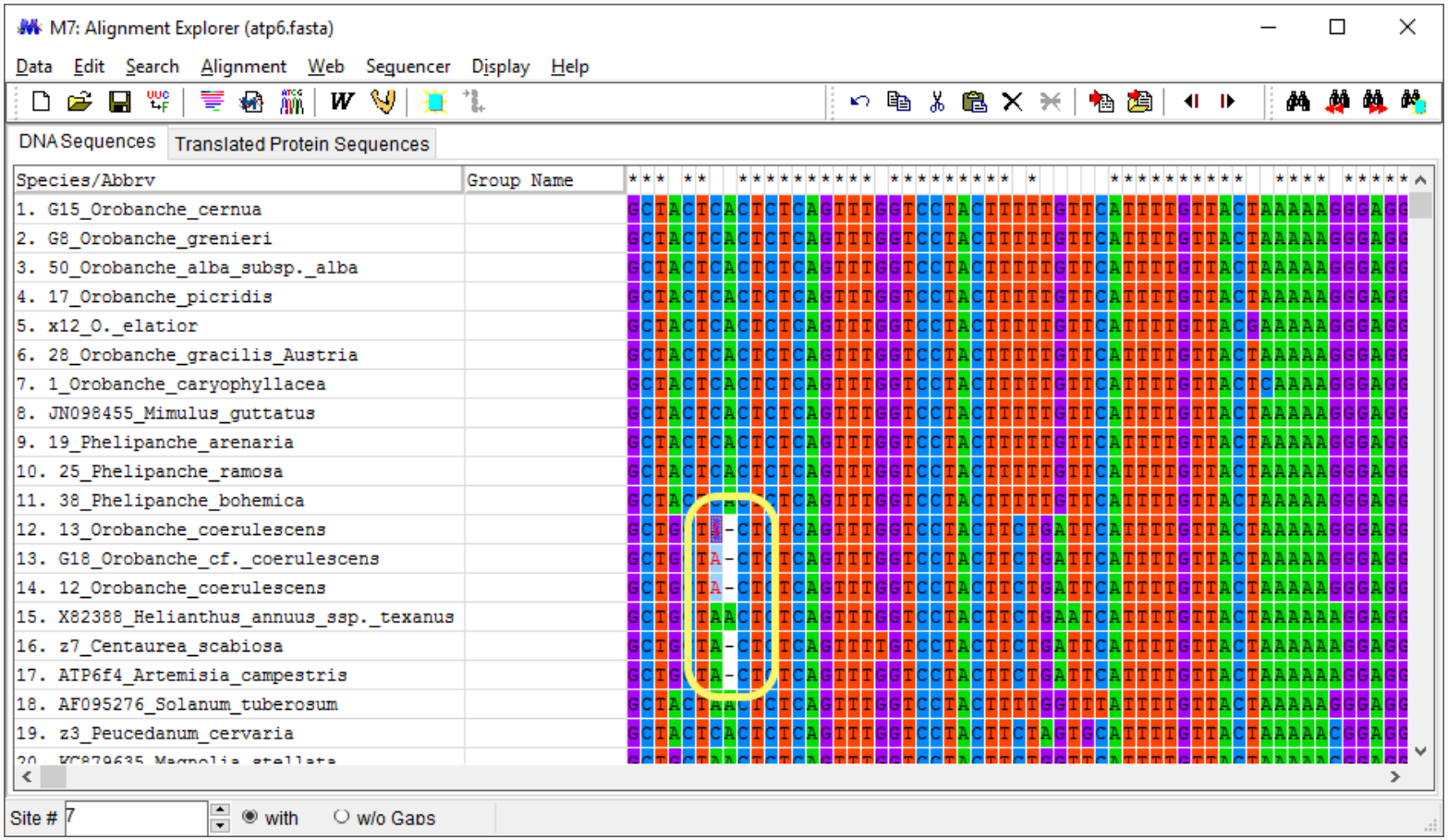
- W tym celu zaznaczamy blok sekwencji po lewej, tak by objął część z brakującymi nukleotydami:



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

Poprawki

- Teraz zwróć uwagę na zaznaczony na żółto fragment:



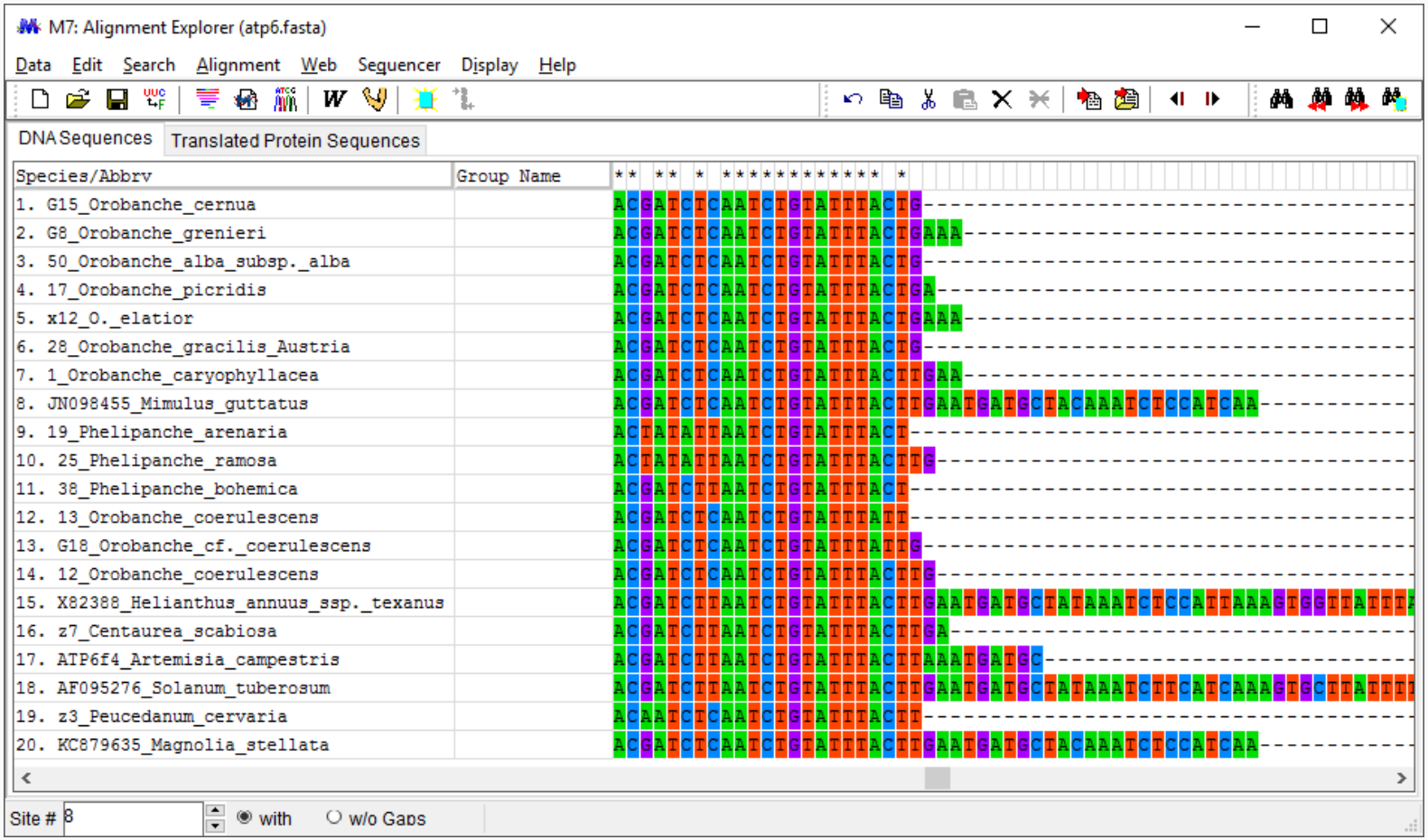
- Wygląda na to, że nukleotydy „A” powinny się zamienić miejscami z indelami (-). Zamień je używając „wytnij” i „wklej”.

Poprawki

- 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 przyjrzeć się ponownie surowym plikom z sekwencjonera.
- Teraz jednak pozostawimy sekwencje takie jakie są.

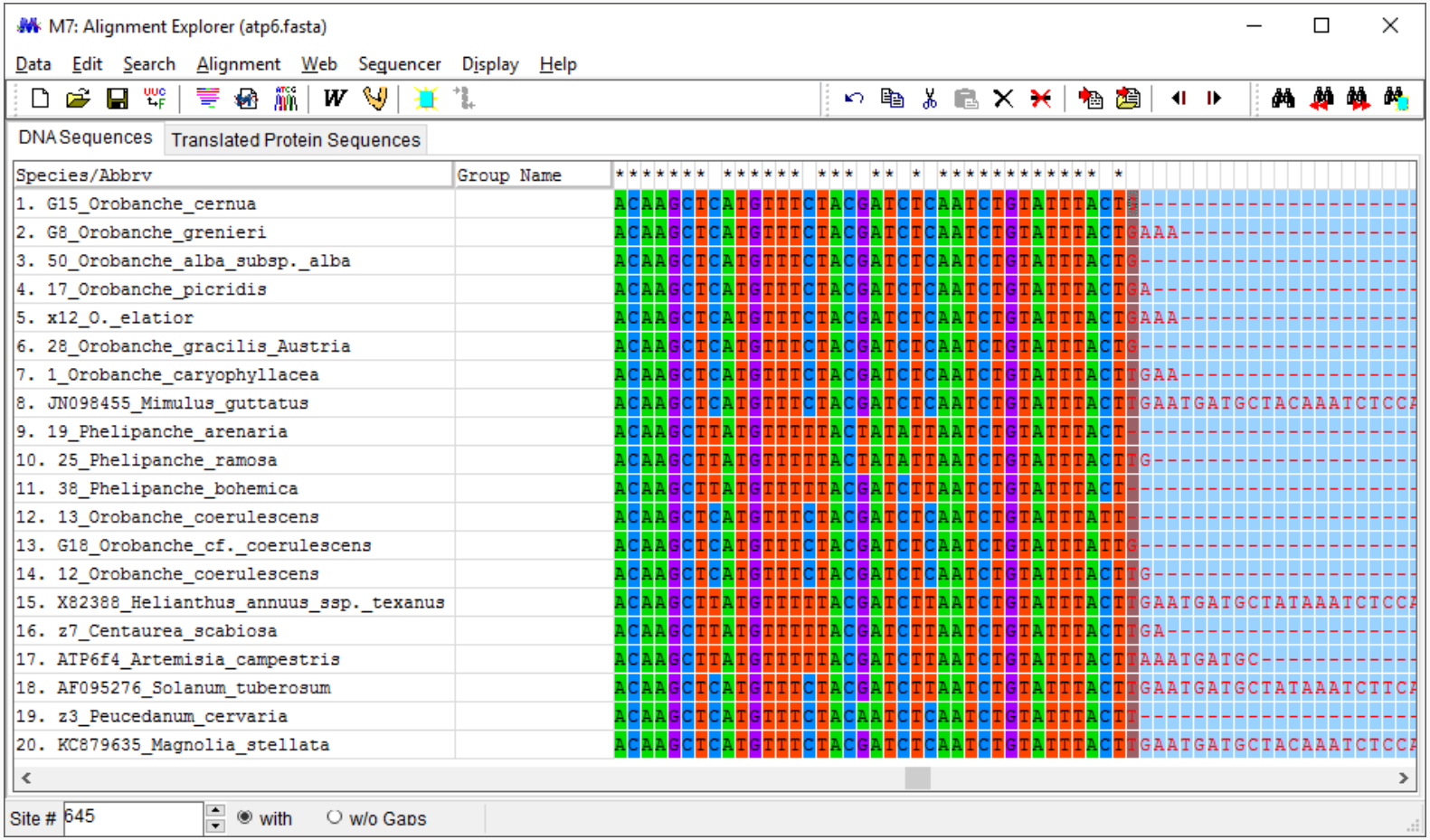
Wyrównywanie końców

- Przejrzyj pozostałą część sekwencji i wykonaj ewentualne poprawki
- Końcowa część sekwencji prezentuje się tak:



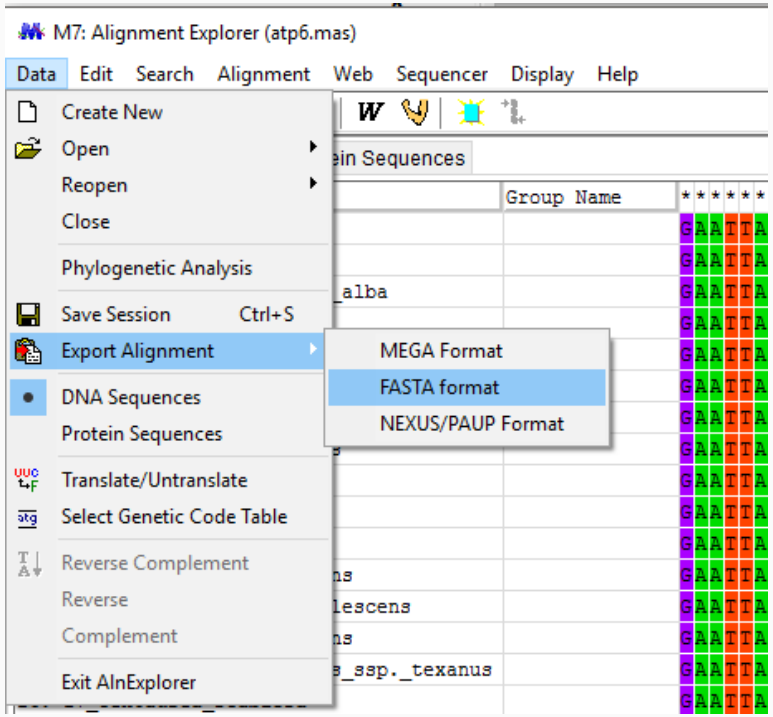
Wyrównywanie końców c.d.

- Trzeba je też wyrównać:



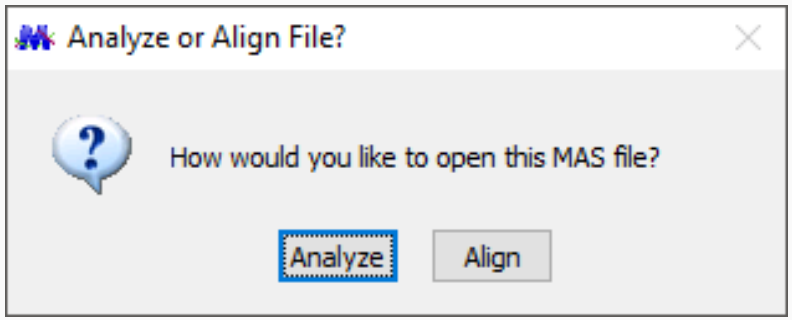
Zapisanie pliku FASTA

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



Zapis i ponowne otwarcie

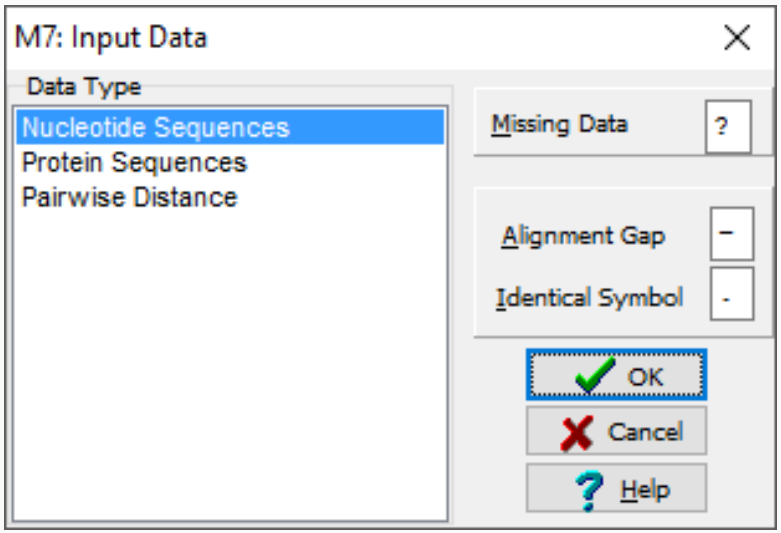
- Zapisujemy sesję w formacie MAS
- Następnie otwieramy go. Pojawia się pytanie:



- Tym razem wybieramy „Analyze”.

Zapis i ponowne otwarcie c.d.

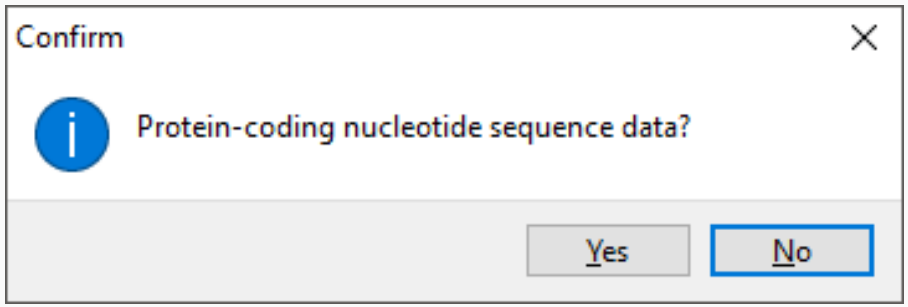
- Pojawia się okienko:



- Wybieramy „Nucleoide sequence”.

Zapis i ponowne otwarcie c.d.

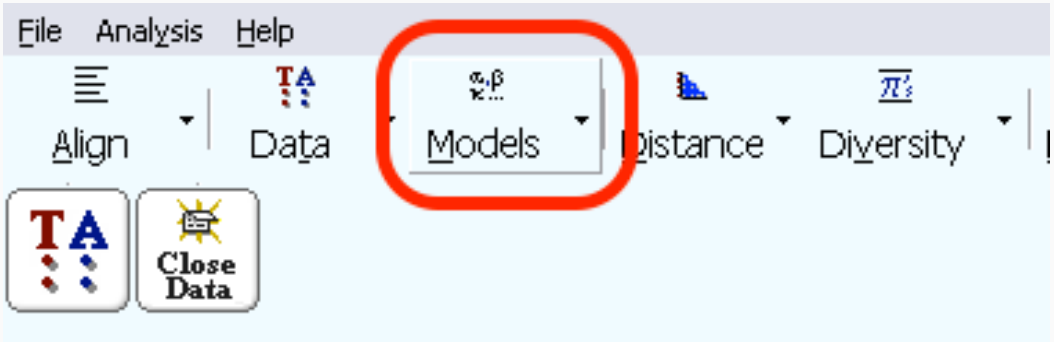
- Teraz pada pytanie czy to sekwencja kodująca:



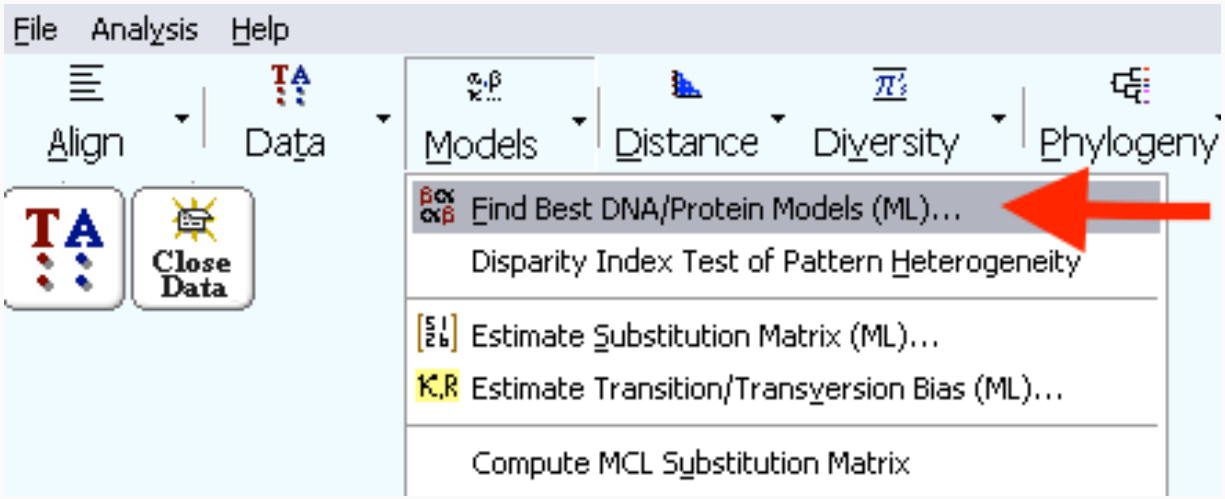
- W zasadzie jest to sekwencja kodująca, ale dla uproszczenia wybieramy opcję „No”.

Dobór Modelu ewolucji molekularnej

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

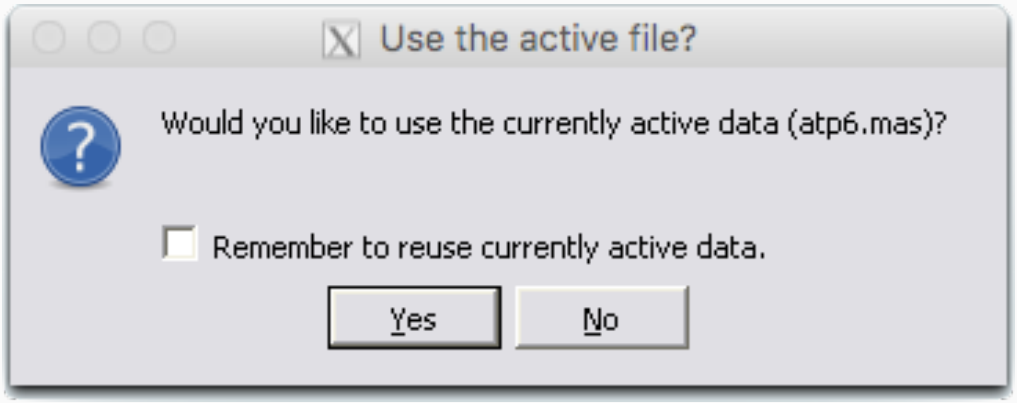


- Następnie wybieramy z rozwijanego menu „Find Best DNA/Protein Models (ML)...”



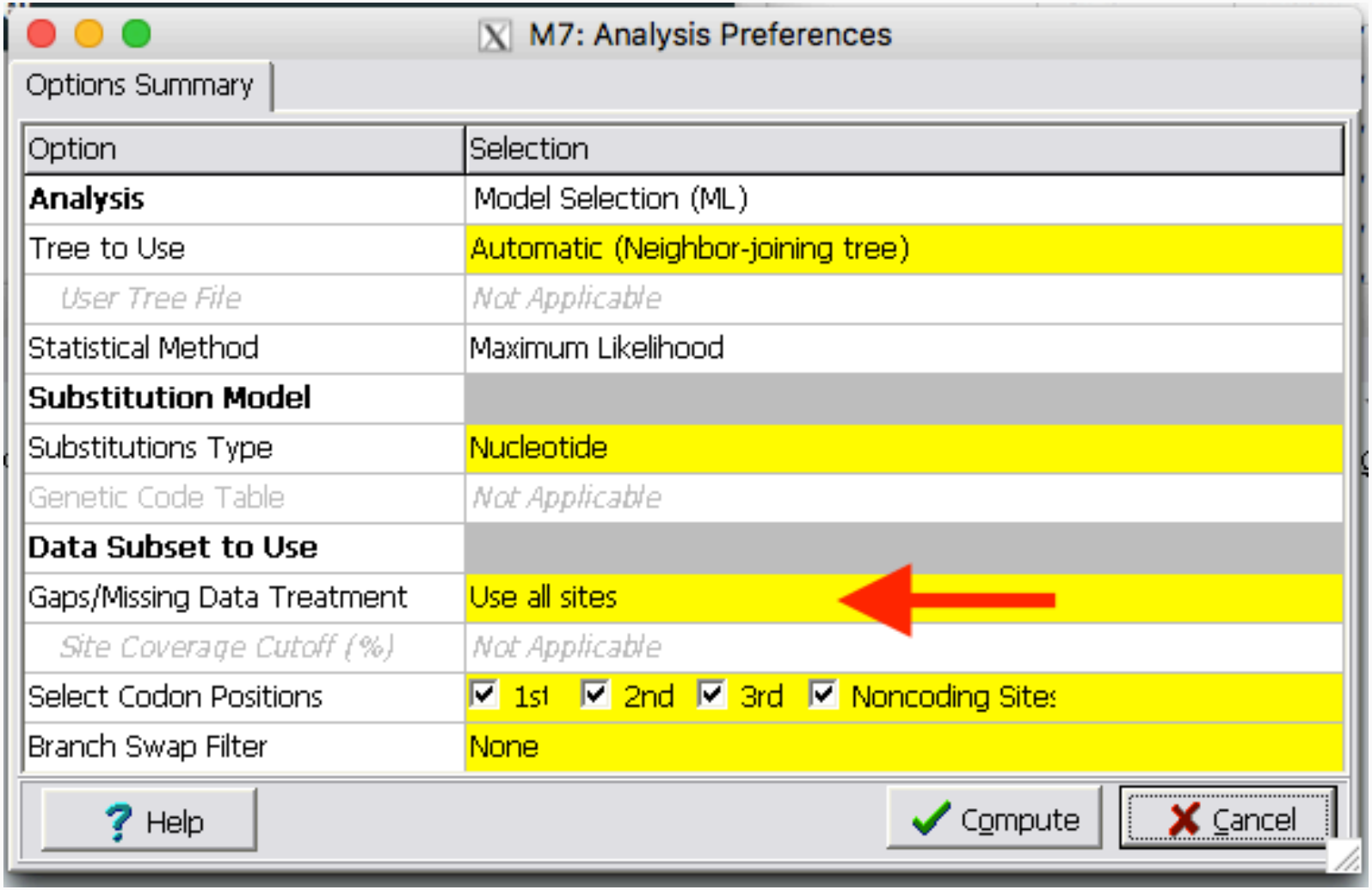
Dobór modelu ewolucji molekularnej

- Zatwierdzamy:



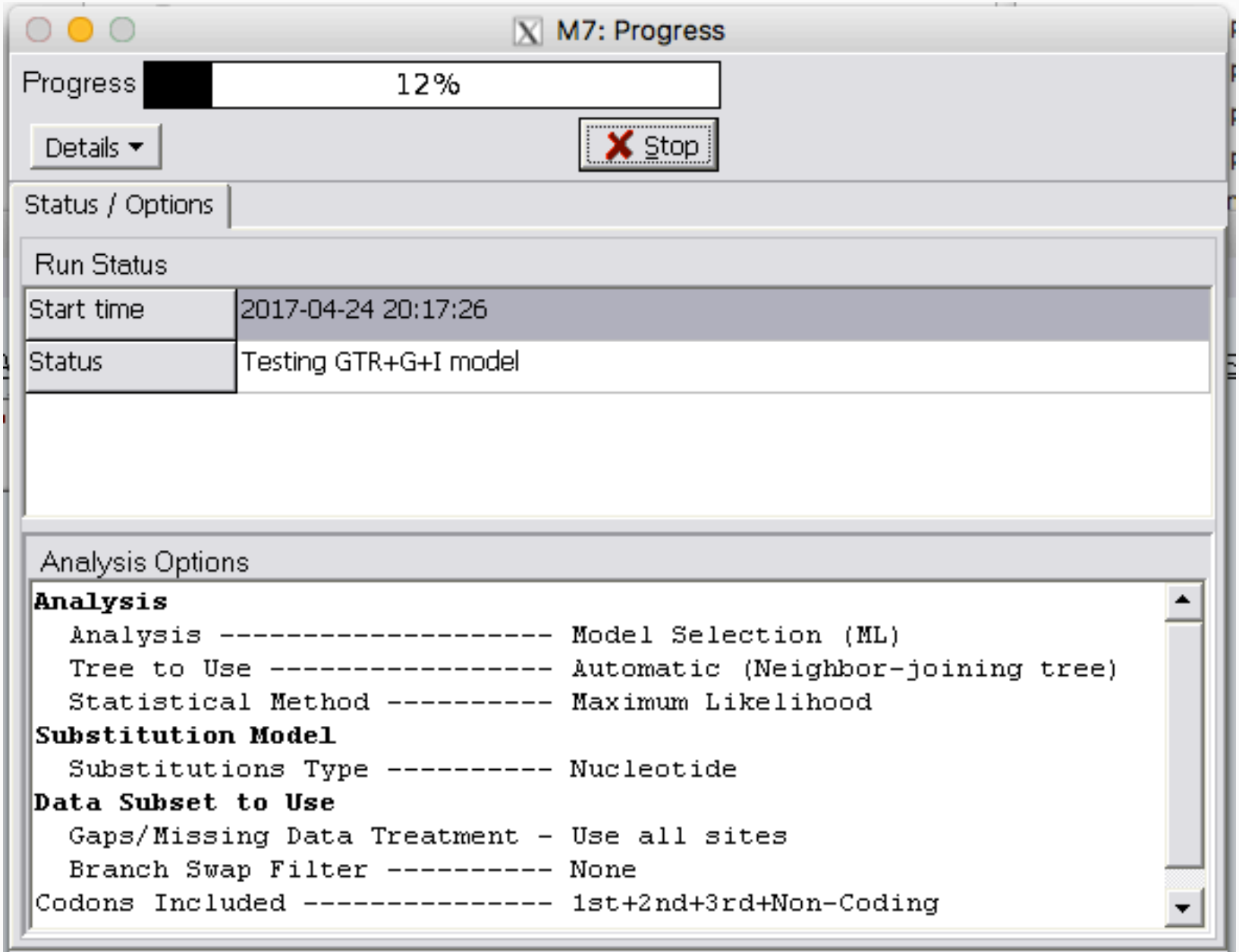
Dobór Modelu ewolucji molekularnej

- 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:



Wybór modelu ewolucji molekularnej

- W końcu uzyskujemy wynik:

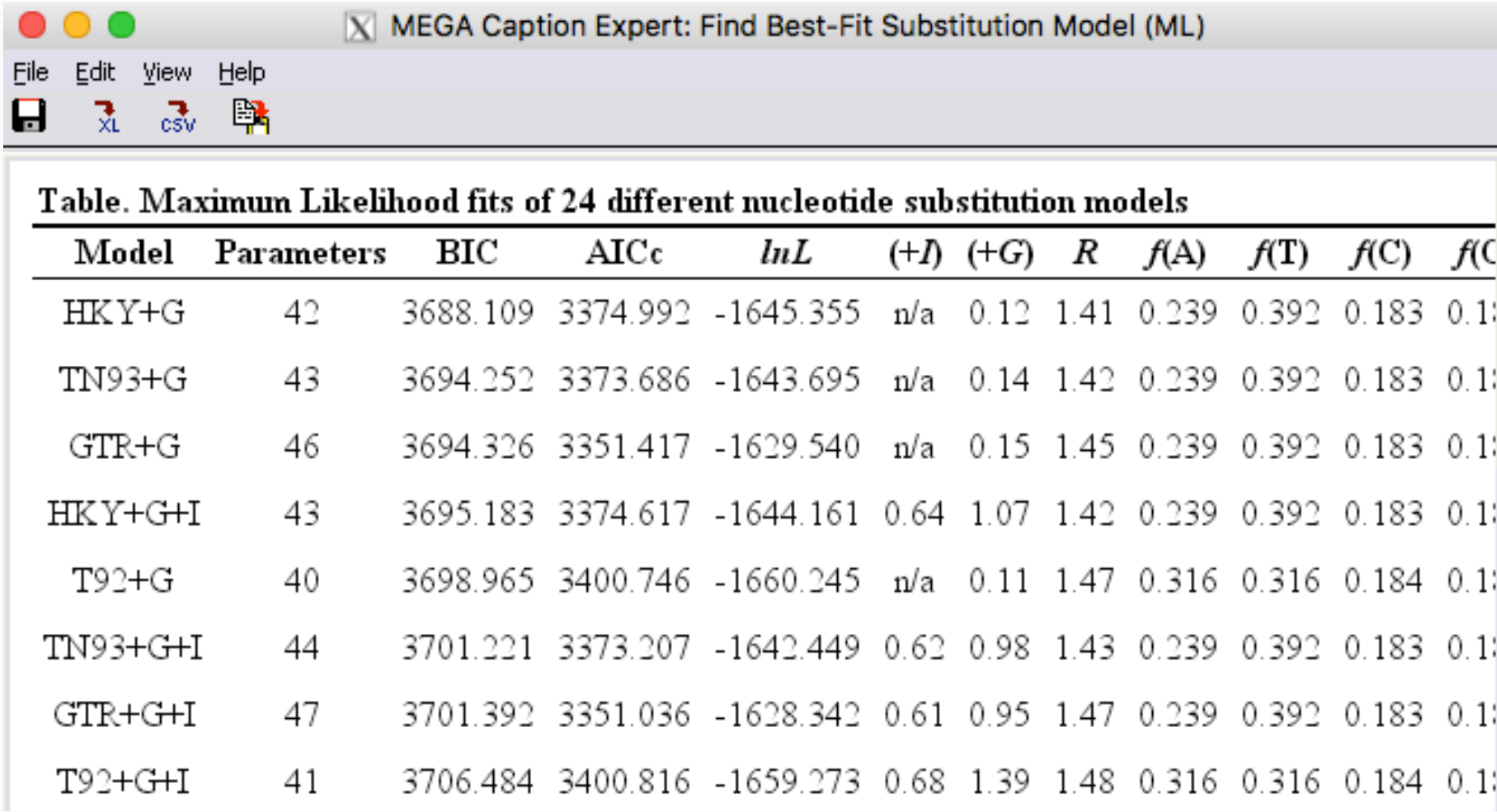


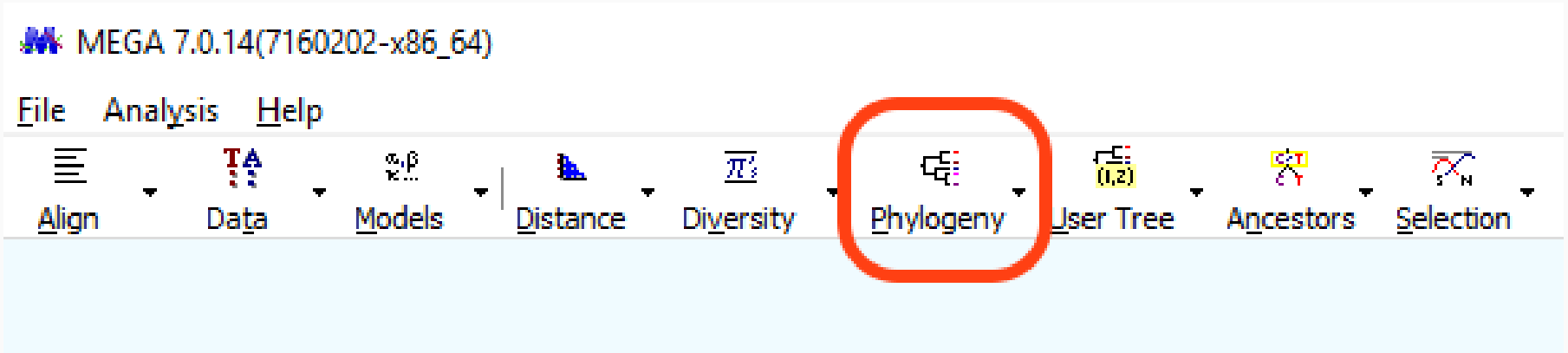
Table. Maximum Likelihood fits of 24 different nucleotide substitution models

Model	Parameters	BIC	AICc	*lnL*	(+*I*)	(+*G*)	*R*	*f*(A)	*f*(T)	*f*(C)	*f*(G
HKY+G	42	3688.109	3374.992	-1645.355	n/a	0.12	1.41	0.239	0.392	0.183	0.1
TN93+G	43	3694.252	3373.686	-1643.695	n/a	0.14	1.42	0.239	0.392	0.183	0.1
GTR+G	46	3694.326	3351.417	-1629.540	n/a	0.15	1.45	0.239	0.392	0.183	0.1
HKY+G+I	43	3695.183	3374.617	-1644.161	0.64	1.07	1.42	0.239	0.392	0.183	0.1
T92+G	40	3698.965	3400.746	-1660.245	n/a	0.11	1.47	0.316	0.316	0.184	0.1
TN93+G+I	44	3701.221	3373.207	-1642.449	0.62	0.98	1.43	0.239	0.392	0.183	0.1
GTR+G+I	47	3701.392	3351.036	-1628.342	0.61	0.95	1.47	0.239	0.392	0.183	0.1
T92+G+I	41	3706.484	3400.816	-1659.273	0.68	1.39	1.48	0.316	0.316	0.184	0.1

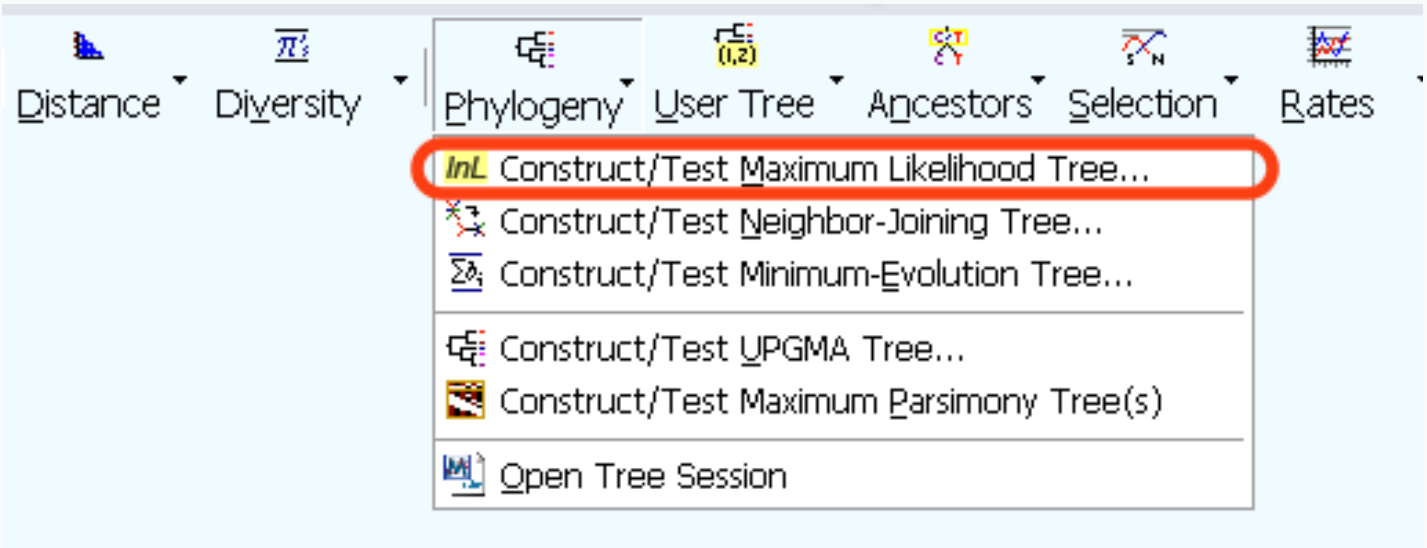
- Do budowy drzewa użyjemy model znajdujący się na pierwszym miejscu na liście, w tym wypadku: **HKY+G**

Budujemy drzewko

- Teraz wybieramy z menu **Phylogeny**

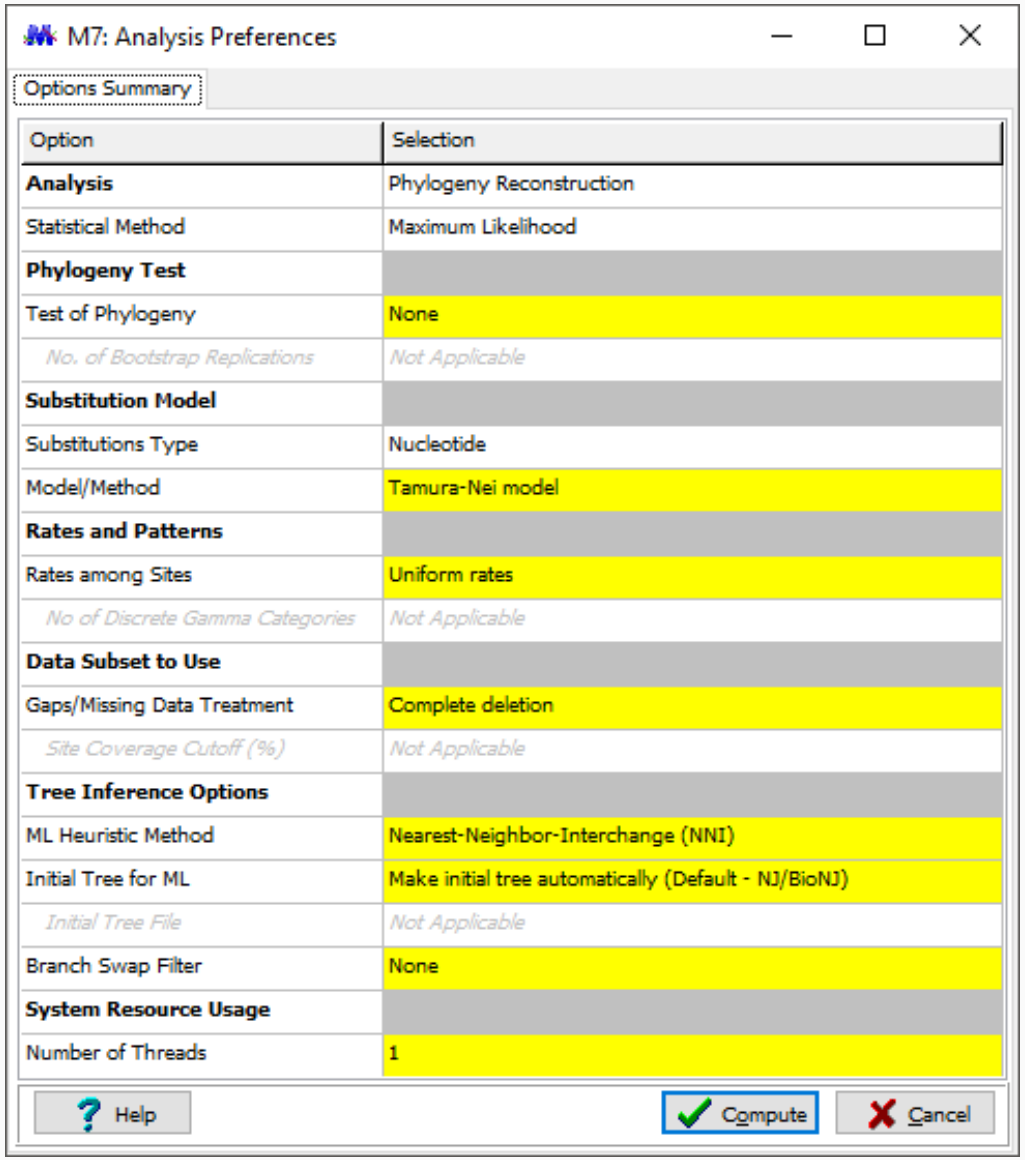


- Z menu wybierz **Maximum Likelihood**



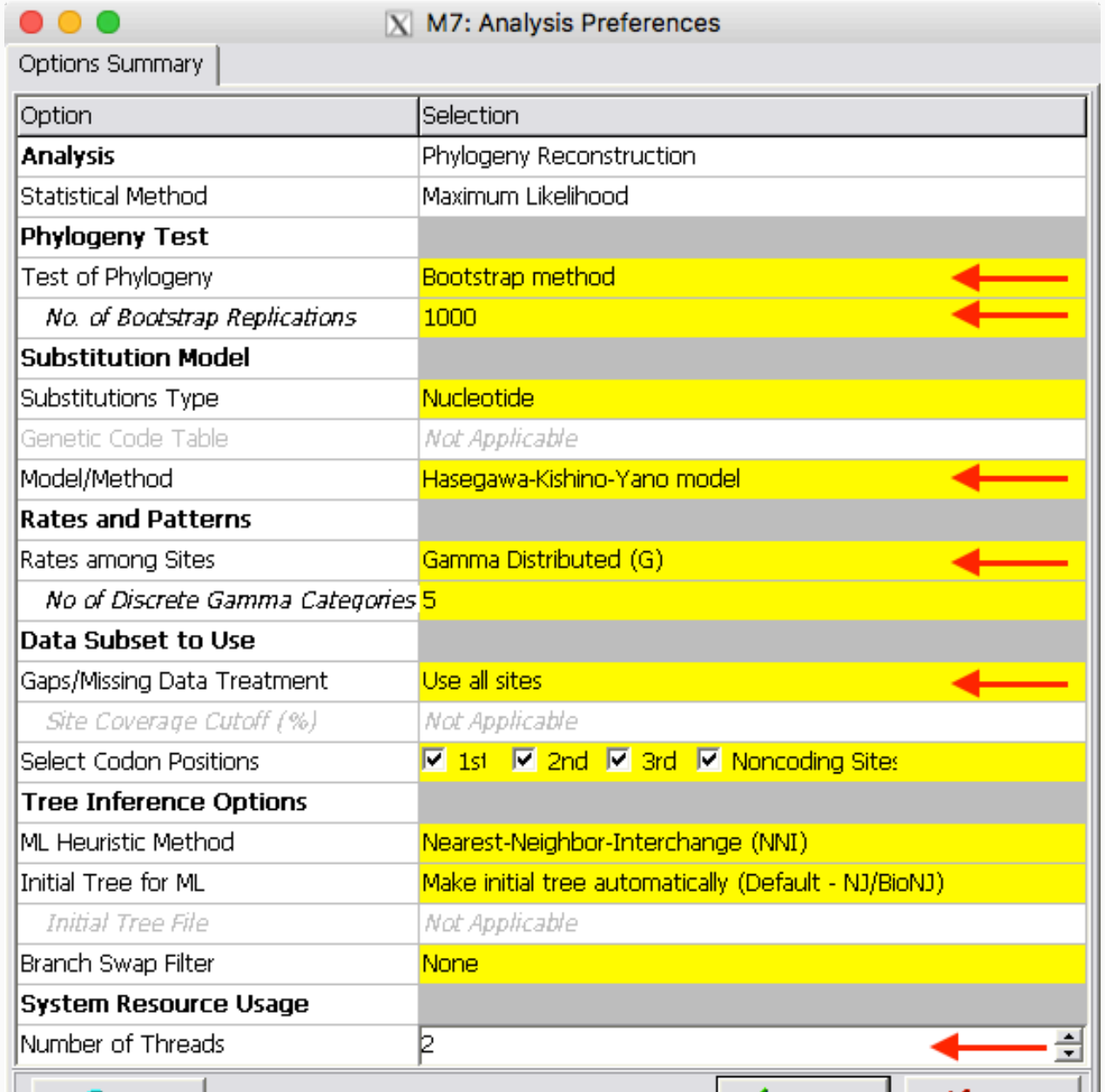
Opcje

- Otwiera się okno z opcjami:



Opcje

- Ustawiamy:

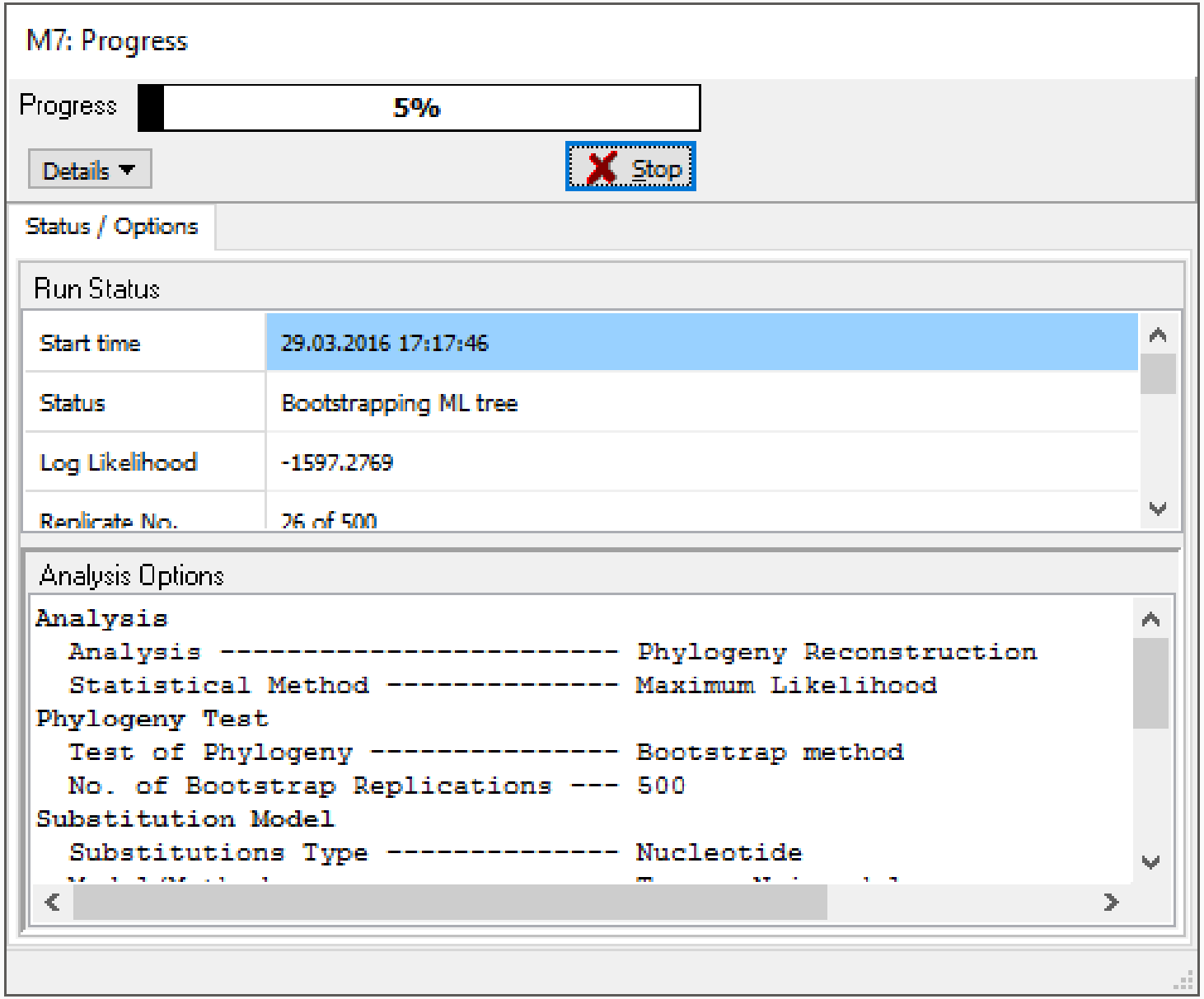


Opcje

- 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.

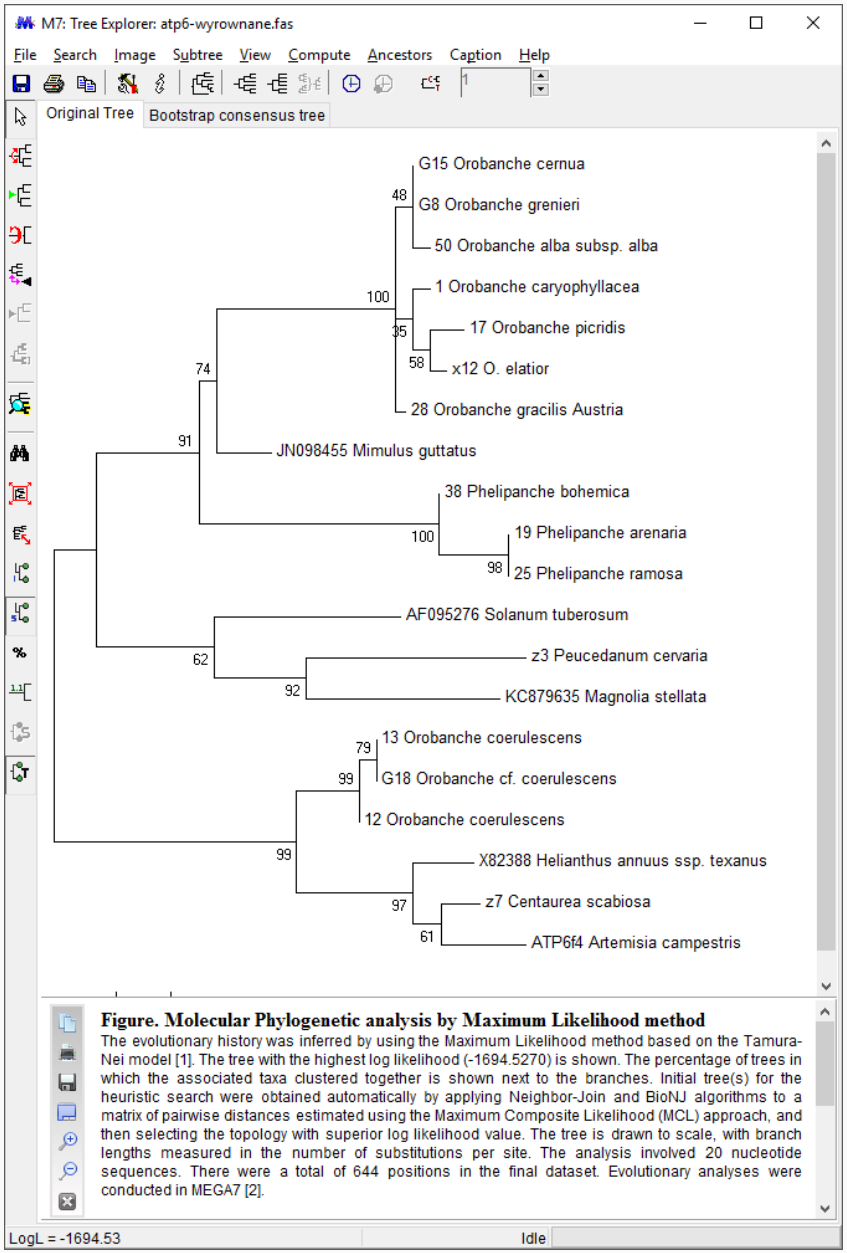
Drzewko w trakcie tworzenia

- Pojawia się okienko z postępem procesu tworzenia drzewa



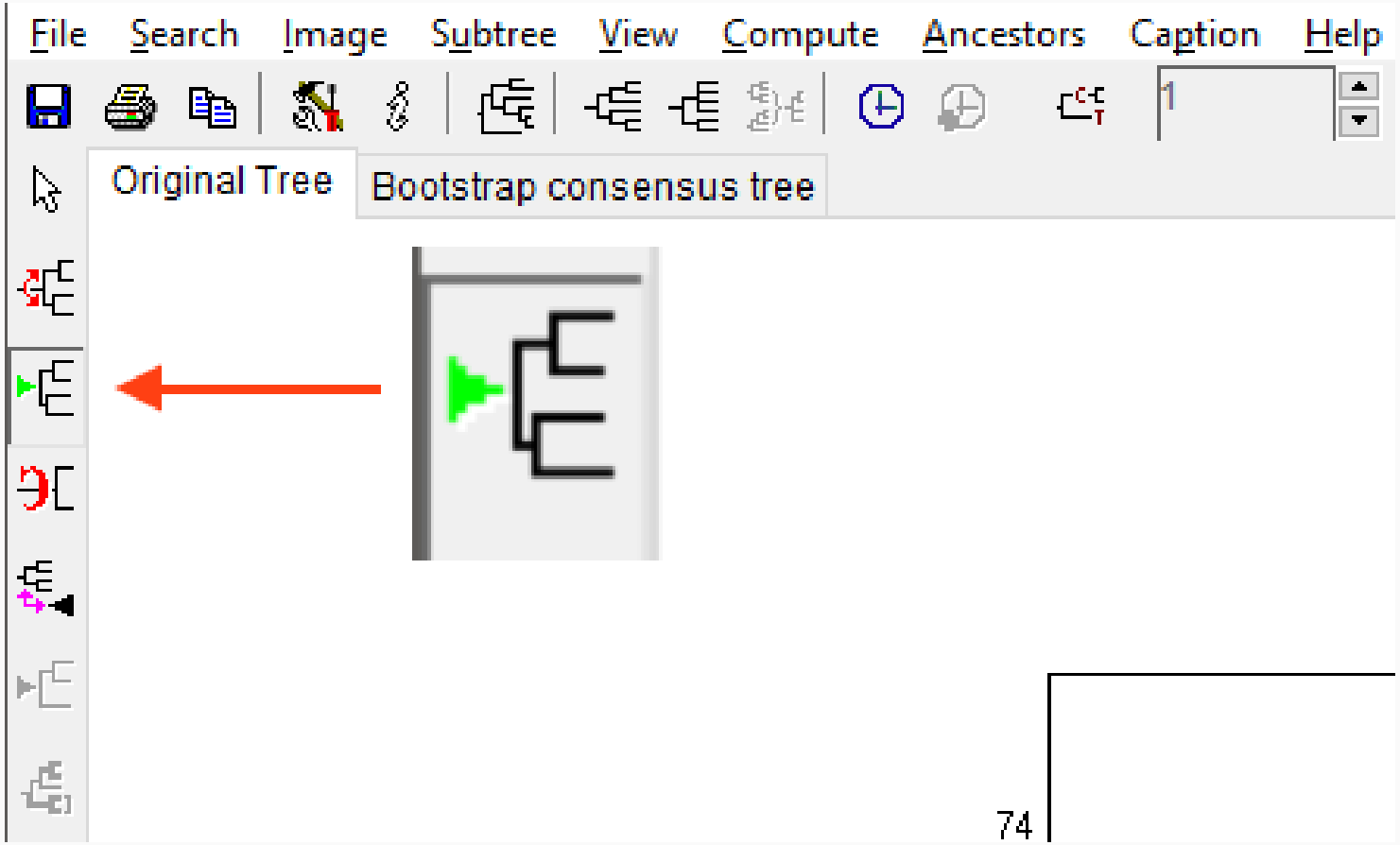
Nieukorzenione drzewo

- W końcu widzimy drzewo, na razie nieukorzenione



Ukorzenianie drzewa

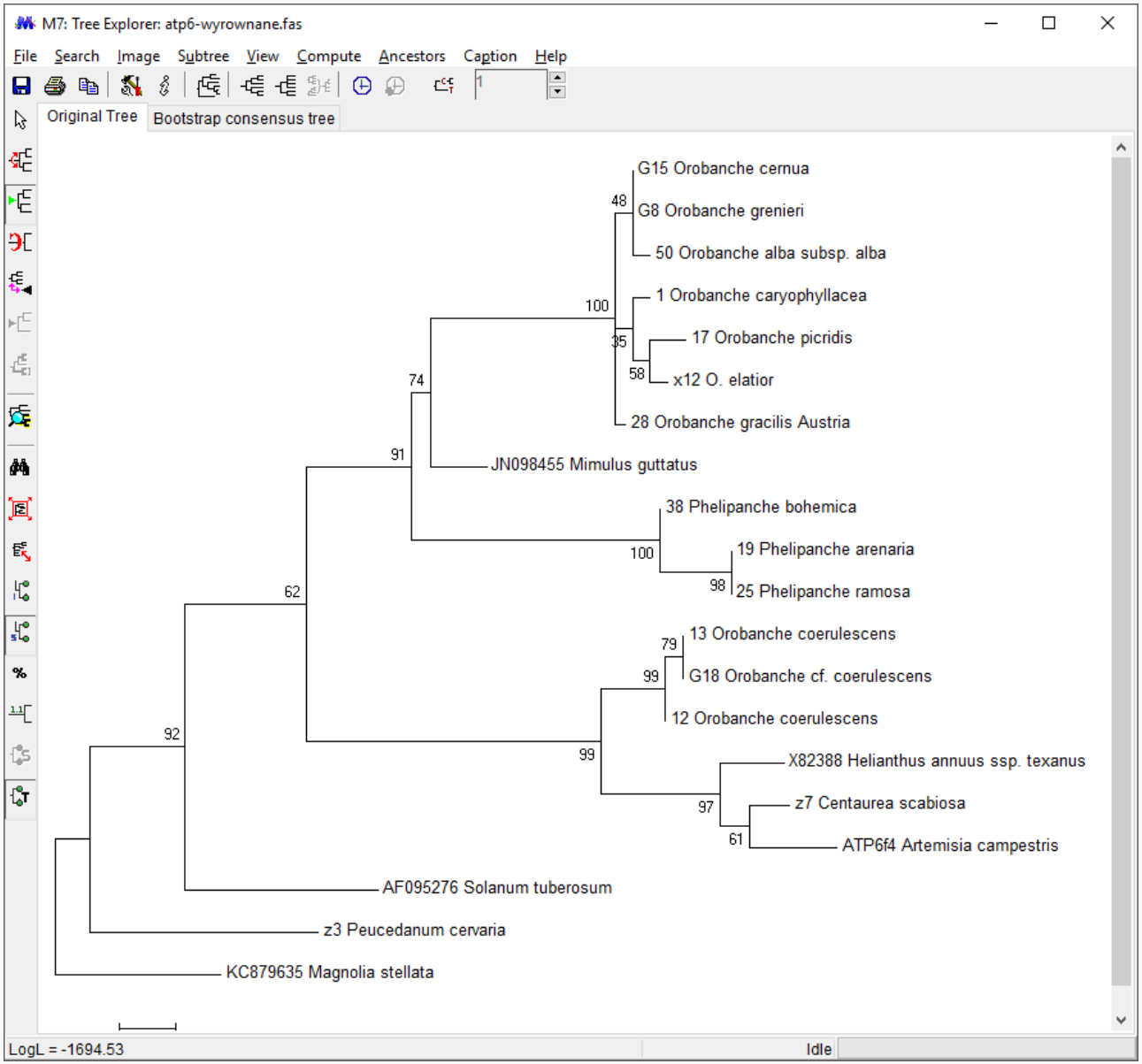
- Teraz należy ukorzenić drzewo
- W tym celu wybieramy odpowiednią ikonę po lewej:



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

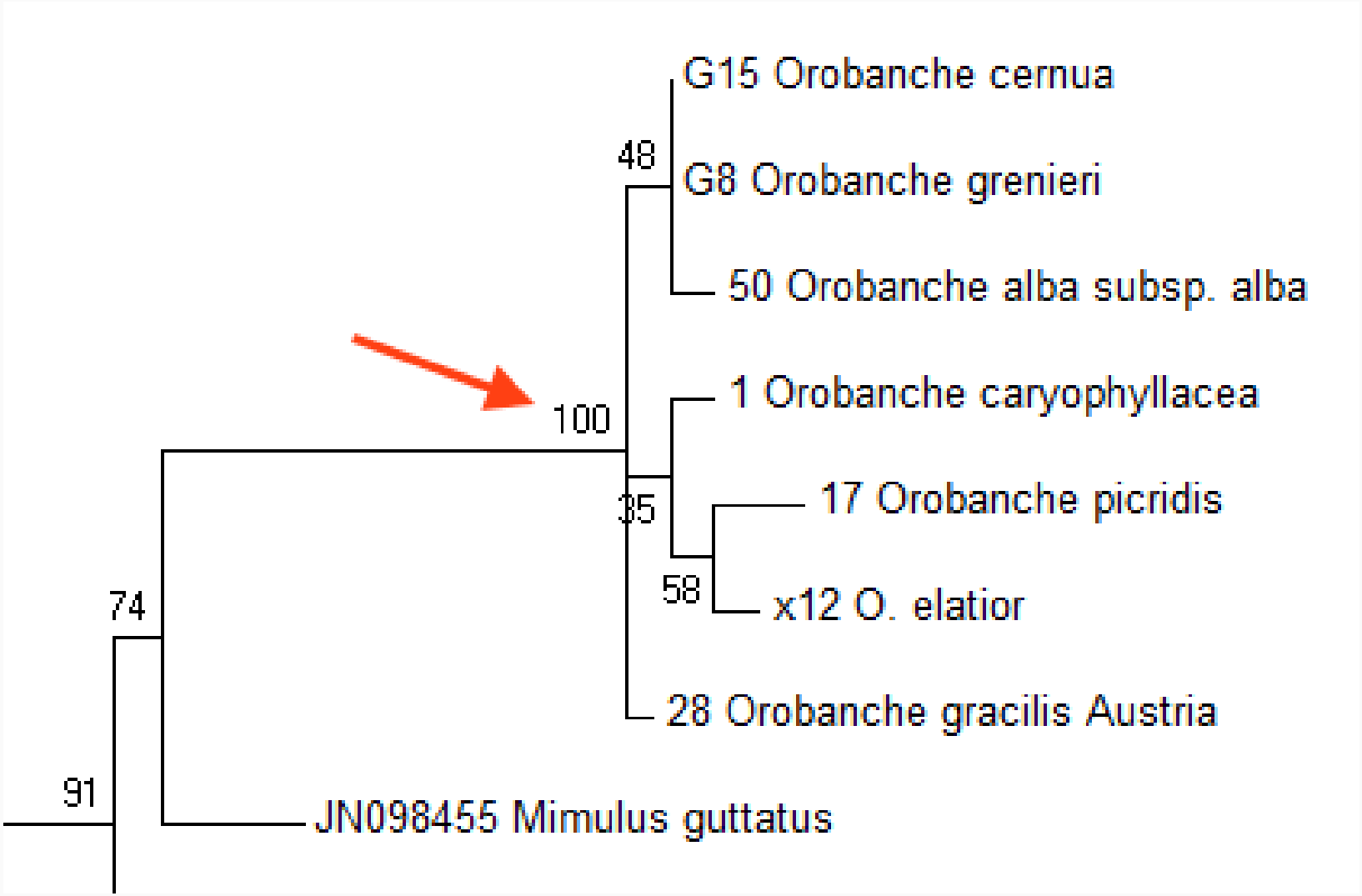
Drzewo ukorzenione

- Teraz drzewo jest ukorzenione:



Wartości bootstrapu

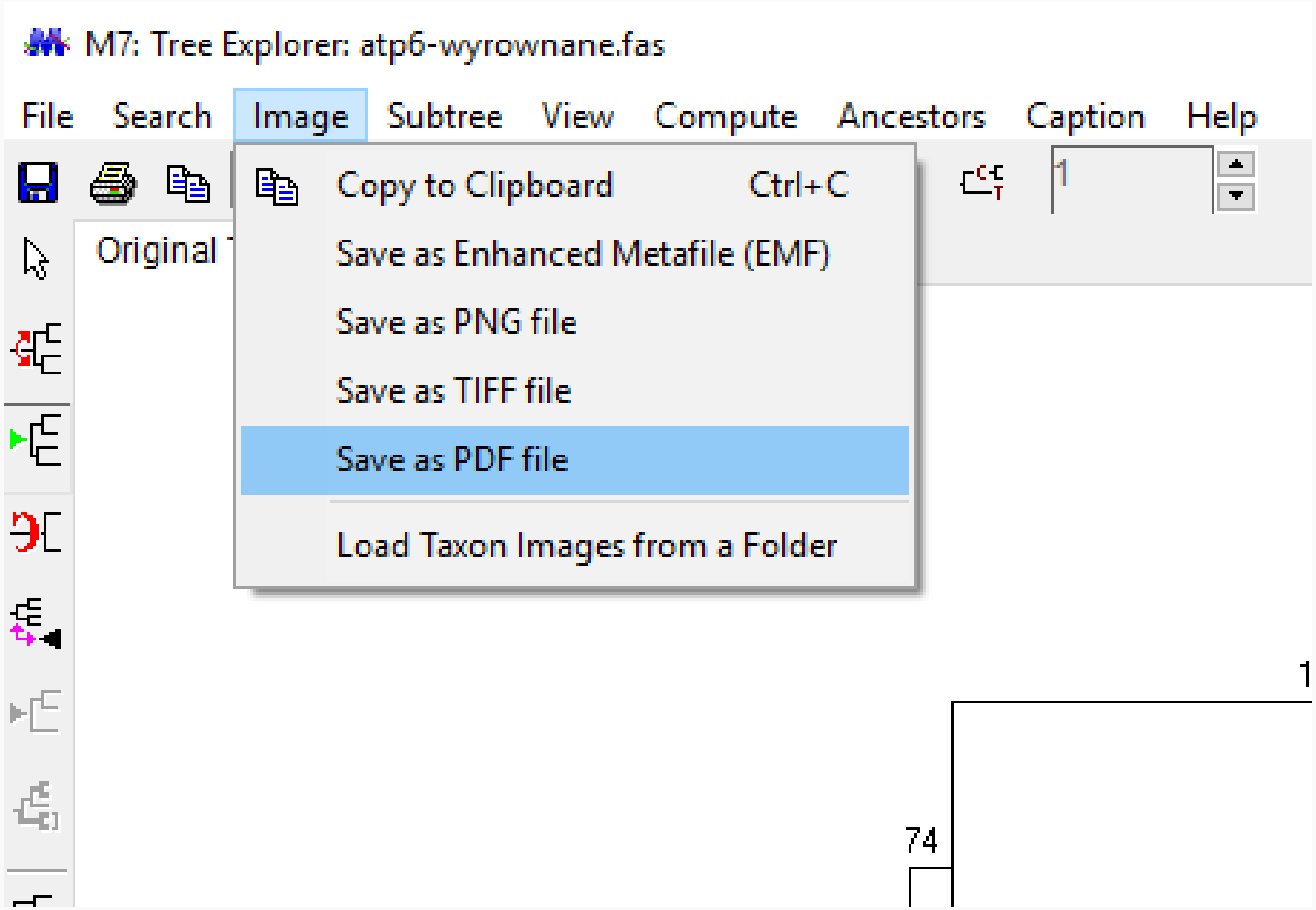
- Na drzewie widoczne są wartości bootstrapu



- Im wyższa wartość bootstrapu tym większa wiarygodność węzła.

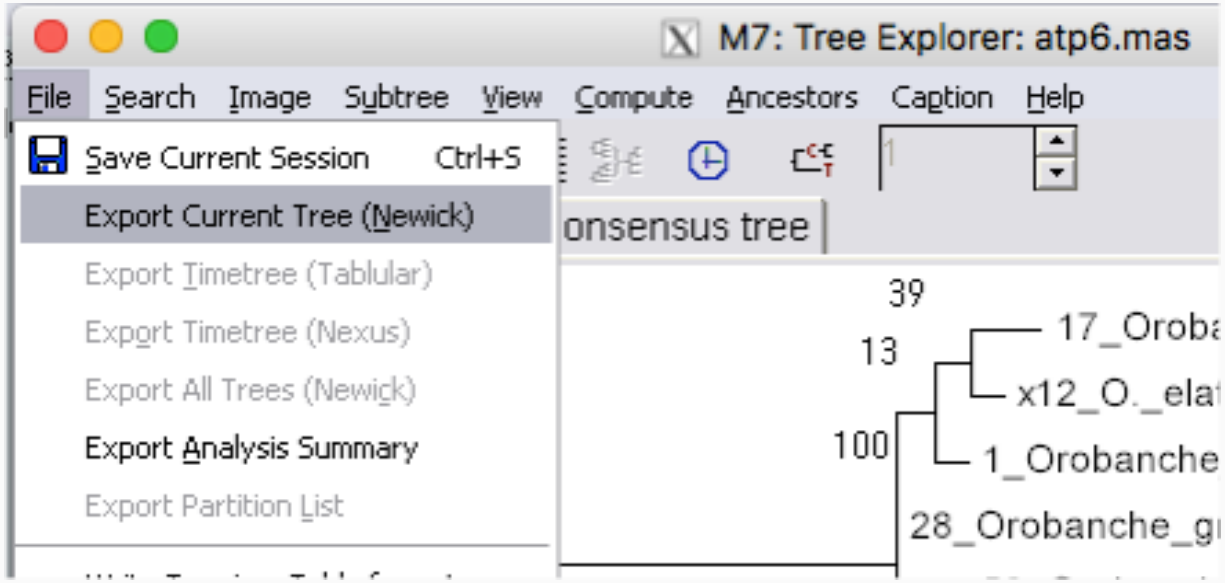
Zapisanie drzewa jako w pliku pdf

- Plik można zapisać np. w formacie pdf:



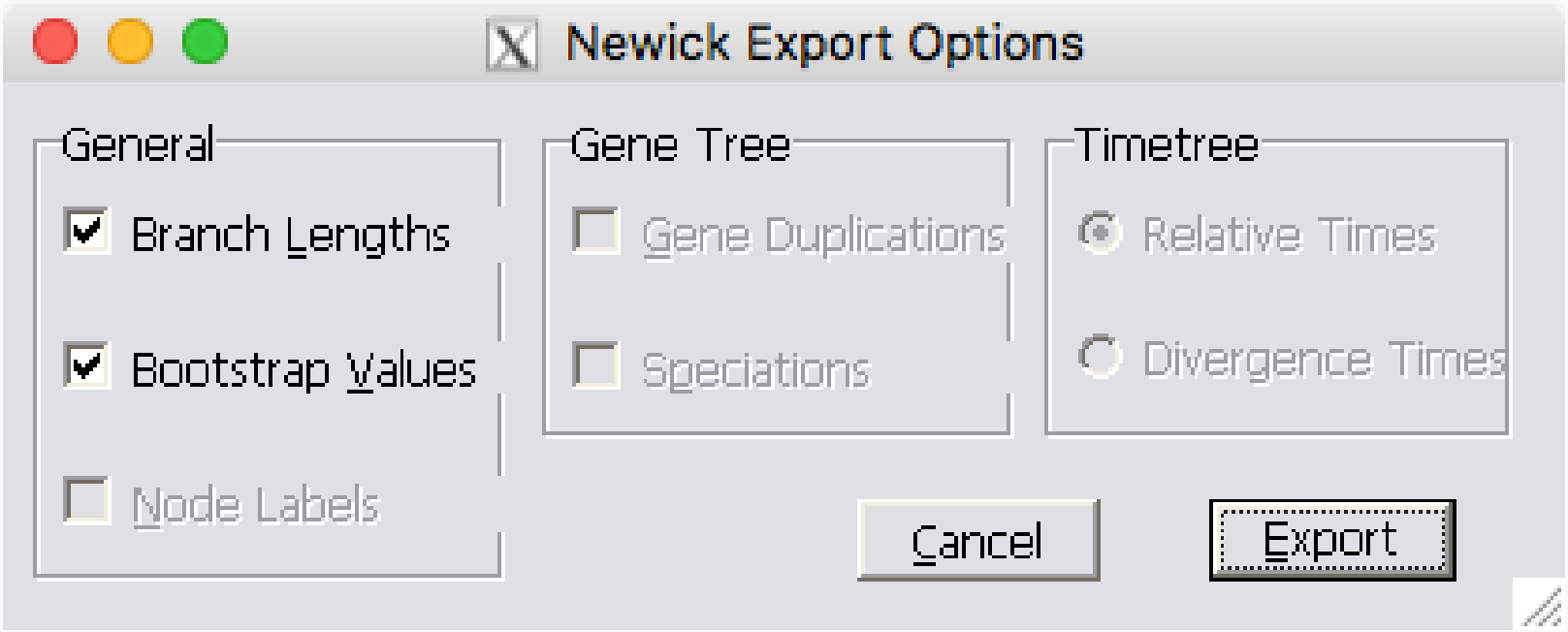
Eksport drzewa do formatu Newick

- 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)”:



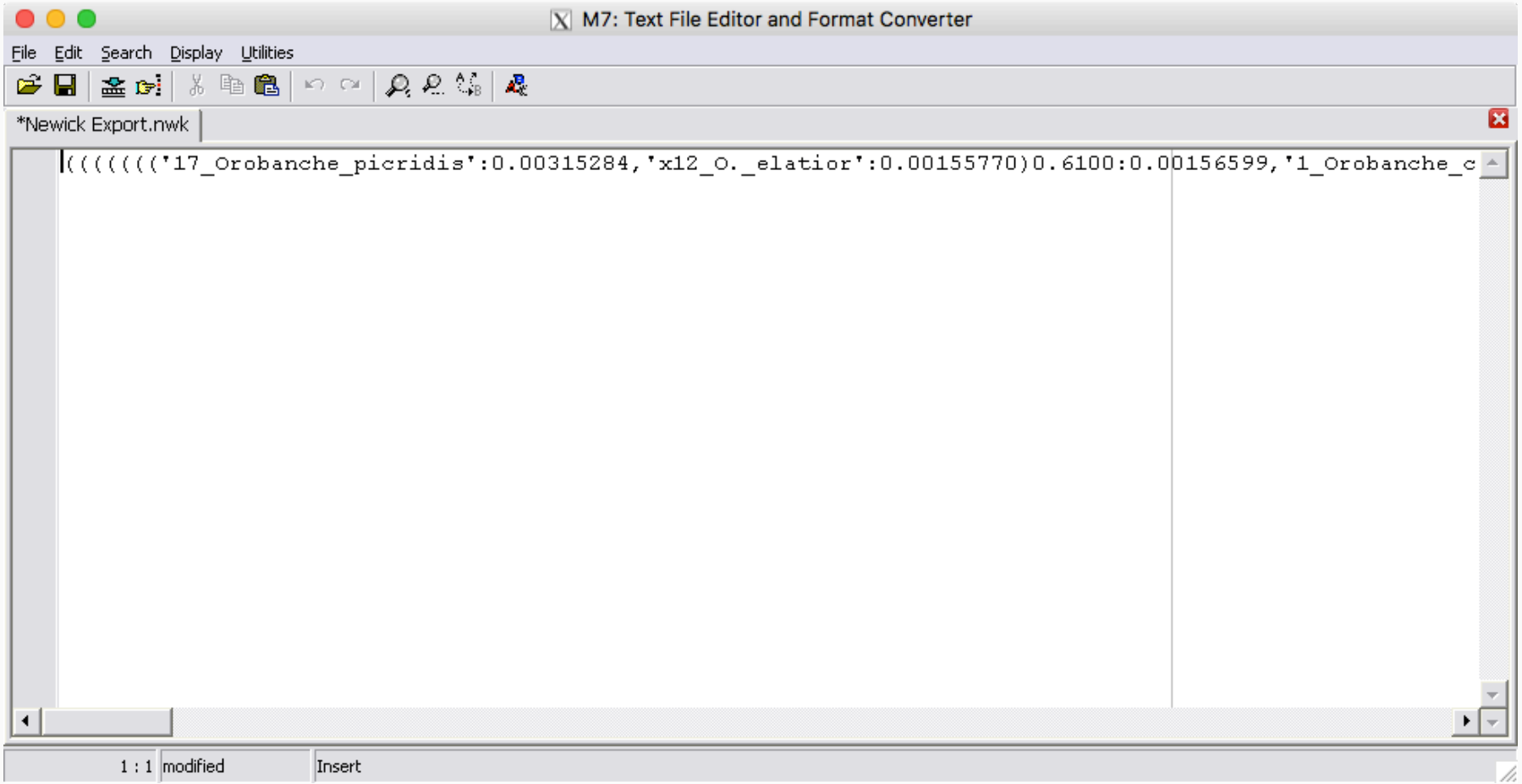
Eksport drzewa do formatu Newick

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



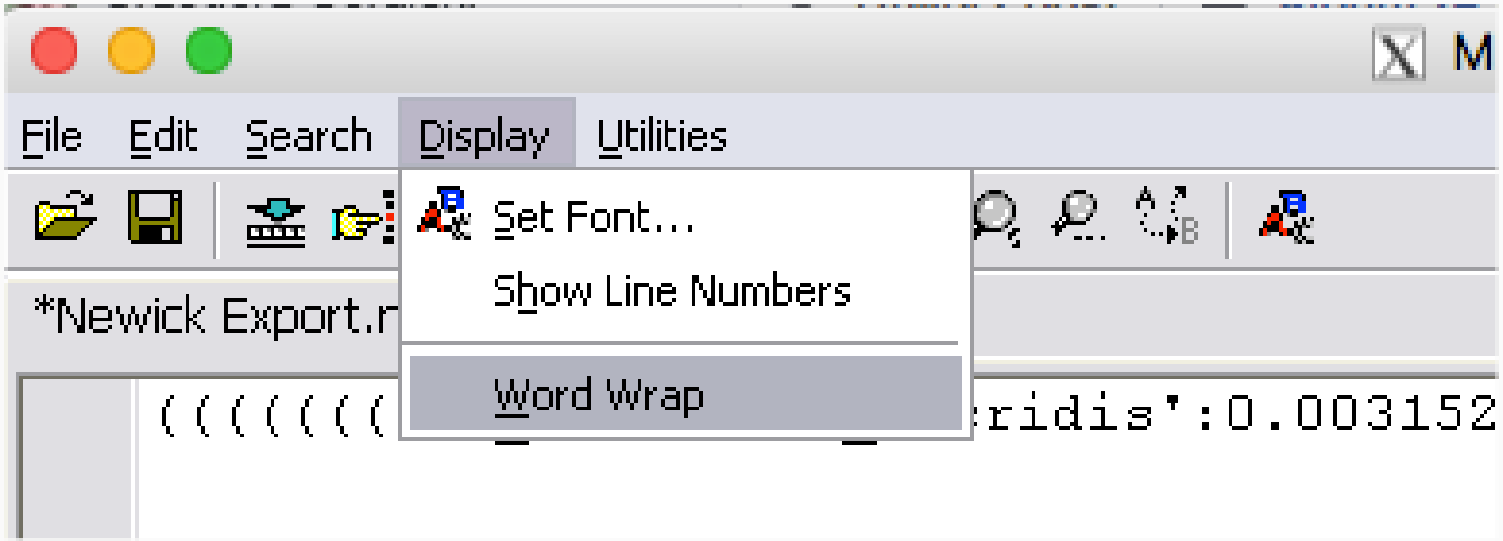
Eksport drzewa do formatu Newick

Pokazuje się wynik:



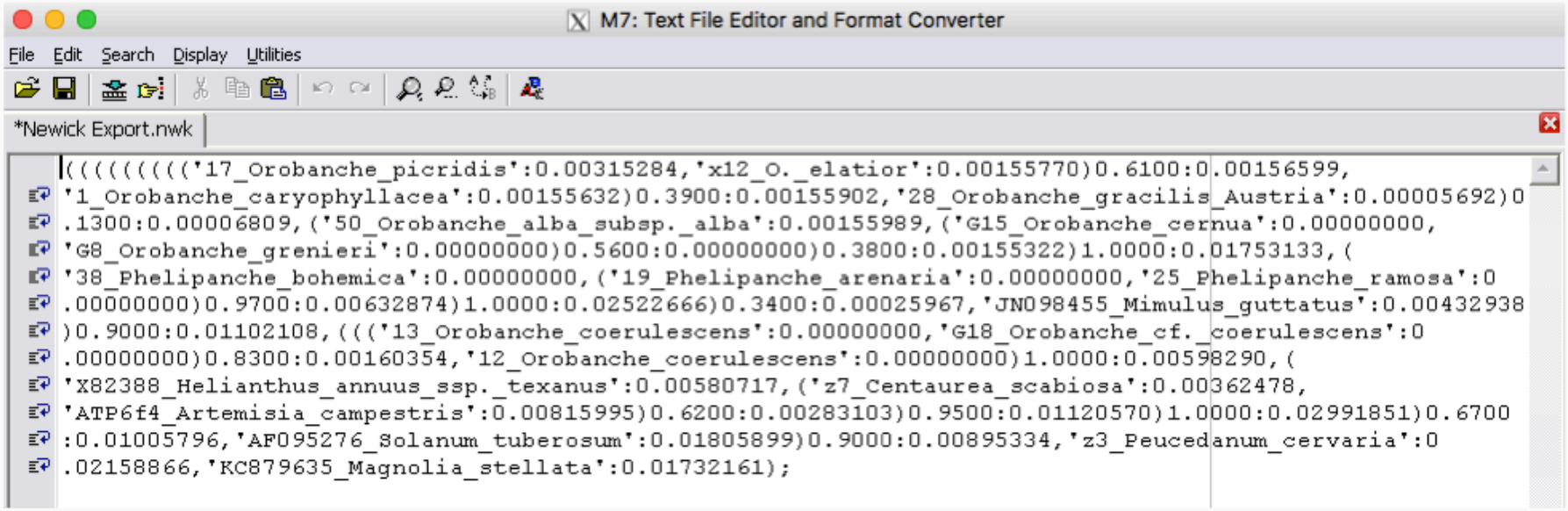
Eksport drzewa do formatu Newick

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



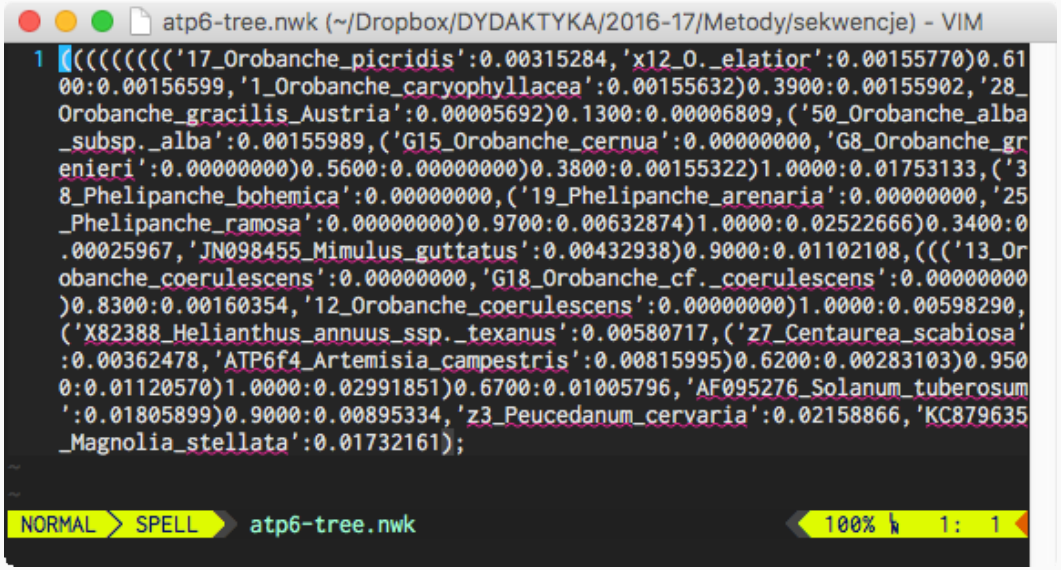
Eksport drzewa do formatu Newick

- Teraz widać strukturę zapisu drzewa bardziej przejrzyście:



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:



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

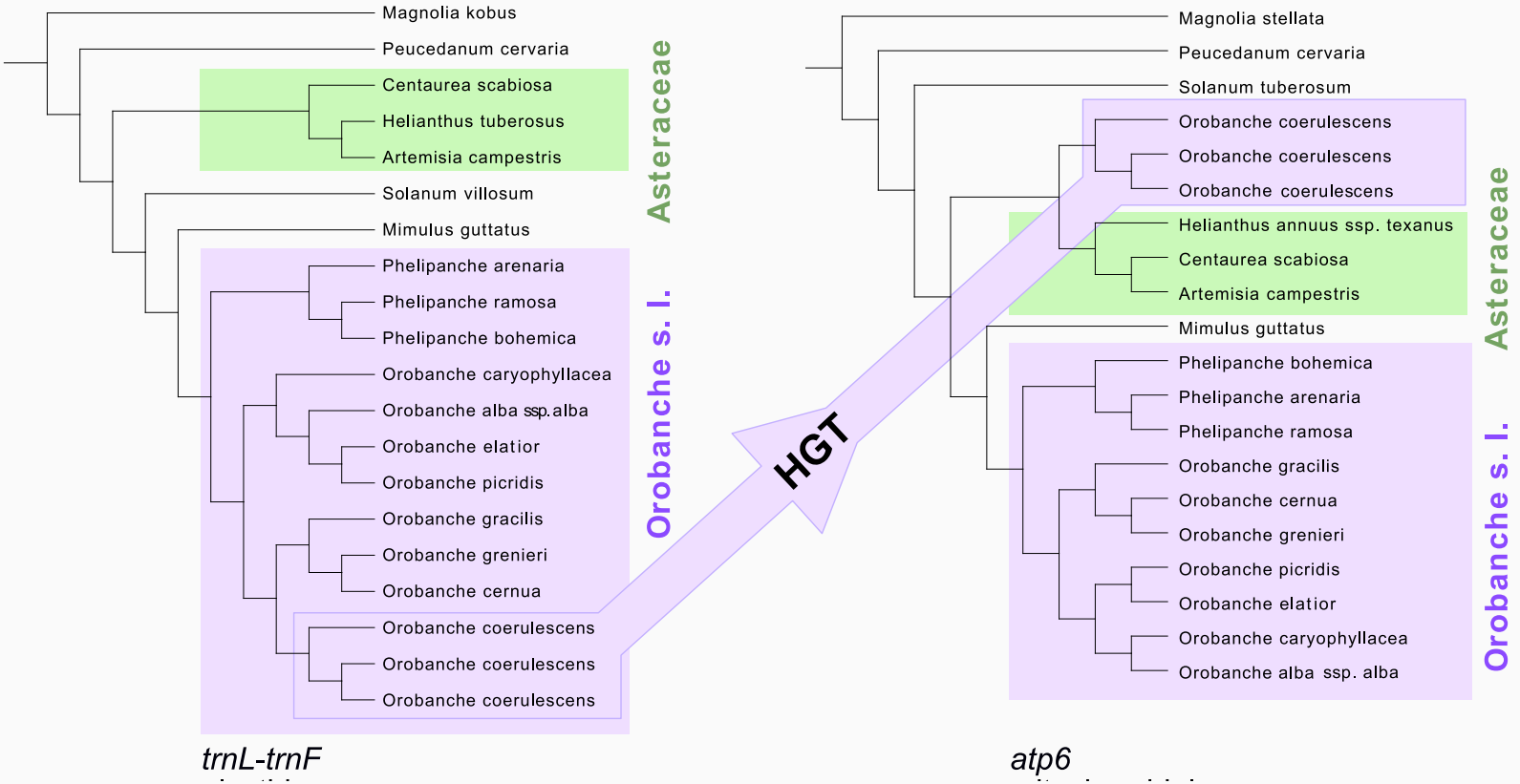
Sekwencje trnL-trnF

- 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?

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



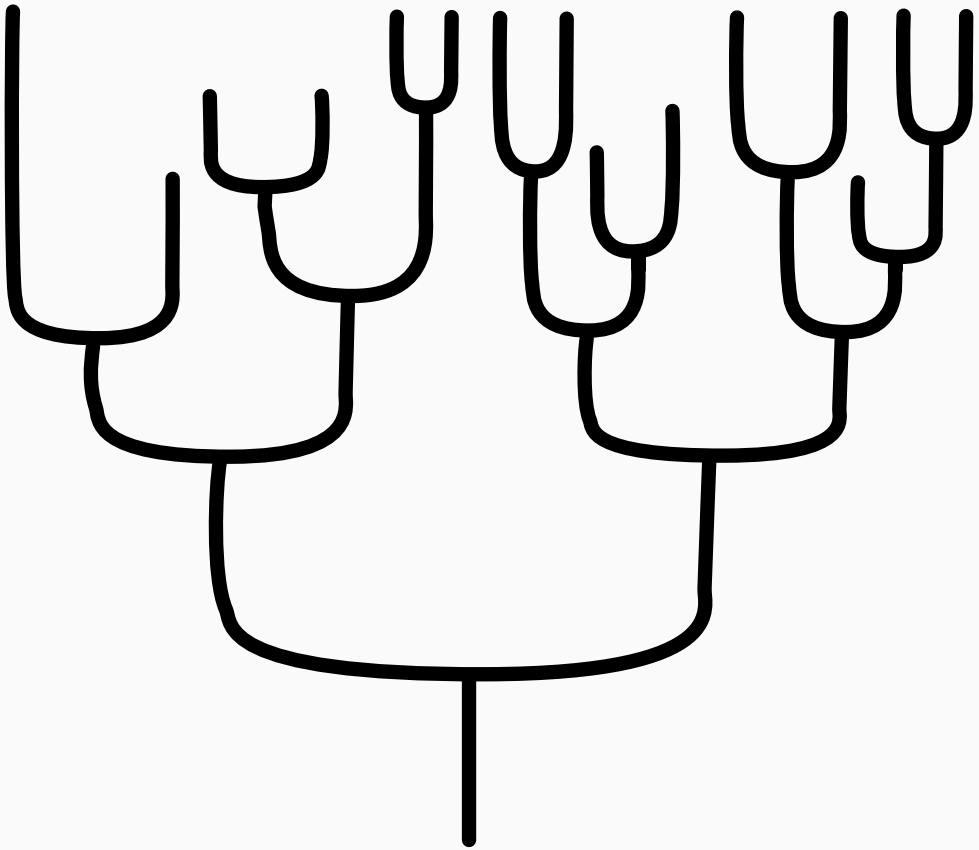
Horyzontalny transfer genów

- 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

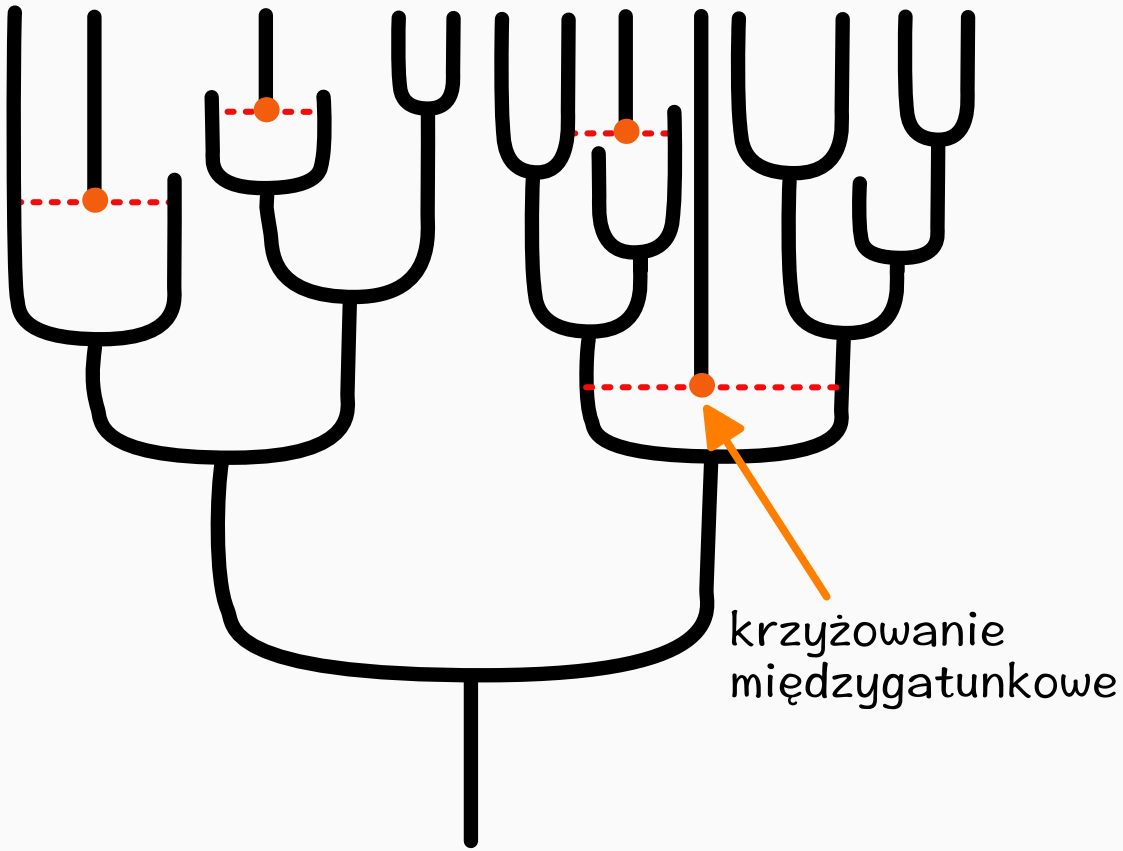
Zaraza żółta (*Orobanche flava*)
w Dolinie Strążyskiej

Drzewo życia



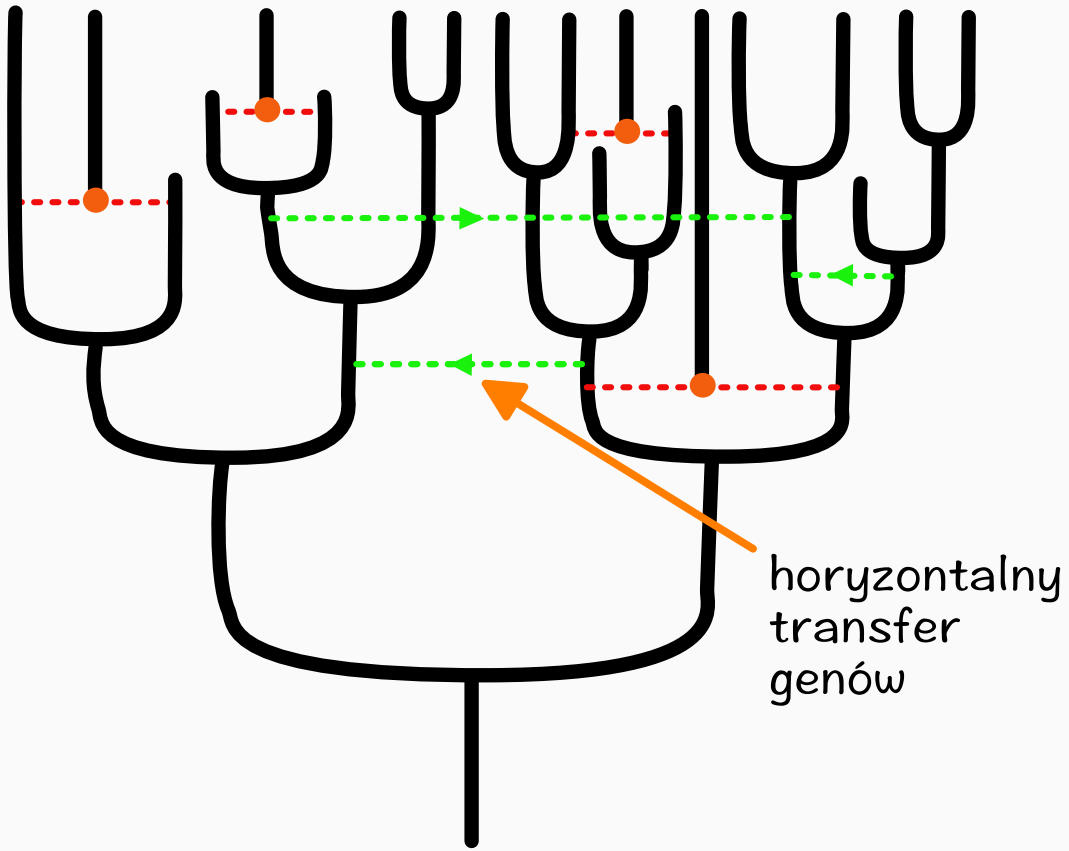
„Tradycyjne” drzewo

Drzewo życia



Drzewo uwzględniające krzyżówki międzygatunkowe

Drzewo życia



Drzewo uwzględniające krzyżówki międzygatunkowe i HGT