Drzewa filogenetyczne i Horyzontalny Transfer Genów

Grzegorz Góralski

Zakład Cytologii i Embriologii Roślin
Instytut Botaniki
Uniwersytet Jagielloński

Spis treści

Drzewa Filogenetyczne

Czym się zajmuje filogenetyka molekularna?

- **Filogenetyka** to nauka zajmująca się badaniem historii ewolucyjnej (filogenezy) organizmów lub ich grup z użyciem różnych metod w tym paleontologicznych, anatomii porównawczej, genetyki itd.
- **Filogenetyka molekularna**, jak wskazuje nazwa, skupia się na badaniach cząsteczek (DNA, białek) w celu rekonstrukcji filogenezy.
- W dalszej części mówiąc o filogenetyce będę miał na myśli głównie filogenetykę molekularną.
- Zwykle badania filogenetyczne zmierzają do stworzenia **drzewa filogenetycznego**, które w formie wizualnej pozwala przedstawić pokrewieństwa taksonów (zwykle gatunków) w badanej grupie, kolejność ich wyodrębniania oraz szacowane różnice genetyczne między nimi.

Etapy tworzenia drzew filogenetycznych

Proces tworzenia drzew filogenetycznych składa się z kilku 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 filogramu/kladogramu
- Poprawki: wskazanie outgrupy, obracanie gałęzi, wybór typu drzewa itp.

Wybór sekwencji do badań

- Pierwszym krokiem w badaniach filogenetycznych jest wybór odpowiednich sekwencji do analiz.
- Tego typu analizy opierają się na założeniu, 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.
- Takie porównania sekwencji mają oczywiście sens tylko wtedy, gdy pochodzą one od wspólnego „molekularnego” przodka, czyli są **homologiczne**. Na tym jednak nie koniec.
- Sekwencje homologiczne można bowiem podzielić na dwie kategorie:
  - **ortologi**: sekwencje, które miały wspólnego przodka zaraz przed procesem specjacji
  - **paralogi**: sekwencje, które powstały w skutek duplikacji, czyli miały wspólnego przodka przed zduplikowaniem.
- Do badań filogenetycznych należy wybierać ortologi.

Wybór sekwencji do badań

- Kolejnym aspektem, który należy wziąć pod uwagę przy wyborze sekwencji jest ich tempo ewolucji.
- Różne sekwencje DNA mają różne tempo gromadzenia mutacji.
- Generalnie niekodujące sekwencje DNA zmieniają się w toku ewolucji dużo szybciej niż geny.
- Przyczyną tej różnicy nie jest różne tempo mutacji ale presja selekcyjna.
- Geny także różnią się między sobą „wrażliwością” na mutacje.
- W niektórych z nich niemal każda zmiana prowadzi do upośledzenia właściwego funkcjonowania kodowanego białka - są to **geny konserwatywne**.
- Inne geny wykazują większą tolerancję.
- Zatem im bardziej gen jest konserwatywny tym mniej różnic zauważymy między sekwencjami pochodzącymi między badanymi organizmami.

Wybór sekwencji do badań

- 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.
- Ogólna zasada jest taka, że im bliżej są one spokrewnione tym bardziej zmienne sekwencje należy wybrać.
- Jeśli wybierze się sekwencję o zbyt małej zmienności, może okazać się, że nie ma różnic między badanymi cząsteczkami u blisko spokrewnionych organizmów albo jest ich zbyt mało aby wyciągnąć sensowne wnioski.
- Mniej oczywiste są konsekwencje wyboru zbyt zmiennej sekwencji. Zbyt wiele zmian może na tyle zatrzeć podobieństwa a także poprzednie mutacje, że sekwencje nie będą się nadawać do badań filogenetycznych.
- Przy wyborze rodzaju sekwencji do badań należy także wziąć pod uwagę aspekty praktyczne związane z sekwencjonowaniem DNA a także dostępność sekwencji w bazach danych (wtedy nie trzeba ich sekwencjonować we własnym zakresie).

Zbieranie sekwencji

- Sekwencje używane w badaniach pochodzą zazwyczaj z dwu źródeł:
  - badania własne
  - bazy danych
- Z punktu widzenia ekonomicznego i praktycznego im więcej sekwencji można pobrać z baz danych tym lepiej.
- Z drugiej strony, sekwencje pochodzące z własnych badań mogą wzbogacić dostępne dla innych badaczy bazy danych, co samo w sobie jest jakimś wkładem w naukę.

Genbank

- Trzy najbardziej znane, dostępne publicznie bazy danych sekwencji DNA (oraz RNA i białek) to:
  - **GenBank** utrzymywany przez National Center for Biotechnology Information (NCBI)
  - DNA DataBank of Japan (DDBJ),
  - The European Nucleotide Archive (ENA)
- Wszystkie trzy bazy współpracują ze sobą w ramach (International Nucleotide Sequence Database Collaboration](http://insdc.org)(INSDC) synchronizując dane. W dalszej części kursu skupimy się na bazie GenBank.
- Baza GenBank pozwala wyszukiwać sekwencje na kilka sposobów.
- Zapewne najczęściej używana jest metoda zbliżona do wyszukiwarek internetowych, polegająca na wpisywaniu tekstu, np. numerów sekwencji, nazw organizmów czy genów w okienko i przeglądaniu wyników wyszukiwania.
- Można przy tym wybierać spośród wielu dostępnych kategorii, m. in. sekwencje nukleotydów, genomy, taksony, białka.

Blast

- GenBank i inne podobne do niego bazy sprawdzają się dobrze, gdy wyszukujemy sekwencje po ich nazwie, opisie czy nazwie taksonu.
- Ale często trzeba podejść do problemu z drugiej strony - mamy sekwencję nukleotydów i chcemy znaleźć inne, podobne do niej.
- Jest tak na przykład gdy nie wiemy czy odpowiada ona jakiemuś konkretnemu genowi albo gdy chcemy sprawdzić u jakiego organizmu występuje sekwencja najbardziej podobna (np. jeśli badamy odcinek DNA niewiadomego pochodzenia).
- W takich sytuacjach z pomocą przychodzi **BLAST** (Basic Local Alignment Search Tool), który umożliwia wyszukiwanie takich samych i podobnych sekwencji nukleotydowych a także białkowych znajdujących się w bazie GenBank.

Zbieranie sekwencji

- Do badań filogenetycznych sekwencje zapisuje się zwykle w plikach tekstowych w formacie **FASTA**.
- Zapis danych w pliku tekstowym ma wiele zalet.
- Może być edytowany w dowolnym edytorze tekstu (np. Vim, Emacs, Notepad++, Atom, TextMate, Jed, Pico), a także łatwo używać do pracy z nimi licznych dostępnych w systemach Uniksowych (np. w Linuksie) narzędzi ułatwiających na przykład wyciąganie z nich konkretnych danych.
- Uwaga: Word NIE jest edytorem tekstu, plik zapisany w formacie Worda NIE jest plikiem tekstowym.

Plik FASTA

Plik zawierający sekwencje nukleotydów w formacie FASTA może wyglądać np. tak:

```
>KC879635_Magnolia_stellata
CTGCTAACTCTCAGTTTGGTCCTACTTCTGGTTCATTTTGTTACTAAAAACGGAGGGGGAA
ACTCAGTACCAAATGCTTGGCAATCCTTGGTAGAGCTTATTCATGATTTCGTGCCGAACCC
GGTAAACGAACAAATAGGTGGTCTTTCCGGAAATGTTCAACAAAAGTTTTCCCCTCGCATC
TCGGTCACTTCTACTTTTTCGTTATTTCGTAATCCCCAGGGTATGATACCTTATAGCTTCA
CAGTCACAAGTCATTTTCTCATTACTTTGGGTCTCTCATTTCCGATTTTTATTGGCATTAC
TATAGTGGGATTTCAAAGAAATGGGCTTCATTTTTTAAGCATCTCATTACCCGCAGGAGTC
CCACTGCCGTTAGC
>JN098455_Mimulus_guttatus
CTACTCACTCTCAGTTTGGTCCTACTTTTTGTTCATTTTGTTACTAAAAAGGGAGGAGGAA
ACTCAGTACCAAATGCTTGGCAATCCTTGGTAGAGCTTATTTATGATTTCGTGCCGAACCT
. . .
```

Format FASTA

- Każda sekwencja nukleotydów poprzedzona jest linią zaczynającą się znakiem >.
- Po tym znaku powinien znajdować się opis sekwencji.
- W powyższym przykładzie jest on dość lakoniczny, zawiera tylko numer dostępowy GenBank oraz nazwę taksonu, ale można tam zawrzeć też dużo więcej informacji.
- Tak na przykład wygląda nagłówek jednej z sekwencji pobranej z bazy GenBank:

```
>KX282989.1 Rumex vesicarius voucher EDNA15-0042869
ribulose-1,5-bisphosphate carboxylase/oxygenase large
subunit (rbcL) gene, partial cds; chloroplast
```

(w jednej linii)

- Po linii z opisem zapisana jest sekwencja nukleotydów lub aminokwasów.
- Może ona znajdować się w jednej lub wielu liniach, zasada jest taka, że wszystkie linie aż do następnego znaku > na początku linii powinny zawierać tylko i wyłącznie sekwencję.
- Niekoniecznie musi ona zawierać wyłącznie litery oznaczające nukleotydy lub aminokwasy, mogą tam także znajdować się dodatkowe oznaczenia np. niejednoznacznych lub nieznanych nukleotydów czy miejsc delecji.

Oznaczenia 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

- Czasem pomiędzy końcem sekwencji a nagłówkiem kolejnej dodawana jest pusta linia co może zwiększać czytelność dla człowieka.
- Format FASTA jest najprostszy i najpopularniejszy ale istnieją także inne, np. phylip, nexus, fastq.

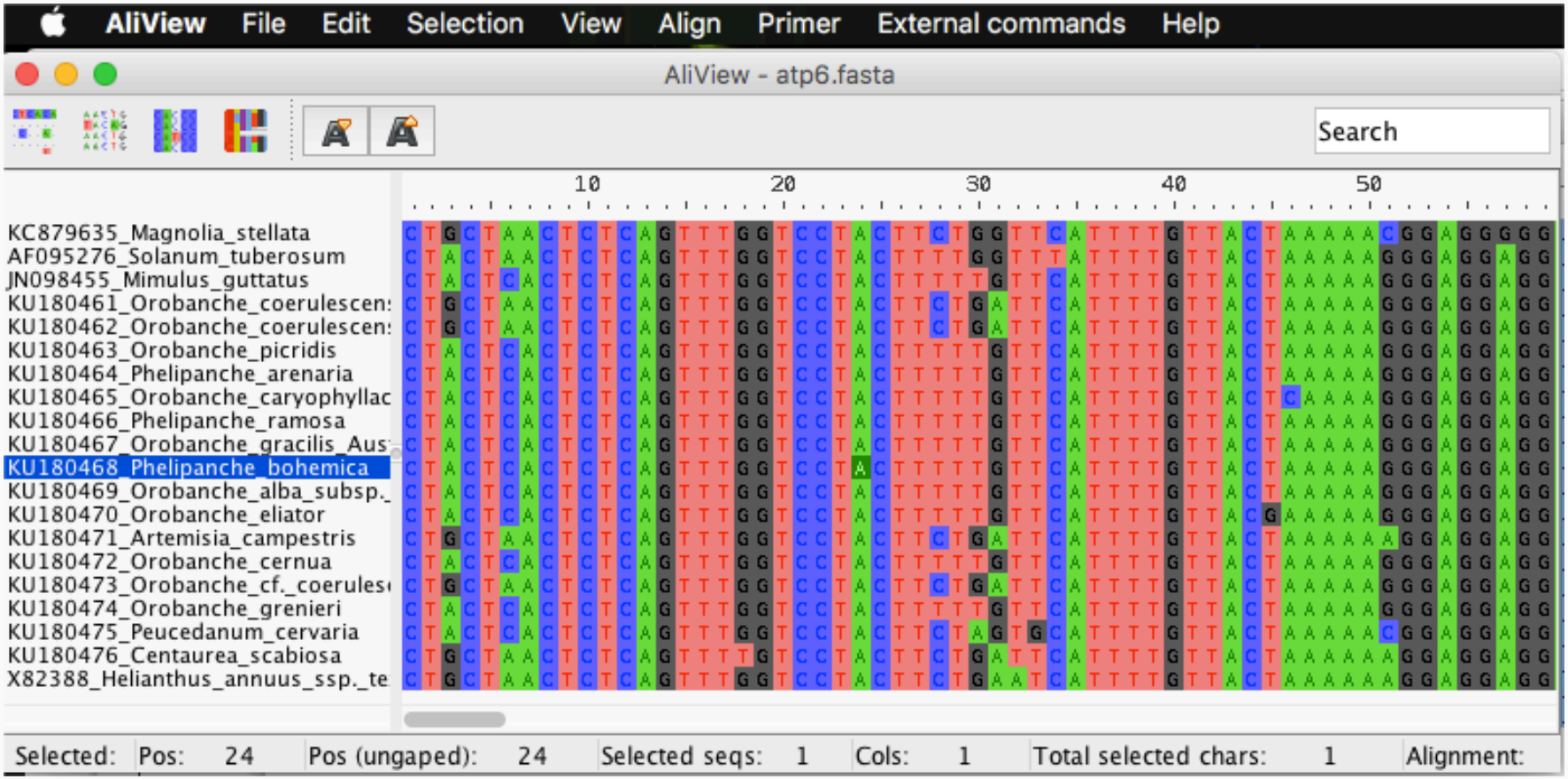
Dopasowanie sekwencji

- Jak wspomniałem wcześniej sekwencje używane do badan filogenetycznych powinny być homologiczne.
- Dopasowanie wybranych sekwencji polega na tym aby ustawione w kolejnych liniach sekwencje miały w kolejnych kolumnach **homologiczne** względem siebie **nukleotydy**.
- Dopasowanie sekwencji składa się zwykle z dwu etapów:
  - Wstępne dopasowanie automatyczne dokonywane przez odpowiednie programy
  - Poprawki dokonywane przez człowieka
- Do wstępnego automatycznego wyrównania stosowanych jest wiele programów, które używają różnych algorytmów.
- W dodatku na sposób i efektywność działania każdego z nich duży wpływ mają parametry, które ustawia się przy ich uruchamianiu.
- Dlatego na pytania w rodzaju „*który program jest najlepszy do dopasowania sekwencji?*” nie ma dobrej odpowiedzi.
- Bardzo duże znaczenie ma tu rodzaj dopasowywanych sekwencji i ustawienia programów.
- Do najpopularniejszych programów tego typu należą m. in. `clustalw`, `muscle`, `mafft`, `probcons`.

Dopasowanie sekwencji

- Jeśli porównywane sekwencje są stosunkowo mało zmienne i nie mają indeli (insercji i/lub delecji) automatyczne dopasowanie może nie wymagać ręcznych poprawek albo są one ograniczone do przycięcia końców sekwencji tak aby miały równą długość.
- Jeśli jednak tak nie jest,etap ręcznych poprawek może być długi i żmudny a końcowy efekt może być w większym lub mniejszym stopniu niepewny.
- Do pracy nad wstępnie wyrównanym zestawem sekwencji używać można edytorów tekstu, najlepiej z odpowiednimi skryptami/wtyczkami ułatwiającymi czytelne przedstawienie wyrównywanych sekwencji co związane jest z ich odpowiednim wyświetleniem oraz zwykle kolorowaniem.
- Częściej jednak, używa się w tym celu dedykowanych programów, które zwykle są wzbogacone w wiele dodatkowych funkcji ułatwiających pracę z plikami FASTA jak wyrównywanie sekwencji, zmiana na sekwencje odwrócone komplementarne, eksport do innych formatów a także dodatkowymi czynnościami jak wyszukiwanie sekwencji w bazach czy tworzenie drzewek filogenetycznych.
- Przykładami są programu AliView i Jalview:

AliView



AliView - wyrównane sekwencje

Jalview

Jalview - wyrównane sekwencje

Wybór modelu ewolucji molekularnej

- Kolejnym etapem w drodze do stworzenia drzewa filogenetycznego powinien być wybór modelu ewolucji molekularnej.
- Modele ewolucji molekularnej a dokładniej modele substytucji (podstawień) nukleotydów, opisują w jaki sposób mogły ewoluować badane sekwencje.
- Jeśli chcemy rozwikłać pokrewieństwa ewolucyjne między badanymi organizmami, co jest zasadniczym celem badań filogenetycznym, powinniśmy dysponować jakąś metodą oceny odległości ewolucyjnych między nimi.
- Organizmy o mniejszej odległości będą uważane za bliżej spokrewnione między sobą niż taksony bardziej od siebie oddalone.
- Najprostszym sposobem, który przychodzi do głowy jest proste porównanie sekwencji i wyliczenie w ilu miejscach się one różnią - im więcej różnic tym większa odległość ewolucyjna.
- Takie podejście co prawda pozwala ocenić różnice między sekwencjami ale niekoniecznie odzwierciedla rzeczywiste odległości ewolucyjne, zwłaszcza jeśli porównywane są sekwencje z dużą liczbą różnic.
- Niekoniecznie jest to intuicyjnie oczywiste ale wynika to ze sposobu w jaki zmieniają się nici DNA w czasie.

Ewolucja sekwencji

- Rozważmy hipotetyczną ewolucję dwu sekwencji, przedstawioną na poniższym rysunku:

Ewolucja sekwencji

Dopasowanie sekwencji

- W sumie w obu sekwencjach doszło do ośmiu substytucji. Dopasujmy teraz obie sekwencje do siebie:

AGACTTGCATC
AGACATGCATG

- Jak widać mutacje są widoczne w czterech pozycjach i sześciu nukleotydach z czego dwa mutowały dwukrotnie.
- Zaznaczmy teraz miejsca gdzie widoczne są różnice:

AGACTTGCATC
AGACATGCATG

- Okazuje się, że sekwencje różnią się tylko w dwu miejscach, mimo że w sumie wydarzyło się w nich osiem mutacji.
- Powyższy przykład pokazuje mechanizm „ukrywania się” mutacji.
- Porównując dwie sekwencje, jeśli widzimy różnicę między nukleotydami w danym miejscu, nie jesteśmy w stanie stwierdzić, czy jest ona wynikiem jednej czy wielu mutacji.

Dopasowanie sekwencji

- Co więcej, następujące po sobie mutacje mogą najpierw sprawić, że nukleotydy będą się różnić a później, że będą takie same (choć niekoniecznie takie jak na początku).
- Im więcej czasu upływa i im więcej zachodzi mutacji w badanych sekwencjach, tym większy odsetek zmian zostaje „zatarty”.
- O ile możemy przyjąć, że liczba mutacji w czasie rośnie w sposób liniowy, to liczba obserwowanych różnic rośnie liniowo tylko na początku (dla małej liczby różnic) a później coraz wolniej, ponieważ coraz więcej zmian wydarza się w tych samych miejscach.
- Liczba różnic zmienia się, dla sekwencji o równych proporcjach rodzajów nukleotydów, do wartości 3/4 liczby nukleotydów, przy czym zmierza do tej granicy coraz wolniej.
- Trzeba też pamiętać o tym, że zasady prawdopodobieństwa wskazują, że dla dwu losowo wybranych sekwencji DNA o tej samej długości 1/4 miejsc powinna być zgodna.

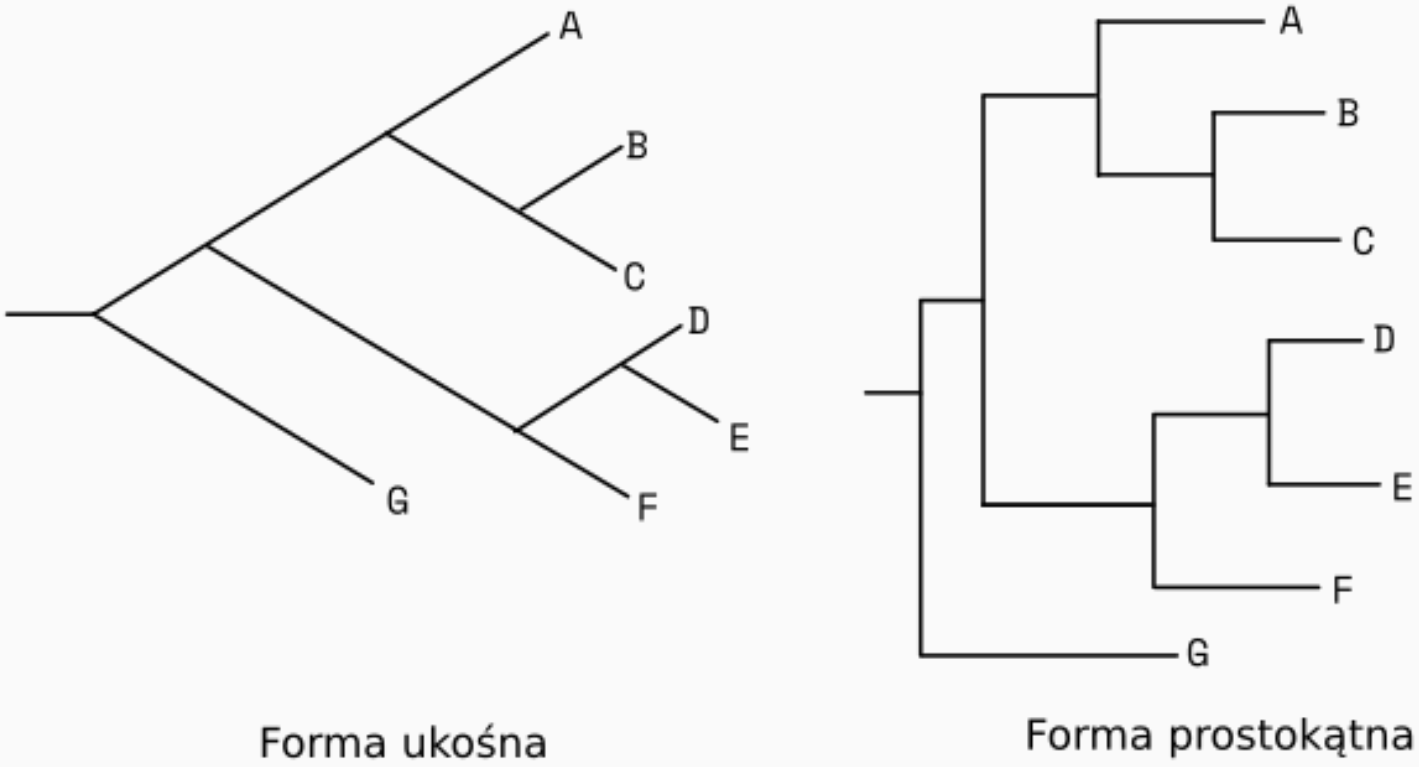
Dopasowanie sekwencji

- Jak widać, prosta metoda obliczania różnic między sekwencjami jest zawodna.
- Konieczne zatem okazało się stworzenie modeli, które w bardziej realistyczny sposób pozwalałyby oszacować odległości ewolucyjne.
- Bardziej złożone modele uwzględniają różne prawdopodobieństwa różnych rodzajów substytucji.
- Do najbardziej znanych modeli ewolucji molekularnej (substytucji) należą: Junkes-Cantor (JC, JC69), Kimura (K80), Felenstein (F81), Hasegawa, Kishino i Yano (HKY, HKY85), Tamura i Nei (TN93), General Time Reversible (GTR).

Struktura drzewa filogenetycznego

- Zanim przejdziemy do algorytmów wykorzystywanych przy konstruowaniu drzew filogenetycznych, zwanych też **dendrogramami**, przyjrzyjmy się pokrótce ich podstawowym formom i strukturze.
- Drzewa filogenetyczne najczęściej przedstawiane są w dwu formach: Ukośnej i prostokątnej.

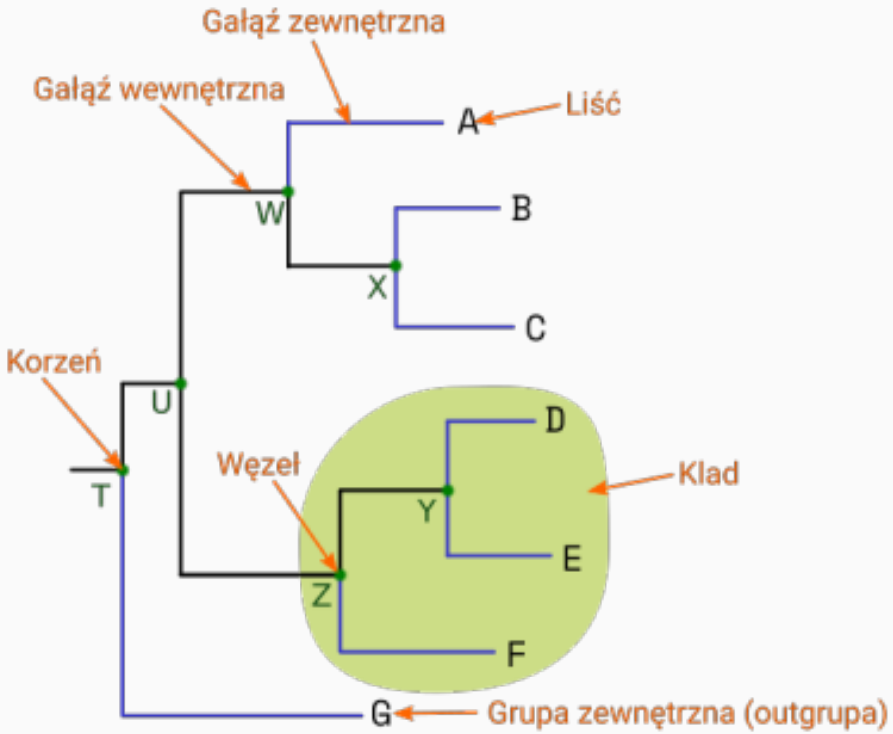
Formy drzew filogenetycznych



Formy drzewa filogenetycznego

Struktura drzewa filogenetycznego

- Podstawowe elementy drzewa filogenetycznego to: liście, gałęzie i węzły.



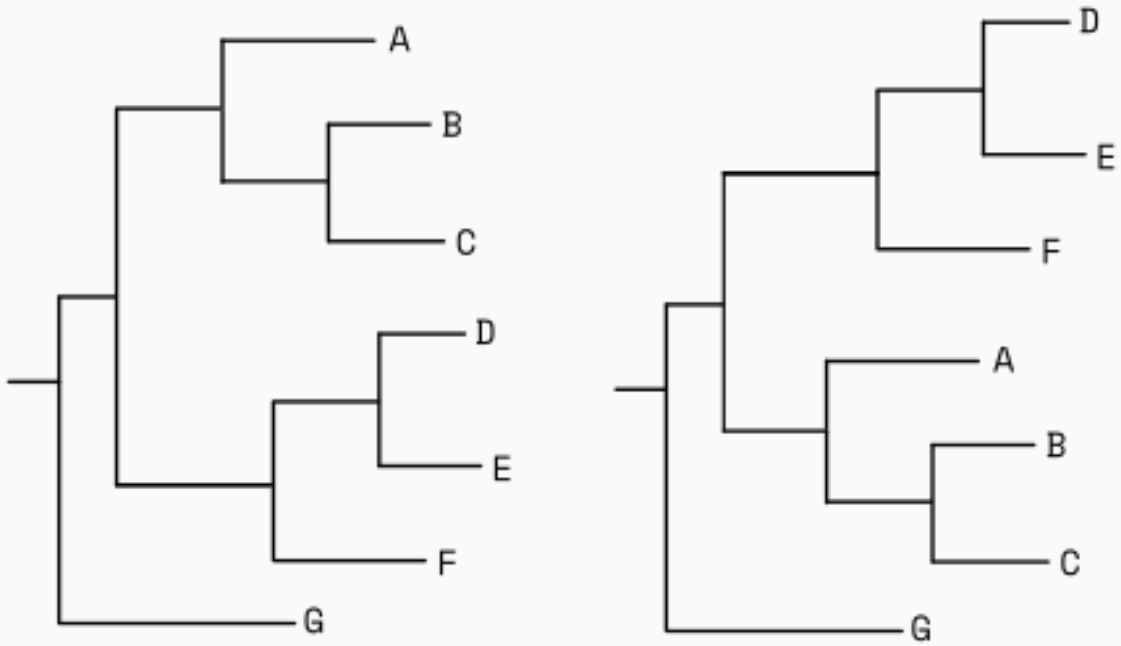
Struktura drzewa filogenetycznego

Struktura drzewa filogenetycznego

- **Gałęzie** pokazują związki pomiędzy nimi. Ich długość może (w zależności od rodzaju drzewa) odpowiadać zmianom 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.

Topologia drzewa

- Wzorzec rozgałęzienia drzewa nazywany jest **topologią drzewa**. Drzewa o takiej samej topologii mogą mieć inną reprezentację graficzną, wynikającą np. z obracania gałęzi względem węzła.
- Przykładowo poniższe dwa drzewa mają taką samą topologię mimo innego wyglądu:

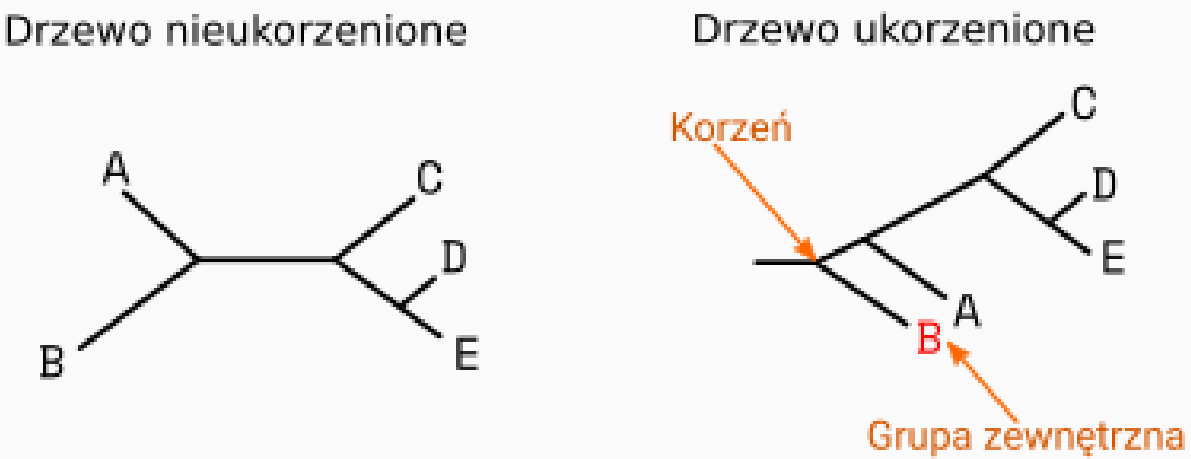


Ukorzenianie drzewa

- **Drzewa nieukorzenione** przedstawiają wzajemne podobieństwa ale nie pozwalają określić w jakiej kolejności poszczególne taksony się od siebie oddzielały.
- **Drzewa ukorzenione** posiadają węzeł, który odpowiada ostatniemu wspólnemu przodkowi badanych taksonów.
- Często wyznacza się go (jest to tzw. „ukorzenianie drzewa”) wskazując na **grupę zewnętrzną**, zwaną także **outgrupą** (ang. *outgroup*).
- Jest to takson (lub grupa taksonów), który jest dalej spokrewniony z pozostałymi badanymi, niż one między sobą.
- Innymi słowy, oddzielił się on najwcześniej podczas ewolucji.
- Przykładowo, gdybyśmy badali genetycznie gatunki *Homo*, grupą zewnętrzną mógłby być szympans.

Ukorzenianie drzewa

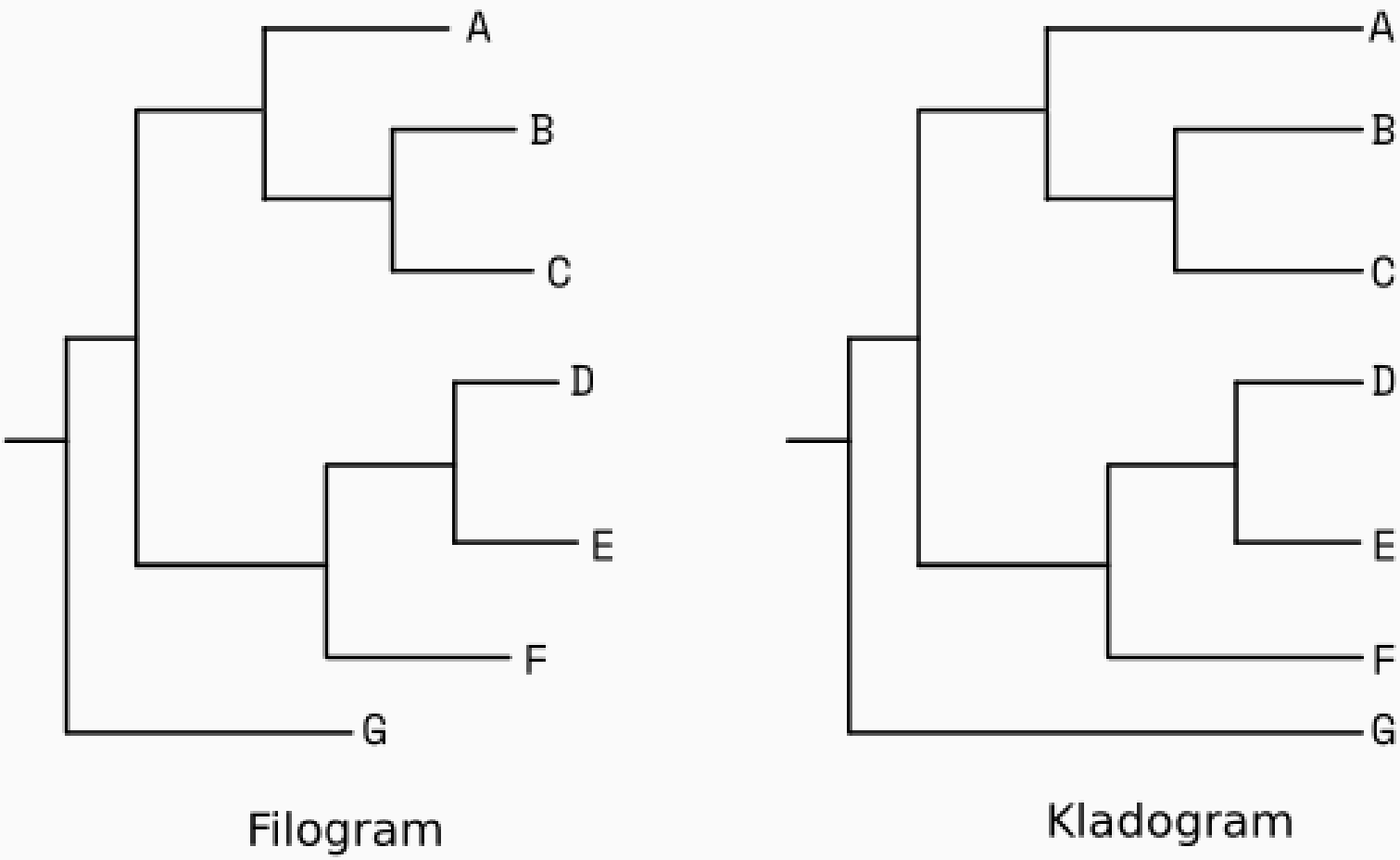
- Ukorzenianie drzewa pozwala ustalić kolejność oddzielania się poszczególnych kladów i liści w toku ewolucji.



Filogramy i kladogramy

- Jak wcześniej wspomniałem, długość gałęzi drzewa może odzwierciedlać odległość ewolucyjną badanych sekwencji, wtedy drzewo nazywamy **filogramem**.
- **Kladogram** natomiast pokazuje jedynie pokrewieństwa między badanymi taksonami.
- Wizualnie można go poznać po tym, że wszystkie gałęzie kończą się wzdłuż jednej linii (pionowej lub poziomej).

Filogram i kladogram



Filogram i kladogram

Konstruowanie drzewa filogenetycznego

- Mając dopasowane sekwencje nukleotydów oraz znaleziony model podstawień nukleotydów można przystąpić do konstruowania drzewa.
- Najbardziej znane metody używane przy konstruowaniu drzew to:
  - UPGMA (Unweighted Pair-Group Method using arithmetic Averages)
  - Metoda najbliższego sąsiada (NJ - NeighborJoining)
  - Metoda największej oszczędności (MP - Maximum Parsimony)
  - Metoda największej wiarygodności (ML - Maximum Likelihood)
  - Metody bayesowskie (Bayesian Methods)
- Możemy je wykorzystać w programach, które zazwyczaj implementują jedną z metod, choć często z pewnymi modyfikacjami i dodatkami.
- Należą do nich `PhyML`, `IQ-tree`, `RAxML`, `PHYLIP`, `PAUP*`, `mrBAYES`, `BEAST`.
- Istnieją także „kombajny”, jak np. `MEGA`, które pozwalają liczyć drzewa na kilka sposobów.

Szacowanie wiarygodności

- Konstruowaniu drzew towarzyszy zazwyczaj szacowanie ich wiarygodności.
- W większości przypadków stosuje się tu metodę **bootstrap** (samopróbkowania) a dla metod bayesowskich wyliczane jest prawdopodobieństwo bayesowskie.
- Samopróbkowanie w podstawowej formie polega na tym, że po utworzeniu optymalnego drzewa, z zestawu dopasowanych sekwencji losuje się kolumny zasad i tworzy się z nich kolejne zestawy „sekwencji” o takiej samej długości jak sekwencje wyjściowe.
- Jest to losowanie ze zwracaniem, co oznacza, że te same kolumny mogą zostać wylosowane wielokrotnie a inne nie pojawiają się w ogóle w generowanych zestawach.

Bootstrap

- Na przykład dla przyrównania:

```
0123456789 CAGTCCGATG
TAATCTGATA
TAGTTTGATA
TAGTTTGATA
TAGTTTGATA
TAGTTTGATA
```

- można stworzyć m. in. takie pseudosekwencje:

```
1735320955 | 8327248441
AATCTGCGCC | TTGAGCTCCA
AATTTATATT | TTAAACTCCA
AATTTGTATT | TTGAGTTTTA
AATTTGTATT | TTGAGTTTTA
AATTTGTATT | TTGAGTTTTA
AATTTGTATT | TTGAGTTTTA
```

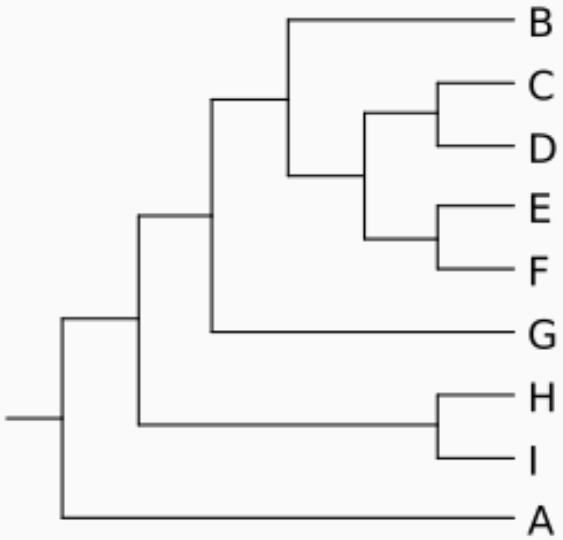
itd...

Bootstrap

- Dla każdego „pseudoprzyrównania” liczone jest drzewo w taki sam sposób jak drzewo główne, a następnie sprawdzana jest obecność poszczególnych kladów na obu drzewach.
- Każdemu kladowi, który występuje na drzewie oryginalnym i wygenerowanym w procesie samopróbkowania przypisywany jest punkt.
- Im większa suma punktów, tym dany klad na drzewie jest bardziej wiarygodny.
- Wartości bootstrap przedstawia się zwykle w zakresie wartości 0-100, przy węzłach, co odpowiada procentowi wygenerowanych drzew w których występował dany klad.
- Liczba bootstrapów, który jest zazwyczaj jednym z parametrów ustawianych w programach generujących drzewa, powinna wynosić minimum 100 a najlepiej osiągać 1000-2000.
- Ponieważ dla każdego zestawu pseudosekwencji generowane jest drzewo, w zależności od stosowanej metody, proces samopróbkowania może zająć mniej lub więcej czasu.

Format Newick i wartości dodatkowe na drzewie

- Po zakończeniu obliczeń otrzymujemy plik tekstowy opisujący relacje pomiędzy badanymi taksonami a także inne parametry drzew (np. wartości bootstap).
- Poniżej znajduje się przykładowy plik kladogramu zapisanego w formacie `newick`:
  `((((B,((C,D),(E,F))),G),(H,I)),A);`
- Po przekształceniu go w formę graficzną uzyskujemy taki obraz:



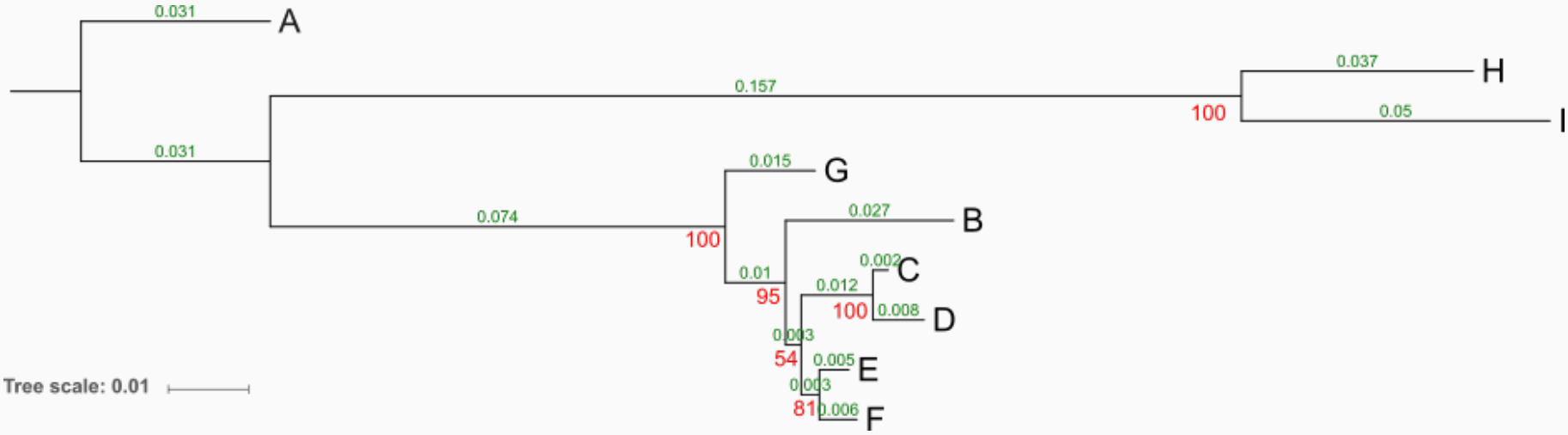
Format Newick

- Porównując powyższy plik w formacie `newick` z kladogramem można zrozumieć zasadę kodowania informacji w pierwszym z nich.
- W parze nawiasów zamykane się taksony należącego do wspólnego kladu.
- W formacie `newick` można też zapisać inne dane, na przykład dotyczące długości gałęzi i wartości bootstrapu:

```
(A:0.0611905636,((B:0.0271634370,((C:0.0024799833,D:0.0082762103)100:0.011585,
(E:0.0047747513,F:0.0060564542)81:0.002943)54:0.002522)95:0.009753,
G:0.0145402289)100:0.073576,(H:0.0374628169,I:0.0498809623)100:0.157039);
```

- Poniżej znajduje się odpowiedni dendrogram:

Oznaczenia na drzewku



Dendrogram z oznaczeniami

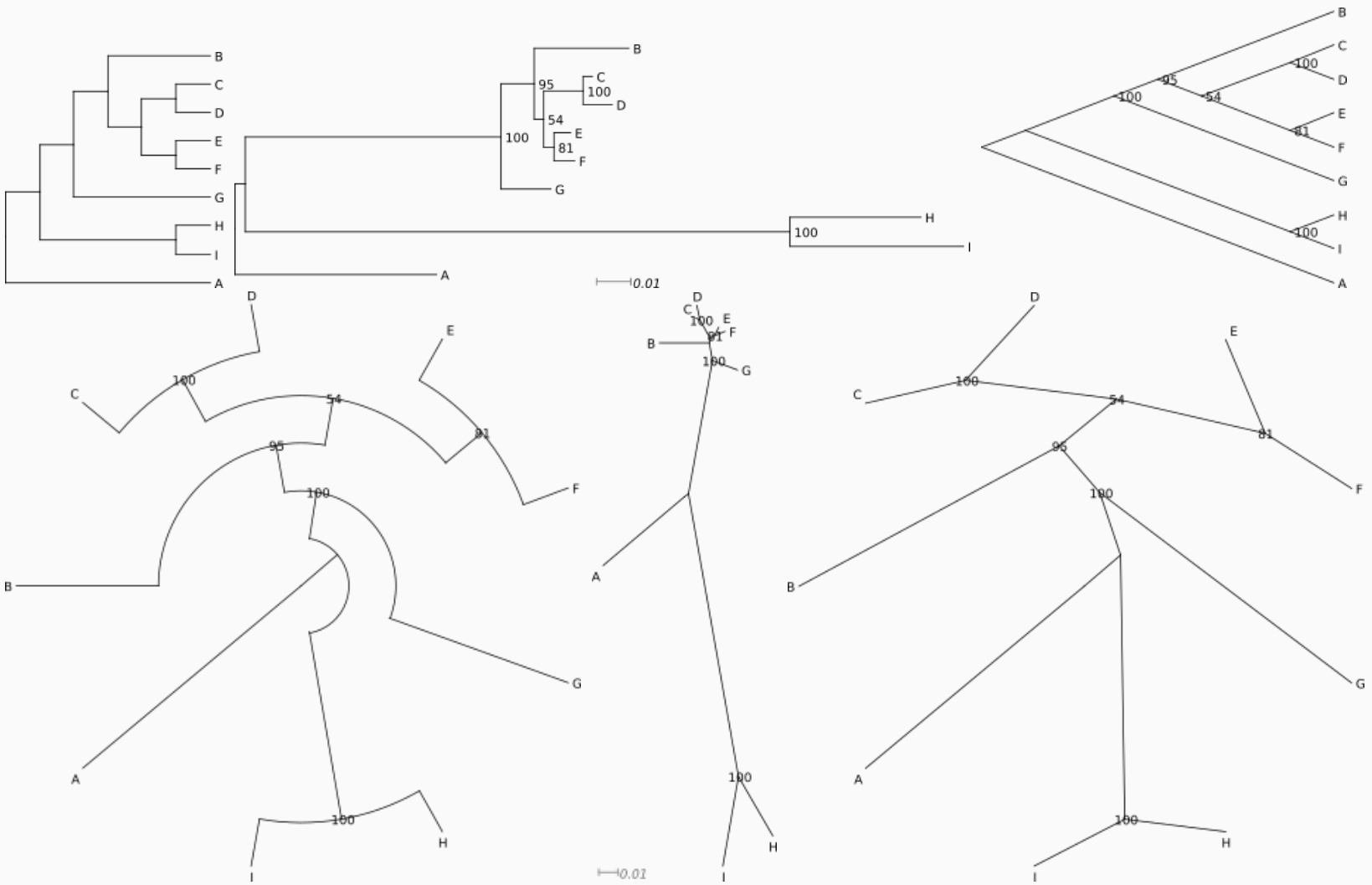
- Na zielono zaznaczono długości gałęzi (odpowiadające liczbie mutacji na miejsce), zaokrąglone do trzech miejsc po przecinku.
- Na czerwono wartości bootstrap.
- W lewym dolnym rogu widać skalę drzewa, którą można odnieść do długości gałęzi (zwłaszcza gdy ich wartości nie są zaznaczone).

Wizualizacja drzew

- Forma graficzna drzewa filogenetycznego jest znacznie bardziej przejrzysta dla człowieka niż prezentowany powyżej zapis tekstowy.
- Pozwała łatwo uchwycić pokrewieństwa i odległości ewolucyjne pomiędzy badanymi organizmami, choć ich prawidłowe odczytanie wymaga jednak nieco wiedzy i wprawy.
- Trzeba pamiętać, że opierając się na tych samych danych, można utworzyć bardzo różnie wyglądające drzewa.
- Poniżej znajduje się zawartość pliku w formacie `newick` opisująca drzewo i kilka przykładów jak można je przedstawić:

```
((((B:0.027163437,((C:0.0024799833,D:0.00827621):0.011585,
(E:0.004774751,F:0.006056454):0.002943):0.002522):0.009753,
G:0.014540229):0.073576, (H:0.037462816,I:0.049880963):
0.157039):0.003059528,A:0.058131035);
```

Różne typy drzew



Różne typy drzew

- Nie ma „najlepszej” formy drzewa.
- To jakiej należy użyć, zależy od tego co i w jaki sposób chcemy pokazać, liczby danych, rodzaju odbiorcy itp.
- W pewnych sytuacjach najlepiej sprawdzi się forma drzewa „prostokątnego” w innej drzewo „okrągłe”.

Horyzontalny Transfer Genów (HGT)

Czym jest HGT?

- Horyzontalny transfer genów (ang. *Horizontal Gene Transfer - HGT*) czasem zwany także poziomym transferem genów (ang. *Lateral Gene Transfer* - LGT) to proces przenoszenia materiału genetycznego pomiędzy organizmami w inny sposób niż rodzic-potomek (pionowy transfer genów, ang *vertical gene transfer* - VGT).

U jakich organizmów występuje HGT?

- Zjawisko po raz pierwszy odkryto w 1951 r. u maczugowca błonicy (*Corynebacterium diphtheriae*). Zauważono, że odpowiedzialny za patogenność gen pochodzenia wirusowego *tox* może przenosić się od bakterii patogennych do niepatogennych.
- W 1959 wykazano, że tą droga mogą się przenosić bakteryjne geny odpowiedzialne za odporność na antybiotyki
- Kolejne badania wskazały na dużą rolę HGT w wymianie materiału genetycznego u prokariontów. Kluczową rolę odgrywają w tej grupie organizmów takie procesy jak koniugacja, transdukcja i transformacja.
- Wykazano także, znaczny wpływ HGT na ewolucję eukariontów.
- Przede wszystkim zwraca się uwagę na rolę tego procesów u protistów.
- Obserwuje się je jednak u pozostałych grup *Eucaryota* i kolejne badania wskazują na istotną rolę w ewolucji tej grupy organizmów.

Pomiędzy jakimi organizmami występuje HGT?

- W przeciwieństwie do przenoszenia genów drogą krzyżowania międzygatunkowego, które ograniczone są do blisko spokrewnionych organizmów, wydaje się, że nie ma wyraźnych granic taksonomicznych dla HGT.
- Znane są transfery pomiędzy różnymi gatunkami bakteriami czy roślin, ale także pomiędzy bakteriami i grzybami, bakteriami i roślinami, bakteriami i zwierzętami, grzybami i zwierzętami czy grzybami i roślinami.
- Wydaje się zatem, że nie istnieją żadne bariery genetyczne „zakazujące” przenoszenia się materiału genetycznego pomiędzy nawet odległymi ewolucyjnie organizmami.
- Dalsze rozważania będą dotyczyć przede wszystkim roślin

W jaki sposób przenoszą się sekwencje DNA?

- Mechanizmy odpowiedzialne za HGT nie są dostatecznie wyjaśnione.
- Zwykle wskazuje się na:
  - Przenoszenie kwasów nukleinowych przez pośredników takich jak wirusy, bakterie, grzyby
  - Transpozony
  - Bezpośrednie pobieranie kwasów nukleinowych (zwłaszcza w układach pasożyt-żywiciel)
- Teoretycznie materiał genetyczny może przenosić się za pomocą fragmentów DNA lub poprzez mRNA, które następnie dzięki odwrotnej transkypcji mógłby z powrotem zostać przekształcony w DNA
- Badania wskazują raczej na tą pierwszą możliwość.

HGT u eukariontów

- Uważa się, że procesowi HGT sprzyja długotrwały fizyczny kontakt pomiędzy organizmami
- Taka sytuacja może dotyczyć np:
  - Endosymbiontów
  - Układów pasożyt-żywiciel
  - Szczepień
  - Wchłaniania jednych organizmów przez inne (pierwotniaki)
  - Jeśli przeniesione sekwencje DNA mają być przekazane następnym pokoleniom, muszą przedostać się do linii generatywnej (o ile organizm nie rozmnaża się bezpłciowo) toteż skutecznemu HGT sprzyja fizyczna komórek rozrodczych i symbiontów lub ich kontakt ze środowiskiem zewnętrznym

HGT w mitochondriach

- Mitochondria wydają się być szczególnie predysponowane do horyzontalnego transferu genów:
  - Posiadają mechanizmy pobierania DNA i RNA z otoczenia.
  - Często ulegają fuzji.
  - Roślinne mitochondria mają system rekombinacji homologicznej.
  - Ich genomy mają strukturę dynamiczną i ulegają rearanżacjom
  - Genomy mitochondriów roślinnych zawierają kilkadziesiąt genów
  - Pomiędzy genami znajdują się niekodujące odcinki w które może się wbudowywać obce DNA.
- U okrytonasiennych obce mtDNA zwykle pochodzi od mitochondriów innych okrytonasiennych ale znajduje się także geny mchów czy glonów.

HGT w jądrach komórkowych i plastydach

- W jądrach komórek roślin okrytonasiennych znaleziono także wiele śladów HGT
- Dotyczą one genów jądrowych a także transpozonów
- Ciekawym przypadkiem jest pasożytnicza roślina *Rafflesia cattlei*, u której znaleziono ponad 30 genów przeniesionych od żywiciela. Przynajmniej niektóre są funkcjonalne.
- Plastydy uważane są za bardzo odporne na takie procesy jak HGT czy IGT (zob. dalej).
- Obce sekwencje plastydowe, znajduje się raczej w innych genomach komórki - mitochondrialnym lub jądrowym

Transfer pomiędzy genomami wewnątrz komórki

- Fragmenty DNA mogą przenosić się z jądra komórkowego jednego organizmu do jądra komórkowego innego organizmu.
- Proces ten może także przebiegać pomiędzy wszystkimi elementami komórki zawierającymi materiał genetyczny: jądrem, mitochondriami i plastydami.
- Przenoszenie fragmentów DNA pomiędzy genomami wewnątrz komórki nazywamy międzygenomowym transferem genów (ang. *Intergenomic Gene Transfer - IGT*)
- Trzeba pamiętać, że genomy mitochondriów i plastydów maja charakter prokariotyczny a jądra (niejako z definicji) eukariotyczny.
- Mitochondria większości zbadanych roślin nasiennych zawierają sekwencje jądrowe i plastydowe.
- Geny mitochondrialne znajduje się także w plastydach, ale rzadko. Różnica wynika prawdopodobnie z tego, że mitochondria, w przeciwieństwie do plastydów, mają efektywne mechanizmy pobierania obcego DNA.
- W jądrach znaleziono wiele genów pochodzenia mitochondrialnego. W takich przypadkach następuje konwersja genów prokariotycznych w eukariotyczne co wiąże się m. in. z tym, że podlegają rekombinacji przy rozmnażaniu płciowym. Przypuszczalnie w tego typu IGT bierze udział RNA jako pośrednik.

Znaczenie HGT w ewolucji roślin

- Dotychczasowe badania wskazują na dużą rolę HGT w ewolucji eukariontów
- Ślady tego procesu znajduje się we wszystkich dużych grupach organizmów
- Odegrał także ważną rolę w ewolucji roślin
- Przykładowo, procesowi przekształcania się wewnątrzkomórkowego prokariotycznego endosymbiontu w chloroplast towarzyszył transfer kilkudziesięciu genów z chlamydii - które w tym czasie także prawdopodobnie były endosymbiontami komórek eukariotycznych.
- Uważa się, że geny pobrane od różnych organizmów miały istotną rolę w nabywaniu wielu ważnych cech umożliwiających m. in. adaptacje roślin do nowych i ekstremalnych warunków, efektywne reakcje na stress, wydajniejszą naprawę DNA, degradację celulozy czy rozwój tkanek przewodzących.

Rośliny pasożytnicze

- Rośliny pasożytnicze są dobrym kandydatem na organizmy pobierające obce DNA, ponieważ bezpośrednio są połączone z żywicielem i pobierają od niego składniki odżywcze.
- Połączenie odbywa się przez haustorium - strukturę która wnika w tkanki korzenia lub pędu gospodarza i pobiera wodę, sole mineralne i inne składniki odżywcze.
- Wyróżnia się dwie podstawowe kategorie pasożytów:
  - hemipasożyty (półpasożyty) - zdolne do prowadzenia własnej fotosyntezy, pobierające od żywiciela głównie wodę i sole mineralne (np. jemioła (*Viscum*), szelężnik (*Rhinanthus*))
  - holopasożyty - niezdolne do fotosyntezy, pobierają od żywiciela także cukry i inne składniki odżywcze (np. zaraza (*Orobanche*), kanianka (*Cuscuta*))
- Bardziej oczywistymi kandydatami na HGT wydają się być oczywiście holopasożyty

HGT u roślin pasożytniczych

- Rzeczywiście, badania wskazują, na stosunkowo liczne przypadki HGT w relacjach pasożyt-żywiciel
- Są więc dobrym modelem do badania tego procesu.
- Przy czym znajduje się nie tylko sekwencje przeniesione od żywiciela do pasożyta ale także od pasożyta do żywiciela.
- Nie jest zaskoczeniem, że głównie dotyczą one sekwencji mitochondrialnych, ale także znajduje się geny jądrowe i plastydowe
- Szacuje się, że u *Raflesiaceae* nawet ok 40% genów mitochondrialnych wykazuje ślady HGT

Wykrywanie HGT

- HGT wykrywa się głównie drogą znajdywania niezgodności na drzewach filogenetycznych.
- Porównuje się drzewo, które przedstawia „prawidłowe” relacje filogenetyczne z drzewem sporządzonym dla badanej sekwencji.
- Jeśli występują niezgodności, mogą one świadczyć o transferze genów.
- Położenie badanej sekwencji na drzewie filogenetycznym może wskazywać na źródło obcej sekwencji,
- Na przykład sekwencja pobrana od pasożyta może wykazywać bliskie podobieństwo do sekwencji żywiciela
- Wtedy można przypuszczać, że została pobrana od żywiciela i została wbudowana w genom pasożyta.

Nasze badania - transfer *atp6* u *Orobanchaceae*

Orobanchaceae

- Rodzina *Orobanchaceae* jest najliczniejszą pod względem pasożytów.
- Zawiera 90 rodzajów i 2060 gatunków obejmujących autotrofy, hemipasożyty oraz holopasożyty
- Zatem stanowi dobry model do badań na pasożytnictwem na różnych etapach jego rozwoju ewolucyjnego
- Uważa się, że półpasożytnictwo w tej rodzinie wyewoluowało raz, natomiast holopasożytnictwo kilkukrotnie
- Najliczniejszymi holopasożytniczymi rodzinami, są blisko spokrewnione *Orobanche* i *Phelipanche* zawierające ok. 150-200 gatunków

Orobanche flava

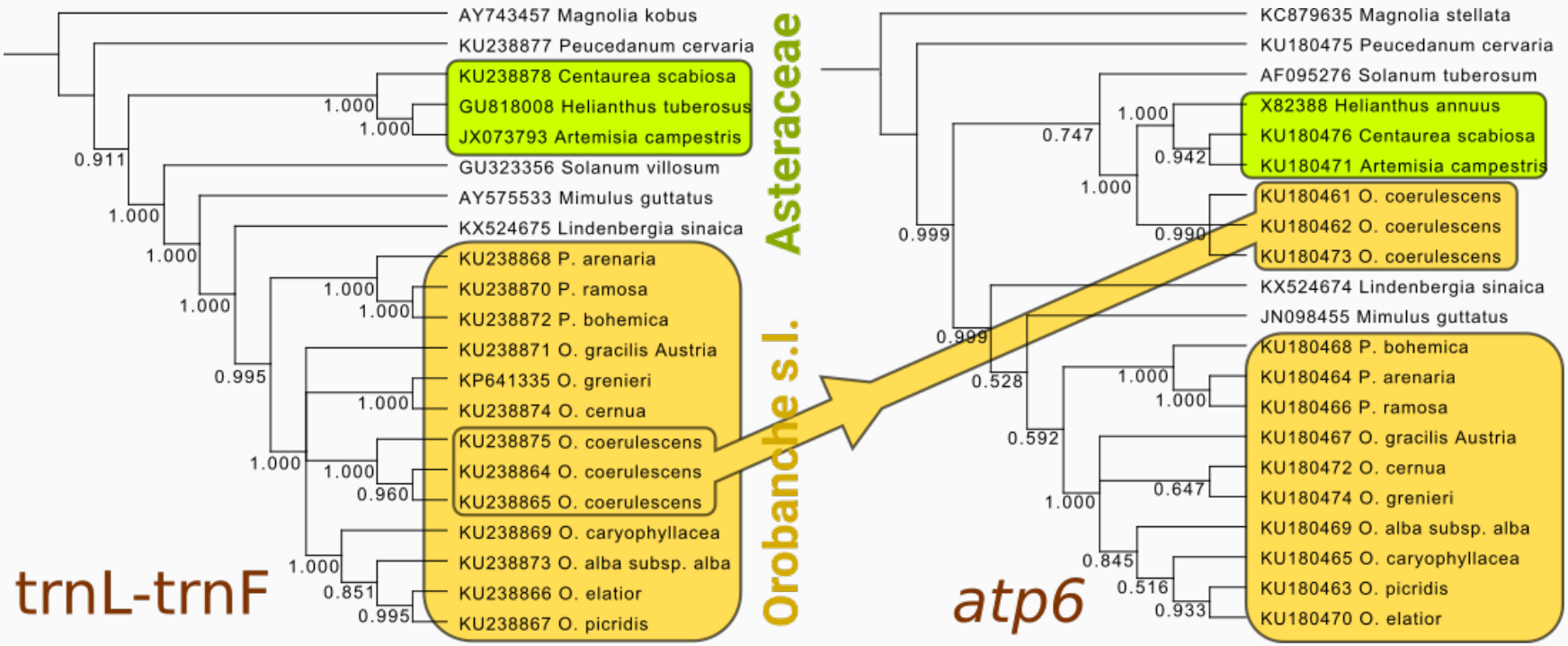
Orobanche flava

HGT u *Orobanchaceae*

- Chociaż *Orobanchaceae* wydaje się być idealnym kandydatem do badań nad HGT, stosunkowo niewiele doniesień na ten temat można znaleźć w literaturze.
- Opublikowano HGT zaledwie w przypadku kilku genów, w tym zaledwie jeden dotyczący genu mitochondrialnego i to znalezionego u żywiciela a nie u pasożyta.
- Nasze badania polegały na sprawdzeniu czy w sekwencjach genów mitochondrialnych nie ma śladów HGT
- Jako referencyjne drzewo filogenetyczne używaliśmy sekwencji *trnL-trnF*, odzwierciedlających „właściwe” relacje filogenetyczne
- Na drzewach znalazły się sekwencje różnych gatunków *Orobanche* i *Phelipanche* a także sekwencje wybranych gatunków z innych grup roślin, włączając żywicieli lub ich krewniaków.
- W przypadku genu *atp6* udało się uzyskać sygnał wskazujący na HGT tego genu.

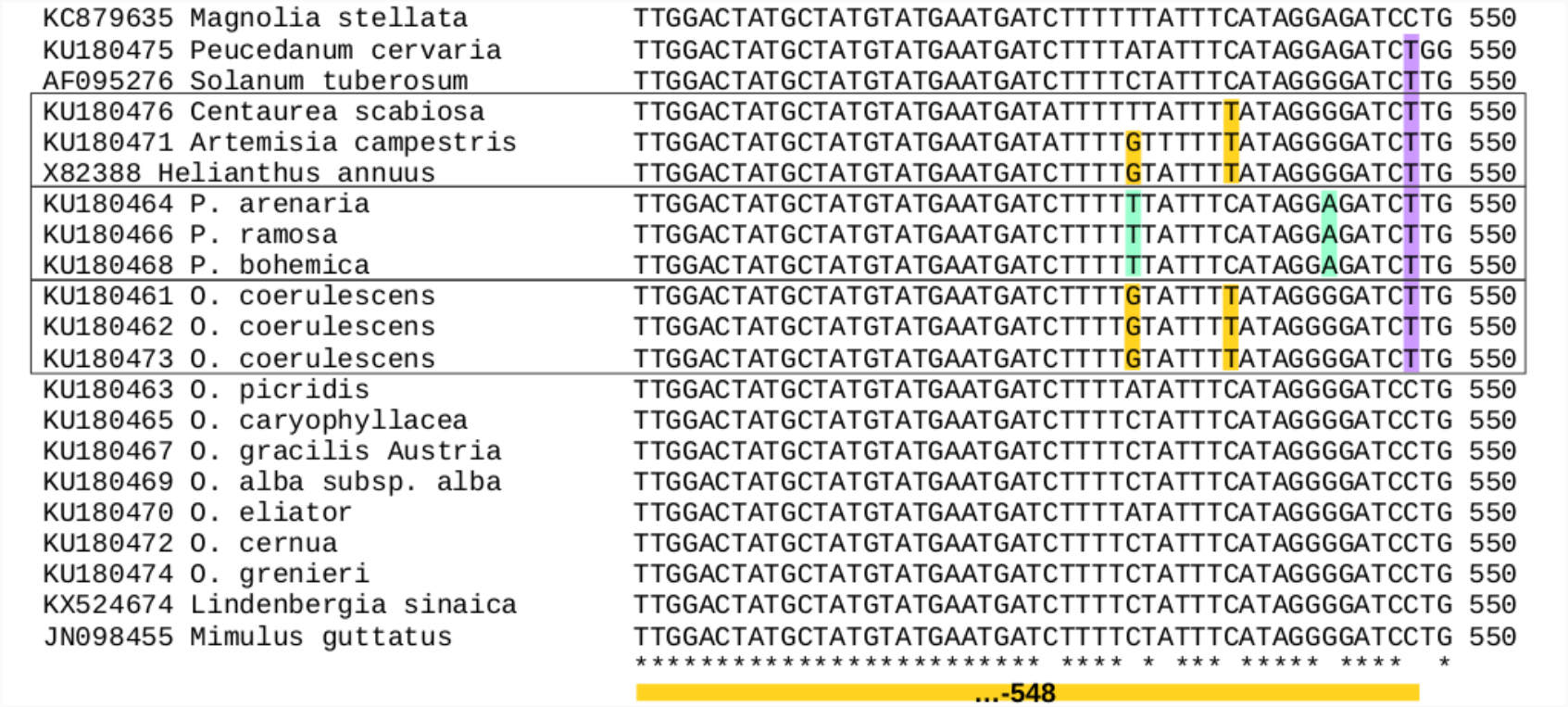
HGT u *Orobanche coerulescens*

- Badanie całej sekwencji wskazało na transfer *atp6* u *Orobanche coerulescens*



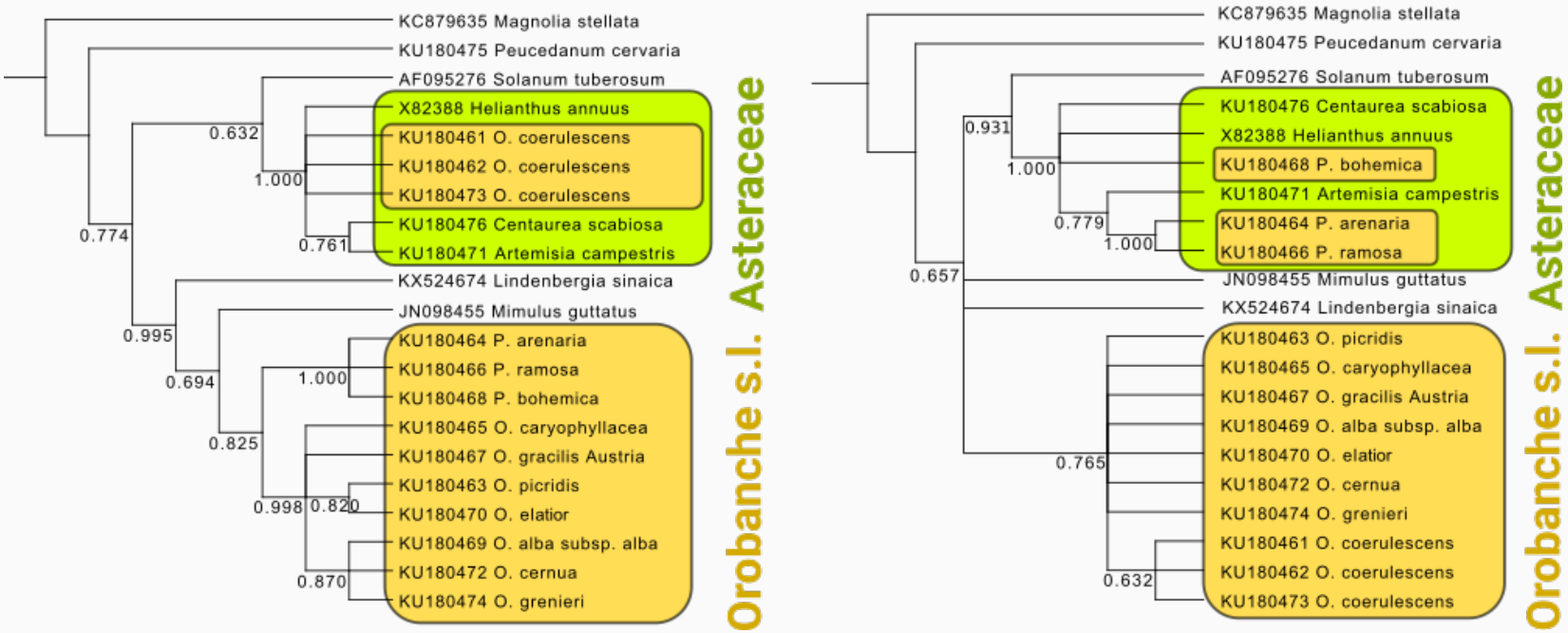
HGT u _O. coerulescens

- Dalej przyjrzeliśmy się bliżej sekwencjom badając które nukleotydy zostały przeniesione a które nie:



Badanie sekwencji

- Takie podejście przyniosło ciekawe rezultaty, okazał się bowiem, że ślad HGT występuje także u *Phelipanche*, że są to najprawdopodobniej dwa różne transfery i w dodatku dotyczą różnych fragmentów *atp6*



HGT u *Orobanche* i *Phelipanche*

Dwa HGT

- HGT1 - wydarzył się stosunkowo niedawno, nie ma go u pokrewnych gatunków. *atp6* w tym przypadku ma charakter hybrydowy - składa się z fragmentów „oryginalnych” i przeniesionych od gospodarza.
- HGT2 - wydarzył się dawno, u wspólnego przodka badanych *Phelipanche*. Obejmuje końcowy fragment genu, nie wiadomo jak daleko sięga.
- Planujemy dalsze badania transferów.

Dziękuję za uwagę.