

wykład 8

# **Analizy filogenetyczne**

dr Jacek Śmietanski

[jacek.smietanski@ii.uj.edu.pl](mailto:jacek.smietanski@ii.uj.edu.pl)

<http://jaceksmietanski.net>

1. Cele i zastosowania
2. Podstawy ewolucyjne
3. Drzewa filogenetyczne
4. Metody konstrukcji drzew
5. Ocena wiarygodności drzewa



# Cele i zastosowania

Analizy filogenetyczne wykorzystuje się aby poznać **pokrewieństwa między gatunkami** oraz innymi jednostkami taksonomicznymi.

Pozwalają określić w jakim stopniu **spokrewnione są różne geny w obrębie genomu**, przyczyniając się do zrozumienia jak powstają i ewoluują rodziny genów.

Mają wielkie znaczenie praktyczne **w medycynie i ochronie zdrowia**. Pozwalają one na określenie pochodzenia szczepów patogenów, śledzenie obecnych i historycznych dróg ich transmisji między regionami geograficznymi czy nosicielami oraz identyfikację źródeł pochodzenia epidemii.

W przypadku szybko mutujących wirusów metody filogenetyczne umożliwiają **śledzenie zmian zachodzących w czasie choroby u jednego pacjenta**.

Wyniki analiz filogenetycznych mogą być wykorzystywane jako **materiał dowodowy w sądach**.

O ile jeszcze niedawno drzewa filogenetyczne budowane z kilkuset sekwencji uchodziły za duże, teraz nierzadko konstruuje się drzewa na podstawie **setek tysięcy sekwencji**.

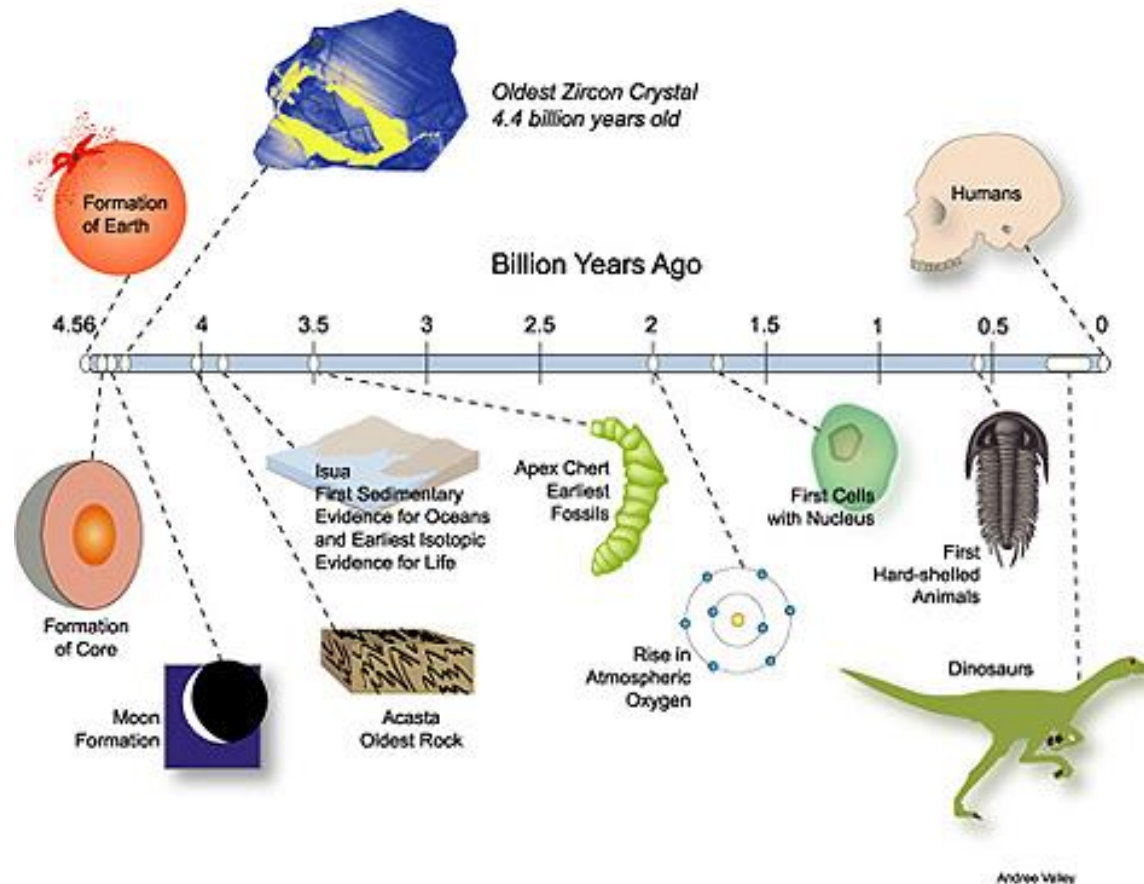
Wymusza to projektowanie coraz bardziej **wydajnych algorytmów** heurystycznych o jak najlepszych własnościach obliczeniowych.

Algorytmy te musi charakteryzować **szybkość działania** na dużych zbiorach danych, ale też wystarczająca **precyzja**.

Zapotrzebowanie na takie algorytmy wśród biologów i medyków jest ogromne, co stawia wielkie wyzwania przed bioinformatykami, ale i daje szansę na ciekawą, stymulującą intelektualnie i użyteczną dla innych pracę.

# Podstawy ewolucyjne

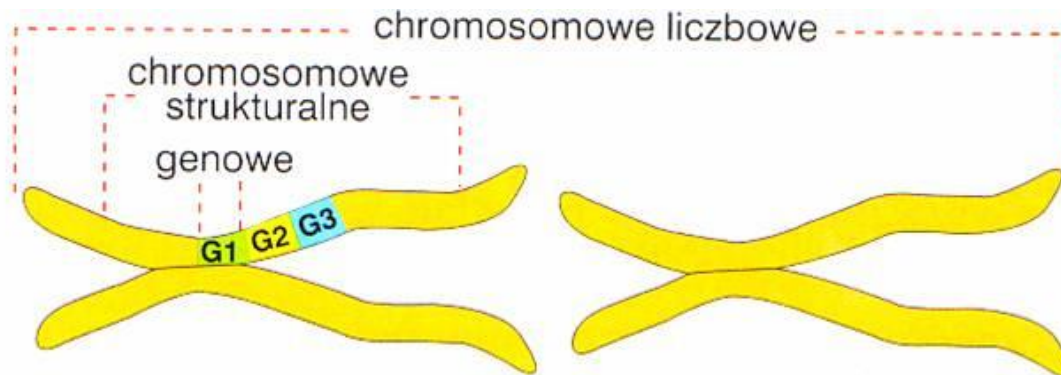
Ewolucja to rozwój formy biologicznej z innych wcześniej istniejących form lub jej powstanie w postaci istniejącej obecnie na skutek działania doboru naturalnego.



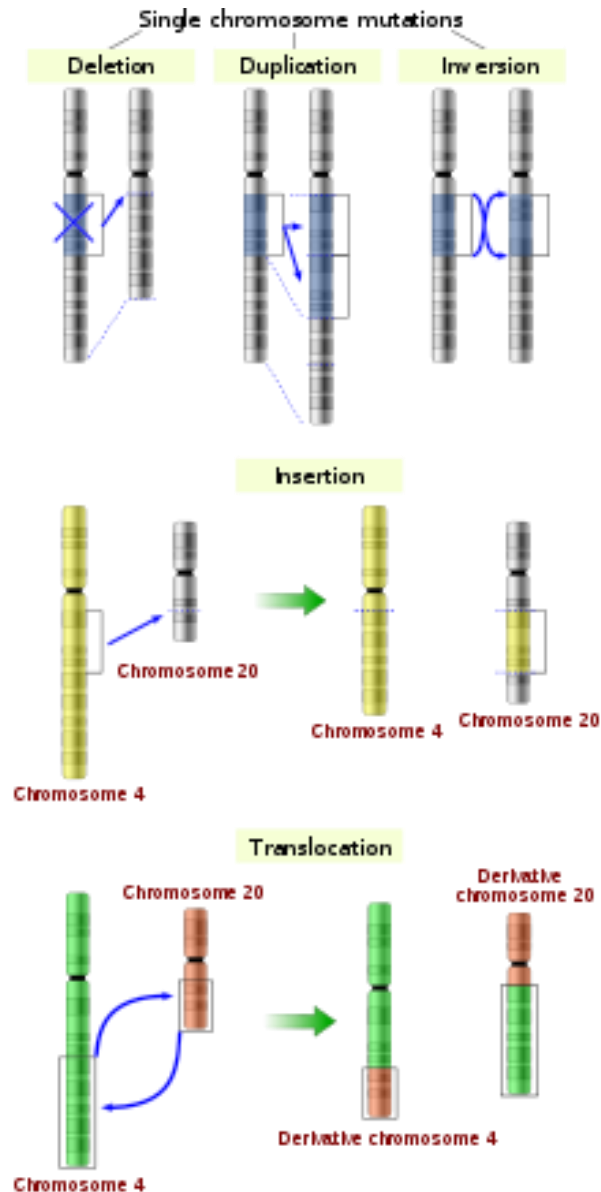
Ewolucja postępuje w wyniku nagromadzenia się mutacji.

# Rodzaje mutacji (podział ze względu na zasięg)

- chromosomowe:
  - liczbowe;
  - strukturalne;
- genowe (punktowe)



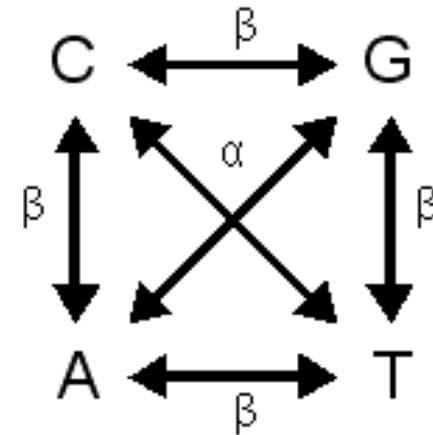
Ryc. Modelowy podział mutacji o różnym zasięgu (G1, G2, G3 – geny)





# Mutacje punktowe - substytucja

$\alpha$  = **tranzycja** ( $A \leftrightarrow G$ ,  $C \leftrightarrow T$ )  
 $\beta$  = **transwersja** (pozostałe)



Wpływ na kodowane białko:

**substytucja cicha** – zamiana nukleotydów nie powodująca zmian w sekwencji aminokwasowej;

**substytucja błędna** (missense) – zamiana kodowanego aminokwasu;

**substytucja nonsensowna** (nonsense) – zamiana kowanego aminokwasu na kodon STOP.

## Wpływ na kodowane białko:

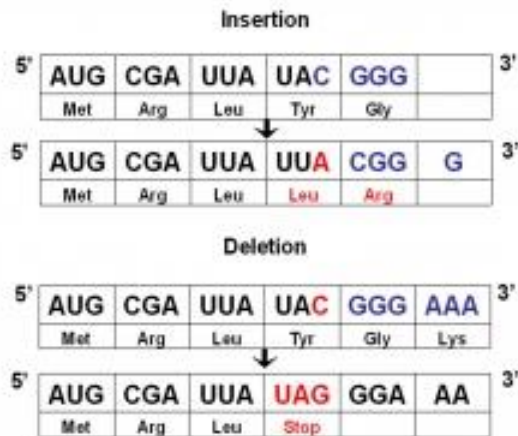
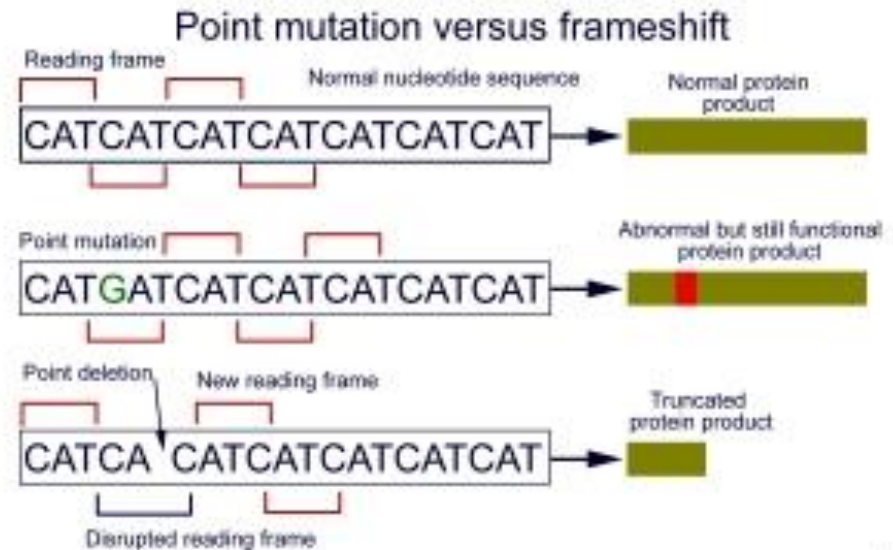


Figure 2. Schematic representation of nucleotide insertion and deletion

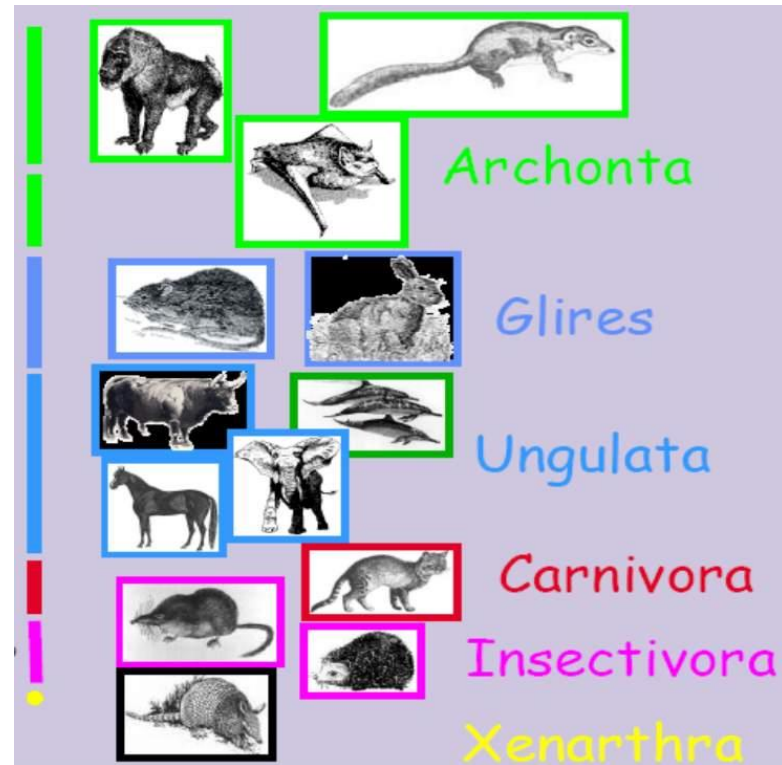


Filogenetyka zajmuje się badaniem ewolucyjnej historii organizmów.

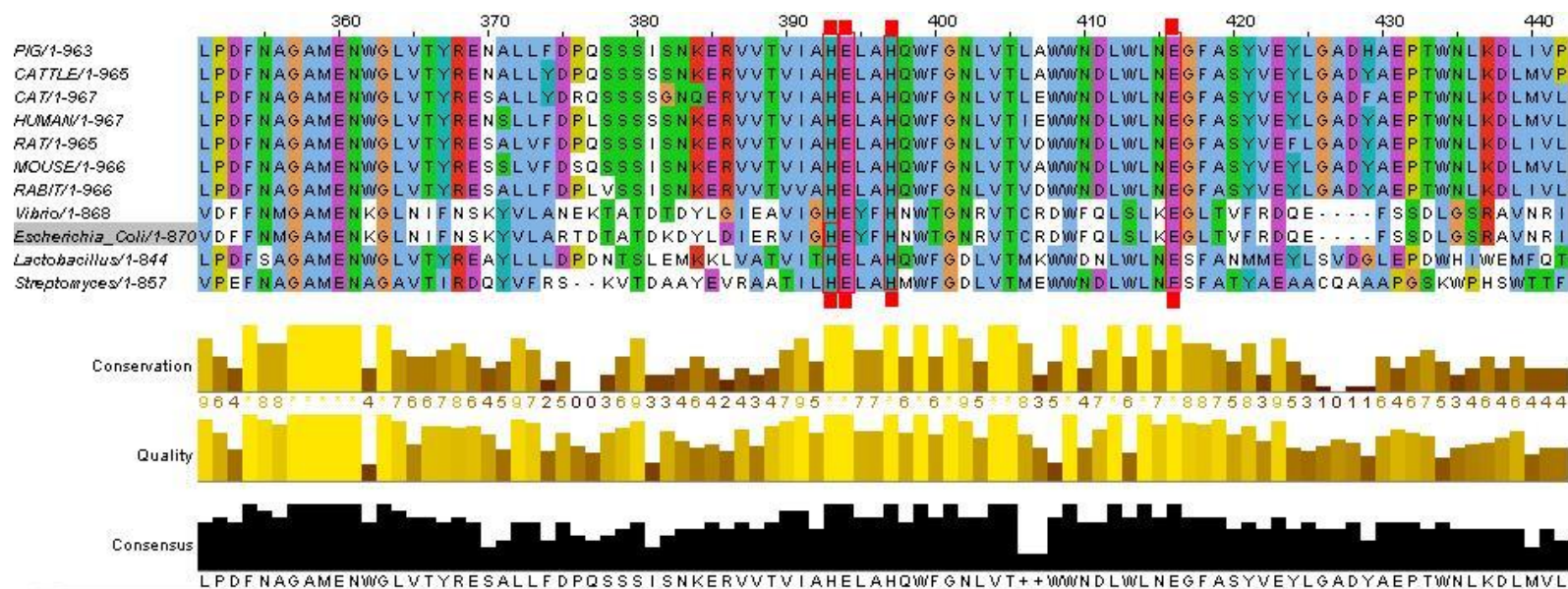
Zadania filogenetyki:

1. Rekonstrukcja ewolucyjnej historii wszystkich organizmów
2. Odkrycie przodka wszystkich organizmów żyjących na Ziemi
3. Segregacja i klasyfikacja organizmów
4. Poznanie mechanizmów ewolucji

Rekonstrukcja historii ewolucji głównie w oparciu o cechy morfologiczne np. długość dzioba u ptaków, nóg itd.

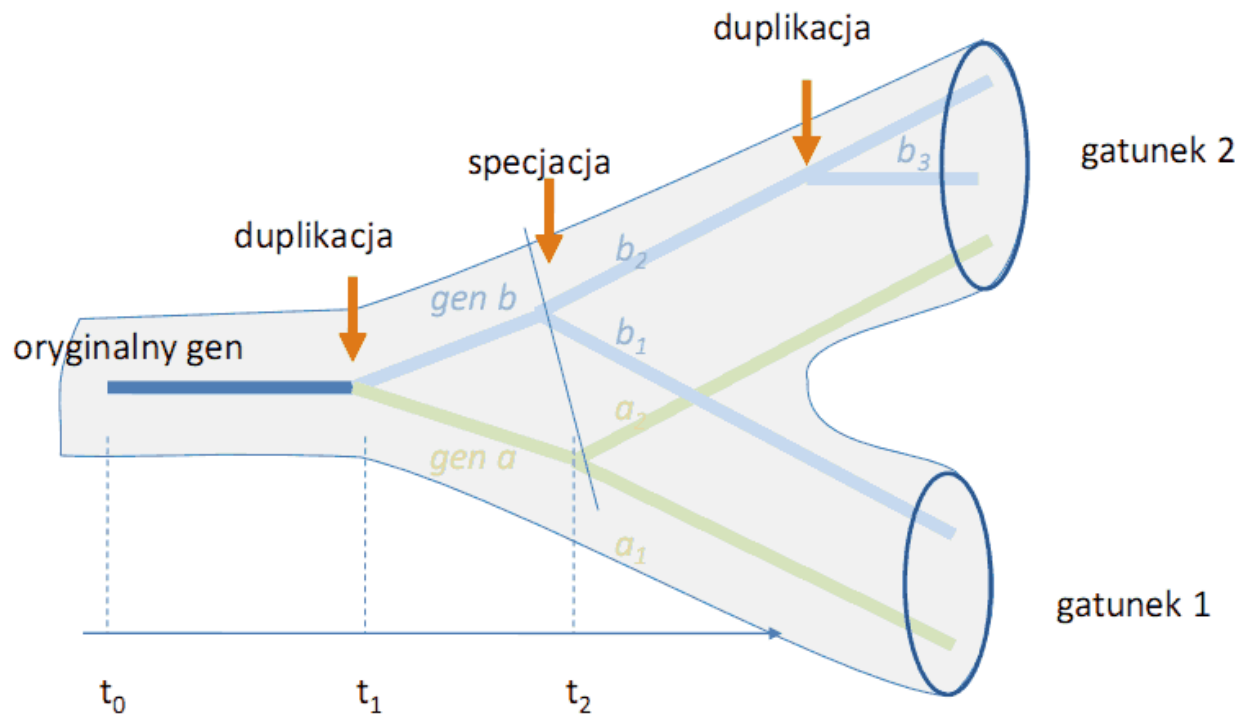


Rekonstrukcja historii ewolucji poprzez porównanie sekwencji nukleotydowych lub aminokwasowych pochodzących z różnych organizmów.



# Homologia

## Podobieństwo cech odziedziczonych po wspólnym przodku



ortologi – homologii powstałe w procesie specjacji; często pełnią podobną funkcję  
paralogi – homologii powstałe w wyniku duplikacji; zazwyczaj pełnią różne funkcje  
ksenologi – homologii nabyte w wyniku poziomego przenoszenia informacji genetycznej



1. Analizowane sekwencje są homologiczne
2. Dywergencja filogenetyczna jest dychotomiczna
3. Niezależna ewolucja każdej pozycji w sekwencji  
(uproszczenie ze względu na procedury obliczeniowe; ew. można użyć macierzy substytucji zależnych od kontekstu)
4. Różnorodność analizowanych sekwencji dostarcza informacji umożliwiającej konstrukcję jednoznacznych drzew filogenetycznych

**takson** – grupa organizmów uznawanych za spokrewnione, wyróżniających się konkretną cechą różniącą je od innych jednostek taksonomicznych;

**klad** - grupa taksonów mających wspólnego przodka, obejmująca wszystkie wywodzące się z niego grupy potomne;

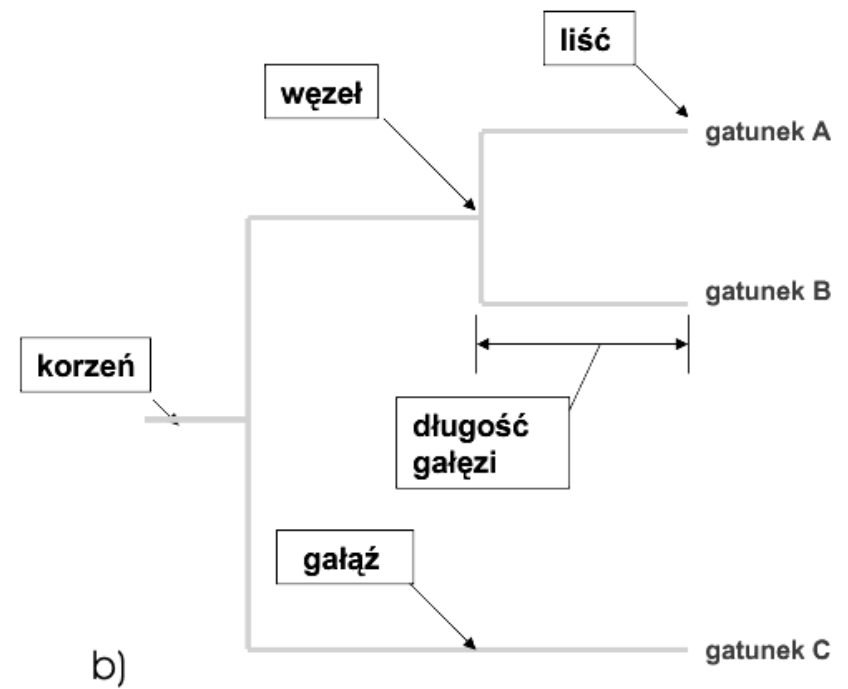
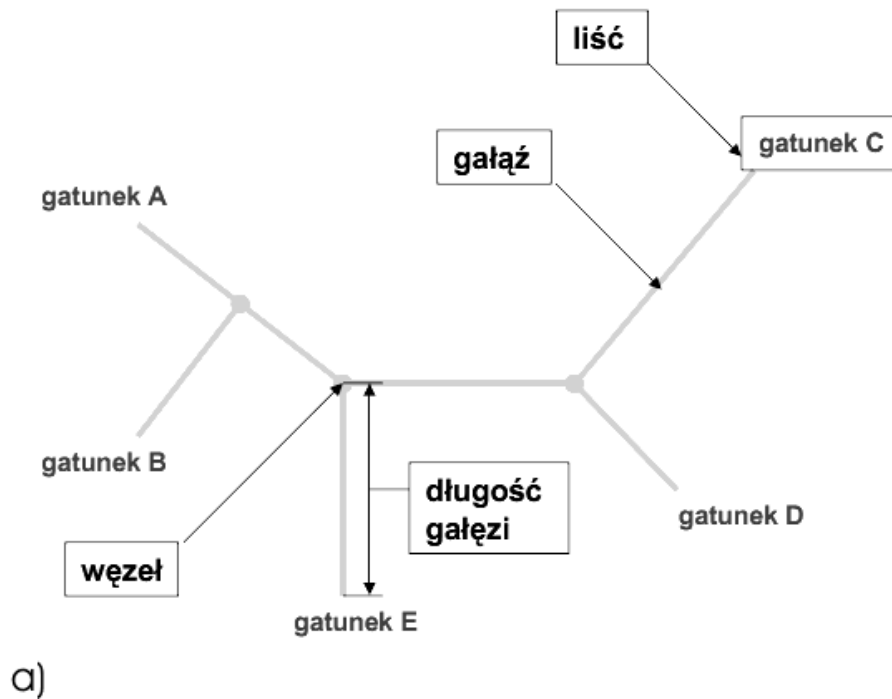
**filogeneza** – nauka badająca drogi rozwoju rodowego, pochodzenie i ewolucję w obrębie jakiejś grupy;

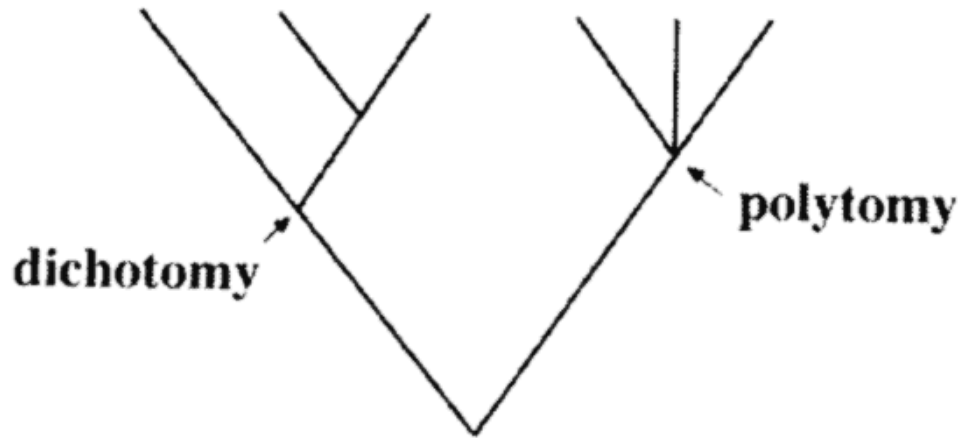
**drzewo filogenetyczne** – ilustracja pochodzenia i relacji pomiędzy poszczególnymi taksonami



# Drzewa filogenetyczne

- a) drzewo nieukorzenione (nie znamy wspólnego przodka)
- b) drzewo ukorzenione

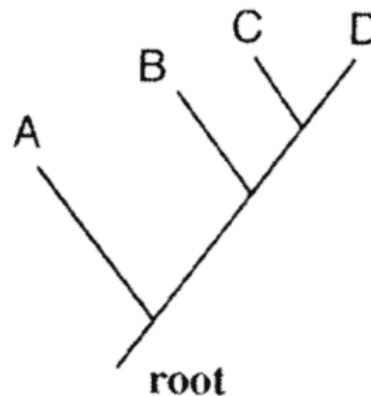
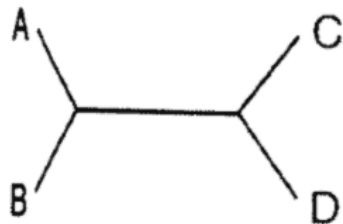




Wielokrotne rozdzielenie (politomia) wynika zwykle z niewystarczającej ilości informacji potrzebnej do rozwiązania struktury drzewa (ale może też być skutkiem procesu radiacji)

Korzeń może nie być znany, bo wspólny przodek wyginął.

- wykorzystanie grupy zewnętrznej (np. sekwencja ptaka przy analizie filogenezy ssaków)
- ukorzenianie w środkowym punkcie – punkt równo oddalony od najbardziej różniących się grup (zgodnie z hipotezą zegara molekularnego).

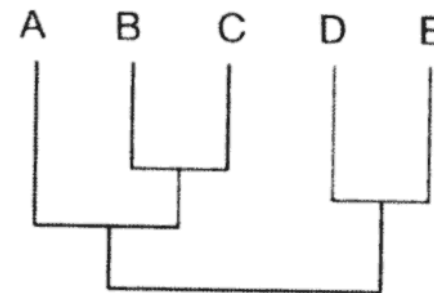
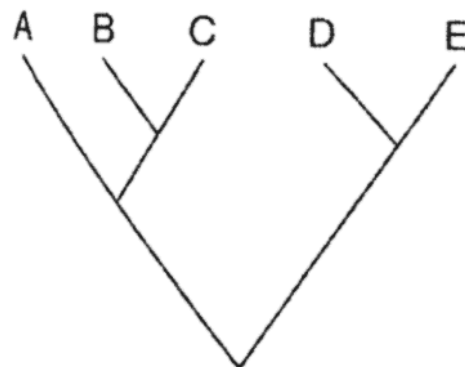


Koncepcja zegara molekularnego (Zuckerlandl i Pauling, 1965) postuluje równe tempo substytucji we wszystkich liniach ewolucyjnych.

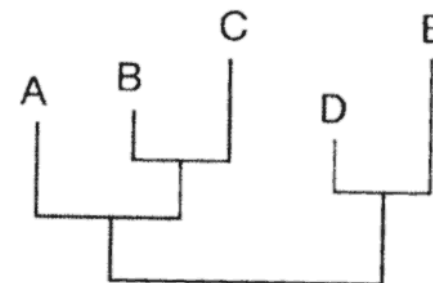
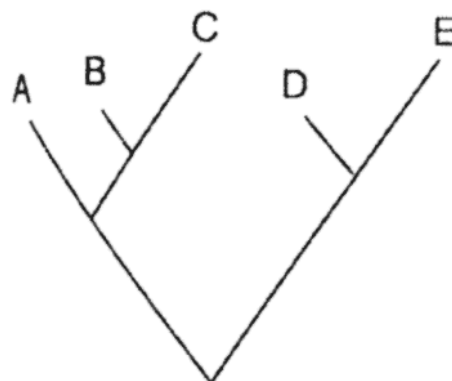
Hipoteza nie do końca prawdziwa, ale wykorzystywana w niektórych algorytmach w celu uproszczenia obliczeń.



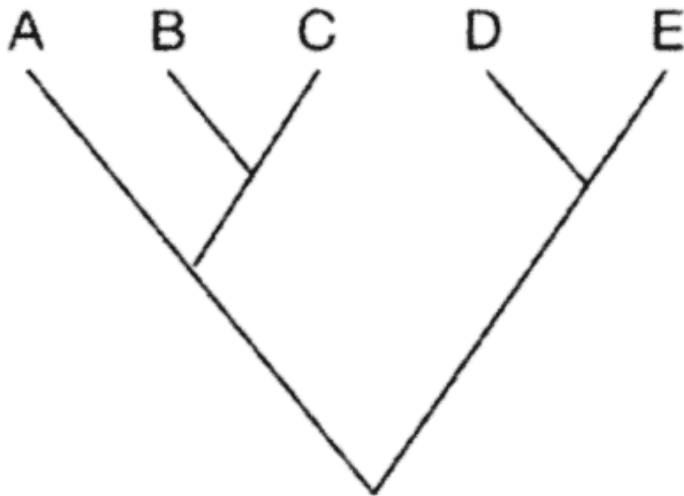
kladogram



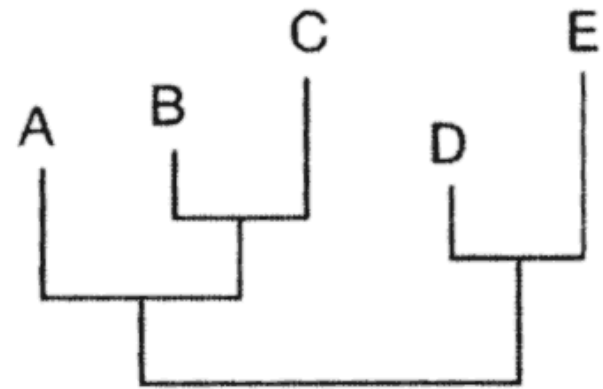
filogram



## Reprezentacja liniowa: format Newick



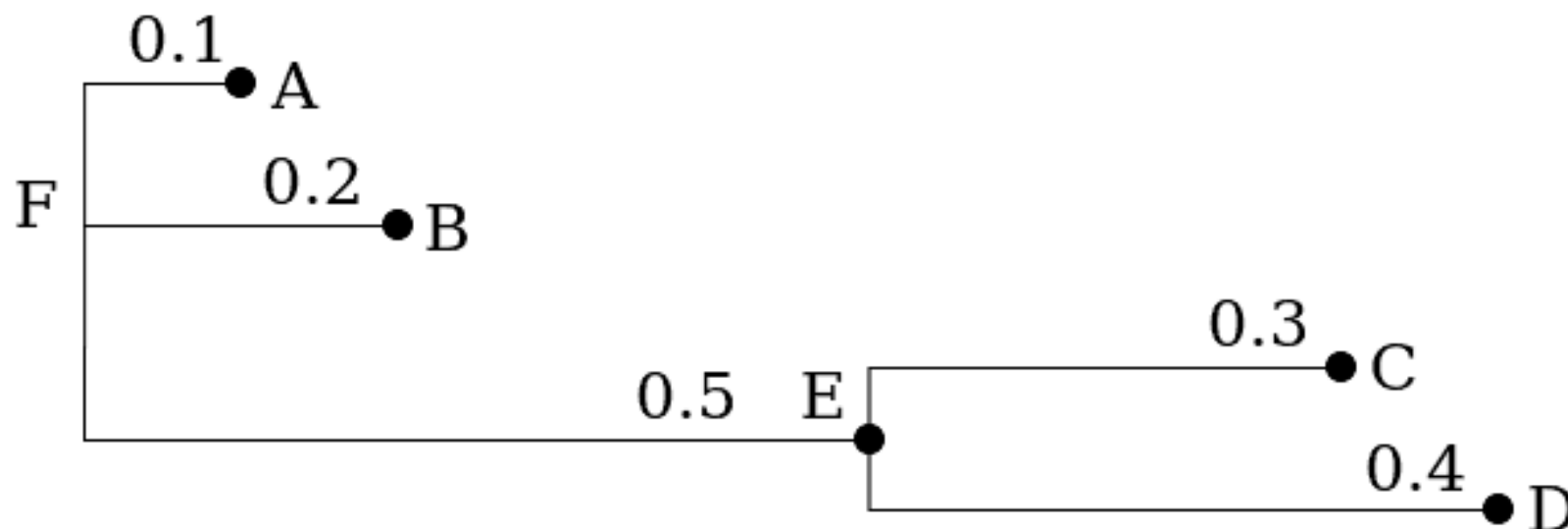
`((B,C),A),(D,E))`



`((B:1,C:2),A:2),(D:1.2,E:2.5))`

# Opis topologii drzewa (2)

The following tree:



could be represented in Newick format in several ways

```
(,,(,));  
(A,B,(C,D));  
(A,B,(C,D)E)F;  
(:0.1,:0.2,(:0.3,:0.4):0.5);  
(:0.1,:0.2,(:0.3,:0.4):0.5):0.0;  
(A:0.1,B:0.2,(C:0.3,D:0.4):0.5);  
(A:0.1,B:0.2,(C:0.3,D:0.4)E:0.5)F;  
((B:0.2,(C:0.3,D:0.4)E:0.5)F:0.1)A;
```

*no nodes are named*

*leaf nodes are named*

*all nodes are named*

*all but root node have a distance to parent*

*all have a distance to parent*

*distances and leaf names (popular)*

*distances and all names*

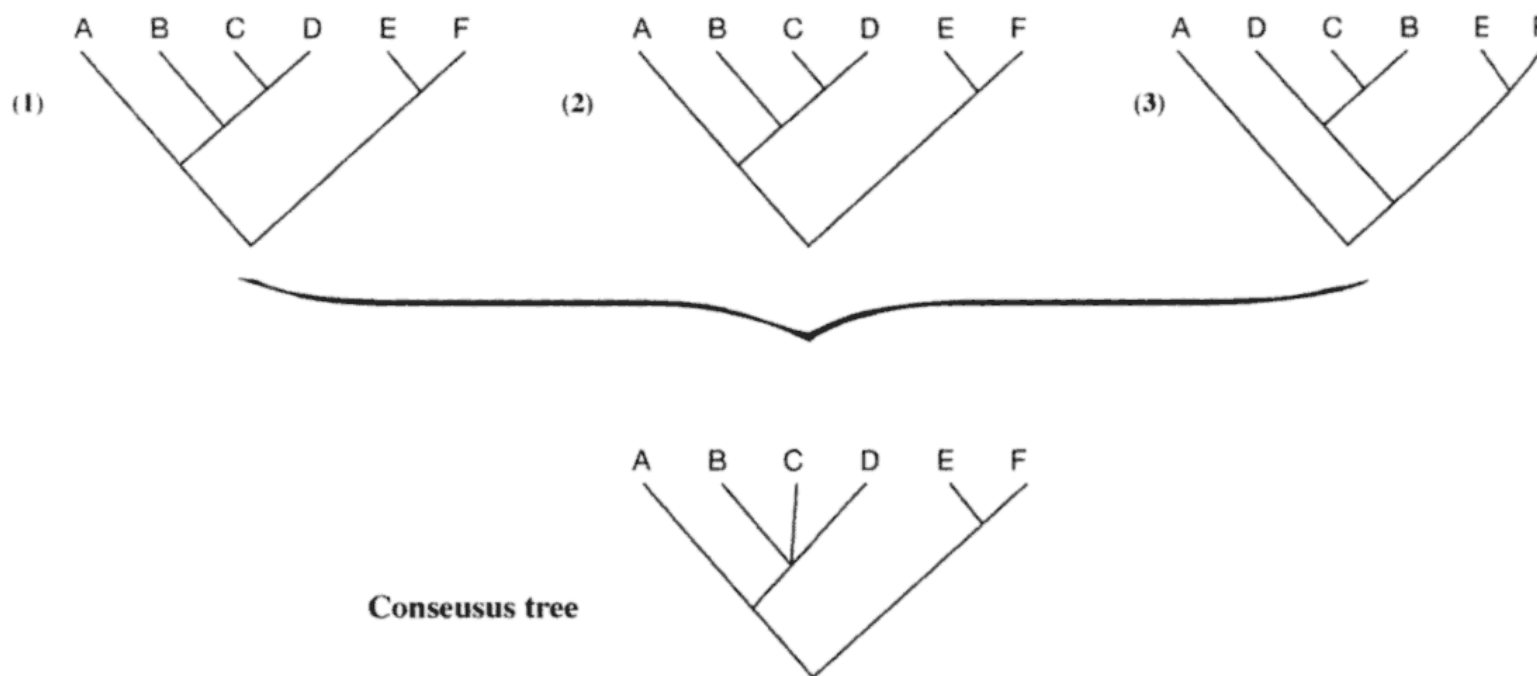
*a tree rooted on a leaf node (rare)*



Tworzone, gdy w wyniku analizy otrzymamy kilka równie dobrych drzew.

Metody:

- ścisły konsensus (węzły niejednoznaczne upraszczamy do politomii)
- reguła większości



Źródło grafiki: Xiong J., Essential bioinformatics

$n$  – liczba taksonów

$N_R$  – liczba drzew ukorzenionych

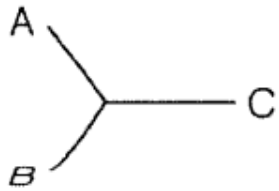
$$N_R = (2n - 3)! / 2^{n-2} (n - 2)!$$

$N_U$  – liczba drzew nieukorzenionych

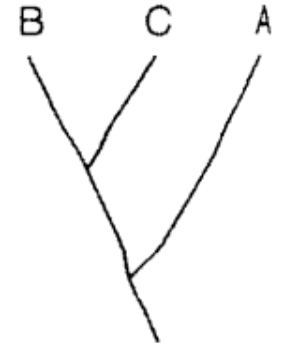
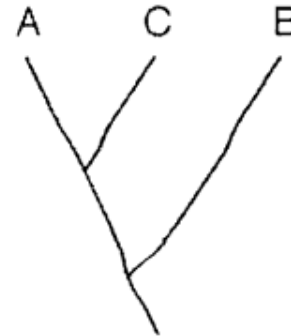
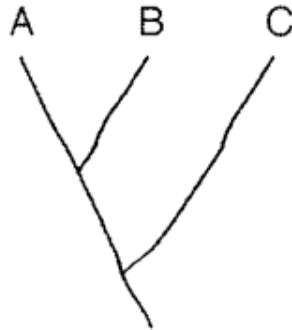
$$N_U = (2n - 5)! / 2^{n-3} (n - 3)!$$

# Możliwe topologie, $n=3$

**Unrooted**

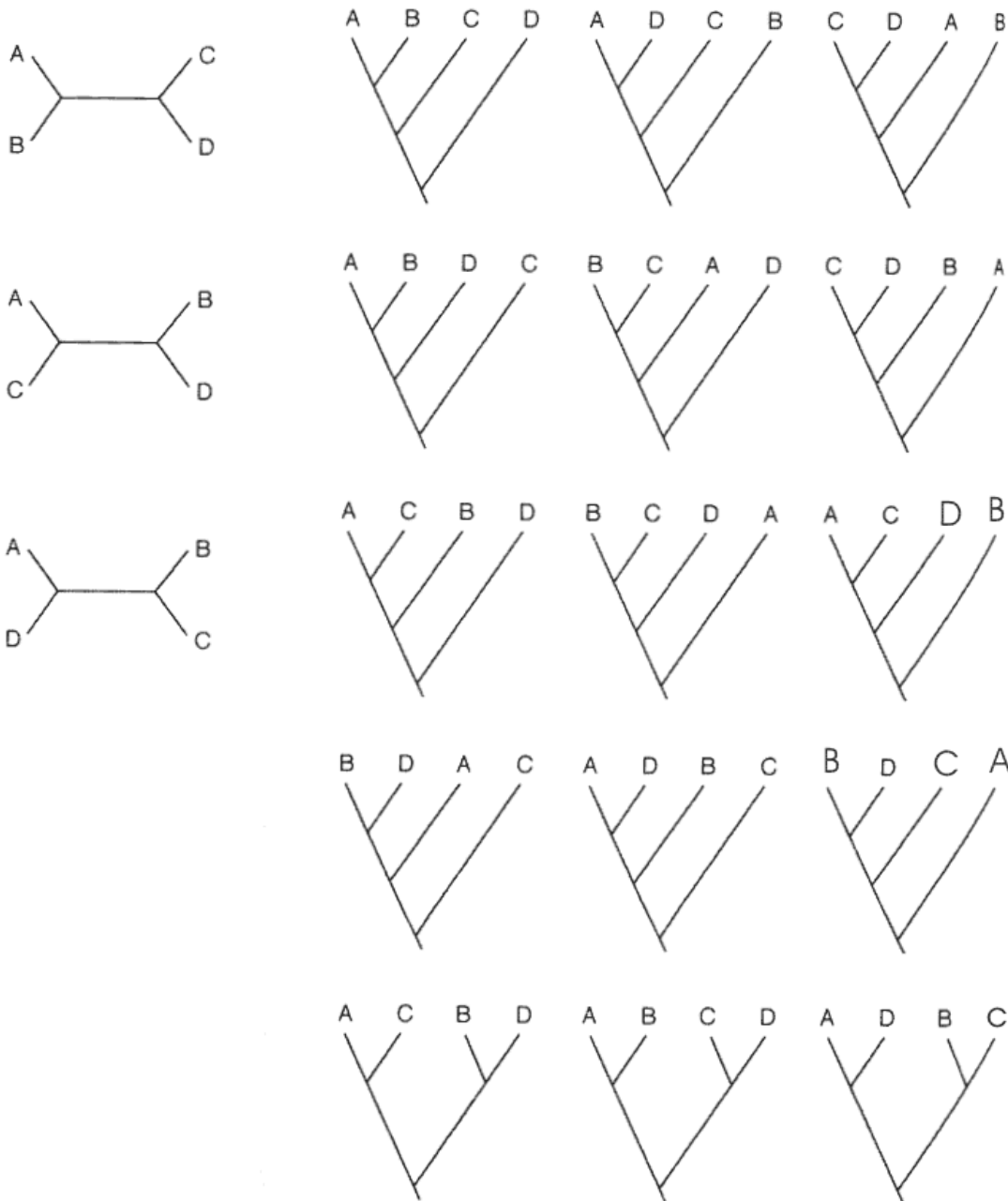


**Rooted**



Źródło grafiki: Xiong J., Essential bioinformatics

# Możliwe topologie, $n=4$



1. Wybór markerów molekularnych (sekwencji)
2. Dopasowanie sekwencji (*multiple sequence alignment*)
3. Wybór modelu ewolucji  
(metoda oceny odległości ewolucyjnej)
4. Konstrukcja drzewa
5. Ocena wiarygodności uzyskanego drzewa



Sekwencje DNA czy białkowe?

DNA ewoluuje szybciej niż białka; analiza DNA ma zastosowanie dla blisko spokrewnionych organizmów, np. z jednej populacji

Zwykle jednak lepiej analizować bardziej konserwowane sekwencje białkowe.

## 2. Dopasowanie sekwencji

- Przyrównujemy wiele sekwencji – brak efektywnego algorytmu dokładnego.
- Błędna konstrukcja przyrównania może prowadzić do kumulacji błędów.
- Zaleca się wykonać kilka niezależnych przyrównań (różnymi metodami) i porównać.
- Pomocą może być wykorzystanie informacji o strukturze drugorzędowej (jeśli jest znana). [np. program Praline]
- Wykorzystać całe przyrównanie czy tylko jego część?

### 3. Wielokrotne substytucje

Np.:

A -> C

A -> T -> G -> C

A -> C -> A -> T -> C -> A -> C

Utrudniają oszacowanie rzeczywistych odległości ewolucyjnych.  
Korekta w oparciu o statystyczne modele substytucji.



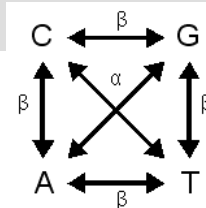
Zakłada, że wszystkie nukleotydy ulegają podstawieniu z jednakowym prawdopodobieństwem.

$d_{AB}$  – odległość ewolucyjna między sekwencjami A i B

$p_{AB}$  – obserwowana odległość sekwencji

$$d_{AB} = -(3/4) \ln[1 - (4/3) p_{AB}]$$

Stosowany dla sekwencji blisko spokrewnionych.



Zakłada, że tempa zachodzenia tranzycji i transwersji są odmienne.

$d_{AB}$  – odległość ewolucyjna między sekwencjami A i B

$p_{ti}$  – obserwowana częstotliwość tranzycji

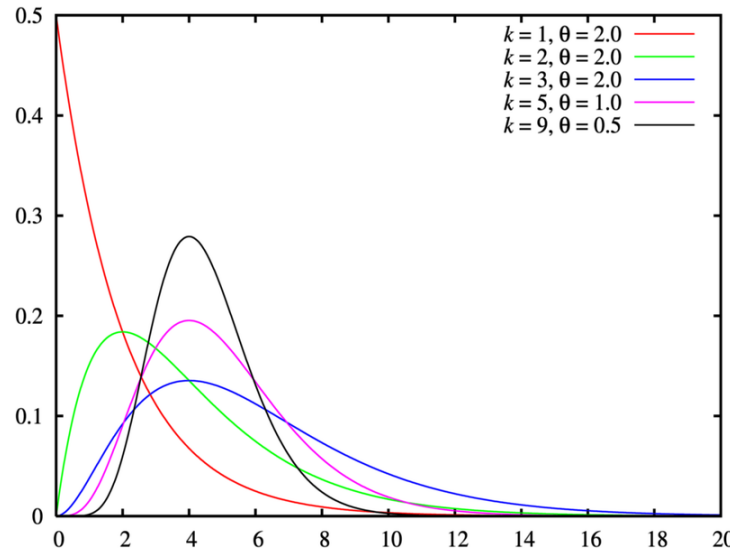
$p_{tv}$  – obserwowana częstotliwość transwersji

$$d_{AB} = -(1/2) \ln(1 - 2p_{ti} - p_{tv}) - (1/4) \ln(1 - 2p_{tv})$$

Istnieją też inne modele, np. TN93, HKY, GTR a także odpowiedniki powyższych opracowane dla sekwencji aminokwasowych.

Dla sekwencji aminokwasowych zazwyczaj stosuje się modele PAM i JTT.

Poszczególne pozycje sekwencji ewoluują w różnym tempie.  
Rozkład miejsc zmiennych odpowiada rozkładowi  $\gamma$



$$\theta^k x^{k-1} \frac{\exp(-x\theta)}{\Gamma(k)}$$

Skorygowane modele substytucji:

$$d_{AB} = (3/4)\alpha[(1 - 4/3 p_{AB})^{-1/\alpha} - 1]$$

$$d_{AB} = (\alpha/2)[1 - 2p_{ti} - p_{tv})^{-1/\alpha} - (1/2)(1 - 2p_{tv})^{-1/\alpha} - 1/2]$$

# Metody konstrukcji drzew

## Metoda obliczeniowa

optymalizacja

analiza klastrow

<ul style="list-style-type: none"><li>• <b>Parsymonia</b></li><li>• <b>Maximum Likelihood</b></li><li>• wnioskowanie Bayesowskie</li></ul>	
<ul style="list-style-type: none"><li>• Minimum Evolution</li><li>• Least Squares</li></ul>	<ul style="list-style-type: none"><li>• <b>UPGMA</b></li><li>• <b>Neighbor-Joining</b></li></ul>

Cechy

Dystanse

Idea:

obliczamy odległości ewolucyjne dla wszystkich par taksonów i konstruujemy macierz odległości.

- a) algorytmy bazujące na klasteryzacji – tworzą drzewo bazując na macierzy odległości, poczynając od najbardziej podobnych sekwencji
- b) algorytmy wykorzystujące kryterium optymalności – porównują wiele alternatywnych topologii drzew

(metoda grupowania nieważonych par z arytmetycznymi średnimi)

1. konstruujemy macierz odległości;
2. grupujemy dwa taksony, których wzajemna odległość jest najmniejsza – w środkowym punkcie między nimi umieszczamy nowy węzeł;
3. tworzymy zredukowaną macierz, gdzie nowy węzeł zastąpił dwa wcześniejsze  
itd.
4. ostatnio dodany takson uznajemy za grupę zewnętrzną i ukorzeniamy drzewo.

- heurystyka: zakładamy, że wszystkie taksony mają stałe tempo ewolucji (to rzadko jest prawdą);  
wszystkie taksony są równoodległe od korzenia
- kolejne etapy analizy oparte są na początkowo wyliczonych wartościach, co może oznaczać utratę ważnych informacji i kumulowanie błędów;
- + metoda intuicyjna i szybka



(metoda łączenia sąsiadów; metoda najbliższego sąsiedztwa)

Metoda podobna do UPGMA, ale nie zakładamy stałego tempa ewolucji.

Odległość nowego węzła od rozważanych taksonów wyznaczamy na podstawie skorygowanych odległości.

$n$  – liczba rozważanych taksonów

$d_{AB}$  – odległość ewolucyjna pomiędzy A i B

$r_i$  – suma odległości i-tego taksonu od wszystkich innych taksonów

$d'_{AB}$  – skorygowana odległość ewolucyjna

$$d'_{AB} = d_{AB} - 1/2 \times (r_A + r_B)$$

$$r_i = \sum d_{ij}$$

$$r'_i = r_i / (n-2)$$

$$d_{AU} = [d_{AB} + (r'_A - r'_B)]/2$$

Przykład na tablicy

Dane wejściowe:

rzeczywiste odległości ewolucyjne pomiędzy badanymi taksonami:

	A	B	C	D
A				
B	0,40			
C	0,35	0,45		
D	0,60	0,70	0,55	

## Algorytm NJ – przykład (2)

	A	B	C	D
A				
B	0,40			
C	0,35	0,45		
D	0,60	0,70	0,55	

$$r_i = \sum d_{ij} \quad r'_i = r_i / (n-2)$$

Metoda NJ jest podobna do UPGMA, ale przed konstrukcją drzewa dokonuje korekcji tempa ewolucji – obliczamy wartości  $r$  i  $r'$ :

$$r_A = AB + AC + AD = 0,4 + 0,35 + 0,6 = 1,35$$

$$r_B = BA + BC + BD = 0,4 + 0,45 + 0,7 = 1,55$$

$$r_C = CA + CB + CD = 0,35 + 0,45 + 0,55 = 1,35$$

$$r_D = DA + DB + DC = 0,6 + 0,7 + 0,55 = 1,85$$

$$r'_A = r_A / (4-2) = 1,35 / 2 = 0,675$$

$$r'_B = r_B / (4-2) = 1,55 / 2 = 0,775$$

$$r'_C = r_C / (4-2) = 1,35 / 2 = 0,675$$

$$r'_D = r_D / (4-2) = 1,85 / 2 = 0,925$$

oraz skorygowane odległości:

$$d'_{AB} = d_{AB} - 1/2 \times (r_A + r_B)$$

$$d'_{AB} = d_{AB} - 1/2 (r_A + r_B) = 0,4 - (1,35 + 1,55) / 2 = -1,05$$

$$d'_{AC} = d_{AC} - 1/2 (r_A + r_C) = 0,35 - (1,35 + 1,35) / 2 = -1$$

$$d'_{AD} = d_{AD} - 1/2 (r_A + r_D) = 0,6 - (1,35 + 1,85) / 2 = -1$$

$$d'_{BC} = d_{BC} - 1/2 (r_B + r_C) = 0,45 - (1,55 + 1,35) / 2 = -1$$

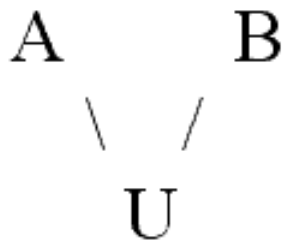
$$d'_{BD} = d_{BD} - 1/2 (r_B + r_D) = 0,7 - (1,55 + 1,85) / 2 = -1$$

$$d'_{CD} = d_{CD} - 1/2 (r_C + r_D) = 0,55 - (1,35 + 1,85) / 2 = -1,05$$

Skorygowane względem tempa ewolucji odległości pozwalają skonstruować nową macierz:

	A	B	C	D
A				
B	-1,05			
C	-1	-1		
D	-1	-1	-1,05	

Wybieramy parę taksonów z najkrótszymi skorygowanymi odległościami. W naszym przypadku mamy dwie takie pary: AB i CD. Wybieramy dowolną z nich, np. AB. Te dwa taksony łączymy w drzewie w węzeł U.



Obliczamy długości gałęzi od A i B do węzła U:

$$d_{AU} = [d_{AB} + (r_A' - r_B')]/2 = [0,4 + (0,675 - 0,775)]/2 = 0,15$$

$$d_{BU} = [d_{AB} + (r_B' - r_A')]/2 = [0,4 + (0,775 - 0,675)]/2 = 0,25$$

Punktem wyjścia do konstrukcji zredukowanej macierzy są rzeczywiste odległości:

$$d_{CU} = (d_{AC} + d_{BC} - d_{AB})/2 = (0,35 + 0,45 - 0,4)/2 = 0,2$$

$$d_{DU} = (d_{AD} + d_{BD} - d_{AB})/2 = (0,6 + 0,7 - 0,4)/2 = 0,45$$

	U	C	D
U			
C	0,20		
D	0,45	0,55	



Na podstawie zredukowanej macierzy oblicza się nowy zestaw wartości  $r$  i  $r'$  i konstruuje skorygowaną macierz odległości:

$$r_C = CU + CD = 0,2 + 0,55 = 0,75$$

$$r_D = DU + CD = 1$$

$$r_U = CU + DU = 0,65$$

$$r_C' = r_C / (3-2) = 0,75$$

$$r_D' = r_D / (3-2) = 1$$

$$r_U' = r_U / (3-2) = 0,65$$

$$d'_{CU} = d_{CU} - \frac{1}{2} (r_C + r_U) = 0,2 - (0,75+0,65)/2 = -0,5$$

$$d'_{DU} = -0,375$$

$$d'_{CD} = -0,325$$

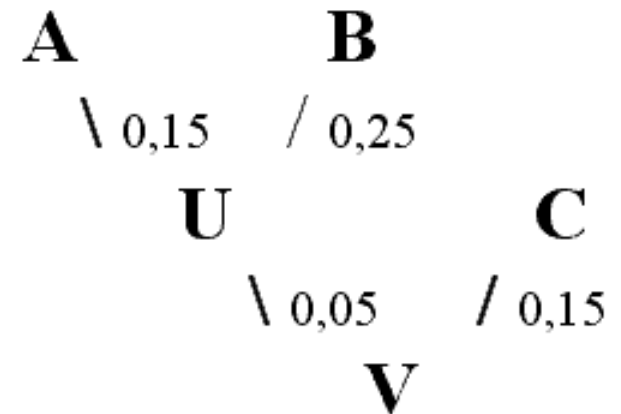
Macierz skorygowana:

	U	C	D
U			
C	-0,5		
D	-0,375	-0,325	

W macierzy tej najkrótsza jest odległość CU i te dwa węzły łączymy w nowy węzeł V i, analogicznie jak poprzednio, obliczamy długości gałęzi:

$$d_{CV} = 0,15$$

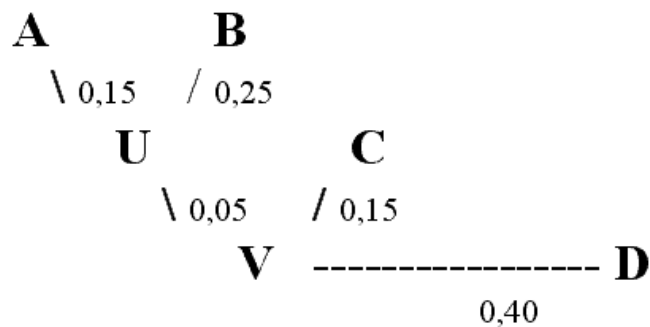
$$d_{UV} = 0,05$$



Ponieważ została nam już tylko jedna para węzłów (V i D), wiadomo, że do drzewa włączona będzie gałąź VD. Nie ma więc potrzeby obliczania kolejnych macierzy. Pozostaje jedynie obliczyć długość gałęzi VD:

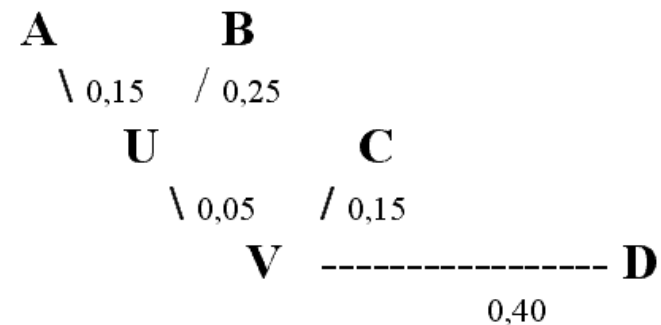
$$d_{VD} = (d_{DU} + d_{DC} - d_{CU})/2 = (0,45 + 0,55 - 0,2) = 0,4.$$

Ostatecznie uzyskujemy drzewo (nieukorzenione lub z korzeniem w D – jeśli takson D stanowił grupę zewnętrzną):



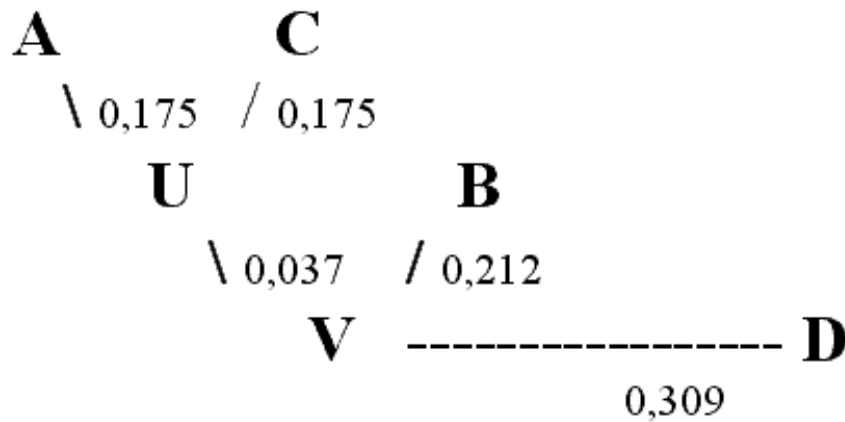
Sumując odległości pomiędzy węzłami możemy stwierdzić, że wyznaczone metodą NJ odległości odpowiadają rzeczywistym odległościom występującym w macierzy wejściowej.

	A	B	C	D
A				
B	0,40			
C	0,35	0,45		
D	0,60	0,70	0,55	



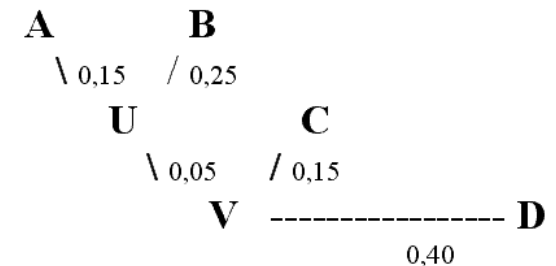
# Algorytm NJ – przykład (11), porównanie z UPGMA

Dla porównania, metoda UPGMA nie zachowuje odległości. Zastosowanie UPGMA dla tych samych danych wejściowych generuje następujące drzewo i macierz odległości zmierzonych na drzewie:



	A	B	C	D
A				
B	0,42			
C	0,35	0,42		
D	0,62	0,62	0,62	

## Metoda NJ:



	A	B	C	D
A				
B	0,40			
C	0,35	0,45		
D	0,60	0,70	0,55	

Umożliwia konstrukcję większej liczby drzew, co pozwala zredukować błędy popełniane w początkowych etapach analizy.

Przegląd wszystkich możliwych topologii drzewa i wybór tej, która minimalizuje kwadrat odchyłeń rzeczywistych odległości ewolucyjnych od wyznaczonych na podstawie drzewa.

Kryterium optymalności:

$$E = \sum_{i=1}^{T-1} \sum_{j=j+1}^T \frac{(d_{ij} - p_{ij})^2}{d_{ij}^2}$$

Szukamy drzewa o minimalnej łącznej długości gałęzi.

Kryterium optymalności:

$$S = \sum b_i$$

$b_i$  – długość i-tej gałęzi





Opierają się bezpośrednio na symbolach w sekwencji, a nie na odległościach.

Zalety:

- możliwość badania dynamiki ewolucyjnej dla każdego znaku (wyznaczenie przypuszczalnej sekwencji przodka)

Zakłada, że najlepszym rozwiązaniem jest to najprostsze, czyli takie drzewo, które wymaga najmniej zmian (substytucji).

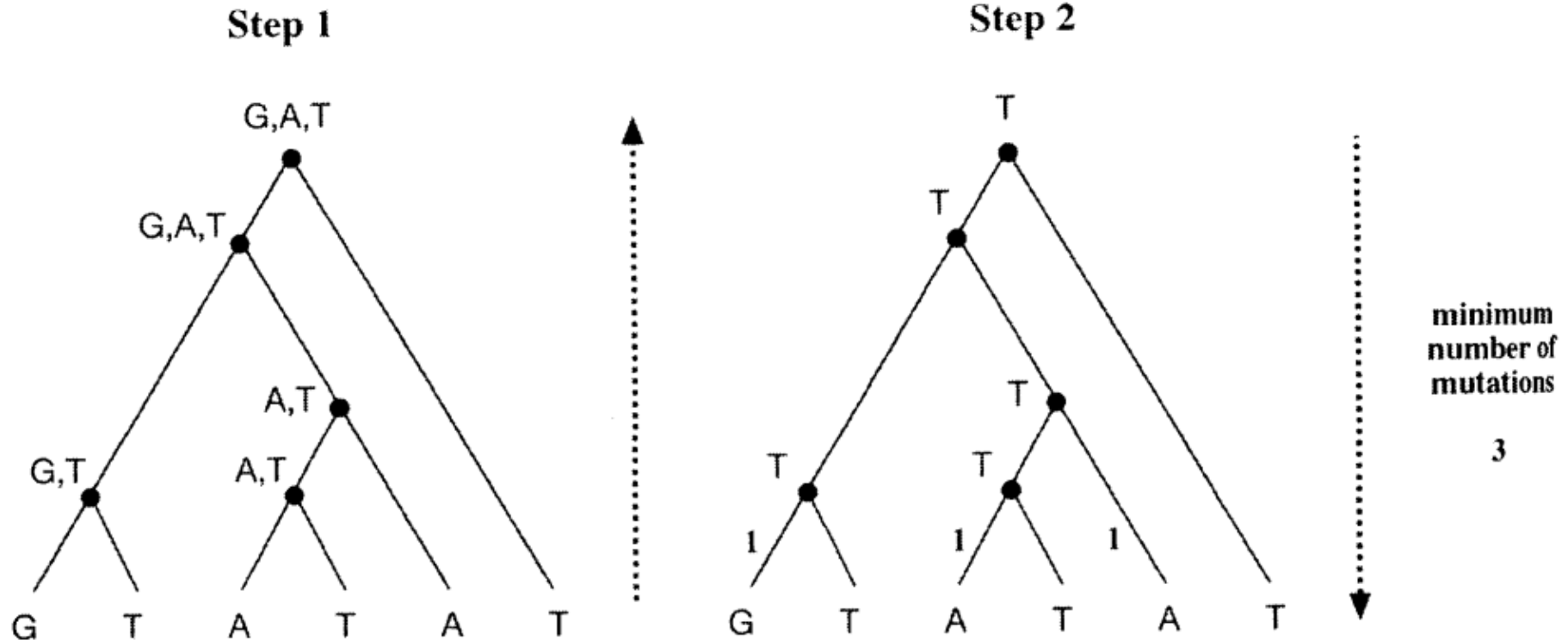
Analizowane są jedynie miejsca w których sekwencje się różnią, pozostałe pozycje są usuwane i nie są dalej wykorzystywane

(założenie to ogranicza w znacznym stopniu zastosowanie tej metody – z powodu występowania wstecznych substytucji przy wysokim poziomie dywergencji sekwencji metoda ta nie jest w stanie określić prawidłowej topologii drzewa).

pozycje, w których występują przynajmniej dwa rodzaje znaków, każdy przynajmniej dwa razy

taxa \ sites	sites							
	1	2	3	4	5	6	7	8
I	A	A	T	T	A	G	C	T
II	G	G	T	C	G	T	A	G
III	A	A	T	G	C	G	C	T
IV	A	G	T	A	A	G	C	A
V	A	C	T	T	C	G	C	G
VI	A	C	A	T	G	G	C	A

1. określenie wszystkich możliwych pierwotnych znaków na węzłach wewnętrznych;
2. identyfikacja tych, które wymagają minimalnej liczby mutacji

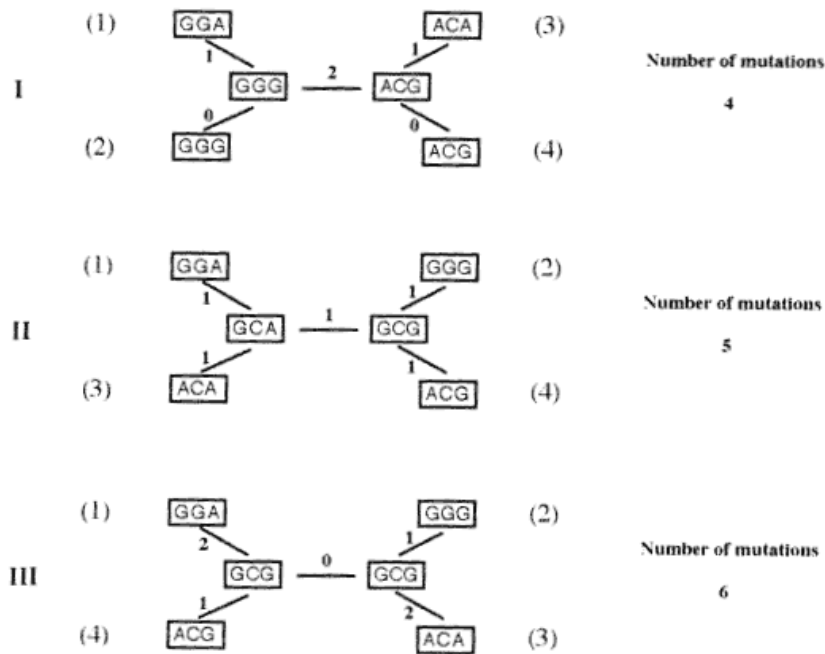


Zastosowane wagi:

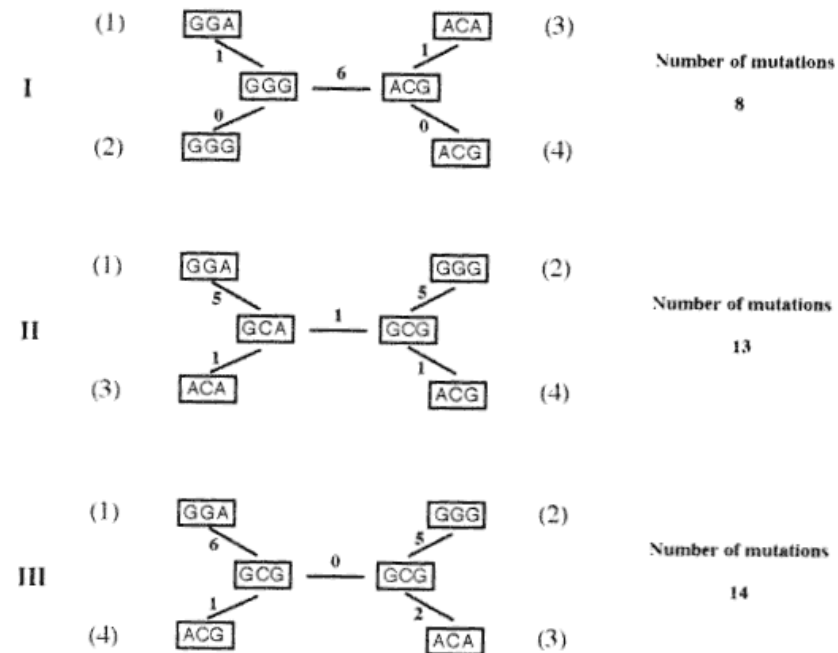
tranzycja: 1

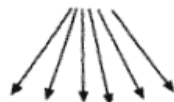
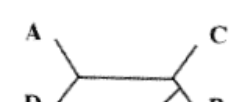
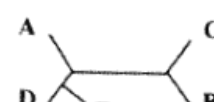
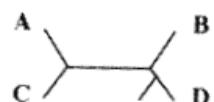
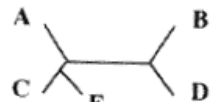
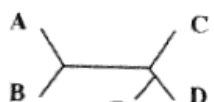
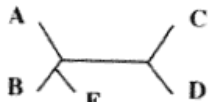
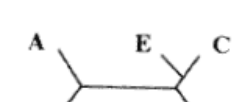
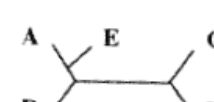
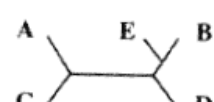
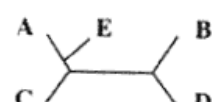
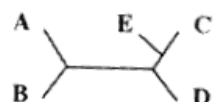
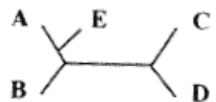
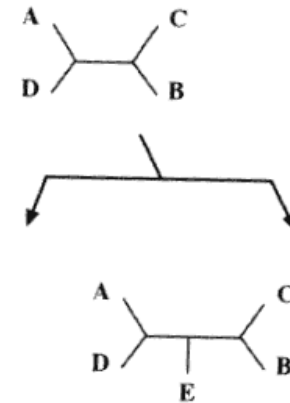
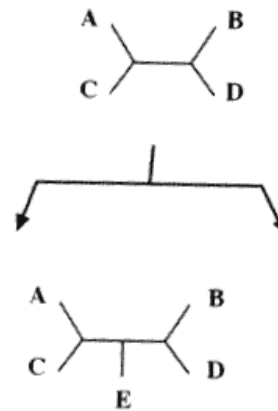
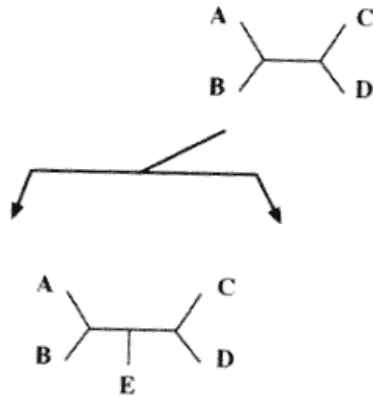
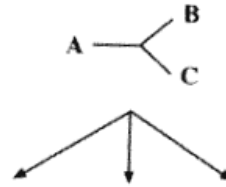
transwersja: 5

## Unweighted parsimony



## Weighted parsimony





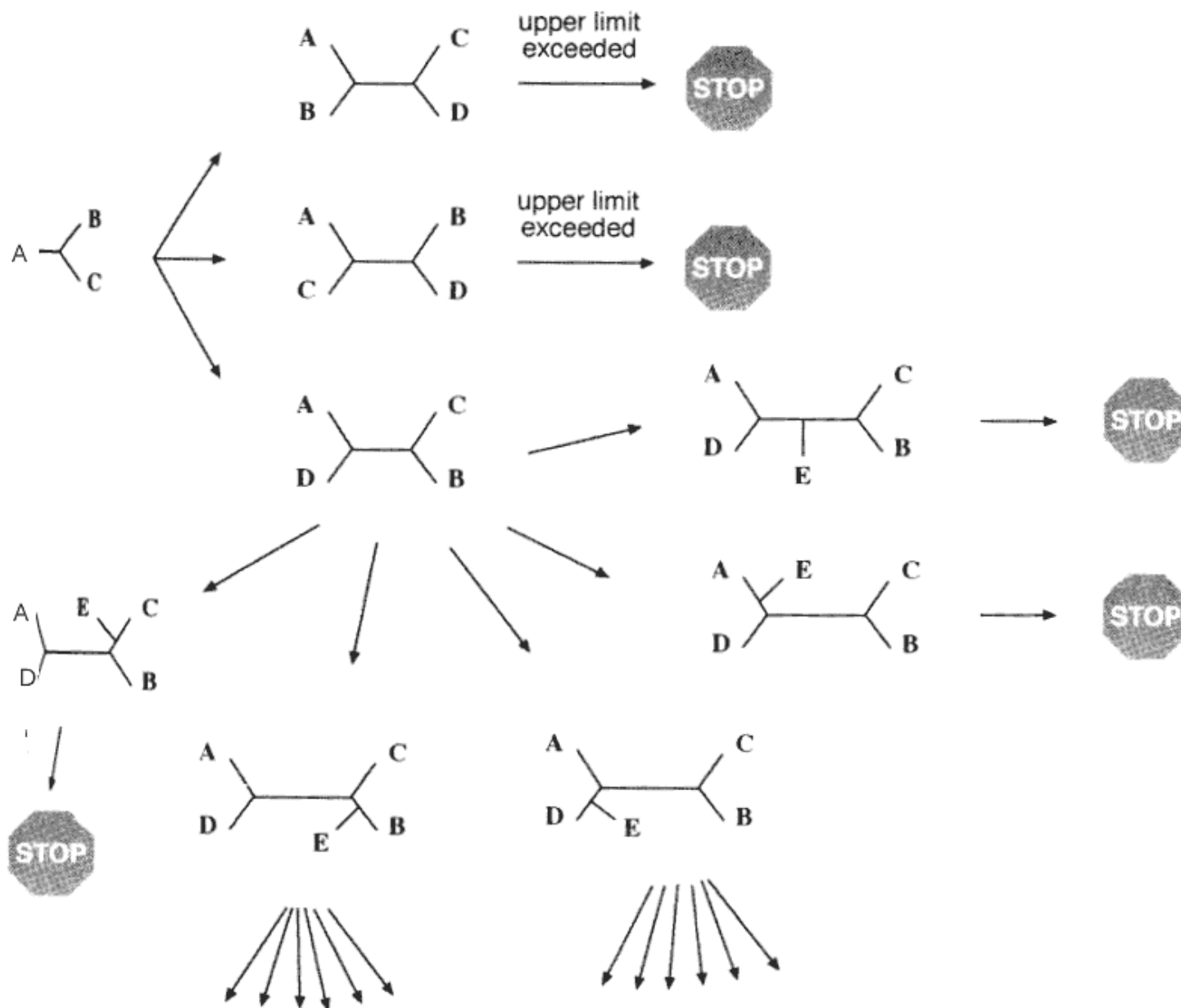
Przeglądanie wszystkich możliwych topologii.

Ograniczenie liczby przeszukiwanych topologii do zadanej z góry liczby zmian w sekwencji.

Wartość ograniczenia można uzyskać na podstawie wyniku szybkiej analizy UPGMA lub NJ.

Praktyczne zastosowanie – do ok. 20 analizowanych taksonów.

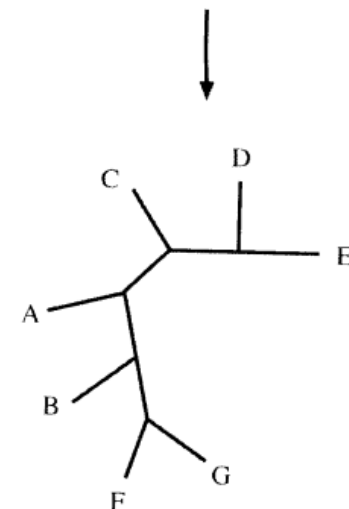
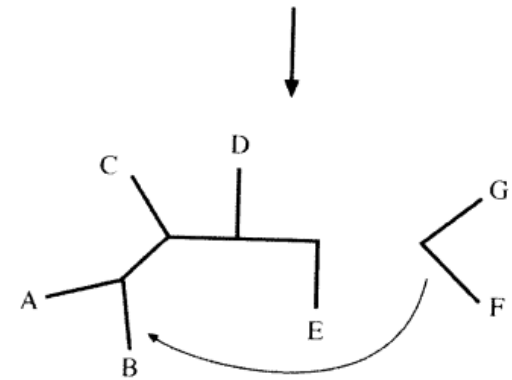
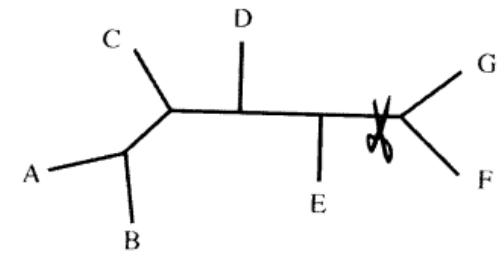
# MP – *branch and bound*





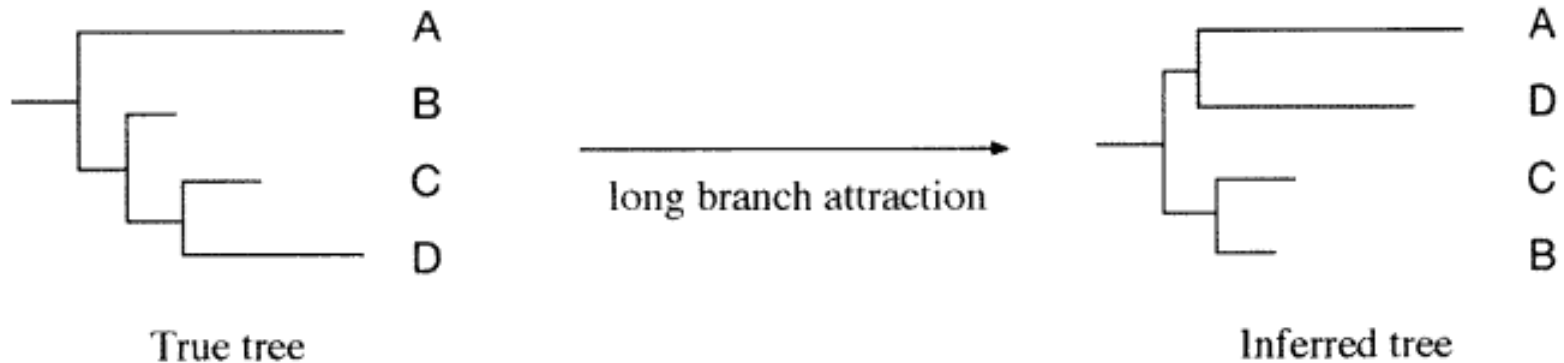
Idea:

Szybka budowa wstępnego przybliżenia (np. metodą NJ) i jego optymalizacja (np. metodą zamiany gałęzi).



Ryzyko utknięcia w minimum globalnym.

Problem przyciągania się długich gałęzi (long-branch attraction) – grupowanie szybko ewoluujących taksonów.



Oblicza prawdopodobieństwa ewolucji sekwencji pierwotnych do węzłów wewnętrznych i do sekwencji badanych.

Wykorzystuje model substytucji.

Analizuje wszystkie znaki w sekwencji.

Metoda dokładna, lecz czasochłonna.

Możliwe uproszczenia:

- metoda kwartetów
- hybryda NJML
- algorytm genetyczny (GA)

AWTY (*Are We There Yet?*)

[http://king2.sc.fsu.edu/CEBProjects/awty/awty\\_start.php](http://king2.sc.fsu.edu/CEBProjects/awty/awty_start.php)

Phylip <http://evolution.genetics.washington>

MrBayes <http://mrbayes.sourceforge.net/>

MacClade <http://macclade.org/index.html>

## Przykład:

```
#NEXUS Begin data;  
Dimensions ntax=4 nchar=15;  
Format datatype=dna symbols="ACTG" missing=? gap=-; Matrix  
Species1 atgctagctagctcg  
Species2 atgcta??tag-tag  
Species3 atgttagctag-tgg  
Species4 atgttagctag-tag ;  
End;
```

## Opis i przykłady:

[http://wiki.christophchamp.com/index.php/NEXUS\\_file\\_format](http://wiki.christophchamp.com/index.php/NEXUS_file_format)  
<http://nexml.org/>

## Narzędzia do konwersji formatów:

<http://www.bugaco.com/bioinf/nexusfasta.php>  
[http://phylosoft.org/forester/applications/phyloxml\\_converter/](http://phylosoft.org/forester/applications/phyloxml_converter/)

# Ocena wiarygodności drzewa

- test bootstrap (konstrukcja drzew dla zaburzonych zbiorów danych)
- test jackknife (konstrukcja drzew dla okrojonych zbiorów danych)
- symulacja bayesowska (MCMC)

- test Kishino-Hasegawy (dla drzew uzyskanych metodą MP)
- test Shimodairo-Hasegawy (dla drzew uzyskanych metodą ML)



