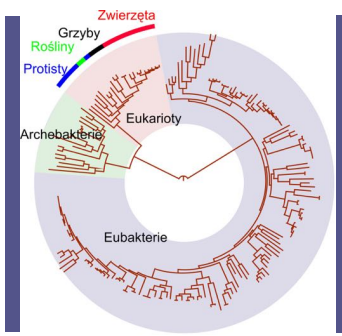


Filogenetyka molekularna I

Krzysztof Spalik
Zakład Filogenetyki Molekularnej i Ewolucji

2

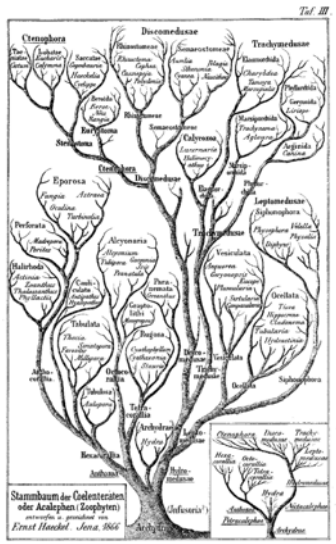
Literatura

- Krzysztof Spalik, Marcin Piwczyński (2009), Rekonstrukcja filogenezy i wnioskowanie filogenetyczne w badaniach ewolucyjnych, Kosmos 58(3-4): 485-498
- John C. Avise (2008), Markery molekularne, historia naturalna i ewolucja, Wydawnictwa Uniwersytetu Warszawskiego, Warszawa

3

Filogeneza a ewolucja

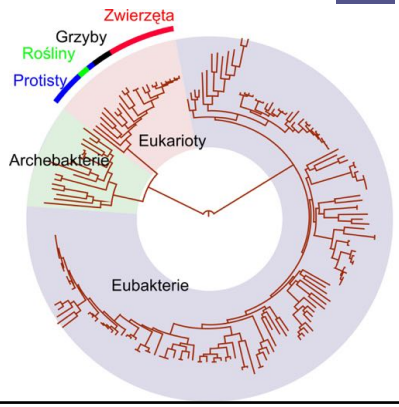
- Od Darwina podstawą rozważań ewolucyjnych stały się drzewa rodowe (jak przedstawione obok drzewo Haeckla)
- Drzewa te jednak tworzono intuicyjnie, a nie za pomocą metod formalnych



4

Drzewo jako hipoteza naukowa

- Drzewo zaproponowane przez badacza odzwierciedla jego wiedzę i poglądy, ale trudno je zweryfikować
- Drzewo uzyskane za pomocą metod formalnych (matematycznych) poddaje się analizie i weryfikacji



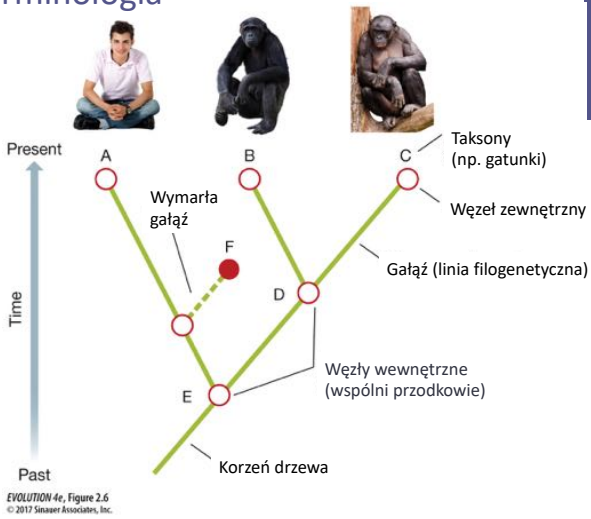
5

Po co nam drzewa rodowe?

- Aby stworzyć hierarchiczny system klasyfikacji organizmów – czyli ich katalog, umożliwiający dostęp do zgromadzonej wiedzy o nich
- Aby badać przebieg ewolucji, np. dokonywać rekonstrukcji zmian cech w czasie

Terminologia

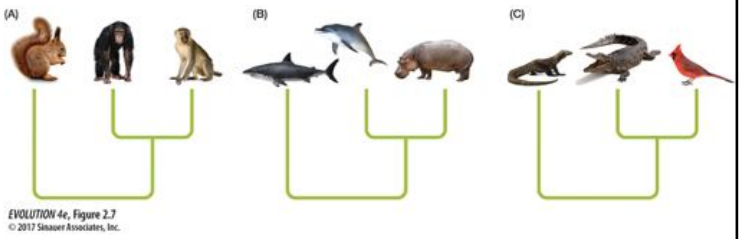
6



Podobieństwo a pokrewieństwo

7

- Podobieństwo organizmów może oddawać ich ewolucyjne pokrewieństwo (A), ale może także wynikać z ewolucji zbieżnej, czyli **konwergencji** (B), albo być świadectwem odległej przeszłości ewolucyjnej – dziedzictwem dawnego wspólnego przodka (C).



8

Filogenetyka (kladystyka)

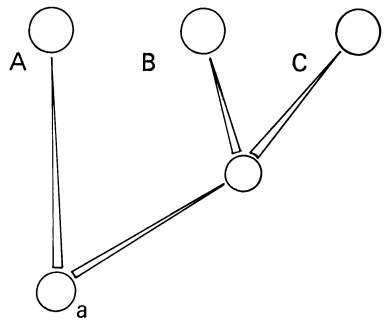
http://en.wikipedia.org/wiki/File:Willi_Hennig.jpg

- Twórcą był niemiecki entomolog Willi Hennig (1913-1976)
- Kladystyka polega na szacowaniu liczby zmian na gałęziach drzewa
- Poszukuje wspólnych cech zaawansowanych ewolucyjnie, identyfikujących gałęzie drzewa filogenetycznego
- Znajduje zastosowanie do wszystkich danych nieciągłych (nie tylko sekwencji)

Ewolucja cech

9

- Rozważamy ewolucję cechy u trzech spokrewnionych gatunków. U ich wspólnego przodka występowała cecha ***a***, natomiast u jego potomków ***a*** lub ***a'***.
- Innymi słowy, ***a*** jest cechą pierwotną, natomiast ***a'*** – pochodną ewolucyjnie



Stany pierwotne i pochodne

10

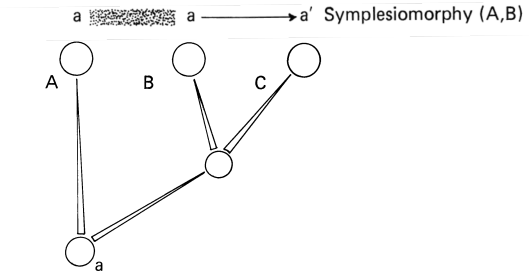
- Stan pierwotny ewolucyjnie to **plezjomorfa**
- Stan pochodny (wtórny ewolucyjnie) to **apomorfa**


Długa szyja plezjozaura to apomorfa – cecha zaawansowana ewolucyjnie

Symplezjomorfia

11

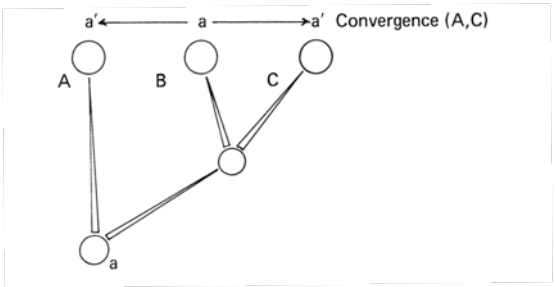
- Symplezjomorfia, czyli posiadanie tej samej cechy pierwotnej ewolucyjnie, nie świadczy o bliskim pokrewieństwie



Konwergencja

12

- Konwergencja, czyli niezależne uzyskanie w toku ewolucji tej samej cechy (wtórnej ewolucyjnie), nie świadczy o bliskim pokrewieństwie



13

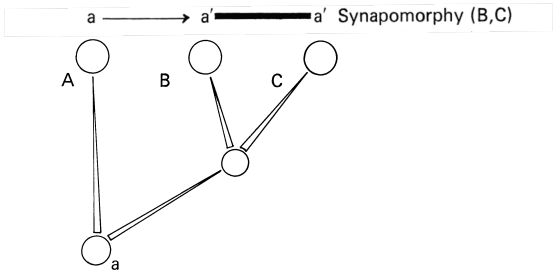
Homoplazje

- Symplezjomorfie i konwergencje, czyli podobieństwa nieświadczące o wspólnym pochodzeniu, określane są łącznie jako homoplazje
- Utrudniają one odtworzenie rzeczywistej filogenezy organizmów

14

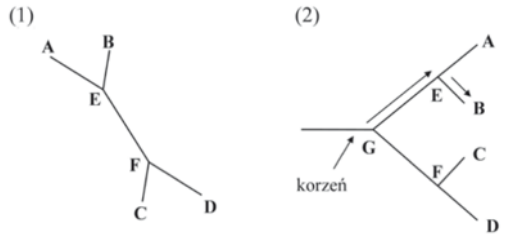
Synapomorfia

- O bliskim pokrewieństwie świadczy jedynie posiadanie tej samej cechy zaawansowanej (wtórnej) ewolucyjnie, czyli synapomorfia



15

Zakorzenianie drzewa filogenetycznego



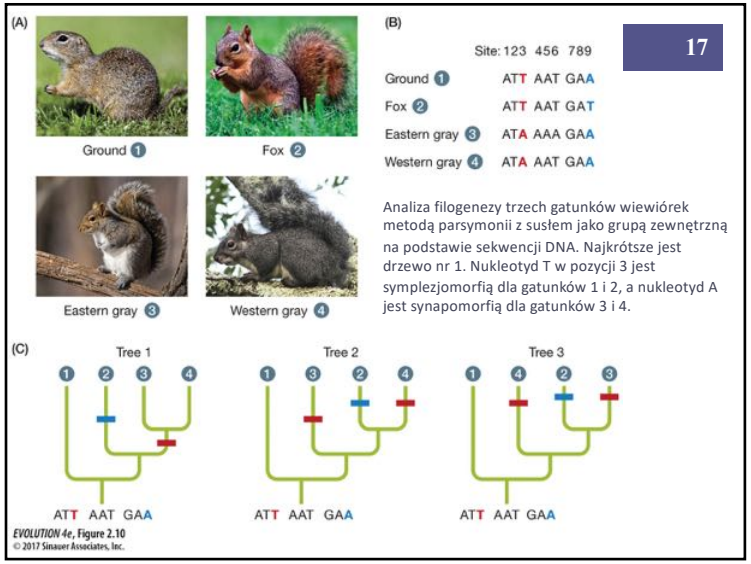
- Oba drzewa mają identyczną topologię, ale dzięki ukorzenieniu drzewa w węźle G znamy kierunek ewolucji

16

Zasada parsymonii

- Spośród wielu możliwych drzew filogenetycznych wybieramy to, które tłumaczy różnorodność cech u badanych organizmów w najbardziej oszczędny sposób, czyli za pomocą najmniejszej liczby zmian ewolucyjnych
- Metoda największej parsymonii (*maximum parsimony*) dostarcza zatem jedynie kryterium wyboru drzewa

17

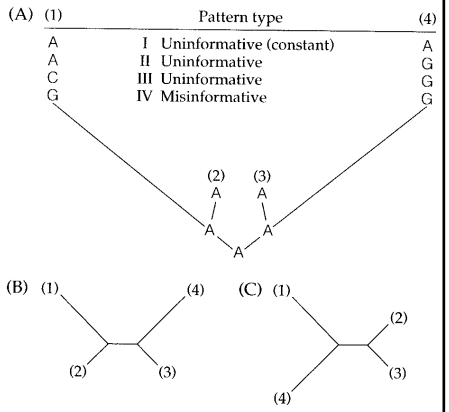


Analiza filogenezy trzech gatunków wiewiórek metodą parsymonii z susłem jako grupą zewnętrzną na podstawie sekwencji DNA. Najkrótsze jest drzewo nr 1. Nukleotyd T w pozycji 3 jest symplezjomorfią dla gatunków 1 i 2, a nukleotyd A jest synapomorfią dla gatunków 3 i 4.

Strefa Felsensteina, czyli kłopoty z parsymonią

18

- Przy nierównym tempie ewolucji w różnych gałęziach i krótkim czasie między radiacjami, metoda parsymonii daje błędne oszacowania
- Zbiór drzew rodowych trudnych do oszacowania za pomocą parsymonii nazywamy „strefą Felsensteina", sam efekt zaś – „przyciąganiem się długich gałęzi"

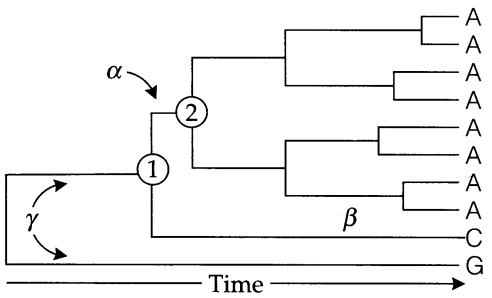


19

Spójność metody szacowania

- Modele podstawiania (substytucji) nukleotydów w DNA umożliwiają oszacowanie liczby wielokrotnych podstawień
- Metody uwzględniające długość gałęzi są spójne, o ile nie ma systematycznego błędu w oszacowaniu długości gałęzi

20



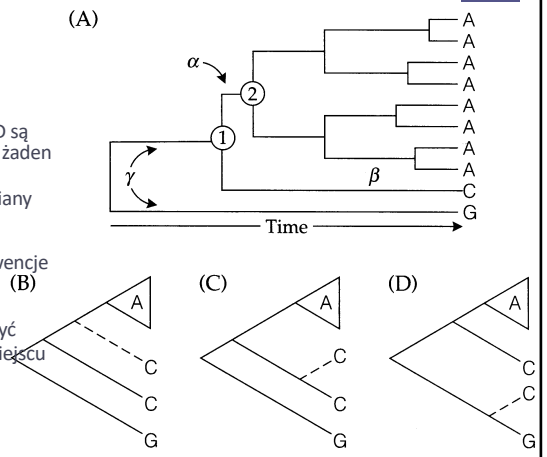
Parsymonia – liczą się tylko cechy wspólne

Zgodnie z zasadą parsymonii, przodek nr 2 miał adeninę w danym miejscu sekwencji, natomiast przodek nr 1 – adeninę, cytozynę lub guaninę

Parsymonia

21

- Jeśli dołączamy sekwencję z cytozyną, to wszystkie trzy warianty B, C i D są równie dobre – żaden nie wymaga dodatkowej zmiany na drzewie
- Natomiast sekwencje z tyminą lub adeniną możemy dołączyć w dowolnym miejscu



Metoda największej wiarygodności

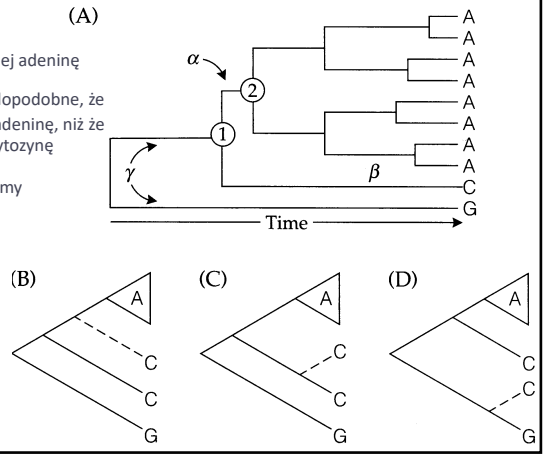
22

- Metoda największej wiarygodności czyni założenia co do tempa podstawiania nukleotydów, typów podstawień i różnic tempa podstawień w różnych gałęziach drzewa i w różnych odcinkach sekwencji
- Najlepsze drzewo to takie, dla którego **najbardziej prawdopodobne jest uzyskanie obserwowanego rozkładu cech** (nukleotydów w sekwencjach)

Metoda największej wiarygodności – liczy się prawdopodobieństwo

23

- Przodek nr 2 miał najprawdopodobniej adeninę
- Jest bardziej prawdopodobne, że przodek nr 1 miał adeninę, niż że miał guaninę lub cytozynę
- Jeśli zatem dołączamy sekwencję z cytozyną, to najbardziej prawdopodobne jest drzewo C



Model ewolucji sekwencji

24

- Model ewolucji sekwencji to macierz tempa podstawień nukleotydów
- Najbardziej ogólny model ma następujące parametry
  - μ: bezpośrednie tempo podstawień
  - a-l: stałe dla każdego typu podstawienia (razem 12)
  - π: frekwencja danego nukleotydu
- Prawie wszystkie modele używane w analizie są szczególnymi przypadkami tego modelu

$$\mathbf{Q} = \begin{pmatrix} -\mu(a\pi_C + b\pi_G + c\pi_T) & \mu a\pi_C & \mu b\pi_G & \mu c\pi_T \\ \mu g\pi_A & -\mu(g\pi_A + d\pi_G + e\pi_T) & \mu d\pi_G & \mu e\pi_T \\ \mu h\pi_A & \mu j\pi_C & -\mu(h\pi_A + j\pi_C + f\pi_T) & \mu f\pi_T \\ \mu i\pi_A & \mu k\pi_C & \mu l\pi_G & -\mu(i\pi_A + k\pi_C + l\pi_G) \end{pmatrix}$$

25

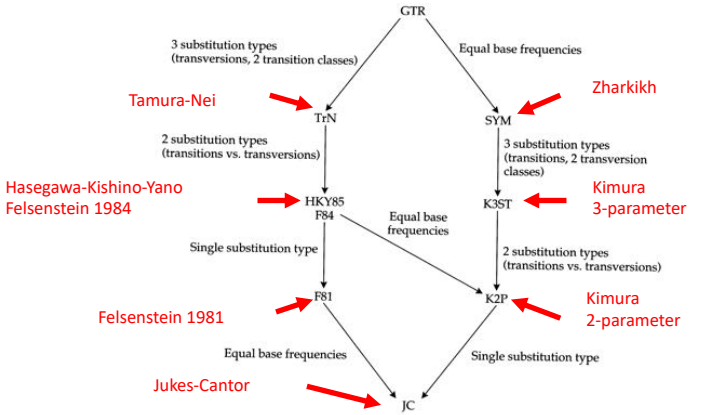
Model GTR

- Model GTR (general time-reversible) zakłada pełną odwracalność podstawień. Innymi słowy, tempo podstawienia np. tyminy przez adeninę jest identyczne jak adeniny przez tyminę. Zamiast 12 typów podstawień mamy zatem 6 (macierz jest symetryczna wzdłuż przekątnej)

$$\mathbf{Q} = \begin{pmatrix} -\mu(a\pi_C + b\pi_G + c\pi_T) & \mu a\pi_C & \mu b\pi_G & \mu c\pi_T \\ \mu a\pi_A & -\mu(a\pi_A + d\pi_G + e\pi_T) & \mu d\pi_G & \mu e\pi_T \\ \mu b\pi_A & \mu d\pi_C & -\mu(b\pi_A + d\pi_C + f\pi_T) & \mu f\pi_T \\ \mu c\pi_A & \mu e\pi_C & \mu f\pi_G & -\mu(c\pi_A + e\pi_C + f\pi_G) \end{pmatrix}$$

26

Najpowszechniej używane modele są szczególnymi przypadkami GTR



27

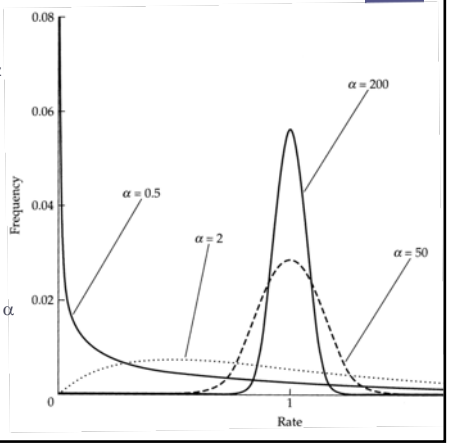
Problemy z niespójnością

- Metoda parsymonii może być niespójna, zwłaszcza w wypadku nierównego tempa ewolucji w poszczególnych gałęziach drzewa rodowego (efekt przyciągania długich gałęzi
- Metody bazujące na modelach substytucji DNA mogą być niespójne, jeśli różne miejsca sekwencji ewoluują w różnym tempie (a nie jest to uwzględnione w modelu)

28

Heterogeniczność tempa podstawień

- Jeśli tempo podstawień jest heterogeniczne, to metoda największej wiarygodności oraz metody odległościowe są niespójne
- Rozwiązaniem jest uwzględnienie różnic tempa podstawiania pozycji
- Zwykle przyjmuje się, że mają one tzw. rozkład γ, charakteryzowany przez współczynnik kształtu rozkładu α



29

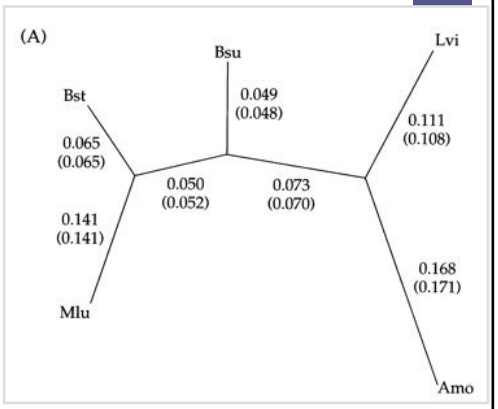
Metody odległościowe

- Bazują najczęściej (ale nie zawsze) na modelach substytucji DNA (np. JC, K2, K3, HKY, F84 itd.)
- Podobnie jak metoda największej wiarygodności, są spójne w tych wypadkach, gdzie parsymonia zawodzi
- Są szybką alternatywą dla metody największej wiarygodności, ale są wrażliwe na wpadnięcie w optimum lokalne

30

Łączenie sąsiadów

- Metoda łączenia sąsiadów (neighbour-joining) bierze pod uwagę zarówno odległość między łączonymi obiektami, jak i ich odległość do wszystkich pozostałych
- Jest metodą spójną. Dla dobrych danych, jej wyniki nie odbiegają znacząco od parsymonii, największej wiarygodności i metody bayesowskiej



31

Twierdzenie Bayesa (prawdopodobieństwo warunkowe)

- P(A) – prawdopodobieństwo zajścia zdarzenia A
- P(H) – prawdopodobieństwo hipotezy H
- P(A|H) – prawdopodobieństwo A przy założeniu, że H jest prawdziwe
- P(H|A) – prawdopodobieństwo H przy założeniu, że zaszło A
- $P(H|A) = P(A|H)P(H) / P(A)$

32

Metoda bayesowska

- Bazuje na twierdzeniu Bayesa (dotyczącym prawdopodobieństwa warunkowego) – w tym wypadku pytamy o prawdopodobieństwo drzewa filogenetycznego przy danym zestawie przyrównanych sekwencji
- Wykorzystuje się nieinformacyjne prawdopodobieństwa a priori dla stawianych hipotez
- W przeszukiwaniu przestrzeni wszystkich możliwych drzew stosuje się zwykle łańcuchy Markowa-Monte Carlo
- Do wyznaczenia prawdopodobieństwa drzew i gałęzi służy rozkład stacjonarny łańcucha MMC

33

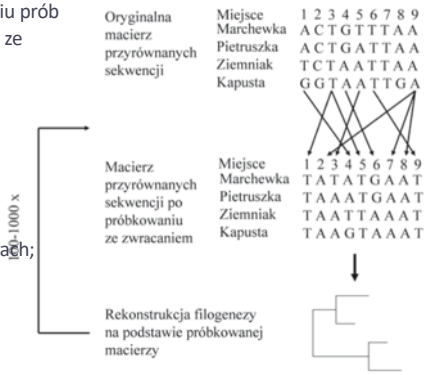
Drzewu można wierzyć, jeśli:

- Metoda szacowania filogenezy jest spójna
- W danych jest silny sygnał filogenetyczny
- Filogeneza genu oddaje filogenezę gatunków
- Zakorzenienie drzewa jest prawidłowe

34

Bootstrap – jak się wyciągnąć z filogenetycznego bagna

- Bootstrap polega na tworzeniu prób pseudo-losowych (losowanie ze zwracaniem)
- Wygenerowane macierze są poddawane takiej samej analizie, jak macierz oryginalna
- Zliczane są wystąpienia takich samych grup na drzewach; przyjmuje się, że dobrze wsparte grupy to takie, które pojawiły się w 95% drzew (przez analogię ze statystyką)



35

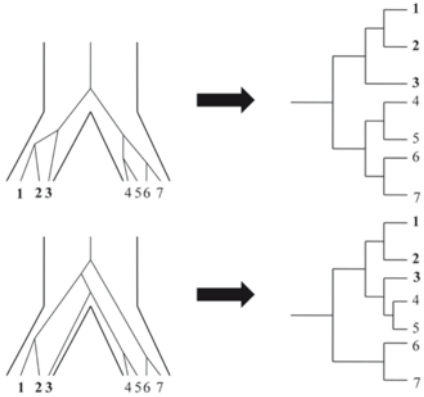
Kiedy filogeneza genu nie oddaje filogenezy gatunków?

- Drzewo genu nie oddaje filogenezy gatunków w wypadku:
  - horyzontalnego przepływu genów (transferu horyzontalnego)
  - rekombinacji genów paralogicznych
  - niepełnego sortowania linii genealogicznych
  - silnego doboru premiującego polimorfizm alleli w loci
  - hybrydyzacji i introgresji
- Zjawiska te można wychwycić porównując drzewa uzyskane za pomocą różnych genów (filogenomika)

36

Filogeneza genu a filogeneza gatunków

- Drzewo genu jest identyczne z drzewem gatunków, jeśli allele uległy pełnemu posortowaniu między gatunkami potomnymi



37

Podsumowanie

- Skuteczność oszacowania filogenezy organizmów zależy od:
  - wyboru odpowiedniego locus lub loci (drzewo genów powinno odpowiadać drzewu gatunków)
  - mocy sygnału filogenetycznego (tempa ewolucji sekwencji w czasie oraz zróżnicowania tego tempa w obrębie sekwencji)
  - spójności metody rekonstrukcji filogenezy (metody mogą być niespójne w wypadku dużego szumu filogenetycznego lub niedoszacowania długości gałęzi)
  - dobrego wyboru grupy zewnętrznej

Drzewo życia Haeckla i współczesne

38