Filogenetyka molekularna - zew rozpaczy

Abstract

W tym dokumencie spróbuję zawrzeć najważniejsze informacje z zakresu filogenetyki molekularnej, które mogą się przydać na egzaminie. Nie potrafię zagwarantować, że posiadając wiedzę z tego dokumentu uda się zdać egzamin, aczkolwiek lepsze to niż nic. Do trzech razy *(w moim przypadku)* sztuka.

> *In the depth of winter, I finally learned that within me there lay an invincible summer.*
>
> — Albert Camus

Wyjaśnij pojęcie mikroewolucji

Ewolucja ponizej poziomu gatunków, np. dryf, selekcja.

Wyjaśnij pojęcie makroewolucji

Ewolucja na poziomie gatunków bądź powyzej.

Kod genetyczny

1. Jest trójkowy - na jeden kodon przypadają trzy nukleotydy.
2. Jest zdegenerowany - aminokwas moze byc kodowany przez rozne kodony / trojki nukleotydow.
3. Jest bezprzecinkowy - nie ma przerw miedzy kodonami.
4. Jest niezachodzacy - jeden nukleotyd moze byc czescia tylko jednego kodonu.
5. Jest uniwersalny - kodony koduja te same aminokwasy u wszystkich organizmow.

Kodony STOP

Kodony STOP to **UAA UAG UGA** - koncza ekspresje bialka.

IUAPC

IUAPC	**Nukleotyd**
A	Adenina
C	Cytozyna
G	Guanina
T badz U	Tymina / Uracyl
R	A / G
Y	C / T
S	C / G
W	A / T
K	G / T
M	A / C
B	C / G / T
D	A / G / T
H	A / C / T
V	A / C / G

IUAPC	Nukleotyd
N	cokolwiek
. lub -	przerwa

Modele ewolucji

Model	Tempo subtytucji	Frekwencja nukleotydow
K2P	rózne	równa
GTR	rózne	różna
HKY	różne	różna
JC	równe	równa

Metody filogenetyczne

- oparte na dystansie
  1. Neighbor Joining
  2. UPGMA
  3. ME
- oparte na podobieństwie
  1. Maximum Parsimony
  2. Maximum Likelihood
  3. Bayesian Inference (bayesowska)

Markery molekularne

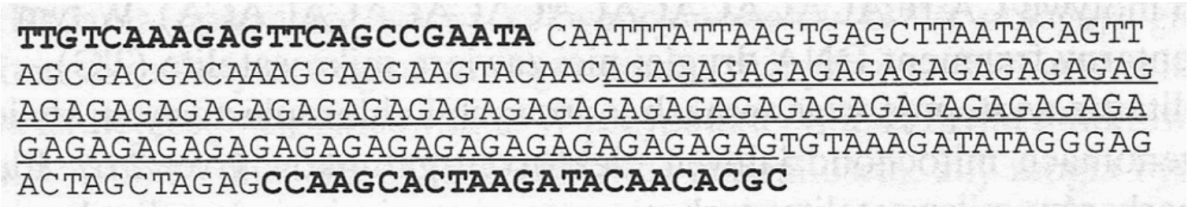


Figure 1: Ze względu na obecność krótkich tandemowych powtórzeń (STR), ten marker to mikrosatelita.

Zapis schematyczny tej mikrosatelity:

5′ - TTGTCAAAGAGTTCAGCCGAATACAATTTATTAAGTG ... [AG]n ... TAAAGATATAGGAGACTAGCTAGAGCCAAGCACTAAGATACAACACGC - 3′

- [AG]n oznacza powtarzającą się sekwencję AG, gdzie n to liczba powtórzeń (w tym przypadku jest ich sporo, można policzyć dokładnie, ale schematycznie wystarczy oznaczyć jako n).
- Sekwencje flankujące przed i po powtórzeniach są istotne dla projektowania primerów do amplifikacji PCR.

Cechy widoczne w sekwencji:

- Region powtarzalny: W środku sekwencji znajduje się wiele powtórzeń AG i AGAG, które są charakterystyczne dla markerów typu mikrosatelitów (krótkie tandemowe powtórzenia, **STR** - *short tandem repeats*).

- Sekwencje flankujące: Powtórzenia AG są otoczone przez unikalne sekwencje DNA na początku i na końcu.

Markery dziedziczone dwurodzicielsko
- Dziedziczone od obojga rodziców.
- Wykorzystywane do badania rekombinacji genetycznej, różnorodności genetycznej i filogenii na poziomie populacji.

Przykłady:
- Alloenzymy:
  - Polimorfizmy w białkach kodowanych przez geny jądrowe.
  - Używane w analizach enzymatycznych, np. elektroforezie.
  - Zastosowania: badania różnorodności genetycznej, porównania populacji.
- nDNA (jądrowy DNA):
  - Zawiera zarówno geny kodujące białka, jak i niekodujące sekwencje.
  - Zastosowania:
    1. Analiza filogenetyczna (np. geny kodujące rRNA).
    2. Badania różnorodności genetycznej.
    3. Analizy związane z rekombinacją genetyczną.

Markery dziedziczone jednorodzicielsko
- Dziedziczone wyłącznie od jednego z rodziców.
- Pozwalają na śledzenie linii matczynej lub ojcowskiej.
- Stabilność w dziedziczeniu (brak rekombinacji lub jej ograniczenie).

Przykłady:
- mtDNA (mitochondrialny DNA):
  - Dziedziczenie matczyne (w większości organizmów).
  - Zastosowania:
    1. Analizy linii matczynej.
    2. Badania różnorodności populacyjnej.
    3. Rekonstrukcja filogenezy.
  - Cechy charakterystyczne:
    1. Wysoka mutowalność w niektórych regionach (np. D-loop).
    2. Brak rekombinacji.
- cpDNA (chloroplastowy DNA):
  - Dziedziczenie głównie matczyne (u większości roślin, choć u niektórych gatunków ojcowskie).
  - Zastosowania:
    1. Analiza filogenetyczna roślin.
    2. Śledzenie migracji roślin.
  - Cechy charakterystyczne:
    1. Relatywnie konserwatywne sekwencje.
    2. Stabilne dziedziczenie.
- Chromosomy haploidalne:

  Np. chromosom Y u organizmów o determinacji płciowej XY (dziedziczenie ojcowskie).
  - Zastosowania:
    1. Analizy linii ojcowskiej.
    2. Rekonstrukcja historii populacji ludzkich i zwierzęcych.

Dwurodzicielskie markery dostarczają informacji o zmienności genetycznej i rekombinacji w obrębie populacji.

Jednorodzicielskie markery pozwalają na śledzenie linii genealogicznych i migracji.

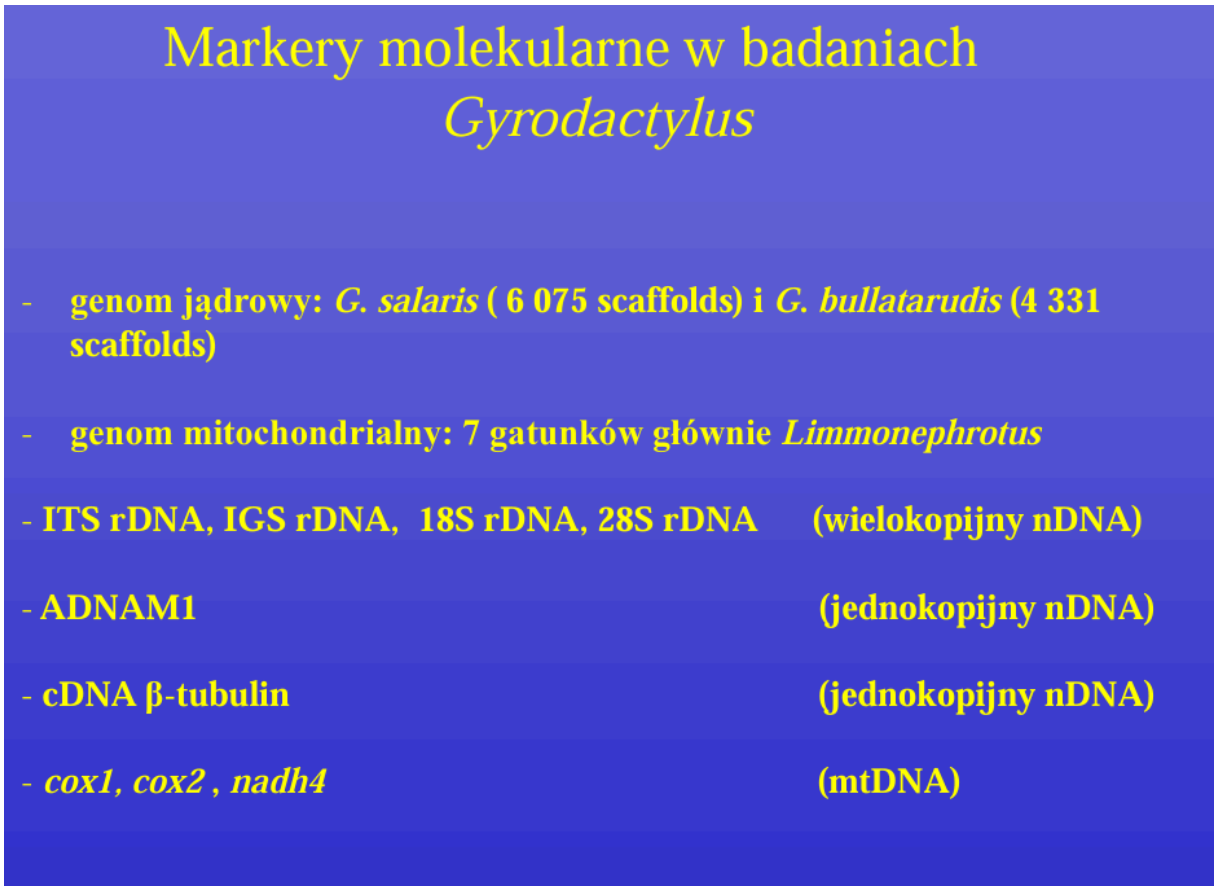
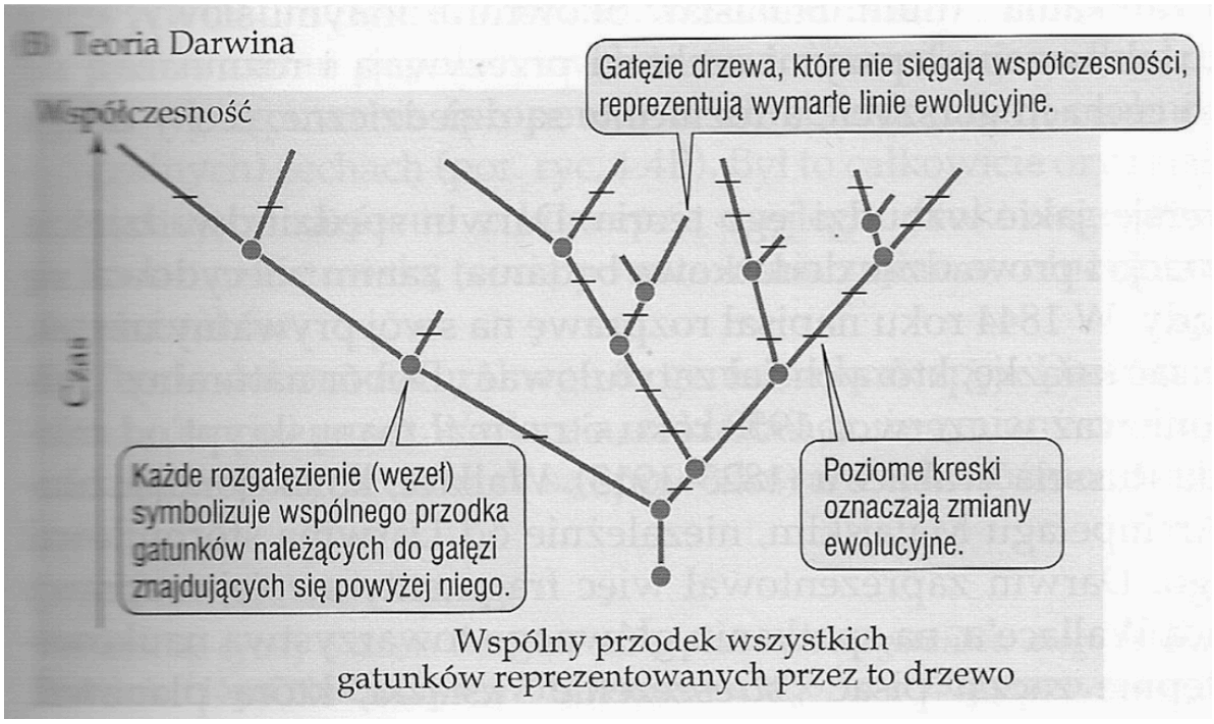
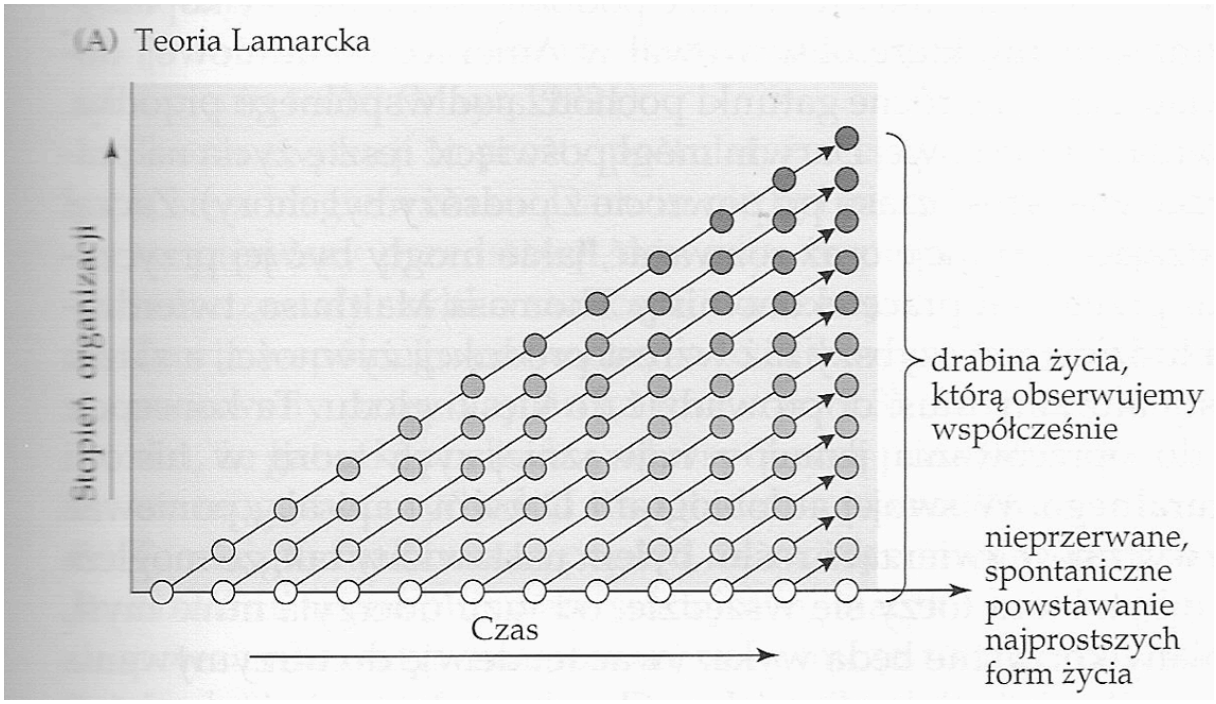


Figure 2: Markery molekularne w badaniach Gyrodactylus

Teroria Darwina



Teoria Lamarcka



Analiza BI z opcja zegara molekularnego

5 kroków analizy:

1. Matrycę .nxs importujemy do programu BEAUTi -> tworzymy plik .xml.
2. Plik .xml importujemy do programu BEAST -> uzyskujemy plik .log., .trees.
3. Plik .log. importujemy do programu Tracer -> weryfikujemy parametry.
4. Plik .trees. importujemy do programu TreeAnnotator -> uzyskujemy plik z drzewem o największej wiarygodności.
5. Plik z drzewem importujemy do programu FigTree -> wizualizujemy drzewo i formatujemy.

Tranzycje i transwersje

- Tranzycje: zmiana puryny na purynę lub pirymidynę na pirymidynę.
- Transwersje: zmiana puryny na pirymidynę lub odwrotnie.

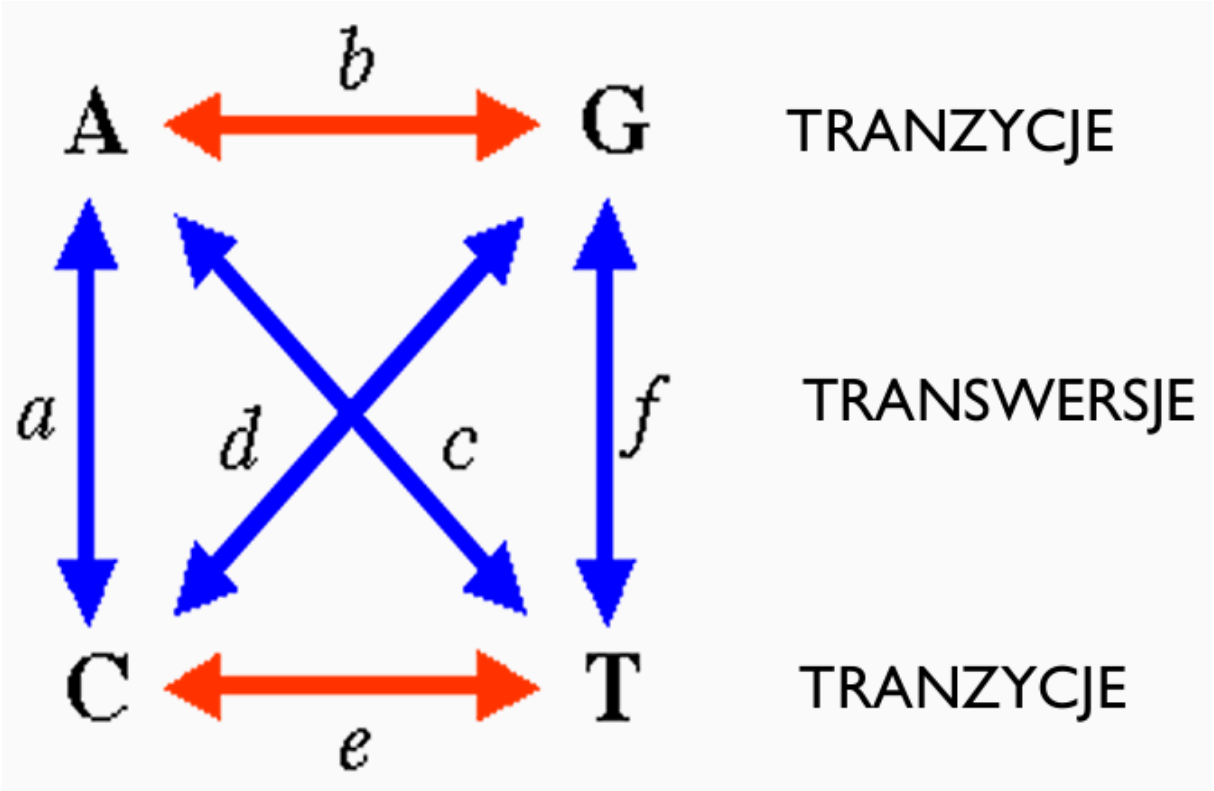


Figure 5: Na niebiesko zaznaczone są transwersje (pionowo i na skoks), na czerwono tranzycje (poziomo).

Miara zmienności genetycznej

	Locus (j)								
i	A	B	C	D	E	F	G	H	h_i
1	*aa*	*aa*	*aa*	*aa*	***ab***	*aa*	*aa*	*aa*	0,125
2	*bb*	***ab***	***ab***	***ab***	*bb*	*aa*	***ab***	*aa*	0,500
3	*cc*	***ac***	***bc***	***bd***	*dd*	***ad***	***cd***	***bc***	0,750
4	*aa*	*aa*	*aa*	*aa*	*aa*	*aa*	*aa*	*aa*	0,000
5	*cc*	*cc*	*cc*	*cc*	*cc*	*cc*	*cc*	*cc*	0,000
h_j	0,0	0,4	0,4	0,4	0,2	0,2	0,4	0,2	$H = 0{,}275$

- h_i - heterozygotyczność osobnika (kolumna)
- h_j - heterozygotyczność locus / populacji (wiersz)
- H - heterozygotyczność populacji. Liczy się jako średnia h_i dla wszystkich osobników w populacji. Bądź jako średnia h_j dla wszystkich locusów w populacji. Obie średnie są sobie równe. Albo jako $\sum_{k=1}^{n} k = \frac{k_1+k_2+...+k_n}{n}$ gdzie n to ilość osobników w populacji (czyli i · j w naszej tabeli to wynosi *40*), a k to suma heterozygotycznych alleli w całej tabeli w naszym przypadku *11*. Co daje nam $\frac{11}{40} = 0.275$
- *a, b, c, d, e* - allele
- *A, B, C, D, E* - locii

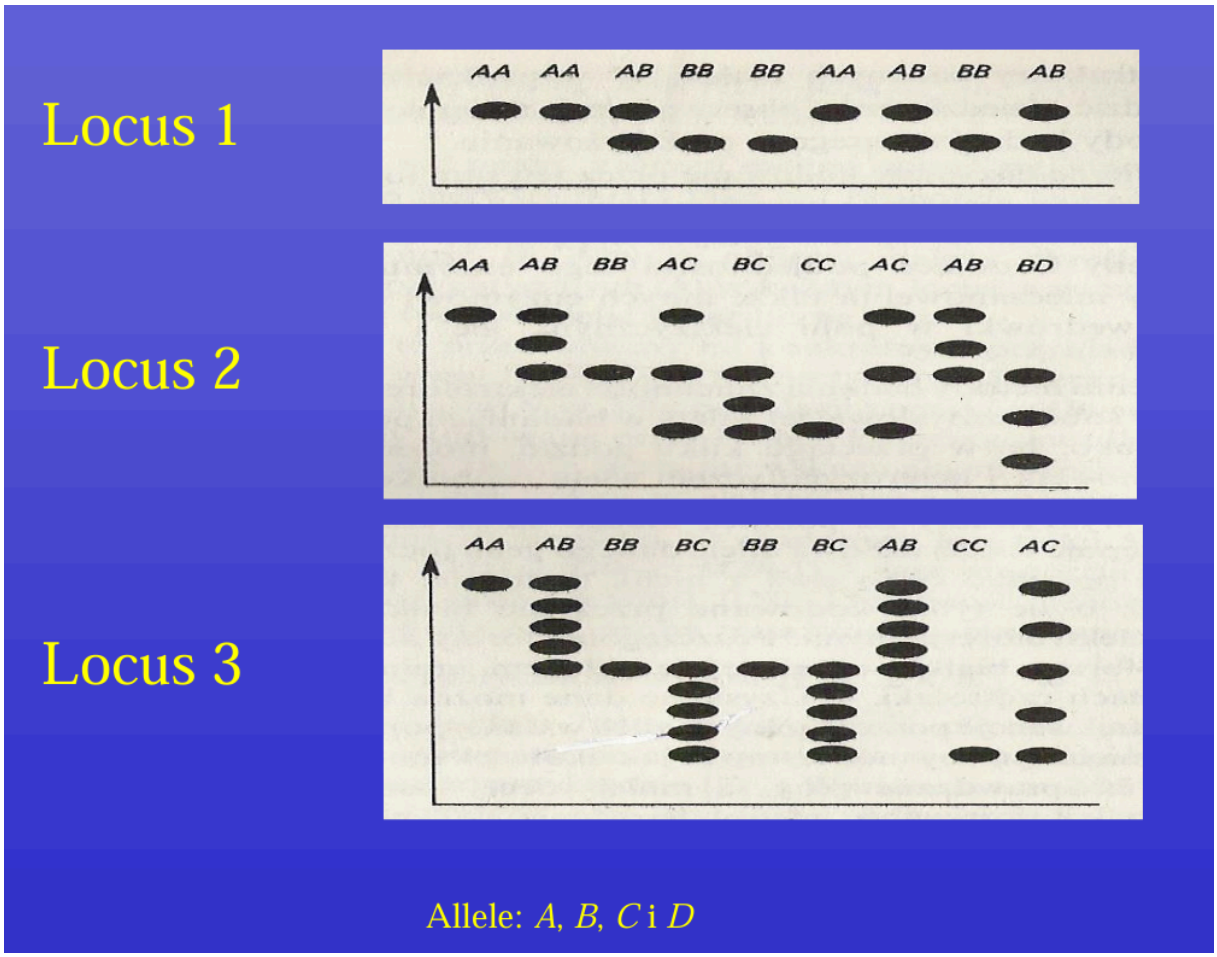
Jak liczymy h_j? Patrzymy ile jest różnych alleli w danym locusie (kolumna) i dzielimy przez liczbę wierszy.

Przykład:
- h_j dla osobnika *B* = $\frac{1+1}{5} = \frac{2}{5} = 0.4$
- h_j dla locus *A* = $\frac{0}{5} = 0$

Teraz policzmy h_i, czyli heterozygotyczność locus / populacji (wiersz).

- h_i dla populacji *1* = $\frac{1}{8} = 0.125$
- h_i dla populacji *3* = $\frac{1+1+1+1+1+1}{8} = \frac{6}{8} = 0.75$
- h_i dla populacji *5* = $\frac{0}{8} = 0$

Teraz zrobimy na innym przykładzie.



h_i

- h_i dla *locus 1* - $\frac{1+1+1}{9} = \frac{3}{9} = \frac{1}{3}$
- h_i dla *locus 2* - $\frac{1+1+1+1+1+1}{9} = \frac{6}{9} = \frac{2}{3}$
- h_i dla *locus 3* - $\frac{1+1+1+1+1}{9} = \frac{5}{9}$

h_j

- h_j dla *osobnika 1* - $\frac{0}{3} = 0$
- h_j dla *osobnika 2* - $\frac{1+1}{3} = \frac{2}{3}$
- h_j dla *osobnika 3* - $\frac{1}{3}$
- h_j dla *osobnika 4* - $\frac{2}{3}$
- h_j dla *osobnika 5* - $\frac{1}{3}$
- h_j dla *osobnika 6* - $\frac{1}{3}$
- h_j dla *osobnika 7* - $\frac{2}{3}$
- h_j dla *osobnika 8* - $\frac{1}{3}$
- h_j dla *osobnika 9* - $\frac{3}{3} = 1$

H

- H = $\frac{\frac{1}{3}+\frac{2}{3}+\frac{5}{9}}{3} = \frac{3+6+5}{27} = \frac{14}{27}$ - sposób pierwszy
- H = $\frac{0+\frac{2}{3}+\frac{1}{3}+\frac{2}{3}+\frac{1}{3}+\frac{1}{3}+\frac{2}{3}+\frac{1}{3}+\frac{3}{3}}{9} = \frac{13}{27}$ - sposób drugi, **UWAGA** tu jest błąd, który wynika z błędu w slajdzie, powinno być $\frac{14}{27}$ ale profesor się pomylił.

Koalescencja

Dla genów dziedziczonych jednorodzicielsko np. *mtDNA* czy *chromosom Y*. Wszystkie kopie genu zbiegają się do jednej kopii – koalescencja.

Odległość genetyczna D i d

- **D** - odległość między sekwencjami jako odsetek różnych nukleotydów w dopasowaniu.
- **d** - średnia liczba podstawień na pojedynczą pozycję w dopasowaniu.

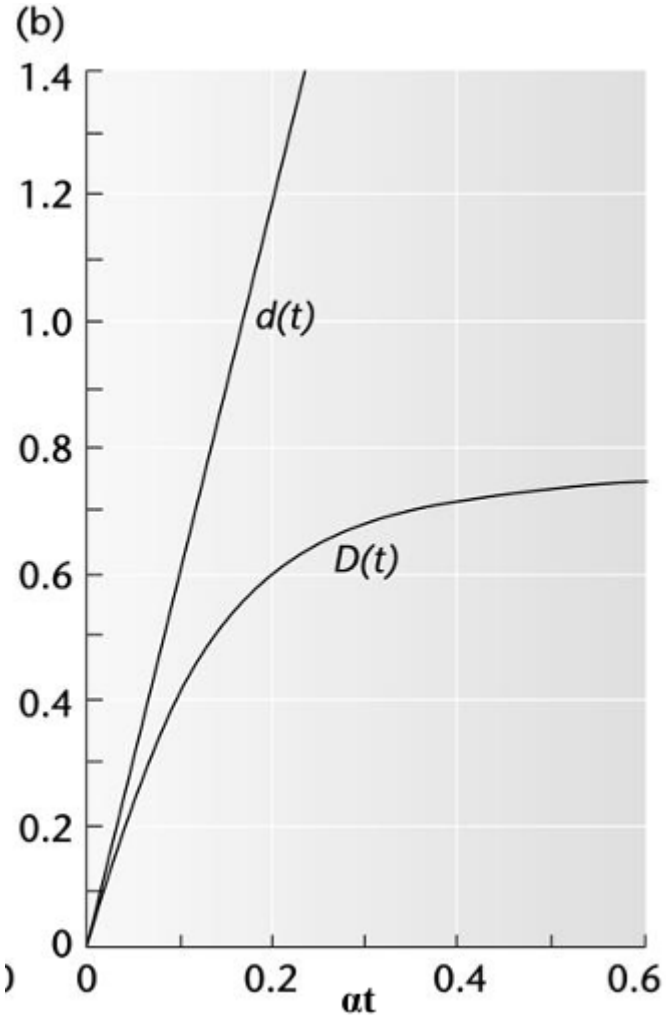


Figure 8: Wykres odległości genetycznej **D** i **d**.

Różnice między **D**, a **d**:
- Główna różnica wynika z faktu, że **D** to miara obserwowalna, natomiast **d** to miara oszacowana, która uwzględnia model ewolucji.
- **D** zawsze jest mniejsze niż **d**, ponieważ **d** uwzględnia *„niewidoczne"* zmiany genetyczne, które nie są bezpośrednio widoczne w sekwencjach porównawczych.
- Na wykresie:
  - Krzywa *D(t)* rośnie wolniej, ponieważ reprezentuje obserwowalną różnicę.
  - Krzywa *d(t)* rośnie szybciej, pokazując bardziej dokładny obraz liczby mutacji.

Odległość **d** w kontekście modelu Jukes-Cantora (JC): Założenia:
- Tempo mutacji jest jednakowe dla wszystkich nukleotydów (oznaczymy je jako μ) i częstość nukleotydów jest jednakowa.

Wzór na odległość **d** w modelu Jukes-Cantora (JC): $d = 2t * 3\mu = 6t\mu$, gdzie:
- t to czas ewolucji.
- μ to tempo mutacji.

inny wzór: $d = -\frac{3}{4} * \ln\left(\frac{1-4}{3D}\right)$, gdzie D to odległość genetyczna.

Odległość **D** w kontekście modelu Kimury 2-parametrowego (K2P): Założenia:
- Tempo podstawień jest różne dla transwersji i tranzycji (oznaczymy je jako α i β). Częstość nukleotydów jest jednakowa.

Wzór na odległość **D**: $D = S + V$, gdzie:
- S - liczba pozycji, w których występuje tranzycja, podzielona przez całkowitą liczbę porównanych pozycji.
- V - liczba pozycji, w których występuje transwersja, podzielona przez całkowitą liczbę porównanych pozycji.

S i V liczymy poprzez porównanie sekwencji.

Wzór na odległość **d**: $d = -\frac{1}{2}\ln(1 - 2S - V) - \frac{1}{4}\ln(1 - 2V)$

d z **D** w miarę się pokrywa do wartości 0.25, dla **D** > 0.5 **d** zaczyna rosnąć szybciej.

Metoda sekwencjonowania Sangera

Wersja krótka:
- Izolacja DNA.
- Przygotowanie reakcji sekwencjonowania.
- Synteza DNA.
- Rozdział fragmentów DNA.
- Odczyt fluorescencji.
- Analiza danych.

Wersja długa:
- **Izolacja DNA**: Na początku izoluje się fragment DNA, który ma być sekwencjonowany. Najczęściej używa się PCR (reakcji łańcuchowej polimerazy) do powielenia tego fragmentu.
- **Przygotowanie reakcji sekwencjonowania**:
  - Do probówki dodaje się:
    - Matrycowy DNA (fragment, który ma być sekwencjonowany).
    - Starter (krótki odcinek DNA komplementarny do sekwencji początkowej matrycy).
    - Polimerazę DNA.
    - Nukleotydy (dNTPs) oraz zmodyfikowane nukleotydy (ddNTPs) oznaczone barwnikami fluorescencyjnymi.
- **Synteza DNA**: Polimeraza DNA zaczyna tworzyć nową nić, korzystając z matrycy i startera. Gdy wbudowany zostanie zmodyfikowany nukleotyd (ddNTP), synteza zostaje zakończona, ponieważ ddNTP nie posiada grupy hydroksylowej (3'-OH), która jest niezbędna do przyłączenia kolejnego nukleotydu.
- **Rozdział fragmentów DNA**: Powstałe fragmenty DNA różnej długości są rozdzielane za pomocą elektroforezy kapilarnej, gdzie krótsze fragmenty migrują szybciej niż dłuższe.
- **Odczyt fluorescencji**: Oznakowane ddNTPs emitują światło o różnych długościach fal w zależności od barwnika, co pozwala na określenie ostatniego nukleotydu w każdym fragmencie.
- **Analiza danych**: Na podstawie sygnałów fluorescencyjnych komputer odczytuje sekwencję DNA.

Organizacja genomu u eukariontów

Uszeregowanie homologiczne (paralogi)

Osobnik 1	LdhA
Osobnik 1	LdhB
Osobnik 1	LdhC
Osobnik 2	LdhA
Osobnik 2	LdhB
Osobnik 2	LdhC
Osobnik 3	LdhA

Osobnik 3	LdhB
Osobnik 3	LdhC

Uszeregowanie ortologiczne (ortologi)

Osobnik 1	LdhA, LdhB, LdhC
Osobnik 2	LdhA, LdhB, LdhC
Osobnik 3	LdhA, LdhB, LdhC

- Paralogi: Duplikacja ⟶ Różnicowanie w obrębie gatunku.
- Ortologi: Specjacja ⟶ Zachowanie funkcji w różnych gatunkach.

Związki filogenetyczne drzewa

- Takson monofiletyczny: takson posiadający wspólnego przodka i grupujący wszystkich potomków taksonu ancestralnego.

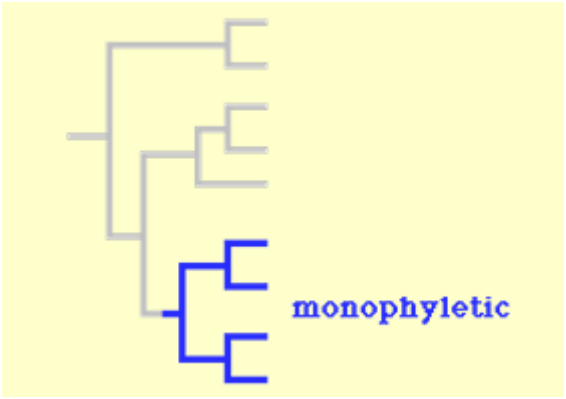


Figure 9: Drzewo filogenetyczne z taksonem monofiletycznym.

- Takson parafiletyczny: takson posiadający wspólnego przodka ale nie grupujący wszystkich potomków taksonu ancestralnego

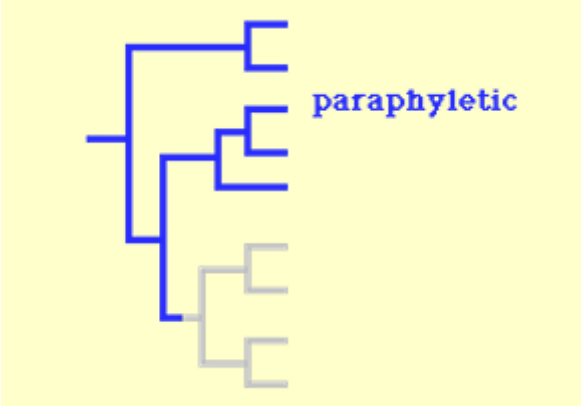


Figure 10: Drzewo filogenetyczne z taksonem parafiletycznym.

- Takson polifiletyczny: takson nie posiadający wspólnego przodka i grupujący potomków kilku taksonów ancestralnych

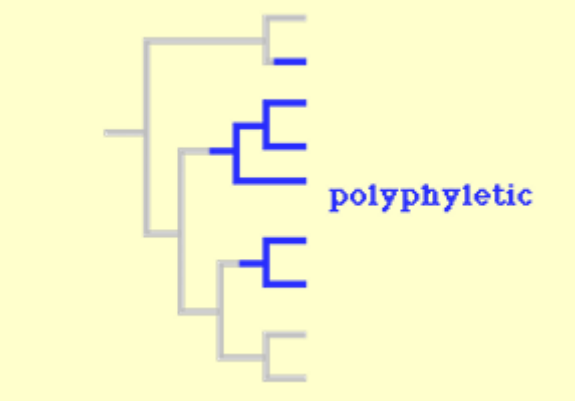


Figure 11: Drzewo filogenetyczne z taksonem polifiletycznym.

Zegar molekularny

Etapy analizy

1. Szacowanie stopnia zróżnicowania sekwencji: obliczanie dystansu genetycznego
2. Kalibracja zegara:
   - przybliżona data rozejścia się dwóch linii genetycznych, powinna być uzyskana z innych danych niż badania molekularne np.:
     1. znane wydarzenie geologiczne
     2. zapisów kopalnych organizmów
3. Określenie tempa substytucji RS:
   - dystans genetyczny podzielony przez czas np.: 2% na mln lat – „uniwersalny” zegar dla mtDNA obliczny z badań nad naczelnymi

R_S vs R_D

Tempo dywergencji RD między dwoma dowolnie wybranymi taksonami jest równe podwojonemu tempu substytucji RS:

$$R_D = 2 * R_S$$

strict vs relaxed

- opcja strict (rygorystyczna) - zakłada równe tempo substytucji wzdłuż gałęzi drzewa filogenetycznego (mało realna, ale lepsza dla nierówno próbkowanych matryc)
- opcja relaxed (rozluźniona) - zakłada różne tempo substytucji wzdłuż gałęzi drzewa filogenetycznego (realna, ale wymaga równomiernie próbkowanych matryc)

Ewaluacja topologii drzewa

- Bootstrap: metoda resamplingu, polega na wielokrotnym losowaniu z powtórzeniami sekwencji z macierzy i ponownym budowaniu drzewa filogenetycznego.
- Posterior probability: prawdopodobieństwo a posteriori, wyznaczane na podstawie analizy bayesowskiej, określa jak bardzo dana gałąź drzewa jest wspierana przez dane.

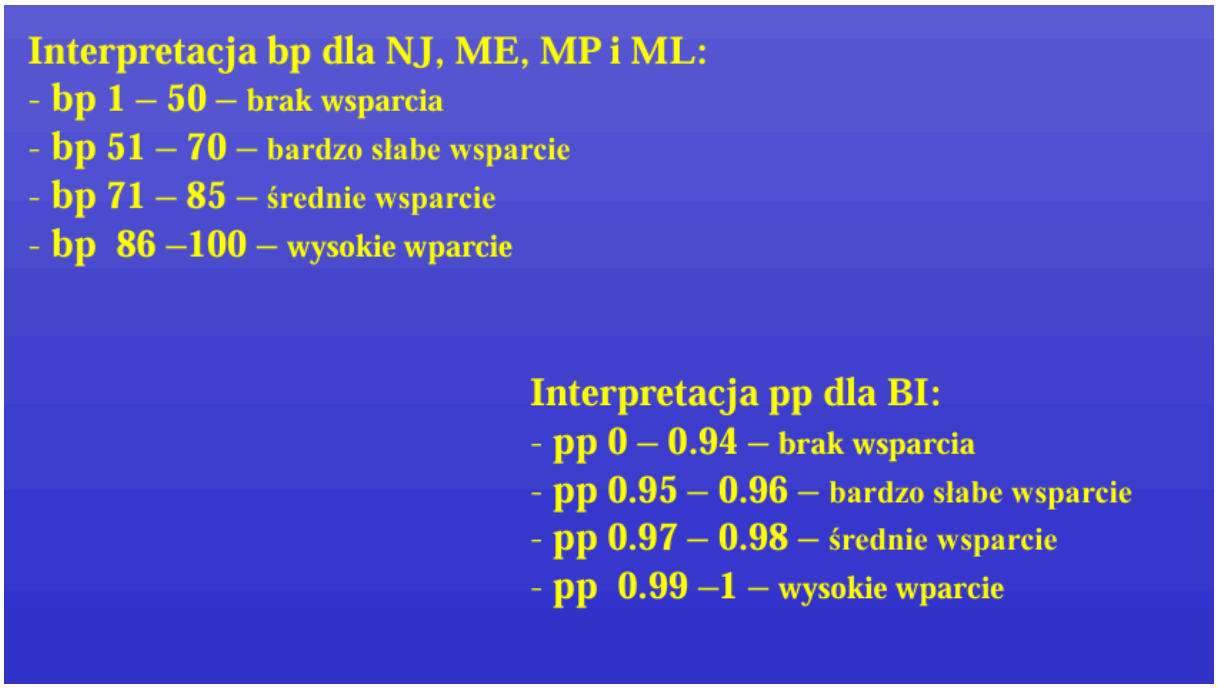


Figure 12: Ewaluacja topologii drzewa filogenetycznego.

Kryterium parsymonii

- Parsymonia: zasada prostoty, wybieramy drzewo, które wymaga najmniejszej liczby zmian.
- Sekwencja jest parsynomicznie informatywna, gdy ma co najmniej dwa różne znaki w różnych organizmach i każdy z tych dwóch znaków musi występować w co najmniej dwóch taksonach.

Przkyład:

Organizm	Sekwencja
A	AGGCT
B	AGGTT
C	AAGCT
D	AAGCT
E	AAGCT

Pozycja 2 jest parsynomicznie informatywna: A i G występują w różnych organizmach.

Pozycja 4 nie jest parsynomicznie informatywna: C i T występują w różnych organizmach, ale tylko w jednym taksonie.

Paralogi/ortologi

Ortologi zawsze tworzą grupę monofiletyczną.

Paralogi nie zawsze mają różne funkcje. Mogą zachować tę samą funkcję, mogą też pod wpływem ewolucji zacząć pełnić różne funkcje.

Zmiany synonimiczne i niesynonimiczne

- Zmiany synonimiczne: zmiany w sekwencji, które nie prowadzą do zmiany aminokwasu.
- Zmiany niesynonimiczne: zmiany w sekwencji, które prowadzą do zmiany aminokwasu.

$\frac{dN}{dS} > 1$ - dodatni dobór, selekcja pozytywna, zmiany niesynonimiczne są częstsze niż synonimiczne.

$\frac{dN}{dS} = 1$ - neutralna ewolucja, zmiany niesynonimiczne i synonimiczne są tak samo częste.

$\frac{dN}{dS} < 1$ - ujemny dobór, selekcja negatywna, zmiany synonimiczne są częstsze niż niesynonimiczne.

Zmiany synonimiczne występują częściej niż niesynonimiczne, więc częściej zdarza się, że $\frac{dN}{dS} < 1$.

Mechanizmy molekularne

- **Horyzontalny transfer genów (HGT)**: przenoszenie genów między organizmami, niezależnie od pokrewieństwa.
- **Duplikacja i pseudogenizacja**: duplikacja genów prowadzi do powstania paralogów, które mogą zacząć pełnić nowe funkcje lub zachować funkcje oryginalne.
- **Niekompletne sortowanie alleli**: w wyniku rekombinacji genów w obrębie populacji, niektóre allelomorfy mogą być przekazywane w sposób niezgodny z drzewem filogenetycznym.

Metoda Bootstrap

1. Robi się zestawienia sekewncji, tyle ile replik bootstrap o długości takiej samej jak orginalne, ale wybierając losowe (z powtórzeniami) kolumny z oryginalnego zestawienia.
2. Robi się drzewo z każdego zestawienia.
3. Dla każdego węzła w orginalnym drzewie sprawdza się ile razy ten węzeł wystąpił w drzewach z zestawień.

Zadanie analizowano sekwencje dwóch białek.

Czy na podstawie dN/dS i tego przez jakie kodony są kodowane można wywnioskować o poziomie ekspresji białka?

Odpowiedź: Tak, wykorzystanie mniejszej ilości różnych kodonów, może świadczyć o większej ekspresji białka.