Wykład 8

Macierze substytucji
Sekwencje pokrewne

Macierze substytucji

PAM - Point Accepted Mutations
Margaret Dayhoff 1978

BLOSUM

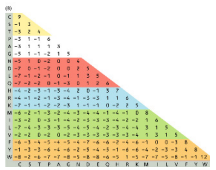
BLOSUM – BLOck SUbistution Matrix

- Hennikof & Hennikof 1992
- Ponad 2000 bloków dopasowanych sekwencji, bez przerw
  - różne poziomy podobieństwa sekwencji
- Sekwencje klastrowane wg podobieństwa

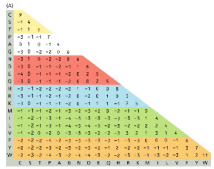
Macierze substytucji – podsumowanie

- PAM
  - utworzone na podstawie globalnego dopasowania podobnych sekwencji
  - najlepiej nadają się do porównywania podobnych białek
    - **PAM-1** gdy oczekujemy **1** mutacji / 100 aminokwasów
    - **PAM-n** gdy oczekujemy **n** mutacji / 100 aminokwasów
- BLOSUM
  - na podstawie lokalnego dopasowania zróżnicowanych sekwencji
  - najlepiej nadają się do porównywania ewolucyjnie odległych białek
    - **BLOSUM 62** gdy średnio **62%** procent aminokwasów zachowane
    - **BLOSUM n** gdy średnio **n%** procent aminokwasów zachowane

Macierze substytucji
zależna od prawdopodobieństwa mutacji



PAM120 przybliżenie z globalnego
Point Accepted Mutations
120 mutacji/100 długości



BLOSUM-62 z lokalnego dopasowania
BLOck SUbstitution Matrix
Co najmniej 62% identyczne

Jeśli koniecznie chcemy je porównać

PAM	BLOSUM
PAM100	BLOSUM90
PAM120	BLOSUM80
PAM160	BLOSUM60
PAM200	BLOSUM52
PAM250	BLOSUM45

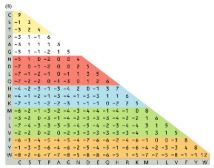
Jak wyznaczamy macierz substytucji

W każdej macierzy substytucji częstotliwość $q_{a,b}$ mutacji z *a* do *b* wyznacza wartości elementów macierzy $s_{a,b}$, ; p_a to prawdopodobieństwo wystąpienia aminokwasu *a* w całej bazie danych sekwencji

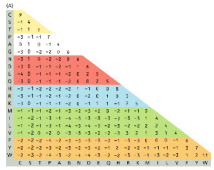
$$s_{a,b} = \frac{\ln(q_{a,b} / p_a p_b)}{\lambda}$$

gdzie λ jest jakimś współczynnikiem, charakterystycznym dla typu macierzy

Macierze substytucji zależna od prawdopodobieństwa mutacji



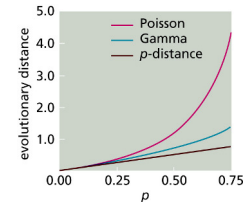
PAM120 przybliżenie z globalnego
Point Accepted Mutations
120 mutacji/100 długości



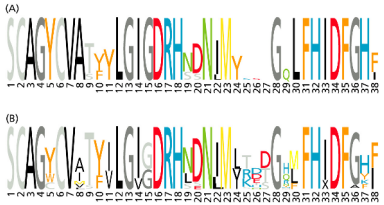
BLOSUM-62 z lokalnego dopasowania
BLOck SUbstitution Matrix
Co najmniej 62% identyczne

Poprawka Poissona

Liczba zaobserwowanych mutacji (d=**p=D/L**) jest zwykle znacznie mniejsza od rzeczywistej liczby mutacji bo bywa kilka na 1 miejsce (np. macierz PAM-120 - 120 mutacji na 100 miejsc)



Gdy mutacje zależą od położenia w sekwencji



Macierze substytucji zależne od pozycji

Position Specific Scoring Matrix (PSSM)

Potrzeba macierzy punktacji zależnej od pozycji)

Macierze substytucji (typu PAM, BLOSUM) mogą być używane do:

- Dopasowywania par sekwencji
- Oceny odległości ewolucyjnej pomiędzy parą białek
- Wyszukania w bazie danych białek podobnych do białek wybranych

Nie są wystarczające do stwierdzenia, czy dane białko jest elementem określonej rodziny (zbioru) białek.

Jak wyznaczamy macierz częstości zależną od pozycji (PSSM)

Załóżmy, że mamy ***N sekwencji*** z jednej rodziny.

Wyznaczamy *częstotliwość* wystąpienia *aminokwasu* ***a*** na *pozycji* ***u*** dla tej rodziny sekwencji: $q_{u,a} = n_{u,a}/N$

$$m_{u,a} = \frac{\log(q_{u,a} / p_a)}{\lambda}$$

p_a to *prawdopodobieństwo* wystąpienia aminokwasu *a* w całej bazie danych sekwencji. Podobnie jak w macierzy substytucji wyznaczamy element macierzy PSSM (λ można tu pominąć, choć stosowane w Psi-Blast):

13

PSSM

Etapy konstrukcji:

- Wybór rodziny do profilu:
- BLAST na podstawie pojedynczej sekwencji
- BLAST z kolejnych sekwencji
- Obliczenie PSSM

14

Entropia

W fizyce ENTROPIA jest miarą gęstości stanów. Stanowi miarę kierunku zachodzenia procesów samorzutnych. Układ fizyczny zawsze dąży do nieporządku, czyli równomiernego rozkładu (P_i) gęstości stanów *i*

$$S = -k_B \sum_i P_i \cdot \ln(P_i)$$ k_B – stała Boltzmanna

ENTROPIA informacyjna (entropia Shannona) jest miarą niepewności informacji

$$H = -\sum_i P_i \cdot \log(P_i)$$

Im bardziej równomierny jest rozkład jakiejś cechy tym mniej informacji mamy o jej potencjalnym wystąpieniu w określonej sytuacji (np. aminokwasu na danej pozycji w sekwencji)

15

Entropia

Istotność informacji w macierzy PSSM dla kolumny ***u*** można ocenić obliczając jej ENTROPIĘ po wszystkich aminokwasach *a* (w bitach to $\log_2$):

$$H_u = -\sum_a q_{u,a} \cdot \log_2(q_{u,a})$$

Przy równomiernym występowaniu wszystkich aminokwasów entropia jest maksymalna.

Max (H_u) = -20 * (1/20 * $\log_2(1/20)$) = **$\log_2$ (20)**

16

Informacja w sekwencji

Informacja I_u zawarta w sekwencji na pozycji *u* może być obliczona jako:

$$I_u = H_{\max} - H_u = \log_2 20 - H_u$$

$H_{max} = \log_2 20$, ale pod warunkiem, że dysponujemy co najmniej 20 sekwencjami. Jeśli sekwencji jest mniej to $H_{max} = \log_2$ (liczba sekwencji).

Maksymalna wartość I_u (mała entropia) oznacza bardzo dobrą konserwację jakiegoś jednego aminokwasu na pozycji *u*

Logo

$$Rozmiar = I_u \cdot q_{u,a}$$

Schneider, Stephens, NAR ,1990

18

Znajdowanie białek z określonej rodziny

CEL: znamy zbiór białek z jednej rodziny i znajdujemy białka pasujące

Metody:
- Dopasowanie do sekwencji konsensusowej (uśrednionej) rodziny
- Dopasowanie do profilu rodziny
- Szukanie „odcisku palca" rodziny
- Metody probabilistyczne (np. ukryte modele Markowa - HMM)

PSI-BLAST

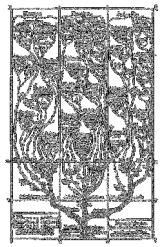
Position-Specific Iterative BLAST
Profil
PSSM - Position-Specific Scoring Matrix

Wykład

Drzewa Filogenetyczne

22

Tradycyjne drzewa pokrewieństwa



Drzewa oparte były o podobieństwa morfologiczne. Nie były binarne.

Hackel. E. 'Monophyletischer Stambaum der Organismen' from 'Generelle Morphologie der Organismen' (1866) with the three branches Plantae, Protista, Animalia

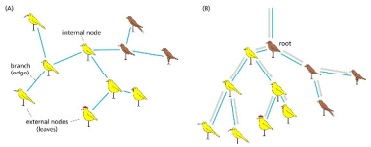
23

Drzewa filogenetyczne

- Drzewo jest grafem, który ma gałęzie, węzły i liście (końcówki, końcowe węzły)
- Drzewo może, ale nie musi mieć korzeń
- Drzewo filogenetyczne to drzewo **binarne**
- Gałęzie w drzewie filogenetycznym są skierowane
- Węzły i liście odpowiadają gatunkom lub sekwencjom molekularnym i pokrewieństwom pomiędzy nimi
- Liście nazywa się w nim **OUT** („Operational Taxonomic Units")

24

Drzewa filogenetyczne



Drzewo może, ale nie musi, mieć korzeń

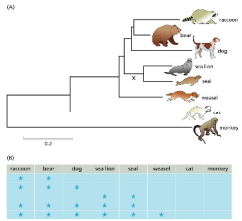
Typy drzew filogenetycznych

Kladogram – nie uwzględnia czasu

Drzewo addytywne. Długości gałęzi proporcjonalne do ilości mutacji na miejsce w sekwencji, skąd można obliczyć odległość ewolucyjną.

Drzewo ultrametryczne. Drzewo addytywne, które dodatkowo ma stałą szybkość mutacji wzdłuż gałęzi (liczy się tylko czas). Skala po prawej pokazuje więc zarówno czas ewolucji, jak i (wyrównaną) ilość mutacji na miejsce.

Tekstowy zapis drzewa

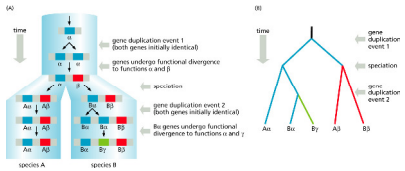


W tabeli mogą się również znaleźć liczby określające długości gałęzi.

Zjawiska, którym podlegają geny (m.in. w sekwencjach homologicznych)

- **Duplikacja** genów (również całych chromosomów i genomów)
- **Podział i ponowne połączenie** fragmentów, ale w innej kolejności
- **Mutacja** genów (wewnątrz gatunku i międzygatunkowo)
- Zanikanie genów po duplikacji (**geny ciche** - pseudogeny)
- **Konwergencja** genów (**homoplazja** sekwencji)
- **Poziomy** (horyzontalny, lateralny) transfer genów – **HGT** (Horizontal Gene Transfer)

Ortologi i Paralogi



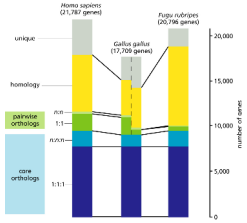
Ortologi / Paralogi

Geny ortologiczne to takie, których rozdzielenie nastąpiło na skutek specjacji (po rozdzieleniu gatunków).

Inaczej mówiąc, w momencie specjacji gen miał tylko jedną kopię, dopiero później gen ewoluował w ramach odrębnych gatunków, w każdym niezależnie

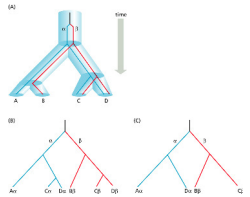
Geny paralogiczne to takie, których rozdzielenie nastąpiło w wyniku duplikacji genu (nie specjacji). Jeden gatunek ma dwie kopie tego samego genu, które ewoluują niezależnie od siebie, ale w ramach tego samego gatunku.

Jaka część genów jest ortologiczna?



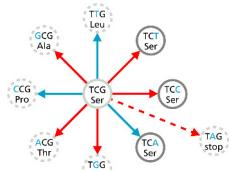
31

Zanikanie (lub nieoznakowanie) genów – wpływ na odległości w drzewie



32

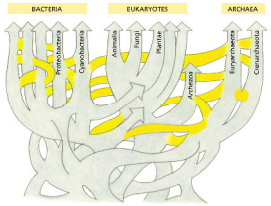
Presja selektywna w pozycji 3 kodonu mutacje synonimowe



Możliwa losowa zawartość nukleotydu na 3 pozycji

33

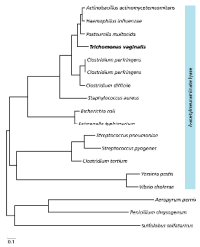
Transfer poziomy genów (HGT), szczególnie bakterie



Geny xenologowe – pozornie b. podobne, jako efekt HGT

34

Transfer poziomy



Rzęsistek pochwowy
Science z 2007 roku doniósł o zsekwencjonowaniu genomu, o wielkości ok. 160 M bp, z czego dwie trzecie to sekwencje powtórzeniowe (repetywne), lub transpozonowe – świadczy to o niedawnym rozszerzeniu genomu, związanym z przystosowaniem do pasożytniczego trybu życia, a wynikłym z poziomego transferu genu od bakterii i późniejszym amplifikacjom odpowiednich rodzin genowych.

Odległa grupa (różne enzymy z jednej superodziny)

35

Jak wyznaczyć dobre drzewo ?

1. Wybrać rodzinę sekwencji **homologicznych** jako OTU

- Sekwencje powinny mieć wystarczająco silny sygnał filogenetyczny. Jeśli sygnał jest słaby (sekwencje są bardzo rozbieżne) to drzewo powstanie, ale nie będzie zbyt informatywne
- Sekwencje muszą pochodzić z genów, które mają odpowiednio dużą, ale nie zbyt dużą szybkość mutacji
- W wybranych genach nie powinno być poziomego tranferu genów
- Sekwencje musza pochodzić z odpowiednich genów (***ortologi***, nie *paralogi*).

36

Geny, które spełniają założenia

- **Prokariota** – sekwencja DNA małej podjednostki rybosomu rRNA (16S RNA). Mimo, że w niektórych genomach pojawia się w kilku kopiach (na tej podstawie wyodrębniono Bacteria i Archea – Carl Woese)

- **Bakterie** – enzymy DNA: GyrA, GyrB, białko chaperonowe HSP60.

- **Zwierzęta** – segment (648 bp) z cytochromu *c* oksydaza I

37

Jak wyznaczyć dobre drzewo

38