Wprowadzenie do Informatyki Biomedycznej

Wykład 2:

Metody dopasowywania sekwencji

Marek Krętowski
pokój 206
e-mail: m.kretowski@pb.edu.pl
http://aragorn.pb.bialystok.pl/~mkret

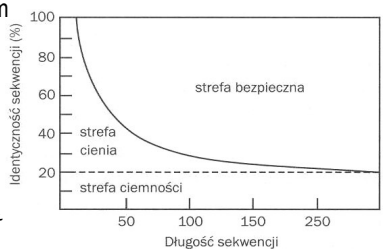
Wersja 1.05

Dopasowywanie sekwencji
(ang. sequence alignment)

- Dopasowywanie (przyrównywanie) sekwencji polega na porównywaniu dwóch (ang. *pair-wise alignment*) lub wielu (ang. *multiple sequence alignment*) sekwencji poprzez poszukiwanie serii pojedynczych znaków lub wzorców znakowych, które występują w tej samej kolejności w sekwencjach
  - w najprostszym przypadku polega na ustawieniu (zapisaniu) dwóch sekwencji obok siebie, tak aby wzajemne przypisanie znaków było jednoznaczne;
  - identyczne lub podobne znaki są umieszczane w tej samej kolumnie, natomiast znaki, których nie można dopasować są ustawiane w jednej kolumnie i oznaczane jako brak dopasowania lub ustawiane naprzeciw przerwy (ang. gap)
- W przypadku optymalnego dopasowania, przerwy i braki dopasowania są tak ustawiane aby liczba znaków odpowiadających sobie (identycznych lub podobnych) była możliwie największa
  - sekwencje, które bez problemu potrafimy w powyższy sposób ustawić określane są mianem podobnych

Homologia a podobieństwo sekwencji

- Jeżeli dwa geny (lub kodowane przez nie białka) pochodzą od wspólnego ewolucyjnie przodka => wykazują homologię (są homologiczne)
  - odróżnienie homologii od podobieństwa jest istotne
  - homologia sekwencji jest wnioskiem o pokrewieństwie, który można wysnuć z podobieństwa
  - jeżeli stopień podobieństwa sekwencji jest odpowiednio wysoki to można wnioskować o ich pokrewieństwie ewolucyjnym
  - ważny jest rodzaj sekwencji (w sekwencjach nukleotydowych łatwiej o przypadkowe podobieństwo) oraz ich długość (krótkie sekwencyjne wymagają ostrzejszych kryteriów przy wnioskowaniu o homologii)
  - w „strefie cienia" - odległe homologie mieszają się z sekwencjami, których podobieństwo jest przypadkowe



Strefy przyrównań sekwencji białkowych
[J. Xiong: Podstawy bioinformatyki, WUW]

Rodzaje dopasowania

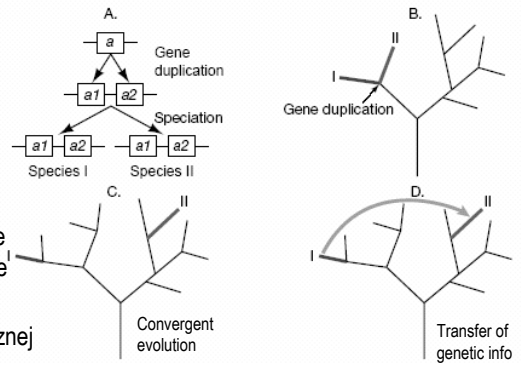
- Dopasowanie globalne - dokonywana jest próba wzajemnego ustawienia całych sekwencji
  - wykorzystująca jak największą liczbę znaków i rozciągająca się pomiędzy końcami każdej z sekwencji
  - stosowane w analizie sekwencji o dużym stopniu podobieństwa i zbliżonej długości
- Dopasowanie lokalne - wychwytywane są odcinki sekwencji zawierające najbardziej "gęste" dopasowanie, tworzą one jedną lub wiele wysp w ramach wzajemnie ustawionych sekwencji
  - stosowane w sytuacji gdy sekwencje są podobne jedynie na pewnym odcinku; gdy sekwencje różnią się długością lub gdy sekwencje współdzielą pewne utrwalone obszary

```
L G P S S K Q T G K G S - - S R I W D N
|       |     | | |       |     |
L N - I T K S A G K G A I M R L G D A
```

```
- - - - - - - T G K G - - - - - - - -
            |   | |
- - - - - - - A G K G - - - - - - - -
```

Znaczenie dopasowywania sekwencji

- Niezbędne w odkrywaniu funkcjonalnej, strukturalnej i ewolucyjnej wiedzy o biologicznych sekwencjach (szczególnie ważne jest optymalne dopasowanie)
  - najbardziej zbliżone sekwencje zapewne mają takie same funkcje (np. podobne funkcje biochemiczne i strukturę 3D w przypadku białek)
  - jeżeli dwie sekwencji z dwóch odmiennych organizmów są zbliżone być może istniała sekwencja je poprzedzająca (przodek w procesie ewolucji)
  - podobieństwo sekwencji nie musi jednak być związane z ewolucją genu przodka - możliwe jest upodobnienie się (takie same funkcje) w procesie ewolucji
  - przeniesienie informacji genetycznej



Metody dopasowywania sekwencji

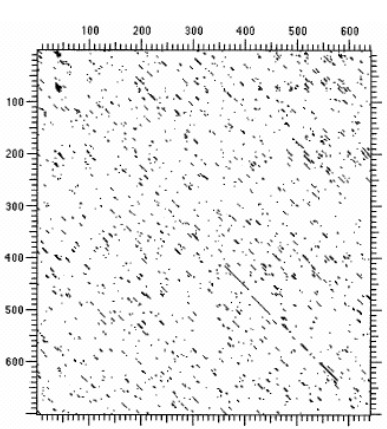
- Metoda kropkowa (ang. dot matrix) - zwykle stosowana jako pierwsza, pokazuje wszystkie możliwe dopasowania w postaci linii ukośnych; pozwala na bezpośrednie wykrywanie wstawek i braków oraz powtórzeń (również odwróconych) - może to być trudniejsze przy wykorzystaniu bardziej "automatycznych" metod
- Programowanie dynamiczne - szeroko stosowana metoda, gwarantuje odnalezienie optymalnego ustawienia względem zadanego systemu oceny; opracowano usprawnienia metody o złożoności bliskiej liniowej (zarówno czas jak i pamięć)
  - globalne dopasowanie (Needlemann-Wunsch, 1970)
  - lokalne dopasowanie(Smith-Waterman, 1981)
- Metody bazujące na słowach (ang. word or k-tuple) - dopasowują sekwencje bardzo szybko; najpierw wyszukiwane są identyczne krótkie odcinki w sekwencjach (nazywane słowami) a następnie są one ustawiane przy użyciu programowania dynamicznego; wykorzystywane do przeszukiwania całych baz sekwencji w celu odnalezienia najbardziej podobnej sekwencji do zadanej
  - FASTA i BLAST (metody heurystyczne)

Metoda kropkowa

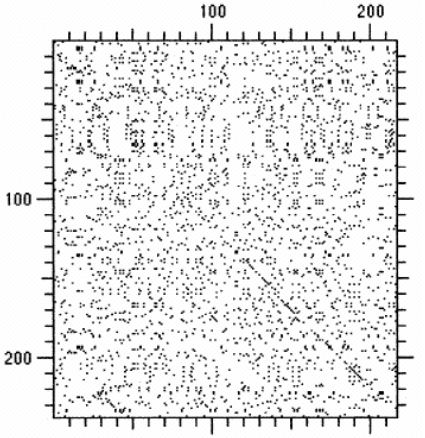
- Podstawową zaletą jest fakt, że wszystkie powiązania pomiędzy sekwencjami są przedstawione (zwizualizowane) pozostawiając badaczowi wybór najbardziej znaczących
- Konstrukcja:
  - pierwsza z sekwencji wypisywana jest u góry od lewej do prawej, druga od góry do dołu po lewej stronie
  - kropka (punkt) pojawia się wszędzie tam, gdzie występuje dopasowanie
- Detekcja dopasowanych regionów może być poprawiona przy użyciu filtracji
  - zamiast pojedynczych znaków porównywane w danym momencie jest okno zawierające zadaną liczbę kolejnych znaków i punkt jest rysowany tylko wtedy, gdy określona minimalna liczba znaków jest zgodna
  - zwykle większe okno używane przy sekwencjach DNA (potencjalnie duża liczba przypadkowych dopasowań, tylko 4 znaki), typowo 10 zgodnych spośród 15

	A	C	G	T	A	T	C	A
A	*				*			*
C		*					*	
T				*		*		
A	*				*			*
T				*		*		
G			*					
T				*		*		
G			*					

Przykłady analizy kropkowej sekwencji DNA



Okno 11, min. zgodność 7



Okno 1, min. zgodność 1

Programowanie dynamiczne w dopasowywaniu sekwencji

- Metoda, która może być stosowana zarówno do sekwencji DNA jak i białek
- Gwarantuje odnalezienie optymalnego dopasowania przy czym uzyskiwane dopasowanie jest zależne od przyjętego systemu oceny
- Wiele różnych ustawień może uzyskiwać zbliżony wynik oceny w stosunku do najlepszego i w takiej sytuacji należy je przeanalizować
- W standardowej formie złożoność metody jest kwadratowa lub nawet wyższa oraz występują znaczne wymagania pamięciowe co ograniczało wykorzystanie metody w przypadku długich sekwencji; zaproponowane zostały jednak usprawnienia znacznie redukujące wymagania, bez poświęcania pewności działania

Ocena dopasowania

- Jakość dopasowania sekwencji jest wyliczana na bazie przyjętego systemu oceny, który
  - preferuje dopasowanie identycznych lub podobnych aminokwasów
  - karze dopasowanie "odległych" aminokwasów oraz występowanie przerw
- Do tworzenia systemów oceny wykorzystane są informacje zaobserwowane w powiązanych białkach i dotyczące różniących ich elementów
  - prawdopodobieństwa wystąpienia konkretnych par aminokwasów w podobnych białkach
  - prawdopodobieństwo przypadkowego wystąpienia pary aminokwasów przy założeniu, że niektóre aminokwasy występują często a inne rzadko
  - prawdopodobieństwo, że wprowadzenie przerwy w jednej z sekwencji pozwoli lepiej dopasować dalsze części sekwencji
- Im uzyskany wynik jest wyższy tym lepiej

Macierze zastępstw aminokwasów (ang. substitution matrix)

- Każda komórka macierzy podaje stosunek zaobserwowanej częstości zastąpienia pomiędzy parami aminokwasów w powiązanych białkach do częstości przypadkowej zamiany wyliczonej na podstawie częstości występowania konkretnych aminokwasów w białkach - tzw. szanse (ang. odds scores)
  - Dayhoff PAM (Percent Accepted Mutation) - rodzina macierzy, w których podane są prawdopodobieństwa zamiany jednego aminokwasu na inny w sekwencjach homologicznych białek w procesie ewolucji; każda z macierzy zawiera prawdopodobieństwa zmian w zadanych okresie czasu ewolucji; wartości zostały uzyskane na podstawie 1572 zmian w 71 grupach sekwencji białek, które mają co najmniej 85% podobieństwa; np.PAM250 (reprezentuje poziom 250% zmian i odpowiada podobieństwo ok. 20%)
  - BLOSUM (Blocks Amino Acid Substitution Matrices) - wartości wyliczone na podstawie wymian aminokwasów w zbiorze ~2000 utrwalonych wzorców (nazywanych blokami); bloki te zostały odnalezione w bazie ponad 500 rodzin powiązanych białek; np. BLOSUM62

Przykłady macierzy zastępstw aminokwasów

Macierz może zawierać wartości dodatnie i ujemne, odzwierciedlające prawdopodobieństwo zastępstwa aminokwasów w powiązanych białkach

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	12																				C
S	0	2																			S
T	−2	1	3																		T
P	−3	1	0	6																	P
A	−2	1	1	1	2																A
G	−3	1	0	−1	1	5															G
N	−4	1	0	−1	0	0	2														N
D	−5	0	0	−1	0	1	2	4													D
E	−5	0	0	−1	0	0	1	3	4												E
Q	−5	−1	−1	0	0	−1	1	2	2	4											Q
H	−3	−1	−1	0	−1	−2	2	1	1	3	6										H
R	−4	0	−1	0	−2	−3	0	−1	−1	1	2	6									R
K	−5	0	0	−1	−1	−2	1	0	0	1	0	3	5								K
M	−5	−2	−1	−2	−1	−3	−2	−3	−2	−1	−2	0	0	6							M
I	−2	−1	0	−2	−1	−3	−2	−2	−2	−2	−2	−2	−2	2	5						I
L	−6	−3	−2	−3	−2	−4	−3	−4	−3	−2	−2	−3	−3	4	2	6					L
V	−2	−1	0	−1	0	−1	−2	−2	−2	−2	−2	−2	−2	2	4	2	4				V
F	−4	−3	−3	−5	−4	−5	−4	−6	−5	−5	−2	−4	−5	0	1	2	−1	9			F
Y	0	−3	−3	−5	−3	−5	−2	−4	−4	−4	0	−4	−4	−2	−1	−1	−2	7	10		Y
W	−8	−2	−5	−6	−6	−7	−4	−7	−7	−5	−3	2	−3	−4	−5	−2	−6	0	0	17	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

PAM250

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	−1	4																			S
T	−1	1	5																		T
P	−3	−1	−1	7																	P
A	0	1	0	−1	4																A
G	−3	0	−2	−2	0	6															G
N	−3	1	0	−2	−2	0	6														N
D	−3	0	−1	−1	−2	−1	1	6													D
E	−4	0	−1	−1	−1	−2	0	2	5												E
Q	−3	0	−1	−1	−1	−2	0	0	2	5											Q
H	−3	−1	−2	−2	−2	−2	1	−1	0	0	8										H
R	−3	−1	−1	−2	−1	−2	0	−2	0	1	0	5									R
K	−3	0	−1	−1	−1	−2	0	−1	1	1	−1	2	5								K
M	−1	−1	−1	−2	−1	−3	−2	−3	−2	0	−2	−1	−1	5							M
I	−1	−2	−1	−3	−1	−4	−3	−3	−3	−3	−3	−3	−3	1	4						I
L	−1	−2	−1	−3	−1	−4	−3	−4	−3	−2	−3	−2	−2	2	2	4					L
V	−1	−2	0	−2	0	−3	−3	−3	−2	−2	−3	−3	−2	1	3	1	4				V
F	−2	−2	−2	−4	−2	−3	−3	−3	−3	−3	−1	−3	−3	0	0	0	−1	6			F
Y	−2	−2	−2	−3	−2	−3	−2	−3	−2	−1	2	−2	−2	−1	−1	−1	−1	3	7		Y
W	−2	−3	−2	−4	−3	−2	−4	−4	−3	−2	−2	−3	−3	−1	−3	−2	−3	1	2	11	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

BLOSUM62

Przykład oceny dopasowania

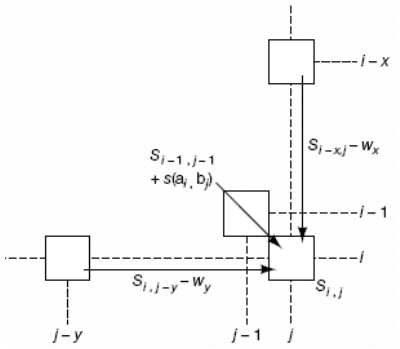
```
sequence 1  V  D  S  -  C  Y
sequence 2  V  E  S  L  C  Y
SCORE       4  2  4 -11  9  7   SCORE = SUM OF AMINO ACID PAIR SCORES
(26)                            MINUS SINGLE GAP PENALTY (11) = 15
```

```
1.  SCORE OF NEW    =   SCORE OF PREVIOUS +  SCORE OF NEW
    ALIGNMENT           ALIGNMENT (A)        ALIGNED PAIR

V  D  S  -  C  Y        V  D  S  -  C           Y
V  E  S  L  C  Y        V  E  S  L  C           Y

      15            =        8             +     7


II. SCORE OF        =   SCORE OF PREVIOUS +  SCORE OF NEW
    ALIGNMENT (A)       ALIGNMENT (B)        ALIGNED PAIR

V  D  S  -  C           V  D  S  -              C
V  E  S  L  C           V  E  S  L              C

      8             =       -1             +     9
```


Algorytm

- Wykorzystywane są 2 pomocnicze macierze
  - macierz "wyników" (ang. scoring matrix) w poszczególnych komórkach zapisywany jest wynik najlepszego możliwego dopasowanie (z uwzględnieniem przerw) osiągniętego na podstawie dotychczas przeanalizowanych sekwencji; wypełniana krokowo na podstawie wyliczonych wcześniej wartości oraz informacji bezpośredniego dopasowania
  - macierz przejść (ang. back-track matrix) - w poszczególnych komórkach zawiera najwyżej ocenione przejście do rozpatrywanej komórki (zapisywane podczas uzupełniania macierzy wyników); możliwa jest sytuacja w której dwa przejścia są jednakowo ocenione
- Po wypełnieniu macierzy, wyszukiwana jest komórka z najwyższym wynikiem w ostatnim wierszu lub kolumnie i posuwając się ku początkowi odtwarzane jest dopasowanie
- Ponadto końcowe przerwy (jedna sekwencja kończy się wcześniej niż druga) mogą być uwzględnione w wyniku

Wyliczanie macierzy wyników



$$S_{ij} = \max \{\ S_{i-1,j-1} + s(a_i b_j),\ \max_{x \geq 1} (S_{i-x,j} - w_x),\ \max_{y \geq 1} (S_{ij-y} - w_y)\ \}$$

	gap	a1	a2	a3	a4
gap	0	1 gap	2 gaps	3 gaps	4 gaps
b1	1 gap	s11	s21		
b2	2 gaps	s12	s22		
b3	3 gaps				
b4	4 gaps				

S_{ij} - wynik dotyczący komórki *i,j*
$s(a_i,b_j)$ - wynik dopasowania znaków a_i i b_j na pozycjach odpowiednio *i,j*
w_x - kara za przerwę długości *x* w pierwszej sekw.
w_y - kara za przerwę długości *y* w drugiej sekw.

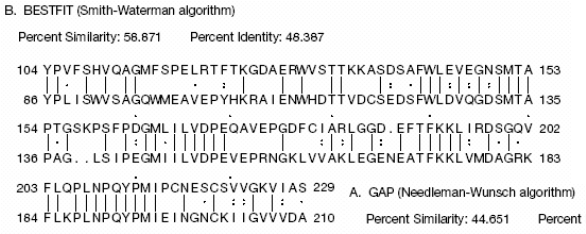
Zmodyfikowany algorytm - dopasowanie lokalne

- Lokalne dopasowania są często bardziej znaczące niż globalne, ponieważ zawierają wzorce zachowane w sekwencjach

$$H_{ij} = \max \{\ H_{i-1,j-1} + s(a_i b_j),\ \max_{x \geq 1} (H_{i-x,j} - w_x),\ \max_{y \geq 1} (H_{i,j-y} - w_y),\ 0\ \}$$

- Różnice dotyczą przede wszystkim wyliczania macierzy wyników:
  - system ocen musi zawierać negatywne wartości dla braków dopasowania
  - gdy wartość macierzy wyników powinna stać się negatywna, wartość jej ustawiana jest na zero (*de facto* oznacza to zakończenie dopasowania w tym miejscu)
- Odcinki lokalnego dopasowania są wyznaczane poprzez prześledzenie przejść rozpoczynając od najwyższych wyników i cofając się do komórki zawierającej zero

Przykład globalnego i lokalnego dopasowania



Oznaczenia:
| - identyczne aminokwasy
: - duże podobieństwo
. - mniejsze podobieństwo

Rozpoczęcie przerwy -11
przedłużenie -8