

Metody bioinformatyki (MBI)

Wykład 2 - podobieństwo sekwencji. Macierze substytucji.

Robert Nowak

2025Z

Powtórzenie : algorytm Needlemana - Wunscha

- ▶ znajduje maksymalną wartość oceny dopasowania
- ▶ znajduje strukturę dla maksymalnej wartości dopasowania

$$F(i, j) = \max \begin{cases} F(i-1, j-1) + e(s_i, t_j) \\ F(i-1, j) + d \\ F(i, j-1) + d \end{cases}$$

	G	A	A	T	T	C	
G	0	-5	-10	-15	-20	-25	-30
A	-5	7	2	-3	-8	-13	-18
A	-10	2	17	12	7	2	-3
T	-15	-3	12	13	20	15	10
T	-20	-8	7	8	21	28	23
A	-25	-13	2	17	16	23	25

	G	A	A	T	T	C	
G	0	-5	-10	-15	-20	-25	-30
A	-5	7	2	-3	-8	-13	-18
T	-10	2	17	12	7	2	-3
T	-15	-3	12	13	20	15	10
A	-20	-8	7	8	21	28	23
A	-25	-13	2	17	16	23	25

Plan wykładu

- ▶ algorytmy programowania dynamicznego do znajdowania podobieństw sekwencji
 - ▶ algorytm znajdujący podobieństwo lokalne Smitha-Waterman
 - ▶ algorytm bez kar dla końców
 - ▶ algorytm dla liniowej złożoności pamięciowej
 - ▶ nieliniowe kary za przerwę
- ▶ algorytmy obliczania macierzy substytucji
 - ▶ macierze BLOSUM
 - ▶ macierze PAM

Podobieństwo globalne i lokalne

- ▶ Globalne – porównywane są całe łańcuchy.
- ▶ Lokalne – porównywane pod-łańcuchy, zastosowanie:
 - ▶ szukanie podobieństw części genów
 - ▶ pomijanie sekwencji niekodujących (introny) przy poszukiwaniu

global



local

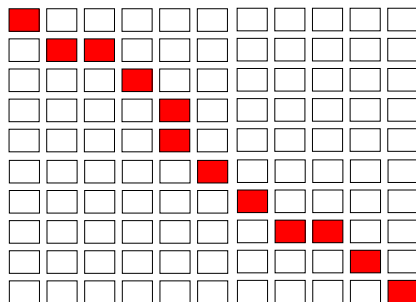


Podobieństwo globalne (algorytm Needlemana-Wunscha)

start: $F(0, 0) = 0$

stop: w komórce (M, N)

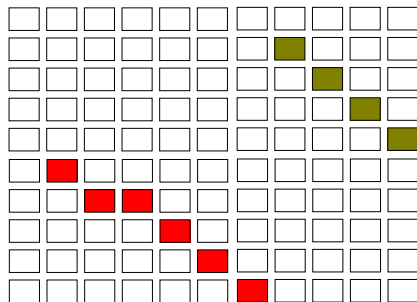
$$F(i, j) = \max \begin{cases} F(i-1, j-1) + e(s_i, t_j) \\ F(i-1, j) + d \\ F(i, j-1) + d \end{cases}$$



Podobieństwo lokalne (algorytm Smitha-Watermana)

start: $F(i,0) = 0, F(0,j) = 0$

$$F(i,j) = \max \begin{cases} F(i-1,j-1) + e(s_i, t_j) \\ F(i-1,j) + d \\ F(i,j-1) + d \\ 0 \end{cases}$$



stop:

- ▶ lokalne podobieństwa o ocenie większej niż z

$$F(i,j) > z$$

- ▶ najlepsze lokalne podobieństwo:

$$\max_{i,j} F(i,j)$$

Podobieństwo lokalne - przykład

BBABB, ABBA;

	A	B
A	2	-1
B	-1	2

; kara za przerwę $d = -3$

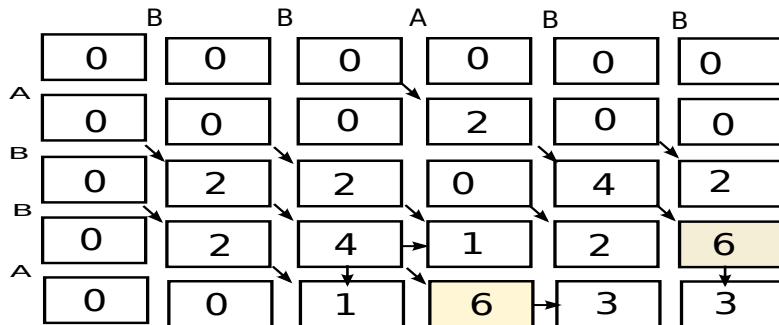
		B	B	A	B	B
A	0	0	0	0	0	0
B	0					
B	0					
B	0					
A	0					
A	0					

Podobieństwo lokalne - przykład

BBABB, ABBA;

	A	B
A	2	-1
B	-1	2

; kara za przerwę $d = -3$

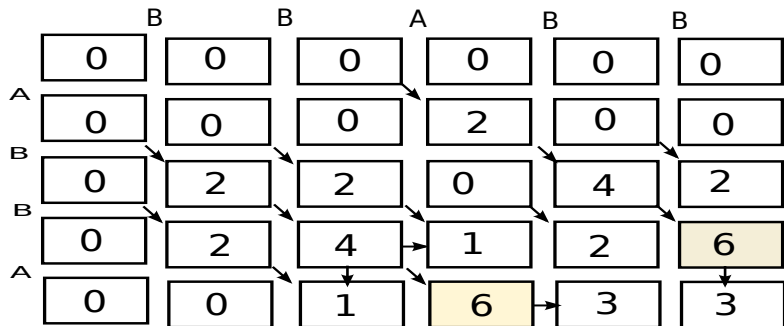


Podobieństwo lokalne - przykład

BBABB, ABBA;

	A	B
A	2	-1
B	-1	2

; kara za przerwę $d = -3$



Rozwiązania:

B B A
B B A

A B B
A B B

podobieństwo globalne - pomijanie przesunięć

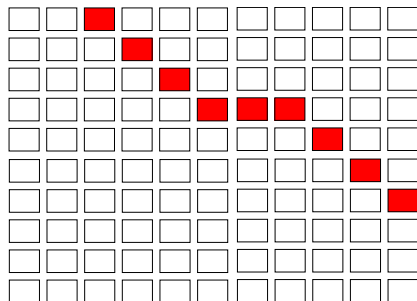
- ▶ nieistotne końcówki sekwencji
- ▶ brak kar za przerwy na początku lub końcu sekwencji

start: $F(0,i) = 0$, $F(j,0) = 0$

$$F(i,j) = \max \begin{cases} F(i-1,j-1) + e(s_i, t_j) \\ F(i-1,j) + d \\ F(i,j-1) + d \end{cases}$$

stop:

$$F(x,y) = \max \begin{cases} \max_{i:0..M} F(i,N) \\ \max_{j:0..N} F(M,j) \end{cases}$$



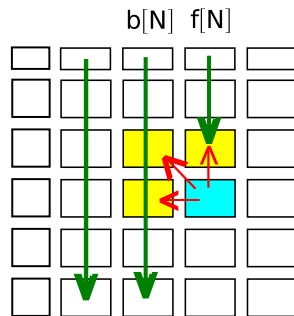
Algorytm o liniowym koszcie pamięciowym

Algorytm Needlemana-Wunscha, złożoność

- ▶ czasowa: $O(m*n)$
- ▶ pamięciowa: $O(m*n)$ - czasami zbyt duża

Maksymalna ocena dopasowania : zależność lokalna, więc można zwalniać pamięć

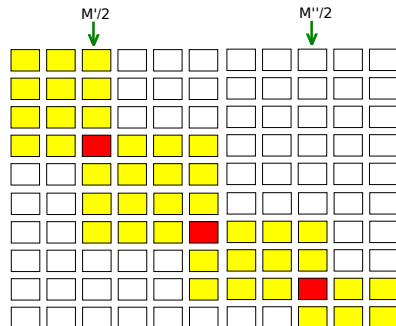
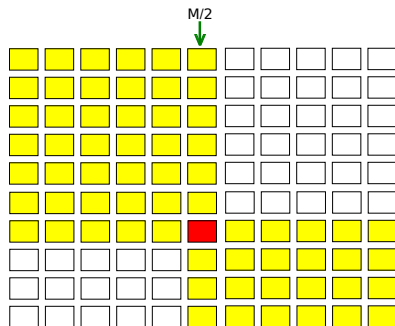
```
int* b = new int[N];  
//inicjacja pierwszej kolumny  
for(int i=1;i<M;++i) {  
    int* f = new int[N];  
    for(int j=0;j<N;++j)  
        f[j] = //oblicza kolumnę f  
    delete [] b; //zwalnia kolumnę  
    b = f;  
}
```



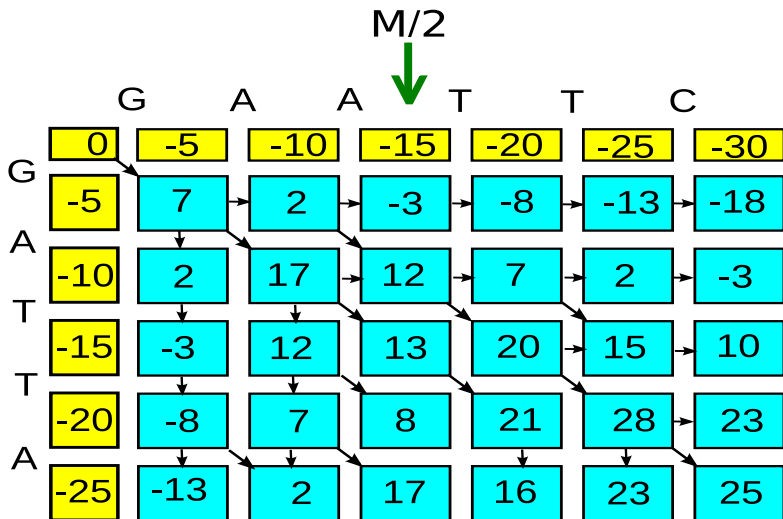
Połączenie o najlepszej ocenie

Jak odtworzyć połączenie (cofanie się w macierzy)?

- ▶ znajdowanie punktu, przez który przechodzi najlepsza ścieżka
- ▶ metoda dziel i zwyciężaj



Znajdowanie punktu na ścieżce



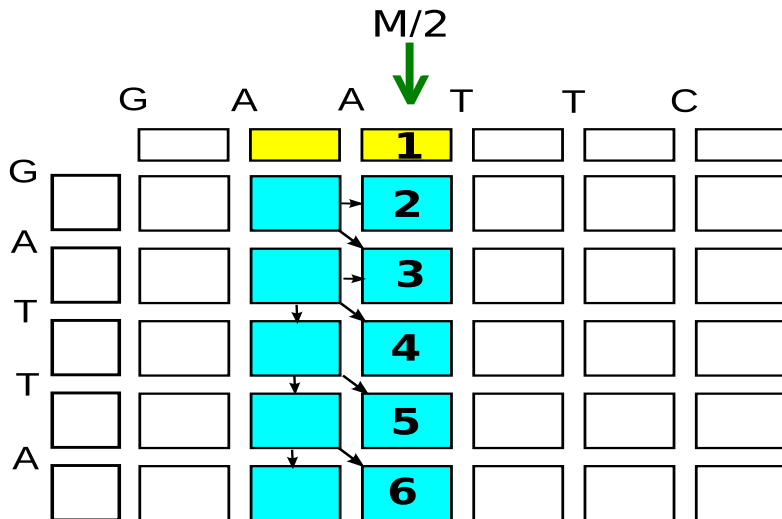
Znajdowanie punktu na ścieżce

M/2

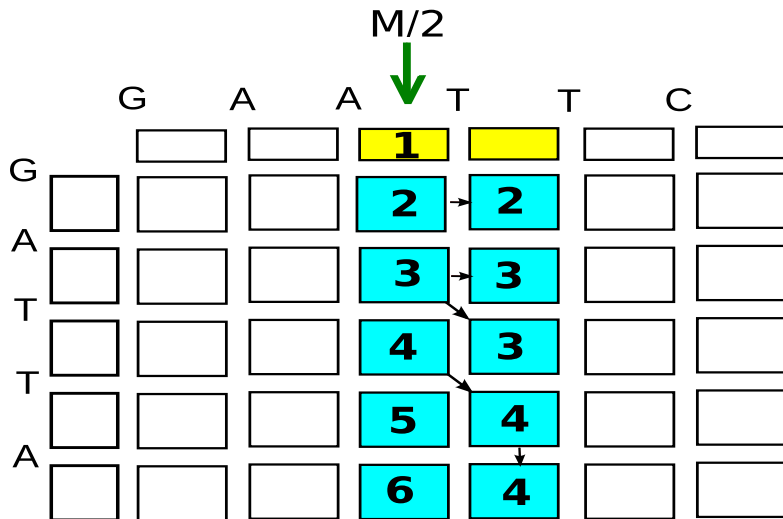


	G	A	A	T	T	C
G		-10	-15			
A		2	-3			
T		17	12			
T		12	13			
A		7	8			
A		2	17			

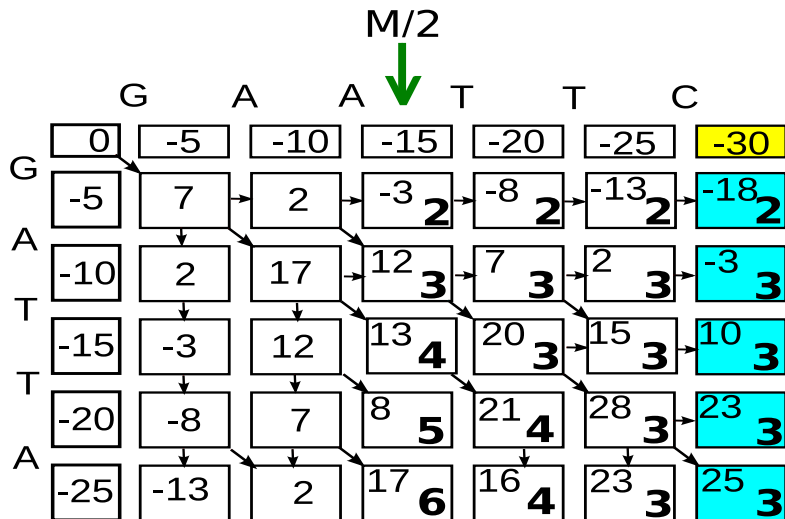
Znajdowanie punktu na ścieżce



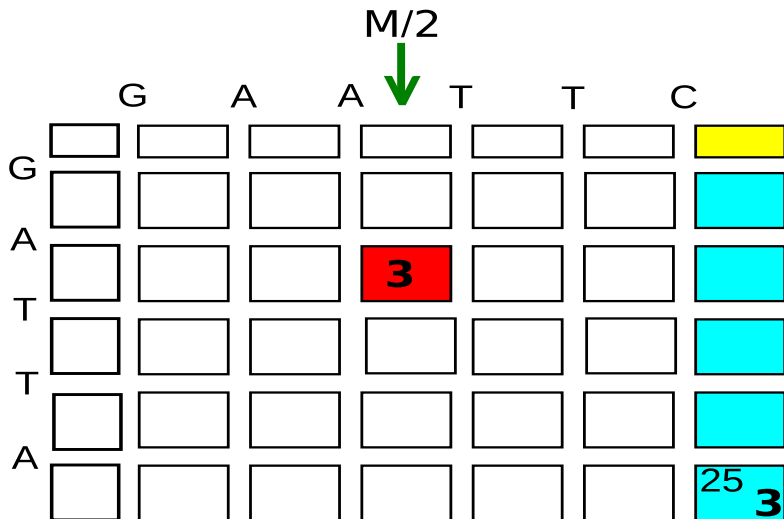
Znajdowanie punktu na ścieżce



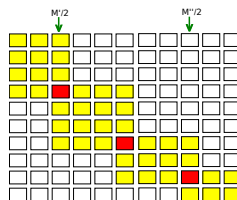
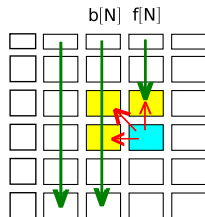
Znajdowanie punktu na ścieżce



Znajdowanie punktu na ścieżce



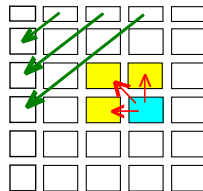
Algorytm w liniowej pamięci: podsumowanie



czas obliczeń: $cMN + cMN/2 + cMN/4 + \dots = 2cMN$

- ▶ złożoność czasowa: $O(M \cdot N)$
- ▶ złożoność pamięciowa: $O(M + N)$

Obliczenia równoległe
(można obliczać cały rząd
jednocześnie):



Kara za przerwę

liniowa kara za przerwę,

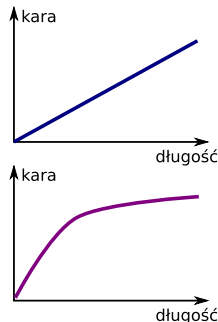
$$\gamma(n) = d * n,$$

bliższy rzeczywistości model (występują wstawienia i usunięcia segmentów o długości > 1),

Algorytm przy ogólnej (stabilizowanej) karze za przerwę:

$$F(i, j) = \max \begin{cases} F(i-1, j-1) + e(s_i, t_j) \\ \max_{k=0, \dots, i-1} F(k, j) + \gamma(i-k) \\ \max_{k=0, \dots, j-1} F(i, k) + \gamma(j-k) \end{cases}$$

Złożoność czasowa takiego algorytmu: $O(n^3)$

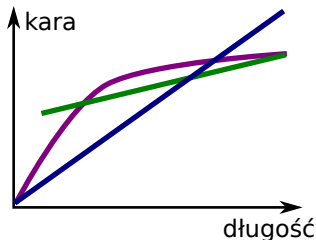
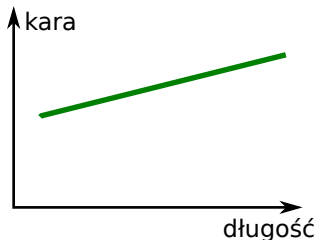


Afiniczna kara za przerwę

$$\gamma(n) = d + (n - 1)c$$

gdzie:

- ▶ d - kara za rozpoczęcie przerwy
- ▶ c - kara za kontynuowanie przerwy ($d < c$)



Algorytm o złożoności czasowej $O(n^2)$

Podobieństwo sekwencji z afiniczną karą za przerwę

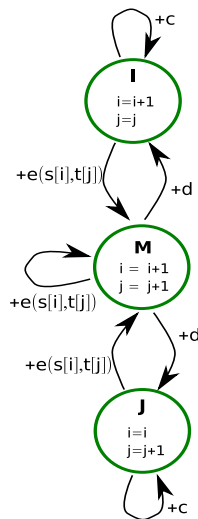
$$\gamma(n) = d + (n - 1)c$$

$$F(i, j) = \max \begin{cases} M(i, j) \\ I(i, j) \\ J(i, j) \end{cases}$$

$$M(i, j) = F(i - 1, j - 1) + e(s_i, t_j)$$

$$I(i, j) = \max \begin{cases} M(i - 1, j) + d \\ I(i - 1, j) + c \\ J(i - 1, j) + d \end{cases}$$

$$J(i, j) = \max \begin{cases} M(i, j - 1) + d \\ I(i, j - 1) + c \\ J(i, j - 1) + c \end{cases}$$



Podobieństwo sekwencji, $\gamma(n) = d + (n - 1)c$, przykład

	G	A	A	T	T	C	
G	0	-2	-4	-6	-8	-10	-12
A	-2						
A	-4						
T	-6						
T	-8						
A	-10						

	A	G	C	T
A	10	-1	-3	-4
G	-1	7	-5	-3
C	-3	-5	9	0
T	-4	-3	0	8

$d = -9$

$c = -2$

Podobieństwo sekwencji, $\gamma(n) = d + (n - 1)c$, przykład

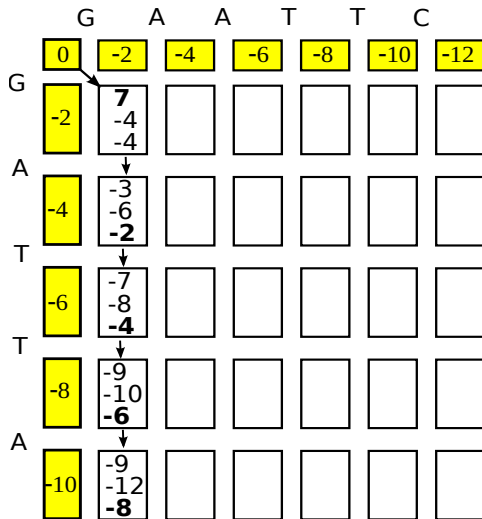
	G	A	A	T	T	C	
G	0	-2	-4	-6	-8	-10	-12
A	-2	7 -4 -4					
T	-4						
T	-6						
A	-8						
A	-10						

	A	G	C	T
A	10	-1	-3	-4
G	-1	7	-5	-3
C	-3	-5	9	0
T	-4	-3	0	8

$d = -9$

$c = -2$

Podobieństwo sekwencji, $\gamma(n) = d + (n - 1)c$, przykład

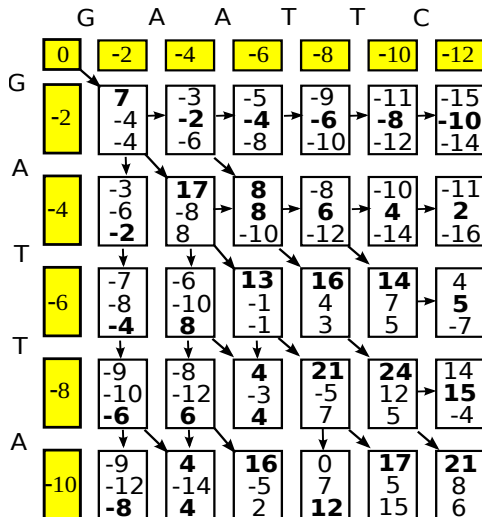


	A	G	C	T
A	10	-1	-3	-4
G	-1	7	-5	-3
C	-3	-5	9	0
T	-4	-3	0	8

$d = -9$

$c = -2$

Podobieństwo sekwencji, $\gamma(n) = d + (n - 1)c$, przykład



	A	G	C	T
A	10	-1	-3	-4
G	-1	7	-5	-3
C	-3	-5	9	0
T	-4	-3	0	8

$d = -9$

$c = -2$

Macierze substytucji

Tworzenie macierzy podobieństwa

Macierz kar i nagród może być tworzona:

- ▶ na podstawie wiedzy eksperckiej,
- ▶ na podstawie modelu probabilistycznego.

Modele probabilistyczne:

- ▶ mutacje (zamiana symboli) są niezależne
- ▶ wprowadzenie przerwy (wstawianie, usuwanie) także niezależne
- ▶ logarytm prawdopodobieństwa : wtedy ocena dla sekwencji jest sumą ocen dla poszczególnych par

Popularne macierze do analizy podobieństwa białek:

- ▶ BLOSUM (Blocks Substitution Matrix),
- ▶ PAM (Point Accepted Mutation).

Macierz podobieństwa: BLOSUM, PAM

Nie uwzględniamy przerw! Oznaczenia:

p_a prawdopodobieństwo występowania symbolu a

q_{ab} prawdopodobieństwo wyst. symboli a i b na tej samej pozycji

Dla $|s| = |t| = n$:

$$P(s, t|random) = \prod_{i=1}^n p_{s_i} * \prod_{j=1}^n p_{t_j} \quad // \text{sekwencje niezależne,}$$

$$P(s, t|match) = \prod_{i=1}^n q_{s_i t_i} \quad // \text{sekwencje zależne,}$$

$$M(s, t) = \log \frac{P(s, t|match)}{P(s, t|random)} = \log \left(\prod_{i=1}^n \frac{q_{s_i t_i}}{p_{s_i} * p_{t_i}} \right) = \sum_{i=1}^n e(s_i, t_i)$$

$$\text{gdzie: } e(s_i, t_i) = \log \left(\frac{q_{s_i t_i}}{p_{s_i} * p_{t_i}} \right)$$

macierze BLOSUM (Henikoff i Henikoff, 1992)

Algorytm uwzględnia duży zbiór sekwencji podobnych w $L\%$, np. BLOSUM62 w 62%, BLOSUM75 w 75%

1. zlicza się częstości A_{ij} występowania symbolu i w danym miejscu na łańcuchu oraz symbolu j w tym samym miejscu w innych łańcuchach
2. szacuje częstość występowania symbolu
3. oblicza funkcję oceny, przybliżając do najbliższej liczby całkowitej

sekwencje:

indeks 0	$A_{CB} = 5$ $A_{BC} = 5$ $A_{CC} = 20$
indeks 1	$A_{BB} = 2$ $A_{BC} = 8$ $A_{CB} = 8$ $A_{CC} = 12$

	A	B	C
A	52	8	10
B	8	58	24
C	10	24	46

BLOSUM - przykład (2)

Estymacja prawdopodobieństwa wystąpienia pary q_{ab}

$$q_{ij} = \frac{A_{ij}}{\sum_{xy} A_{xy}}$$

A_{ij}	A	B	C
A	52	8	10
B	8	58	24
C	10	24	46

$$\sum A_{ij} = 240$$

q_{ij}	A	B	C
A	0.217	0.033	0.042
B	0.033	0.242	0.1
C	0.042	0.1	0.192

Estymacja prawdopodobieństwa wystąpienia symbolu p_i

$$p_i = q_{ii} + \sum_{j \neq i} \frac{q_{ij}}{2}$$

$$p_a = 0.292; p_b = 0.375; p_c = 0.333;$$

BLOSUM - przykład (3)

obliczanie macierzy $e(i, j) = \log\left(\frac{q_{ij}}{p_i p_j}\right)$

najczęściej $e(i, j) = 2 \log_2\left(\frac{q_{ij}}{p_i p_j}\right)$ zaokrąglając do liczb całkowitych

q_{ij}	A	B	C
A	0.217	0.033	0.042
B	0.033	0.242	0.1
C	0.042	0.1	0.192
p_j	0.292	0.375	0.333

$e(i, j)$	A	B	C
A	3	-3	-2
B	-3	2	-1
C	-2	-1	2

BLOSUM62

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

BLOSUM 62

<http://www.ncbi.nlm.nih.gov>

Macierze PAM (Dayhoff, Schwartz, Orcutt, 1978)

PAM (Point Accepted Mutation)

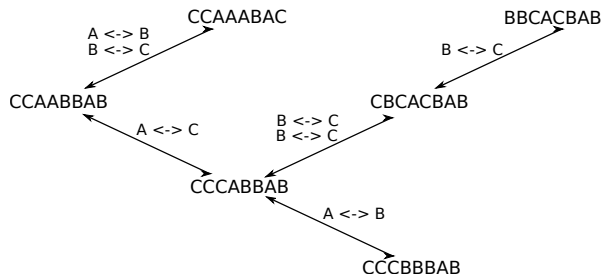
- ▶ inaczej dopasowuje sekwencje, korzysta z drzewa filogenetycznego
- ▶ wykorzystuje podobieństwa dla blisko spokrewnionych białek
- 1. zliczanie podstawienia aminokwasów w drzewie filogenetycznym
- 2. szacowanie prawdopodobieństwa zmiany $i \rightarrow j$ dla małych przedziałów czasowych
- 3. normalizacja macierzy, tworzenie PAM1
- 4. estymacja prawdopodobieństwa dla bardziej oddalonych białek

PAM - przykład

sekwencje:

CCAAABAC
CCAABBAB
BBCACBAB
CBCACBAB
CCCABBAB
CCCBABAB

Drzewo filogenetyczne



Macierz podstawień:

A_{ij}	A	B	C
A	-	2	1
B	2	-	4
C	1	4	-

$$N_A = 14, N_B = 18, N_C = 16$$

$$N = 48$$

$$p_A = 0.292, p_B = 0.375, p_C = 0.333$$

PAM - przykład (2)

Macierz podstawień:

A_{ij}	A	B	C
A	-	2	1
B	2	-	4
C	1	4	-

$$N_A = 14, N_B = 18, N_C = 16$$

$$N = 48$$

$$p_A = 0.292, p_B = 0.375, p_C = 0.333$$

szacowanie prawdopodobieństwa zmiany $i \rightarrow j$

$$P_{i \rightarrow j}(\delta t) = P(j|i) = \frac{A_{ij}}{N_i}$$

$P_{i \rightarrow j}$	A	B	C
A	-	2/14	1/14
B	2/18	-	4/18
C	1/16	4/16	-

macierz nie jest symetryczna (bo $N_i \neq N_j$ dla $i \neq j$)

PAM - przykład (2)

Macierz M_{ij} dla PAM1, odpowiada okresowi ewolucji, w którym podmieniono średnio 1% symboli, więc z definicji

$$\sum_i p_i * \sum_{j \neq i} M_{ij} = 0.01$$

Zakłada się, że dla małych t prawdopodobieństwo podmiiany rośnie liniowo, czyli $M_{ij} = \lambda P_{i \rightarrow j}$

$$\begin{aligned} \sum_i p_i * \sum_{j \neq i} M_{ij} &= \sum_i \sum_{j \neq i} p_i * \lambda P_{i \rightarrow j} = \sum_i \sum_{j \neq i} \frac{N_i}{N} \lambda \frac{A_{ij}}{N_i} = \\ &= \lambda \frac{\sum_i \sum_{j \neq i} A_{ij}}{N} = 0.01 \text{ więc } \lambda = 0.01 \frac{N}{\sum \sum A_{ij}} \end{aligned}$$

PAM - przykład (3)

Macierz podstawień:

$N = 48$

A_{ij}	A	B	C
A	-	2	1
B	2	-	4
C	1	4	-

$P_{i \rightarrow j}$	A	B	C
A	-	2/14	1/14
B	2/18	-	4/18
C	1/16	4/16	-

$$\lambda = 0.01 \frac{N}{\sum \sum A_{ij}} = 0.01 \frac{48}{14} = 0.034$$

$$M_{ii} = 1 - \sum_{j \neq i} M_{ij}$$

M_{ij}	A	B	C
A	0.9927	0.0049	0.0024
B	0.0038	0.9886	0.0076
C	0.0021	0.0086	0.9893

PAM - przykład (4)

M_{ij}	A	B	C
A	0.9927	0.0049	0.0024
B	0.0038	0.9886	0.0076
C	0.0021	0.0086	0.9893
p_j	0.292	0.375	0.333

$$q_{ij} = M_{ij} * p_i$$

$$e(i, j) = \log\left(\frac{q_{ij}}{p_i p_j}\right) = \log\left(\frac{M_{ij}}{p_j}\right)$$

$$e(i, j) = 2 \log_2 \frac{q_{ij}}{p_i p_j}$$

PAM1	A	B	C
A	4	-13	-14
B	-13	3	-11
C	-14	-11	3

PAM - przykład (5)

Ekstrapolacja modelu na większe odległości ewolucyjne:
dla t równego przyjętej jednostce (takiej w której podmianie ulega 1% symboli)

$$P(j|i, t = 1) = M_{ij}$$

Przejście $A \rightarrow B$ w dwóch krokach to

$$\begin{array}{l} A \rightarrow A \rightarrow B \\ A \rightarrow B \rightarrow B \\ A \rightarrow C \rightarrow B \end{array}$$

dlatego

$$P(j|i, t = 2) = \sum_k M_{ik} M_{kj} = (M_{ij})^2$$

$$P(j|i, t = n) = (M_{ij})^n$$

Macierz PAM250

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

Obliczone przez skrypt: <http://www.bioinformatics.nl/cgi-bin/pam/csh>

Wybór macierzy substytucji

Musimy założyć odległość w czasie dla sekwencji, które porównujemy lub wyszukujemy

- ▶ podobne sekwencje nieznacznie oddalone w czasie mają q_{ij} dla $i \neq j$ bardzo małe, $e(i,j)$ ma duże liczby ujemne poza przekątną
- ▶ podobne sekwencje znacznie oddalonych w czasie mają q_{ij} bliskie $p_i * p_j$, wartości $e(i,j)$ nieznacznie oscylują wokół zera

Wyższe numery w macierzach PAM oznaczają większy dystans ewolucyjny, natomiast w macierzach BLOSUM – mniejszy dystans.

BLOSUM80	PAM1
BLOSUM62	PAM120
BLOSUM45	PAM250

Dziękuję