

wykład 7

Dopasowania wielosekwencyjne

dr Jacek Śmietański

jacek.smietanski@ii.uj.edu.pl

<http://jaceksmietanski.net>

1. Dopasowania wielosekwencyjne – wprowadzenie
2. Algorytmy MSA (optymalny i suboptymalne)
3. Wykorzystanie metod klasteryzacji
4. Dostępne narzędzia
5. Motywy, wzorce, profile

Dopasowania wielosekwencyjne

Definicja:

Niech S_0, S_1 – sekwencje nad alfabetem Σ o długościach odpowiednio l_0 i l_1 . **Globalnym dopasowaniem pary sekwencji** (*global pairwise sequence alignment*) nazywamy macierz M o wymiarach $2 \times n$, gdzie $n \geq \max\{l_0, l_1\}$ taką, że

$\forall 0 \leq k \leq 1, 0 \leq i \leq n-1$:

$M[k][i] = -$ lub $M[k][i] = S_k[p], p \in \{0, \dots, l_k - 1\}$

$M[k][i] = - \Rightarrow M[1-k][i] \neq -$

$\forall i < j \ (M[k][i] = S_k[p] \text{ i } M[k][j] = S_k[q] \Rightarrow p < q)$

$\forall 0 \leq p \leq l_k - 1 \ \exists j \in \{0, \dots, n-1\} : M[k][j] = S_k[p]$

Przykład: $S_0 = \text{TAGACTAG}$
 $S_1 = \text{ACGTATG}$

T	A	-	G	A	C	T	A	-	G
-	A	C	G	-	-	T	A	T	G

T	A	G	A	C	T	A	-	G
A	C	G	-	-	T	A	T	G

Niech $\mathbf{S}=\{\mathbf{S}_0, \dots, \mathbf{S}_{t-1}\}$ – zbiór t -sekwencji nad alfabetem Σ takich, że $|\mathbf{S}_i|=l_i$ dla $i \in \{0, \dots, t-1\}$.

Globalnym dopasowaniem zbioru sekwencji (*global multiple sequence alignment*) nazywamy macierz \mathbf{M} o wymiarach $t \times n$, gdzie $n \geq \max\{l_0, \dots, l_{t-1}\}$ taką, że

$\forall 0 \leq k \leq t-1, 0 \leq i \leq n-1$:

$M[k][i] = -$ lub $M[k][i] = S_k[p], p \in \{0, \dots, l_k-1\}$

$\exists r \in \{0, \dots, n-1\} : M[r][i] \neq -$

$\forall 0 \leq i < j \leq n-1 \quad (M[k][i] = S_k[p] \wedge M[k][j] = S_k[q] \Rightarrow p < q)$

$\forall 0 \leq p \leq l_k-1 \exists j \in \{0, \dots, n-1\} : M[k][j] = S_k[p]$

(**MSA** = *multiple sequence alignment*)

Definicja:

Lokalnym dopasowaniem zbioru sekwencji (*local multiple sequence alignment*) nazywamy globalne dopasowanie zbioru podsekwencji $\mathbf{S}_k[i_k \dots, j_k] \forall 0 \leq k \leq t-1, 0 \leq i_k \leq j_k \leq l_k$

1. Określanie powiązań filogenetycznych między sekwencjami.
2. Poszukiwanie odległych homologów.
3. Poszukiwanie wspólnych, konserwowanych wzorów, motywów i domen w sekwencjach, odpowiedzialnych za odpowiednie funkcje biochemiczne lub strukturę przestrzenną.
4. Grupowanie białek w rodziny o wspólnej funkcji biochemicznej lub historii ewolucyjnej. Identyfikowanie członków rodzin białek.
5. Identyfikowanie zachodzących fragmentów sekwencji powstałych w wyniku losowego sekwencjonowania genomów i ułatwienie ich składania w jedną całą sekwencję.

$$\text{score}(M) = \sum_{i=0}^{n-1} \delta_{\text{obj}}(M[0][i], \dots, M[t-1][i])$$

wartość MSA = suma wartości dopasowań na poszczególnych pozycjach

$$\delta_{\text{obj}}(M[0][i], \dots, M[t-1][i]) = \sum_{r=0}^{k-2} \sum_{s=r+1}^{k-1} \delta_{\text{pair}}(M[r][i], M[s][i])$$

sumujemy wartości dopasowań dla wszystkich możliwych par sekwencji

$$\delta_{\text{pair}}(M[j][i], M[k][i]) = \begin{cases} g, & \text{gdy } M[j][i] = - \text{ lub } M[k][i] = - \\ \alpha, & \text{gdy } M[j][i] = M[k][i] \\ \beta, & \text{gdy } M[j][i] \neq M[k][i] \end{cases}$$

wartość dopasowania dla pojedynczej pary sekwencji liczymy jak w algorytmie N-W

Przykład – dopasowanie czterech sekwencji

$$M_3$$

	0	1	2	3	4	5	6	7	8	9	10	11
0	P	Y	R	F	T	-	-	-	I	K	S	M
1	P	Y	K	F	-	-	-	S	I	K	S	M
2	P	Y	M	Y	-	-	-	S	S	E	S	M
3	P	M	D	D	N	P	F	S	F	Q	S	M

A global MSA of $S = \{\text{PYRFTIKSM}, \text{PYKFSIKSM}, \text{PYMYSSSESM}, \text{PMDDNPFSFQSM}\}$.

$$M_{0,1}$$

P	Y	R	F	T	-	I	K	S	M
P	Y	K	F	-	S	I	K	S	M

$$M_{0,2}$$

P	Y	R	F	T	-	I	K	S	M
P	Y	M	Y	-	S	S	E	S	M

$$M_{0,3}$$

P	Y	R	F	T	-	-	-	I	K	S	M
P	M	D	D	N	P	F	S	F	Q	S	M

$$M_{1,2}$$

P	Y	K	F	S	I	K	S	M
P	Y	M	Y	S	S	E	S	M

$$M_{1,3}$$

P	Y	K	F	-	-	-	S	I	K	S	M
P	M	D	D	N	P	F	S	F	Q	S	M

$$M_{2,3}$$

P	Y	M	Y	-	-	-	S	S	E	S	M
P	M	D	D	N	P	F	S	F	Q	S	M

$$\delta_{obj}(M[0][i], \dots, M[t-1][i]) = \sum_{r=0}^{k-1} \delta_{pair}(M[r][i], cons(i))$$

$$cons(i) = \arg \max_{c \in \Sigma \cup \{-\}} \left(\sum_{r=0}^{k-1} \delta_{pair}(M[r][i], c) \right)$$

M_3

	0	1	2	3	4	5	6	7	8	9	10	11
0	P	Y	R	F	T	-	-	-	I	K	S	M
1	P	Y	K	F	-	-	-	S	I	K	S	M
2	P	Y	M	Y	-	-	-	S	S	E	S	M
3	P	M	D	D	N	P	F	S	F	Q	S	M

A global MSA of $S = \{\text{PYRFTIKSM}, \text{PYKFSIKSM}, \text{PYMYSSSESM}, \text{PMDDNPFSFQSM}\}$.

Sekwencja konsensusowa dla powyższego przykładu: **PYKF---SIKSM**

Punktacja dla powyższej sekwencji konsensusowej:

$$4 + 2 - 2 + 0 - 4 - 2 - 2 + 1 + 0 + 0 + 4 + 4 = 5$$

Algorytmy MSA (optymalne i suboptymalne)

L – długość sekwencji

Optymalne dopasowanie dwóch sekwencji:

macierz punktacji kwadratowa, wymagany czas: L^2

Optymalne dopasowanie trzech sekwencji:

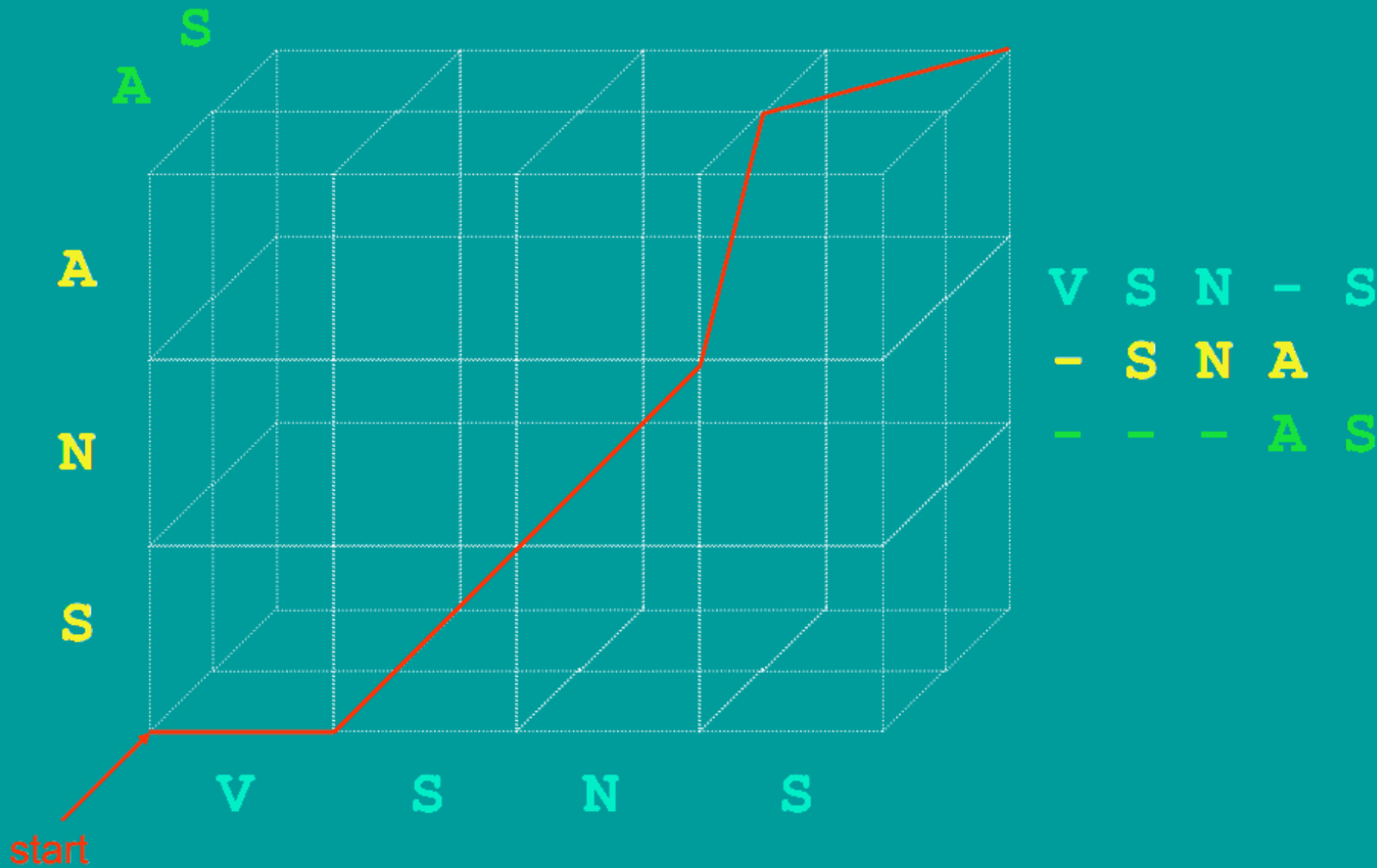
macierz punktacji sześcienna; wymagany czas: L^3

Optymalne dopasowanie N sekwencji:

macierz punktacji N-wymiarowa: wymagany czas: L^N
(rośnie wykładniczo ze wzrostem liczby sekwencji N)



Szukanie optymalnego dopasowania w objętości sześcianu.



Macierz punktacji będzie trójwymiarowa.

Konstrukcja macierzy punktacji:

$$H[i][j][k] = \max \begin{cases} H[i-1][j-1][k-1] + \delta_{\text{obj}}(S_0[i], S_1[j], S_2[k]) \\ H[i-1][j-1][k] + \delta_{\text{obj}}(S_0[i], S_1[j], -) \\ H[i-1][j][k-1] + \delta_{\text{obj}}(S_0[i], -, S_2[k]) \\ H[i][j-1][k-1] + \delta_{\text{obj}}(-, S_1[j], S_2[k]) \\ H[i-1][j][k] + \delta_{\text{obj}}(S_0[i], -, -) \\ H[i][j-1][k] + \delta_{\text{obj}}(-, S_1[j], -) \\ H[i][j][k-1] + \delta_{\text{obj}}(-, -, S_2[k]) \end{cases}$$

Warunki początkowe:

$$H[i][j][k] = \begin{cases} g \cdot k, & \text{gdy } i = 0, j = 0, k \geq 0 \\ g \cdot j, & \text{gdy } i = 0, j > 0, k = 0 \\ g \cdot i, & \text{gdy } i > 0, j = 0, k = 0 \end{cases}$$

k – liczba sekwencji

l – średnia długość sekwencji

Programowanie dynamiczne (optymalne)

liczba komórek macierzy: $O(l^k)$

liczba porównań dla każdej komórki: $O(2^k)$

koszt obliczenia punktacji pojedynczej komórki: $O(k)$

Złożoność obliczeniowa:

$O(k 2^k l^k)$

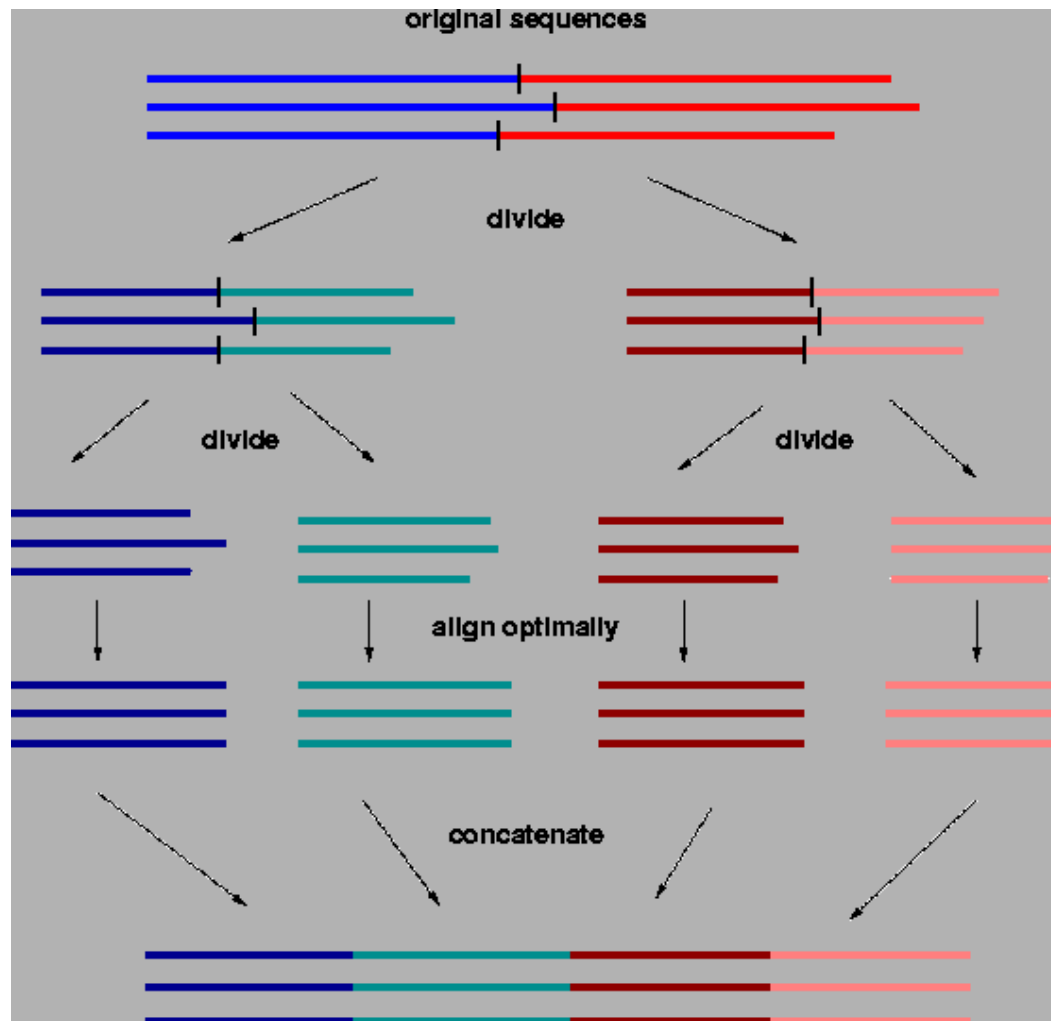
Wniosek:

DP dla MSA może być stosowane dla co najwyżej kilku sekwencji.

W praktyce używa się metod heurystycznych (suboptymalnych) o złożoności wielomianowej, zwykle pomiędzy $O(k^2l)$ a $O(k^2l^2)$.

DCA (*Divide-and-Conquer Multiple Sequence Alignment*)

<http://bibiserv.techfak.uni-bielefeld.de/dca/>



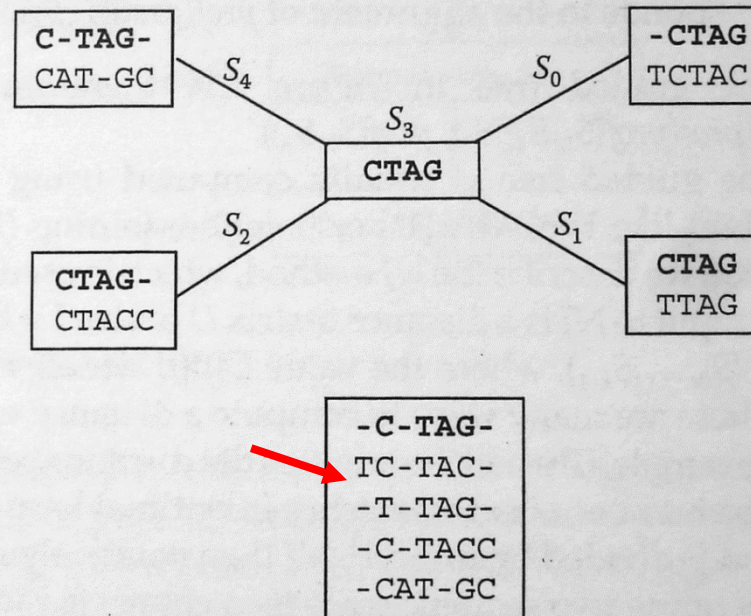
1. Obliczamy optymalne dopasowanie dla każdej pary sekwencji (niezależnie od siebie).
2. Na podstawie uzyskanych rezultatów wyznaczamy sekwencję centralną (sekwencja o najwyższej wartości dopasowania).
3. MSA uzyskujemy na podstawie optymalnego dopasowania każdej sekwencji do sekwencji centralnej.

Metoda suboptymalna.

Metoda gwiazdy - przykład

$S_0 = \text{TCTAC}$
 $S_1 = \text{TTAG}$
 $S_2 = \text{CTACC}$
 $S_3 = \text{CTAG}$
 $S_4 = \text{CATGC}$

	S_0	S_1	S_2	S_3	S_4	Sum
S_0		+1	+2	+1	-1	+3
S_1	+1		-1	+2	-2	0
S_2	+2	-1		+1	0	+2
S_3	+1	+2	+1		0	+4
S_4	-1	-2	0	0		-3

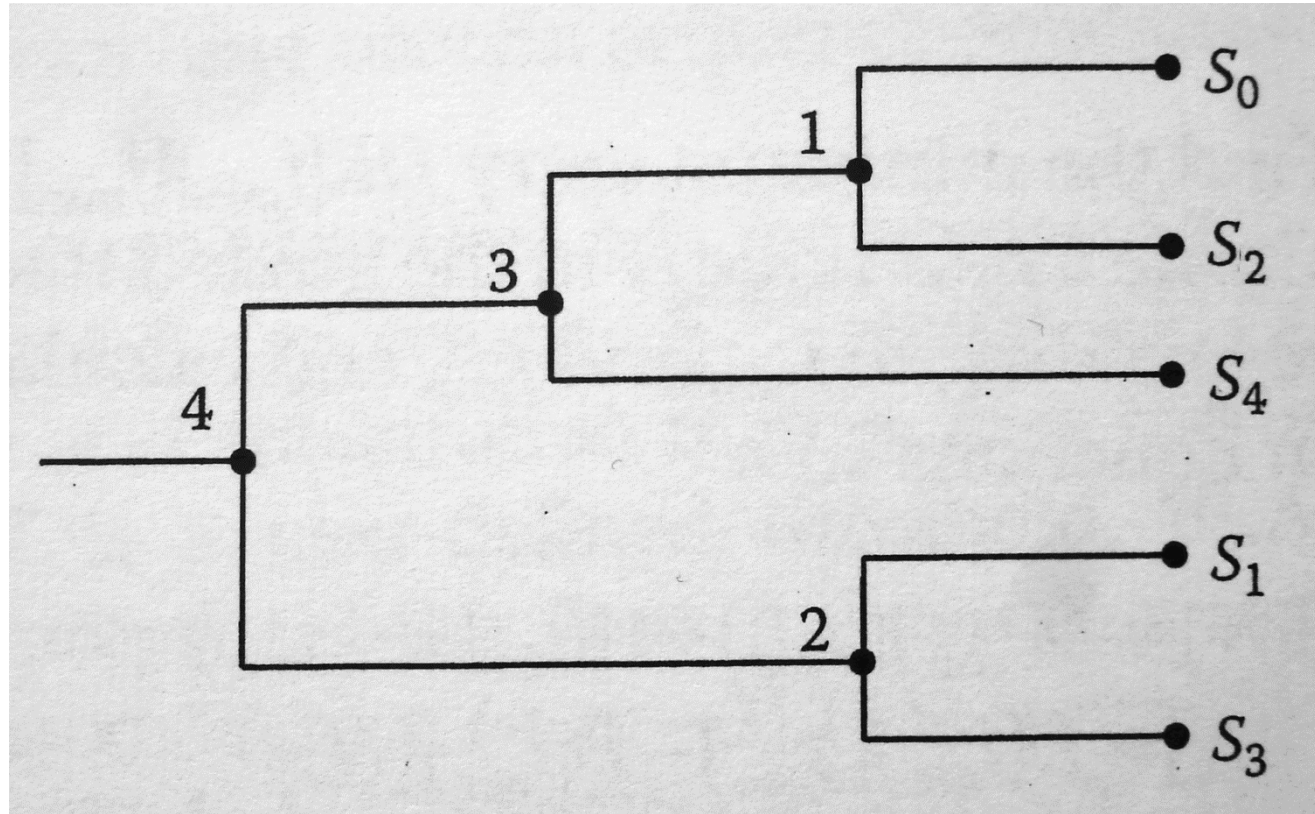


zgodność +1
niezgodność -1
przerwa: -1

Wykorzystanie metod klasteryzacji

Opiera się na konstrukcji drzewa naprowadzającego (przewodniego;
guided tree):

$S_0 = \text{TCTAC}$
 $S_1 = \text{TTAG}$
 $S_2 = \text{CTACC}$
 $S_3 = \text{CTAG}$
 $S_4 = \text{CATGC}$



- 1,2. Na podstawie optymalnego dopasowania pary sekwencji wyznaczamy jego profil.
3. Dopasowanie sekwencja – profil.
4. Dopasowanie profil – profil.

Metody klasteryzacji bazujące na odległości.

Clustal używa metody NJ (*neighbor-joining*)

D – macierz odległości

(wymiar $k \times k$, k – liczba dopasowywanych sekwencji)

$D[i][j]$ – odległość między S_i a S_j

Macierz D może być wyznaczana na różne sposoby.

Metodą NJ iteracyjnie wybieramy wejście w D . Odpowiadające mu sekwencje zostają połączone węzłem w drzewie.

Odpowiednie wiersze w D zostają połączone, tworząc nową, mniejszą macierz dla kolejnego kroku iteracji.

$D[i][j]$ – liczba zgodności w najlepszym lokalnym dopasowaniu na ścieżce wychodzącej z S_i i S_j , podzielona przez $\min\{l_i, l_j\}$.

Iteracja:

1. Wyznaczenie macierzy DR (*rate-corrected distance matrix*)

$$DR[i][j] = D[i][j] - (r[i] + r[j]) \quad \text{with} \quad r[q] = \frac{1}{k-2} \sum_{p=0}^{k-1} D[q][p]$$

2. Znajdujemy najmniejszą wartość w DR.

3. Tworzymy nowy wierzchołek w drzewie, łączący wejścia odpowiadające i_{\min} i j_{\min} , przeliczamy odległości:

$$D[i_{\min}][N] = (D[i_{\min}][j_{\min}] + r[i_{\min}] - r[j_{\min}]) / 2$$

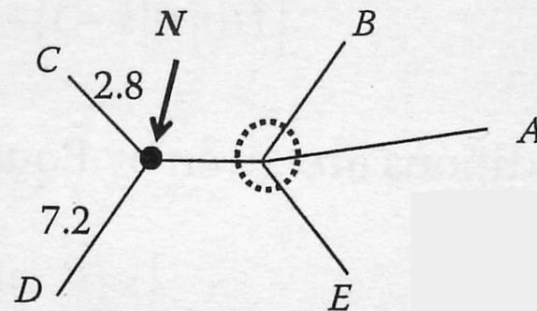
$$D[j_{\min}][N] = (D[i_{\min}][j_{\min}] - r[i_{\min}] + r[j_{\min}]) / 2$$

$$D[x][N] = (D[i_{\min}][x] + D[j_{\min}][x] - D[i_{\min}][j_{\min}]) / 2$$

4. Zastępujemy wiersze i kolumny i_{\min} i j_{\min} w D nowym wierszem i kolumną reprezentującymi odległości do N .

Metoda NJ (Clustal) - przykład

	$D[][]$						$r[]$		$DR[][]$				
A	0	3	6	9	12		10		0				
B	3	0	5	9	11		9.3		-16.3	0			
C	6	5	0	10	13		11.3		-15.3	-15.7	0		
D	9	9	10	0	19		15.7		-16.7	-16.0	-17.0	0	
E	12	11	13	19	0		18.3		-16.3	-16.7	-16.7	-15.0	0



	$D[][]$			
A	0	3	2.5	12
B	3	0	2	11
N	2.5	2	0	11
E	12	11	11	0

Dla danej sekwencji S o długości l i MSA o wymiarze kn definiujemy macierz częstości P :

$$P[c][j] = \frac{|\{i | M[i][j] = c\}|}{k}$$

Wartość dopasowania litery c z profilem w kolumnie j określa równanie:

$$t(c, j) = \begin{cases} \alpha \cdot P[c][j] + \beta \cdot \sum_{b \in \Sigma \setminus \{c\}} P[b][j] + g \cdot P[-][j] & \text{if } c \neq - \\ g \cdot \sum_{b \in \Sigma} P[b][j] & \text{if } c = - \end{cases}$$

Wartość dopasowania sekwencja-profil = suma wartości dla poszczególnych kolumn.

Dopasowanie sekwencja-profil (przykład)

M_1

A	G	C	-	A
A	G	A	G	A
A	T	C	G	-
C	G	-	G	C

P

A	0.75	0.00	0.25	0.00	0.50
C	0.25	0.00	0.50	0.00	0.25
G	0.00	0.75	0.00	0.75	0.00
T	0.00	0.25	0.00	0.00	0.00
-	0.00	0.00	0.25	0.25	0.25

M_2

A	A	C	-	G	C
1	-	2	3	4	5

$$\begin{aligned}
 \text{score}(M_2) &= t(A,1) + g + t(C,2) + t(-,3) + t(G,4) + t(C,5) \\
 &= (2 \cdot 0.75 - 3 \cdot 0.25) + (-1) + (-3 \cdot (0.75 + 0.25)) + (-1 \cdot (0.25 + 0.75)) + (2 \cdot 0.75 - 1 \cdot 0.25) + \\
 &\quad (2 \cdot 0.25 - 3 \cdot 0.5 - 1 \cdot 0.25) = -4.25
 \end{aligned}$$

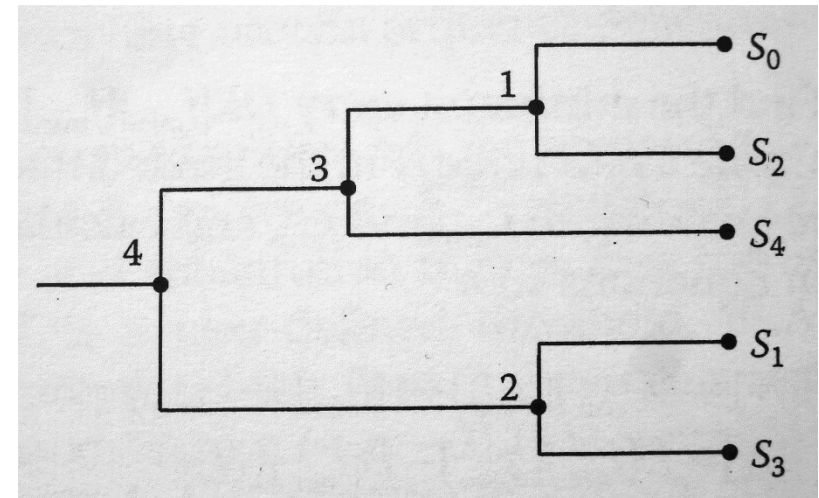
A sequence-profile alignment M_2 of the DNA sequence AACGC to the letter frequency matrix P of the MSA M_1 . The score of M_2 is -3.4875 using the scoring scheme $\alpha = +2$, $\beta = -3$, and $g = -1$.

$$H[i][j] = \max \begin{cases} H[i-1][j-1] + t(S_0[i-1], j-1) \\ H[i-1][j] + g \\ H[i][j-1] + t(-, j-1) \end{cases}, \text{ for all } 1 \leq i \leq l \text{ and } 1 \leq j \leq n$$

Warunki początkowe: $H[i][0] = ig$, $H[0][j] = H[0][j-1] + t(-, j-1)$

Złożoność: $O(l n |\Sigma|)$

1. Przyrównanie sekwencji parami i utworzenie na ich podstawie macierzy odległości
2. Obliczenie drzewa naprowadzającego
3. Przyrównanie dwóch najbliższych sekwencji (profilu) (programowanie dynamiczne)
4. Dwie przyrównywane sekwencje (profile) zastępujemy nowym profilem i powtarzamy procedurę od punktu 1. Kończymy, gdy pozostanie nam jeden wspólny profil.



1. Zamiast wyznaczać profile, możemy po każdym przyrównaniu określić sekwencję konsensusową.
2. Zwykle dopasowujemy sekwencje aminokwasowe (a nie nukleotydowe) – dlaczego?
3. W dopasowaniu wykorzystywane są macierze substytucji. Wybór macierzy może się zmieniać dynamicznie, w zależności od wyznaczonej odległości pomiędzy rozważanymi sekwencjami.
4. Znaczenie ważenia – zwiększenie wiarygodności przyrównania sekwencji, które uległy dywergencji.

1. Metoda oparta na globalnym przyrównaniu, dlatego analizowane sekwencje powinny być podobnej długości.
2. Kumulacja błędów – jeśli pojawi się na wczesnym etapie, kolejne przyrównania bazują na błędnych sekwencjach.

Dostępne narzędzia

Starsze programy:

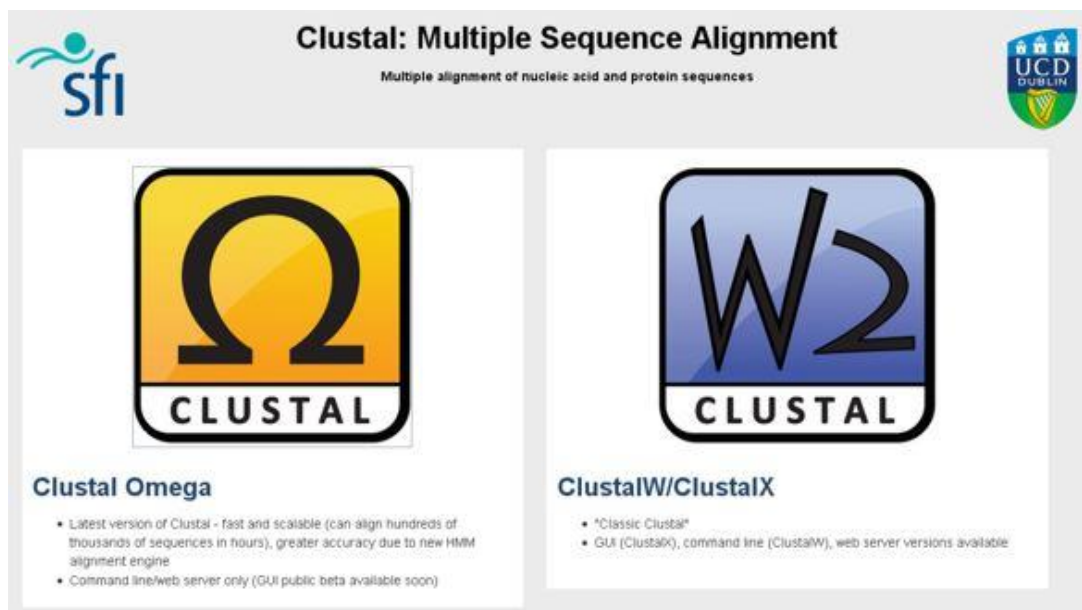
ClustalW – standardowa implementacja

ClustalX – interfejs graficzny

Nowa wersja (uwzględniająca HMM przy dopasowywaniu profili):

Clustal Omega

www.clustal.org



The screenshot shows the Clustal website with the title "Clustal: Multiple Sequence Alignment" and the subtitle "Multiple alignment of nucleic acid and protein sequences". It features two main logos: Clustal Omega (a yellow square with a black Omega symbol) and ClustalW/ClustalX (a blue square with a black W and a 2). Below each logo is a description of the software.

Clustal Omega

- Latest version of Clustal - fast and scalable (can align hundreds of thousands of sequences in hours), greater accuracy due to new HMM alignment engine
- Command line/web server only (GUI public beta available soon)

ClustalW/ClustalX


- "Classic Clustal"
- GUI (ClustalX), command line (ClustalW), web server versions available

<http://tcoffee.crg.cat/>







Swiss Institute of Bioinformatics



[HOME](#) | [references](#) | [help](#) | 

A collection of tools for Computing, Evaluating and Manipulating Multiple Alignments of DNA, RNA, Protein Sequences and Structures

[Try our new Beta Tcoffee server !](#)

Mirror sites:       

ALIGNMENT				
TCOFFEE	Regular	Advanced	cite	?
EXPRESSO(3DCoffee)	Regular	Advanced	cite	?
MCOFFEE	Regular	Advanced	cite	?
RCOFFEE	Regular	Advanced	cite	?
COMBINE	Regular	Advanced	cite	?
EVALUATION				
CORE	Regular	Advanced	cite	?
iRMSD-APDB	Regular	Advanced	cite	?
PROCESSING				
PROTOGENE	Regular	Advanced	cite	?

Algorytm podobny do Clustal (metoda progresywna).

Jest dokładniejszy (dopasowanie początkowe wybierane jest spośród wielu przyrównań), ale przez to wolniejszy niż Clustal.

Przeprowadza zarówno globalne, jak i lokalne dopasowania.

Posiada predefiniowane parametryzacje dedykowane dla DNA, RNA, Białek oraz odmiany celowane, np. SARA-Caffee dopasowuje sekwencje RNA, dla których znana jest struktura trzeciorzędowa.

<http://www.ebi.ac.uk/Tools/msa/muscle/>

Metoda progresywna, dopasowanie trójetapowe:

1. *Draft* – szybkie wstępne zestawienie
2. *Improved* – dokładniejsze, z wykorzystaniem odległości Kimury
3. *Refinement* – optymalizacja

Zwykle daje dokładniejsze wyniki niż Clustal, jest też szybki.

<http://www.ibi.vu.nl/programs/pralinewww/>

PRALINE multiple sequence alignment



• SOAP service (WSO) now available.

[PRALINE sample output](#)

[References and FAQs](#)

PRALINE is a multiple sequence alignment program with many options to optimise the information for each of the input sequences: e.g. homology-extended alignment, predicted secondary structure and/or transmembrane structure information and iteration capabilities.

Paste in your PROTEIN sequences in FASTA format (MAX 500 sequences, length 2000):

Or upload a FASTA file (MAX 500 sequences, length 2000):

Enter a name for your job:

Options

Exchange weights matrix: Associated gap penalties:

[Help](#) Open Extension [Help](#)

Progressive alignment strategy: [Help](#)

☒ PSI-BLAST pre-profile processing (Homology-extended alignment)

PSI-BLAST iterations at an E-value cut-off of DB

☐ Global pre-profile processing

at a score cut-off of iterations

☐ Standard progressive strategy

Structural features: [Help](#)

DSSP-defined secondary structure search: ☒ YES ☐ NO [Help](#)

Secondary structure prediction: [Help](#)

Transmembrane structure prediction: [Help](#)

Output customization:

Tree representation final alignment: ☐ YES ☒ NO [Help](#)

Customize alignment colours: ☐ YES ☒ NO [Help](#)

File format final alignment: ☐ No file ☐ MSF ☒ FASTA [Help](#)

E-mail

E-mail me when my job is done at:

Submit

Ocenia profile metodą najbliższego sąsiada.

Pozwala wykorzystać informację o strukturze drugorzędowej.

Dokładniejszy (i wolniejszy) niż ClustalW.

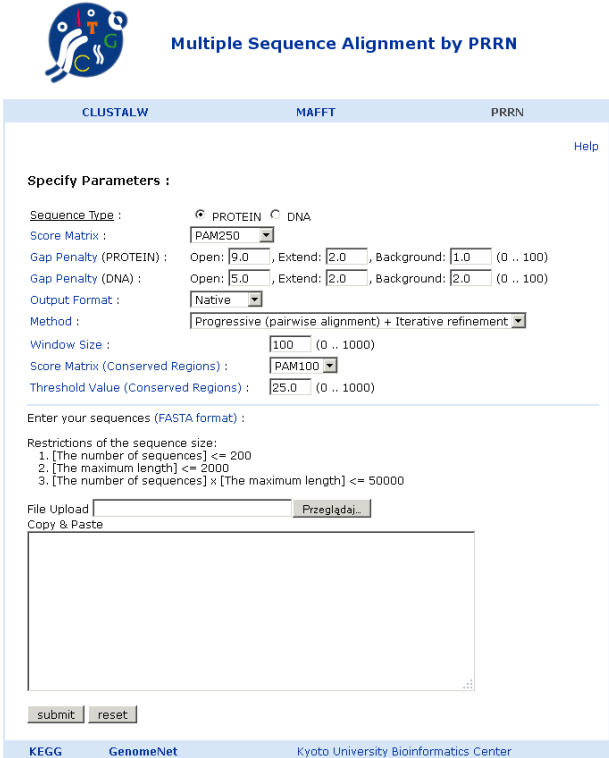
Idea:

tworzymy dopasowanie (początkowo niskiej jakości),
które w kolejnych krokach poprawiamy.

Przykład narzędzia:

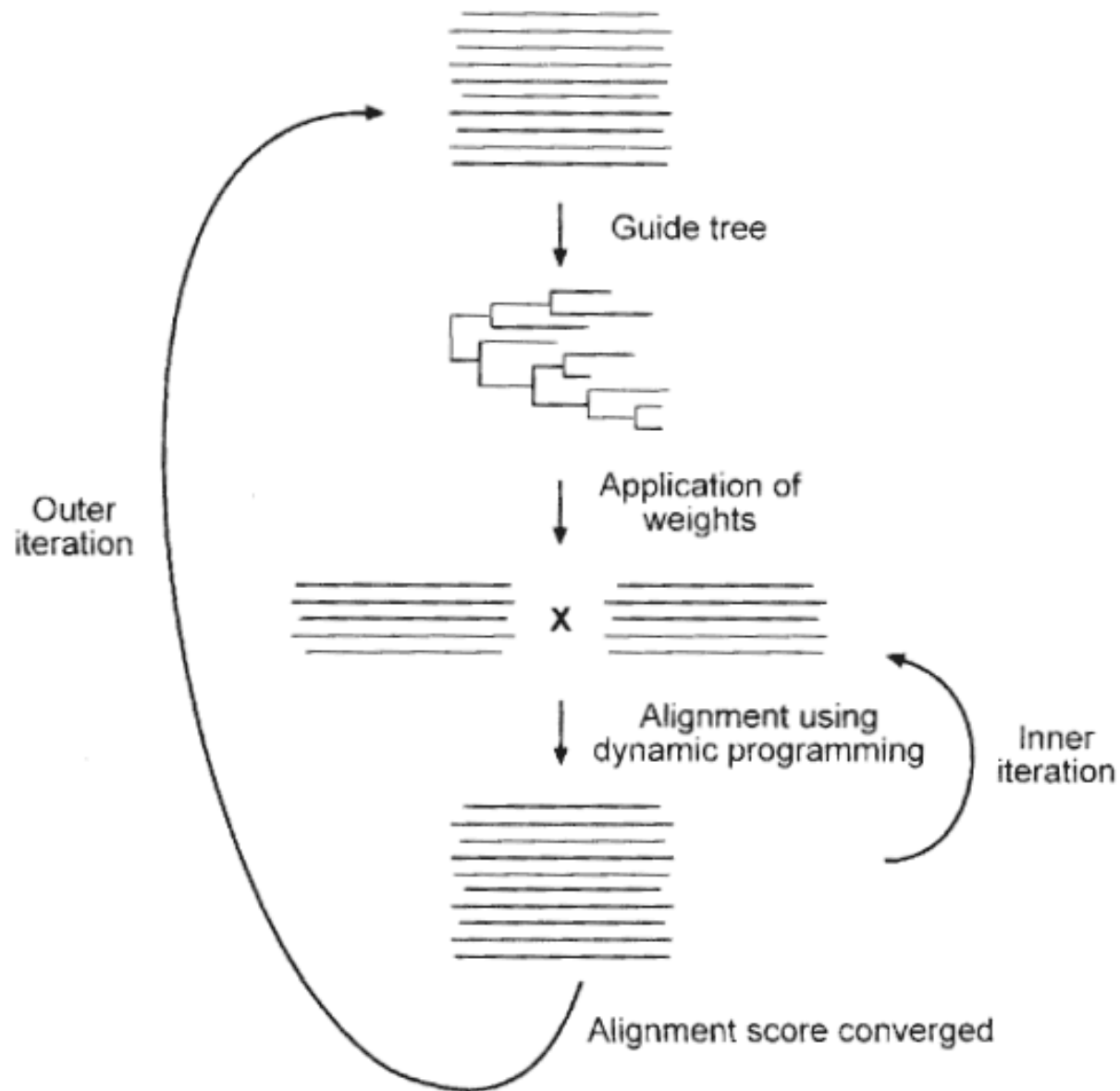
PRRN

(<http://www.genome.jp/tools/prrn/>)



The screenshot shows the web interface for 'Multiple Sequence Alignment by PRRN'. At the top, there is a logo with a stylized 'P' and 'R' and the title 'Multiple Sequence Alignment by PRRN'. Below the title, there are three tabs: 'CLUSTALW', 'MAFFT', and 'PRRN', with 'PRRN' being the active tab. A 'Help' link is located on the right. The main section is titled 'Specify Parameters :'. It contains several input fields and dropdown menus: 'Sequence Type' (radio buttons for PROTEIN and DNA, with PROTEIN selected), 'Score Matrix' (a dropdown menu showing PAM250), 'Gap Penalty (PROTEIN)' (Open: 9.0, Extend: 2.0, Background: 1.0), 'Gap Penalty (DNA)' (Open: 5.0, Extend: 2.0, Background: 2.0), 'Output Format' (a dropdown menu showing Native), 'Method' (a dropdown menu showing Progressive (pairwise alignment) + Iterative refinement), 'Window Size' (100), 'Score Matrix (Conserved Regions)' (PAM100), and 'Threshold Value (Conserved Regions)' (25.0). Below these parameters, there is a section 'Enter your sequences (FASTA format) :'. It includes 'Restrictions of the sequence size' with three bullet points: 1. [The number of sequences] <= 200, 2. [The maximum length] <= 2000, and 3. [The number of sequences] x [The maximum length] <= 50000. There are 'File Upload' and 'Copy & Paste' options. The 'File Upload' option has a text input field and a 'Przełóż' button. The 'Copy & Paste' option has a large text area. At the bottom, there are 'submit' and 'reset' buttons. The footer of the interface includes 'KEGG', 'GenomeNet', and 'Kyoto University Bioinformatics Center'.

MSA – metoda iteracyjna



Zastosowanie:

sekwencje wykazujące jedynie miejscowe podobieństwo

Metoda:

dopasowanie lokalne – identyfikacja wspólnego dla wszystkich sekwencji bloku niezawierającego przerw.

Przykład narzędzia:

DIALIGN

<http://www.genomatix.de/cgi-bin/dialign/dialign.pl>

Wynik dopasowania – prezentacja graficzna

```
AAB24882      TYHMCQFHCRCYVNNHSGEKLIECNERSKAFSCPSHLQCHKRRQIGEKTHEHNQCGKAFPT 60
AAB24881      -----YECNQCGKAFAQHSSLKCHYRTHIGEKPYECNQCGKAFSK 40
                ****: .***: * *:*** * :****.:* *****..

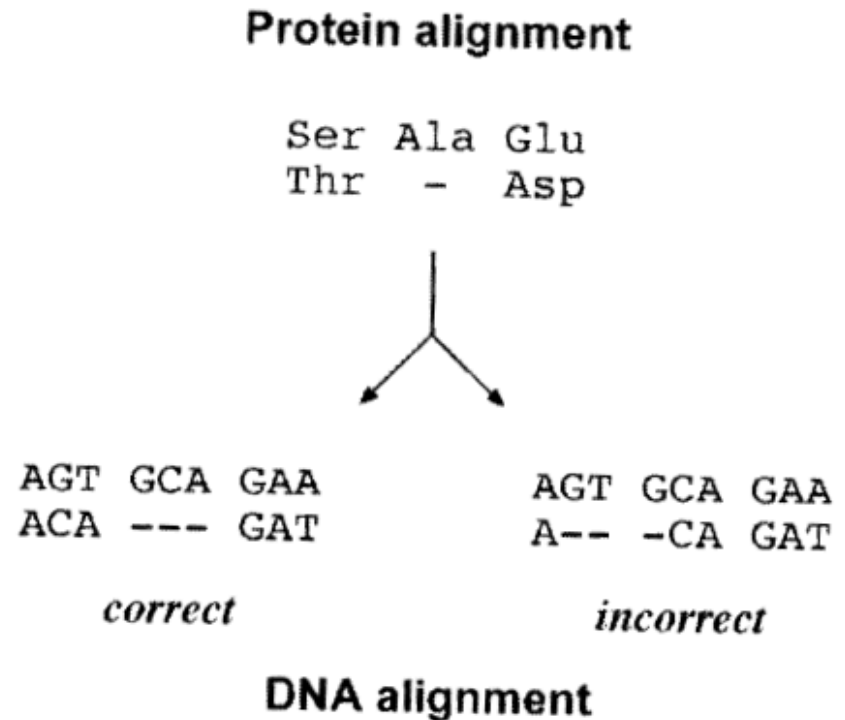
AAB24882      PSHLQYHERTHTGKPYECHQCGQAFKKCSLLQRHKRTHTGKPYE-CNQCGKAFAQ- 116
AAB24881      HSHLQCHKRTHTGKPYECNQCGKAFSQHGLLQRHKRTHTGKPYMNVINMVKPLHNS 98
                **** *:*****:***:**.: .*****: *: : :
```

- * (gwiazdka) – zgodność
- : (dwukropek) – substytucje konserwatywne
- . (kropka) – podstawienie semikonserwatywne

Dodatkowe oznaczenia kolorami.

Dopasowanie na poziomie białek
charakteryzuje się większą czułością.

Na poziomie DNA może wystąpić
niedopasowanie kodonów.



Czasem może nastąpić konieczność ręcznej poprawy wygenerowanego automatycznie dopasowania MSA.

Może w tym pomóc np. BioEdit:

<http://www.mbio.ncsu.edu/bioedit/bioedit.html>

lub JalView:

<http://www.jalview.org/examples/applets.html>

Jalview comes in two distinct flavours.

The main application allows connection to multiple web services provided as well as a host of additional useful features such as printing, making images annotating your alignment. You need to download and install this software.

The applet version runs in web browsers and is a useful interactive display of features and annotations files. It does not have the full functionality of the main application, such as loading images, saving files, running web service jobs due to security restrictions in

For more information on how to use the applet in your website, see [full list c](#)

Pressing one of the buttons below will load up a cut down version of Jalview browser.

Ferredoxins, chloroplast precursor related UniRef
(15 sequences x 150 residues)

Start Jalview

User Defined Colours
an associated Newick
tree file

Start Jalview

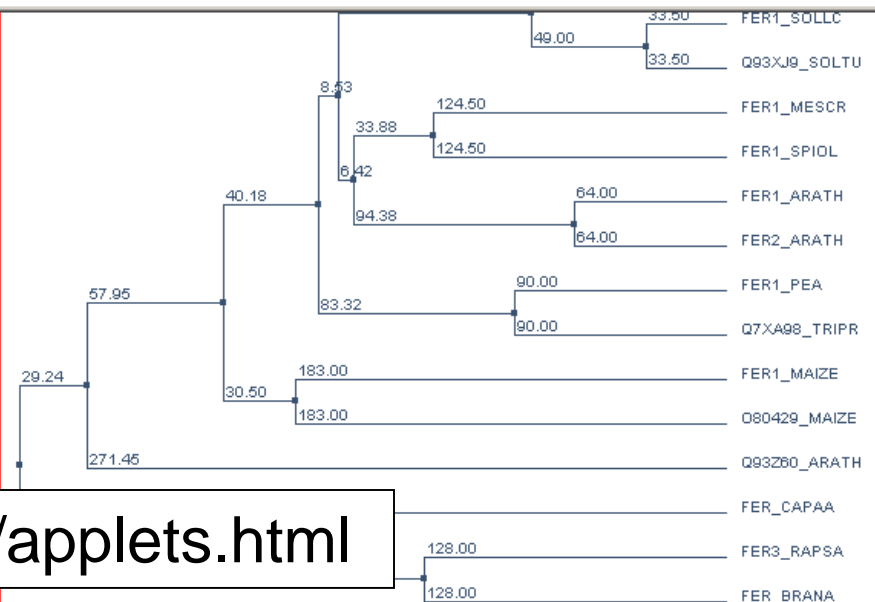
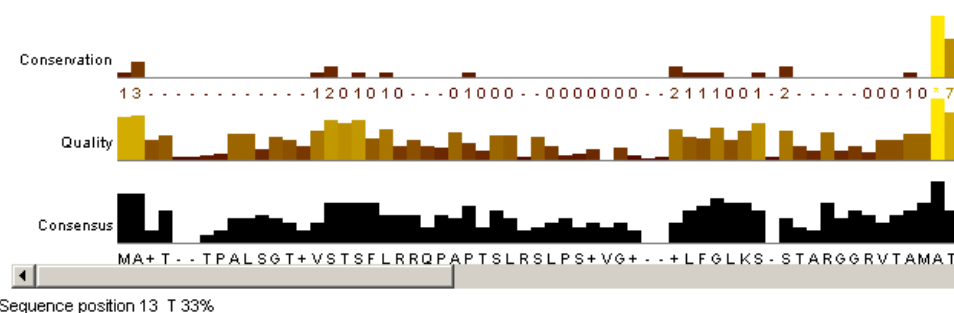
Displays a features file
the alignment

Start Jalview

Associates PDB file 1
with sequence FER1

Start Jalview

Displays a Multiple Sequence Alignment & JNet Prediction for a Sequence

[illegible]

<http://www.jalview.org/examples/applets.html>

Motywy, wzorce, profile

Motyw (*motif*) – często powtarzający się wzorzec sekwencji mający (potwierdzone bądź hipotetyczne) znaczenie biologiczne.

(W podobny sposób definiuje się motywy strukturalne (przestrzenne), przy czym budujące je aminokwasy nie muszą sąsiadować w sekwencji)

Wzorzec (*pattern*) – opisuje motyw za pomocą sekwencji konsensusowej (często wykorzystywane są wyrażenia regularne)

Profil – opisuje motyw ilościowo – wykorzystuje macierze PSSM.

Baza danych wzorców i profili:

<http://prosite.expasy.org/>

Wyszukiwanie wg motywów: konstrukcja wyrażenia regularnego

Alignment	Regular Expression or Pattern
ADLGAVFALCDRYFQ SDVGPRSCFCERFYQ ADLGRTQNRCDRYYQ ADIGQPHSLCERYFQ	[AS]-D-[IVL]-G-x(4)-{PG}-C-[DE]-R-[FY]2-Q

{X} – dowolny aminokwas za wyjątkiem X

[XY] – X albo Y

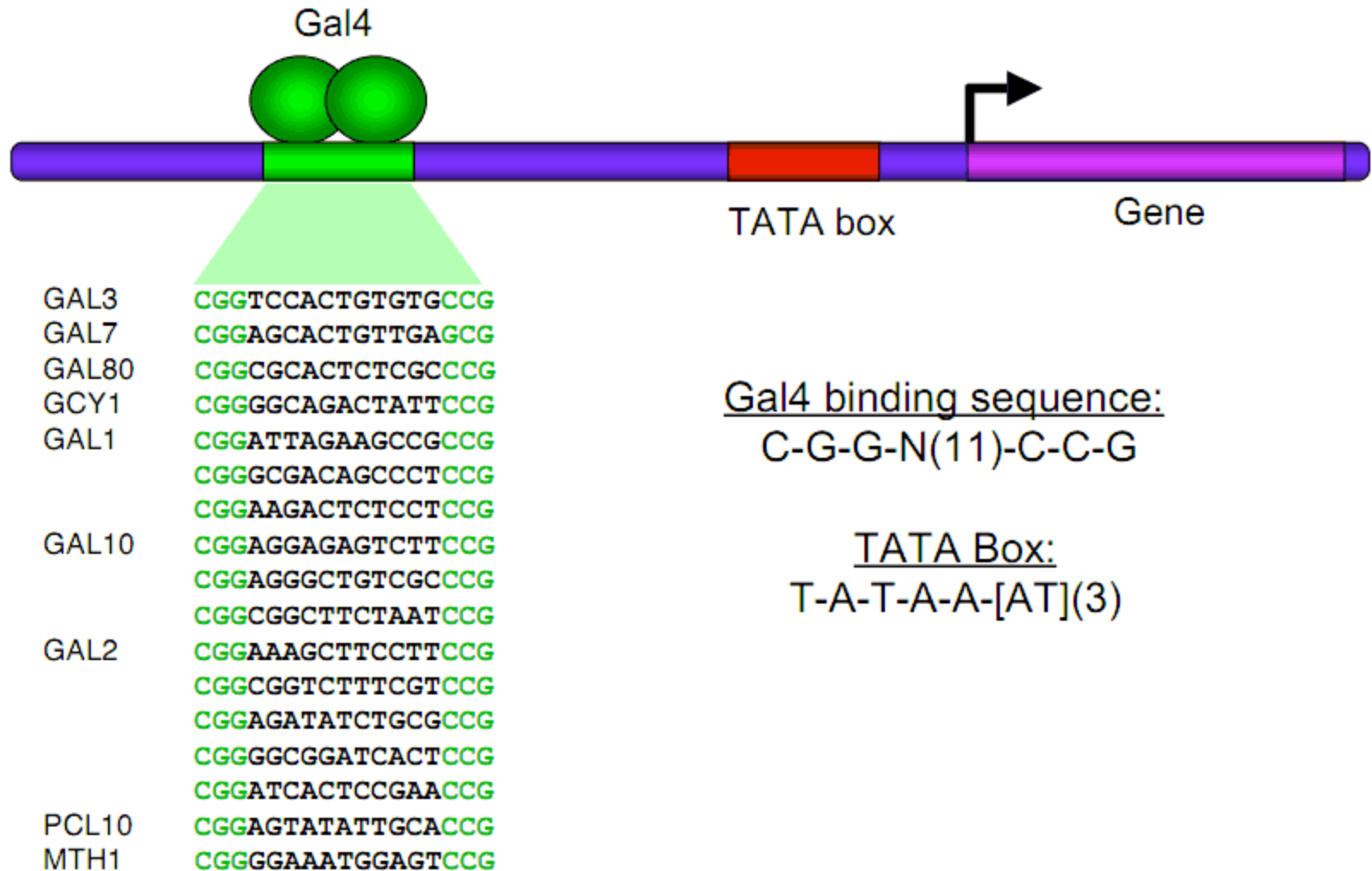
x – dowolny aminokwas

cyfra oznacza liczbę powtórzeń

Konserwatywne motywy w sekwencjach białkowych mają zwykle ważną biologiczną funkcję.

Mogą zatem wskazywać np. miejsce aktywne enzymu lub miejsce wiązania innego białka, jonu metalu itp.

Przykład motywu – miejsce wiązania czynników transkrypcyjnych



- Alignment of transcription factor binding sites

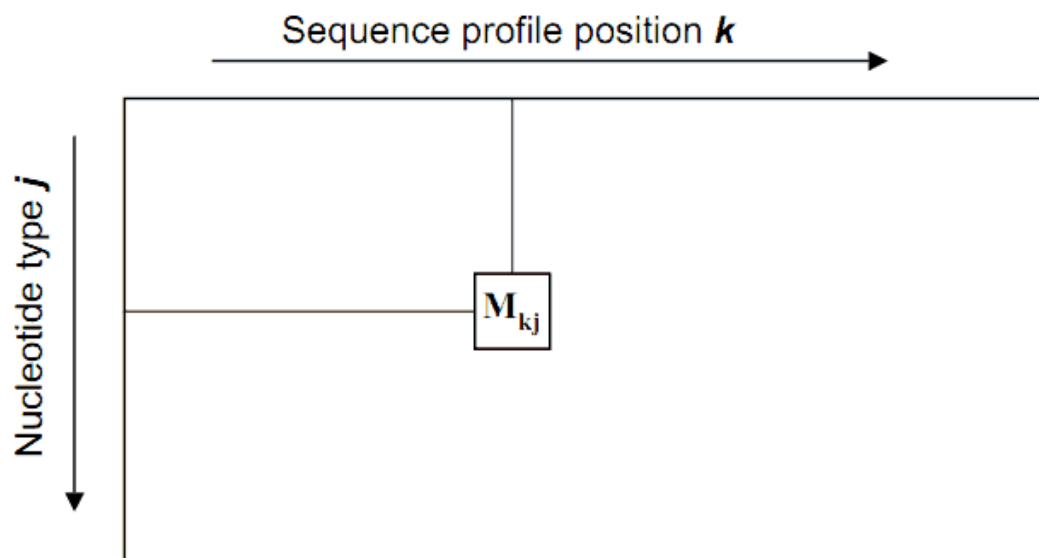
Motif

```
CCAAATTAGGAAA  
CCTATTAAGAAAA  
CCAAATTAGGAAA  
CCAAATTCGGATA  
CCCATTTGAAAA  
CCTATTTAGTATA  
CCAAATTAGGAAA  
TCTATTTTGGAAA  
CCAATTTTCAAAA  
CCAAATTGGCAAA
```

Consensus: YCHAWTWNSNAWA or CCHAWTTNGNAWA



Macierze ocen specyficzne względem pozycji.
Macierz tworzona jako macierz podobieństwa, ale podstawienia aminokwasów są zależne od pozycji względem wzorca).

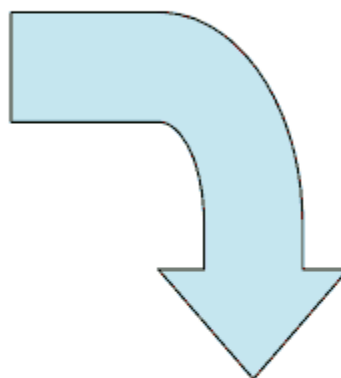


$$M_{kj} = \log \left(\frac{p_{kj}}{p_j} \right)$$

- p_{kj} = probability of nucleotide j at position k in the profile
- p_j = "background" probability of nucleotide j in genome sequence

Alignment of Transcription factor consensus binding sequence:

CCAAATTAGGAAA
CCTATTAAGAAAA
CCAAATTAGGAAA
CCAAATTCGGATA
CCCATTTCGAAAA
CCTATTTAGTATA
CCAAATTAGGAAA
CCAAATTGGCAAA
TCTATTTTGGAAA
CCAATTTTCAAAA



Alignment Matrix:

Position k =	1	2	3	4	5	6	7	8	9	10	11	12	13
A:	0	0	6	10	5	0	1	5	0	3	10	8	10
C:	9	10	1	0	0	0	0	2	1	1	0	0	0
G:	0	0	0	0	0	0	0	1	9	5	0	0	0
T:	1	0	3	0	5	10	9	2	0	1	0	2	0
Consensus:	C	C	[ACT]	A	[AT]	T	T	N	G	N	A	[AT]	A

Macierze PSSM – przykład 2

Position⇒	1	2	3	4	5
Sequence 1	C	C	G	T	L
Sequence 2	C	G	H	S	V
Sequence 3	G	C	G	S	L
Sequence 4	C	G	G	T	L
Sequence 5	C	C	G	S	S

Position⇒ 1 2 3 4 5

Prob(C)	0.8	0.6	-	-	-
Prob(G)	0.2	0.4	0.8	-	-
Prob(H)	-	-	0.2	-	-
Prob(S)	-	-	-	0.6	0.2
Prob(T)	-	-	-	0.4	-
Prob(L)	-	-	-	-	0.6
Prob(V)	-	-	-	-	0.2

...

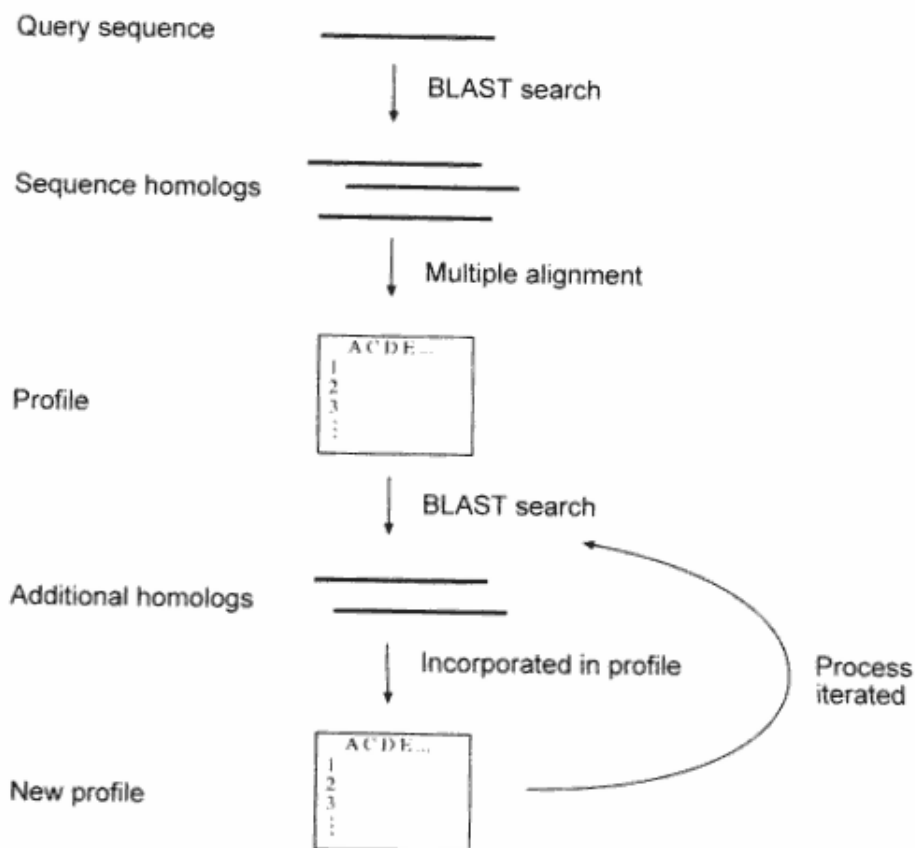
Macierze PSSM – przykład 2, oszacowanie prawdopodobieństwa

	C	G	G	S	V
Position⇒	1	2	3	4	5
Prob(C)	0.8	0.6	-	-	-
Prob(G)	0.2	0.4	0.8	-	-
Prob(H)	-	-	0.2	-	-
Prob(S)	-	-	-	0.6	0.2
Prob(T)	-	-	-	0.4	-
Prob(L)	-	-	-	-	0.6
Prob(V)	-	-	-	-	0.2

...

$$0.8 * 0.4 * 0.8 * 0.6 * 0.2 = 0.031$$

Specyficzny dla pozycji iterowany BLAST.
Pozwala identyfikować słabe, lecz istotne biologicznie podobieństwa.



Lokalne dopasowanie przekształcane jest w profil,
który jest wykorzystywany do wyszukiwania kolejnych sekwencji.

