Porównywanie sekwencji

Homologia, podobieństwo i analogia

dopasowanie sekwencji

- **Dopasowanie/porównywanie**
- **Uliniowienie**
- **Alignment**

W bioinformatyce, dopasowanie sekwencji jest sposobem dopasowania struktur pierwszorzędowych DNA, RNA, lub białek do zidentyfikowania regionów wykazujących podobieństwo, mogące być konsekwencją funkcjonalnych, strukturalnych, lub ewolucyjnych powiązań pomiędzy sekwencjami. Zestawione sekwencje nukleotydów lub aminokwasów są zazwyczaj przedstawione jako wiersze macierzy. Pomiędzy reszty wprowadzane są przerwy, tak że reszty zbliżonych do siebie sekwencji tworzą kolejne kolumny.

Jeśli dwie dopasowywane sekwencje mają wspólne pochodzenie, niedopasowania mogą być interpretowane jako mutacje punktowe, a przerwy jako indele (mutacje polegające na delecji lub insercji), które zaszły w jednej lub obu liniach od czasu, kiedy obie sekwencje uległy rozdzieleniu. W przypadku dopasowywania sekwencji białek, stopień podobieństwa pomiędzy aminokwasami zajmującymi konkretną pozycję, może stanowić zgrubną miarę tego, jak konserwatywny jest dany region lub motyw . Brak substytucji lub obecność jedynie konserwatywnych substytucji (tj. zamiany reszty na inną, ale o podobnych właściwościach chemicznych) w określonym regionie sekwencji sugeruje, że jest on ważny strukturalnie lub funkcjonalnie. Dopasowywanie sekwencji może być także stosowane dla sekwencji pochodzenia poza biologicznego, np. danych finansowych lub sekwencji występujących w językach naturalnych.

Masur i inni, Dopasowanie sekwencji, Wikipedia 11.2009

Skąd te różnice

różnice między sekwencjami świadczą o mutacjach, które zaszły po rozdzieleniu się sekwencji od wspólnego przodka

```
AGA--TTGATACCCA
|||  ||  |   | |
AGACATTAA---CTA
```

```
AGA--TTGATACCCA
Insercja  +TAC
Delecja   -CA
```

```
AGACATTAA---CTA
G->A C->T
substytucje
```

```
      TAC
      ↓
AGACATTGACCA
```


Etapy dopasowywania sekwencji

```
1 zestawienie (0 identycznych, 0% podobieństwa)

X = długość sekwencji (30)
MHSSIVLATVLFVAIASASKTRELCMKSLV
                          MHVSIVLATVLFVAIASAS
                           y = długość sekwencji (20)
```

```
2 zestawienie (0 identycznych, 0% podobieństwa)

MHSSIVLATVLFVAIASASKTRELCMKSLV
                          MHVSIVLATVLFVAIASAS
```

```
3 zestawienie (1 identyczna, 33% podobieństwa)

MHSSIVLATVLFVAIASASKTRELCMKSLV
                          MHVSIVLATVLFVAIASAS
```

4 zestawienie (0 identycznych, 0% podobieństwa)

```
MHSSIVLATVLFVAIASASKTRELCMKSLV
                        MHVSIVLATVLFVAIASAS
```

5 zestawienie (0 identycznych, 0% podobieństwa)

```
MHSSIVLATVLFVAIASASKTRELCMKSLV
                         MHVSIVLATVLFVAIASAS
```

6 zestawienie (2 identyczne, 33% podobieństwa)

```
MHSSIVLATVLFVAIASASKTRELCMKSLV
                       MHVSIVLATVLFVAIASAS
```

7 zestawienie (0 identycznych, 0% podobieństwa)

```
MHSSIVLATVLFVAIASASKTRELCMKSLV
                      MHVSIVLATVLFVAIASAS
```


X-2 zestawienie (3 identyczne, 15% podobieństwa)

```
MHSSIVLATVLFVAIASASKTRELCMKSLV
  MHVSIVLATVLFVAIASAS
```

x zestawienie (19 identycznych, 95% podobieństwa)

```
MHSSIVLATVLFVAIASASKTRELCMKSLV
MHVSIVLATVLFVAIASAS
```

X+1 zestawienie (1 identyczna, 5,26% podobieństwa)

```
 MHSSIVLATVLFVAIASASKTRELCMKSLV
MHVSIVLATVLFVAIASAS
```

X+2 zestawienie (3 identyczne, 16,67% podobieństwa)

```
  MHSSIVLATVLFVAIASASKTRELCMKSLV
MHVSIVLATVLFVAIASAS
```


X+Y-4 zestawienie (1 identycznych, 25% podobieństwa)

```
                MHSSIVLATVLFVAIASASKTRELCMKSLV
MHVSIVLATVLFVAIASAS
```

X+Y-3 zestawienie (1 identycznych, 33,3% podobieństwa)

```
                 MHSSIVLATVLFVAIASASKTRELCMKSLV
MHVSIVLATVLFVAIASAS
```

X+Y-2 zestawienie (0 identyczne, 0% podobieństwa)

```
                  MHSSIVLATVLFVAIASASKTRELCMKSLV
MHVSIVLATVLFVAIASAS
```

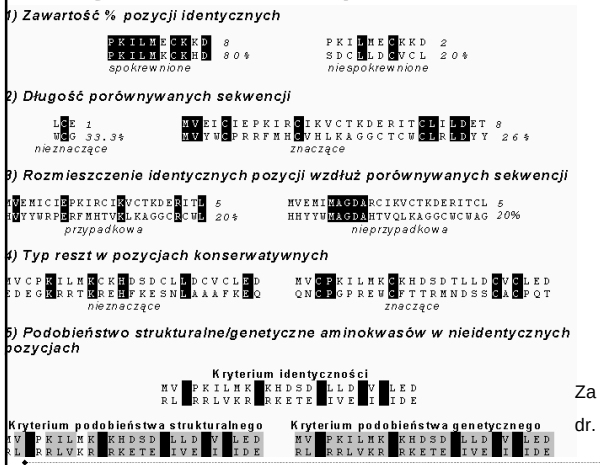
X+Y-1 zestawienie (0 identycznych, 0% podobieństwa)

```
                   MHSSIVLATVLFVAIASASKTRELCMKSLV
MHVSIVLATVLFVAIASAS
```


Etapy dopasowywania sekwencji

Za zgodą
dr. Jacka Leluka

Kryteria szacowania podobieństwa sekwencji



Za zgodą

dr. Jacka Leluka

Kryteria szacowania podobieństwa sekwencji

- Procent identyczności (względny udział odpowiadających sobie pozycji obsadzonych tymi samymi resztami)
- Długość porównywanych sekwencji (liczba porównywanych pozycji)
- Rozmieszczenie identycznych pozycji wzdłuż porównywanych sekwencji
- Typ reszt okupujących pozycje konserwatywne (sekwencje białkowe)
- Relacje genetyczne/strukturalne między resztami znajdującymi się w odpowiadających sobie nieidentycznych pozycjach (sekwencje białkowe)

Local vs. Global

Global alignment – znajduje najlepsz
dopasowanie dla **CAŁYCH**
dwóch sekwencji
(Needleman-Wunsch algorithm)

```
ADLGAVFALCDRYFQ
||||     |||| |
ADLGRTQN-CDRYYQ
```

Global alignment: forces alignment in regions which differ

Local alignment – poszukuje podobi
regionów we **FRAGMENTACH**
sekwencji
(Smith-Waterman algorithm)

```
ADLG    CDRYFQ
||||    |||| |
ADLG    CDRYYQ
```

Local alignment will return only **regions** of good alignment

Pairwise alignment

```
AAGCTGAATTCGAA
AGGCTCATTTCTGA
```

Tylko jeden możliwy alignment

```
AAGCTGAATT-C-GAA
AGGCT-CATTTCTGA-
```

This alignment includes:
2 mismatches
4 indels (gap)
10 perfect matches

Kilka możliwych rozwiązań:

```
AAGCTGAATTCGAA
AGGCTCATTTCTGA
```

```
A-AGCTGAATTC--GAA
AG-GCTCA-TTTCTGA-
```

```
AAGCTGAATT-C-GAA
AGGCT-CATTTCTGA-
```

scoring system:

- Perfect match: +1
- Mismatch: -2
- Indel (gap): -1 *(kara za przerwy)*

```
AAGCTGAATT-C-GAA
AGGCT-CATTTCTGA-
```

```
A-AGCTGAATTC--GAA
AG-GCTCA-TTTCTGA-
```

Score: = (+1)x10 + (-2)x2 + (-1)x4 = 2

Score: = (+1)x9 + (-2)x2 + (-1)x6 = -1

Zadanie 1

- Jaki jest **score** tego alignmentu??

dopasowanie:	+1
niedopasowanie:	-1
przerwa:	-2

```
---bardzo---lubiebioinformatyke
   ||||||   ||||*|||||||||||||*
niebardzonielubiębioinformatyki
```


Kara za przerwy (gap costs)

Kara za otwarcie przerwy – G
Kara za przedłużenie przerwy – L

Kara = G + Ln

gdzie:
n – długość przerwy

Standardowo:
G = 10 - 15
L = 1 - 2

Zadanie 2

Kara za otwarcie przerwy – G
Kara za przedłużenie przerwy – L

Kara = G + Ln
gdzie:
n – długość przerwy

```
-GAGCTGAA-----GAA
AGAGCTCAATTTCTGA-
```

Standardowo dla aa:
G = 10 - 15
L = 1 - 2

G = 10
L = 1

Kara = (10 + 5*1),
czy
Kara = (10 + 1*1) + (10 + 5*1) + (10 + 1*1)

Zadanie 3

Wiemy, że w toku ewolucji z danej sekwencji wyskoczyła jedna cała stosunkowo duża domena. Jakie wartości G i L dla kary za przerwy należy ustawić?

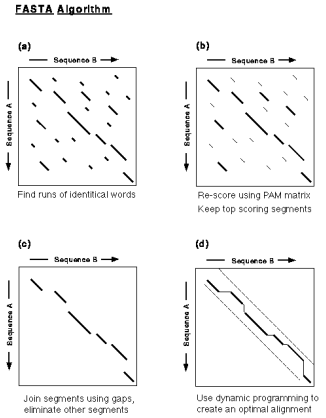
```
nielubiebardzo------bioinformatyki
   ||||||||||       ||||||||||||*
---lubiebardzobardzobioinformatyke
```


Metody dopasowania

dopasowanie par sekwencji *(pairwise alignment)*

1. **Metody słów (k - tuple methods)** - szybkie metody stosowane przy przeszukiwaniu baz danych sekwencji z wykorzystaniem programów FASTA i BLAST
2. **Macierz punktowe** - dot matrix, dotplot
3. **Programowanie dynamiczne (DP)**

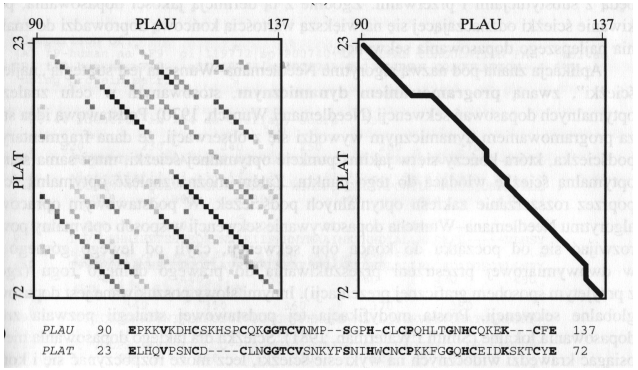
1. „słowa” - FASTA



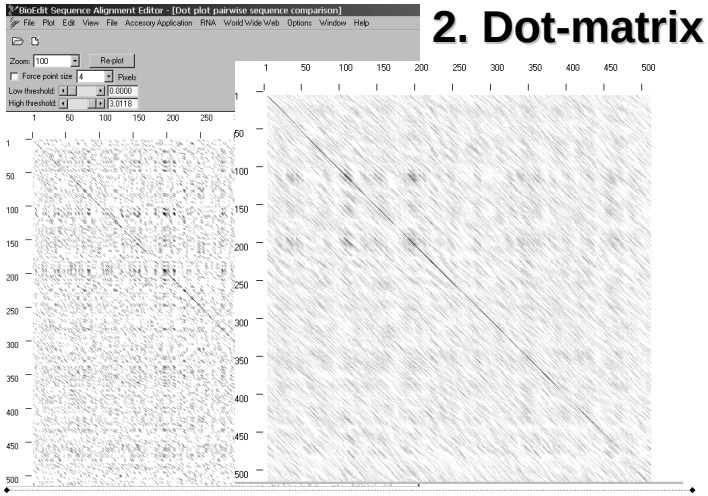
1. „słowa” - BLAST vs. FASTA

BLAST	FASTA
może podawać więcej niż jeden region o wysokiej punktacji	podaje tylko jedno najlepsze dopasowanie
lepszy dla sekwencji białek niż DNA	lepszy dla sekwencji DNA niż białek
szybszy niż FASTA	wolniejszy niż BLAST
mniej czuły niż FASTA przy użyciu domyślnych ustawień	bardziej czuły niż BLAST
daje gorsze rozróżnienie między prawdziwymi i fałszywymi homologami	daje lepsze rozróżnienie między prawdziwymi i fałszywymi homologami

2. Macierze punktowe

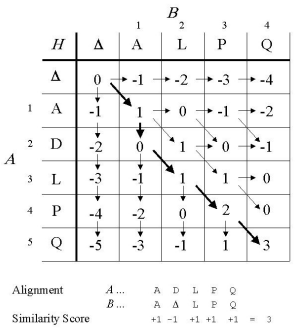


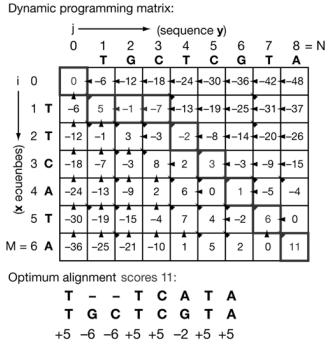
2. Dot-matrix



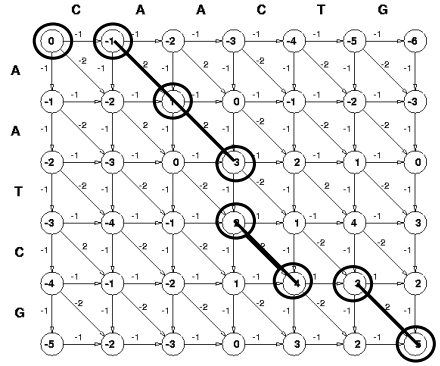
3. Programowanie dynamiczne

opiera się na podziale rozwiązywanego problemu na podproblemy względem kilku parametrów.





3. Programowanie dynamiczne



Scoring matrix

- Reprezentuje system punktowania jako tabela lub macierz $n \times n$ (n jest liczbą liter, które zawiera alfabet. n=4 dla DNA, n=20 dla białek)
- Macierz punktowania jest symetryczna

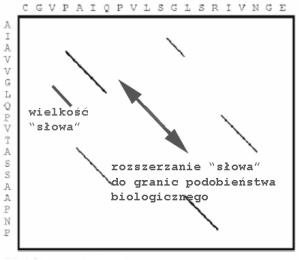
	A	G	C	T
A	**2**			
G	**-6**	**2**		
C	**-6**	**-6**	**2**	
T	**-6**	**-6**	**-6**	**2**

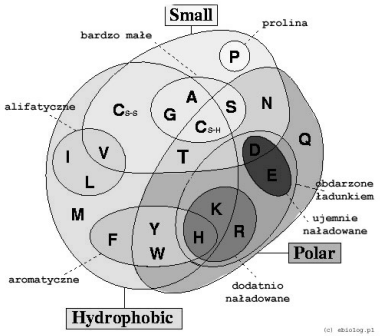
Mismatch

Match

Podobieństwa biochemiczne i biofizyczne aminokwasów

Diagram Venn-a





Macierze substytucji (podstawień)

- Jak za pomocą liczby określić podobieństwa biochemiczne i biofizyczne poszczególnych aminokwasów tak, aby liczba ta wyrażała jednocześnie realny wpływ na całe białko podstawienia danego aminokwasu w łańcuchu polipeptydowym i była uniwersalna dla wszystkich sekwencji?

- Przede wszystkim należy bazować na danych empirycznych
- Należy stworzyć alignment bardzo wielu blisko spokrewnionych sekwencji – na tyle podobnych, aby bez wątpliwości można było jednoznacznie i precyzyjnie określić częst amino

W kolumnie 4 E i D występują z częstotliwością w 4/8

PAM Matrix – Point/Percent Accepted Mutations

n**PAM** (*n Percent Accepted Mutations*)

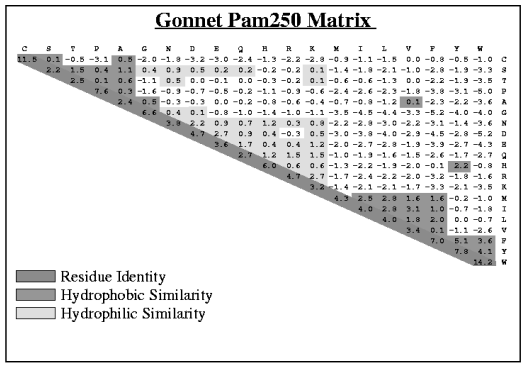
S_1, S_2 różnią się o jednostkę n PAM, jeśli S_2 można otrzymać z S_1 w ciągu akceptowalnych mutacji punktowych takich, że średnia liczba nieletalnych mutacji na 100 wynosi n. Najpopularniejsza jest 250PAM.

- Based on a database of 1,572 changes in 71 groups of closely related proteins (85% identity)
  - Alignment was easy

PAM Matrices

- Family of matrices PAM 80, PAM 120, PAM 250
- The number on the PAM matrix represents evolutionary distance
- Larger numbers are for larger distances

PAM



PAM - limitations

- Only one original dataset - PAM 1
- Examining proteins with few differences (85% identity)
- Bazuje głównie na małych białkach globularnych więc macierz jest nieco stronnicza

BLOSUM

- Henikoff i Henikoff (1992) stworzyli zestaw matryc bazujących na większej ilości danych empirycznych

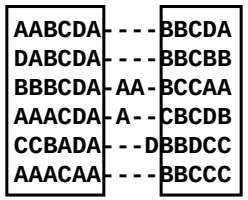
  BLOSSUMn (*Block Substitution Matrix* n)

  Oparta na bazie białek BLOCKS, gdzie są one podzielone na grupy tak, że dwa białka są zaliczane do jednej, jeśli można przejść od jednego do drugiego używając białek pośrednich tak, że dwa każde kolejne białka w tym przejściu maja skład identyczny w co najmniej $n\%$.
  Popularne są BLOSUM 50, BLOSUM 62.

- BLOSUM observes significantly more replacements than PAM, even for infrequent pairs

BLOSUM: Blocks Substitution Matrix

- Based on BLOCKS database
  - ~2000 blocks from 500 families of related proteins
  - Families of proteins with identical function



- Blocks are short conserved patterns of 3-60 aa long without gaps

BLOSUM

- Each block represent sequences alignment with different identity percentage
- For each block the amino-acid substitution rates were calculated to create BLOSUM matrix

BLOSUM Matrices

- BLOSUM*n* is based on sequences that shared at least *n* percent identity
- BLOSUM62 represents closer sequences than BLOSUM45

BLOSUM (62)

	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	

BLOSUM / PAM

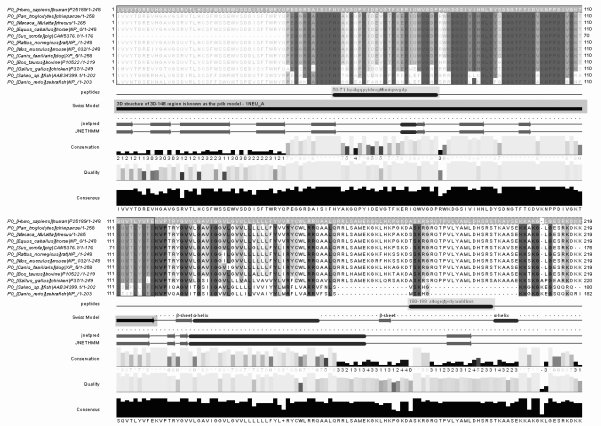
Wszystkie macierze na podstawie danych empirycznych	Tylko PAM1 na podstawie danych empirycznych, pozostałe macierze z interpolacji
Opracowywane na podstawie sekwencji o dalszym pokrewieństwie	Opracowane na podstawie bardzo blisko spokrewnionych sekwencji
Podobieństwo sekwencji rośnie wraz ze wzrostem indeksu	Podobieństwo sekwencji maleje wraz ze wzrostem indeksu
Bezpośrednie podobieństwo sekwencji tu i teraz	Poniekąd reprezentuje dystans ewolucyjny (model ewolucyjny akceptowanych mutacji punktowych)
Macierz symetryczna (im wyższa wartość tym łatwiejsza substytucja)	Macierz symetryczna (im wyższa wartość tym łatwiejsza substytucja)
Nie uwzględnia bezpośrednio ani właściwości fizykochemicznych aminokwasów, ani podobieństwa genetycznego (podobieństwa kodonów)	Nie uwzględnia bezpośrednio ani właściwości fizykochemicznych aminokwasów, ani podobieństwa genetycznego (podobieństwa kodonów)

PAM vs. BLOSUM

PAM100 ~ BLOSUM90
PAM120 ~ BLOSUM80
PAM160 ~ BLOSUM60
PAM200 ~ BLOSUM52
PAM250 ~ BLOSUM45

Sekwencje bardziej odległe

Multiple alignment



Pozycje konserwatywne

- Kolumny, gdzie wszystkie sekwencje zawierają takie same aminokwasy lub nukleotydy (lub w większości takie same – pozycje konserwatywne) są bardzo ważne (kluczowe) dla funkcji lub struktury.

```
VTIS CTGS SSNIGAG-NHVK WYQQ L P G
VTIS CTGS SSNIGS--ITVN WYQQ L P G
LRLS CTGS GFIFSS--YAMY WYQQ A P G
LSLT CTGS GTSFDD-QYYST WYQQ P P G
```