

# Porównywanie sekwencji

Homologia, podobieństwo i analogia

# dopasowanie sekwencji

- Dopasowanie/porównywanie
- Uliniowanie
- Alignment

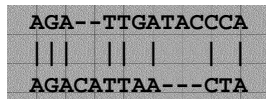
W bioinformatyce, dopasowanie sekwencji jest sposobem dopasowania struktur pierwszorzędowych DNA, RNA, lub białek do zidentyfikowania regionów wykazujących podobieństwo, mogące być konsekwencją funkcjonalnych, strukturalnych, lub ewolucyjnych powiązań pomiędzy sekwencjami. Zestawione sekwencje nukleotydów lub aminokwasów są zazwyczaj przedstawiane 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 indela (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

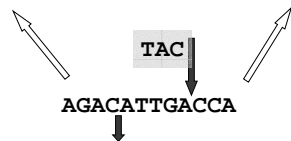
# Skąd te różnice

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



AGA--TTGATACCCA  
Insercja +TAC  
Delecja -CA

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



# Etapy dopasowywania sekwencji

1 zestawienie (0 identycznych, 0% podobieństwa)

```

X = długość sekwencji (30)
MHSSIVLATVLFVVAIASASKTRELCKMSLV
                                MHVSIVLATVLFVVAIASAS
                                y = długość sekwencji (20)
  
```

2 zestawienie (0 identycznych, 0% podobieństwa)

```

MHSSIVLATVLFVVAIASASKTRELCKMSLV
                                MHVSIVLATVLFVVAIASAS
  
```

3 zestawienie (1 identyczna, 33% podobieństwa)

```

MHSSIVLATVLFVVAIASASKTRELCKMSLV
                                MHVSIVLATVLFVVAIASAS
  
```

4 zestawienie (0 identycznych, 0% podobieństwa)

```
MHSSIVLATVLFVAIASASKTRELCKMSLV
                MHVSIVLATVLFVAIASAS
```

5 zestawienie (0 identycznych, 0% podobieństwa)

```
MHSSIVLATVLFVAIASASKTRELCKMSLV
                MHVSIVLATVLFVAIASAS
```

6 zestawienie (2 identyczne, 33% podobieństwa)

```
MHSSIVLATVLFVAIASASKTRELCKMSLV
                MHVSIVLATVLFVAIASAS
```

7 zestawienie (0 identycznych, 0% podobieństwa)

```
MHSSIVLATVLFVAIASASKTRELCKMSLV
                MHVSIVLATVLFVAIASAS
```

Wstęp do bioinformatyki Wykład 3 Biologia UWM Dr Jan Paweł Jastrzębski Slajd nr 5

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

```
MHSSIVLATVLFVAIASASKTRELCKMSLV
                MHVSIVLATVLFVAIASAS
```

x zestawienie (19 identycznych, 95% podobieństwa)

```
MHSSIVLATVLFVAIASASKTRELCKMSLV
                MHVSIVLATVLFVAIASAS
```

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

```
MHSSIVLATVLFVAIASASKTRELCKMSLV
                MHVSIVLATVLFVAIASAS
```

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

```
MHSSIVLATVLFVAIASASKTRELCKMSLV
                MHVSIVLATVLFVAIASAS
```

Wstęp do bioinformatyki Wykład 3 Biologia UWM Dr Jan Paweł Jastrzębski Slajd nr 6

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

```
MHSSIVLATVLFVAIASASKTRELCKMSLV
                MHVSIVLATVLFVAIASAS
```

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

```
MHSSIVLATVLFVAIASASKTRELCKMSLV
                MHVSIVLATVLFVAIASAS
```

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

```
MHSSIVLATVLFVAIASASKTRELCKMSLV
                MHVSIVLATVLFVAIASAS
```

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

```
MHSSIVLATVLFVAIASASKTRELCKMSLV
                MHVSIVLATVLFVAIASAS
```

Wstęp do bioinformatyki Wykład 3 Biologia UWM Dr Jan Paweł Jastrzębski Slajd nr 7

## Etapy dopasowywania sekwencji

<pre> 1 RVC PKILMECKKSDSDCLAEICILEHGYCG 2 RVC PKILMECKKSDSDCLAEICILEHGYCG 3 RVC PKILMECKKSDSDCLAEICILEHGYCG 4 RVC PKILMECKKSDSDCLAEICILEHGYCG 5 RVC PKILMECKKSDSDCLAEICILEHGYCG 6 RVC PKILMECKKSDSDCLAEICILEHGYCG 7 RVC PKILMECKKSDSDCLAEICILEHGYCG 8 RVC PKILMECKKSDSDCLAEICILEHGYCG 9 RVC PKILMECKKSDSDCLAEICILEHGYCG 10 RVC PKILMECKKSDSDCLAEICILEHGYCG 11 RVC PKILMECKKSDSDCLAEICILEHGYCG 12 RVC PKILMECKKSDSDCLAEICILEHGYCG 13 RVC PKILMECKKSDSDCLAEICILEHGYCG 14 RVC PKILMECKKSDSDCLAEICILEHGYCG 15 RVC PKILMECKKSDSDCLAEICILEHGYCG 16 RVC PKILMECKKSDSDCLAEICILEHGYCG 17 RVC PKILMECKKSDSDCLAEICILEHGYCG 18 RVC PKILMECKKSDSDCLAEICILEHGYCG 19 RVC PKILMECKKSDSDCLAEICILEHGYCG 20 RVC PKILMECKKSDSDCLAEICILEHGYCG 21 RVC PKILMECKKSDSDCLAEICILEHGYCG 22 RVC PKILMECKKSDSDCLAEICILEHGYCG 23 RVC PKILMECKKSDSDCLAEICILEHGYCG 24 RVC PKILMECKKSDSDCLAEICILEHGYCG 25 RVC PKILMECKKSDSDCLAEICILEHGYCG 26 RVC PKILMECKKSDSDCLAEICILEHGYCG 27 RVC PKILMECKKSDSDCLAEICILEHGYCG 28 RVC PKILMECKKSDSDCLAEICILEHGYCG 29 RVC PKILMECKKSDSDCLAEICILEHGYCG 30 RVC PKILMECKKSDSDCLAEICILEHGYCG 31 RVC PKILMECKKSDSDCLAEICILEHGYCG 32 RVC PKILMECKKSDSDCLAEICILEHGYCG 33 RVC PKILMECKKSDSDCLAEICILEHGYCG 34 RVC PKILMECKKSDSDCLAEICILEHGYCG 35 RVC PKILMECKKSDSDCLAEICILEHGYCG 36 RVC PKILMECKKSDSDCLAEICILEHGYCG 37 RVC PKILMECKKSDSDCLAEICILEHGYCG 38 RVC PKILMECKKSDSDCLAEICILEHGYCG 39 RVC PKILMECKKSDSDCLAEICILEHGYCG 40 RVC PKILMECKKSDSDCLAEICILEHGYCG 41 RVC PKILMECKKSDSDCLAEICILEHGYCG 42 RVC PKILMECKKSDSDCLAEICILEHGYCG 43 RVC PKILMECKKSDSDCLAEICILEHGYCG 44 RVC PKILMECKKSDSDCLAEICILEHGYCG 45 RVC PKILMECKKSDSDCLAEICILEHGYCG 46 RVC PKILMECKKSDSDCLAEICILEHGYCG 47 RVC PKILMECKKSDSDCLAEICILEHGYCG 48 RVC PKILMECKKSDSDCLAEICILEHGYCG 49 RVC PKILMECKKSDSDCLAEICILEHGYCG 50 RVC PKILMECKKSDSDCLAEICILEHGYCG 51 RVC PKILMECKKSDSDCLAEICILEHGYCG 52 RVC PKILMECKKSDSDCLAEICILEHGYCG 53 RVC PKILMECKKSDSDCLAEICILEHGYCG 54 RVC PKILMECKKSDSDCLAEICILEHGYCG 55 RVC PKILMECKKSDSDCLAEICILEHGYCG 56 RVC PKILMECKKSDSDCLAEICILEHGYCG 57 RVC PKILMECKKSDSDCLAEICILEHGYCG 58 RVC PKILMECKKSDSDCLAEICILEHGYCG 59 RVC PKILMECKKSDSDCLAEICILEHGYCG 60 RVC PKILMECKKSDSDCLAEICILEHGYCG 61 RVC PKILMECKKSDSDCLAEICILEHGYCG 62 RVC PKILMECKKSDSDCLAEICILEHGYCG 63 RVC PKILMECKKSDSDCLAEICILEHGYCG 64 RVC PKILMECKKSDSDCLAEICILEHGYCG 65 RVC PKILMECKKSDSDCLAEICILEHGYCG 66 RVC PKILMECKKSDSDCLAEICILEHGYCG 67 RVC PKILMECKKSDSDCLAEICILEHGYCG 68 RVC PKILMECKKSDSDCLAEICILEHGYCG 69 RVC PKILMECKKSDSDCLAEICILEHGYCG 70 RVC PKILMECKKSDSDCLAEICILEHGYCG 71 RVC PKILMECKKSDSDCLAEICILEHGYCG 72 RVC PKILMECKKSDSDCLAEICILEHGYCG 73 RVC PKILMECKKSDSDCLAEICILEHGYCG 74 RVC PKILMECKKSDSDCLAEICILEHGYCG 75 RVC PKILMECKKSDSDCLAEICILEHGYCG 76 RVC PKILMECKKSDSDCLAEICILEHGYCG 77 RVC PKILMECKKSDSDCLAEICILEHGYCG 78 RVC PKILMECKKSDSDCLAEICILEHGYCG 79 RVC PKILMECKKSDSDCLAEICILEHGYCG 80 RVC PKILMECKKSDSDCLAEICILEHGYCG 81 RVC PKILMECKKSDSDCLAEICILEHGYCG 82 RVC PKILMECKKSDSDCLAEICILEHGYCG 83 RVC PKILMECKKSDSDCLAEICILEHGYCG 84 RVC PKILMECKKSDSDCLAEICILEHGYCG 85 RVC PKILMECKKSDSDCLAEICILEHGYCG 86 RVC PKILMECKKSDSDCLAEICILEHGYCG 87 RVC PKILMECKKSDSDCLAEICILEHGYCG 88 RVC PKILMECKKSDSDCLAEICILEHGYCG 89 RVC PKILMECKKSDSDCLAEICILEHGYCG 90 RVC PKILMECKKSDSDCLAEICILEHGYCG 91 RVC PKILMECKKSDSDCLAEICILEHGYCG 92 RVC PKILMECKKSDSDCLAEICILEHGYCG 93 RVC PKILMECKKSDSDCLAEICILEHGYCG 94 RVC PKILMECKKSDSDCLAEICILEHGYCG 95 RVC PKILMECKKSDSDCLAEICILEHGYCG 96 RVC PKILMECKKSDSDCLAEICILEHGYCG 97 RVC PKILMECKKSDSDCLAEICILEHGYCG 98 RVC PKILMECKKSDSDCLAEICILEHGYCG 99 RVC PKILMECKKSDSDCLAEICILEHGYCG 100 RVC PKILMECKKSDSDCLAEICILEHGYCG </pre>	<p>0,0%</p> <p>0,0%</p> <p>0,0%</p> <p>0,0%</p> <p>25,0%</p> <p>0,0%</p> <p>1,36%</p> <p>62,1%</p> <p>5,17,2%</p> <p>2,6,9%</p> <p>1,33,3%</p> <p>0,0%</p> <p>0,0%</p> <p>0,0%</p> <p>7,3%</p>
--	--

Za zgodą  
dr. Jacka Leluka

Wstęp do bioinformatyki Wykład 3 Biologia UWM Dr Jan Paweł Jastrzębski Slajd nr 8

## Kryteria szacowania podobieństwa sekwencji

### 1) Zawartość % pozycji identycznych

PKILMCKKRD 9  
 PKILMCKKRD 20%  
 spokrewnione

PKIMECKKD 2  
 SDCILDVCL 20%  
 niespokrewnione

### 2) Długość porównywanych sekwencji

LDE 1  
 MGG 39.3%  
 nieznaczące

MVDIEPKIRCIKVKCTKDERITCLIDET 8  
 MGVUPRRFMHCVHLKAGGCTCWLRUDTY 26%  
 znaczące

### 3) Rozmieszczenie identycznych pozycji wzdłuż porównywanych sekwencji

MVMICIPKIRCIKVKCTKDEITCL 5  
 MYYVRFPRFMHTVCLKAGGCCLV 20%  
 przypadkowa

MVEIMAGDARCIKVKCTKDERITCL 5  
 HHYYMAGDHTVQLKAGGCWVAG 20%  
 nieprzypadkowa

### 4) Typ reszt w pozycjach konserwatywnych

MVEIPKILMCKKRDSDCLDVCVCLD  
 EDDEGRRTREDFKESNLAARFKEQ  
 nieznaczące

MVEIPKILMCKKRDSDTLLDVCVLED  
 QNPGPREWCFPTTMMNDSSAEPQT  
 znaczące

### 5) Podobieństwo strukturalne/genetyczne aminokwasów w nieidentycznych pozycjach

Kryterium identyczności  
 MVPKILMKRHDSDDLDDVLEDE  
 RLRLVLRKRRKETEIVFVIDE

Za zgodą  
dr. Jacka Leluka

Kryterium podobieństwa strukturalnego  
 MVPKILMKRHDSDDLDDVLEDE  
 RLRLVLRKRRKETEIVFVIDE

Kryterium podobieństwa genetycznego  
 MVPKILMKRHDSDDLDDVLEDE  
 RLRLVLRKRRKETEIVFVIDE

Wstęp do bioinformatyki Wykład 3  
 Biologia UWM  
 Dr Jan Paweł Jastrzębski Slajd nr 9

## 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 najlepsze dopasowanie dla **CAŁYCH** dwóch sekwencji

(Needleman-Wunsch algorithm)

ADLGAVFALCDRYFQ  
||| |||  
ADLGRTQN-CDRYQ

Global alignment:  
forces alignment in regions which differ

**Local alignment** – poszukuje podobieństw regionów we **FRAGMENTACH** sekwencji

(Smith-Waterman algorithm)

ADLG CDRYFQ  
||| |||  
ADLG CDRYQ

Local alignment will return only regions of good alignment

## Pairwise alignment

AAGCTGAATTCGAA  
AGGCTCATTCTGA

Tylko jeden możliwy alignment

AAGCTGAATT-C-GAA  
AGGCT-CATTCTGA-

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

## Kilka możliwych rozwiązań:

AAGCTGAATTCGAA  
AGGCTCATTCTGA

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

AAGCTGAATT-C-GAA  
AGGCT-CATTCTGA-



## scoring system:

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

AAGCTGAATT-C-GAA  
AGGCT-CATTCTGA-

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

$$\text{Score} = (+1) \times 10 + (-2) \times 2 + (-1) \times 4 = 2$$

$$\text{Score} = (+1) \times 9 + (-2) \times 2 + (-1) \times 6 = -1$$



## 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

$$\text{Kara} = G + Ln$$

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-
```

G = 10  
 L = 1

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

Kara =  $(10 + 1*1) + (10 + 5*1) + (10 + 1*1)$

Standardowo dla aa:

G = 10 - 15

L = 1 - 2

## 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?

```
nie lubie bardzo ----- bioinformatyki
||||| ||||||| ||||||| ||||||| |||*
--- lubie bardzo bardzo bioinformatyke
```

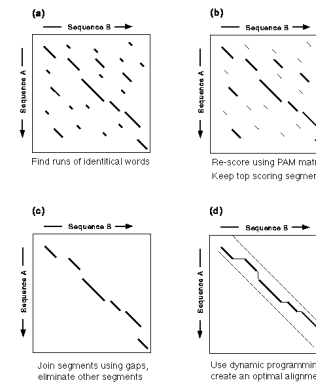
## Metody dopasowania

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

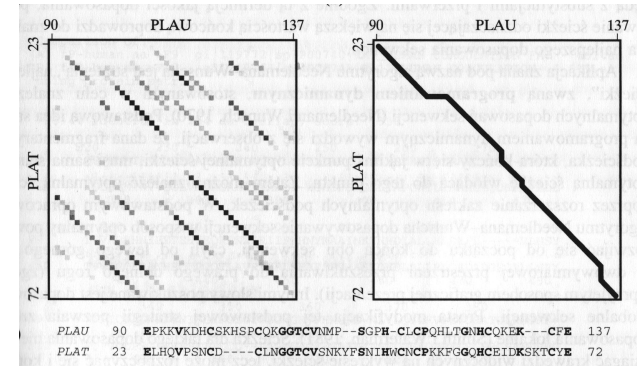
FASTA Algorithm



# 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

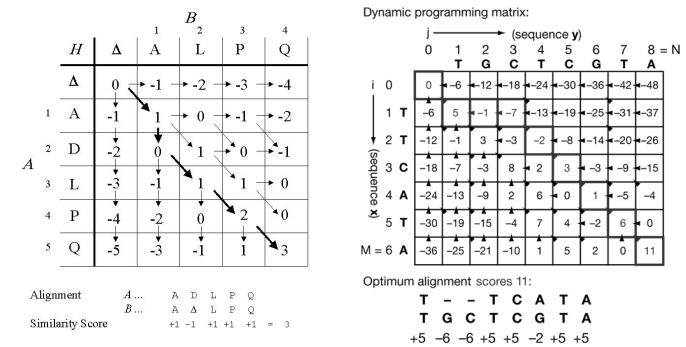


## 2. Dot-matrix

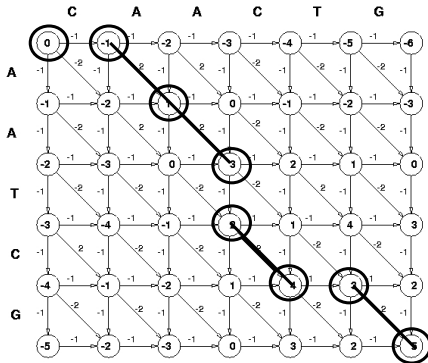
Wstęp do bioinformatyki Wykład 3 Biologia UWM Dr Jan Paweł Jastrzębski Slajd nr 23

# 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

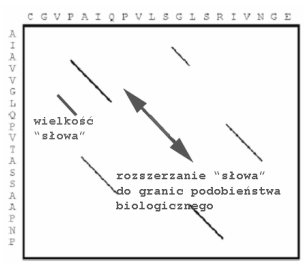
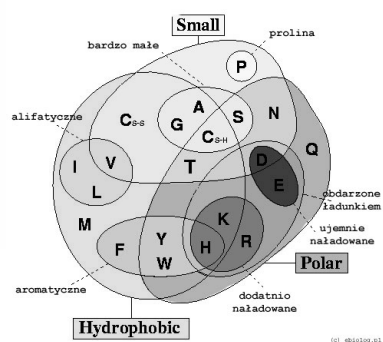


Diagram Venn-a



### 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 można było jednoznacznie i precyzyjnie określić częstotliwość występowania w kolumnie 4 E i D występują z częstotliwością w 4/8

M	G	Y	D	E
M	G	Y	D	E
M	G	Y	D	E
M	G	Y	D	E
M	G	Y	D	E
M	G	Y	D	E
M	A	Y	E	E
M	A	Y	E	E

## PAM Matrix – Point/Percent Accepted Mutations

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

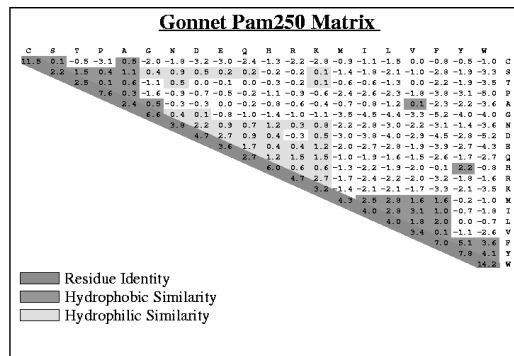
$S_1, S_2$  różnią się o jednostkę *n* PAM, jeśli  $S_2$  można otrzymać z  $S_1$  w ciągu akceptowalnych mutacji punktowych takich, że średnia liczba nieletalnych mutacji na 100 wynosi *n*. Najpopularniejsza jest 250PAM.

- 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

### BLOSSUM<sub>n</sub> (*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 mają 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

AABCDA	---	BBCDA
DABCD	---	BBCBB
BBBCDA	AA-	BCCAA
AAACDA	A--	CBDCB
CCBADA	---	DBBDCC
AAACAA	---	BBCCC

- 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<sub>n</sub> 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																			
S	-1	4																		
T	-1	1	5																	
P	-3	-1	-1	7																
A	0	1	0	-1	4															
G	-3	0	-2	-2	0	6														
N	-3	1	0	-2	-2	0	6													
D	-3	0	-1	-1	-2	-1	1	6												
E	-4	0	-1	-1	-1	-2	0	2	5											
Q	-3	0	-1	-1	-1	-2	0	0	2	5										
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8									
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5								
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5							
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5						
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	1	4						
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	2	2	4					
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	1	3	1	4				
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6		
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7	
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11

## 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)

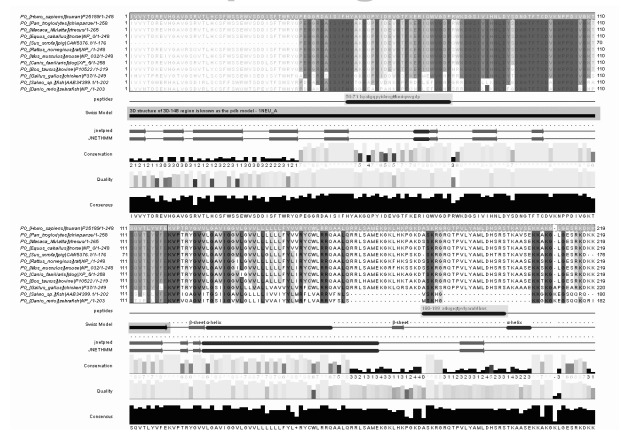
## PAM vs. BLOSUM



- 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.

```
VTISCTGSSNIGAG-NHVKWYQQLPG  
VTISCTGSSNIGS--ITVNWYQQLPG  
LRLSCTGSGFIFSS--YAMYWYQQAPG  
LSLTCTGSGTSFDD-QYYSTWYQQPPG
```