# Podstawy bioinformatyki dla biotechnologów

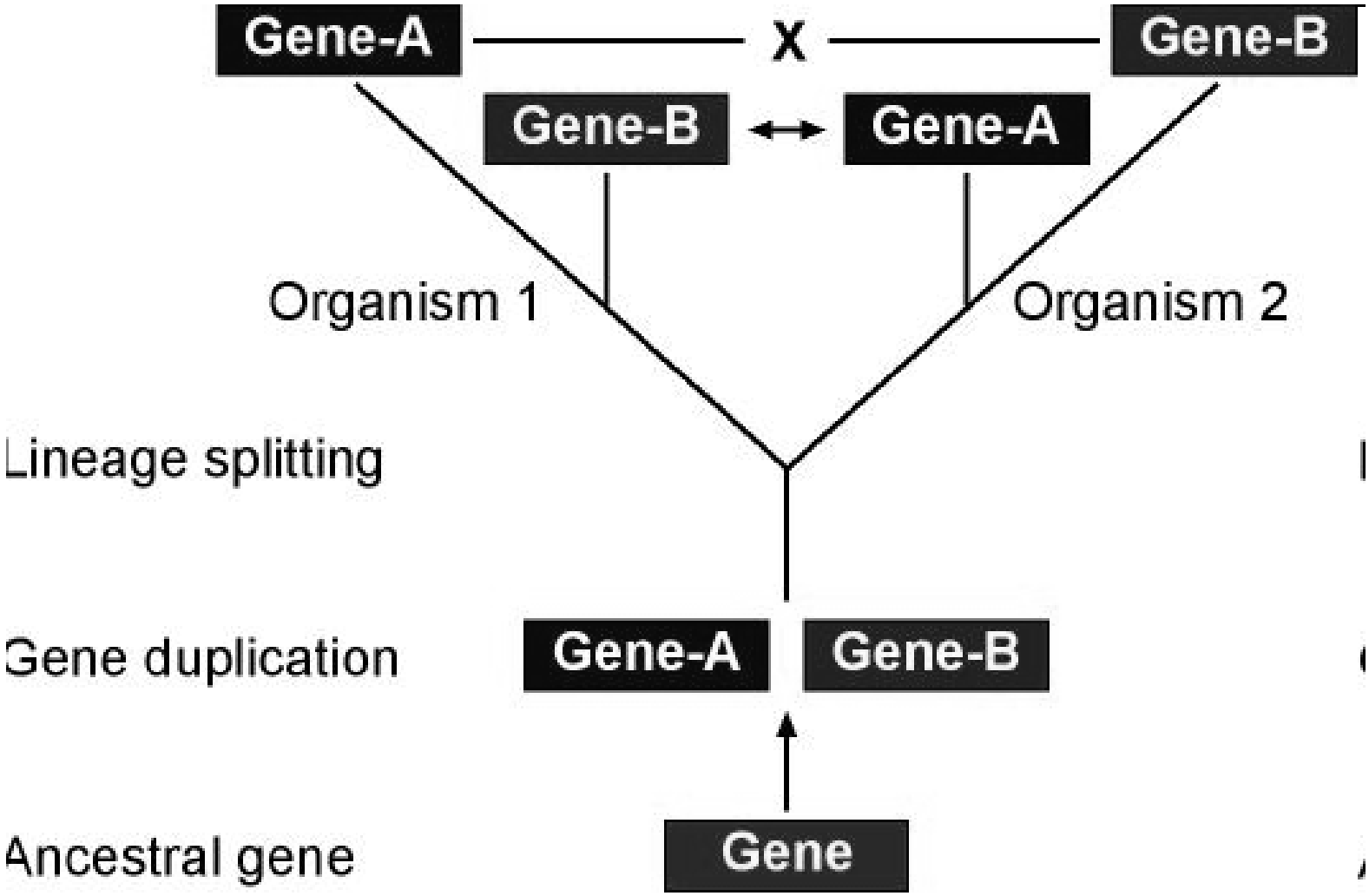
Wykład 3

alignment

# Porównywanie sekwencji

Homologia, podobieństwo i analogia

# Duplikacja, specjacja

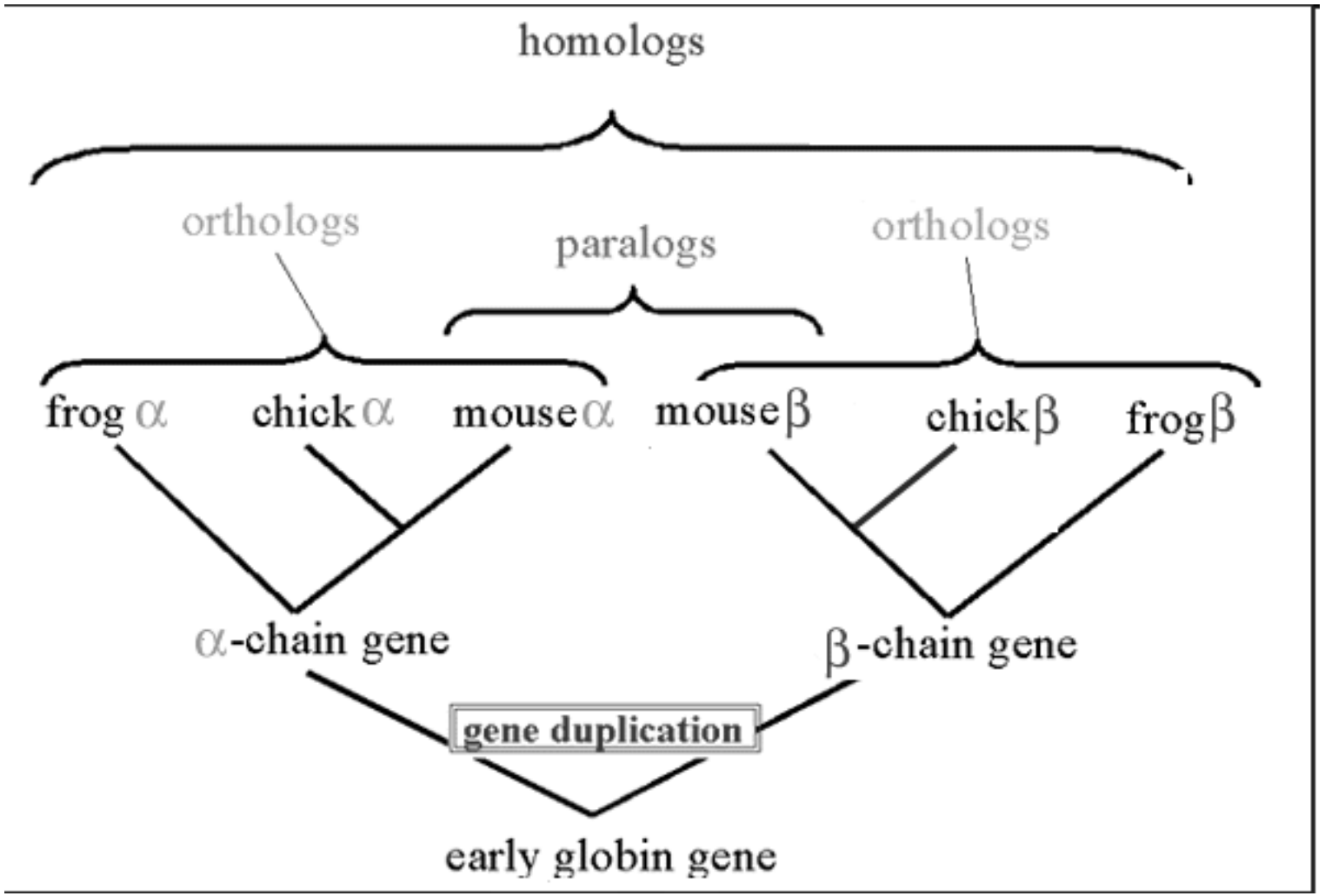


# Homologi

- **Ortologi** – homologiczne geny, których rozdzielenie nastąpiło na skutek specjacji, czyli rozdzielenia gatunków, lub rzadziej horyzontalnego transferu genu. Geny ortologiczne mają zwykle taką samą, albo zbliżoną funkcję.

- **Paralogi –** geny pochodzące od wspólnego przodka, rozdzielone w wyniku duplikacji genu. Paralogi mają często różne funkcje w organizmie. Przykładem mogą być mioglobina i hemoglobina u człowieka.

# Homo-, para-, orto-, analogi



# dopasowanie sekwencji

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

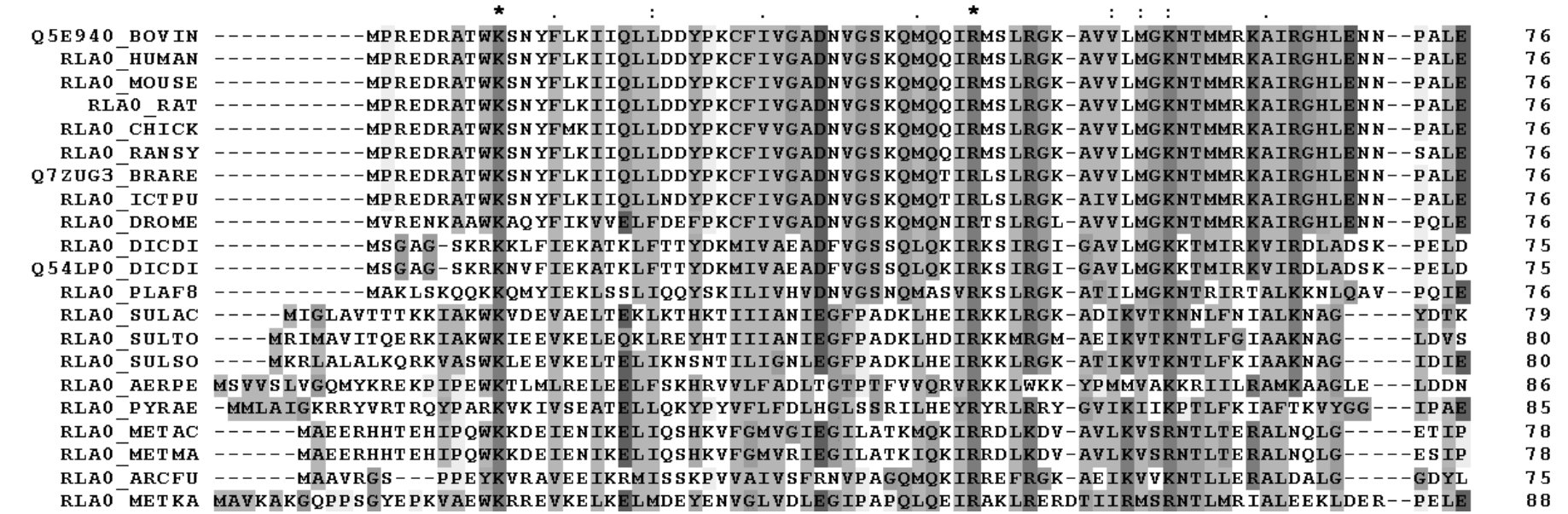
**W bioinformatyce, dopasowanie sekwencji jest sposobem dopasowania struktur pierwszorzędowych DNA, RNA, lub białek do zidentyfikowania regionów wykazujących podobieństwo, mogące być konsekwencją funkcjonalnych, strukturalnych, lub ewolucyjnych powiązań pomiędzy sekwencjami. Zestawione sekwencje nukleotydów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

# alignment

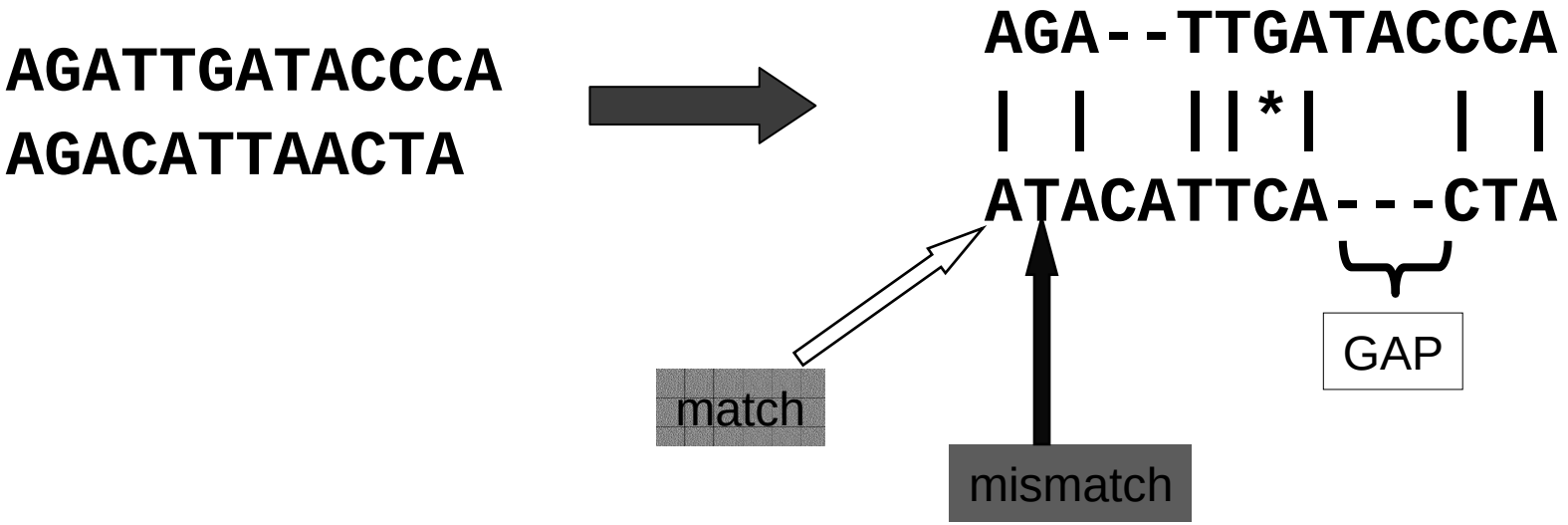
```
AAB24882        TYHMCQFHCRYVNNHSGEKLYECNERSKAFSCPSHLQCHKRRQIGEKTHEHNQCGKAFPT 60
AAB24881        --------------------YECNQCGKAFAQHSSLKCHYRTHIGEKPYECNQCGKAFSK 40
                                    ****: .***:  * *:** * :****.:* *******..

AAB24882        PSHLQYHERTHTGEKPYECHQCGQAFKKCSLLQRHKRTHTGEKPYE-CNQCGKAFAQ- 116
AAB24881        HSHLQCHKRTHTGEKPYECNQCGKAFSQHGLLQRHKRTHTGEKPYMNVINMVKPLHNS 98
                 **** *:***********:***:**.: .****************    :  *.: :
```



# alignment

Ułożenie dwóch lub więcej sekwencji biopolimerów (DNA, RNA lub białka) w celu zidentyfikowania regionów podobieństwa istotnego ze względów ewolucyjnych, strukturalnych lub funkcjonalnych (procedura oraz jej efekt).

- dwie sekwencje - pairwise alignment
- wiele sekwencji - multiple sequence alignment



# Znaczenie dopasowania

**Podobieństwo porównywanych sekwencji (*similarity*) może świadczyć o:**

- podobnej strukturze białek
- podobnej funkcji sekwencji
- wspólnej historii ewolucyjnej sekwencji

**Podobieństwo porównywanych sekwencji (*similarity*) może wynikać z:**

- **homologii** - pochodzeniu sekwencji (homologicznych) od wspólnego przodka; sekwencje mogą, ale nie muszą pełnić te same funkcje
- **konwergencji** - podobne motywy, które wyewoluowały w obu sekwencjach (analogicznych) niezależnie; np. chymotrypsyna i subtylizyna - różna struktura 3D, ale podobne centrum aktywne (histydyna, seryna, kwas asparaginowy)

{... Problem rozróżnienia odległej homologii od analogii }

# Skąd te różnice

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

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

**AGA--TTGATACCCA**

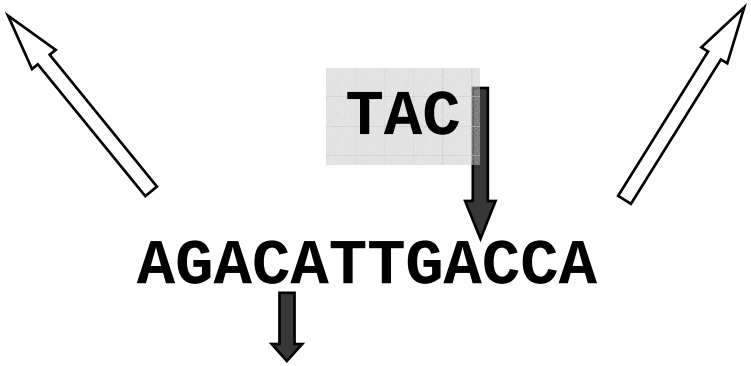
**Insercja +TAC**

**Delecja -CA**

**AGACATTAA---CTA**

**G->A C->T**

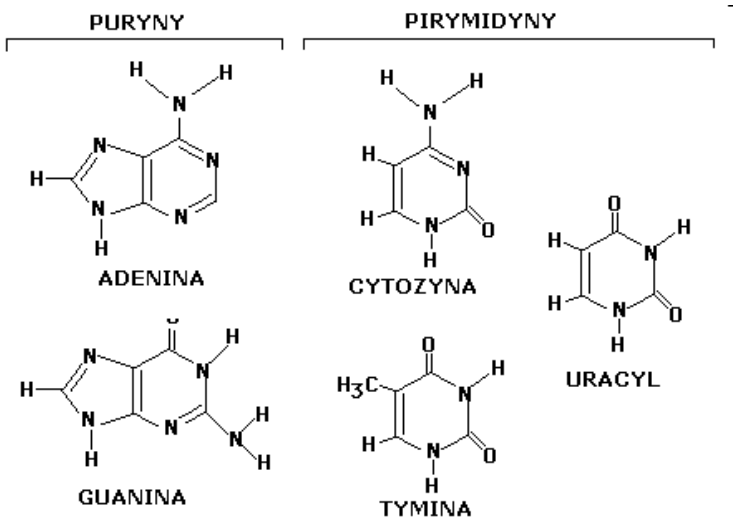
**substytucje**



# Substytucje nukleotydowe

- **Tranzycja** - okres przejściowy między systemem politycznym, który był, a tym który nastąpi. Proces ten jest krótszy i łatwiejszy od konsolidacji systemu politycznego. Tranzycja kończy się gdy pojawiają się ogólne ramy funkcjonowania nowego systemu. Przykładem są wszystkie państwa byłego bloku wschodniego, w tym Polska. *(Czy o to chodzi?)*

- **Transwersja** - mutacja genowa, punktowa zmiana chemiczna w obrębie nici DNA, w której zasada purynowa ulega zamianie na pirymidynową lub odwrotnie. Mutacja taka może nie spowodować żadnej zmiany lub zmianę kodu genetyczego (UUU -> UUA) albo też skróconą syntezę białka (UCG -> UCA).



# Zastosowanie alignmentu

- poszukiwaniu oraz określaniu funkcji i struktury (białek) dla „nowych” sekwencji (nieznanych nam do tej pory)

- określaniu powiązań filogenetycznych między sekwencjami - homologii między sekwencjami oraz w analizach ewolucyjnych

# Metody dopasowania

**dopasowanie par sekwencji *(pairwise alignment)***

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

- **dopasowanie wielu sekwencji *(multiple alignment)***

# 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

### 1) Zawartość % pozycji identycznych

```
PKILMECKKD  8
PKILMKCKHD  80%
spokrewnione

PKILMECKKD  2
SDCLLDCVCL  20%
niespokrewnione
```

### 2) Długość porównywanych sekwencji

```
LCE  1
WCG  33.3%
nieznaczące

MVEICIEPKIRCIKVCTKDERITCLILDET  8
MVYWCPRRFMHCVHLKAGGCTCWCLRLDYY  26%
znaczące
```

### 3) Rozmieszczenie identycznych pozycji wzdłuż porównywanych sekwencji

```
MVEMICIEPKIRCIKVCTKDERITL  5
HVYYWRPERFMHTVKLKAGGCRCWL  20%
przypadkowa

MVEMIMAGDARCIKVCTKDERITCL  5
HHYYWMAGDAHTVQLKAGGCWCWAG  20%
nieprzypadkowa
```

### 4) Typ reszt w pozycjach konserwatywnych

```
MVCPKILMKCKHDSDCLLDCVCLED
EDEGKRRTKREHFKESNLAAAFKEQ
nieznaczące

MVCPKILMKCKHDSDTLLDCVCLED
QNCPGPREWCFTTRMNDSSCACPQT
znaczące
```

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

**Kryterium identyczności**

```
MV PKILMK KHDSD LLD V LED
RL RRLVKR RKETE IVE I IDE
```

**Kryterium podobieństwa strukturalnego**

```
MV PKILMK KHDSD LLD V LED
RL RRLVKR RKETE IVE I IDE
```

**Kryterium podobieństwa genetycznego**

```
MV PKILMK KHDSD LLD V LED
RL RRLVKR RKETE IVE I IDE
```

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)

# Procedura oszacowania stopnia podobieństwa porównywanych sekwencji

Bardzo często oszacowanie stopnia podobieństwa porównywanych sekwencji sprowadzane jest jedynie do określenia względnego udziału pozycji identycznych. Pozostałe kryteria analizy zazwyczaj nie są w ogóle brane pod uwagę (np. bezwzględna długość sekwencji, dystrybucja identycznych pozycji wzdłuż łańcucha). Podejście takie jest niekompletne i stwarza ryzyko błędnej interpretacji otrzymanych wyników.

Przedstawiona niżej metoda oparta jest na prawdopodobieństwie przypadkowego pojawienia się zadeklarowanego stopnia identyczności. Uwzględnia ona podstawowe parametry mające znaczenie dla opisu faktycznego związku między porównywanymi sekwencjami.

**Liczbę wszystkich możliwych stopni identyczności** dla danych dwóch sekwencji opisuje poniższe r

$$T = x^{2n} = \sum_{a=0}^{n} \binom{n}{a} x^a \left(x(x-1)\right)^{n-a}$$

Gdzie:

x – ilość rodzajów jednostek występujących w sekwencjach (20 dla białek; 4 dla kwasów nukleinowych)

n – długość sekwencji (liczba porównywanych par pozycji)

a – ilość pozycji identycznych

# Local vs. Global

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

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

Global alignment: forces alignment in regions which differ

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

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

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

# Global - local

**Globalne dopasowanie** dla pary $S_1,\ S_2 \in \mathfrak{S}^\star$
to każda taka para $(S_1^\Upsilon, S_2^\Upsilon) \in (\mathfrak{S} \cup \text{`}-\text{'})^\star \times (\mathfrak{S} \cup \text{`}-\text{'})^\star$,
która spełnia warunki:

- $S_1$ otrzymuje się z $S_1^\Upsilon$ przez usunięcie wszystkich '$-$',
- $|S_1^\Upsilon| = |S_2^\Upsilon|$,
- $\forall i \in 1 \ldots |S_1^\Upsilon| : (S_A^\Upsilon(i) = \text{`}-\text{'}) \ \Rightarrow \ S_B^\Upsilon(i) \neq \text{`}-\text{'}$

**Lokalne dopasowanie** sekwencji $S_1, S_2$ to każde globalne uliniowienie dla pewnych podciągów $s_1 \in S_1^\star, s_2 \in S_2^\star$.

# 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:

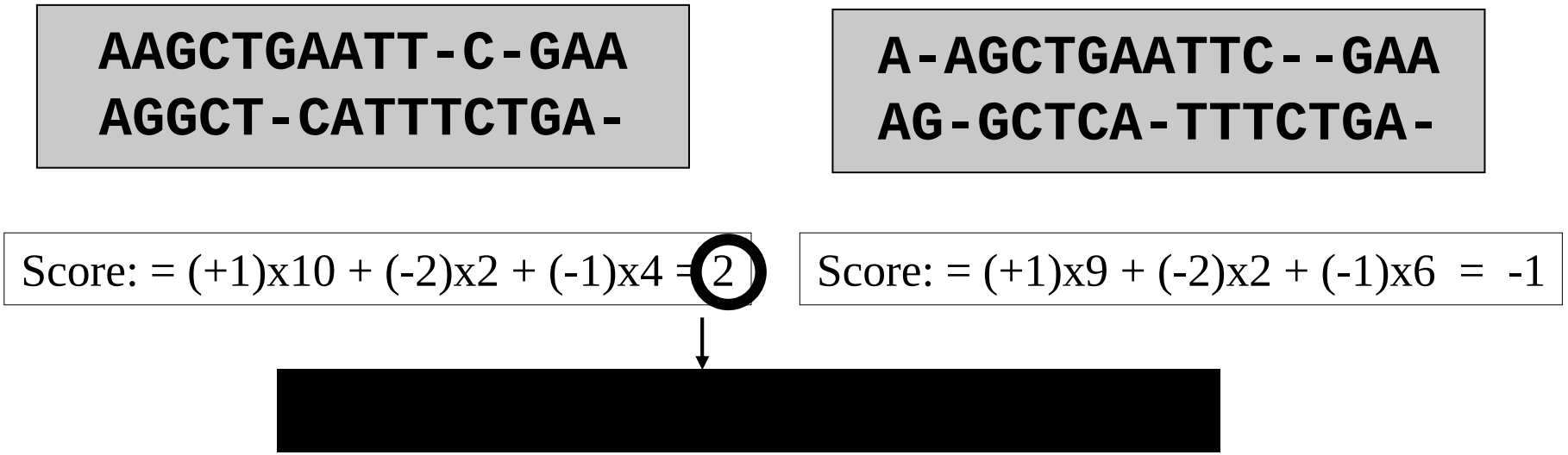
```
AAGCTGAATTCGAA
AGGCTCATTTCTGA
```

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

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

# scoring system:

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



# 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-**

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
```

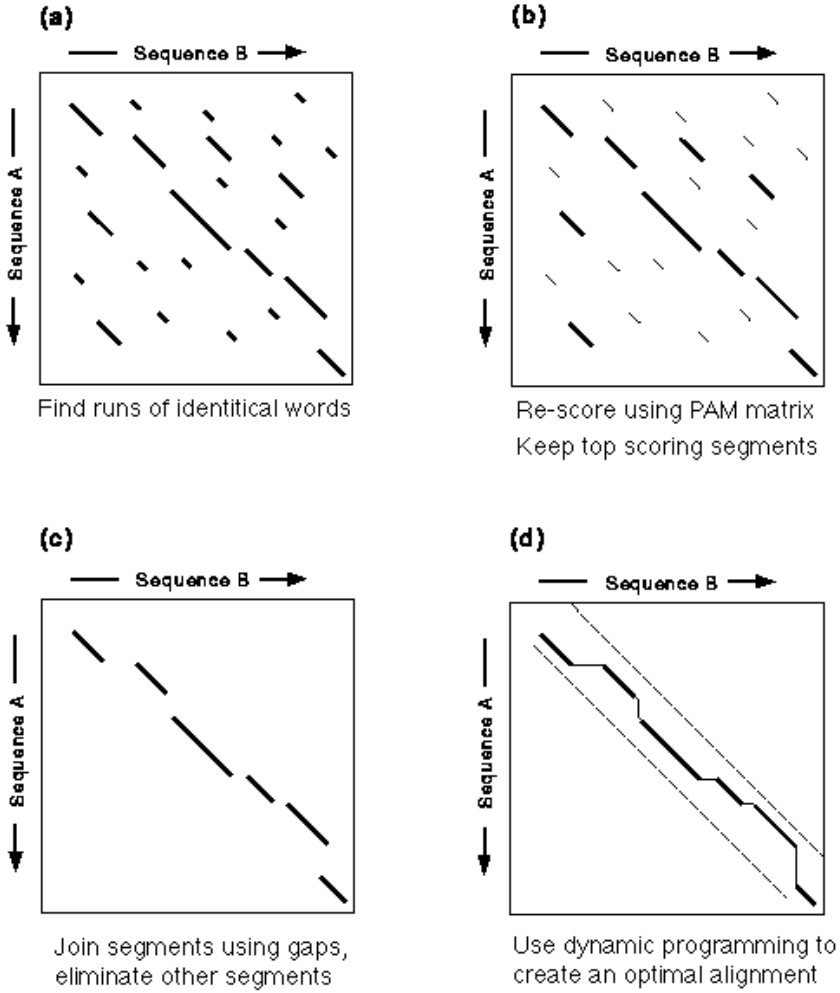
# Metody dopasowania

**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

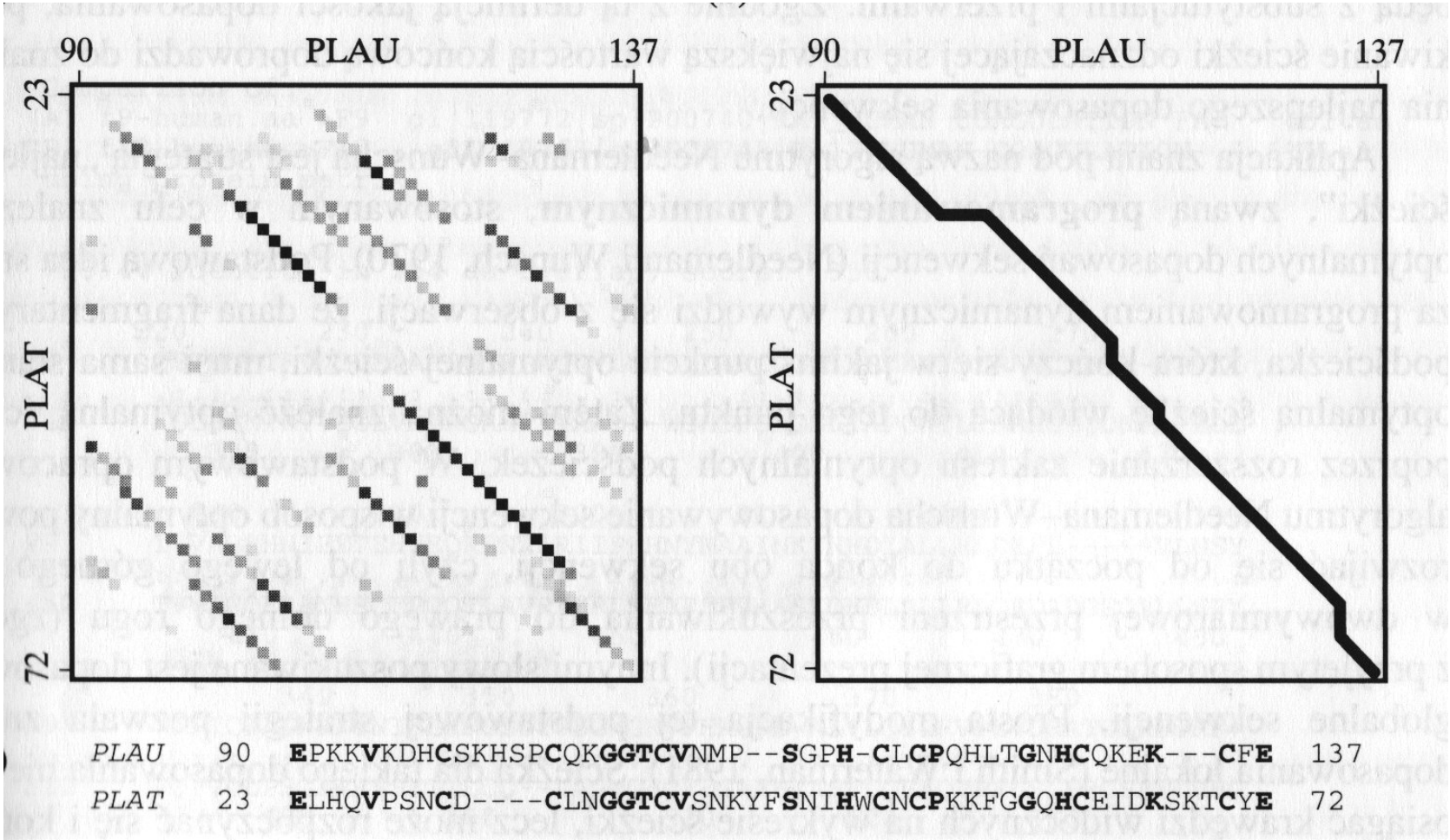
**FASTA Algorithm**



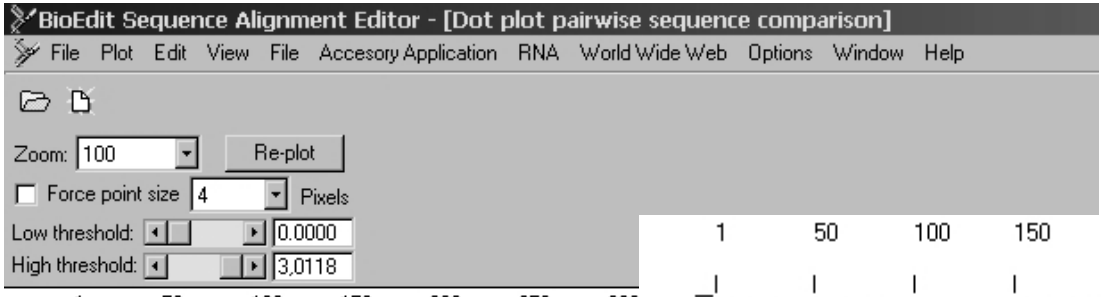
# 1. „słowa” - BLAST vs. FASTA

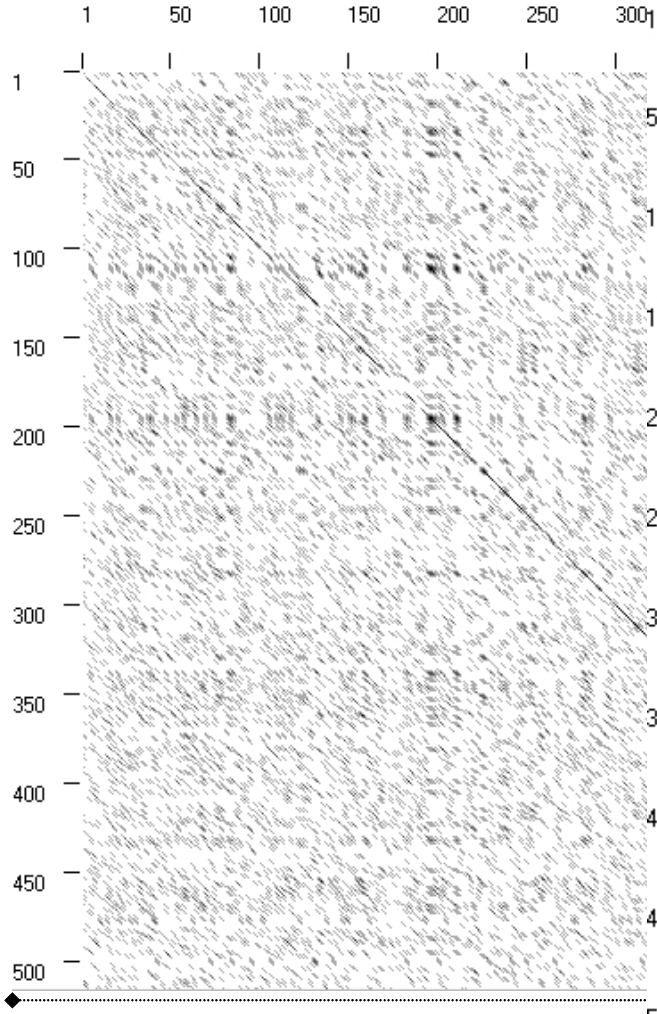
| BLAST | FASTA |
|---|---|
| może podawać więcej niż jeden region o wysokiej punktacji | podaje tylko jedno najlepsze dopasowanie |
| lepszy dla sekwencji białek niż DNA | lepszy dla sekwencji DNA niż białek |
| szybszy niż FASTA | wolniejszy niż BLAST |
| mniej czuły niż FASTA przy użyciu domyślnych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

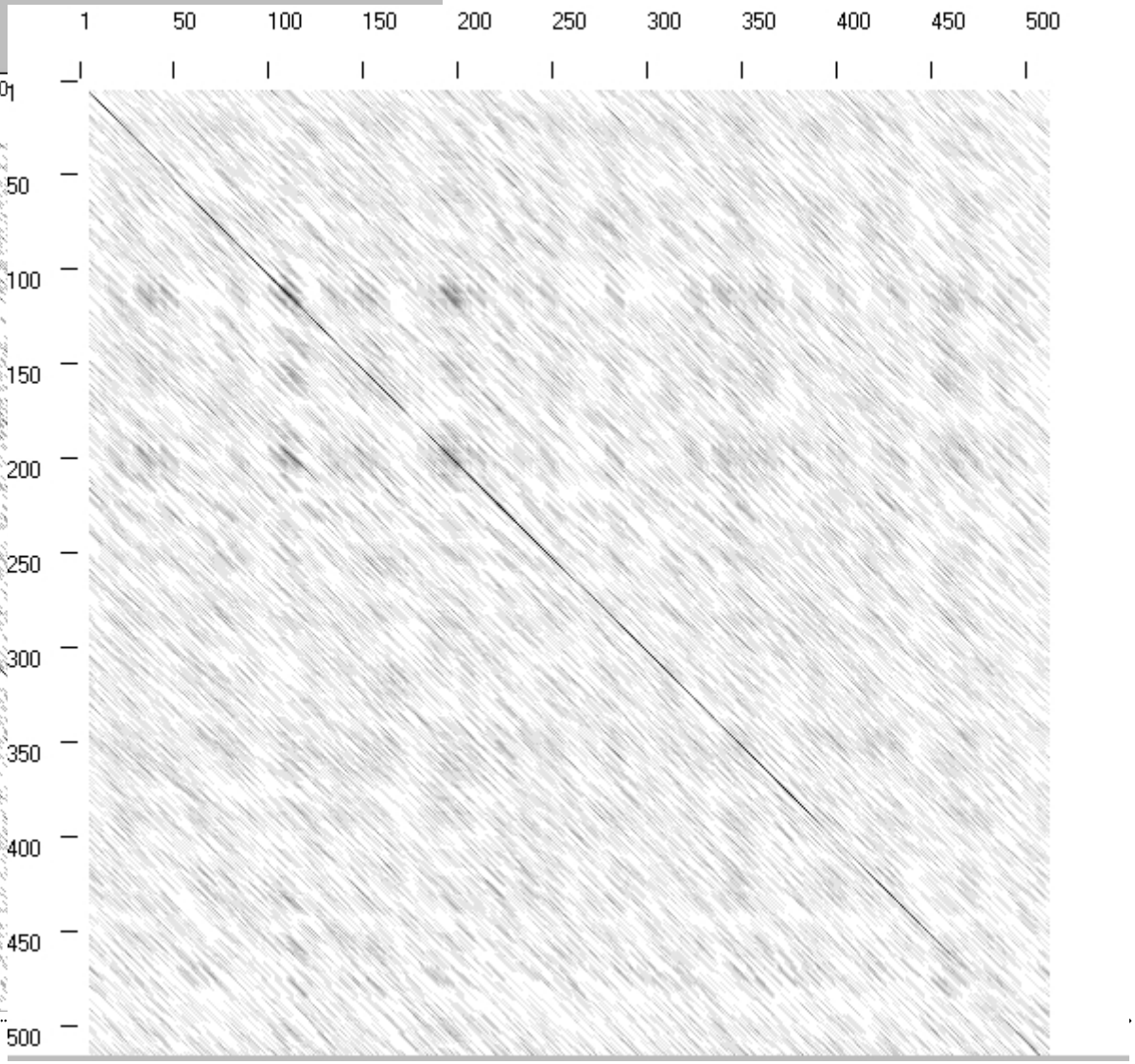
# 2. Macierze punktowe



# 2. Dot-matrix


BioEdit Sequence Alignment Editor - [Dot plot pairwise sequence comparison]
File Plot Edit View File Accesory Application RNA World Wide Web Options Window Help
Zoom: 100
Re-plot
Force point size 4 Pixels
Low threshold: 0.0000
High threshold: 3,0118


1 50 100 150 200 250 300
1 50 100 150 200 250 300 350 400 450 500


1 50 100 150 200 250 300 350 400 450 500
1 50 100 150 200 250 300 350 400 450 500

# 3. Programowanie dynamiczne

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

| | | B | | | | |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | |
| H | Δ | A | L | P | Q | |
| Δ | 0 | -1 | -2 | -3 | -4 | |
| 1 A | -1 | 1 | 0 | -1 | -2 | |
| 2 D | -2 | 0 | 1 | 0 | -1 | |
| A 3 L | -3 | -1 | 1 | 1 | 0 | |
| 4 P | -4 | -2 | 0 | 2 | 0 | |
| 5 Q | -5 | -3 | -1 | 1 | 3 | |

Alignment

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| A ... | A | D | L | P | Q | | |
| B ... | A | Δ | L | P | Q | | |
| Similarity Score | +1 | −1 | +1 | +1 | +1 | = | 3 |

Dynamic programming matrix:

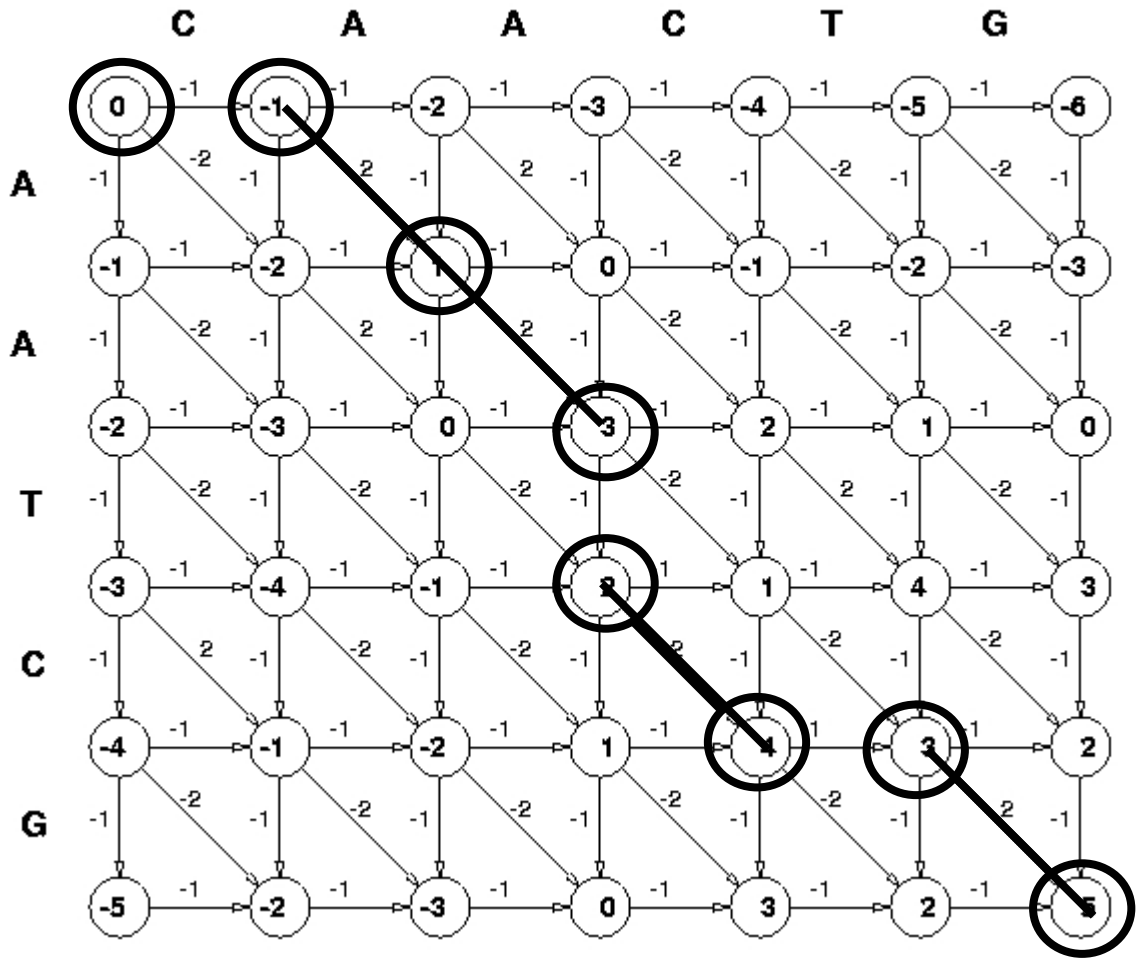
| | | j → (sequence y) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 0 | 1 T | 2 G | 3 C | 4 T | 5 C | 6 G | 7 T | 8 = N A |
| i | 0 | 0 | −6 | −12 | −18 | −24 | −30 | −36 | −42 | −48 |
| | 1 T | −6 | 5 | −1 | −7 | −13 | −19 | −25 | −31 | −37 |
| | 2 T | −12 | −1 | 3 | −3 | −2 | −8 | −14 | −20 | −26 |
| | 3 C | −18 | −7 | −3 | 8 | 2 | 3 | −3 | −9 | −15 |
| | 4 A | −24 | −13 | −9 | 2 | 6 | 0 | 1 | −5 | −4 |
| | 5 T | −30 | −19 | −15 | −4 | 7 | 4 | −2 | 6 | 0 |
| M = 6 | A | −36 | −25 | −21 | −10 | 1 | 5 | 2 | 0 | 11 |

(sequence x)

Optimum alignment scores 11:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| T | – | – | T | C | A | T | A |
| T | G | C | T | C | G | T | A |
| +5 | −6 | −6 | +5 | +5 | −2 | +5 | +5 |

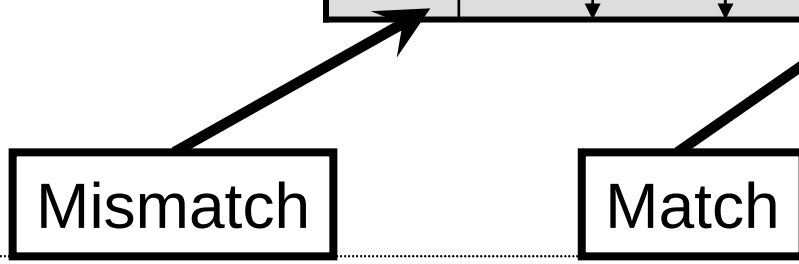
# 3. Programowanie dynamiczne



# 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** |



# Podobieństwa biochemiczne i biofizyczne aminokwasów

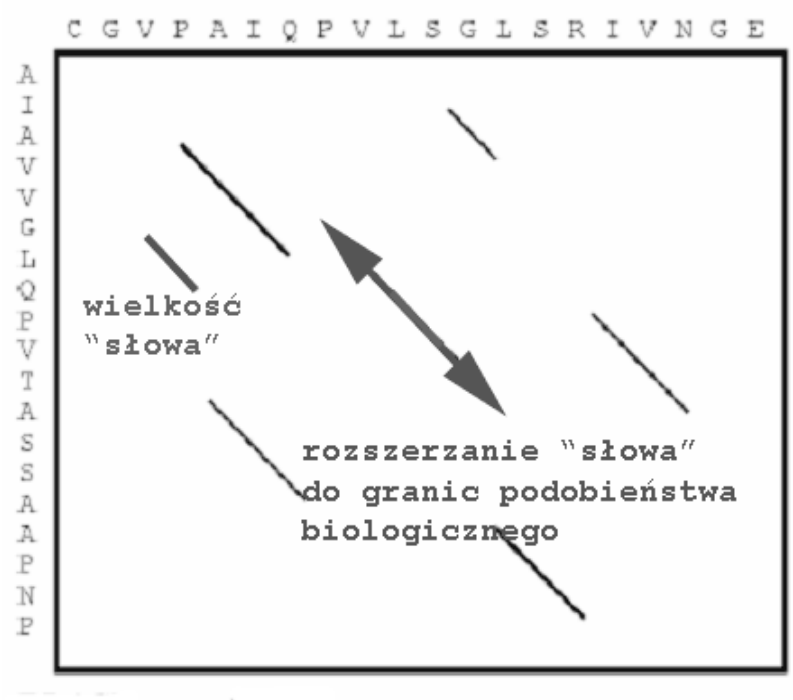
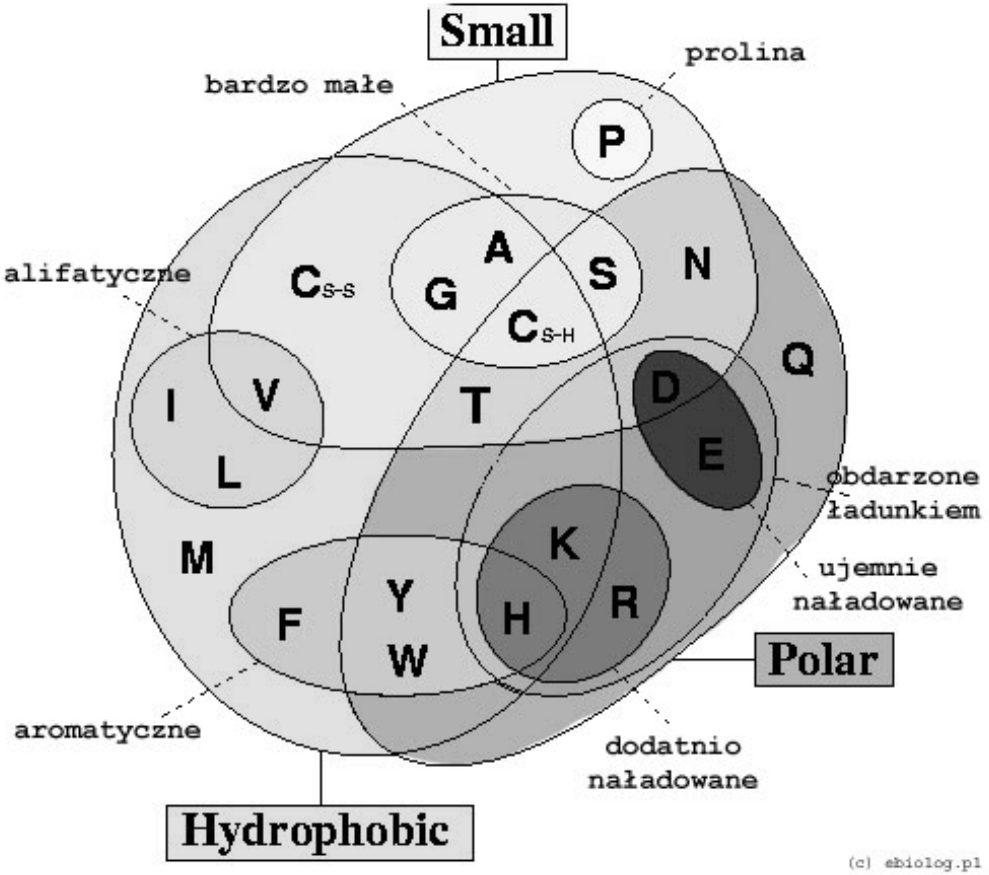


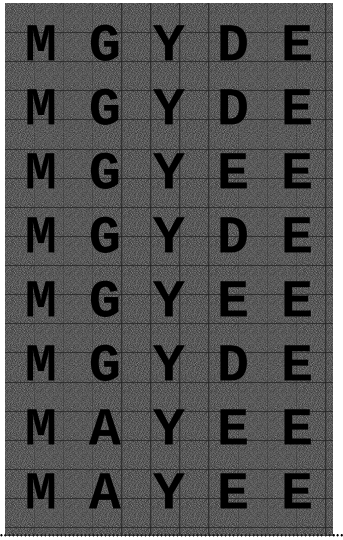

Diagram Venn-a



# 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 częstotliwość substytucji poszczególnych aminokwasów w konkretnych pozycjach.




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

# PAM Matrix – Point/Percent Accepted Mutations

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

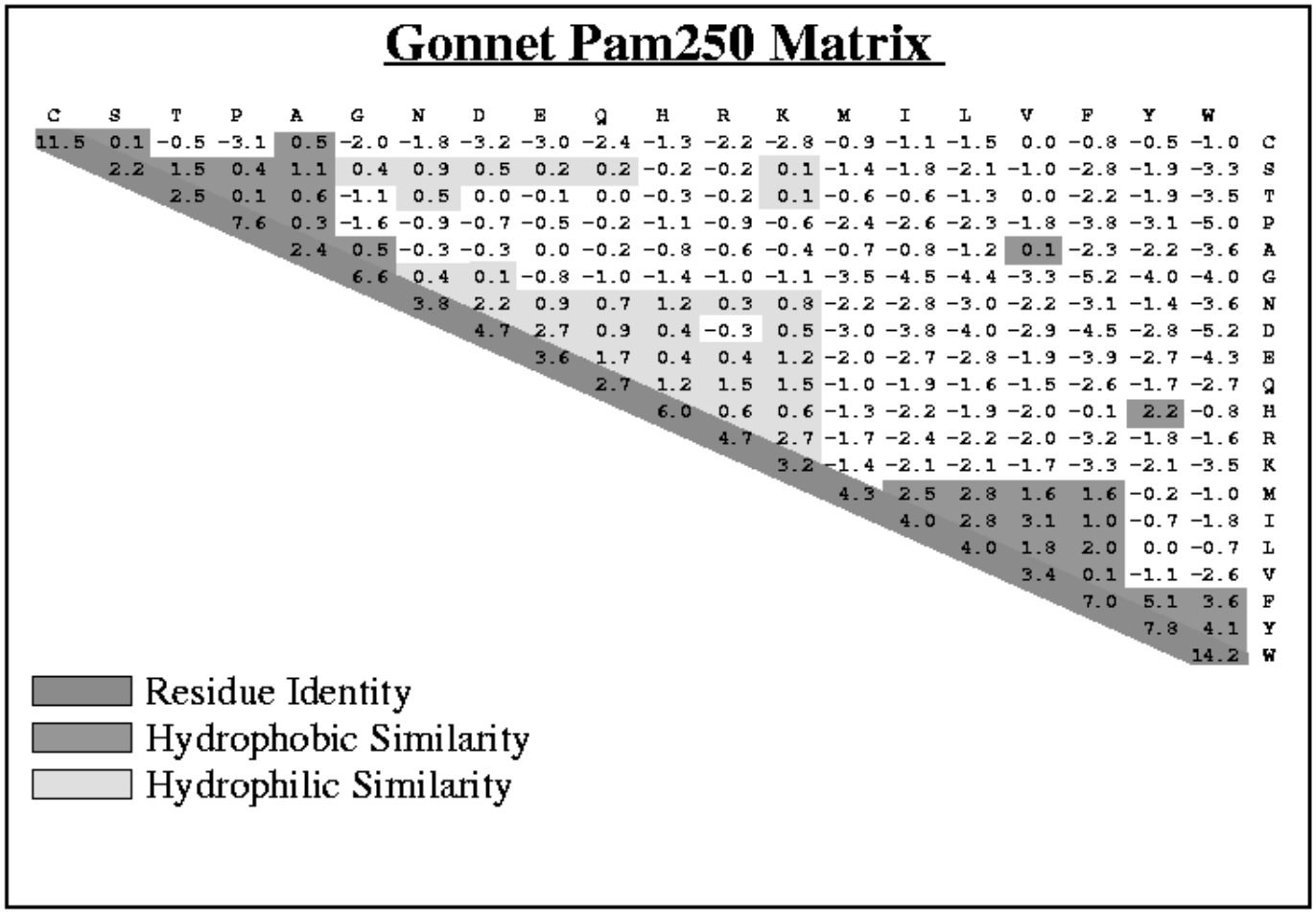
$S_1, S_2$ różnią się o jednostkę $n$ PAM, jeśli $S_2$ można otrzymać z $S_1$ w ciągu akceptowalnych mutacji punktowych takich, że średnia liczba nieletalnych mutacji na 100 wynosi $n$. Najpopularniejsza jest 250PAM.

- 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


Gonnet Pam250 Matrix
C S T P A G N D E Q H R K M I L V F Y W
11.5 0.1 -0.5 -3.1 0.5 -2.0 -1.8 -3.2 -3.0 -2.4 -1.3 -2.2 -2.8 -0.9 -1.1 -1.5 0.0 -0.8 -0.5 -1.0 C
2.2 1.5 0.4 1.1 0.4 0.9 0.5 0.2 0.2 -0.2 -0.2 0.1 -1.4 -1.8 -2.1 -1.0 -2.8 -1.9 -3.3 S
2.5 0.1 0.6 -1.1 0.5 0.0 -0.1 0.0 -0.3 -0.2 0.1 -0.6 -0.6 -1.3 0.0 -2.2 -1.9 -3.5 T
7.6 0.3 -1.6 -0.9 -0.7 -0.5 -0.2 -1.1 -0.9 -0.6 -2.4 -2.6 -2.3 -1.8 -3.8 -3.1 -5.0 P
2.4 0.5 -0.3 -0.3 0.0 -0.2 -0.8 -0.6 -0.4 -0.7 -0.8 -1.2 0.1 -2.3 -2.2 -3.6 A
6.6 0.4 0.1 -0.8 -1.0 -1.4 -1.0 -1.1 -3.5 -4.5 -4.4 -3.3 -5.2 -4.0 -4.0 G
3.8 2.2 0.9 0.7 1.2 0.3 0.8 -2.2 -2.8 -3.0 -2.2 -3.1 -1.4 -3.6 N
4.7 2.7 0.9 0.4 -0.3 0.5 -3.0 -3.8 -4.0 -2.9 -4.5 -2.8 -5.2 D
3.6 1.7 0.4 0.4 1.2 -2.0 -2.7 -2.8 -1.9 -3.9 -2.7 -4.3 E
2.7 1.2 1.5 1.5 -1.0 -1.9 -1.6 -1.5 -2.6 -1.7 -2.7 Q
6.0 0.6 0.6 -1.3 -2.2 -1.9 -2.0 -0.1 2.2 -0.8 H
4.7 2.7 -1.7 -2.4 -2.2 -2.0 -3.2 -1.8 -1.6 R
3.2 -1.4 -2.1 -2.1 -1.7 -3.3 -2.1 -3.5 K
4.3 2.5 2.8 1.6 1.6 -0.2 -1.0 M
4.0 2.8 3.1 1.0 -0.7 -1.8 I
4.0 1.8 2.0 0.0 -0.7 L
3.4 0.1 -1.1 -2.6 V
7.0 5.1 3.6 F
7.8 4.1 Y
14.2 W
Residue Identity
Hydrophobic Similarity
Hydrophilic Similarity

# 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

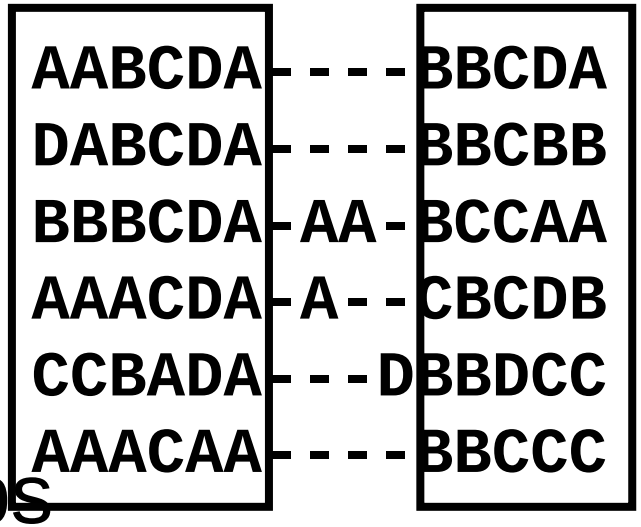
**BLOSSUM**$n$ (*Block Substitution Matrix* $n$)

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**

# Uwarunkowania genetyczne substytucji aminokwasowych

# Podstawy genetyczne algorytmów do zestawień aminokwasów?

| Replacement | PAM250 | BLOSUM62 |
|---|---|---|
| Arg/Lys | 3 | 2 |
| Lys/Gln | 1 | 1 |
| Arg/Gln | 1 | 1 |
| Lys/Glu | 0 | 1 |
| Arg/Glu | -1 | 0 |

# Algorytm semihomologiczny

# Dot matrix pairwise alignment

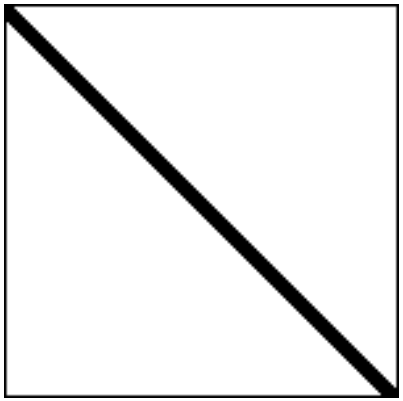
## Internal homology (gene multiplication)

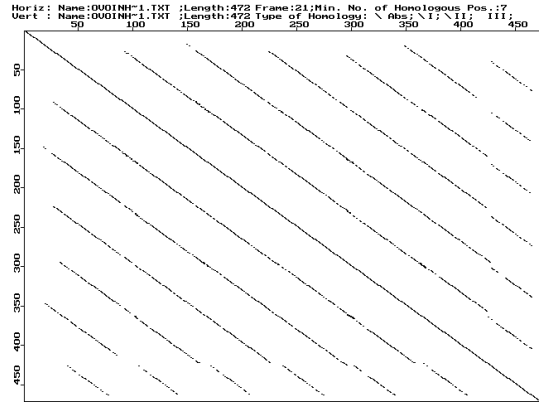
**BLAST 2 SEQUENCES**

**SEMIHOM**

Chicken ovoinhibitor precursor

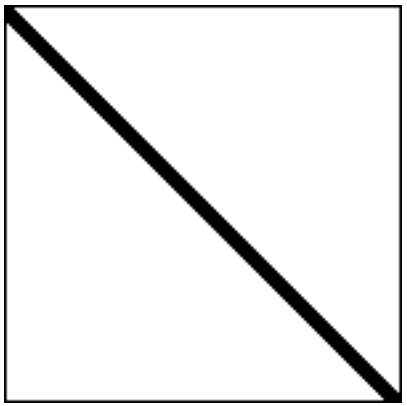
(7 domains)

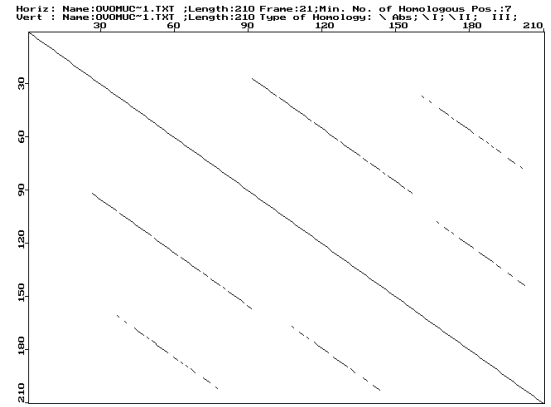




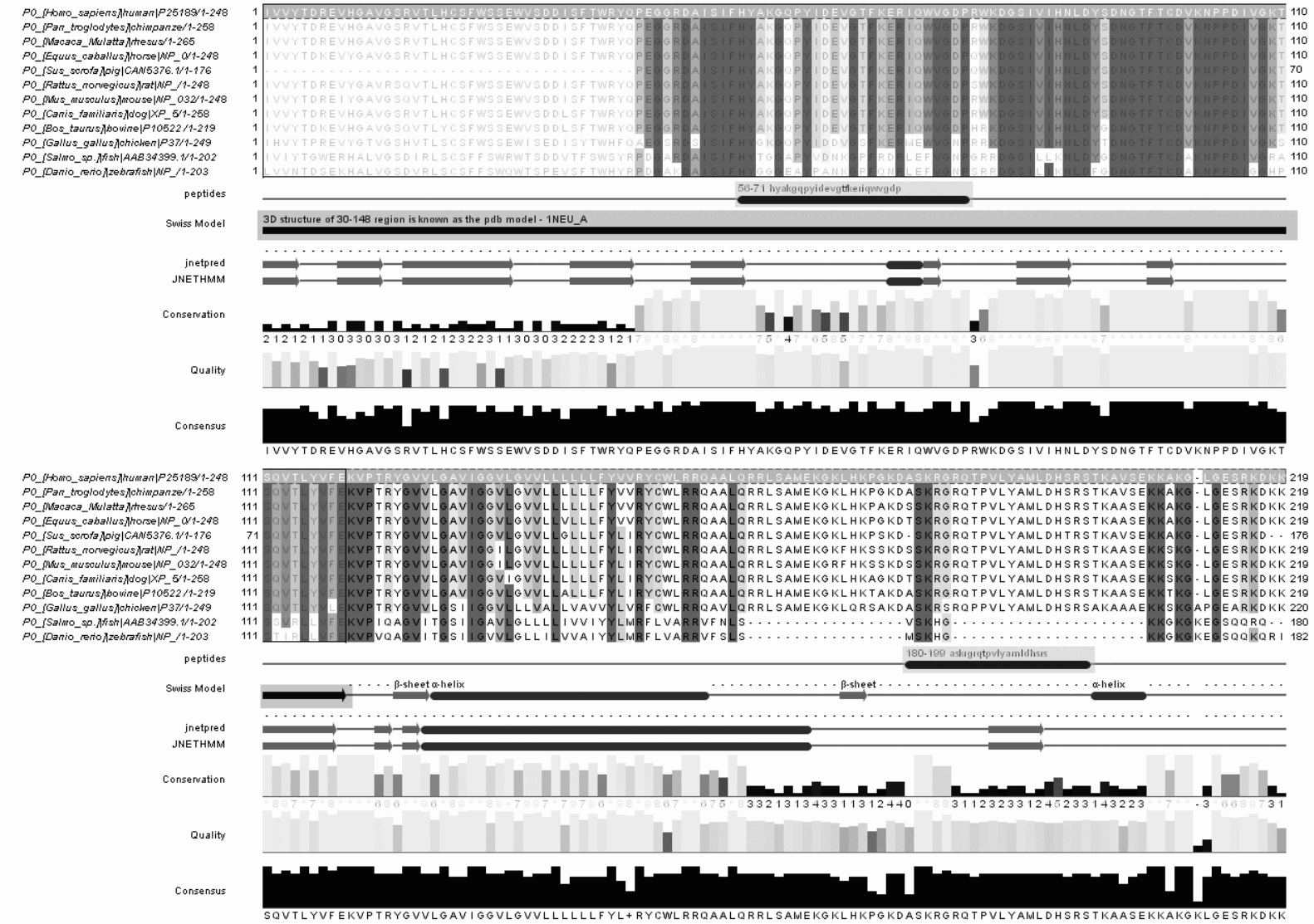

Chicken ovomucoid precursor

(3 domains)





# Multiple alignment



```
VTISCTGSSSNIGAG-NHVKWYQQLPG
VTISCTGTSSNIGS--ITVNWYQQLPG
LRLSCSSSGFIFSS--YAMYWVRQAPG
LSLTCTVSGTSFDD--YYSTWVRQPPG
PEVTCVVVDVSHEDPQVKFNWYVDG--
ATLVCLISDFYPGA--VTVAWKADS--
AALGCLVKDYFPEP--VTVSWNSG---
VSLTCLVKGFYPSD--IAVEWWSNG--
```

**Tak jak pairwise alignment ALE zestawienie *n* sekwencji zamiast *2***

**W rzędach ustawione są poszczególne sekwencje**
**W kolumnach ustawia się „te same”/”odpowiadające sobie” pozycje (pozycje konserwatywne); grupy pozycji konserwatywnych tworzą bloki konserwatywne (w blokach dozwolone są mutacje – insercje, delecje, substytucje – reprezentowane jako przerwy lub różne pozycje w kolumnach)**

# MSA & Evolution

## MSA może dawać obraz sił kształtujących ewolucję !!!

- Ważne aminokwasy lub nukleotydy (pozycje w sekwencjach) mutują „niechętnie”
- Mniej ważne pozycje dla struktury i funkcji mogą wykazywać większą zmienność w kolumnach porównywanych sekwencji

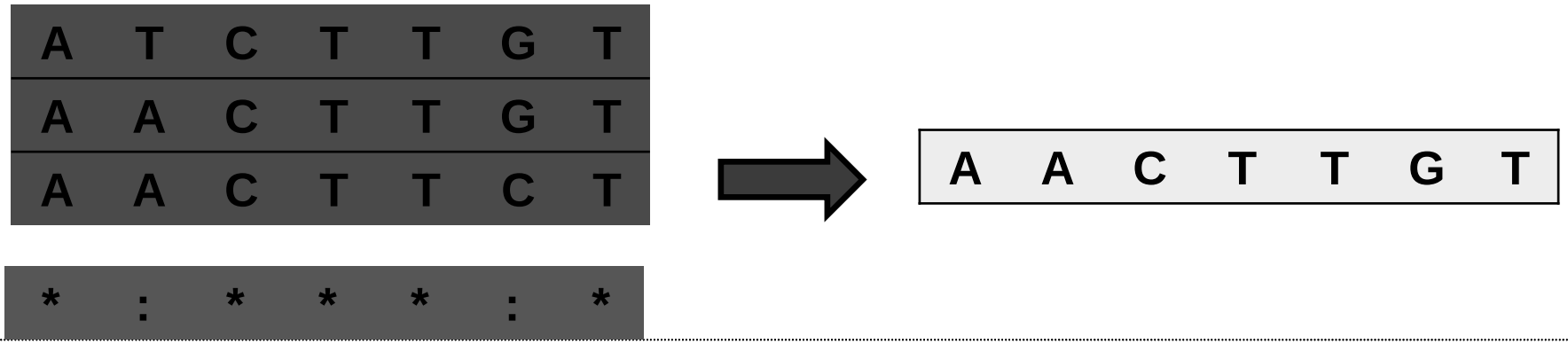
# Pozycje konserwatywne

- 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 LP G
VTIS CTGS SSNIGS--ITVN WYQQ LP G
LRLS CTGS GFIFSS--YAMY WYQQ AP G
LSLT CTGS GTSFDD-QYYST WYQQ PP G
```

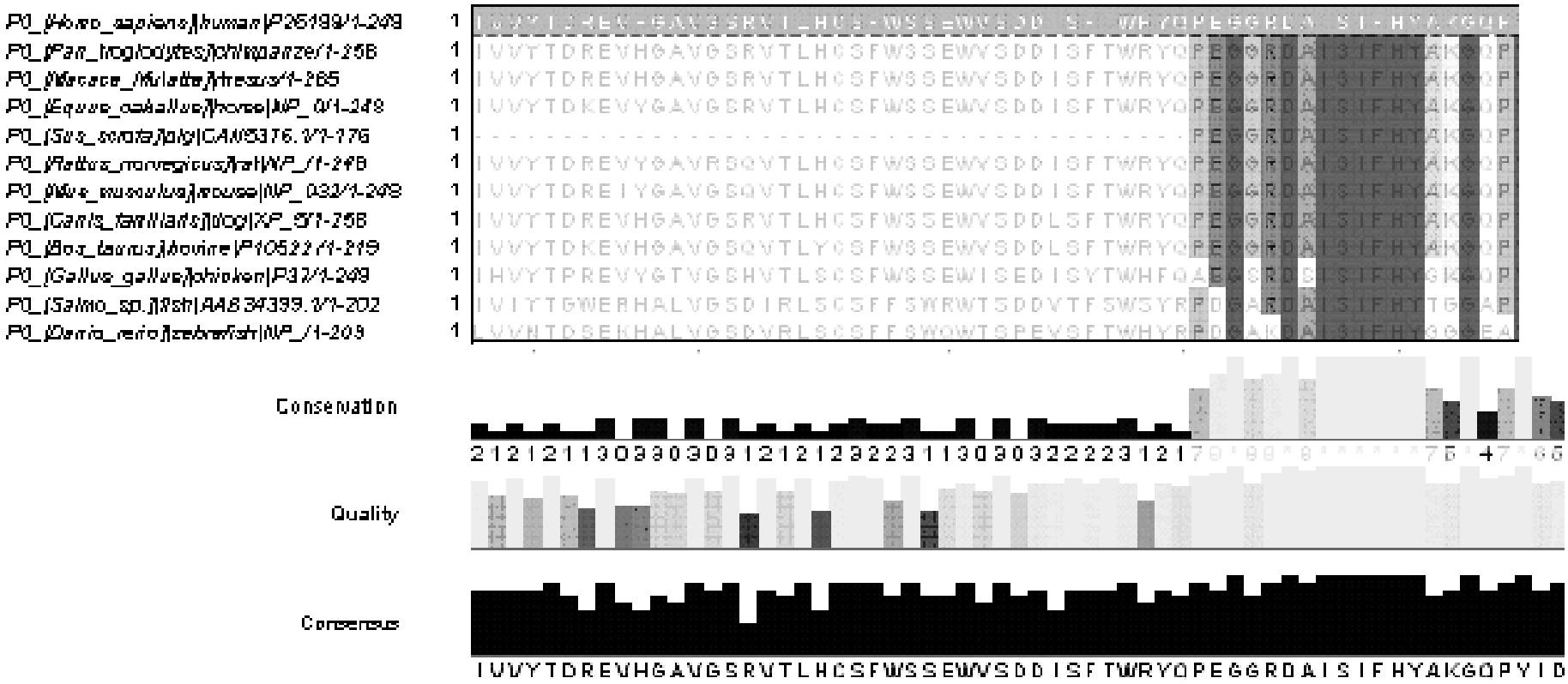
# Sekwencja konsensusowa

- W **sekwencji konsensusowej** zachowane są pozycje o największej częstotliwości występowania w każdej z kolumn alignmentu (The consensus sequence holds the most frequent character of the alignment at each column)
- Jest to sposób reprezentowania wyników multiple alignment, gdzie pokrewne sekwencje są porównywane każda do każdej, aby odnaleźć funkcjonalnie podobne motywy sekwencji (domeny białek). Sekwencja konsensusowa obrazuje które pozycje są konserwatywne, a które zmienne.



# Sekwencja konsensusowa

...******  ....  ********. ..
...  .....  .....



# Alignment methods

- Progressive alignment (Clustal)
- Iterative alignment (mafft, muscle)

- All methods today are an approximation strategy (**heuristic algorithm**), yield a possible alignment, but not necessarily the best one

# Praca domowa

- **iteracja** (np. pętle w programowaniu)
- **heurestyka** (głównie w informatyce)
- **Alignment progresywny**

# Jak wyświetlić na ekranie liczby od 1 do 5000 za pomocą 2 linijek kodu?

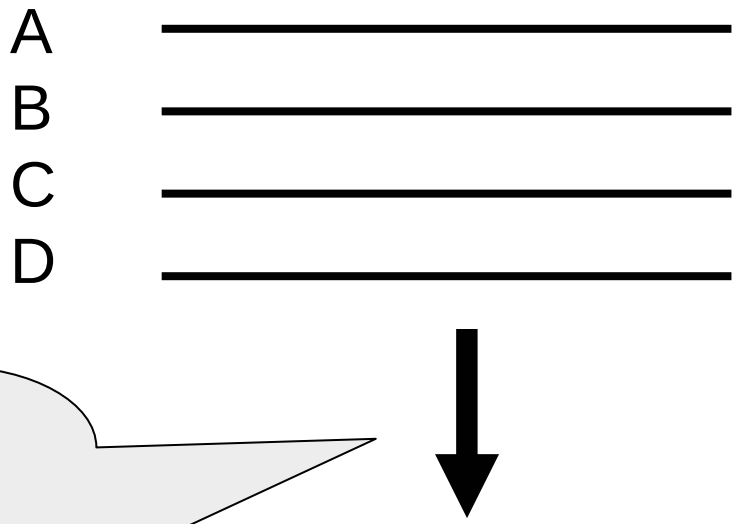
```
<?php
  for( $x = 1; $x = 5000; $x++ )
  echo $x. "<br /> ;
?>
```

```
dla (zmiennej x początkowo równej 1; aż do momentu kiedy
      zmienna x osiągnie wartość równą 5000; z każdym
      krokiem powiększając wartość zmiennej x o +1)
wyświetl wartość zmiennej i przejdź do nowej linijki;
```

**Iteracja** (łac. iteratio ‘powtórzenie’) to czynność powtarzania (najczęściej wielokrotnego) tej samej instrukcji (albo wielu instrukcji) w pętli. Mianem iteracji określa się także operacje wykonywane wewnątrz takiej pętli.

# Progressive alignment

**First step:**




**Compute the pairwise alignments for all against all (6 pairwise alignments) the similarities are stored in a table**

| | A | B | C | D |
|---|---|---|---|---|
| A | | | | |
| B | 11 | | | |
| C | 3 | 1 | | |
| D | 2 | 2 | 10 | |

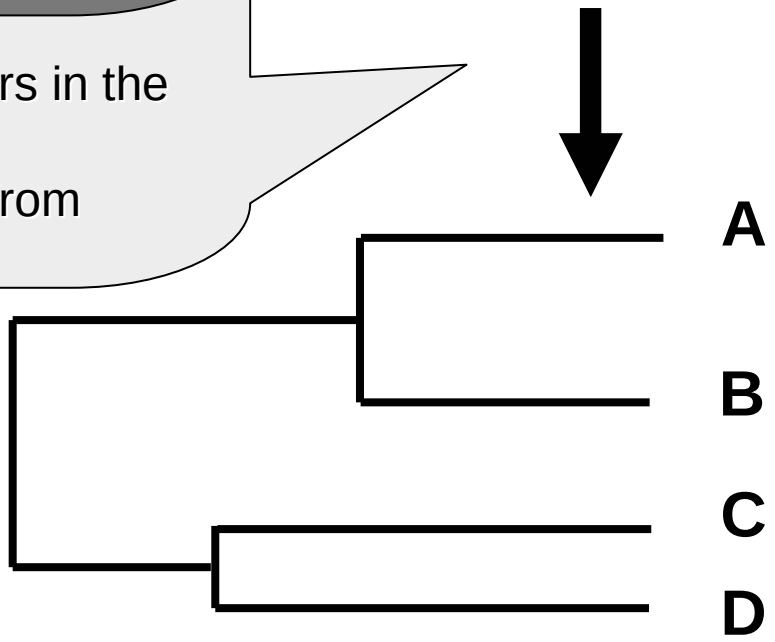
# Second step:

**The guide tree is imprecise and is NOT the tree which truly describes the relationship between the sequences!**

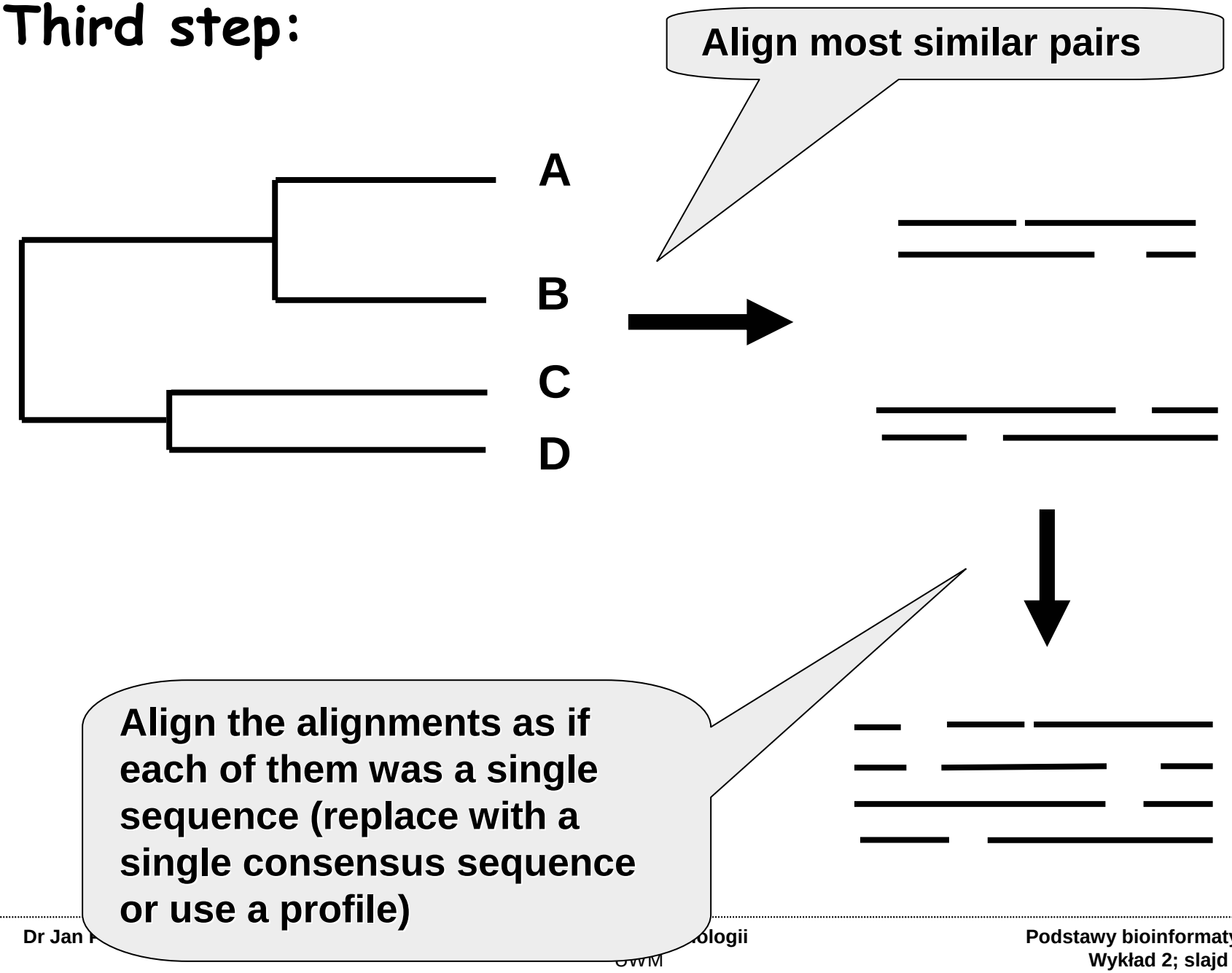
| | A | B | C | D |
|---|---|---|---|---|
| A | | | | |
| B | 11 | | | |
| C | 3 | 1 | | |
| D | 2 | 2 | 10 | |

sequences are to be aligned

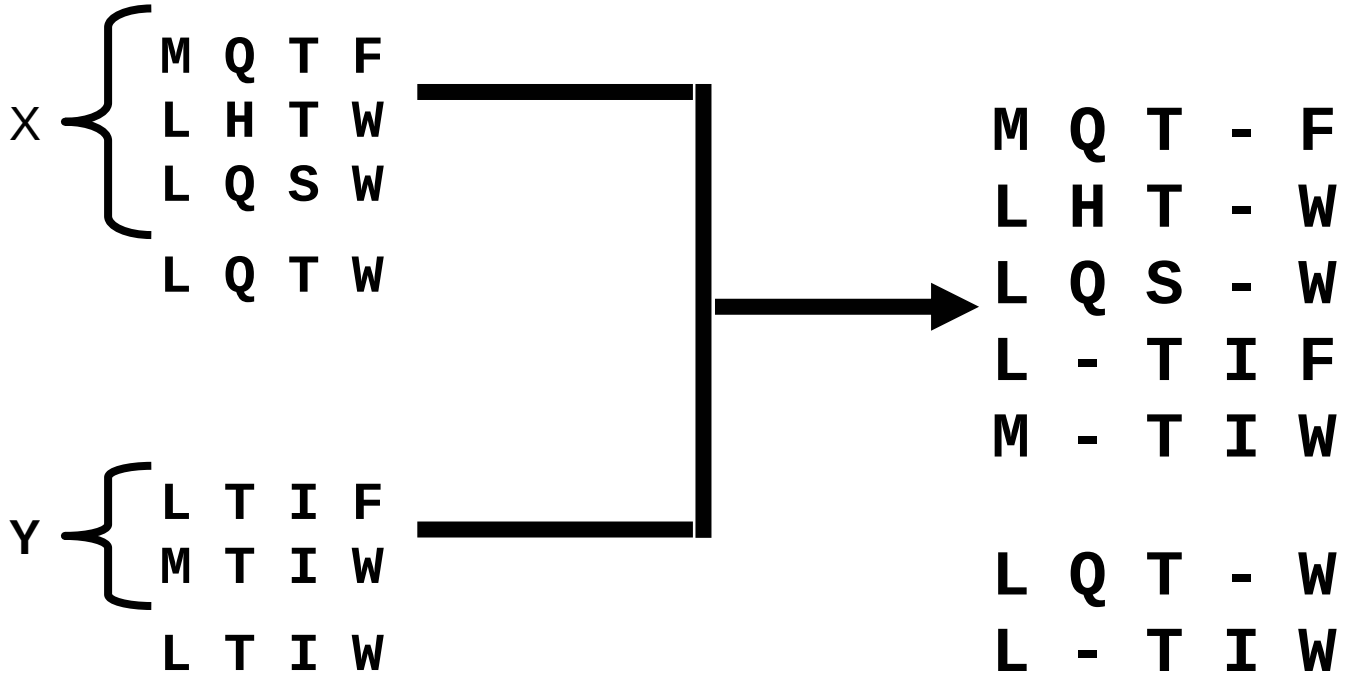
- similar sequences are neighbors in the tree
- distant sequences are distant from each other in the tree



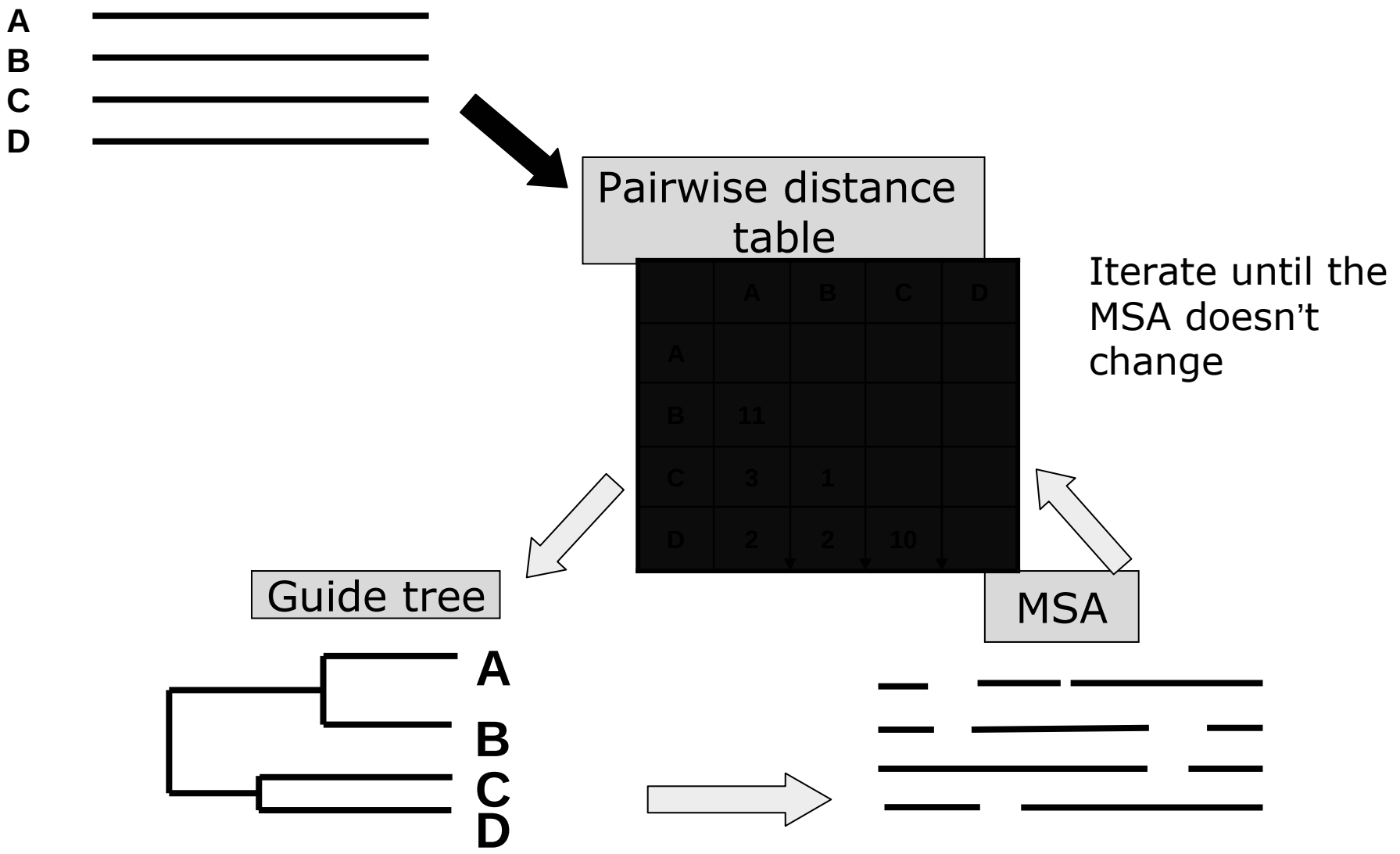
# Third step:


Align most similar pairs
A
B
C
D
Align the alignments as if each of them was a single sequence (replace with a single consensus sequence or use a profile)

# Alignment of alignments



# Iterative alignment



# Searching for remote homologs

- Sometimes BLAST isn’t enough.
- Large protein family, and BLAST only gives close members. We want more distant members
- PSI-BLAST
- Profile HMMs

# Profile

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

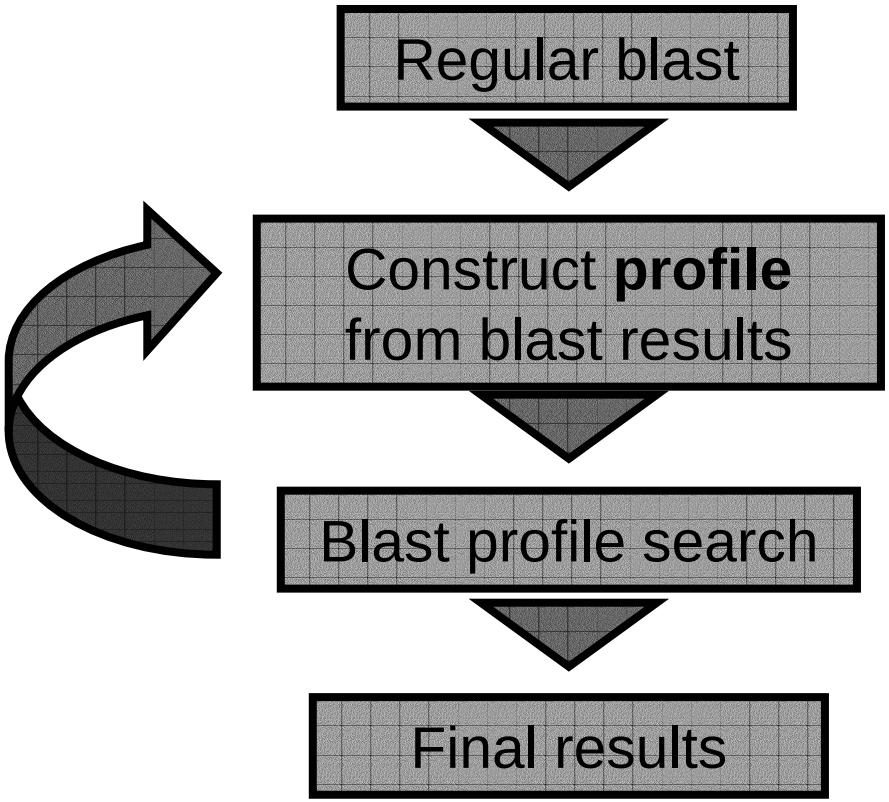
| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| A | 1 | 0.67 | 0 | 0 | . | . |
| T | 0 | 0.33 | 1 | 1 | . | . |
| C | 0 | 0 | 0 | 0 | . | . |
| G | 0 | 0 | 0 | 0 | . | . |

Profile =

PSSM – Position Specific Score Matrix

# PSI-BLAST

- **P**osition **S**pecific **I**terated BLAST



# PSI-BLAST

- zalety: PSI-BLAST looks for seq.s that are close to ours, and learns from them to extend the circle of friends
- wady: if we found a WRONG sequence, we will get to unrelated sequences (contamination). This gets worse and worse each iteration

# Profile HMM

- Similar to PSI-BLAST: also uses a profile
- Takes into account:
  - Dependence among sites (if site *n* is conserved, it is likely that site *n+1* is conserved → part of a domain
  - The probability of a certain column in an alignment

# PSI BLAST vs profile HMM

| PSI BLAST | Profile HMM |
| --- | --- |
| Less exact<br>Faster | More exact<br>Slower |