

Niniejsze opracowanie zostało stworzone przez dr Magdę Mielczarek, pracownika Uniwersytetu Przyrodniczego we Wrocławiu w ramach wykonywania obowiązków związanych z kształceniem studentów i jest przeznaczone dla studentów Biologii (Wydział Biologii i Hodowli Zwierząt) na potrzeby dydaktyczne bez prawa do dalszego rozpowszechniania.

DOPASOWANIE SEKWENCJI

Wykład 5

DOPASOWANIE SEKWENCJI

Dopasowanie = przyrównanie (sequence alignment)

Poszukiwanie ciągów znaków (zasad nukleotydowych lub reszt aminokwasowych), które posiadają to samo ułożenie w przyrównywanych sekwencjach.

Stwierdzone różnice pomiędzy sekwencjami świadczą o mutacjach, które zaszły po rozdzieleniu się sekwencji od wspólnego przodka.

DOPASOWANIE SEKWENCJI

KATEGORIA STRUKTURALNA

- Identyfikacja regionów o zakonserwowanej strukturze - korespondują z grupami czynnymi białek

KATEGORIA FUNKCJONALNA

- podobne funkcje
- Identyfikacja funkcji nowych sekwencji – poznanie funkcji nowo zsekwencjonowanych regionów

KATEGORIA EWOLUCYJNA

- pochodzenie od wspólnego przodka
- Identyfikacja powiązań ewolucyjnych pomiędzy osobnikami

Podobieństwo sekwencji może wynikać z homologii lub homoplazji (konwergencji).

DOPASOWANIE (LICZBA SEKWENCJI)

PODWÓJNE

**Pair-wise alignment
(dwie sekwencje)**

WIELOKROTNE

**Multiple sequence
alignment (wiele sekwencji)**

DOPASOWANIE (POKRYCIE SEKWENCJI)

GLOBALNE

- Obejmuje całe sekwencje
- Sekwencje podobne na całej długości
- Bliskie pokrewieństwo, podobna długość

Global	FTFTALILLAVAV
	F--TAL-LLA-AV
Local	FTFTALILL-AVAV
	--FTAL-LLAAV--

LOKALNE

- Identyfikacja podobnych regionów w obrębie przyrównanych sekwencji
- Sekwencje o różnych długościach
- Sekwencje o niskim poziomie podobieństwa
- Wiele możliwych dopasowań

DOPASOWANIE SEKWENCJI NUKLEOTYDÓW



WYBÓR NAJLEPSZEGO DOPASOWANIA

g c t g a a c g

c t a t a a t c

- - - - - g c t g a a c g

c t a t a a t c - - - - -

g c t g a a c g

c t a t a a t c

g c t g a - a - - c g

- - c t - a t a a t c

g c t g - a a - c g

- c t a t a a t c -

WYBÓR NAJLEPSZEGO DOPASOWANIA

Algorytmy dopasowania sekwencji:

- macierze punktowe (dot-matrix)
- programowanie dynamiczne (dynamic programming)
- metody słowne (k – tuple methods)

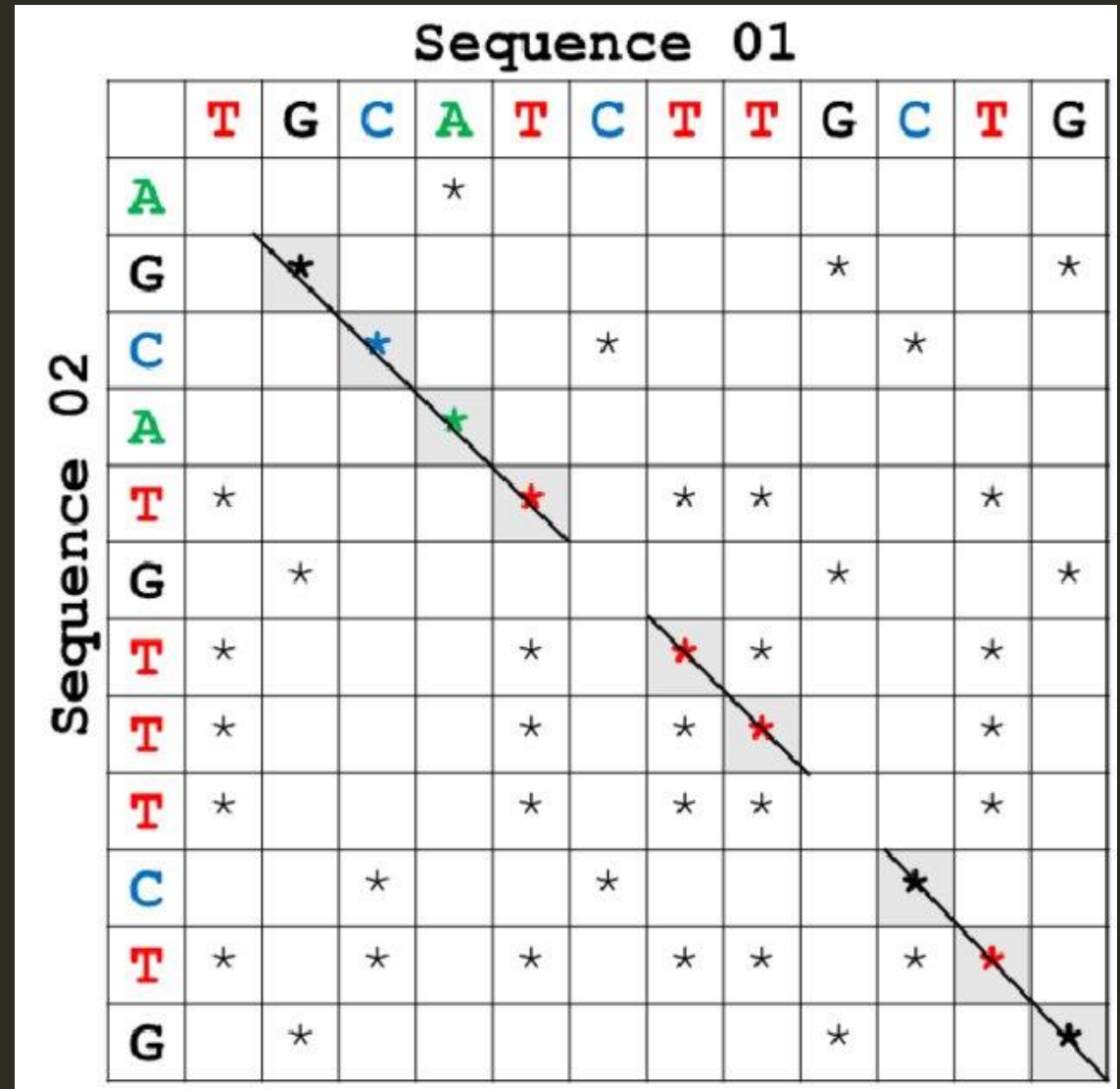
DOPASOWANIE SEKWENCJI

Algorytmy dopasowania sekwencji:

- **macierze punktowe (dot-matrix)**
- programowanie dynamiczne (dynamic programming)
- metody słowne (k – tuple methods)

DOT-MATRIX

Diniz WJ, Canduri F. REVIEW-ARTICLE
Bioinformatics: an overview and its
applications. Genet Mol Res. 2017 Mar
15;16(1). doi: 10.4238/gmr16019645.
PMID: 28301675.



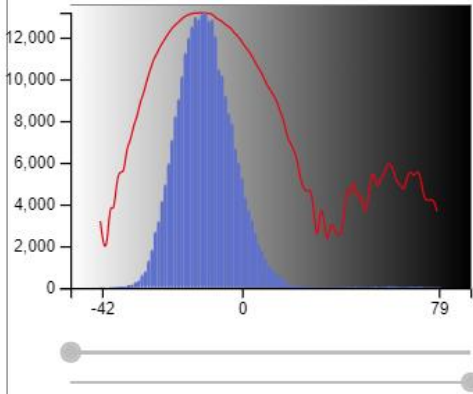
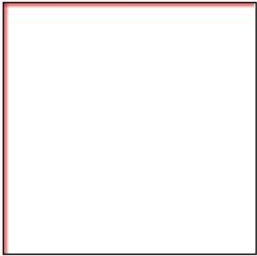
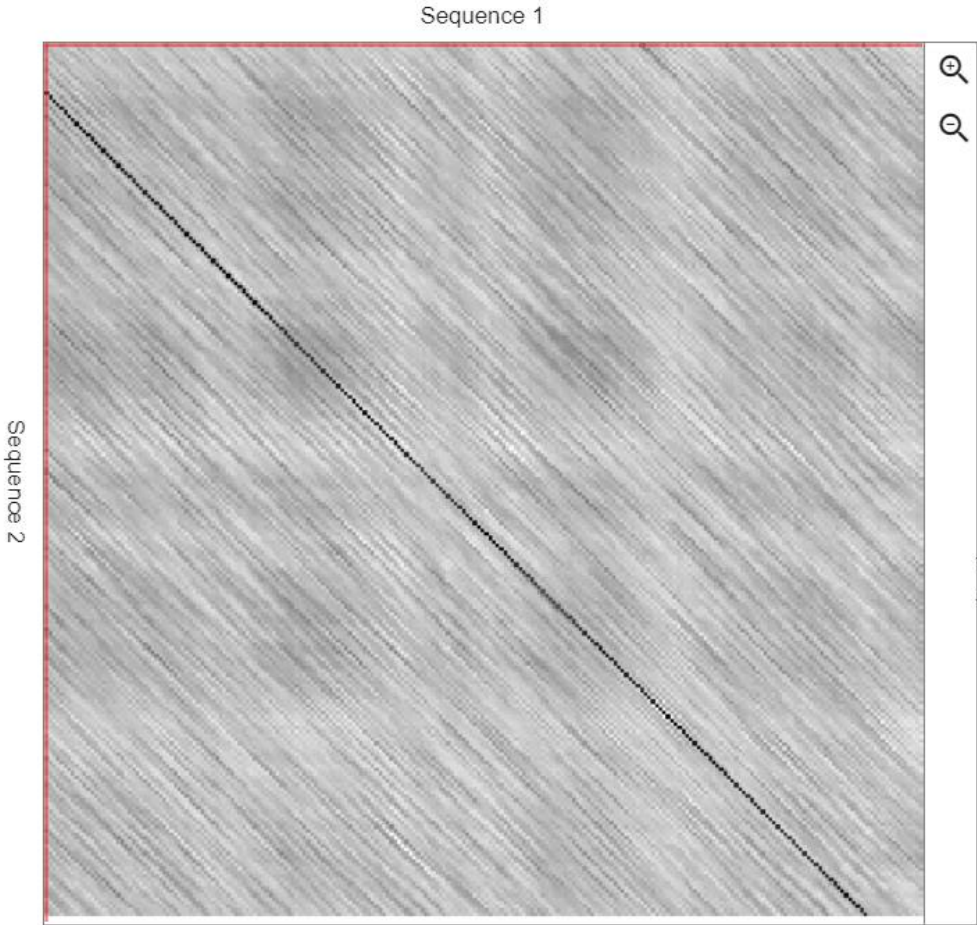
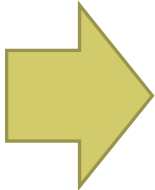
SEQUENCE 1

SEQUENCE 2

Window size
15

Scoring matrix
BLOSUM 62





[246 x 244] # Score at (1:M, 1:R) : -8



DOT-MATRIX

DOPASOWANIE: Leptyna *Homo sapiens* ↔ Leptyna *Mus musculus*

NCBI Resources How To				
Gene		Gene	leptin	
Name/Gene ID	Description	Location	Aliases	MIM
☐ LEPR ID: 3953	leptin receptor [<i>Homo sapiens</i> (human)]	Chromosome 1, NC_000001.11 (65420652..65641559)	CD295, LEP-RD, OB-R, OBR, LEPR	601007
☐ LEP ID: 3952	leptin [<i>Homo sapiens</i> (human)]	Chromosome 7, NC_000007.14 (128241201..128257629)	LEPD, OB, OBS	164160
☐ lepa ID: 106561227	leptin a [<i>Salmo salar</i> (Atlantic salmon)]	Chromosome ssa10, NC_027309.1 (86214874..86217578)	leptin	
☐ lepa ID: 111953061	leptin a [<i>Salvelinus alpinus</i> (Arctic char)]	Chromosome LG26, NC_036866.1 (27908024..27909782)	LEP, leptin	

LEP leptin [*Homo sapiens* (human)]

Gene ID: 3952, updated on 7-Mar-2021

Download Datasets

Summary

Official Symbol LEP provided by HGNC

Official Full Name leptin provided by HGNC

Primary source [HGNC:HGNC:6553](#)

See related [Ensembl:ENSG00000174697](#) [MIM:164160](#)

Gene type protein coding

RefSeq status REVIEWED

Organism [Homo sapiens](#)

Lineage Eukaryota; Metazoa; Chordata; Mammalia; Primates; Hominidae; Homo

Also known as OB; OBS; LEPD

Summary This gene encodes a protein that is involved in energy homeostasis. Circulating leptin promotes energy expenditure, promotes responses, hematopoietic responses, and causes severe obesity and diabetes mellitus development.

Expression Restricted expression to adipocytes

Orthologs [mouse](#) [all](#)

NEW

Try the new [Gene tab](#)

Try the new [Transcript tab](#)

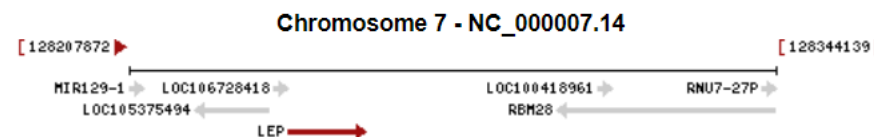
Genomic context

Location: 7q32.1

See LEP in [Genome Data Viewer](#)

Exon count: 3

Annotation release	Status	Assembly	Chr	Location
109.20210226	current	GRCh38.p13 (GCF_000001405.39)	7	NC_000007.14 (128241201..128257629)
105.20201022	previous assembly	GRCh37.p13 (GCF_000001405.25)	7	NC_000007.13 (127881331..127897682)



Genomic regions, transcripts, and products

Go to [reference sequence details](#)

Genomic Sequence: [NC_000007.14](#) Chromosome 7 Reference GRCh38.p13 Primary Assembly

Go to nucleotide: [Graphics](#) [FASTA](#) [GenBank](#)

Nucleotide

Nucleotide

Advanced

FASTA

Homo sapiens chromosome 7, GRCh38.p13 Primary Assembly

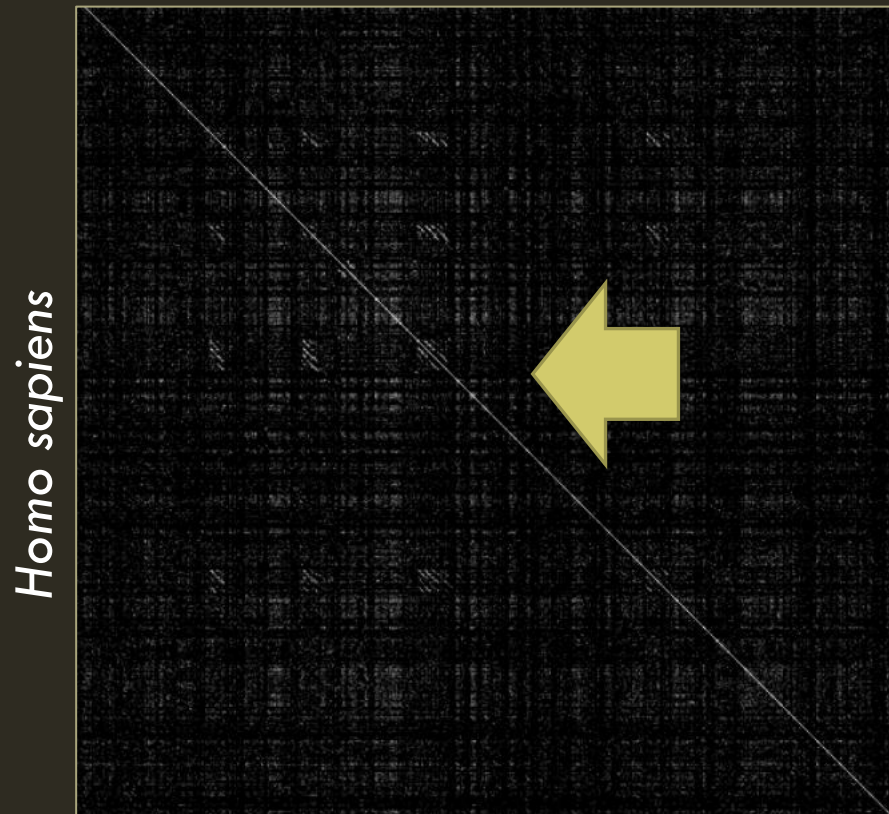
NCBI Reference Sequence: NC_000007.14

[GenBank](#) [Graphics](#)

```
>NC_000007.14:128241201-128257629 Homo sapiens chromosome 7, GRCh38.p13 Primary Assembly
ATGCGCCGGGGCCTGCGGGGCAGTTGCGCAAGTTGTGATCGGGCCGCTATAAGAGGGGCGGGCAGGCATG
GAGCCCCGTAGGAATCGCAGCGCCAGCGGTTGCAAGGTAAGGCCCCGGCGCGCTCCTTCCTCCTTCTCTG
CTGGTCTTTCTTGGCAGGCCACAGGGCCCCACACAACCTCTGGATCCCGGGGAAACTGAGTCAGGAGGGAT
GCAGGGCGGATGGCTTAGTTCTGGAATATGATAGCTTTGTACCGAGTTCTAGCCAGATAGAAGGTTACCG
GGAGCTGGGGAGCGTTGGATTTGCTGCTGGGCTGTGCCGGTGCCAGAAAGGCAGGACCTTGCAGAACCAG
CCAGGTCCTTGGGAGACTGTCAGACCCACCAACCTGGTGGCATTTCGAGAGCTGAGATGCATTGGAAATT
GCCTTGGGCACATCCCCAAAGATCAGGATGTCCCACCCAGTCTGAAGGAGATAAAGTTGGGGGTAGGAG
AGACGCAGATGCAAGTGATCAGTCTCAGTCCCAGACATTGCCTTGCTCTGCGGGTAGGAATTCAGGATTC
ATTTTCCAGGGAAGTTCTGACCTCTGAATGAGAGGGGCTGTGTAAGGCCAATGCCTGGGAGGAAGGCAA
GGATGAGTAGAGGTGGGGGGAACAAGTGTCAGGAAGACTCAAATCTTCCAGAGAAATTGTGCAGGGTC
TTACCAGATCTGTCTCAAAGCCATGCAAATTGCCTTCTTTGCAATGCATACAATGAGGTGTCTCTGGGG
GTCAGAACTGGTTATTAGGGAACCTTAGCCAGGACTGCTAAATACGCGCTGTTGGCCACCAGGCTCAC
CTATAGCCTTCCTTCAGTCTGGGCTTGGTTTGGATTTCACTGTGGGTGCCATCGCCTTTACACTCCTGTT
TCTATAGTTTAAAGATAGTGGTGCTTTGGGAAAGTGACTCCTTAAATACAGTTAGGTCCAAGTGAGACAA
GTGGCCTGGCTGTCAATTCAGAATAGCAGCTTCCAAGAGGTGATTAATTTCTGTTGGAAGGGTGATCTTT
GGGGAGGTGGGTGAAGAGCAGAGACTTGGTGGTACC GTTCCAGGAGCACAGGCTCTCTTCTTTGCAGTG
CAGAATGACCTCTGGCAGCCGGAGTTGTGTTTGTCTGTAGGATTCTGAGGTGGGCCATGGGCAGCTGGA
ACTGGGGAATTTTGCCAACTCTTTTCATATTAGGATTGTCTGCAGAACCAGATATGGAGGCTTCTAGCAA
CGTGAGTGCTCCTGTTCTAATGCCCTTAGAAAACAAGAAGGCCACACTGATCATTCTCTCACTTAGGCAG
GGAGACAAGGCAAGAGAGAAAACAGTGGATGCTTTTAGGTTCTTTCCCTTCCCAAGCAGTTGTGGACATTG
GGCTGAGGGGAACATTTCCACATTGGCTAAAGGAGCGTCCTCCTCATATTTGTACATTTTATACCCAAA
ATAACTCTTCTTGGTATTTGGGGAAATATTTTCTCCCCGTCCATTCCAGGAAATGGCTCCAAGTGCCAA
GGACAGAGCCAGGGAAGTTGCAATGAATTCCTGCCCCGTAGCCCCAGGCAGATGCCTTGACAGTCTGAGT
```

DOT-MATRIX

DOPASOWANIE: Leptyna *Homo sapiens* ↔ Leptyna *Homo sapiens*
Homo sapiens



DOT-MATRIX

DOPASOWANIE: Leptyna *Homo sapiens* ↔ Leptyna *Mus musculus*

NCBI Resources How To			
Gene		Gene	leptin
☐ lepa	leptin a [<i>Tachysurus fulvidraco</i> (yellow catfish)]		leptin
☐ Lep	leptin [<i>Mus musculus</i> (house mouse)]	Chromosome 6, NC_000072.7 (29060220..29073875)	ob, obese
☐ Lep	leptin [<i>Rattus norvegicus</i> (Norway rat)]	Chromosome 4, NC_051339.1 (57661127..57675262)	OB, obese
☐ Lepr	leptin receptor [<i>Mus musculus</i> (house mouse)]	Chromosome 4, NC_000070.7 (101574393..101676375)	LE, LEPROTb, Modb, Modb1, O, OB-, OB-RGRP, Obr, db, diabetes, obese-, obese-like, obl, Lepr

Lep leptin [*Mus musculus* (house mouse)]

Gene ID: 16846, updated on 16-Mar-2021

Download Datasets

Summary

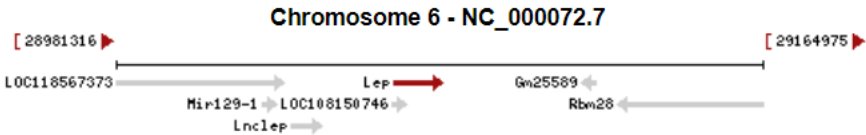
- Official Symbol Lep provided by MGI
- Official Full Name leptin provided by MGI
- Primary source MGI:MGI:104663
- See related Ensembl:ENSMUSG00000059201
- Gene type protein coding
- RefSeq status VALIDATED
- Organism Mus musculus
- Lineage Eukaryota; Metaz Muroidea; Murida
- Also known as ob; obese
- Expression Biased expression
- Orthologs human all
- NEW Try the new Gen
- Try the new Tran

Location: 6 A3.3; 6 12.3 cM

Exon count: 3

See Lep in Genome Data Viewer

Annotation release	Status	Assembly	Chr	Location
109	current	GRCm39 (GCF_000001635.27)	6	NC_000072.7 (29060220..29073875)
108.20200622	previous assembly	GRCm38.p6 (GCF_000001635.26)	6	NC_000072.6 (29060221..29073876)
Build 37.2	previous assembly	MGSCv37 (GCF_000001635.18)	6	NC_000072.5 (29010221..29023876)



Genomic regions, transcripts, and products

Go to reference details

Genomic Sequence: NC_000072.7 Chromosome 6 Reference GRCm39 C57BL/6J

Go to nucleotide: Graphics FASTA GenBank

Nucleotide

Nucleotide ▼

Advanced

FASTA ▼

Mus musculus strain C57BL/6J chromosome 6, GRCm39

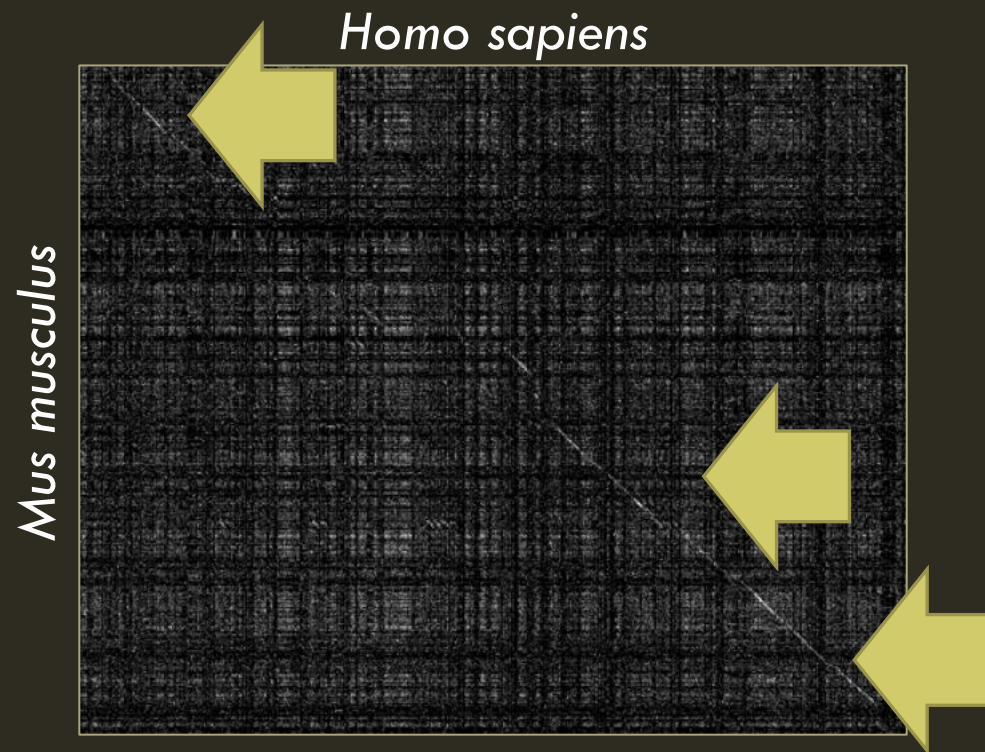
NCBI Reference Sequence: NC_000072.7

[GenBank](#) [Graphics](#)

```
>NC_000072.7:29060220-29073875 Mus musculus strain C57BL/6J chromosome 6, GRCm39
GAGGGATCCCTGCTCCAGCAGCTGCAAGGTAAGGCCCGGGCGCGCTACTTTCTCTCCACCAGTCTTTC
TAATAGCACCCCATCCAGCTCTGGAAATTAGAGAACTGAGGCAAGAAGGAGGTCATGTGGACAGCTTGG
TGTTGAATTCAGTAGTTTTGCAGCGAGGGACTCTGCAGACAGAGGGAAGGCTGTTGGGAGCTGAAAGGCA
TTAAATTCCTGGCTGGGCCGGCCTGTGCCCACAGGGCACCAACCTGTTCTTGCAAAGCCAGCCAGGGCTC
CAAGCCTGTCTTTGGTCTGCAGACTTGAGATGGTTAGGTTTCCACAGCGTTCAGATGGAAAAAGCAAGG
CTGGACGTGTAGTGAGGCAGCCACAGCTTTCGCCTTCTGAAGGCCAGATACAACCTTGCTTTGCTGTT
GAGAGTTTCATTTCCAGTGAGGGCTACCTAGTACTAAAACTACTGTTGAATACAGGTGGGTAGAAAAAC
TGTCAGGAAGAACCCAGTCCCCTCCCACCTTACCCCCAACTTGAGCGTGCACACACAAACACACACACA
CACAAACCAACAAACCATCATCAACAATAAAAAAGCTAGGGCTCTCTCCACATCTGTCCCAAAGCCATTC
AAATTCTCTTCTTTGGGGAACAGAACAGCAGATTGTGGAGAAGGCTTGTAACCTGCACAGTTTTTGTG
TAGGCATTTCTACCCAAGATGGTTAAATACACGTTGCCAGCCTGAGGTCCTGCATGCCCTCCTTCAGC
TTTCTTTCTTGCTGGGGTAGTATTACTGCATTGTGGGTATGGTGCCCTTACACTACTCTATGTCTTACT
GGGTGCATTAGATATGCCACGATGCCCTGTGTACATTGCTGCTTGGGAGAGGACAAAGCAAGCCAGTCAT
TTCTATTGGGAGCATGGGACTCAGGAGAGGCAGGCAGAGGGCAGGGATTTGGTGGCACCATCCTGGGAGC
ATAGGCTCTCTGAGCACTCTTTGCAAGGTGGGAATGGAATCCAGTAGTGAGAGTCATATTTTTCGTAGA
GTTTGTGATGGGCAGTGGTCAGCTGGGGCTGGGGGAGAGCATCTCTGACCTATTGCTGATTTGTACGGTG
CATGGAACCAAAC TTGAAAGGAGGCATGTCTAACAGTCACTCTCTGCTGACTATAAGGACCACACTGAT
CATGTATTTTCATTGAGGCACGAGAGAAGCACCTCTCTCTCTCCCCCTCAGACTTCTTTTTTTGGTTCTTT
CCTTTCTCAAGCACTCAAGATACTGGGCTTAGGGGAACATGCCCGTGTGGCTAAAGAAGCATCCTTCTC
ATATTTGGTAGATCTCAAACCCAAGTAACAAGTTCTTCTTGGTATCTGGGGGGAATATTTTCTCCTTCC
CCCCACCCCCCGTGTTCAGGAAAAAGCTCAAAGAGCCAAGGACAGAGCCAGGGATGCTGTAATAAATT
CTTGTCATCAGCCCCGGGCAGATGCCTCGCATGTCTGAGTGACCCATGTGGGATGCAGAGATGGCTGCC
GGAGACAATTCCCTTTACAGAGCACCAGTCATTGTTGCACCCCATACCCTGTATCCCTCTAGAGGCACAT
```

DOT-MATRIX

DOPASOWANIE: Leptyna *Homo sapiens* ↔ Leptyna *Mus musculus*



DOT-MATRIX

DOPASOWANIE: Leptyna *Homo sapiens* ↔ Leptyna *Takifugu rubripes*



DOPASOWANIE:

Leptyna *Homo sapiens* ↔ Leptyna *Mus musculus*

leptin Human | 40

```
GTAGGAATCGCAGCGCCAGCGGTTGCAAGGTAAGGCCCGGCGCGCTCCTTCCTCCTTCTCTGCTGGTCTTTCTTTGGCAGGCCAC  
GAGGGATCCCTGCTCCAGCAGCTGCAAGGTAAGGCCCGGGCGCGCTACTTTCTCCTCCACCAGTCTTCTAATAGCACCCCATC
```

leptin Mouse | 40

DOPASOWANIE:

Leptyna *Homo sapiens* ↔ Leptyna *Takifugu rubripes*

takifugu | 40

```
GGCACACGCTCGAACACCACTAAGGAAGCTTTCAAGAGCAACACCTGGACAAAAAGGTGCCGACATGGATCACATTCTGGCCCTT  
GTAGGAATCGCAGCGCCAGCGGTTGCAAGGCCCAAGAAGCCCATCCTGGGAAGGAAAATGCATTGGGGAACCTGTGCGGATTCT
```

human | 40

DOPASOWANIE SEKWENCJI

Algorytmy dopasowania sekwencji:

- macierze punktowe (dot-matrix)
- **programowanie dynamiczne (dynamic programming)**
- metody słowne (k – tuple methods)

PROGRAMOWANIE DYNAMICZNE

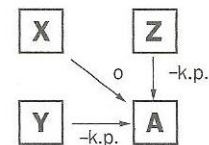
Analiza dwuwymiarowej macierzy z naniesionymi wartościami ocen dopasowania danych reszt
→ wyznaczenie ścieżki

Ścieżka reprezentująca najlepsze przyrównanie, ma najwyższą możliwą wartość punktacji – najwyższy score.

Identyczne reszty = **1**

Brak skojarzenia = **0**

Kara = **-1**



A jest maksymalną oceną pobraną z jednej z trzech komórek, do której dodano wartość dopasowania aktualnej pozycji

	A	T	T	G	C
A	1	0	0	0	0
G					
G					
C					



	A	T	T	G	C
A	1	0	0	0	0
G	0	1			
G					
C					



	A	T	T	G	C
A	1	0	0	0	0
G	0	1	1	2	
G					
C					



	A	T	T	G	C
A	1	0	0	0	0
G	0	1	1		
G					
C					



	A	T	T	G	C
A	1	0	0	0	0
G	0	1	1	2	2
G	0	1	1	3	3
C	0	1	1	3	4



	A	T	T	G	C
A	1	0	0	0	0
G	0	1	1	2	2
G	0	1	1	3	3
C	0	1	1	3	4

końcowe przyrównanie

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

DOPASOWANIE SEKWENCJI

Algorytmy dopasowania sekwencji:

- macierze punktowe (dot-matrix)
- programowanie dynamiczne (dynamic programming)
- **metody słowne (heurystyczne)**

WYKORZYSTANIE METODY HEURYSTYCZNEJ

- Sprawdzenie tylko części przyrównań, które byłyby analizowane zwykłymi metodami programowania dynamicznego – znaczne przyspieszenie!
- Zmniejszenie czułości i specyficzności
- Nie gwarantuje wygenerowania poprawnego przyrównania
- Użyteczna, gdy pełny algorytm jest z przyczyn technicznych zbyt kosztowny lub gdy jest nieznany.

LOKALNE PRZYRÓWNANIE SEKWENCJI NUKLEOTYDÓW

BLAST

- Basic Local Alignment Search Tool
- Kilka rodzajów BLAST:
 - BLASTN → przyrównanie nukleotydów
- Przeszukuje połączone światowe bazy danych sekwencji kwasów nukleinowych (lub białkowych) wykorzystując podaną sekwencję jako „zapytanie”
- Poszukuje wysoko punktowanych, ciągłych segmentów pokrewnych sekwencji

blast.ncbi.nlm.nih.gov

The screenshot displays the NCBI BLAST web interface. At the top, the NIH and NCBI logos are visible, along with the text "U.S. National Library of Medicine" and "National Center for Biotechnology Information". The main heading is "BLAST" with a subheading "blastn suite". Below this, there are tabs for "blastn", "blastp", "blastx", "tblastn", and "tblastx", with "blastn" currently selected. The "Enter Query Sequence" section includes a text input field for "Enter accession number(s), gi(s), or FASTA sequence(s)", a "Clear" button, and a "Query subrange" section with "From" and "To" input fields. A large yellow arrow points to the "Enter accession number(s)..." field. Below the input field, there is a section for "Or, upload file" with a "Wybierz plik" button and a "Nie wybrano pliku" status. A "Job Title" input field is also present. The "Choose Search Set" section includes a "Database" dropdown menu set to "Nucleotide collection (nr/nt)", an "Organism" input field, and several checkboxes for "Exclude" and "Limit to" options. At the bottom, there is an "Entrez Query" input field and links to "YouTube" and "Create custom database".

BLAST

Etapy działania:

- Tworzenie listy słów analizowanej sekwencji (X bp)
- Wyszukanie słów w sekwencyjnej bazie danych (identyfikacja sekwencji, z którymi zapytanie będzie skojarzone)
- Ocena skojarzonych słów (macierz substytucji) i wyznaczenie dla nich score
- Przyrównanie parami poprzez rozszerzanie słów w obu kierunkach, ocena (wartość graniczna).

BLAST

1. Query: MRD**PYN**KLIS

2. Scan every three residues to be used in searching BLAST word database.

3. Assuming one of the words finds matches in the database.

Query	PYN	PYN	PYN	PYN	...
Database	PYN	PFN	PFQ	PFE	...

4. Calculate sums of match scores based on BLOSUM62 matrix.

Query	PYN	PYN	PYN	PYN	...
Database	PYN	PFN	PFQ	PFE	...
Sum of score	20	16	10	10	...

5. Find the database sequence corresponding to the best word match and extend alignment in both directions.

Query	M R D	PYN	K L I S
Database	M H E	PYN	D V P W

←
→

extension to left extension to right

6. Determine high scored segment above threshold (22).

Query	M R D	PYN	K L I S
Database	M H E	PYN	D V P W
	5 0 2	20	-1 1 -3 -3

5 0 2 20 -1 1 -3 -3

HSP, total score 24

Figure 4.1: Illustration of the BLAST procedure using a hypothetical query sequence matching with a hypothetical database sequence. The alignment scoring is based on the BLOSUM62 matrix (see Chapter 3). The example of the word match is highlighted in the box.

GLOBALNE PRZYRÓWNANIE SEKWENCJI NUKLEOTYDÓW

Clustal Omega

Multiple Sequence Alignment (MSA)

[Job Dispatcher](#) [Help & Privacy](#) [Your Jobs](#) **Input form**

[Feedback](#)

Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments between three or more sequences. For the alignment of two sequences please instead use our pairwise sequence alignment tools.

Input sequence ⓘ

Sequence Type

☐ Protein ☒ DNA ☐ RNA

Paste your sequence here - or use the example sequence

Clustal Omega

Multiple Sequence Alignment (MSA)

[Job Dispatcher](#) [Help & Privacy](#) [Your Jobs](#) **Input form**

[Feedback](#)

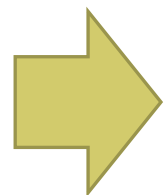
Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments between three or more sequences. For the alignment of two sequences please instead use our pairwise sequence alignment tools.

Input sequence ⓘ

Sequence Type

☐ Protein ☒ DNA ☐ RNA

Paste your sequence here - or use the example sequence



Wybierz plik leptin.txt

[Use the example](#)

[Clear sequence](#)

[More example inputs](#)

Parameters

OUTPUT FORMAT ⓘ

ClustalW with character counts

[More options](#) ▼

Tool output

[Download](#)

CLUSTAL 0(1.2.4) multiple sequence alignment

Takifugu_rubripes	-----	0
Homo_sapiens	GTAGGAATCGCAGCGCCAGCGTTGCAAGGTAAGGCCCC-GGCGCGCTCCTTCCTCCTTC	59
Mus_musculus	-GAGGGATCCCTGCTCCAGCAGCTGCAAGGTAAGGCCCGGGCGCGCTACTTTCTCC---	56
Takifugu_rubripes	-----	0
Homo_sapiens	TCTGCTGGTCTTTCTTGGCAGGCCACAGGGCCCCACACAACCTCTGGATCCCGGGGAAACT	119
Mus_musculus	TCCACCAGTCTTTCT-----AATAGCACCCCATCCAGCTCTGGAAATTAGAGAAACT	108
Takifugu_rubripes	-----	0
Homo_sapiens	GAGTCAGGAGGGATGCAGGGCGGATGGCTTAGTTCTGGACTATGATAGCTTTGTACCGAG	179
Mus_musculus	GAGGCAAGAAGGAGGTCATGTGGACAGCTTGGTGTGAATTCAG-TAGTTTTGCAGCGAG	167
Takifugu_rubripes	-----	0
Homo_sapiens	-----TTCTAGCCAGATAGAAGGTTACCGGGAGCTGGGGAGCGTTGGATTTGCTGCTGGG	234
Mus_musculus	GGACTCTGCAGACAGAGGGAAGGCTGTTGGGAGCTGAAAGGCATTAAATTCCTGGCTGGG	227

Clustal Omega

Multiple Sequence Alignment (MSA)

[Job Dispatcher](#) [Help & Privacy](#) [Your Jobs](#) [Input form](#)

Results for Job ID

clustalo-I20240525-160427-0487-30792591-p1m

Copy

[Tool Output](#)

[Alignments](#)

[Guide Tree](#)

[Phylogenetic Tree](#)

[Results Viewers](#)

[Result Files](#)

Percent Identity Matrix

clustalo-I20240525-154337-0259-35892461-p1m

Submission Details

clustalo-I20240525-154337-0259-35892461-p1m.submission

The alignment in FASTA format

clustalo-I20240525-154337-0259-35892461-p1m.fasta

#

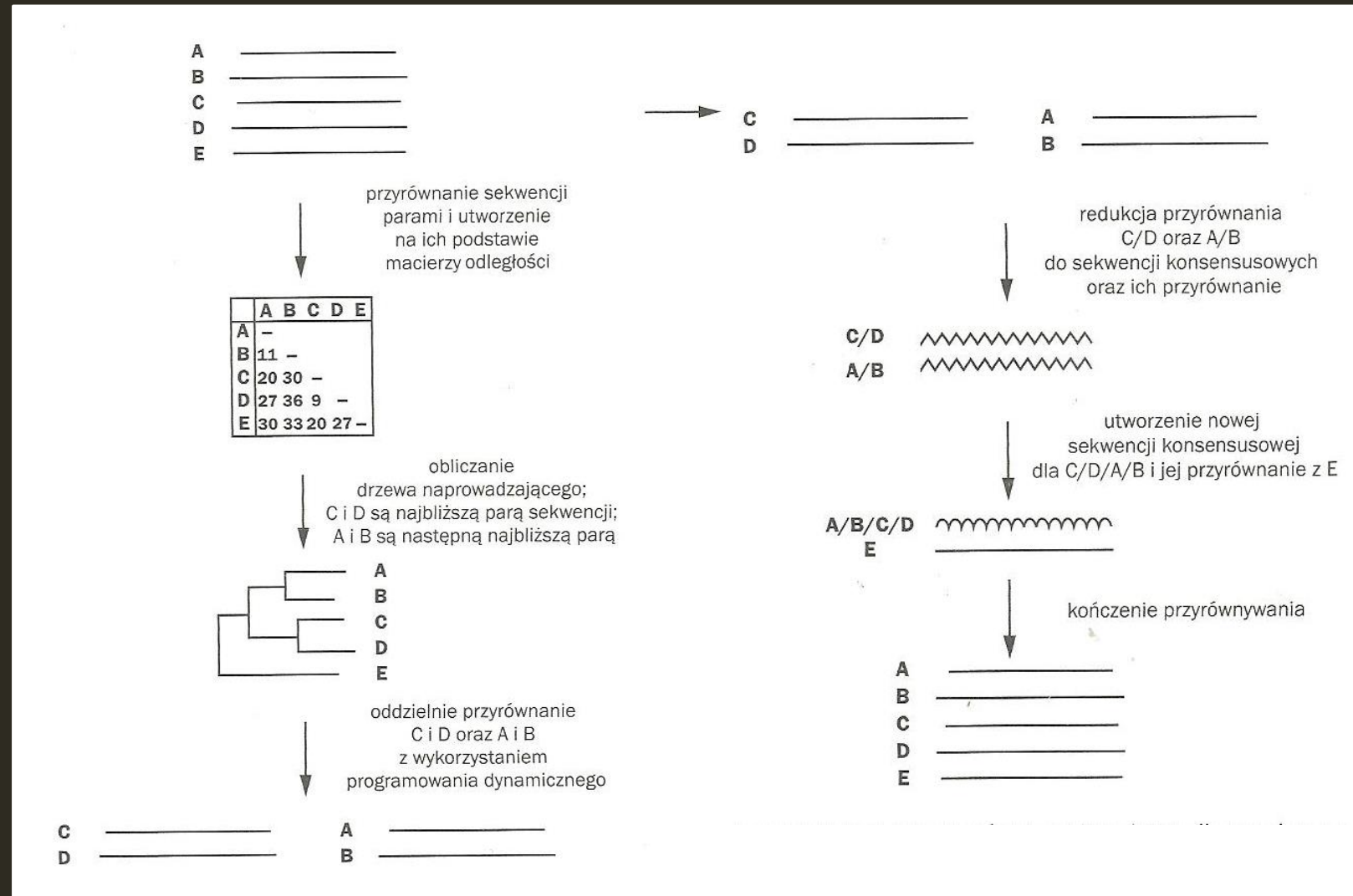
Percent Identity Matrix - created by Clustal2.1

1: Takifugu_rubripes	100.00	47.22	44.92
2: Homo_sapiens	47.22	100.00	63.65
3: Mus_musculus	44.92	63.65	100.00

Download

RODZINA PROGRAMÓW CLUSTAL

Metoda progresywna
→ etapowe składanie
przryównania na
podstawie
podobieństwa par
sekwencji



DOPASOWANIE SEKWENCJI

Algorytmy dopasowania sekwencji:

- macierze punktowe (dot-matrix)
- programowanie dynamiczne (dynamic programming)
- metody słowne (heurystyczne)

Może istnieć kilka „najlepszych dopasowań”

Miary określające jakość dopasowania

- wizualizacja podobieństwa (dot-matrix)
- miary zróżnicowania sekwencji
- miary podobieństwa sekwencji

MIARA ZRÓŻNICOWANIA SEKWENCJI NUKLEOTYDÓW

ODLEGŁOŚĆ HAMMINGA:

- Liczba pozycji różniących się literami
- Łańcuchy o równych długościach

g c t g a a c g

c t a t a a t c

= 6

ODLEGŁOŚĆ LEVENSHTEINA:

- Minimalna liczba operacji potrzebna do zmiany jednego łańcucha w drugi
- Dozwolone operacje: insercja, delecja, mutacja punktowa - zmiana litery
- Łańcuchy o różnych długościach

c - - t g a a c

c t a t a a t c

= 4

MIARA PODOBIEŃSTWA SEKWENCJI NUKLEOTYDÓW

PROSTA: np.

Różne litery $\rightarrow -1$

Jednakowe litery $\rightarrow +1$

g c t g a a c g

c t a t a a t c

-1-1-1-1+1+1-1-1

= -4

ZŁOŻONA np.:

Transwersja $(a,g) \leftrightarrow (t,c) \rightarrow 0$

Tranzycja $(a \leftrightarrow g \text{ lub } c \leftrightarrow t) \rightarrow 1$

Jednakowe $\rightarrow 2$

g c t g a a c g

c t a t a a t c

+0+1+0+0+2+2+1+0

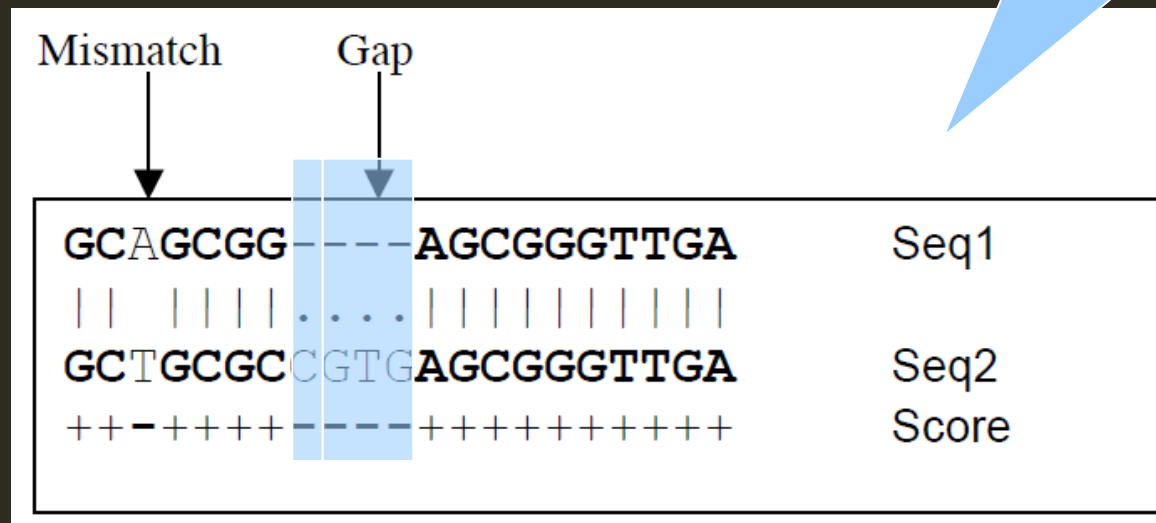
= 6

TABELE KODOWANIA ZDARZEŃ

TABELE KODOWANIA ZDARZEŃ

- Jednakowe litery
- Różne litery
- Rozpoczęcie przerwy
- Kontynuacja przerwy

Penalizacja afiniczna
→ łatwiej poszerzyć
przerwę niż
otworzyć nową
→ -12/-1 pkt



DOPASOWANIE SEKWENCJI AMINOKWASÓW



MIARY PODOBIEŃSTWA SEKWENCJI AMINOKWASÓW

KODOWANIE ZDARZEŃ → odzwierciedlenie właściwości fizykochemicznych

Sekwencje różnicują się na drodze ewolucji:

- uwzględnienie prawdopodobieństwa poszczególnych mutacji
- wśród wszystkich białek o znanej sekwencji i odległości ewolucyjnej

Konsekwencje metaboliczne mutacji:

- uwzględnienie czy zmiana aminokwasu powoduje zmianę konformacji = funkcji białka

Najczęściej wykorzystywane tabele

- PAM
- BLOSUM

PAM

Percent Accepted Mutation

- Opracowane modele ewolucyjne
- Akceptowana mutacja punktowa oznacza zastąpienie jednego aminokwasu w sekwencji innym. Aminokwasem, który jest „akceptowany” przez ewolucję to taki, dla którego mutacja powstała i się utrwała.

Tabele opracowano dla różnych odległości ewolucyjnych między sekwencjami, np.

- PAM1 – sekwencje różniące się 1% aminokwasów = blisko spokrewnionych
- PAM250 – sekwencje różniące się 80% aminokwasów = daleko spokrewnionych

TABELA PAM250

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

Wartości > 0 : częstość zamiany większa niż wynikałoby to z przypadku

Wartość $= 0$: częstość zamiany losowa

Wartości < 0 : częstość zamiany mniejsza niż wynikałoby to z przypadku

BLOSUM

Blocks Substitution Matrix

Tabele opracowano na podstawie danych z **bazy danych** BLOCKS wykorzystując prawdopodobieństwo mutacji w **zakonserwowanych** regionach

... przy różnych założeniach dotyczących interpretacji sekwencji o dużym podobieństwie

- BLOSUM62— sekwencje o ponad 62% aminokwasów identycznych są traktowane jako wspólna grupa ewolucyjna
- BLOSUM45— sekwencje o ponad 45% aminokwasów identycznych są traktowane jako wspólna grupa ewolucyjna

TABELA BLOSUM45

G	7																			
P	-2	9																		
D	-1	-1	7																	
E	-2	0	2	6																
N	0	-2	2	0	6															
H	-2	-2	0	0	1	10														
Q	-2	-1	0	2	0	1	6													
K	-2	-1	0	1	0	-1	1	5												
R	-2	-2	-1	0	0	0	1	3	7											
S	0	-1	0	0	1	-1	0	-1	-1	4										
T	-2	-1	-1	-1	0	-2	-1	-1	-1	2	5									
A	0	-1	-2	-1	-1	-2	-1	-1	-2	1	0	5								
M	-2	-2	-3	-2	-2	0	0	-1	-1	-2	-1	-1	6							
V	-3	-3	-3	-3	-3	-3	-3	-2	-2	-1	0	0	1	5						
I	-4	-2	-4	-3	-2	-3	-2	-3	-3	-2	-1	-1	2	3	5					
L	-3	-3	-3	-2	-3	-2	-2	-3	-2	-3	-1	-1	2	1	2	5				
F	-3	-3	-4	-3	-2	-2	-4	-3	-2	-2	-1	-2	0	0	0	1	8			
Y	-3	-3	-2	-2	-2	2	-1	-1	-1	-2	-1	-2	0	-1	0	0	3	8		
W	-2	-3	-4	-3	-4	-3	-2	-2	-2	-4	-3	-2	-2	-3	-2	-2	1	3	15	
C	-3	-4	-3	-3	-2	-3	-3	-3	-3	-1	-1	-1	-2	-1	-3	-2	-2	-3	-5	12
G	P	D	E	N	H	Q	K	R	S	T	A	M	V	I	L	F	Y	W	C	

INTERPRETACJA WARTOŚCI W TABELACH PAM I BLOSUM

I	-4	-2	-4	-3	-2	-3	-2	-3	-3	-2	-1	-1	2	3	5					
L	-3	-3	-3	-2	-3	-2	-2	-3	-2	-3	-1	-1	2	1	2	5				
F	-3	-3	-4	-3	-2	-2	-4	-3	-2	-2	-1	-2	0	0	0	1	8			
Y	-3	-3	-2	-2	-2	2	-1	-1	-1	-2	-1	-2	0	-1	0	0	3	8		
W	-2	-3	-4	-3	-4	-3	-2	-2	-2	-4	-3	-2	-2	-3	-2	-2	1	3	15	
C	-3	-4	-3	-3	-2	-3	-3	-3	-3	-1	-1	-1	-2	-1	-3	-2	-2	-3	-5	12
	G	P	D	E	N	H	Q	K	R	S	T	A	M	V	I	L	F	Y	W	C

1. Mutacja Izoleucyna ↔ Valina jest bardziej prawdopodobna = **+3**, niż Izoleucyna ↔ Cysteina = **-3**
2. Mutacja Izoleucyna ↔ Valina zachodzi $10^{0.3} = 1.995$ częściej niż średni poziom mutacji dla wszystkich aminokwasów
3. Mutacja Izoleucyna ↔ Cysteina zachodzi $10^{-0.3} = 0.501$ rzadziej niż średni poziom mutacji dla wszystkich aminokwasów

DOPASOWANIE SEKWENCJI AMINOKWASÓW

Beta-2-mikroglobulina *Homo sapiens* ↔ Beta-2-mikroglobulina *Mus musculus*

Gen zaliczany do tzw. Houskeeping genes → wysoki poziom zakonserwowania sekwencji

Baza danych UniProt:

Homo sapiens

<http://www.uniprot.org/uniprot/P61769>

Mus musculus

<http://www.uniprot.org/uniprot/Q91XJ8>

Function

Names & Taxonomy

Subcellular Location

Disease & Variants

PTM/Processing

Expression

Interaction

Structure

Family & Domains

Sequence

Similar Proteins

P61769 · B2MG_HUMAN

ProteinⁱBeta-2-microglobulin

Amino acids119 (go to sequence)

GeneⁱB2M

Protein existenceⁱEvidence at protein level

Statusⁱ UniProtKB reviewed (Swiss-Prot)

Annotation scoreⁱ

5/5

OrganismⁱHomo sapiens (Human)

Entry

View

Viewer

408

Feature viewer

Genomic coordinates

Publications

External links

History

BLAST

Download

Add

Functionⁱ

Component of the class I major histocompatibility complex (MHC) class II molecules. Exogenously applied M.tuberculosis protein levels do not change).

Download

Datasetⁱ

Entry

Format

Text

Text

FASTA (canonical)

FASTA (canonical & isoform)

JSON

XML

RDF/XML

GFF

Clustal Omega

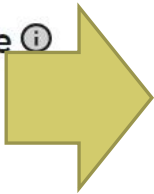
Multiple Sequence Alignment (MSA)

[Job Dispatcher](#) [Help & Privacy](#) [Your Jobs](#) **Input form**

[Feedback](#)

Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments between three or more sequences. For the alignment of two sequences please instead use our pairwise sequence alignment tools.

Input sequence ⓘ

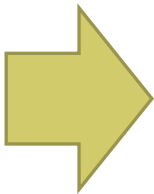


Sequence Type

☒ Protein ☐ DNA ☐ RNA

Paste your sequence here - or use the example sequence

```
>sp|P61769|B2MG_HUMAN Beta-2-microglobulin OS=Homo sapiens OX=9606 GN=B2M PE=1 SV=1
MSRSVALAVLALLSLSGLEAIQRTPKIQVYSRHPAENGKSNFLNCYVSGFHPSDIEVDLL
KNGERIEKVEHSDLSFSKDWSFYLLYYTEFTPTEKDEYACRVNHVTLSPKIVKWDRDM
>sp|P01887|B2MG_MOUSE Beta-2-microglobulin OS=Mus musculus OX=10090 GN=B2m PE=1 SV=2
MARSVTLVFLVLVSLTGLYAIQKTPQIQVYSRHPPENGKPNILNCYVTQFHPPHIEIQML
KNGKKIPKVEMSDMSFSKDWSFYILAHTFTPTETDTYACRVKHASMAEPKTVYWDRDM
```



Parameters

OUTPUT FORMAT ⓘ

ClustalW with character counts



[More options](#) ▼



Tool output

CLUSTAL 0(1.2.4) multiple sequence alignment

Download

```
sp|P61769|B2MG_HUMAN      MSRSVALAVLALLSLSGLEAIQRTPKIQVYSRHPAENGKSNFLNCYVSGFHPSDIEVDLL 60
sp|P01887|B2MG_MOUSE      MARSVTLVFLVLVSLTGLYAIQKTPQIQVYSRHPPENGKPNILNCYVTQFHPPHIEIQML 60
*:***:*..*.*:**:** ***:**:***** ***** *:*****: *** .**:::*

sp|P61769|B2MG_HUMAN      KNGERIEKVEHSDLSFSKDWSFYLLYYTEFTPTEKDEYACRVNHVTLSPKIVKWDRDM 119
sp|P01887|B2MG_MOUSE      KNGKKIPKVEMSDMSFSKDWSFYILAHTFTPTETDTYACRVKHASMAEPKTVYWDRDM 119
***::* *** **:*****:* :*****.* *****:*.:::*** * *****
```

KODY DOPASOWANIA:

- [*] identyczne
- [:] substytucja konserwatywna $\approx$ wymiana w obrębie grupy o podobnej strukturze i właściwościach
- [.] substytucja semikonserwatywna $\approx$ wymiana w obrębie grupy tylko o podobnej strukturze

DOPASOWANIE SEKWENCJI AMINOKWASÓW

Beta-2-mikroglobulina *Homo sapiens* ↔ Beta-2-mikroglobulina *Mus musculus* ↔ *Danio rerio* ↔ *Gallus gallus*

Gen zaliczany do tzw. Houskeeping genes → wysoki poziom zakonserwowania sekwencji

Download

sp Q04475 B2MG_DANRE	MRALITFALLCLL - YITVQGKVSTPKVHVYSHFPGEYGKPNLTICVSSFHPPDISIELL	59
sp P21611 B2MG_CHICK	MGKAAAVLVTLVALLGLAQADLTPKVQVYSRFPASAGTKNVLNCFAAGFHPPKISITLM	60
sp P61769 B2MG_HUMAN	MSRSVALAVLALLSLSGLEAIQRTPKIQVYSRHPAENGKSNFLNCYVSGFHPSDIEVDLL	60
sp P01887 B2MG_MOUSE	MARSVTLVFLVLVSLTGLYAIQKTPQIQVYSRHPPENGKPNILNCYVTQFHPPHIEIQML	60
	* :...: *: :	**::**:*.* . *. * * *::: *** .*.: ::
sp Q04475 B2MG_DANRE	KNGQVMSDTKQTDLAFEKGWQHFLTksVAFTPEKGDEYTCSVRH - -MKETKKFSWEPMN	110
sp P21611 B2MG_CHICK	KDGVPMEGAQYSDMSFNDDWTFQRLVHADFTPSSGSTYACKVEHETLKEPQVYKWDFEF	110
sp P61769 B2MG_HUMAN	KNGERIEKVEHSDLFSKDWsfYLLYYTEFTPTeKDEYACRVNHVtLSQPkiVKwDRDM	110
sp P01887 B2MG_MOUSE	KNGKKIPKvEMSDMSfSKDWsfYILAHTeFTPTeTDYAcRVkHASMAEPktvywDRDM	110
	: : .: :*:~*..* *	. *** . .*:~*~*.~* : : : ~*: ::

clustalo-I20240525-154337-0259-35892461-p1m.tree

Download

clustalo-I20240525-154337-0259-35892461-
p1m.phyloree

Download

clustalo-I20240525-154337-0259-35892461-p1m.pim

Download

clustalo-I20240525-154337-0259-35892461-
p1m.submission

Download

clustalo-I20240525-154337-0259-35892461-p1m.fa

Download

```
#
#
# Percent Identity Matrix - created by Clustal2.1
```

1:	sp	Q04475	B2MG_DANRE	100.00	37.07	43.10	40.52
2:	sp	P21611	B2MG_CHICK	37.07	100.00	43.70	42.86
3:	sp	P61769	B2MG_HUMAN	43.10	43.70	100.00	68.07
4:	sp	P01887	B2MG_MOUSE	40.52	42.86	68.07	100.00

Tool output

Download

CLUSTAL O(1.2.4) multiple sequence alignment

```
sp|Q04475|B2MG_DANRE      MRALITFALLCLL-YITVQGKVSTPKVHVYSHFPG EYGKPN TLICYVSSFHPPDISIELL 59
sp|P21611|B2MG_CHICK      MGKAAAVVLVTLVALLGLAQADLTPKVQVYSRFPASAGTKNVLNCF AAGFHPPKISITLM 60
sp|P61769|B2MG_HUMAN      MSRSVALAVLALLSLSGLEAIQRTPKIQVYSRHPAENGKSNFLN CYVSGFHPSDIEVDLL 60
sp|P01887|B2MG_MOUSE      MARSVTLVFLVLVSLTGLYAIQKTPQIQVYSRHPPENGKPNILN CYVTQFHPPHIEIQML 60
*      :...: *:      :      *****.* . *. * * *...: *** .*: ::

sp|Q04475|B2MG_DANRE      KNGQVMSDTKQTDLAFEKGWQFHLTKSVAFTPEKGD EYTCVRH--MKETKKFSWEPNM 116
sp|P21611|B2MG_CHICK      KDGVPMEGAQYSDMSFNDDWTFQRLVHADFTPSSG STYACKVEHETLKEPQVYKWDPEF 119
sp|P61769|B2MG_HUMAN      KNGERIEKVEHSDLSFSKDWSFYLLYYTEFTPTEKDEYACRVNHV TLSQPKIVKWDRDM 119
sp|P01887|B2MG_MOUSE      KNGKKIPKVEMSDMSFSKDWSFYILAHTFTPTETD TYACRVKHASMAEPKTVYWRDM 119
*: * : .: :*: :*...* * . *** . . *: * *. * : : : *: ::
```

Guide Tree

clustalo-I20240525-154337-0259-35892461-p1m.tree

Download

Phylogenetic Tree

clustalo-I20240525-154337-0259-35892461-p1m.phylotree

Download

Percent Identity Matrix

clustalo-I20240525-154337-0259-35892461-p1m.pim

Download

Submission Details

clustalo-I20240525-154337-0259-35892461-p1m.submission

Download

The alignment in FASTA format
converted by Secret

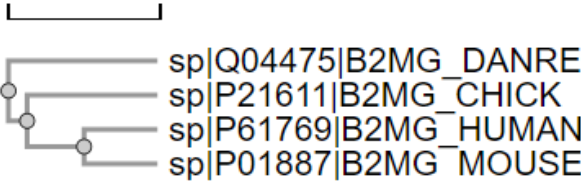
clustalo-I20240525-154337-0259-35892461-p1m.fasta

Download



Guide Tree

Phylogenetic Tree



BLASTP

BLAST[®] » blastp suite

[Home](#) [Recent Results](#) [Saved Strategies](#) [Help](#)

Standard Protein BLAST

blastn

blastp

blastx

tblastn

tblastx

BLASTP programs search protein databases using a protein query. [more...](#)

[Reset page](#)

[Bookmark](#)

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [?](#) [Clear](#)

Query subrange [?](#)

Or, upload file

Wybierz plik

Nie wybrano pliku [?](#)

Job Title

Enter a descriptive title for your BLAST search [?](#)

☐ Align two or more sequences [?](#)

New columns added to the Description Table

Click 'Select Columns' or 'Manage Columns'.



Choose Search Set

Database

Non-redundant protein sequences (nr) [?](#)

Organism

Optional

Enter organism name or id--completions will be suggested

☐ exclude

[Add organism](#)

[Restore default search parameters](#)

General Parameters

Max target sequences

100 ▾

Select the maximum number of aligned sequences to display ?

Short queries

☒ Automatically adjust parameters for short input sequences ?

Expect threshold

0.05 ?

Word size

6 ▾ ?

Max matches in a query range

0 ?

Scoring Parameters

Matrix

BLOSUM62 ▾ ?

Gap Costs

Existence: 11 Extension: 1 ▾ ?

Compositional adjustments

Conditional compositional score matrix adjustment ▾ ?

Filters and Masking

Filter

☐ Low complexity regions ?

Mask

☐ Mask for lookup table only ?

☐ Mask lower case letters ?

BLAST

Search **database nr** using **Blastp (protein-protein BLAST)**

☐ Show results in a new window