

# Sequence Alignment

# Why do we want to compare sequences?

- **Evolutionary relationships**

- Phylogenetic trees can be constructed based on comparison of the sequences of a molecule (example: 16S rRNA) taken from different species
- Residues conserved during evolution play an important role

- **Prediction of protein structure and function**

- Proteins which are very similar in sequence generally have similar 3D structure and function as well
- By searching a sequence of unknown structure against a database of known proteins the structure and/or function can in many cases be predicted

# Using sequence alignment to search databases

- The most common usage of pairwise sequence alignment is searching databases for related sequences
- Although the alignments themselves may be unreliable the alignment scores gives a lot of information about which sequences are related and which are not
- Having a set of related sequences is a lot more informative than just one sequence - even if nothing is known about the related sequences

# What is sequence alignment?

- Sequence alignment is a way of arranging the sequences of DNA, RNA or protein to identify regions of similarity that may be a consequence of functional, structural or evolutionary relationships between the sequences.
- The procedure of comparing two (pair-wise alignment) or more multiple sequences is to search for a series of individual characters or patterns that are in the same order in the sequences.
- There are two types of alignment: local and global.

# Global alignment vs Local alignment

- Global alignment is attempting to match as much of the sequence as possible.

The tool for Global alignment is based on Needleman-Wunsch algorithm.

**-Entire sequence of each protein or DNA sequence**

- Local alignment is to try to find the regions with highest density of matches. The tool for local alignment is based on Smith-Waterman.

**-Focuses on the region of greatest similarity between two sequences**

- Both algorithms are derivatives from the basic dynamic programming algorithm.

```

L G P S S K Q T G K G S - S R I W D N
L N - I T K S A G K G A I M R L G D A
- - - - - T G K G - - - - -

```

Global alignment

```

- - - - - A G K G - - - - -

```

Local alignment

# Local or global alignment

- Generally local alignment is used for performing database searches
  - For most cases you would be interested in knowing if any parts of your sequences look like something else
  - The protein sequence databases have not been split into domains
- It is not always the optimal thing to do but ...
  - In the case where the complete sequence should match the local alignment score will be almost identical to the global one
  - If you really want a global alignment you can make it afterwards

# Why do sequence alignment?

- Sequence alignment is useful for discovering structural, functional and evolutionary information in biological sequences.
- Sequences that are very much alike may have similar secondary and 3D structure, similar function and likely a common ancestral sequence. It is extremely unlikely that such sequences obtained similarity by chance.
  - For DNA molecules with  $n$  nucleotides such probability is very low  $P = 4^{-n}$ .
  - For proteins with  $n$  nucleotides, the probability even much lower  $P = 20^{-n}$ .
- Sequence alignment makes the following tasks easy: 1. annotation of new sequences; 2. modelling of protein structures; 3. design and analysis of gene expression experiments

# Alignement evaluation

## What is a good alignment ?

- We need a way to evaluate the biological meaning of a given alignment
- Intuitively we "know" that the following alignment:

is better than:

```
CGAGGCACAACGTCA
||| ||| |||||
CGATGCAAGACGTCA
```

```
ATTGGACAGCAATCAGG
|      ||  |      |
ACGATGCAAGACGTCAG
```

- We can express this notion more rigorously, by using a [scoring system](#).

# Scoring system

## Simple alignment scores

- A simple way (but not the best) to score an alignment is to count 1 for each match and 0 for each mismatch.

CGAGGCACAACGTCA
CGATGCAAGACGTCA

⇒ Score: 12

ATTGGACAGCAATCAGG
ACGATGCAAGACGTCAG

⇒ Score: 5

# Terminologies of sequence comparison

- **Sequence identity** -- exactly the same Amino Acid or Nucleotide in the same position.
- **Sequence similarity** -- Substitutions with similar chemical properties.
- **Sequence homology** -- general term that indicates evolutionary relatedness among sequences; we usually measure of percentage identity of sequence homology
- **Pairwise alignment** -- used to find the best-matching piecewise (local) or global alignments of two query sequences. Pairwise alignments can only be used between two sequences at a time.
- **Multiple sequence alignment** -- try to align all of the sequences in a given query set.

# Gaps

## Insertions or deletions

- Proteins often contain regions where residues have been **inserted** or **deleted** during evolution
- There are constraints on where these insertions and deletions can happen (between structural or functional elements like: alpha helices, active site, etc.)

## Gaps in alignments

```
GCATGCATGCAACTGCAT
| | | | | | | |
GCATGCATGGGCAACTGCAT
```

can be improved by inserting a **gap**

```
GCATGCATG--CAACTGCAT
| | | | | | | | | | | |
GCATGCATGGGCAACTGCAT
```

- Pairwise alignment is useful as a way to identify mutations that have occurred during evolution and have caused divergence of the sequences of two proteins.
- The most common mutations are **substitution, insertions** and **deletions**.
- Insertions and deletions occur when residues are added or removed and are typically represented by dashes that are added to one or other sequence.
- Insertions and deletions are referred to as gaps in the alignment.

# Importance of Similarity

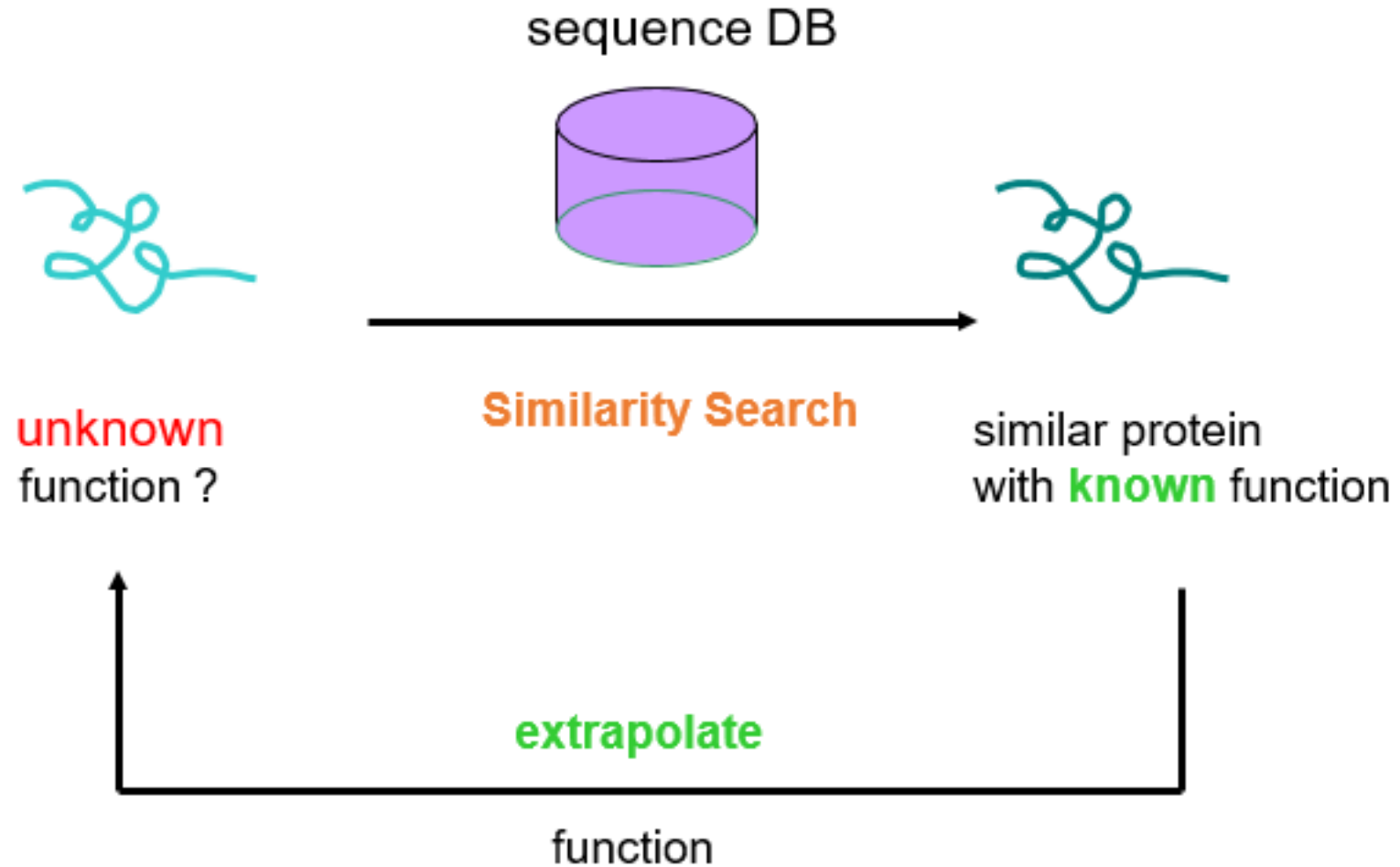
For sequences which are more than 100 amino acids (or nucleotides) long: They can be considered as homologues if 20% of the aa are identical (70% of nucleotide for DNA). Lower than this is called the **twilight zone**.

**Twilight zone** = protein sequence similarity between ~0-20% identity: is **not** statistically **significant**, i.e. could have arisen by chance.

## Beware:

- E-value (*Expectation value*) : which tells you how likely it is that the similarity between your sequence and a database sequence is due to chance (the lower the better)
- Length of the segments similar between the two sequences
- The number of insertions/deletions

# Importance of Similarity Database Search





# A multiple sequence alignment of globins

```
HBB_HUMAN  -----VHLTPEEKSAVTALWGKVN--VDEVGGEALGRLLVVYPWTQRFFESFGDLST
HBB_HORSE  -----VQLSGEKA AVLALWDKVN--EEEVGGEALGRLLVVYPWTQRFFDSFGDLSN
HBA_HUMAN  -----VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLS-
HBA_HORSE  -----VLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHF-DLS-
MYG_PHYCA  -----VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKT
GLB5_PETMA PIVDTGSVAPLSAAEKTKIRSAWAPVYSTYETSGVDILVKFFTSTPAAQEFFPKFKGLTT
LGB2_LUPLU -----GALTESQAALVKSWEEFNANIPKHTHRFFILVLEIAPAAKDLFSFLKGTSE
          *:  :  :  *  .                :  .:  *  :  *  :  .
```

```
HBB_HUMAN  PDAVMGNPKVKAHGKKVLGAFSDGLAHLDN-----LKGTFATLSELHCDKLHVDPENFRL
HBB_HORSE  PGAVMGNPKVKAHGKKVLHSFGEGVHHLDN-----LKGTFAALSELHCDKLHVDPENFRL
HBA_HUMAN  ----HGSAQVKGHGKKVADALTNVAHVDD-----MPNALSALSDLHAHKLRVDPVNFKL
HBA_HORSE  ----HGSAQVKAHGKKVGDALTLAVGHLDL-----LPGALSNLSDLHAHKLRVDPVNFKL
MYG_PHYCA  EAEMKASEDLKKHGVTVLTALGAILKKKGH-----HEAELKPLAQSHATKHKIPIKYLEF
GLB5_PETMA ADQLKKSADVRWHAERIINAVNDAVASMDDT--EKMSMKLRDLSGKHAKSFQVDPQYFKV
LGB2_LUPLU VP--QNNPELQAHAGKVFKLVEAAIQLQVTGVVVTDATLKNLGSVHVSKGVAD-AHFPV
          .  .:: *  :  .                :  *  *  .                :  .
```

# Why multiple alignment is better

- More sequences contain more information
- Multiple sequence alignment allows us to compare all related proteins simultaneously
- It allows us to identify features that are conserved among the sequences
- Using a multiple sequence alignment (a profile) one can find more related sequences than by simple pairwise comparison

- Align

NP\_005359.1

NP\_999401

by using BLAST Alignment tool

# Align two (or more) sequences by using BLAST

The screenshot shows the NCBI BLAST website. At the top, there is a navigation bar with the NIH logo, "U.S. National Library of Medicine", and "NCBI National Center for Biotechnology Information". The URL in the browser is "blast.ncbi.nlm.nih.gov/Blast.cgi". Below the navigation bar, there is a search bar with the text "BLAST finds regions of similarity between biological sequences. more...". To the right of the search bar, there is a "Sign in to NCBI" link. Below the search bar, there is a "New" banner for "SmartBLAST". The main content area is divided into several sections: "BLAST Assembled Genomes" with a search box and a "GO" button, and a list of organisms; "Basic BLAST" with a "Choose a BLAST program to run." section listing various BLAST programs; and "Specialized BLAST" with a "Choose a type of specialized search" section listing various specialized search options. On the right side, there are three sidebar sections: "Your Recent Results", "News", and "Tip of the Day". The footer contains copyright information and links to "Disclaimer", "Privacy", "Accessibility", "Contact", and "Send feedback".

blast.ncbi.nlm.nih.gov/Blast.cgi

NIH U.S. National Library of Medicine NCBI National Center for Biotechnology Information Sign in to NCBI

BLAST® Home Recent Results Saved Strategies Help

BLAST finds regions of similarity between biological sequences. [more...](#)

**New** Try [SmartBLAST](#) for an improved protein-protein search

### BLAST Assembled Genomes

Find Genomic BLAST pages:

Enter organism name or id—completions will be suggested  **GO**

- [Human](#)
- [Rabbit](#)
- [Zebrafish](#)
- [Mouse](#)
- [Chimp](#)
- [Clawed frog](#)
- [Rat](#)
- [Guinea pig](#)
- [Arabidopsis](#)
- [Cow](#)
- [Fruit fly](#)
- [Rice](#)
- [Pig](#)
- [Honey bee](#)
- [Yeast](#)
- [Dog](#)
- [Chicken](#)
- [Microbes](#)

### Basic BLAST

Choose a BLAST program to run.

<a href="#">nucleotide blast</a>	Search a <b>nucleotide</b> database using a <b>nucleotide</b> query <i>Algorithms:</i> blastn, megablast, discontinuous megablast
<a href="#">protein blast</a>	Search <b>protein</b> database using a <b>protein</b> query <i>Algorithms:</i> blastp, psi-blast, phi-blast, delta-blast
<a href="#">blastx</a>	Search <b>protein</b> database using a <b>translated nucleotide</b> query
<a href="#">tblastn</a>	Search <b>translated nucleotide</b> database using a <b>protein</b> query
<a href="#">tblastx</a>	Search <b>translated nucleotide</b> database using a <b>translated nucleotide</b> query

### Specialized BLAST

Choose a type of specialized search (or database name in parentheses.)

- Get faster protein results with a graphical view using [SmartBLAST](#)
- Make specific primers with [Primer-BLAST](#)
- Cluster multiple sequences together with their database neighbors using [MOLE-BLAST](#)
- Find [conserved domains](#) in your sequence (cds)
- Find sequences with similar [conserved domain architecture](#) (cdart)
- Search sequences that have [gene expression profiles](#) (GEO)
- Search [immunoglobulins and T cell receptor sequences](#) (IgBLAST)
- [Screen sequence for vector contamination](#) (vecscreen)
- [Align two \(or more\) sequences using BLAST](#) (bl2seq)
- Search [protein](#) or [nucleotide](#) targets in PubChem BioAssay
- Search [SRA by experiment](#)
- Constraint Based Protein [Multiple Alignment Tool](#)
- Needleman-Wunsch [Global Sequence Alignment Tool](#)
- Search [RefSeqGene](#)
- Search [trace archives](#)
- Search bacterial and fungal rRNA sequences with [Targeted Loci BLAST](#)

### Your Recent Results **New**

[All Recent results...](#)

### News

[Searching Whole Genome Shotgun sequences](#)

It is now much easier to search WGS (Whole Genome Shotgun) with stand-alone BLAST on your own computer.

Wed, 20 Jan 2016 10:00:00 EST

[More BLAST news...](#)

### Tip of the Day

[Use Genomic BLAST to see the genomic context](#)

If you are interested in the evolution of a particular gene or gene family it is often interesting to examine the intro-exon structure even across species.

[More tips...](#)

BLAST is a registered trademark of the National Library of Medicine.

Copyright | Disclaimer | Privacy | Accessibility | Contact | Send feedback

NCBI | NLM | NIH | DHHS

# Align nucleotide sequence/DNA-blastn

blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE\_TYPE=BlastSearch&BLAST\_SPEC=blast2seq&LINK\_LOC=blasttab&LAST\_PAGE=blastf

NIH U.S. National Library of Medicine NCBI National Center for Biotechnology Information Sign in to NCBI

BLAST® >> blastn suite Home Recent Results Saved Strategies Help

### Align Sequences Nucleotide BLAST

blastn blastp blastx tblastn tblastx

BLASTN programs search nucleotide subjects using a nucleotide query. [more...](#) [Reset page](#) [Bookmark](#)

#### Enter Query Sequence

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

NM\_000184.2  **1. NM\_000184.2**

Or, upload file  Dosya seçilmedi [?](#)

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

Align two or more sequences [?](#)

#### Enter Subject Sequence

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

NM\_008220.5  **2. NM\_008220.5**

Or, upload file  Dosya seçilmedi [?](#)

#### Program Selection

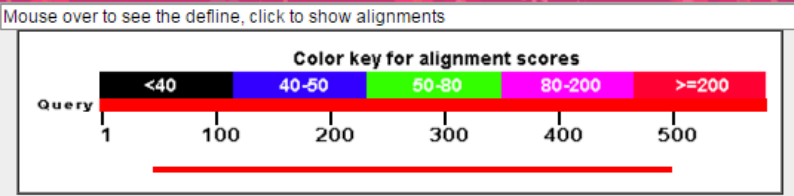
Optimize for

- Highly similar sequences (megablast)
- More dissimilar sequences (discontiguous megablast)
- Somewhat similar sequences (blastn)

Choose a BLAST algorithm [?](#)

**BLAST** Search nucleotide sequence using Blastn (Optimize for somewhat similar sequences)  
 Show results in a new window

[+ Algorithm parameters](#) **Note:** Parameter values that differ from the default are highlighted in yellow and marked with ♦ sign



[+ Dot Matrix View](#)

[- Descriptions](#)

Sequences producing significant alignments:

Select: [All](#) [None](#) Selected:0

[↑ Alignments](#) [Download](#) [GenBank](#) [Graphics](#)

Description	Max score	Total score	Query cover	E value	Ident	Accession
<input type="checkbox"/> <a href="#">Mus musculus hemoglobin, beta adult t chain (Hbb-bt), mRNA</a>	351	351	77%	7e-101	77%	<a href="#">NM_008220.5</a>

[- Alignments](#)

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

[Next](#) [Previous](#) [Descriptions](#)

Mus musculus hemoglobin, beta adult t chain (Hbb-bt), mRNA  
 Sequence ID: [reflNM\\_008220.5](#) Length: 646 Number of Matches: 1

Range 1: 47 to 500 [GenBank](#) [Graphics](#) [Next Match](#) [Previous Match](#)

Score	Expect	Identities	Gaps	Strand
351 bits(388)	7e-101	350/454(77%)	0/454(0%)	Plus/Plus
Query 46	CAGACGCCATGGGTCAATTCACAGAGGAGGACAAGGCTACTACACAAGCCGTGGGGCA	105		
Sbjct 47	CAGACATCATGGTGCACCTGACTGATGCTGAGAAAGGCTGCTCTCTGGCCGTGGGGAA	106		
Query 106	AGGTGAATGTGGAAGATGCTGGAGGAGAAAACCTGGGAAGGCTCCTGGTTGTCTACCCAT	165		
Sbjct 107	AGGTGAACGCCGATGAAGTTGGTGGTGGGCCCTGGCAGGCTGCTGGTTGTCTACCCCT	166		
Query 166	GGACCCAGAGGTTCTTTGACAGCTTTGGCAACCTGTCCTCTGCCTCTGCCATCATGGGCA	225		
Sbjct 167	GGACCCAGCGGTACTTTGATAGCTTTGGAGACCTATCCTCTGCCTCTGCTATCATGGGTA	226		
Query 226	ACCCCAAAGTCAAGGCCACATGGCAAGAAAGGTGCTGACTTCTTGGGAGATGCCATAAAGC	285		
Sbjct 227	ATGCCAAAGTGAAGGCCATGGCAAGAAAGTGATAACTGCCTTTAACGATGGCCTGAATC	286		
Query 286	ACCTGGATGATCTCAAGGGCACCTTTGCCAGCTGAGTGAACGCACTGTGACAAGCTGC	345		
Sbjct 287	ACTTGGACAGCCTCAAGGGCACCTTTGCCAGCCTCAGTGAGCTCCACTGTGACAAGCTGC	346		
Query 346	ATGTGGATCCTGAGAACTTCAAGCTCCTGGGAAATGTGCTGGTGACCGTTTGGCAATCC	405		
Sbjct 347	ATGTGGATCCTGAGAACTTCAAGCTCCTGGGCAATATGATCGTGATTGTGCTGGGCCACC	406		
Query 406	ATTCGGCAAAGAATTCAACCCTGAGGTGCAGGCTTCTGGCAGAAGATGGTGACTGGAG	465		
Sbjct 407	ACCTGGCAAGGATTTCAACCCTGAGGTGCAGGCTTCTGGCAGAAGATGGTGACTGGAG	466		
Query 466	TGGCCAGTGCCCTGTCTCCAGATACCACTGAGC	499		
Sbjct 467	TGGCTGCTGCCCTGGCTCACAAGTACCACTAAGC	500		

**Related Information**

- [Gene-associated gene details](#)
- [GEO Profiles](#)-microarray expression data
- [Map Viewer](#)-aligned genomic context

# Align a.a sequence/protein-blastp

blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE\_TYPE=BlastSearch&BLAST\_SPEC=blast2seq&LINK\_LOC=blasttab&LAST\_PAGE=bl...

NIH U.S. National Library of Medicine NCBI National Center for Biotechnology Information Sign in to NCBI

BLAST® >> blastp suite Home Recent Results Saved Strategies Help

### Align Sequences Protein BLAST

blastn blastp **blastx** tblastn tblastx

BLASTP programs search protein subjects using a protein query. [more...](#) [Reset page](#) [Bookmark](#)

#### Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [Clear](#) Query subrange [v](#)

AAF87098.1 From  To  **1. AAF87098.1**

Or, upload file  Dosya seçilmedi [v](#)

Job Title  Enter a descriptive title for your BLAST search [v](#)

Align two or more sequences [v](#)

#### Enter Subject Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [Clear](#) Subject subrange [v](#)

BAB03272.1 From  To  **2. BAB03272.1**

Or, upload file  Dosya seçilmedi [v](#)

#### Program Selection

Algorithm  blastp (protein-protein BLAST) Choose a BLAST algorithm [v](#)

Search **protein sequence** using **Blastp (protein-protein BLAST)**  Show results in a new window

[+ Algorithm parameters](#) **Note: Parameter values that differ from the default are highlighted in yellow and marked with ♦ sign**

Download GenPept Graphics

Next Previous Descriptions

liver transporter 1st-1 [Mus musculus]

Sequence ID: [dbj|BAB03272.1](#)

[See 4 more title\(s\)](#)

Range 1: 1 to 689 GenPept Graphics

Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps
1166 bits(3016)	0.0	Compositional matrix adjust.	560/691(81%)	613/691(88%)	6/691(0%)
Query 1	MDHTQQSRKAAEAQPSRQKTRFC	MD TQ KAA QP RQ++TR CDGF++FLAALSFSYICKALGGVVMKSSITQIERRFD	60		
Sbjct 1	MDQTQHPKAA--QPLRQEKTRH	MDQTQHPKAA--QPLRQEKTRHCDGFRIFLAALSFSYICKALGGVIMKSSITQIERRFD	58		
Query 61	IPSSISGLIDGGFEIGNLLVIVFV	IPSSISGLIDGGFEIGNLLVIVFVSYFGSKLHRPKLIGGCFIMGIGSILTALPHFFMGY	120		
Sbjct 59	IPSSISGLIDGGFEIGNLLVIVFV	IPSSISGLIDGGFEIGNLLVIVFVSYFGSKLHRPKLIGTGCFIMGIGSILTALPHFFMGY	118		
Query 121	YKYAKENDIGSLGNSTLTFCINQ	Y+YA ENDI SL NSTLTC +NQ TS TG SPEI+EKGCEKG S+ WIYVLMGNMMLRGI	180		
Sbjct 119	YRYATENDISSLHNSTLTCLVNQ	YRYATENDISSLHNSTLTCLVNQTTSLTGTSPREIEMKGEKGSNSYTWIYVLMGNMMLRGI	178		
Query 181	GETPIVPLGISYLDFAKEGHTSM	GETPIVPLG+SY+DDFAKEG++SM+LGTLHTIAMIGPILGFMSSVFAKIYVDVGYVDLN	240		
Sbjct 179	GETPIVPLGVSYIDFAKEGNSSMY	GETPIVPLGVSYIDFAKEGNSSMYLGTLHTIAMIGPILGFMSSVFAKLYVDVGYVDLR	238		
Query 241	SVRITPNARWVGAWLSFIVNGLL	SVRITP DARWVGAWL FIVNGLLCI SIPFFFLPKIPKRSQ+ERKNS SLH KTDE	300		
Sbjct 239	SVRITPQDARWVGAWLGFIVNGL	SVRITPQDARWVGAWLGFIVNGLLCIICISIPFFFLPKIPKRSQKERKNSASLHVLTDE	298		
Query 301	EKKHMTNLTKQEEQDPSNMTGFL	+K +TN T QE+Q P+N+TGFL SLRSILTNE YVIFLILTLQ+S FIGSFTYLFKFI	360		
Sbjct 299	DKNPVTNPTTQEQAPANLTGFL	DKNPVTNPTTQEQAPANLTGFLWSLRSILTNEQYVIFLILTLQISSFIGSFTYLFKFI	358		
Query 361	EQQFGRTASQANFLLGIITIPTMA	EQQFG+TASQANFLLG+ITIPTMA+ MFLGGY++K+ KLT +GI KFVFFI+++AY F	420		
Sbjct 359	EQQFGRTASQANFLLGVITIPTMA	EQQFGRTASQANFLLGVITIPTMASGMFLGGYLIKRLKLTLLGITKFVFFITTTMAYVYFL	418		
Query 421	LYFPLLCENKPFAGLTLYDGMNP	YF L+CENK FAGLTLYDGMNPVDSHIDVPLSYCNSDCDKNQWEPICGENGVTYIS	480		
Sbjct 419	SYFLLICENKAFAGLTLYDGMNP	SYFLLICENKAFAGLTLYDGMNPVDSHIDVPLSYCNSDCIDKNQWEPVCGENGVTYIS	478		
Query 481	PCLAGCKSFRGDKKPNTEFYDCS	PCLAGCKSFRGDKK N EFYDCSC+S S GN+SA LGCEPR CKCT YFYFI QV	536		
Sbjct 479	PCLAGCKSFRGDKKLMNIEFYDC	PCLAGCKSFRGDKKLMNIEFYDCSCVSGSGFQKGNHSARLGECPDCKCTKYFYFITFQV	538		
Query 537	TVSFFTAMGSPSLILILMKSQPEL	+SFFTA+GS SL+LIL++SVQPELKSL MGFHSL++R LGGILAP+YYGA IDRTC+KW	596		
Sbjct 539	IISFFTALGSTSLMLILIRSVQPE	IISFFTALGSTSLMLILIRSVQPELKSLMGFHSLLVVRTLGGILAPVYYGALIDRTC+KW	598		
Query 597	SVTSCGKRGACRLYNSRLFGFSY	SVTSCG RGACRLYNSRLFG Y+GL++ALKTP L LYV LIY KRK KRNDNK LENG	656		
Sbjct 599	SVTSCGARGACRLYNSRLFGMIY	SVTSCGARGACRLYNSRLFGMIYVGLSIALKTPILLLYVALIYVMKRKMKRNDNKILENG	658		
Query 657	RQFTDEGNPDSVKNNGYCVPYDE	R+FTDEGNP+ VN NGY CVP DE+++ETPL	687		
Sbjct 659	RKFTDEGNPEPVNNNGYCVPSDE	RKFTDEGNPEPVNNNGYCVPSDEKNSSETPL	689		

Related Information

- [Gene](#) - associated gene details
- [Map Viewer](#) - aligned genomic context
- [Identical Proteins](#) - Identical proteins to NP\_065241.1

% 81 Identical  
% 88 Similar

+ shows similar substitutions

- Performing multiple sequence alignments is useful in identifying regions of similarity.
- These may represent functional, structural, or evolutionary relationships between species.

‡ What do you think? Discuss in small groups.

- Let's say we want to understand the functional similarities/differences between two enzymes. Would it be more informative to align DNA sequences or amino acid sequences? Why?

# Clustal Omega

(<http://www.ebi.ac.uk/Tools/msa/clustalo/>)

- Is a multiple DNA, RNA, or protein sequences can be aligned using this tool to identify areas of similarities. These similarities may be associated with specific features that are highly conserved which in turn can aid in classifying sequences.

# Clustal Omega

Tools > Multiple Sequence Alignment > Clustal Omega

## Multiple Sequence Alignment

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](#).

**Important note:** This tool can align up to 4000 sequences or a maximum file size of 4 MB.

### STEP 1 - Enter your input sequences

Enter or paste a set of

PROTEIN

sequences in any supported format:

[Empty text area for pasting sequences]

Or, upload a file:  No file chosen

[Use a example sequence](#) | [Clear sequence](#) | [See more example inputs](#)

No file chosen

### STEP 2 - Set your parameters

OUTPUT FORMAT

ClustalW with character counts

Alignment of 2 human zinc finger proteins by ClustalW

-ABB24881.1

-ABB24882.1

# Clustal Omega

[Input form](#) | [Web services](#) | [Help & Documentation](#) | [Bioinformatics Tools FAQ](#)

Tools > Multiple Sequence Alignment > Clustal Omega

## Results for job clustalo-I20190326-074836-0915-36895123-p1m

[Alignments](#) | [Result Summary](#) | [Phylogenetic Tree](#) | [Submission Details](#)

[Download Alignment File](#) | [Show Colors](#) | [View result with Jalview](#) | [Send to Simple Phylogeny](#) | [Send to MView](#)

CLUSTAL O(1.2.4) multiple sequence alignment

```
AAB24881.1 -----YECNQCGKAFAQHSSLKCHYRTHIGEKPYECNQCGKAFSK 40
AAB24882.1 TYHMCQFHCRYVNNHSGEKLYECNERSKAFSCPSHLQCHKRRQIGEKTHEHNQCGKAFPT 60
                ****: .***: * *:* * :**** :* ***** .

AAB24881.1 HSHLQCHKRTHHTGKPYECNQCGKAFSQHGLLQRHKRTHHTGKPYMNVINMVKPLHNS 98
AAB24882.1 PSHLQYHERTHHTGKPYECHQCGQAFKKCSLLQRHKRTHHTGKPYECNQCGKA-FAQ- 116
                *** *:*****:***:*. : .*****
                : :
```

*PLEASE NOTE: Showing colors on large alignments is slow.*

10 20 30 40 50 60 70 80 90 100 110

AAB24881.1/1-98 .....YECNQCGKAF AQHSS LKCHYRTH IGEKPYECNQCGKAF SKHSHLQCHKRTH TGEKPYECNQCGKAF SQHG LLQRHKRTH TGEKPYMNV I NMVKPLHNS

AAB24882.1/1-116 TYHMCQFHCRYVNNHSG EKLYECNERSKAF SCP SHLQCHKRRQ IGEKTHEHNQCGKAF PTPSHLQYHERTH TGEKPYECHQCQGAFAFKCSLLQRHKRTH TGEKPYECNQCGKA-FAQ-



- Align;
  - NP\_000509.1
  - XP\_508242.1
  - NP\_001257813.1
  - NP\_058652.1
  - NP\_990820.1

# BIOL312

Protein Structure and Domain

# Protein Structure

Four levels of structure;

-**Primary**: a.a. sequence of the polypeptide

-**Secondary**: H-bonded 3D local conformation E.g.:  $\alpha$ -helix,  $\beta$ -pleated sheet

-**Tertiary**: folded structure in 3D space. Formed by the interactions between the side chain R-groups; ionic interaction.

\*chaperone molecules

\*domains

-**Quaternary**: over-all structure of multimeric proteins.

(a) Primary structure

MVHLTPEEKSAVTALWGVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSD  
GLAHLNDLKGTFATLSELHCDKLHVDPENFRLLGNVLCVLAHFGKEFTPPVQAAAYQKVVAGVANALAHKYH

(b) Secondary structure

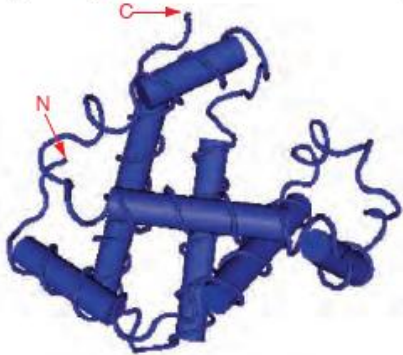
	10	20	30	40	50	60	70				
UNK_257900	MVHLTPEEKSAVTALWGVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLG										
DSC	cccc	hhhhhhhhhhhh	cccc	hhhhhhhhhh	cccc	hhhhhhhh	cccccccccccc	hhhhhhhhhh			
MLRC	cccc	hhhhhhhhhh	cccc	cccc	hhhhhh	eecccc	hhhhh	cccccccccccc	cccc	hhhhh	
PHD	cccc	hhhhhhhhhh	cccc	hhhc	hhhhhh	eecccc	hhhhhhhh	cccc	hhhhh	ee	hhhhhhhhhh
Sec. Cons.	cccc	hhhhhhhhhh	cccc	cccc	hhhhhh	eecccc	hhhhhhhh	cccc	cccccccccccc	hhhhhhhhhh	

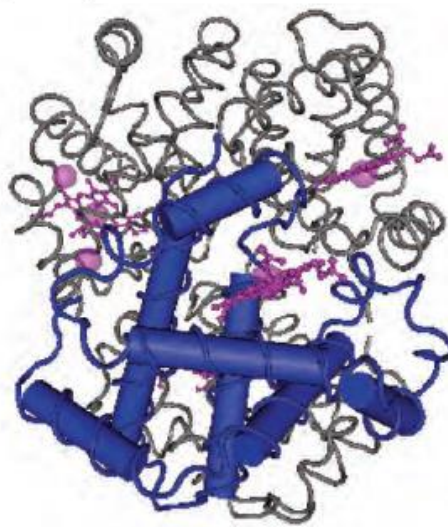
	80	90	100	110	120	130	140	
UNK_257900	AFSDGLAHLNDLKGTFATLSELHCDKLHVDPENFRLLGNVLCVLAHFGKEFTPPVQAAAYQKVVAGVAN							
DSC	hhhhhhhhhhhhhhhhhhhhhhhhhhhhhh	cccc	hhhhhhhhhhhhhhhhhhhh	cccc	cccc	hhhhhhhhhhhhhhhh	hhhhhhhhhhhhhhhh	
MLRC	hhhhhhhhhhhhhhhhhhhhhhhhhhhhhh	cccc	cccc	hhhhhhhhhhhhhhhhhh	cccc	hhhhhhhhhhhhhhhh	hhhhhhhhhhhhhhhh	
PHD	hhhhhhhhhhhhhhhhhhhhhhhhhhhhhh	cccc	hhhhhhhhhhhhhhhhhh	cccc	cccc	hhhhhhhhhhhhhhhh	hhhhhhhhhhhhhhhh	
Sec. Cons.	hhhhhhhhhhhhhhhhhhhhhhhhhhhhhh	cccc	cccc	hhhhhhhhhhhhhhhh	cccc	cccc	hhhhhhhhhhhhhhhh	

UNK\_257900 ALAHKYH  
DSC hhhhccc  
MLRC hhhhccc  
PHD hhhhccc  
Sec. Cons. hhhhccc

(c) Tertiary structure



(d) Quaternary structure



a) The primary structure of a protein refers to the linear polypeptide chain of amino acids.

b) The secondary structure includes elements such as alpha helices and beta sheets.

c) The tertiary structure is the three-dimensional structure of the protein chain.

d) The quaternary structure includes the interactions of the protein with other subunits and heteroatoms.

**Domains:** In folded conformation, most proteins contain specific domains that are discrete structural and functional units of the protein.

These domains are coded by short a.a. sequences and have been linked with specific functions.

- Certain domains have been associated with certain protein families.
- Open reading frames can help us assign them into protein families or understand their functions.
- Domains can help make evolutionary connections between species and understand homology between proteins.

# 1- UniProt (www.uniprot.org)

- The mission of UniProt is to provide the high-quality and freely accessible resource of protein sequence and functional information.

The screenshot displays the UniProt website interface. At the top, there is a navigation bar with the UniProt logo, a search bar containing 'UniProtKB', and a search button. Below the search bar, there are links for 'BLAST', 'Align', 'Retrieve/ID mapping', and 'Peptide search'. On the right side of the navigation bar, there are links for 'Help' and 'Contact'.

The main content area features a mission statement: "The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information."

The page is organized into several sections:

- UniProtKB (UniProt Knowledgebase):** This section is divided into two sub-sections: 'Swiss-Prot (559,228)', which is manually annotated and reviewed, and 'TrEMBL (146,106,279)', which is automatically annotated and not reviewed. It also mentions records with information extracted from literature and curator-evaluated computational analysis.
- UniRef:** A section for 'Sequence clusters'.
- UniParc:** A section for 'Sequence archive'.
- Proteomes:** A section featuring icons for a fly, a person, and a protein structure.
- Supporting data:** A large purple section containing links to 'Literature citations', 'Cross-ref. databases', 'Taxonomy', 'Diseases', 'Subcellular locations', and 'Keywords'.
- News:** A section with social media icons (Blog, Twitter, Facebook, RSS) and news items such as 'Forthcoming changes' (no changes planned), 'UniProt release 2019\_02' (removal of cross-references to CleanEx and change of URIs for Orphanet), and 'UniProt release 2019\_01' (engaging and disengaging: CRISPR rings and cross-references to JPOST). A 'News archive' link is also present.

At the bottom of the page, there are three main sections: 'Getting started' with a 'Text search' link, 'UniProt data' with a 'Download latest release' link, and 'Protein spotlight' with a 'Paths Of Discomfort' link.

# UniProt

([www.uniprot.org](http://www.uniprot.org))

- The mission of UniProt is to provide the high-quality and freely accessible resource of protein sequence and functional information.

# Search for: MAPK1 in human

www.uniprot.org

UniProt

UniProtKB

Advanced Search

BLAST Align Retrieve/ID mapping Help Contact

The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

### UniProtKB

UniProt Knowledgebase

Swiss-Prot (550,740)  
Manually annotated and reviewed.

TrEMBL (63,039,659)  
Automatically annotated and not reviewed.

### UniRef

Sequence clusters

### UniParc

Sequence archive

### Proteomes

### Supporting data

- Literature citations
- Taxonomy
- Subcellular locations
- Cross-ref. databases
- Diseases
- Keywords

### News

Forthcoming changes  
Planned changes for UniProt

UniProt release 2016\_03  
From the Zika forest to the Amazon, news from a viral wanderer | Cross-references to EPD and TopDownProteomics

UniProt release 2016\_02  
Another one (antibiotic) bites the dust | Cross-references to SwissPalm and Gramene | Removal

News archive

## Getting started

- [Text search](#)  
Our basic text search allows you to search all the resources available
- [BLAST](#)  
Find regions of similarity between your sequences
- [Sequence alignments](#)  
Align two or more protein sequences using the Clustal Omega program
- [Retrieve/ID mapping](#)  
This tool merges the "Retrieve" and "ID Mapping" tools



## UniProt data

- [Download latest release](#)  
Get the UniProt data
- [Statistics](#)  
View Swiss-Prot and TrEMBL statistics
- [How to cite us](#)  
The UniProt Consortium
- [Submit your data](#)  
Submit your sequences and annotation updates
- [SPARQL](#)  
Query UniProt data using a SQL like graph query language

## Protein spotlight

### The Art Of Biocuration

March 2016

Museums have their curators. Art galleries too. Their job is to look after collections they are knowledgeable about and present them to an audience in a way that makes sense and is informative. Biocurators do the same. Ever since the advent of computers and advanced technology in the life sciences, the quantity of biological data has grown exponentially and been stored in databases...