

# Identifying and Aligning Homologs

Fran Lewitter

Bioinformatics and Research  
Computing @ Whitehead Institute

6/6/12

Simian sarcoma virus onc gene, v-sis, is derived from the gene (or genes) encoding a platelet-derived growth factor.

Doolittle RF, Hunkapiller MW, Hood LE, Devare SG, Robbins KC, Aaronson SA, Antoniades HN. *Science* 221:275-277, 1983.

Fig. 1. Sequence similarity between p28sis and PDGF. The p28sis sequence is from (10); the PDGF sequences are from (18, 20). Residue identity between the p28sis and PDGF sequences is indicated by the solid lines between the sequences. A question mark indicates that no amino acid sequence assignment has yet been made for that position; the brackets indicate no sequence is yet available for the included segments. The box around p28sis positions 65 and 66 indicates a possible proteolytic processing position for generation of a fragment of p28sis corresponding to PDGF-2. Single letter abbreviations for the amino acid residues are as follows: A, alanine; C, cysteine; D, aspartic acid; E, glutamic acid; F, phenylalanine; G, glycine; H, histidine; I, isoleucine; K, lysine; L, leucine; M, methionine; N, asparagine; P, proline; Q, glutamine; R, arginine; S, serine; T, threonine; V, valine; W, tryptophan; Y, tyrosine.

p28sis	1	MTLTWGGDPIPEELYKMLSGHSIRSFDDLQRLQLGGDSGKEDGAELDNLMT	50
p28sis	51	RSHSGGELESLARGKPSLGSLSVAEPAMIAECKTRTEVFETSRRLIDRTN	100
PDGF-2	1	SLGSLTIAEPAMIAECKTRTEVFETSRRLIDRTN	34
PDGF-1	1	SIEEAVPAVCKTRIVIVYEISRAELD???	28
p28sis	101	ANFLVWPPCVEVQRCSGCCNNRNVQCRPTQVQLAPVQVAKIEIVAKKPIF	150
PDGF-2	35	??????PPCVEVKRCTGCCNNRNVKCAPSQVQLAP?GVAKIEIVAK[	80
PDGF-1	29	ANFL[	32
p28sis	151	KKATVTLEDHLACKCEIVAAARAVTRSPGTSQEGRAKTTQSAVTIARTVAV	200
PDGF-2			
PDGF-1			
p28sis	201	RRPPKGKHKCKKHTHDKTALKETLGA	226
PDGF-2			
PDGF-1			

# What I hope you'll learn

- What we can learn from sequence alignments
- Fundamentals of alignments
- Tools for building alignments

HUMFBRAA: Human fibrinogen A-alpha-chain mRNA, complete cds. [ss-mRNA]  
 EMBL ID: HSFBRRA  
 ACCESSION NUMBERS: J00128  
 DATE: updated 85-12-16

REFERENCES: [1] (bases 1 to 2223) Kant, J.A., Lord, S.T. and Crabtree, G.R.; "Partial mRNA sequences for human A-alpha, B-beta, and gamma fibrinogen chains: Evolutionary and functional implications"; Proc Nat Acad Sci USA 80, 3953-3957 (1983)

KEYWORDS: fibrin; fibrinogen.

SOURCE: Human: liver DNA and cDNA to mRNA. Homo sapiens

COMMENT: An alternate signal peptide start is found at bases 64-66, a consensus polyadenylation signal is found at bases 2180-2185.

FEATURES: key from to description  
 pept 55 1989 A-alpha fibrinogen  
 sigp 55 111 A-alpha fibrinogen signal peptide  
 matp 112 1941 A-alpha fibrinogen mature peptide

SITES: key site span description  
 key 1231 1 A-af 39-bp tandem repeat unit end (8  
 refnumbr 1 1 numbered 25 in [1] copies)  
 mRNA/mRNA 1 0 A-af mRNA unsequenced/sequenced  
 ->pept 55 1 A-af cds signal pept start  
 pept/pept 112 0 A-af cds signal pept end/mature pept  
 start  
 ->rpt 920 1 A-af 39-bp tandem repeat unit start  
 (8 copies)

ORIGIN: Unreported.

SEQUENCE: 2223 bp 648 a 494 c 553 g 528 t  
 1 aatcctttct ttcagctgga gtgtctcag gagccagccc caaccttaga aaagatgttt tccatgagga tctgtctgct ggtcctaagt gtggtgagca  
 101 cagcatggac tgcagatagt ggtgaagggt actttctagc tgaaggagga ggcgtgctgt gcccaagggt tgtggaaga catcaatctg cctgcaaga  
 201 ttcagactgg cccttctgct ctgatgaaga ctggaactac aaatgccctt ctggctgcag gatgaagggt ttgattgatg aagtcacatc agattttaca  
 301 aacagaataa ataagotcaa aaattcacta ttggaatctc agaagaacaa taaggattct cattctgtga ccactaatat aatggaaatt ttgagaggcg  
 401 atttttcttc agccaataac cgtgataata cctacaaccc agtctcagag gatctgagaa gcagaattga agtctctgaag cgcaaaagtc tagaaaaagt  
 501 acagcatatc cagcttctgc aaaaaaatgt tagagctcag ttggttgata tgaacagact ggaggtggac attgatatta agatccgato ttgtcagagg  
 601 tcatgcagta gggttttagc tegtgaagta gatctgaagg actatgaaga tcagcagaag caacttgaa aggtcattgc caaagactta cttccctota  
 701 gagataggca acacttaacca ctgatcaaaa tgaanaccagt tccagacttg gttcccgaa attttaagag ccagcttcag aaggtacccc cagagtggaa  
 801 ggcatttaaca gacatgcgc agatgagaat ggagttagag agacctgtgt gaaatgatag tactogagga ggtccactt cttatggaa cggatcagag  
 901 acggaaagcc caaggaaacc tagcagtgtc ggaagctgga actctgggag ctctggacct ggaagtactg gaaacgaaa ccctgggagc tctgggactg  
 1001 gagggactgc aacctggana cctggaagct ctggacctgg aagtactgga agctggaaact ctgggagctc tggaaactgga accaaaacc  
 1101 tgggagccct agacctggtg gtaccggaaac ctggaatcct ggcagctctg aacgcggaag tctgggacac tggactctgt agagctctgt atctggtagt  
 1201 actggacaat ggcactctga atctggaagt tttaggccc agatgccagg ctctgggaaac ggcaggccca acaaccaga ctggggcaca ttgaaagagg  
 1301 tgtcaggaaa tgtaagtcca gggacaagga gagagtacca cacagaaaaa ctggtcactt ctaaggaga taaagagctc aggaactggt aagagaaggt  
 1401 cactctctgt agcacaca ccaagcgtct ttcagctctc aaaaacctta ctaagactgt tattggtcct gatgtcaca aagaagtta caaagaagt  
 1501 gtgacctcag aagatggttc tgaactgtcc gagggcaatg atttaggcac ttctcaccta tgttaggaga gtttgcagt gagactgagt ctaggggcto  
 1601 aagctgcctt cttcgacact gctcaactg gaaaaacatt cccagggttc ggatagctga attcccttc cgtggttaaa cttcaagttc cagcaaaaca  
 1701 agaactctgc atcttcacaa atacaagga atccagttt catcacctg gagctataaa atggcagatg aggcocggaag tgaagccgat catgaaggaa  
 1801 ttaactagta gcacaggtta caacagagga gactccacat ttgaaagcaa cacacttctc cttggggan gcttccctg tccccctaga ctaagttaaa  
 1901 cactatgac caagagagga catgctaaat ctgcacctgt cagaggatc tgttacagt cattgaaact acacttttt ggtctgttt tgtgctagac  
 2001 tatttttgca cagtgttccc atggccctt gcaatttctt cttactctc aaatgctaa cagtacagga ccatgactca ataaatacat gttaaatgga  
 2101 tgttaagtcc ttggggcgag ggcctttgtc tgtctcatct ctgtattccc  
 2201 tgaatgaatt cctctgaaac tct

tggtg  
tctga  
aaat  
cctg  
cagaa  
ctgt  
actga  
cagat  
ctago  
acctg  
agtac  
aatct  
agga  
acac  
ctgac  
ctgct  
aatac  
tacaac  
ggcag  
ccatgg  
agggcc

-12-16

Sci

1 at

end (8  
ept

gtggga  
gcaaga  
ttttaca  
agaggcg  
aaaaagt  
tccaggg  
ccctota  
agtgga  
atcagag  
ggagctg  
aaaaacc  
tctgagt  
tgaagag  
gagaggt  
agaagt  
cctgatg  
agggcto  
gcaacaa  
agttaba  
tgttaga  
taattga

85-08-01

s 11.

# Topics to Cover

---

- Introduction
  - Why do alignments?
  - Definitions
  - Scoring alignments
- Pairwise Alignment methods
- Multiple sequence alignments
- Pre-computed alignment resources

# Why do alignments

---

- Use sequence similarity to infer homology and/or structural similarity between 2 or more genes/proteins
- Identify more conserved regions of a protein, potentially identifying regions of most functional importance
- Compare and contrast homologs (perhaps into groups) based on shared positions or regions
- Infer evolutionary distance from sequence dissimilarity



# Evolutionary Basis of Sequence Alignment

---

- *Similarity* - observable quantity, such as percent identity
- *Homology* - conclusion drawn from data that two genes share a common evolutionary history; no metric is associated with this
  - *Paralog* – genes related by duplication
  - *Ortholog* – genes related by speciation



# More Definitions

---

- An *alignment* is a mutual arrangement of two sequences, which exhibits where the two sequences are similar, and where they differ.
- An *optimal alignment* is one that exhibits the most correspondences and the least differences. It is the alignment with the highest score. May or may not be biologically meaningful.

# Alignment Concepts

- ***Global alignment*** - Needleman-Wunsch (1970) maximizes the number of matches between the sequences along the entire length of the sequences.
- ***Local alignment*** - Smith-Waterman (1981) produces the highest scoring regional match between two sequences.
- ***Insertion and Deletions*** (indels)
- ***Affine gap costs*** - a scoring system for gaps within alignments that charges a penalty for the existence of a gap and an additional per-residue penalty proportional to the gap's length

# Global vs Local Alignment

---

## GLOBAL

L	G	P	S	<u>S</u>	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

## LOCAL

—	—	—	—	—	—	—	T	G	K	G	—	—	—	—	—	—	—	—
—	—	—	—	—	—	—	A	G	K	G	—	—	—	—	—	—	—	—

From Mount, *Bioinformatics*, 2004, pg 71

# Possible Alignments

---

A: T C A G A C G A G T G  
B: T C G G A G C T G

I.     T C A G A C G A G T G  
       T C G G A - - G C T G

II.    T C A G A C G A G T G  
       T C G G A - G C - T G

III.   T C A G A C G A G T G  
       T C G G A - G - C T G

# Nucleotide vs Protein

---

- If comparing protein coding genes, use protein sequences because of less noise
- If protein sequences are very similar, it might be more instructive to use DNA sequences
- If interested in DNA alignment of coding sequences, first do a protein alignment and use it as a template for aligning DNA sequences

# AA Scoring Matrices

Part of PAM 250 Matrix

	C	S	T	P	A	G	N
C	12						
S	0	2					
T	-2	1	3				
P	-3	1	0	6			
A	-2	1	1	1	2		
G	-3	1	0	-1	1	5	
N	-4	1	0	-1	0	0	

•**PAM** - point accepted mutation  
based on *global* alignment  
[evolutionary model]

$$\text{Log-odds} = \frac{\text{pair in homologous proteins}}{\text{pair in unrelated proteins by chance}}$$

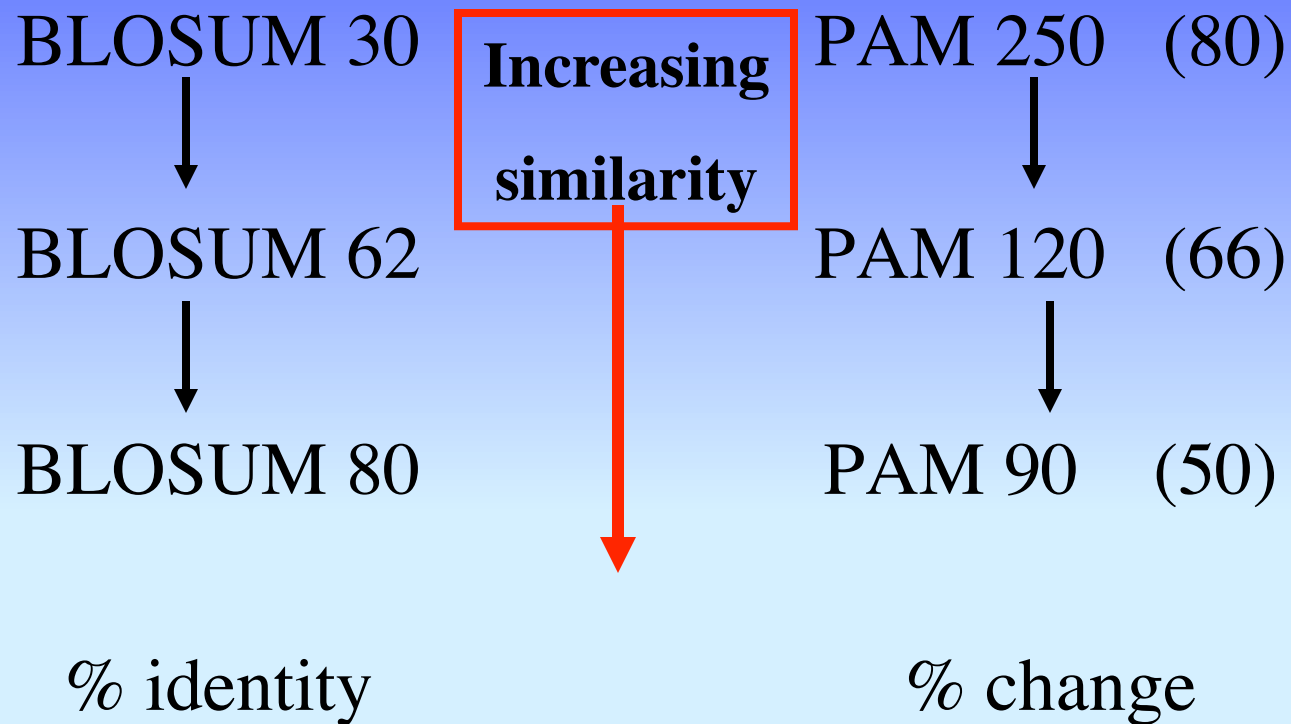
Part of BLOSUM 62 Matrix

	C	S	T	P	A	G	N
C	9						
S	-1	4					
T	-1	1	5				
P	-3	-1	-1	7			
A	0	1	0	-1	4		
G	-3	0	-2	-2	0	6	
N	-3	1	0	-2	-2	0	

•**BLOSUM** - block substitutions  
based on *local* alignments [similarity  
among conserved sequences]

$$\text{Log-odds} = \frac{\text{obs freq of aa substitutions}}{\text{freq expected by chance}}$$

# Substitution Matrices





# Scoring for BLAST Alignments

Score = 94.0 bits (230), Expect = 6e-19

Identities = 45/101 (44%), Positives = 54/101 (52%), Gaps = 7/101 (6%)

Query: 204 YTGPFCDV----DTKASCYDGRGLSYRGLARTTSLGAPCQPWASEATYRNVTAEQ---AR 256

Y+ FC + + CY G G +YRG T SGA C PW S V Q A+

Sbjct: 198 YSSEFCSTPACSEGNSDCYFGNGSAYRGTHSLTESGASCLPWNSMILIGKVYTAQNPSAQ 257

Query: 257 NWGLGGHAFCRNPNDIRPWCFVLNRDRLSWEYCDLAQCQT 297

GLG H +CRNPD D +PWC VL RL+WEYCD+ C T

Sbjct: 258 ALGLGKHNYCRNPDGDAKPWCHVLKNRRLTWEYCDVPSCST 298

Based on  
BLOSUM62

Position	1:	Y	-	Y	=	7
Position	2:	T	-	S	=	1
Position	3:	G	-	S	=	0
Position	4:	P	-	E	=	-1
		.	.	.		
Position	9:	-	-	P	=	-11
Position	10:	-	-	A	=	-1
		.	.	.		
					Sum	230

# What's significant?

---

- High confidence -  $>40\%$  identity for long alignments (Rost, 1999 found that sequence alignments unambiguously distinguish between protein pairs of similar and non-similar structure when the pairwise sequence identity  $>40\%$ )
- “Twilight zone” – blurry - 20-35% identity
- “Midnight zone” -  $<20\%$  identity

# Topics to Cover

---

- Introduction
- Pairwise Alignment methods
  - Dot plot analysis
  - Exhaustive methods; Dynamic programming algorithm (Smith-Waterman (Local), Needleman-Wunsch (Global))
  - Heuristic methods; Approximate methods; word or k-tuple (FASTA, BLAST, BLAT)
- Multiple sequence alignments
- Pre-computed alignment resources

# Comparing two sequences

---

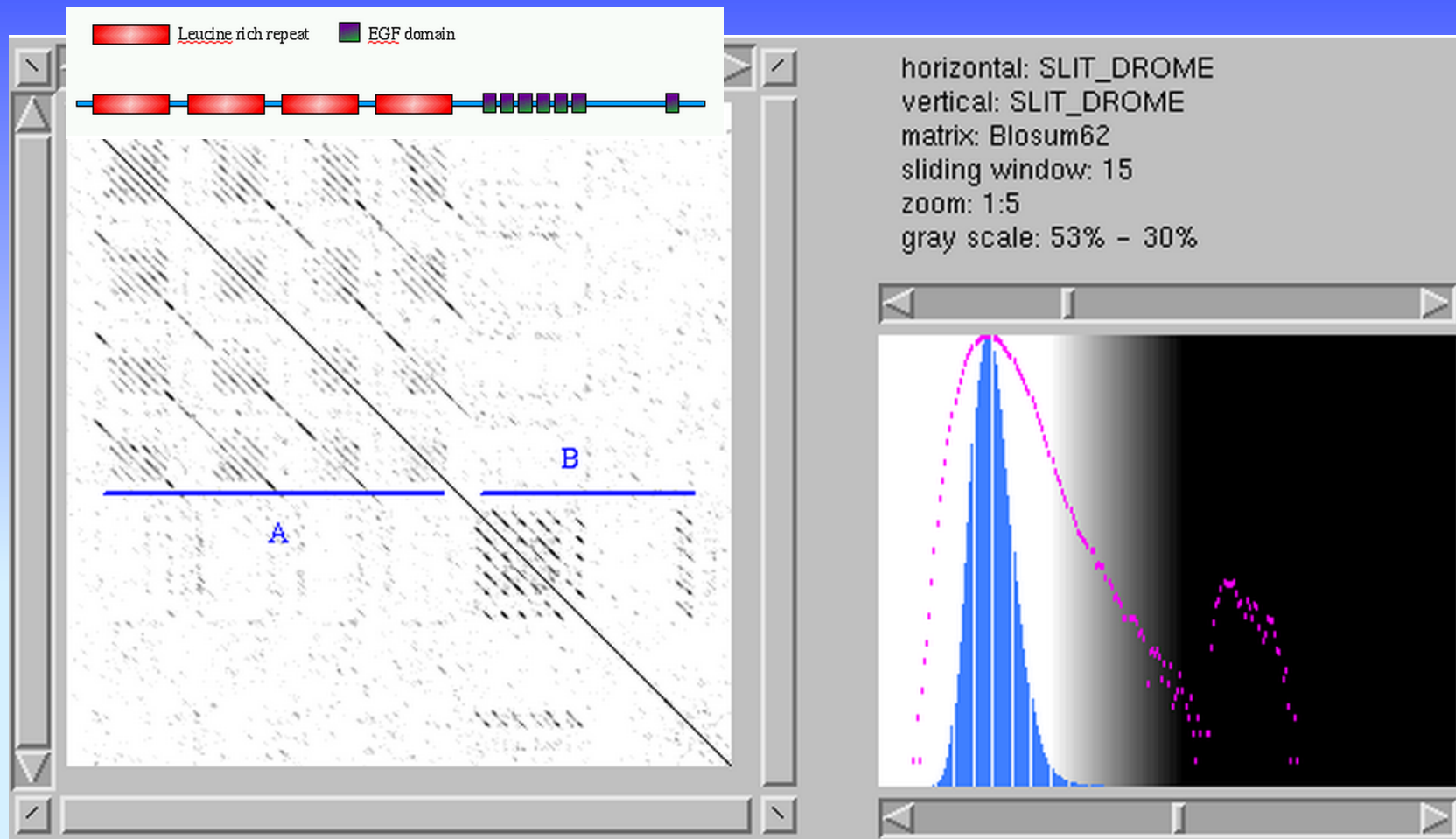
- DOTLET – Dot Plot (<http://myhits.isb-sib.ch/cgi-bin/dotlet>)
- NCBI  
BLAST 2 Sequences ([www.ncbi.nlm.nih.gov/blast/bl2seq/wblast2.cgi](http://www.ncbi.nlm.nih.gov/blast/bl2seq/wblast2.cgi))
- EBI – (<http://www.ebi.ac.uk/Tools/psa/>)
  - GLOBAL
    - needle (EMBOSS) - Needleman - Wunsch
    - stretcher (EMBOSS) – modification of N-W
  - LOCAL
    - water (EMBOSS) - Smith-Waterman
    - matcher (EMBOSS) - uses algorithm based on LALIGN

# Dot Plot

---

- Graphical way of looking at alignment of 2 sequences
- Look at structure of a sequence by doing a self comparison
- Method first described by Gibbs and McIntyre (1970)
- Can find direct or inverted repeats in sequences

# Dot Matrix Comparison



<http://myhits.isb-sib.ch/cgi-bin/dotlet>

Identifying and Aligning Homologs (Whitehead Institute)

# NCBI - Blast2seq

NCBI/BLAST/blastp suite

Align Seq Algorithm parameters

blastn blastp blastx tblastn tblastx

Enter Query Sequence

BLASTP programs search

Enter accession number(s), gi(s), or FASTA sequence(s)

>gi|15235708|ref|NP\_195491.1| 1-aminocyclopropane-1-carl  
[Arabidopsis thaliana]  
MGLLSKKASCNTHGQDSSYFWGWEEYKPNPYDEIKNPDGIIQMGLAEN  
ANFQR  
EGQSIFRELALFQDYHGLPSFKNAMADFMSENRCNRVSFNPNKLVLTAC

Or, upload file Choose File No file chosen

Job Title gi|15235708|ref|NP\_195491.1| 1-aminocyclopropane-1-carl  
Enter a descriptive title for your BLAST search

☒ Align two or more sequences

Enter Subject Sequence

Enter accession number, gi, or FASTA sequence

>gi|62751809|ref|NP\_001015526.1| 1-aminocyclopropane-1-  
like protein 1 [Bos taurus]  
MFTLPQKEFRMTTACPGSDSIQDLPSNKGDLERECRKPDQKLLKFYC  
RGSV  
IKWFWDSAEEGYRTYHMDYEDDKNPSGIINLGTSENKLCFDLLSRRLS

Or, upload file Choose File No file chosen

Program Selection

Algorithm ☒ blastp (protein-protein BLAST)  
Choose a BLAST algorithm

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 10

Word size 3

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

Identifying and A



# NCBI - Blast2seq

NCBI/BLAST/blastp suite-2sequences/ Formatting Results - VJUKYHEG112

[Edit and Resubmit](#) [Save Search Strategies](#) [Formatting options](#) [Download](#)

Blast 2 sequences

gi|15235708|ref|NP\_195491.1| 1-aminocyclopropane-1-carboxylate...

<b>Query ID</b>	Id 31635	<b>Subject ID</b>	31637
<b>Description</b>	gi 15235708 ref NP_195491.1  1-aminocyclopropane-1-carboxylate synthase 8 [Arabidopsis thaliana]	<b>Description</b>	gi 62751809 ref NP_001015526.1  1-aminocyclopropane-1-carboxylate synthase-like protein 1 [Bos taurus]
<b>Molecule type</b>	amino acid	<b>Molecule type</b>	amino acid
<b>Query Length</b>	469	<b>Subject Length</b>	502
		<b>Program</b>	BLASTP 2.2.26+ <a href="#">Citat</a>

Other reports: [Search Summary](#) [Taxonomy reports](#) [Multiple alignment](#)

**Graphic Summary**

Distribution of 1 Blast Hits on the Query Sequence

Mouse over to see the define, click to show alignments

Color key for alignment scores

<40	40-50	50-80	80-200	>=200
-----	-------	-------	--------	-------

Query

1 90 180 270 360 450

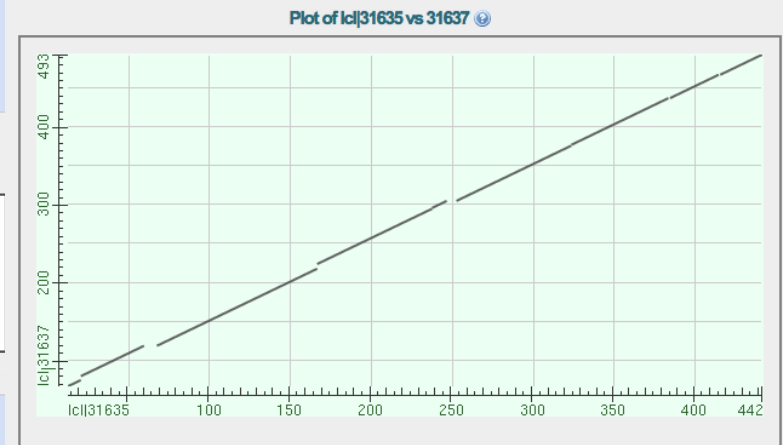
**Dot Matrix View**

**Descriptions**

Legend for links to other resources: [U](#) UniGene [E](#) GEO [G](#) Gene [S](#) Structure [M](#) Map Viewer [P](#) PubChem BioAssay

Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
31637	gi 62751809 ref NP_001015526.1  1-aminocyclopropane-1-	221	221	91%	1e-69	31%	



Identifying and Aligning Homologs (Whitehead Institute)

# NCBI - Blast2seq

**Alignments**

>lcl|31637 gi|62751809|ref|NP\_001015501| 1-aminocyclopropane-1-carboxylate synthase-like protein 1 [Bos taurus]  
Length=502

Score = 221 bits (563), Expect = 1e-69, Method: Compositional matrix adjust.  
Identities = 137/442 (31%), Positives = 216/442 (49%), Gaps = 29/442 (7%)

Query 14 QDSSSYFW-----GWEEYEKNPYDEIKNPDGIIQMGLAENQLSFDLIESWLAKNPDAANF 68  
G +FW G+ Y + YDE KNP GII +G +EN+L FDL+ L+  
Sbjct 68 GSVIKWFWDSEEGYRTYHMDYEDDKNPSGIINLGTSENKLCFDLLSRRLS----- 119

Query 69 QREGQSIFRELALFQDYHGLPSFKNAMADFMSENNGNRVSNPNKLVLTAGATPANETLM 128  
Q + + L + D+ G + +A F+S + P +V+ G L  
Sbjct 120 QSDMLQVEPALLQYPDWRGHLFLREEVARFLSFYCRSPAPLKPENVVVLNGCASLFSALA 179

Query 129 FCLADPGDAFLLPTPYYPGFDRLKWRGAEIVPIQCKS-----ANGFRITKVALEEAY 182  
L + G+AFL+P PYY + + +V + S F++T LE A  
Sbjct 180 TVLCEAGEAFLIPAPYYGAITQHVYLYGNVRLVCVYLDSEVTGLETRPFQLTVEKLEMAL 239

Query 183 EQAQKLNKLVKGVLTITNPSNPLGTTTTTTELNLHLLDFISRKKIHLISDEIYSGTVF-TNP 241  
+ A +KVKG+++ NP NPLG + EL L+F R ++H++ DE+Y +VF +  
Sbjct 240 QGANSEGVKVKGLILINPQNPLGDIYSPGELQEYLEFAKRHELHVMVDEVYMLSVFEESA 299

Query 242 GFISVMEVLKDRKLENTDVFDRVHIVYSLSKDLGLPGFRVGVIIYSNDDFVVSAAATKMSSF 301  
G+ SV+ LE R H++++ SKD G+ G R G +Y+ + V +A + +  
Sbjct 300 GYRSVL-----SLERLPDPQRTTHVMWATSKDFGMSGLRFGTLYTENWAVATAVASLCRY 353

Query 302 GLISSQTQYLLSALLSDKTFTKN-YLEENQIRLKNRHKLVSGLEAAGIECLKSNAGLFC 360  
+S QY ++ LL D + YL EN RLK H + L A GI + AG F  
Sbjct 354 HGLSGLVQYQMAQLLRDHDWINQVYLPENHARLKAHTYVSEDLRALGIPFVSRGAGFFI 413

Query 361 WVDMRHLKSNTFEAEIELWKKIYVEVKLNISPGSSCHCNEPGWFRVCFANLSEETLKVA 420  
WVD+R L TFE E+ LW++ + E K+ +S G + C EPGWFR+ F++ + L +  
Sbjct 414 WVDLRKYLPEATFEEVLLWRRFL-ENKVLLSFGKAFCCKEPGWFRLVFSDKTHR-LHLG 471

Query 421 LDRLKRFVDGSPSPTRRSQSEHQ 442  
+ R+++ ++G HQ  
Sbjct 472 MQVRVQVLEGQPQLADGAPPHQ 493

Identifying and Aligning Homologs (Whitehead Institute)

# Pairwise Alignment Tools at EBI

EBI > Tools > Pairwise Sequence Alignment

## Pairwise Sequence Alignment

**Pairwise Sequence Alignment** is used to identify regions of similarity that may indicate functional, structural and/or evolutionary relationships between two biological sequences (protein or nucleic acid).

By contrast, **Multiple Sequence Alignment (MSA)** is the alignment of three or more biological sequences of similar length. From the output of MSA applications, homology can be inferred and the evolutionary relationship between the sequences studied.

### Global Alignment

Global alignment tools create an end-to-end alignment of the sequences to be aligned. There are separate forms for protein or nucleotide sequences.

**Needle** ⓘ  
(EMBOSS)      EMBOSS Needle creates an optimal global alignment of two sequences using the Needleman-Wunsch algorithm.

**Stretcher** ⓘ  
(EMBOSS)      EMBOSS Stretcher uses a modification of the Needleman-Wunsch algorithm that allows larger sequences to be globally aligned.

### Local Alignment

Local alignment tools find one, or more, alignments describing the most similar region(s) within the sequences to be aligned. There are separate forms for protein or nucleotide sequences.

**Water** ⓘ  
(EMBOSS)      EMBOSS Water uses the Smith-Waterman algorithm (modified for speed enhancements) to calculate the local alignment of two sequences.

**Matcher** ⓘ  
(EMBOSS)      EMBOSS Matcher identifies local similarities between two sequences using a rigorous algorithm based on the LALIGN application.

**LALIGN** ⓘ      LALIGN finds internal duplications by calculating non-intersecting local alignments of protein or DNA sequences.

[http://www.ebi.ac.uk/Tools/pairwise/aligning\\_homologs/](http://www.ebi.ac.uk/Tools/pairwise/aligning_homologs/) Identifying and Aligning Homologs (Whitehead Institute)

# Needle (global)

EBI > Tools > Pairwise Sequence Alignment > EMBOSS Needle

## EMBOSS Needle - Pairwise Sequence Alignment

EMBOSS Needle reads two **protein** or **nucleotide** input sequences and writes their optimal global sequence alignment to file.

### Use this tool

#### STEP 1 - Enter your protein sequences

Enter or paste your first **protein** se

```
>gi|15235708|ref|NP_195491.1|
thaliana]
MGLLSKKASCNTHGQDSSYFWGWE
EGQSIFRELALFQDYHGLPSFKNAMA
PTPYYPGFDRLDKWRTGAEIVPIQCK
TELNHLLDFISRKKIHLISDEIYSGTVF
VGVIYSNDDFVSAATKMSSFGLISSi
```

Or, **upload** a file:

**AND**

Enter or paste your second **protein**

```
>gi|62751809|ref|NP_00101552
taurus]
MFTLPQKEFRMTTACPGSDSIQDLPs
IKWFWDSAEEGYRTYHDEYDEDKs
FLREEVARFLSFYCRSPAPLKPENVV
LVCVYLDSEVTGLETRPFQLTVEKLEI
ELHVMVDEVYMLSVFEESAGYRSVLS
```

Or, **upload** a file:

#### STEP 2 - Set your pairwise alignment options

**MATRIX**

BLOSUM62

**GAP OPEN**

10

**GAP EXTEND**

0.5

**OUTPUT FORMAT**

pair

**END GAP  
PENALTY**

false

**END GAP  
OPEN**

10

**END GAP  
EXTEND**

0.5

#### STEP 3 - Submit your job

☐ Be notified by email *(Tick this box if you want to be notified by email when the results are available)*

Submit

#### STEP 2 - Set your pairwise align

The default settings will fulfill the ne

*(Click here, if you want to view or change the default settings.)*

#### STEP 3 - Submit your job

☐ Be notified by email *(Tick this box if you want to be notified by email when the results are available)*

Submit

Identifying and Aligning Homologs (Whitehead Institute)

# Needle (global)

EBI > Tools > Pairwise Sequence Alignment > EMBOSS Needle

## EMBOSS Needle - Alignment

Alignment Submission Details Submit Another Job

### Alignment

View Alignment File

```
#####
# Program: needle
# Rundate: Sun 20 May 2012 23:24:21
# Commandline: needle
#
# -auto
# -stdout
# -asequence emboss_needle-I20120520-232408-0601-60957810-o-
# -bsequence emboss_needle-I20120520-232408-0601-60957810-o-
# -datafile EBLOSUM62
# -gapopen 10.0
# -gapextend 0.5
# -endopen 10.0
# -endextend 0.5
# -aformat3 pair
# -sprotein1
# -sprotein2
# Align_format: pair
# Report_file: stdout
#####

#=====
#
# Aligned sequences: 2
# 1: NP_195491.1
# 2: NP_001015526.1
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 545
# Identity: 140/545 (25.7%)
# Similarity: 223/545 (40.9%)
# Gaps: 119/545 (21.8%)
# Score: 575.5
#
#=====
```

```
NP_195491.1      1  -----
NP_001015526.1  1  MFTLPQKEFRMTTACPGSDSIQDLPSNKGDLEREC SRKPDQKLLKFYGV 50
NP_195491.1      1  ----MGLLSKKASCNTHGQDSSYFW-----GWEEYEKNPYDEIKNPDGII 41
NP_001015526.1  51  GDPAAELSSSSPYLSSRGSVIKWFWD SAEEGYRTYHMDYDEDEDKNPSGII 100
NP_195491.1      42  QMGLAENQLSFDLIESWLAKNPDAANFQREGQSIFRELALFQDYHGLPSF 91
NP_001015526.1  101  NLGTSENKLCFDLLSRRLS-----QSDMLQVEPALLQYPDWRGHLFL 142
NP_195491.1      92  KNAMADFMSEN RGNRVSNPNKLVLTAGATPANETLMFCLADPGDAFLLP 141
NP_001015526.1  143  REEVARFLSFYCRSPAPLKPENVVV LNCASLFSALATVLC EAGEAFLIP 192
NP_195491.1      142  TPYPYGFDRDLKWRGTAEI VPIQCKS-----ANGFRITKVALEEAEYQA 185
NP_001015526.1  193  APYYGAITQHVVLYGNVRLVCVYLDSEVTGLETRPFLQTV EKLEMALQGA 242
NP_195491.1      186  QKLNKLVKGV LITNPSNPLGTTTTRTELNHLDFISRKKIHLISDEIYSG 235
NP_001015526.1  243  NSEGVKVKGLILINPQNPLGDIYSPGELQEYLEFAKRHELHVMVDEVYML 292
NP_195491.1      236  TVF-TNPGFISVMEVLKDRKLENTDVFDRVHIVYSLSKDLGLPGFRVGV 284
NP_001015526.1  293  SVFEESAGYRSVL-----SLERLPDPQRTHVMWATSKDFGMSGLRFGTL 336
NP_195491.1      285  YSNDDFVVSAAATKMSSFGLISSQTQYLLSALLSDKTFTKN-YLEENQIRL 333
NP_001015526.1  337  YTENWAVATAVASLCRYHGLSGLVQYQMAQLLRDHDWINQVYLPENHARL 386
NP_195491.1      334  KNRHKKLVSGLEAAGIECLKSNAGLFCWVDMRHLLKSNTFEAEIELWKKI 383
NP_001015526.1  387  KAAHTYVSEDLRALGIPFVSRGAGFFIWVDLRKYLPEATFE EEVLLWRRF 436
NP_195491.1      384  VYEVKLNISPSSSCHCNEPGWFRVCFANLSEETLKVALDRLKRFVDG--- 430
NP_001015526.1  437  L-ENKVLVSFGKA FECKEPGWFRLVFSDKTHR-LHLGMQRVRQVLEGQPO 484
NP_195491.1      431  -----PSPTRRSQSEHQRLKNLRKMVSNWVFRLSFHDREPEER 469
NP_001015526.1  485  LADGAPPHQIQEPQGPHR----- 502
```

Identifying and Aligning Homologs (Whitehead Institute)

# Water (local)

EBI > Tools > Pairwise Sequence Alignment > EMBOSS Water

## EMBOSS Water - Pairwise Sequence Alignment

EMBOSS Water uses the Smith-Waterman algorithm (modified for speed enhancements) to calculate the local alignment of two **protein** or **nucleotide** sequences.

**Use this tool**

**STEP 1 - Enter your protein sequences**

Enter or paste your first **protein** sequence in any **supported** format:

```
>gi|15235708|ref|NP_195491.1| 1-aminocyclopropane-1-carboxylate synthase 8 [Arabidopsis thaliana]
MGLLSKKASCNTHGQDSSYFWGWEEYEKNPYDEIKNPDGIIQMGLAENQLSFDLIESWLAKNPDAANFQR
EGQSIFRELALFQDYHGLPSFKNAMADFMSENRRVSNPNKLVLTAGATPANETLMFCLADPGDAFLL
PTPYYPGFDRLDKWRTGAEIVPIQCKSANGFRITKVALEEAYEQAKLNLKVKGVLITNPSNPLGTTTTTR
TELNHLLDFISRKKIHLISDEISGTVFTNPGFISVMEVLKDRKLENTDVFDRVHIVYSLSKDLGLPGFR
VGVIYSNDDFVVSAAATKMSSFGLISSQTQYLLSALLSDKFTTKNYLEENQIRLKNRHKKLVSGLEAAGIE
```

Or, upload a file:  No file chosen

**AND**

Enter or paste your second **protein** sequence in any **supported** format:

```
>gi|62751809|ref|NP_001015526.1| 1-aminocyclopropane-1-carboxylate synthase-like protein 1 [Bos taurus]
MFTLPQKEFRMTTACPGSDSIQDLPSNKGDLERECRSRKPQDKLLKFYGVGDPAELSSSSPYLSSRGSV
IKWFWDSAEEGYRTYHMDYDEDEKNPSGIINLGTSENKLCFDLLSRRLSQSDMLQVEPALQYPDWRGHL
FLREEVARFLSFYCRSPAPLKPENVVVLNCCASLFSALATVLCCEAEAFIPAPYGAITQHVVLYGNVR
LVCVYLDSEVTGLETRPFQLTVEKLEMALQGANSEGVKVGILINPQNPLGDIYSPGELQEYLEFAKRH
ELHVMVDEVYMLSVFEESAGYRSVLSLERLPDPQRTHVMWATSKDFGMSGLRFGTLYTENWAVATAVASL
```

Or, upload a file:  No file chosen

**STEP 2 - Set your pairwise alignment options**

MATRIX	GAP OPEN	GAP EXTEND	OUTPUT FORMAT
<input type="text" value="BLOSUM62"/>	<input type="text" value="10"/>	<input type="text" value="0.5"/>	<input type="text" value="pair"/>

**STEP 3 - Submit your job**

☐ Be notified by email (Tick this box if you want to be notified by email when the results are available)

Identifying and Aligning Homologs (Whitehead Institute)



[EBI](#) > [Tools](#) > [Pairwise Sequence Alignment](#) > [EMBOSS Water](#)

## EMBOSS Water - Alignment

[Alignment](#)
[Submission Details](#)
[Submit Another Job](#)

### Alignment

[View Alignment File](#)

```
#####
# Program: water
# Rundate: Sun 20 May 2012 23:28:18
# Commandline: water
#   -auto
#   -stdout
#   -asequence emboss_water-I20120520-232815-0355-55247721-oy.asequence
#   -bsequence emboss_water-I20120520-232815-0355-55247721-oy.bsequence
#   -datafile EBLOSUM62
#   -gapopen 10.0
#   -gapextend 0.5
#   -aformat3 pair
#   -sprotein1
#   -sprotein2
# Align_format: pair
# Report_file: stdout
#####
#=====
#
# Aligned_sequences: 2
# 1: NP_195491.1
# 2: NP_001015526.1
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 453
# Identity:      139/453 (30.7%)
# Similarity:    220/453 (48.6%)
# Gaps:          29/453 ( 6.4%)
# Score: 578.5
#
#
#=====
```

[illegible]

## Identifying and Aligning Homologs (Whitehead Institute)



# Topics to Cover

---

- Introduction
- Pairwise Alignment methods
- Multiple sequence alignments
- Pre-computed alignment resources

# Multiple Sequence Alignment

- Additional sequences can help resolve ambiguities found in a pairwise comparison
- Remove uninformative sequences
- Dynamic programming techniques require prohibitively large computer resources
- Tree or hierarchical methods (successive pairwise alignments), Consistency-based methods, Template-based methods

# Multiple Sequence Alignment

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

EBI > Tools > Multiple Sequence Alignment

## Multiple Sequence Alignment

**Multiple Sequence Alignment (MSA)** is generally the alignment of three or more biological sequences (protein or nucleic acid) of similar length. From the output, homology can be inferred and the evolutionary relationships between the sequences studied.

By contrast, **Pairwise Sequence Alignment** tools are used to identify regions of similarity that may indicate functional, structural and/or evolutionary relationships between two biological sequences.

### Clustal Omega



New MSA tool that uses seeded guide trees and HMM profile-profile techniques to generate alignments (protein only). Suitable for medium-large alignments.

[Launch Clustal Omega](#)

### ClustalW2



Popular MSA tool that uses tree-based progressive alignments. Suitable for medium alignments.

[Launch ClustalW2](#)

### DbClustal



Create a Multiple Sequence Alignment from a protein BLAST result using the DbClustal program.

[Launch DbClustal](#)

### Kalign



Very fast MSA tool that concentrates on local regions. Suitable for large alignments.

[Launch Kalign](#)

### MAFFT



MSA tool that uses Fast Fourier Transforms. Suitable for medium-large alignments.

[Launch MAFFT](#)

### MUSCLE



Accurate MSA tool, especially good with proteins. Suitable for medium alignments.

[Launch MUSCLE](#)

### MView



Transform a Sequence Similarity Search result into a Multiple Sequence Alignment or reformat a Multiple Sequence Alignment using the MView program.

[Launch MView](#)

### T-Coffee



Consistency-based MSA tool that attempts to mitigate the pitfalls of progressive alignment methods. Suitable for small alignments.

[Launch T-Coffee](#)

Identifying and Aligning Homologs (Whitehead Institute)

# Clustal Omega

[EBI](#) > [Tools](#) > [Multiple Sequence Alignment](#) > [Clustal Omega](#)

## 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.

### Use this tool

### STEP 1 - Enter your input sequences

Enter or paste a set of **PROTEIN** sequence in any supported form

>gi|187936925|ref|NP\_001120691.1| 1-aminocyclopropane-1-carboxylate  
[Homo sapiens]  
MFTLPQKDFRAPSTCLGPTCMQDLCSSHGDELEGECRSKLDQKLPELRGVG  
KFWWDSAEEGYRTYHMDYEDEDKNPSGIINLTSENKLCFDLLSWRLSQRD  
LREEVAKFLSFYCKSPVPLRPENVVLNGGASLFSALATLVCEAGAEFLIPTPY  
AYVYLDSEVTGLDTRPFQLTVEKLEMLALREHSEGVKKVKGILISPAQNPLGDV  
LHVIVDEVYMLSVFEKSGVGRYSLSRLPDPQRTHVMVATSSKDFGMSGLRL

Or, upload a file:  No file chosen

## STEP 2 - Set your parameters

DEALIGN INPUT SEQUENCES	OUTPUT ALIGNMENT FORMAT	MULTI-SEQUENCE CLUSTERING TREE
no	Clustal	yes

NUMBER of COMBINED ITERATIONS	MAX GUIDE TREE ITERATIONS	MAX HMM ITERATIONS
default(0)	default	default

### STEP 3 - Submit your job

☐ Be notified by email (Tick this box if you want to be notified by email when the results are available)

Submit

Defines the type of the sequences to be aligned

Type	Value
PROTEIN	protein
DNA	dna

```
gi|187936925|ref|NP_001120691.1|
gi|332836227|ref|XP_003313044.1|
gi|109106658|ref|XP_001113725.1|
gi|62751809|ref|NP_001015526.1|
gi|34850059|ref|NP_899043.1|
gi|363734384|ref|XP_421096.3|
gi|12212835|ref|NP_001014345.1|
gi|15235708|ref|NP_195491.1|
gi|115455083|ref|NP_001051142.1|
```

```
gi | 187936925 | ref | NP_001120691.1 |
gi | 323836227 | ref | XP_003313044.1 |
gi | 109106658 | ref | XP_001137225.1 |
gi | 62751809 | ref | NP_001015526.1 |
gi | 34805509 | ref | NP_899043.1 |
gi | 363734384 | ref | XP_421096.3 |
gi | 62122835 | ref | NP_001014345.1 |
gi | 15235708 | ref | NP_195491.1 |
gi | 115455083 | ref | NP_001051142.1 |
```

```
gi | 187936925 | ref | NP_001120691.1 |
gi | 323836227 | ref | XP_003313044.1 |
gi | 109106658 | ref | KP_001137225.1 |
gi | 62751809 | ref | NP_001015526.1 |
gi | 34850059 | ref | NP_899043.1 |
gi | 363734384 | ref | XP_421096.3 |
gi | 62122835 | ref | NP_001014345.1 |
gi | 15235708 | ref | NP_195491.1 |
gi | 115455083 | ref | NP_001051142.1 |
```

ELQEYLVFAKRRHLHVIVDEVYMLSVFEKS-VGYRSVLSL-----ERLPDPQRTTHVMWA  
 ELQEYLVFAKRRHLHVIVDEVYMLSVFEKS-VGYRSVLSL-----ERLPDPQRTTHVMWA  
 ELQEYLVFAKRRHLHVIVDEVYMLSVFEES-AGYRSVLSL-----ERLPDPQRTTHVMWA  
 ELQEYLEFAKRRHLHVMDEVYMLSVFEES-AGYRSVLSL-----ERLPDPQRTTHVMWA  
 ELQDPLRFAMRHLKHLVDEVYMLSVFEES-LGYRSVLSL-----ERLPDPQRTTHVMWA  
 ELRDYLEFAKRRHLHVIVDGIYMLSVFDES-ATFHSVLGM-----DRLPDQRTTHVMWG  
 EMTGFLQFAKMQHLHVIVDEIYMLSVFGKE-HTFRSVLSL-----DGLPDPQRTTHVMWG  
 ELNHLLDFTSRKKIHLISDEIYSGTVFTNP--GFISVMEVLKDKKLENTDVFDRVHIVYS  
 DLETIVDFVAAKGIHLISDEIYAGTAFAEPPAGFVSALEVAGRDGGADVSDRVHVVYS

TSKDFGMSGSLRFGLTYLENQDVATAVASLCRYHGLSGLVQYQMAQLLRDR-DWINQVYPL  
TSKDFGMSGSLRFGLTYLENQDVATAVASLCRYHGLSGLVQYQMAQLLRDR-DWINQVYPL  
TSKDFGMSGSLRFGLTYLENQDVATAVASLCRYHGLSGLVQYQMAQLLRDR-DWINQVYPL  
TSKDFGMSGSLRFGLTYLENQDVATAVASLCRYHGLSGLVQYQMAQLLRDH-DWINQVYPL  
TSKDFGMSGSLRFGLTYLENQDVATAVASLCRYHGLSGLVQYQMAQLLRDH-DWISQVYPL  
ISKDFAVSGIRFGLTYLENQDVANAVASLCRYHGVCGPVQHKAQLLRDRGDGSGVYPLR  
VSKDFAMAGMRVGTIYSENKDLVQALDQCGFHGVPQYQMAQLLRDR-DWLNSEFLP  
LSKDLGLPGFRVGVIIYSNDDFVVSAAATKMSSSFLISSQTQYLLSALLSDK--FTTKNYLE  
LSKDLGLPGFRVGAIVSANAAVVSAATKMSSSFLVSSQTQYLLAALLGDR--DFTTSYVA  
\*\*\*\*\*

ENHARLKAATHYVSEELRALGIPFLSRGAGFFIWDLRKYLPGKTFEEMLLWRRFL-DN  
ENHARLKAATHYVSEELRLGIPFLSRGAGFFIWDLRKYLPGKTFEEMLLWRRFL-DN  
ENHARLKAATHYVSEELRALGIPFLSRGAGFFIWDLRKYLPGKTFEEMLLWRRFL-DN  
ENHARLKAATHYVSEDLRALGIPFVSRGAGFFIWDLRKYLPEATFEEVLLWRRFL-EN  
ENHARLKAATHYVSEELRALGIPFVSRGAGFFIWDLRKYLPGKTFEEALLWRQFL-DN  
ANHARLKAATHYVTDCLKLGVFLNRNAGGFFIWDLRKYLRTGTTFEEMLLWRRFL-DN  
ENKRRLKAEHYLTELKKLDIPFPHRGAGFFIWDLSKFLKEKTFAEELCVWR CFL-KH  
ENQIRLKNRHKDLVGLSEAGTIGCLSNAGLFCWVDMHLLKNTPEAELELNKKIVYEV  
ENKRRIRKEHDLVDGLREIGTIGCLSNAGLFCWVDMHSLMRSSPAGELMKWVVFEV

VIENNA	VIENNA alignment format	vienna
--------	-------------------------	--------

Default value is: *Clustal [clustal]*

## Identifying and Aligning Homologs (Whitehead Institute)

# Muscle

EBI > Tools > Multiple Sequence Alignment > MUSCLE

## MUSCLE - Multiple Sequence Alignment

MUSCLE stands for **M**Ultiple **S**equence **C**omparison by **L**og- **E**xpectation. MUSCLE is claimed to achieve both better average accuracy and better speed than [ClustalW2](#) or [T-Coffee](#), depending on the chosen options.

### Use this tool

#### STEP 1 - Enter your input sequences

Enter or paste a set of sequences in any supported format

```
>gi|187936925|ref|NP_001120691.1| 1-aminocycl  
[Homo sapiens]  
MFTLPQKDFRAPTTCLGPTCMQDLGSSHGDELEGECSRKI  
KWFWDSAEEGYRTYHMDEYDEDKNPSGIINLGTSENKLCF  
LREEVAKFLSFYCKSPVPLRPENVVVLNGGASLFSALATVLC  
AYVYLDSEVTGLDTRPFQLTVEKLEMALREAHSEGKVKVG  
LHVIVDEVYMLSVFEKSVGYRSVLSLERLPDPQRTHVMWA
```

Or upload a file:  No file chosen

#### STEP 2 - Set your Parameters

OUTPUT FORMAT:

OUTPUT TREE  OUTPUT ORDER

#### STEP 3 - Submit your job

☐ Be notified by email (Tick this box if you want to be notified by email when the results are available)

EBI > Tools > Multiple Sequence Alignment > MUSCLE

## MUSCLE Results

### Result files

#### Input Sequences

[muscle-l20120517-195822-0596-64663385-pg.input](#)

#### Tool Output

[muscle-l20120517-195822-0596-64663385-pg.output](#)


#### Alignment in FASTA format

[muscle-l20120517-195822-0596-64663385-pg.fasta](#)

#### JalView

Identifying and Aligning Homologs (Whitehead Institute)

# T-Coffee




- Help
- T-Coffee website
- Jalview
- Programmatic Access
- Download

---

- Related Applications
  - Pairwise Sequence Alignment
  - Multiple Sequence Alignment
  - Phylogeny

---

**T-Coffee related literature** 

Search for T-Coffee related literature in Medline... [more](#)

EBI > Tools > Multiple Sequence Alignment > T-Coffee

**T-Coffee - Multiple Sequence Alignment**

T-Coffee is a multiple sequence alignment program. Its main characteristic is that it will allow you to combine results obtained with several alignment methods.

**Use this tool**

**STEP 1 - Enter your input sequences**

Enter or paste a set of sequences in any supported format:

Or upload a file:  No file chosen

**STEP 2 - Set your Parameters**

**MATRIX** **ORDER**

**STEP 3 - Submit your job**

☐ Be notified by email (*Tick this box if you want to be notified by email when the results are available*)

Identifying and Aligning Homologs (Whitehead Institute)

# M-Coffee

<http://tcoffee.crg.cat/>



## M-Coffee

*Aligns DNA, RNA or Proteins by combining the output of popular aligners*

### Sequences input

Paste or upload your set of sequences in FASTA format

#### Sequences to align

[Click here to use the sample file](#)

```
>gi|187936925|ref|NP_001120691.1| 1-aminocyclopropa  
synthase-like protein 1 [Homo sapiens]  
MFTLPQKDFRAPTTCLGPTCMQDLGSSHGEDLEGECSRKLDQKLPELRGVG  
KWFWDSEEGYRTYHMDYDEDEKNPSGIIINLGTSENKLCFDLLSWRLSQRD  
LREEVAKFLSFYCKSPVPLRPENVVVLNGGASLFSALATVLCAGEAFLIP  
AYVYLDSEVTGLDTRPFQLTVEKLEMALEAHSEGVKVKGLILISPQNPLGI  
LHVIVDEVYMLSVFEKSVGYRSVLSLERLPDPQTHVMWATSKDFGMSGLRI  
RYHGLSGLVQYQMAQLLRDRDWINQVYLPENHARLKAHTYVSEELRALGII
```

- OR - [Click here to upload a file](#)

[Hide advanced options](#)

### Alignment Computation

M-Coffee computes its alignments by combining a collection of Multiple Alignments named a Library. In this section you want to combine into the library. You can choose pairwise and multiple sequence alignment methods. The standard sequence alignment methods.

#### Pairwise Methods

- ☐ Mlalign\_id\_pair ☐ Mclustalw\_pair ☐ Mpoa\_pair ☐ Mprobe  
☐ Mproba\_pair ☐ Mmafft\_pair ☐ Mdialigntx\_pair ☐ Mslow\_  
☐ Mmuscle\_pair

#### Multiple Methods

- ☒ Mpcma\_msa ☒ Mmafft\_msa ☒ Mclustalw\_msa ☒ Mdial  
☒ Mmuscle\_msa ☒ Mprobcons\_msa ☒ Mt\_coffee\_msa

### Output options

Use this section to control the output format.

- Alignment format ☒ score\_html ☒ clustalw\_aln ☐ pir\_aln ☐ pir\_seq ☐ gcg  
☒ fasta\_aln ☐ score\_ascii ☐ msf\_aln ☒ phylip

Case

Residue number

outorder

Alignment length

Your email address

[Submit](#)

[Reset](#)

Identifying and Aligning Homologs (Whitehead Institute)



# M-Coffee

## M-Coffee alignment result

### MSA

The multiple sequence alignment result as produced by T-coffee.

T-COFFEE, Version 9.02.r1228 (2012-02-16 18:15:12 - Revision 1228 - Build 336)  
Cedric Notredame  
SCORE=93

\* **BAD AVG 6000**

```
gi|187936925|re : 95
gi|332836227|re : 96
gi|109106658|re : 94
gi|62751809|ref : 96
gi|34850059|ref : 96
gi|363734384|re : 86
gi|62122835|ref : 89
gi|15235708|ref : 88
gi|115455083|re : 90
cons : 93
```

```
gi|187936925|re : 1 ..... 0
gi|332836227|re : 1 ..... 0
gi|109106658|re : 1 ..... 0
gi|62751809|ref : 1 ..... 0
gi|34850059|ref : 1 ..... 0
gi|363734384|re : 1 ..... 0
gi|62122835|ref : 1 ..... 0
gi|15235708|ref : 1 ..... 0
gi|115455083|re : 1 ..... 0
cons : 1 ..... 63
```

```
gi|187936925|re : 1 ..... 0
gi|332836227|re : 1 ..... 0
gi|109106658|re : 1 ..... 0
gi|62751809|ref : 1 ..... 0
gi|34850059|ref : 1 ..... 0
gi|363734384|re : 1 ..... 0
gi|62122835|ref : 1 ..... 0
gi|15235708|ref : 1 ..... 0
gi|115455083|re : 1 ..... 0
cons : 1 ..... 64
```

```
gi|187936925|re : 1 ..... 0
gi|332836227|re : 1 ..... 0
gi|109106658|re : 1 ..... 0
gi|62751809|ref : 1 ..... 0
gi|34850059|ref : 1 ..... 0
gi|363734384|re : 1 ..... 0
gi|62122835|ref : 1 ..... 0
gi|15235708|ref : 1 ..... 0
gi|115455083|re : 1 ..... 0
cons : 1 ..... 127
```

```
gi|187936925|re : 1 ..... 0
gi|332836227|re : 1 ..... 0
gi|109106658|re : 1 ..... 0
gi|62751809|ref : 1 ..... 0
gi|34850059|ref : 1 ..... 0
gi|363734384|re : 1 ..... 0
gi|62122835|ref : 1 ..... 0
gi|15235708|ref : 1 ..... 0
gi|115455083|re : 1 ..... 0
cons : 1 ..... 127
```

<http://tcoffee.crg.cat/>

```
gi|187936925|re : 1 ..... 19
gi|332836227|re : 1 ..... 19
gi|109106658|re : 1 ..... 19
gi|62751809|ref : 1 ..... 19
gi|34850059|ref : 1 ..... 19
gi|363734384|re : 1 ..... 19
gi|62122835|ref : 1 ..... 19
gi|15235708|ref : 1 ..... 19
gi|115455083|re : 1 ..... 19
cons : 1 ..... 19
```

253

```
gi|187936925|re : 20 ..... 59
gi|332836227|re : 20 ..... 59
gi|109106658|re : 20 ..... 59
gi|62751809|ref : 20 ..... 59
gi|34850059|ref : 20 ..... 59
gi|363734384|re : 20 ..... 59
gi|62122835|ref : 20 ..... 59
gi|15235708|ref : 20 ..... 59
gi|115455083|re : 20 ..... 59
cons : 20 ..... 59
```

316

```
gi|187936925|re : 60 ..... 119
gi|332836227|re : 60 ..... 119
gi|109106658|re : 60 ..... 119
gi|62751809|ref : 60 ..... 119
gi|34850059|ref : 60 ..... 119
gi|363734384|re : 60 ..... 119
gi|62122835|ref : 60 ..... 119
gi|15235708|ref : 60 ..... 119
gi|115455083|re : 60 ..... 119
cons : 60 ..... 119
```

379

```
gi|187936925|re : 120 ..... 173
gi|332836227|re : 120 ..... 173
gi|109106658|re : 120 ..... 173
gi|62751809|ref : 120 ..... 173
gi|34850059|ref : 120 ..... 173
gi|363734384|re : 120 ..... 173
gi|62122835|ref : 120 ..... 173
gi|15235708|ref : 120 ..... 173
gi|115455083|re : 120 ..... 173
cons : 120 ..... 173
```

442

```
gi|187936925|re : 174 ..... 236
gi|332836227|re : 174 ..... 236
gi|109106658|re : 174 ..... 236
gi|62751809|ref : 174 ..... 236
gi|34850059|ref : 174 ..... 236
gi|363734384|re : 174 ..... 236
gi|62122835|ref : 174 ..... 236
gi|15235708|ref : 174 ..... 236
gi|115455083|re : 174 ..... 236
cons : 174 ..... 236
```

505

```
gi|187936925|re : 505 ..... 567
gi|332836227|re : 505 ..... 567
gi|109106658|re : 505 ..... 567
gi|62751809|ref : 505 ..... 567
gi|34850059|ref : 505 ..... 567
gi|363734384|re : 505 ..... 567
gi|62122835|ref : 505 ..... 567
gi|15235708|ref : 505 ..... 567
gi|115455083|re : 505 ..... 567
cons : 505 ..... 567
```

Identifying and Aligning Homologs (Whitehead Institute)

# Other Considerations

---

## Web vs Command Line

- More options to change parameters
- Process lots of alignments in one command

## Which web page to use?

- Alignment home page vs uniform interface

## Alternatives

- Use subsequences or subset of sequences

## Realign by hand

# Doing Lots of Alignments

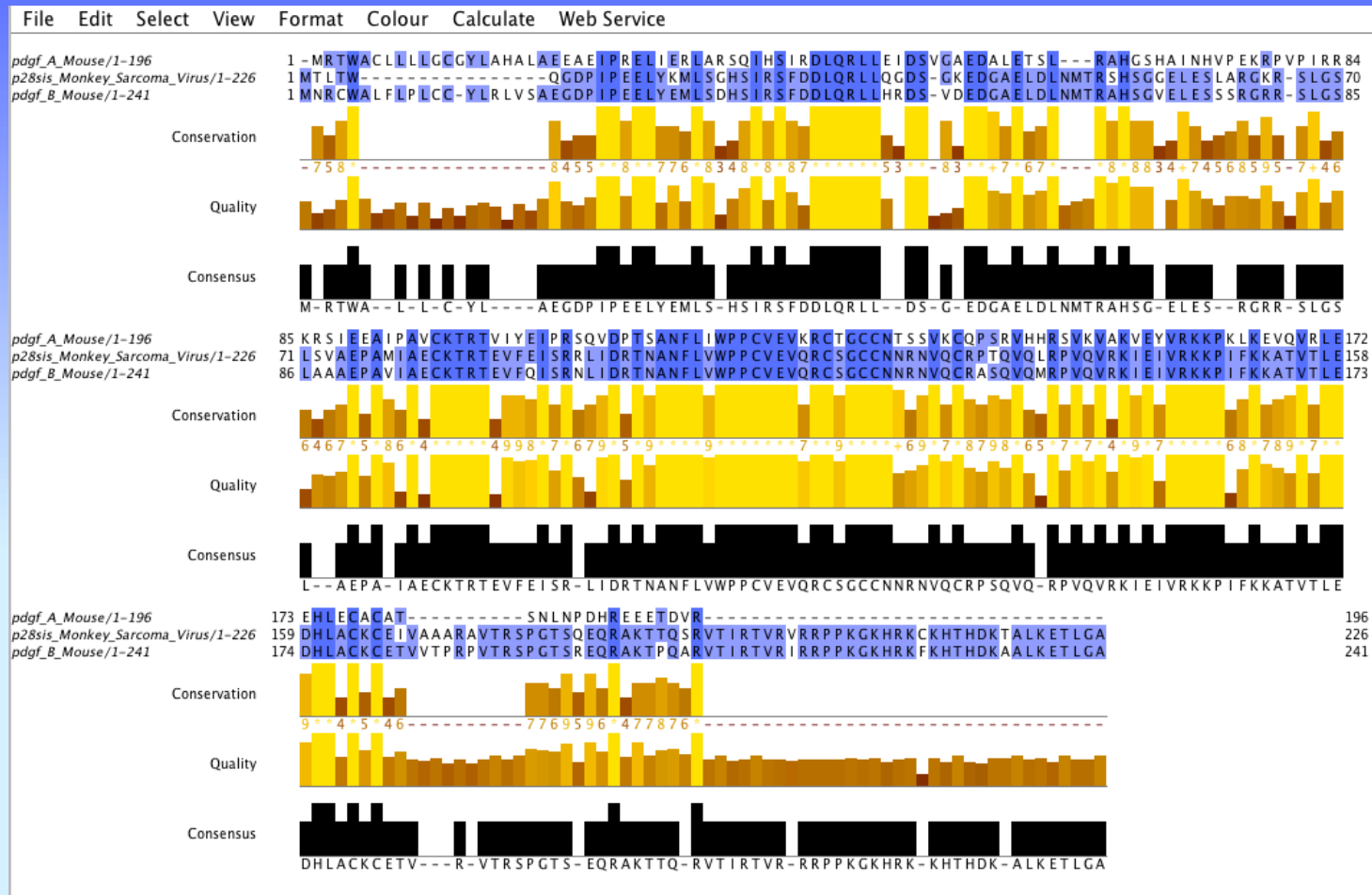
---

Our favorite method is to use the T-COFFEE suite (more specifically, M-Coffee) to run multiple alignment methods and then create a consensus alignment, a sort of a meta-alignment. This can be done with a single command like

```
t_coffee my_proteins.fa  
        -method=t_coffee_msa,mafft_msa,probcons_msa,muscle_msa  
        -output=fasta_aln
```

The final consensus alignment will appear in the file `my_proteins.fasta_aln`, which can then be viewed in JalView.

# Jalview (<http://www.jalview.org>)



# Identifying and Aligning Homologs (Whitehead Institute)

# Topics to Cover

---

- Introduction
- Pairwise Alignment methods
- Multiple sequence alignments
- Pre-computed alignment resources
  - Homologene (NCBI)
  - Ensembl (EBI)

# HOMOLOGENE: <http://www.ncbi.nlm.nih.gov/homologene>

NCBI HomoloGene Discover Homologs

Search: HomoloGene for 1-aminocyclopropane-1-carboxylate synthas Go Clear Save Search

Limits Preview/Index History Clipboard Details

Display: HomoloGene Show: 20 Send to: All: 1 Fungi: 0 Mammals: 0

**1: HomoloGene:86007. Gene conserved in Eukaryota** [Download](#) [Links](#)

**Genes**  
Genes identified as putative homologs of one another during the construction of HomoloGene.

- ACCSL\_H.sapiens 1-aminocyclopropane-1-carboxylate synthase homolog (Arabidopsis)(non-functional)-like
- ACCSL\_P.trogodytes 1-aminocyclopropane-1-carboxylate synthase homolog (Arabidopsis)(non-functional)-like
- ACCSL\_M.mulatta 1-aminocyclopropane-1-carboxylate synthase homolog (Arabidopsis)(non-functional)-like
- ACCSL\_B.taurus 1-aminocyclopropane-1-carboxylate synthase homolog (Arabidopsis)(non-functional)-like
- Accs1\_M.musculus 1-aminocyclopropane-1-carboxylate synthase homolog (Arabidopsis)(non-functional)-like
- Accs1\_R.norvegicus 1-aminocyclopropane-1-carboxylate synthase homolog (Arabidopsis)(non-functional)-like
- MGG\_07740\_M.oryzae hypothetical protein
- ACS2\_A.thaliana 1-aminocyclopropane-1-carboxylate synthase 2
- ACS1\_A.thaliana 1-aminocyclopropane-1-carboxylate synthase-like protein 1
- ACS6\_A.thaliana 1-aminocyclopropane-1-carboxylate synthase 6
- Os04g0578000\_O.sativa hypothetical protein

**Proteins**  
Proteins used in sequence comparisons and their conserved domain architectures.

- NP\_001027025.2 568 aa
- XP\_508381.2 568 aa
- XP\_001113751.1 568 aa
- XP\_002693605.2 576 aa
- NP\_001028624.2 580 aa
- NP\_001103064.1 617 aa
- XP\_367836.1 466 aa
- NP\_171655.1 496 aa
- NP\_191710.1 488 aa
- NP\_192867.1 495 aa
- NP\_001053637.1 483 aa

**Protein Alignments**  
Protein multiple alignment, pairwise similarity scores and evolutionary distances.

[Show Multiple Alignment](#)

[Show Pairwise Alignment Scores](#)

Pairwise alignments generated using BLAST

Regenerate Alignments

NP\_001027025.2 (H.sapiens) XP\_508381.2 (P.trogodytes)

**Conserved Domains**  
Conserved Domains from CDD found in protein sequences by rpsblast searching.

PLN02376 (PLN02376)  
1-aminocyclopropane-1-carboxylate synthase.

Aminotran\_1\_2 (pfam00155)  
Aminotransferase class I and II.

AAT\_I (cl00321)  
Aspartate aminotransferase (AAT) superfamily (fold type I) of pyridoxal phosphate (PLP)-dependent enzymes. PLP combines with an alpha-amino acid to form a compound called a Schiff base or aldimine intermediate which is essential for the reaction.

NCBI HomoloGene Discover Homologs

HOME SEARCH SITE MAP PubMed All Databases Human Genome

Search: HomoloGene for

**HomoloGene Downloader**

Homologene:75335. Gene conserved in Eukaryota

Download Protein sequences (in FASTA format)

Include 0 bp upstream of gene

Include 0 bp downstream of gene



Select which sequences should be included

[Select All](#) [Unselect All](#)

Species	Gene	mRNA	Protein
<input checked="" type="checkbox"/> H.sapiens	ACCS	NM_001127219.1	NP_001120691.1
<input checked="" type="checkbox"/> P.trogodytes	ACCS	XM_003312996.1	XP_003313044.1
<input checked="" type="checkbox"/> M.mulatta	ACCS	XM_001113725.2	XP_001113725.1
<input checked="" type="checkbox"/> C.lupus	ACCS	XM_849057.2	XP_854150.2
<input checked="" type="checkbox"/> B.taurus	ACCS	NM_001015526.1	NP_001015526.1
<input checked="" type="checkbox"/> M.musculus	Accs	NM_183220.2	NP_899043.1
<input checked="" type="checkbox"/> R.norvegicus	Accs	XM_001074505.2	XP_001074505.2
<input checked="" type="checkbox"/> G.gallus	ACCS	XM_421096.3	XP_421096.3
<input checked="" type="checkbox"/> D.reio	accs	NM_001014323.1	NP_001014345.1
<input checked="" type="checkbox"/> A.thaliana	ACS8	NM_119939.2	NP_195491.1
<input checked="" type="checkbox"/> A.thaliana	ACS4	NM_127846.1	NP_179866.1
<input checked="" type="checkbox"/> A.thaliana	ACS9	NM_114830.1	NP_190539.1
<input checked="" type="checkbox"/> A.thaliana	ACS5	NM_125977.2	NP_201381.1
<input checked="" type="checkbox"/> A.thaliana	ACS11	NM_116873.1	NP_567330.1
<input checked="" type="checkbox"/> O.sativa	Os03g0727600	NM_001057677.1	NP_001051142.1

Identifying and Aligning Homologs (Whitehead Institute)

# Ensembl (www.ensembl.org/)

[BLAST/BLAT](#) | [BioMart](#) | [Tools](#) | [Downloads](#)[BLAST/BLAT](#) | [BioMart](#) | [Tools](#) | [Downloads](#) | [Help & Documentation](#) | [Blog](#) | [Mirrors](#)

Human (GRCh37) ▼

Search Ensembl

New Search

Configure this page

Manage your data

Export data

Bookmark this page

Search:  for


e.g. [BRCA2](#) or [rat X:100000..200000](#) or [coronary heart disease](#)


### Browse a Genome


The Ensembl project produces genome databases for vertebrates and other eukaryotic species, and makes this information freely available online.

Click on a link below to go to the species' home page.

**Popular genomes** ([Log in to customize this list](#))

[H](#)  
GR

[M](#)  
NC

[Z](#)  
V

### Results Summary

Your search of Human with 'msh2' returned the following results:

By Feature type	
Total	550
► Domain	1
► Family	1
► Gene	1
► Somatic mutation	51
► Transcript	10
► Variation	485
► Variation phenotype	1

By Species	
Total	550
► Human	550

### Result in Detail

**1 Gene matches your query ('msh2') in Human**

**MSH2**

**Description** mutS homolog 2, colon cancer, nonpolyposis type 1 (E. coli) [Source:HGNC Symbol; Ensembl/Havana merge]

**Gene ID** [ENSG00000095002](#)

**Location** [2:47630108-47789450:1](#)

**Variations** [Variation Table](#)

**Source** e67

### Search for a DNA or protein sequence

**BLAT**

[a you want](#)

atabase, using the Perl API

[abases via FTP](#)

XL and other formats

[h BioMart](#)

ences or tables in text, html, o

our [FAQs](#) or [glossary](#)

All geno

Select

Institute)

# Ensembl

**e!Ensembl** east [BLAST/BLAT](#) | [BioMart](#) | [Tools](#) | [Downloads](#) | [Help & Documentation](#) | [Blog](#) | [Mirrors](#)

Human (GRCh37) ▼ Location: 2:47,630,108-47,789,450 Gene: MSH2

**Gene-based displays**

- Gene summary
- Splice variants (13)
- Supporting evidence
- Sequence
- External references
- Regulation
- Comparative Genomics
  - Genomic alignments
- Gene Tree (image)
  - Gene Tree (text)
  - Gene Tree (alignment)
- Orthologues (61)
- Paralogues
- Protein families (1)
- Phenotype
- Genetic Variation
  - Variation Table
  - Variation Image
  - Structural Variation
- External Data
  - Personal annotation
- ID History
  - Gene history

[Configure this page](#)

[Manage your data](#)

[Export data](#)

[Bookmark this page](#)

**Gene: MSH2** ENSG00000095002

**Description** mutS homolog 2, colon cancer, nonpolyposis type 1 (E. coli) [Source:HGNC Symbol;Acc:7325]

**Location** [Chromosome 2: 47,630,108-47,789,450](#) forward strand.

**Transcripts** ☐ This gene has 13 transcripts

Show **All** entries Show/hide columns Filter

Name	Transcript ID	Length (bp)	Protein ID	Length (aa)	Biotype	CCDS
MSH2-001	<a href="#">ENST00000233146</a>	3307	<a href="#">ENSP00000233146</a>	934	Protein coding	<a href="#">CCDS1834</a>
MSH2-002	<a href="#">ENST00000406134</a>	3628	<a href="#">ENSP00000384199</a>	921	Protein coding	-
MSH2-003	<a href="#">ENST00000454849</a>	560	<a href="#">ENSP00000411482</a>	140	Protein coding	-
MSH2-201	<a href="#">ENST00000394792</a>	2917	<a href="#">ENSP00000378272</a>	341	Protein coding	-
MSH2-202	<a href="#">ENST00000413880</a>	2537	<a href="#">ENSP00000402969</a>	720	Protein coding	-
MSH2-203	<a href="#">ENST00000419559</a>	2918	<a href="#">ENSP00000391195</a>	810	Protein coding	-
MSH2-204	<a href="#">ENST00000422810</a>	1184	<a href="#">ENSP00000389452</a>	168	Protein coding	-
MSH2-205	<a href="#">ENST00000432737</a>	2917	<a href="#">ENSP00000395752</a>	533	Protein coding	-
MSH2-206	<a href="#">ENST00000448533</a>	2918	<a href="#">ENSP00000415023</a>	410	Protein coding	-
MSH2-207	<a href="#">ENST00000453755</a>	2917	<a href="#">ENSP00000390362</a>	485	Protein coding	-
MSH2-208	<a href="#">ENST00000543555</a>	3018	<a href="#">ENSP00000442697</a>	868	Protein coding	-
MSH2-005	<a href="#">ENST00000461394</a>	416	No protein product	-	Processed transcript	-
MSH2-006	<a href="#">ENST00000467323</a>	313	No protein product	-	Processed transcript	-

**Transcript and Gene level displays**

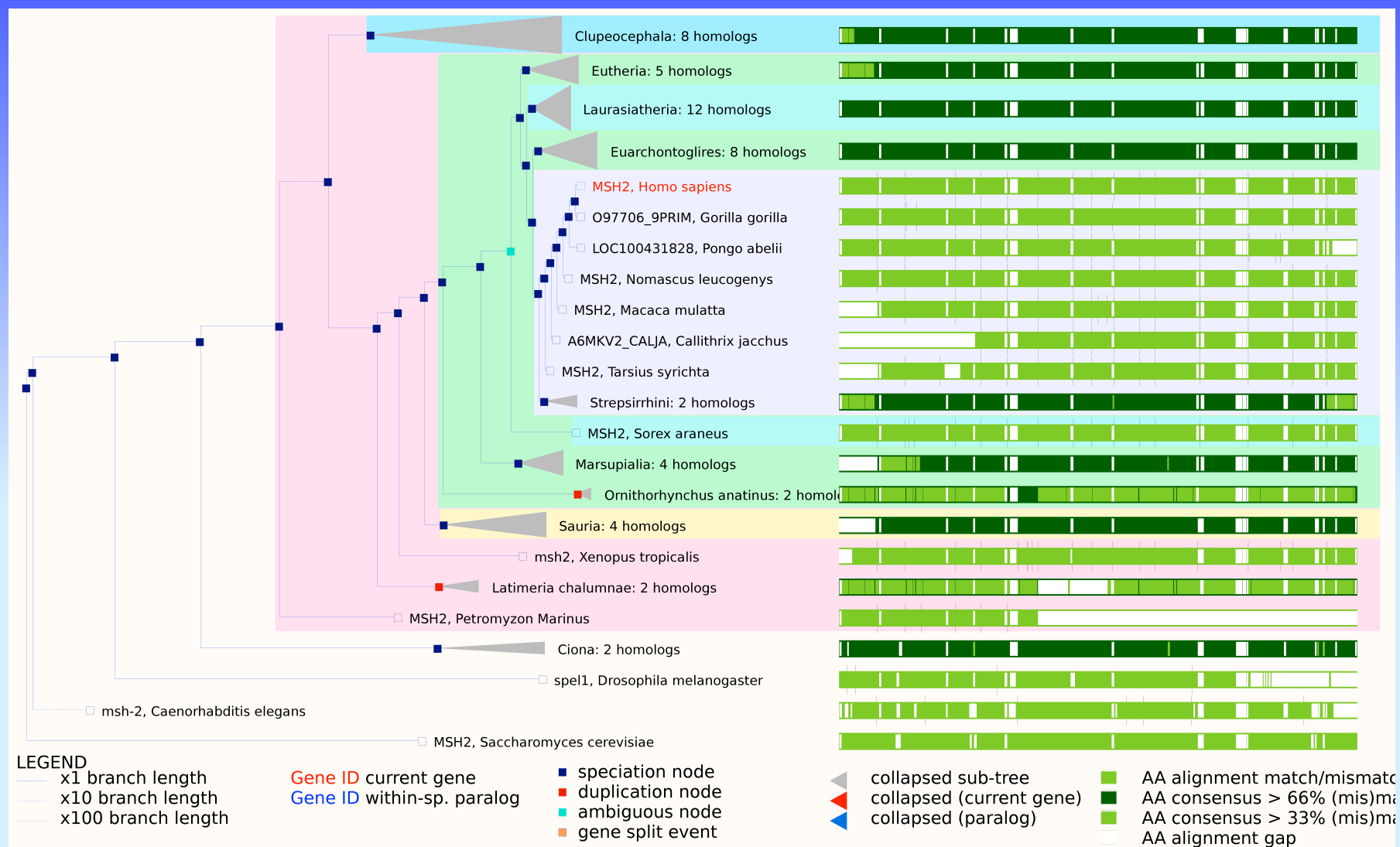
In Ensembl we provide displays at two levels:

- Transcript views which provide information specific to an individual transcript such as the cDNA and CDS sequences

Identifying and Aligning Homologs (Whitehead Institute)



# Genome Research 2008: PMID 19029536



# What I hope you learned

- What we can learn from sequence alignments
- Fundamentals of alignments
- Tools for building alignments

# References

---

1. Web links within handout plus
  - Local EMBOSS - <http://iona.wi.mit.edu/bio/tools/emboss/>
  - Local NCBI - <http://tak.wi.mit.edu/blast/>
2. SOPs - <https://gir.wi.mit.edu/trac/wiki/barc/SOPs>
3. What's on tak - <https://tak.wi.mit.edu/trac/wiki/Packages>
4. Help pages for each application
  - Click, click, click
5. Genome Technology - <http://jura.wi.mit.edu/bio/education/GT/>
6. Marketa Zvelebil and Jeremy O. Baum – Understanding Bioinformatics, Garland Science 2008

# Journal Articles

---

- ◆ Orthologs, paralog, and evolutionary genomics. *Annu Rev Genet.* 2005;39:309-38. PMID: 16285863
- ◆ Ensembl- *NAR* 2012, PMID: 22086963
- ◆ Homologene – *NAR* 2012, PMID: 22140104
- ◆ Clustal Omega – *Mol Syst Biol*, PMID: 21988835
- ◆ T-Coffee – *NAR* 2011, PMID: 21558174
- ◆ Challenges for MSA in hi-thruput era – *Bioinformatics* 2009, PMID: 19648142
- ◆ MSA - *Curr Opin Struct Biol.* 2006, PMID: 16679011