

**EBI web resources III:
Web-based tools in
Europe (EBI, ExPASy,
EMBOSS, DTU)**

Yanbin Yin

Fall 2015

Homework assignment 4

1. Download <http://cys.bios.niu.edu/yyin/teach/PBB/purdue.cellwall.list.lignin.fasta> to your computer
2. Select a C3H protein and a F5H protein from the above file and calculate the sequence identity between them using the Water server at EBI.
3. Perform a multiple sequence alignment using MAFFT with all FASTA sequences in the file
4. Built a phylogeny with the alignment using the "A la Carte" mode at <http://www.phylogeny.fr/>
5. Build another phylogeny starting from the unaligned sequences using the "one-click" mode at <http://www.phylogeny.fr/>; **if you encounter any error reports, try to figure out why and how to solve it** (hint: skip the Gblocks step).

Write a report (in **word or ppt**) to include all the operations, screen shots and the final phylogenies from step 3 and 4.

Due on 10/13 (send by email)

Office hour:

Tue, Thu and Fri 2-4pm, MO325A

Or email: yyin@niu.edu

Outline


- Hands on exercises!

Pairwise alignment (including database search) tools



To the bottom of the page

Bioinformatics services




Services

Research at EMBL-EBI



Research

Bioinformatics training




Training

Industry Programme




Industry

ELIXIR




European Coordination

EMBL Alumni Relations



EMBL ALUMNI

EMBL Alumni Programme

 [EMBO Practical Course on Analysis of High-Throughput Sequencing Data](#)
Oct 20 2014 -Oct 25 2014
Registration deadline: Aug 15 2014

 [Ensembl Browser Workshop - University of Colorado, Oct 2014](#)
Oct 23 2014

[See all courses and conferences](#)
[See other events at EMBL-EBI](#)

- EMBL-EBI 
- News
 - Brochures
 - Contact us
 - Intranet

Services

- By topic
- By name (A-Z)
- Help & Support

Research

- Overview
- Publications
- Research groups
- Postdocs & PhDs

Training

- Overview
- Train at EBI
- Train outside EBI
- Train online
- Contact organisers

Industry

- Overview
- Members Area
- Workshops
- SME Forum
- Contact Industry programme

About us

- Overview
- Leadership
- Funding
- Background
- Collaboration
- Jobs
- People & groups

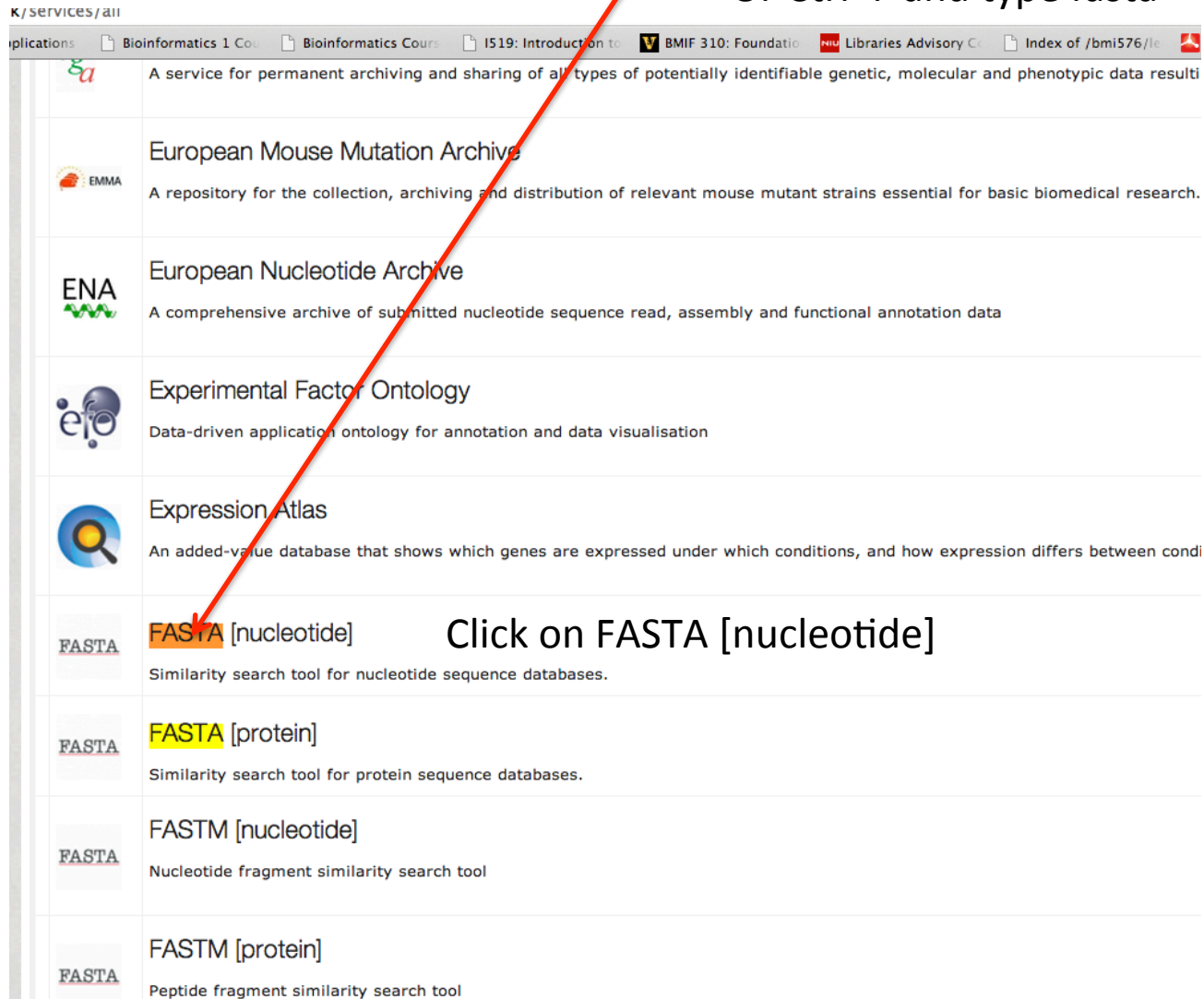


Click by names (A-Z)

We are gonna try FASTA tool










This is a very long list of tools
Scroll down to find FASTA

Or Ctrl+F and type fasta



κ/services/all

Applications: Bioinformatics 1 Cou Bioinformatics Cours IS19: Introduction to BMIF 310: Foundatio nu Libraries Advisory Co Index of /bmi576/le

	A service for permanent archiving and sharing of all types of potentially identifiable genetic, molecular and phenotypic data result
	European Mouse Mutation Archive A repository for the collection, archiving and distribution of relevant mouse mutant strains essential for basic biomedical research.
	European Nucleotide Archive A comprehensive archive of submitted nucleotide sequence read, assembly and functional annotation data
	Experimental Factor Ontology Data-driven application ontology for annotation and data visualisation
	Expression Atlas An added-value database that shows which genes are expressed under which conditions, and how expression differs between condi
	FASTA [nucleotide] Click on FASTA [nucleotide] Similarity search tool for nucleotide sequence databases.
	FASTA [protein] Similarity search tool for protein sequence databases.
	FASTM [nucleotide] Nucleotide fragment similarity search tool
	FASTM [protein] Peptide fragment similarity search tool

Click Genomes
We're gonna search Arabidopsis genome

FASTA

Protein **Nucleotide** Genomes Proteomes Whole Genome Shotgun Web services Help & Documentation

Tools > Sequence Similarity Searching > FASTA

Nucleotide Similarity Search

This tool provides sequence similarity searching against nucleotide databases using the FASTA suite of programs. FASTA provides a heuristic search with a nucleotide query against the DNA database for searching with a protein query. Optimal searches are available with SSEARCH (local), GGSEARCH (global) and GLSEARCH (global query).

STEP 1 - Select your databases

NUCLEOTIDE DATABASES

110 Databanks Selected

X Clear Selection

▼ ENA Sequence (formerly EMBL-Bank)

- ▶ ENA Sequence Release
- ENA Sequence Updates
- EMBL Coding Sequence

▶ Others

▶ IMGT

▶ Patents

STEP 2 - Enter your input sequence

Enter or paste a sequence in any supported format:

or upload a file: No file chosen

Genomes Similarity Search

This tool provides sequence similarity searching against complete genomes databases using the FASTA suite of programs. FASTA provides a heuristic search with a nucleotide translate the DNA database for searching with a protein query. Optimal searches are available with SSEARCH (local).

STEP 1 - Select your databases

GENOME DATABASES

0 Databank Selected X Clear Selection

- ▶ **Eukaryota**
- ▶ Archaea
- ▶ Bacteria
- ▶ Phage

Click on this little arrow

Choose Arabidopsis

Change here to protein

STEP 2 - Enter your input sequence

Enter or paste a **PROTEIN** sequence in any supported format:

or Upload a file: No file chosen

1 Databank Selected

▼ Eukaryota

- Anopheles gambiae str. PEST
- Arabidopsis thaliana
- Ashbya gossypii ATCC 10895
- Aspergillus fumigatus Af293
- Aspergillus nidulans FGSC A4

STEP 3 - Set your parameters

PROGRAM

TFASTX

The default settings will fulfill the needs of most users and, for that reason, are not visible.

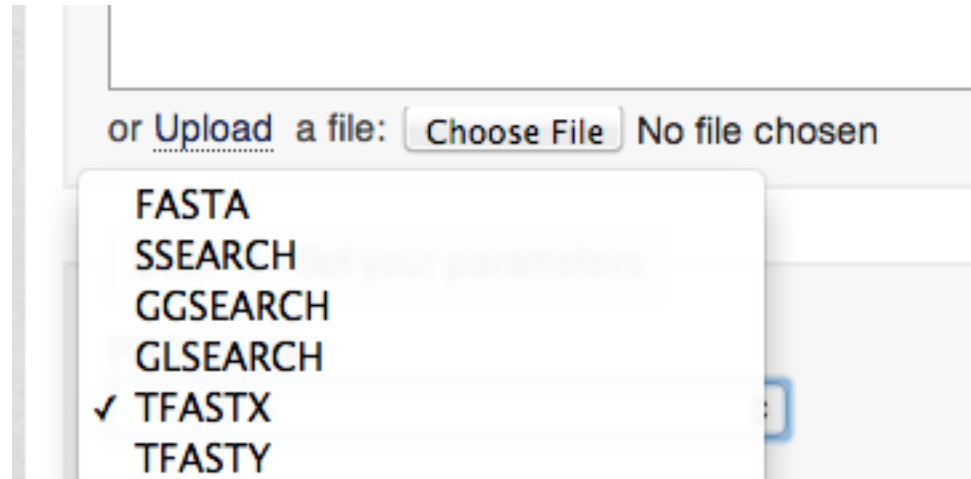
(Click here, if you want to view or change the default settings.)

STEP 4 - Submit your job

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

Go to <http://cys.bios.niu.edu/yyin/teach/PBB/cesa-pr.fa>
and copy the first seq (CesA) and paste here

<http://www.ebi.ac.uk/Tools/sss/fasta/help/index-genomes.html>



Program Name	Description	Abbreviation
FASTA	Scan a protein or DNA sequence library for similar sequences.	fasta
SSEARCH	Compare a protein or DNA sequence to a sequence database using the Smith-Waterman algorithm.	ssearch
GGSEARCH	Compare a protein or DNA sequence to a sequence database using a global alignment (Needleman-Wunsch)	ggsearch
GLSEARCH	Compare a protein or DNA sequence to a sequence database with alignments that are global in the query and local in the database sequence (global-local).	glsearch
TFASTX	Compare a protein sequence to a DNA sequence database, calculating similarities with frameshifts to the forward and reverse orientations.	tfastx
TFASTY	Compare a protein sequence to a DNA sequence database, calculating similarities with frameshifts to the forward and reverse orientations.	tfasty

Tfastx: allow frame shift between codons

Tfasty: also allow frame shift within codons

Tolerate sequence errors

Good for finding pseudogenes

Should be finished very quickly

Graphical presentation of the output

Raw output (plain text)

Tools > Sequence Similarity Searching > FASTA
Results for job fasta-120140922-203900-0063-55691713-oy

Summary Table Tool Output Visual Output Submission Details

Selection:
Select All Invert
Clear

Apply to selection:

Annotations:
Show Hide

Alignments:
Show Hide

Entries:
Download in
fasta format

Tools:
Launch
Custal Omega

Align.	DB:ID	Source	Length	Score	Identities %	Positives %	E()
<input checked="" type="checkbox"/>	EM_PLN:AB016893	STD:Arabidopsis thaliana genomic DNA, chromosome 5, P1 clone:MYH9. <i>Cross-references and related information in:</i> Nucleotide sequences Genomes Protein families Literature Samples & ontologies Protein sequences	82390	3862	59.4	67.9	2.4E-78
<input checked="" type="checkbox"/>	EM_PLN:AB025637	STD:Arabidopsis thaliana genomic DNA, chromosome 5, P1 clone:MVP7. <i>Cross-references and related information in:</i> Nucleotide sequences Genomes Protein families Samples & ontologies Protein sequences	39645	4258	64.2	70.7	1.4E-75
<input checked="" type="checkbox"/>	EM_PLN:AL161595	STD:Arabidopsis thaliana DNA chromosome 4, contig fragment No. 91 <i>Cross-references and related information in:</i> Nucleotide sequences Genomes Protein families Macromolecular structures Samples & ontologies Protein sequences	198151	4660	70.0	74.7	7.3E-75
<input checked="" type="checkbox"/>	EM_PLN:AB006703	STD:Arabidopsis thaliana genomic DNA, chromosome 5, P1 clone:MRH10. <i>Cross-references and related information in:</i> Nucleotide sequences Genomes Protein families Literature Macromolecular structures Samples & ontologies Protein sequences	71522	2493	47.3	65.0	6.3E-64
<input checked="" type="checkbox"/>	EM_PLN:AL391142	STD:Arabidopsis thaliana DNA chromosome 5, BAC clone T10B6 (ESSA project)	33563	3814	58.8	74.7	5.8E-61

Show alignment

Show EMBL format of the subject (hit)

Results for job fasta-l20140922-203900-0063-55691713-oy

Summary Table **Tool Output** Visual Output Submission Details

Download Download in XML format Send to MView

We are at the raw output view

```
      60          70          80          90
AT2G2  ECAFPTCRPCYEYERREGNQACPQCTRYKRIKG-----
      : : : : : : : : : : : : : : : : : :
EM_PLN ECAFPVCRPCYEYERREGNQSCPQCKTRYKRIKGHTYLFQQVLTFV*SYVSDKGFLLFSG
      26680      26710      26740      26770      26800      26830

      100         110         120         130         140
AT2G2  SPRVEGDEEDDDIDDLEHEF-Y---GMDPEHVTE-AALYYMR-LNTGRGTDEVSHLYSA-
      : : : : : : : : : : : : : : : : : : : : : : : : : : : :
EM_PLN SPRVEGDEEDDGIDDLDFEFDYSRSGLESETFSRRNSEFDLA/LCSTWLTDSFVNLWRGG
      26860      26890      26920      26950      26980      27010

      150         160         170         180         190
AT2G2  ---SPGSEVPLLT--YCD-----EDSDMYSDRHALIVPPSTGLGNRVHHPFTDS---
      : : . . : : : : : : : : : : : : : : : : : : : : : :
EM_PLN WFIFSPITSTVLLSLFYCSFWLFF*DVEISSDSHALIVSPSPGHIHRVHQPHFPDPAGKF
      27040      27070      27100      27130      27160      27190

      200         210         220
AT2G2  -----FASIHTRPMVPQKDLTVYGYGSVAWKDRMEVWKKQQ
      . . : : : : : : : : : : : : : : : : : : : : : :
EM_PLN CFLIWGDVFHLLLLFVVTKLLVCYVTAHPRPMVPQKDLAVYGYGSVAWKDRMEEWKRKQ
      27220      27250      27280      27310      27340      27370

      230         240         250         260
AT2G2  IEKLQVVKNERVND-GDGDGFIVDELD-DPGLPM-----MDEGRQP
      : : : : : . . : : : : : : : : : : : : : : : : :
EM_PLN NEKYQVVKHGDGSSLGDDADIPM*DNDPNTVFCHVSLLCFDIYEFCCSCFRMDEGRQP
      27400      27430      27460      27490      27520      27550
```

In the alignment, look for
/ frame shift
\ frame shift
* stop codon

In order to align the query protein to the subject genomic DNA, reading frames have to move 1 or 2 base ahead (1 base insertion or 2 base insertion)

BLAST gives shorter alignment because its alignment breaks where it sees frame shifts

Download [GenBank](#) [Graphics](#) Sort by: E value

Arabidopsis thaliana chromosome 1, complete sequence

Sequence ID: [ref|NC_003070.9|](#) Length: 30427671 Number of Matches: 13

Range 1: 8335055 to 8335714 [GenBank](#) [Graphics](#)

Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
303 bits(777)	3e-149	Compositional matrix adjust.	140/220(64%)	169/220(76%)	28/220(12%)	+2

Features: [putative mannan synthase 3](#)
[putative mannan synthase 3](#)

Query	228	VNSDECLLTRMQEMSLDYHFTVEQEVGSSSTHAFFGFNGTAGIWRIAAINEAGGWKDRITV	287
Sbjct	8335055	VNANECLMTRMQEMSLNYHFVAEQESGSSIHAFVGFNGTAGVWRIAAALNEAGGWKDRITV	8335234
Query	288	EDMDLAVRASLRGWKFLYLGLD-----QVKSELPSTF	319
Sbjct	8335235	EDMDLAVRA L GWKF+Y+ D+ QVK+ELPSTF	8335414
Query	320	RAFRFQQHRWSCGPANLFRKMVMEIVRNKKVRFWKKVYVIYSFFFVRKIIAHWVTFCFYC	379
Sbjct	8335415	KAYRFQQHRWSCGPANLWRKMTMEILQNKVSAWKKLYLIYNFFFIRKIVVHIFTFVFC	8335594
Query	380	VVLPLTILVPEVKVPIWGSVYIPSIITILNSVGTTPRSIHL	419
Sbjct	8335595	LILPTTVLFPPELVQPKWATVYFPTTITILNAIATPR*QHL	8335714

Range 2: 8334112 to 8334624 [GenBank](#) [Graphics](#)

Next Match Previous Match First Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
137 bits(345)	3e-149	Compositional matrix adjust.	75/171(44%)	106/171(61%)	34/171(19%)	+1

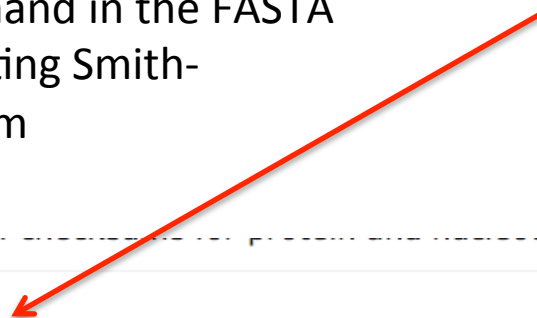
Features: [putative mannan synthase 3](#)
[putative mannan synthase 3](#)

Query	12	ETFDGV-RMEITGQLGMIWELVKAPVIVPLLQLAVYICLLMSVMLLCERVYMGIVIVLVK	70
Sbjct	8334112	DTTDGVVRSIIGEIIYIWKQTRIFVFIPIKCLVTICLVMSLLLFIERVYMSIVVVFVK	8334291
Query	71	LFWKKPKDKRYKFEPIDHDE-ELGSSNFPVVLVQIPMFNEREVY-----	112
Sbjct	8334292	L + P+K +K+EPI+DD+ EL ++N+P+VL+QIPM+NE+EV	8334471




SSEARCH is a command in the FASTA package implementing Smith-Waterman algorithm

Go back to the tool A-Z page:
<http://www.ebi.ac.uk/services/all>

Ctrl+F and type ssearch



Can only do pr-pr or nt-nt search

	SSEARCH [nucleotide] Smith-Waterman sequence similarity search against nucleotide databases.
	SSEARCH [protein] Smith-Waterman sequence similarity search against protein databases.
	Systems Biology Ontologies Controlled vocabularies and ontologies for problems in systems biology.
	T-Coffee Consistency-based multiple sequence alignment tool that attempts to mitigate the pitfalls of progressive ali

Ssearch can only do protein-protein or nucleotide-nucleotide searches
Slower but most accurate

Go back to the tool A-Z page:
<http://www.ebi.ac.uk/services/all>

Ctrl+F and type emboss



EBI Search

The EBI Search presents search results that are up-to-date with the data resources and provides an easy inter-domain



EMBOSS Tools

Selected **EMBOSS** tools for sequence analysis, providing: pairwise sequence alignment, sequence format conversion, sequence statistics.



ENA Sequence Search

EMBOSS contain hundreds of computer programs for sequence analysis



ENA/SVA - Sequence Version Archive

Historical repository of EMBL-Bank entries

EMBOSS: European Molecular Biology Open Software Suite

EMBOSS: The European Molecular Biology Open Software Suite (2000)
Rice, P., Longden, I. and Bleasby, A.
Trends in Genetics 16, (6) pp276--277

Needleman-wunsch algorithm
Smith-Waterman algorithm

Let's try needle first

Tools > EMBOSS Programs

Selected EMBOSS tools for sequence analysis

Pairwise Sequence Alignment

Needle ?

Create an optimal global alignment of two sequences using the Needleman-Wunsch algorithm

[Protein](#) [Nucleotide](#)

Stretcher ?

Improved version of the Needleman-Wunsch algorithm that allows larger sequences to be globally aligned

[Protein](#) [Nucleotide](#)

Water ?

Use the Smith-Waterman algorithm to calculate the local alignment of two sequences

[Protein](#) [Nucleotide](#)

Sequence Statistics

Pepinfo ?

Create a variety of plots that display different amino acid properties, such as hydrophathy or charged residues, and their position in the sequence

[Launch Pepinfo](#)

Pepstats ?

Calculate properties of protein sequences such as molecular weight

[Launch Pepstats](#)

Pepwindow ?

Draw a hydropathy plot for protein sequences

[Launch Pepwindow](#)

Cpgplot ?

[Launch Cpgplot](#)

Equivalent to the `bl2seq` command of the BLAST package

Global vs local alignment:

- in a local alignment, you try to match your query with a substring (a portion) of your subject (reference)
- in a global alignment you perform an end to end alignment with the subject

Local Alignment

```
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'  
      |||  |||||  |||||  
5' TACTCACGGATGAGGTACTTTAGAGGC 3'
```

Global Alignment

```
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'  
      |||||  |||||  |||||  
5' ACTACTAGATT----ACGGATC--GTACTTTAGAGGCTAGCAACCA 3'
```


Go to <http://cys.bios.niu.edu/yyin/teach/PBB/cesa-pr.fa>

Copy & paste Cesa

Copy & paste Csla

STEP 1 - Enter your protein sequences

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



Or, upload a file: No file chosen

AND

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



Or, upload a file: No file chosen

STEP 2 - Set your pairwise alignment options

The default settings will fulfill the needs of most users and, for that reason, are not visible.

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

Csla: 539 aa

Cesa: 1089 aa

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)*

Not a database search, so no E-value is reported

This is needle output

```
#=====  
#  
# Aligned_sequences: 2  
# 1: cesA  
# 2: cs1A  
# Matrix: EBLOSUM62  
# Gap_penalty: 10.0  
# Extend_penalty: 0.5  
#  
# Length: 1161  
# Identity: 125/1161 (10.8%)  
# Similarity: 219/1161 (18.9%)  
# Gaps: 700/1161 (60.3%)  
# Score: 49.5  
#  
#  
#=====  
cesA 1 MNTGGRLIAGSHNRNEFVLINADDTARIRSABELSGQTCKICRDEIELTD 50  
cs1A 1 ----- 0  
cesA 51 NGEPFACNECAFPTCRPCYEYERREGNQACPQCGTRYKRIKGS PRVEGD 100  
cs1A 1 ----- 0  
cesA 101 EEDDDIDDLEHEFYGMDPEHVTEAALYYMRLNTGRGTDEVSHLYSASPGS 150  
cs1A 1 ----- 0  
cesA 151 EVPLLTyceDSDMYSDRHALIVPPSTGLGNRVHVPFTDSFASIHTRPM 200  
cs1A 1 ----- 0  
cesA 201 VPQKDLTVYGYGSVAWKDRMEVWKKQQIEKLQVVKNERVNDGDGDGFIVD 250  
cs1A 1 ----- 0  
cesA 251 ELDDPGLPMMDEGRQPLSRKLPPIRSSRINPYRMLIFCRLAILGLFFHYRI 300  
cs1A 1 ----- 0  
cesA 301 LHPVNDAFGLWLTSVICEIWFVAVSWILDQFPKWYPIERETYLDRLSLRYE 350  
cs1A 1 ----- 0  
cesA 351 KEGKPSSELAPVDVVFVSTVDPLKEPPLITANTVLSILAVDYPVEKVACYVS 400  
cs1A 1 ----- 0  
cesA 401 DDGAAML-----TFEA--LSYTAEFARKW-----VP-----FCKK 428  
 .||:.. .||: .:..|:.....| | | .|..  
cs1A 1 MDGVSPKfVLPETFDGVRMEITGQLGMIWELVKAPVIVPLLQLAVYICLL 50
```

- gap
. Negative score
: positive score
| identical

This is different from what BLAST shows the alignment

```

#=====
#
# Aligned_sequences: 2
# 1: cesA
# 2: cslA
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 207
# Identity:      38/207 (18.4%)
# Similarity:   76/207 (36.7%)
# Gaps:         48/207 (23.2%)
# Score: 73.5
#
#=====

```

This is water output

The best way to find the optimally aligned regions and calculate the similarity between two sequences

```

cesA      779 GWIYGSVTEDILTGFKMHCHGWRSVYCMKRAAFKGSAPINLSDRLHQVL      828
      ||.....||:.....||::|.....|...|.....|..
cslA      280 GWKDRTTVEDMDLAVRASLRGWKFLYLGDLQV--KSELPSTFRAFRFQQH      327

cesA      829 RWALGSVEIF-----LSRHCPIWYGYGGGLKW-----LERFSYINSVVY      867
      ||:|...:|      :.:.....:      |      :..|.....:..
cslA      328 RWSCGPANLFRKMVMEIVRNKKVRF-----WKKVYVIYSFFFVRKIIA      370

cesA      868 PWTSLPLLVYCSLPAICLLTGKFIVPEISNYAGILFLLMFMSIAVTGILE      917
      .|      :.:|.....:||  .:|||:..      :...|:..:|:..
cslA      371 HW-----VTFCFYCVVLPLT--ILVPEVK-----VPIWGSVYIPSIIT      406

cesA      918 MQWGKIGIDDWRNEQFWVI--GGVSSH-----LFALFQGLLKVLGVST      960
      :  ...:|.....:|:~  ..:|.  |..||:      ||:..
cslA      407 I-LNSVGTTPRSIHLLFYWILFENVMSLHRTKATLIGLFE-----AGRAN      449

cesA      961 NFTVTSK      967
      .:.||:|
cslA      450 EWVVTAK      456

```

```

#-----
#-----

```

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

New DELTA-BLAST, a more sensitive protein-protein search

BLAST Assembled Genomes

Find Genomic BLAST pages:

Enter organism name or id--completions will be suggested

- [Human](#)
- [Rabbit](#)
- [Zebra](#)
- [Mouse](#)
- [Chimp](#)
- [Clawed](#)
- [Rat](#)
- [Guinea pig](#)
- [Arabi](#)
- [Cow](#)
- [Fruit fly](#)
- [Rice](#)
- [Pig](#)
- [Honey bee](#)
- [Yeast](#)
- [Dog](#)
- [Chicken](#)
- [Micro](#)

Basic BLAST


Choose a BLAST program to run.

- [nucleotide blast](#) Search a **nucleotide** database using a **nucleotide** query
Algorithms: blastn, megablast, discontinuous megablast
- [protein blast](#) Search **protein** database using a **protein** query
Algorithms: blastp, psi-blast, phi-blast, delta-blast
- [blastx](#) Search **protein** database using a **translated nucleotide** query
- [tblastn](#) Search **translated nucleotide** database using a **protein** query
- [tblastx](#) Search **translated nucleotide** database using a **translated nucleotide** query

Now try blast 2 seq

Specialized BLAST

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

- Make specific primers with [Primer-BLAST](#)
- Search [trace archives](#)
- 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](#)

Choose blastp

Go to <http://cys.bios.niu.edu/yyin/teach/PBB/cesa-pr.fa>

Copy & paste CesaA

Copy & paste Csla

NCBI/BLAST/blastp suite

Align Sequences Protein BLAST

blastn **blastp** blastx tblastn tblastx

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

Enter Query Sequence

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

Query subrange

From

To

Or, upload file No file chosen

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

From

To

Or, upload file No file chosen

Program Selection

Algorithm blastp (protein-protein BLAST)

Choose a BLAST algorithm

BLAST

Search protein sequence using Blastp (protein-protein BLAST)

Show results in a new window

This blast2seq output

Fragmented alignments

Download [Graphics](#) Sort by: E value

AT5G22740.1|AT5G22740.1|cslA

Sequence ID: lcl|253675 Length: 534 Number of Matches: 4

Range 1: 280 to 337 [Graphics](#)

▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
23.1 bits(48)	0.051	Compositional matrix adjust.	14/60(23%)	22/60(36%)	2/60(3%)

Query	779	GWYIGSVTEDILTGFKMHCHGWRSVYCMKRAAFKGSAPINLSDRLHQVLRWALGSVEIF	838
		GW + ED+ + GW+ +Y + K P Q RW+ G +F	
Sbjct	280	GWKDRTTVEDMDLAVRASLRGWKFLYLGLQV--KSELPSTFRAFQHRWSCGPANLF	337

Range 2: 360 to 376 [Graphics](#)

▼ Next Match ▲ Previous Match ▲ First Match

Score	Expect	Method	Identities	Positives	Gaps
17.3 bits(33)	3.4	Compositional matrix adjust.	8/17(47%)	9/17(52%)	4/17(23%)

Query	414	YTAEFARK----WVPFC	426
		Y+ F RK WV FC	
Sbjct	360	YSFFFVRKIIAHWVTF	376

Range 3: 168 to 189 [Graphics](#)

▼ Next Match ▲ Previous Match ▲ First Match

Score	Expect	Method	Identities	Positives	Gaps
16.5 bits(31)	5.8	Compositional matrix adjust.	5/22(23%)	11/22(50%)	0/22(0%)

Query	777	EIGWYIGSVTEDILTGFKMHCH	798
		+G+ G++ E + + HC	
Sbjct	168	RVGYKAGALKEGLKRSYVKHCE	189

Range 4: 454 to 472 [Graphics](#)

▼ Next Match ▲ Previous Match ▲ First Match

Score	Expect	Method	Identities	Positives	Gaps
15.8 bits(29)	9.4	Compositional matrix adjust.	7/19(37%)	10/19(52%)	0/19(0%)

Query	25	TARIRSAEELSGQTCKICR	43
		TA++ S + G T I R	
Sbjct	454	TAKLGSGQSAKNTKGIKR	472

Multiple sequence alignment tools

Foundation for many other further
analyses: phylogeny, evolution,
motif, protein family etc.

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

The MSA page shows nine tools and we're gonna try Clustal Omega, MAFFT and MUSCLE

Tools > 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 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. 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](#)

WebPRANK

The EBI has a new phylogeny-aware multiple sequence alignment program which makes use of evolutionary information to help place insertions and deletions. Try it out at [WebPRANK](#).

Clustal W and Clustal X version 2.0

MA Larkin, G Blackshields, [NP Brown](#), R Chenna... - ..., 2007 - Oxford Univ Press

Summary: The **Clustal W** and **Clustal X** multiple sequence alignment programs have been completely rewritten in C++. This will facilitate the further development of the alignment algorithms in the future and has allowed proper porting of the programs to the latest ...

Cited by 11069 Related articles All 28 versions Web of Science: 8296 Cite Save

CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment, position-specific gap penalties and weight matrix

JD Thompson, [DG Higgins](#), TJ Gibson - Nucleic acids research, 1994 - Oxford Univ Press

Abstract The sensitivity of the commonly used progressive multiple sequence alignment method has been greatly improved for the alignment of divergent protein sequences. Firstly, individual weights are assigned to each sequence in a partial alignment in order to ...

Cited by 47406 Related articles All 56 versions Web of Science: 40197 Cite Saved

[HTML] Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega

F Sievers, [A Wilm](#), [D Dineen](#), TJ Gibson... - Molecular systems ..., 2011 - msb.embopress.

Abstract Multiple sequence alignments are fundamental to many sequence analysis methods. Most alignments are computed using the progressive alignment heuristic. These methods are starting to become a bottleneck in some analysis pipelines when faced with ...

Cited by 806 Related articles All 16 versions Web of Science: 474 Cite Save More

MUSCLE: multiple sequence alignment with high accuracy and high speed

[RC Edgar](#) - Nucleic acids research, 2004 - Oxford Univ Press

Abstract We describe **MUSCLE**, a new computer program for creating multiple sequence alignments of protein sequences. Elements of the algorithm include fast distance estimation using kmer counting, progressive alignment using a new profile function we call the log-expectation ...

Cited by 9879 Related articles All 59 versions Web of Science: 7476 Cite Save

MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform

[K Katoh](#), K Misawa, K Kuma, T Miyata - Nucleic acids research, 2002 - Oxford Univ Press

Abstract A multiple sequence alignment program, **MAFFT**, has been developed. The CPU time is drastically reduced as compared with existing methods. **MAFFT** includes two novel techniques. (i) Homologous regions are rapidly identified by the fast Fourier transform (...

Cited by 2752 Related articles All 14 versions Web of Science: 2011 Cite Saved

MAFFT version 5: improvement in accuracy of multiple sequence alignment on CPU

[K Katoh](#), K Kuma, H Toh, T Miyata - Nucleic acids research, 2005 - Oxford Univ Press

Abstract The accuracy of multiple sequence alignment program **MAFFT** has been improved. The new version (5.3) of **MAFFT** offers new iterative refinement options, H-INS-i, F-INS-i and G-INS-i, in which pairwise alignment information are incorporated into objective function. ...

Cited by 2186 Related articles All 23 versions Web of Science: 1701 Cite Saved

Recent developments in the MAFFT multiple sequence alignment program

[K Katoh](#), H Toh - Briefings in bioinformatics, 2008 - Oxford Univ Press

Abstract The accuracy and scalability of multiple sequence alignment (MSA) of DNAs and proteins have long been and are still important issues in bioinformatics. To rapidly construct a reasonable MSA, we developed the initial version of the **MAFFT** program in 2002. MSA ...

Cited by 1465 Related articles All 15 versions Web of Science: 1079 Cite Save More

You can always check the help page

This is Clustal Omega page

Input form | Web services | Help & Documentation

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 of two sequences please instead use our [pairwise sequence alignment tools](#).

STEP 1 - Enter your input sequences

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

Go <http://cys.bios.niu.edu/yyin/teach/PBB/cesa-pr.fa>
and copy paste all the 9 protein seq here

Or, upload a file: **Choose File** No file chosen

Then submit

STEP 2 - Set your parameters

OUTPUT FORMAT **Clustal w/o numbers**

The default settings will fulfill the needs of most users and, for that reason, are not visible.

More options... *(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

Results for job clustalo-I20140924-162924-0015-60793060-oy

Alignments Result Summary Phylogenetic Tree Submission Details

Download Alignment File Show Colors Send to ClustalW2_Phylogen

This is called clustal format of MSA

Color AA based on chemical properties, e.g. acidic AA in blue

Check the evolutionary relatedness

Get text format summary of the results

CLUSTAL O(1.2.1) multiple sequence alignment

```

AT5G22740.1|AT5G22740.1|cs1A
AT2G24630.1|AT2G24630.1|cs1C
AT2G21770.1|AT2G21770.1|cesA
AT1G02730.1|AT1G02730.1|cs1D
os_42915|LOC_Os07g36610.1|cs1F
AT1G55850.1|AT1G55850.1|cs1E
AT4G23990.1|AT4G23990.1|cs1G
AT2G32530.1|AT2G32530.1|cs1B
os_25268|LOC_Os04g35020.1|cs1H

```

```

-----
-----
-----
MVKSAASQSPSPVTITVTPCKGSGDRSLGLTSPIPRASVITNQNSPLSSRATRRTSISGG
-----
-----
-----

```

```

AT5G22740.1|AT5G22740.1|cs1A
AT2G24630.1|AT2G24630.1|cs1C
AT2G21770.1|AT2G21770.1|cesA
AT1G02730.1|AT1G02730.1|cs1D
os_42915|LOC_Os07g36610.1|cs1F
AT1G55850.1|AT1G55850.1|cs1E
AT4G23990.1|AT4G23990.1|cs1G
AT2G32530.1|AT2G32530.1|cs1B
os_25268|LOC_Os04g35020.1|cs1H

```

```

-----
-----
-----
-----MNTGGRLIAGS
NRRSNGDEGRYCSMSVEDLTAETTNSCVLSYTVHIPPPTPDHQTVFASQSESEDEMLKGN
-----
-----
-----

```

```

AT5G22740.1|AT5G22740.1|cs1A
AT2G24630.1|AT2G24630.1|cs1C
AT2G21770.1|AT2G21770.1|cesA
AT1G02730.1|AT1G02730.1|cs1D
os_42915|LOC_Os07g36610.1|cs1F
AT1G55850.1|AT1G55850.1|cs1E
AT4G23990.1|AT4G23990.1|cs1G
AT2G32530.1|AT2G32530.1|cs1B
os_25268|LOC_Os04g35020.1|cs1H

```

```

-----
-----
HNRNEFVL-----INA----DDTARIRSAEELSGQTCKICRDEIELTDNGEPFIA
SNQKSFLSGTIFGGFKSVTRGHVIDCSMDRADPEKKSQGICWLKGC-----EKVVHG
-----
-----
-----

```

Results for job clustalo-I20140924-162924-0015-60793060-oy

Alignments Result Summary Phylogenetic Tree Submission Details

Input Sequences

clustalo-I20140924-162924-0015-60793060-oy.input

Tool Output

clustalo-I20140924-162924-0015-60793060-oy.output

Alignment in CLUSTAL format

clustalo-I20140924-162924-0015-60793060-oy.clustal

Phylogenetic Tree

clustalo-I20140924-162924-0015-60793060-oy.ph

Percent Identity Matrix

clustalo-I20140924-162924-0015-60793060-oy.pim

Jalview



```
(
(AT5G22740.1|AT5G22740.1|cs1A:0.28854,
AT2G24630.1|AT2G24630.1|cs1C:0.29272)
:0.19734,
(AT2G32530.1|AT2G32530.1|cs1B:0.30390,
os_25268|LOC_Os04g35020.1|cs1H:0.31325)
:0.03405)
:0.00339,
(AT2G21770.1|AT2G21770.1|cesA:0.27403,
AT1G02730.1|AT1G02730.1|cs1D:0.24931,
os_42915|LOC_Os07g36610.1|cs1F:0.28135)
:0.03248)
:0.04783,
(AT1G55850.1|AT1G55850.1|cs1E:0.30074,
AT4G23990.1|AT4G23990.1|cs1G:0.32201)
:0.03279);
```

Txt format to describe relatedness and can be visualized graphically as a tree graph

Both can be copy paste to notepad and save as plain text file

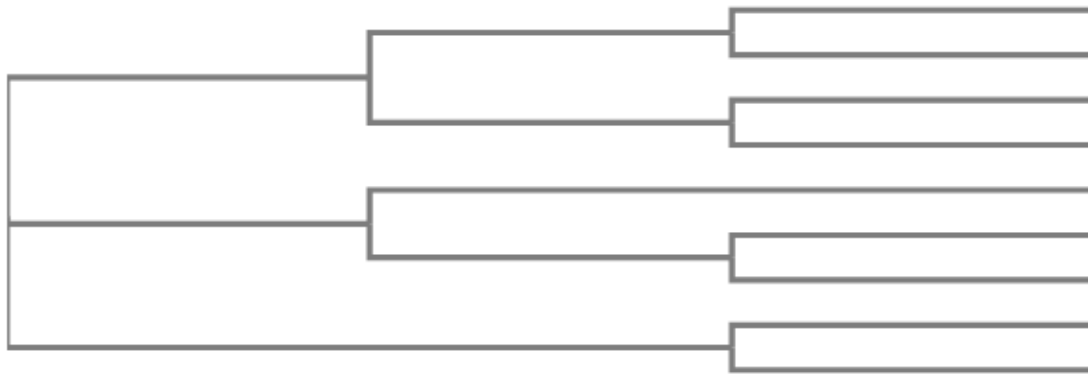
Matrix tells how similar each pair of seqs is

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

1:	AT5G22740.1	AT5G22740.1	cs1A	100.00	41.87	18.40	17.26	16.84	18.86	15.24	16.45	17.00
2:	AT2G24630.1	AT2G24630.1	cs1C	41.87	100.00	18.57	17.92	15.33	15.82	15.18	16.61	17.70
3:	AT2G21770.1	AT2G21770.1	cesA	18.40	18.57	100.00	47.11	38.52	35.00	32.64	32.63	32.96
4:	AT1G02730.1	AT1G02730.1	cs1D	17.26	17.92	47.11	100.00	46.93	33.76	30.81	30.80	32.18
5:	os_42915 LOC_Os07g36610.1	cs1F	16.84	15.33	38.52	46.93	100.00	31.82	25.98	29.66	30.97	
6:	AT1G55850.1	AT1G55850.1	cs1E	18.86	15.82	35.00	33.76	31.82	100.00	37.73	33.73	29.46
7:	AT4G23990.1	AT4G23990.1	cs1G	15.24	15.18	32.64	30.81	25.98	37.73	100.00	31.29	30.15
8:	AT2G32530.1	AT2G32530.1	cs1B	16.45	16.61	32.63	30.80	29.66	33.73	31.29	100.00	38.29
9:	os_25268 LOC_Os04g35020.1	cs1H	17.00	17.70	32.96	32.18	30.97	29.46	30.15	38.29	100.00	

Phylogram

Branch length: Cladogram Real



AT5G22740.1|AT5G22740.1|csiA 0.28854
AT2G24630.1|AT2G24630.1|csiC 0.29272
AT2G32530.1|AT2G32530.1|csiB 0.3039
os_25268|LOC_Os04g35020.1|csiH 0.31325
AT2G21770.1|AT2G21770.1|cesA 0.27403
AT1G02730.1|AT1G02730.1|csiD 0.24931
os_42915|LOC_Os07g36610.1|csiF 0.28135
AT1G55850.1|AT1G55850.1|csiE 0.30074
AT4G23990.1|AT4G23990.1|csiG 0.32201

You can always check the help page

This is MAFFT page



[Tools](#) > [Multiple Sequence Alignment](#) > MAFFT

Multiple Sequence Alignment

MAFFT (**M**ultiple **A**lignment using **F**ast **F**ourier **T**ransform) is a high speed multiple sequence alignment program.

STEP 1 - Enter your input sequences

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



Go <http://cys.bios.niu.edu/yyin/teach/PBB/cesa-pr.fa>
and copy paste all the 9 protein seq here

Or upload a file: Choose File No file chosen

STEP 2 - Set your Parameters

OUTPUT FORMAT Pearson/FASTA



Change here to ClustalW

The default settings will fulfill the needs of most users and, for that reason, are not visible.

More options... *(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

Results for job mafft-l20140924-181554-0300-30983931-es

Alignments

Result Summary

Guide Tree

Phylogenetic Tree

Submission Details

Download Alignment File

Show Colors

Send to ClustalW2_Phylogeny

Force the first M residue aligned

CLUSTAL format alignment by MAFFT L-INS-1 (v6.850b)

IDs were truncated

```
AT2G21770.1|AT2 M-----NTGGRLIAGSHNRNEFVLI
AT1G02730.1|AT1 MVKSAASQSPSPVTITVTPCKGSGDRSLGLTSPVIPRASVITNQNSPLSSRATRRTSISSG
os_42915|LOC_Os MA-----LSPAAAGRTG----
AT1G55850.1|AT1 MVN-----
AT4G23990.1|AT4 MYQ-----
AT2G32530.1|AT2 MA-----
os_25268|LOC_Os MA-----
AT5G22740.1|AT5 M-----
AT2G24630.1|AT2 MAPRFD FSDLWAKETRRGTPVVVKM-----
*
```

```
AT2G21770.1|AT2 N-----
AT1G02730.1|AT1 NRRSNGDEGRYCSMSVEDLTAETTNSCVLSYTVHIPPTPDHQT V FASQESEDEMLKGN
os_42915|LOC_Os -----
AT1G55850.1|AT1 -----
AT4G23990.1|AT4 -----
AT2G32530.1|AT2 -----
os_25268|LOC_Os -----
AT5G22740.1|AT5 -----
AT2G24630.1|AT2 -----
```

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 and faster alignment than other commonly chosen options.

STEP 1 - Enter your input sequences

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

Go <http://cys.bios.niu.edu/yyin/teach/PBB/cesa-pr.fa>
and copy paste all the 9 protein seq here



Or upload a file: No file chosen

STEP 2 - Set your Parameters

OUTPUT FORMAT:

Change here to ClustalW



The default settings will fulfill the needs of most users and, for that reason, are not visible.

(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)*

Tools > Multiple Sequence Alignment > MUSCLE

Results for job muscle-l20140924-181930-0252-89208085-pg

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

[Download Alignment File](#)
[Show Colors](#)
[Send to ClustalW2_Phylogeny](#)

CLUSTAL multiple sequence alignment by MUSCLE (3.8)

```

AT5G22740.1|AT5G22740.1|cs1A -----
AT2G24630.1|AT2G24630.1|cs1C -----MAPRFDSDLWAKETRRG-----
os_25268|LOC_Os04g35020.1|cs1H -----MAVVAAAAATGST-----
AT2G32530.1|AT2G32530.1|cs1B -----MADSSSSL-----
AT2G21770.1|AT2G21770.1|cesA -----MNTGGRLIAGSHNRNEFVLINADDTARI-----
AT1G02730.1|AT1G02730.1|cs1D MVKSAASQSPSPVTITVTPCKGSGDRSLGLTSPIPRASVITNQNSPLSSRATRRTSISSG
os_42915|LOC_Os07g36610.1|cs1F -----MALSPAAAGRTG-----
AT4G23990.1|AT4G23990.1|cs1G -----MYQVSLKQFVFLKIKSTTM-----
AT1G55850.1|AT1G55850.1|cs1E -----MVNKDDRI-----

```

```

AT5G22740.1|AT5G22740.1|cs1A -----
AT2G24630.1|AT2G24630.1|cs1C -----
os_25268|LOC_Os04g35020.1|cs1H -----
AT2G32530.1|AT2G32530.1|cs1B -----
AT2G21770.1|AT2G21770.1|cesA -RSAEELSGQTCKICRDEIELTDNGEPFIACNECAFPTC---RPCYEYERREGNQACPQC
AT1G02730.1|AT1G02730.1|cs1D NRRSNGDEGRYCSMSVEDLTAETTNSVCVLSYTVHIPPPTPDHQTVFASQESEDEMLKGN
os_42915|LOC_Os07g36610.1|cs1F -RNNNDAG-----
AT4G23990.1|AT4G23990.1|cs1G -----
AT1G55850.1|AT1G55850.1|cs1E -----

```

So which MSA tool should I use?

accuracy

MAFFT > Clustal Omega > MUSCLE >> ClustalW

speed

Table I BALiBASE results

Aligner	Av score (218 families)	BB11 (38 families)	BB12 (44 families)	BB2 (41 families)	BB3 (30 families)	BB4 (49 families)	BB5 (16 families)	Tot time (s)	Consistency
MSAprobs	0.607	0.441	0.865	0.464	0.607	0.622	0.608	12 382.00	Yes
Probalign	0.589	0.453	0.862	0.439	0.566	0.603	0.549	10 095.20	Yes
MAFFT (auto)	0.588	0.439	0.831	0.450	0.581	0.605	0.591	1475.40	Mostly (203/218)
Probcons	0.558	0.417	0.855	0.406	0.544	0.532	0.573	13 086.30	Yes
Clustal Ω	0.554	0.358	0.789	0.450	0.575	0.579	0.533	539.91	No
T-Coffee	0.551	0.410	0.848	0.402	0.491	0.545	0.587	81 041.50	Yes
Kalign	0.501	0.365	0.790	0.360	0.476	0.504	0.435	21.88	No
MUSCLE	0.475	0.318	0.804	0.350	0.409	0.450	0.460	789.57	No
MAFFT (default)	0.458	0.258	0.749	0.316	0.425	0.480	0.496	68.24	No
FSA	0.419	0.270	0.818	0.187	0.259	0.474	0.398	53 648.10	No
Dialign	0.415	0.265	0.696	0.292	0.312	0.441	0.425	3977.44	No
PRANK	0.376	0.223	0.680	0.257	0.321	0.360	0.356	128 355.00	No
ClustalW	0.374	0.227	0.712	0.220	0.272	0.396	0.308	766.47	No

The figures are total column scores produced using bali score on core columns only. The average score over all families is given in the second column. The results for BALiBASE subgroupings are in columns 3–8. The total run time for all 218 families is given in the second last column. The last column indicates whether the method is consistency based.

Molecular Systems Biology 7:539, 2011

<http://mafft.cbrc.jp/alignment/software/about.html>

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

Visualize alignment

[Tools](#) > [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. 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](#)

WebPRANK

The EBI has a new phylogeny-aware multiple sequence alignment program which makes use of evolutionary information to help place insertions and deletions. Try it out at [WebPRANK](#).

You can always check the help page

This is Mview page

Input form | **Web services** | **Help & Documentation**

Tools > Multiple Sequence Alignment > MView

A multiple alignment viewer

MView reformats the results of a sequence database search (BLAST, FASTA, etc) or a multiple alignment (MSF, PIR, CLUSTAL, etc) adding optional HTML markup to control colouring layout. MView is not a multiple alignment program, nor is it a general purpose alignment editor.

STEP 1 - Enter your input alignment

Enter or paste a sequence alignment in any supported format:

CLUSTAL O(1.2.1) multiple sequence alignment

AT5G22740.1|AT5G22740.1|csIA -----
AT2G24630.1|AT2G24630.1|csIC -----
AT2G21770.1|AT2G21770.1|csesA -----

Upload a file: No file chosen

STEP 2 - Set input parameters

INPUT FORMAT

STEP 3 - Set output parameters

The default settings will fulfill the needs of most users and, for that reason, are not visible.

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

STEP 4 - Submit your job

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

Go <http://cys.bios.niu.edu/yyin/teach/PBB/cesa-pr.faaln> and copy paste the MSA built above

MView reformats the results of a sequence database search (BLAST, FASTA, etc) or a multiple alignment (MSF, PIR, CLUSTAL, etc) **adding optional HTML mark-up to control colouring and web page layout**. MView is not a multiple alignment program, nor is it a general purpose alignment editor

401

```

1 AT5G22740.1|AT5G22740.1|cslA 100.0% -----
2 AT2G24630.1|AT2G24630.1|cslC 31.3% GKNAKQVTWVLLLKAHKAVGCLTWVATVFWWSLLGSKVRRRLSFTHPLGSERLGRDGLWFSAI--KLFLVA-----SLAI
3 AT2G21770.1|AT2G21770.1|cesA 7.9% IRSSRINPYRMLIFCRLLAIG---LFFHYRIL-----H---PVNDAFGLWLTSVICEIWFVAVSWILDQFPKW
4 AT1G02730.1|AT1G02730.1|cslD 7.2% VSAAIISPYRLLIALRLVALG---LFLTWRVR-----H---PNREAMWLGWMSSTCELWFALSLLDQLPKL
5 os_42915|LOC_Os07g36610.1|cslF 9.0% VSGVLLHPYRLLTLVRLIAVV---LFLAWRLK-----H---RDSAMWLLWVISIAGDFWFGVTVLLNQASKL
6 AT1G55850.1|AT1G55850.1|cslE 11.3% RRTGRVIAYRFFSASVFCVIC---LIWFYRIG-----EIGNRVTLDRLIWFVMPFIVEIWFGLYVWVTVQSSRW
7 AT4G23990.1|AT4G23990.1|cslG 9.1% HPCRRTIPYRIYAVFHTCGII---ALMYHHVH-----SL---LTANTTLITSLLLSDIVLAFMWATTTSLRY
8 AT2G32530.1|AT2G32530.1|cslB 9.4% YK--NYFLRVVDLTILGFLF---SLLLYRIL-----L---MNQNNVWVWVAFVLCESFFSFIWLLITSIKW
9 os_25268|LOC_Os04g35020.1|cslH 9.0% LGRRAAWAWRLAGLAVLLLLL---ALLALRLL-----RHH--GGAGGGGVVWRVALVCEAWFAALCALNVSARKW
consensus/100%
consensus/90%
consensus/80%
consensus/70%
h.....hhhh.h.hhhhhh .hhhhplh .....s.lh.h.....hhu.....th
htt.h..sarhh.hhhhhhlh hhhha+lh ... .stshhla.h.hhs-hhhuh.hhh...+h

```

481

```

1 AT5G22740.1|AT5G22740.1|cslA 100.0% -----
2 AT2G24630.1|AT2G24630.1|cslC 31.3% LAFELVAYRQW-HYFKNENLHIPT---SKLEIQSLLHLFYVVGWLSLRADYLAPEIKALSKFCIVLFLVQSV--DRLI-
3 AT2G21770.1|AT2G21770.1|cesA 7.9% YPIERETYLDRLSLRYEKEG---KPS---ELAP---V-DVFVSTVDLKEPPITANTVLSILAVDYPV--ERVAC
4 AT1G02730.1|AT1G02730.1|cslD 7.2% CPVNRLTDLGVLKERFESLENLRNEKGRS---DLPE---I-DVVFVSTADEKEPPITANTVLSILAVDYPV--ERLAC
5 os_42915|LOC_Os07g36610.1|cslF 9.0% NPVKRVPDLSLLRRRFFD---G---GLPE---I-DVFINTVDVDEPMIYTMNSLILATDYPA--DRHAA
6 AT1G55850.1|AT1G55850.1|cslE 11.3% NPVWRFPFSRLSRRYG-----S---DLPR---I-DVVFCTADEVIEPPILVVNTVLSVLTALDYPP--EKLAV
7 AT4G23990.1|AT4G23990.1|cslG 9.1% KPVRRTEYPEKYAA-EP-----E---DFPK---I-DVVFICTADYKEPPMVMVNTALSVAYEYPS--DKISV
8 AT2G32530.1|AT2G32530.1|cslB 9.4% SPASYKSYPERLD-----ERVH---DLPS---V-DMFVTADVREPPILVANTLLSLAVNYPA--NKLAC
9 os_25268|LOC_Os04g35020.1|cslH 9.0% SPVRFVTRPENLVAEGRTES-TTAAEYEG---ELPA---V-DMLVTTADALEPPITVNTVLSLALDYPRAGERLAC
consensus/100%
consensus/90%
consensus/80%
consensus/70%
.....t.....th.....h..hh..hhss...E.h.hhs.hh.lh.h.....p+h..
.....t.....th.....h..hh..hhss...E.h.hhs.hh.lh.h.....p+h..
.sh.h...phh.....t...-lst.l.shalsshshht.l.hshsphh.lhuh.hss -+lh
.hphhs..cth.....t -lss l thalsshshshhhssNollshhalsYps -+lus

```

561

```

1 AT5G22740.1|AT5G22740.1|cslA 100.0% -----
2 AT2G24630.1|AT2G24630.1|cslC 31.3% -----
3 AT2G21770.1|AT2G21770.1|cesA 7.9% -----
4 AT1G02730.1|AT1G02730.1|cslD 7.2% -----
5 os_42915|LOC_Os07g36610.1|cslF 9.0% -----
6 AT1G55850.1|AT1G55850.1|cslE 11.3% -----
7 AT4G23990.1|AT4G23990.1|cslG 9.1% -----
8 AT2G32530.1|AT2G32530.1|cslB 9.4% -----
9 os_25268|LOC_Os04g35020.1|cslH 9.0% -----
consensus/100%
consensus/90%
consensus/80%
consensus/70%

```

alcohol	=>	o	{ S, T }
aliphatic	=>	l	{ I, L, V }
aromatic	=>	a	{ F, H, W, Y }
charged	=>	c	{ D, E, H, K, R }
hydrophobic	=>	h	{ A, C, F, G, H, I, K, L, M, R, T, V, W, Y }
negative	=>	-	{ D, E }
polar	=>	p	{ C, D, E, H, K, N, Q, R, S, T }
positive	=>	+	{ H, K, R }
small	=>	s	{ A, C, D, G, N, P, S, T, V }
tiny	=>	u	{ A, G, S }
turnlike	=>	t	{ A, C, D, E, G, H, K, N, Q, R, S, T }

Another MSA visualization tool: ESPrnt <http://esprnt.ibcp.fr/ESPrnt/ESPrnt/>

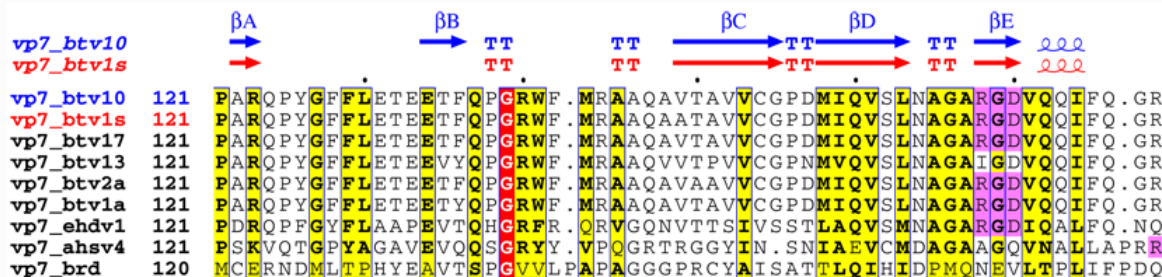
ESPrnt 3.0



NEWS
September 9, 2014: if you want to relax, the ESPrnt 'Rebcross' game is online!
• July 1, 2014: ESPrnt3 paper is published in the 2014 Web Server Issue of Nucleic Acids Research (freely accessible online).
• January 23, 2014: T-Coffee is now interfaced with ESPrnt
• December 23, 2013: we are pleased to announce the release of a major update: ESPrnt version 3.0!

What is ESPrnt?

- ESPrnt, 'Easy Sequencing in PostScript', is a program which renders sequence similarities and secondary structure information from aligned sequences for analysis and publication purpose.



Excerpt from a generated ESPrnt figure (full size in PDF)

- Key features:
 - ESPrnt is a utility, whose output is a PostScript / PDF / PNG or TIFF file of aligned sequences with graphical enhancements.
 - Its main input is a file of pre-aligned sequences in Clustal, FASTA, MultAlin, NPS@ or ProDom format.
 - The program calculates a similarity score for each residue of the aligned sequences.
 - Optional files allow further rendering.
- A typical ESPrnt figure shows:
 - aligned sequences,
 - similarities,
 - consensus,
 - accessibility,
 - hydropathy,
 - secondary structures elements,
 - intermolecular contacts,
 - user-supplied markers.

Click here to start

EScript 3.0

SUBMIT | **DISPLAY** | **MODE** | **LAYERS** | **SESSION** | **TIME** | **EXIT**
RESULTS | DOC | BEG | ADV | EXP | -1 | +1 | LOAD | SAVE | ██████████

ENDscript / EScript uses popup windows to display results - please be sure to disable popup blockers before submitting a job

Copy <http://cys.bios.niu.edu/yyin/teach/PBB/cesa-pr.fa.aln> and paste to a txt file and save it on your desktop, and upload to

Aligned Sequences

ALN file No file chosen
Example file • Tutorial

Number sequences

Secondary structure depiction

Top secondary structures | Parameters

Inputfile Upload a file below OR click here: **PDB**
 No file chosen

Chain ID | **Relative accessibility**

Sec. structure labels: $\alpha 1, \beta 1, \alpha 2, \beta 2, \dots$
Sec. structure labels: $\alpha A, \beta A, \alpha B, \beta B, \dots$
Sec. structure labels: $\alpha 1, \beta A, \alpha 2, \beta B, \dots$

After the file is uploaded

A new window popped out, view in PDF

The screenshot shows a web browser window titled "Results and logs - Google Chrome" with the URL `espript.ibcp.fr/ESPr ipt/temp/1611179935/reslauncher-1411615826.html`. The main content area is titled "RESULTS" and contains a message: "Here are the results of your ESPript job. You can now adjust settings on the main form and submit it again." Below this, there is a table of download options:

PostScript file	4 page(s) [104 Kb]
PDF file	pdf [18 Kb]

A red arrow points from the text "A new window popped out, view in PDF" to the "PDF file" row in the table. At the bottom of the page, there is a "Remarks" section with a citation: "When publishing data resulting from usage of this server, please use the following citation: Robert, X. and Gouet, P. (2014) 'Deciphering key features in protein structures with the new ENDscript server'. Nucl. Acids Res. 42(W1), W320-W324 - doi: 10.1093/nar/gku316 (freely accessible online)."

We just tried the very basic function. This web server has many more useful functions such as displaying secondary structures along with MSA. To learn more: http://esprict.ibcp.fr/ESPrict/ESPrict/esp_tutorial.php

	150	160	170	180
AT5G22740.1 AT5G22740.1 cs1A	TVK.QMVEV.....ECQRWASKGIN	TRYQIREN	RVGY....	KAGALKR
AT2G24630.1 AT2G24630.1 cs1C	SIQ.ELIRD.....EVTKWSQKGVN	IIRHRLV	RTGY....	KAGNLKSA
AT2G21770.1 AT2G21770.1 cesA	CDMD.....GNEIPR	LVYVSREK	PGFDHKK	KAGAMNSL
AT1G02730.1 AT1G02730.1 cs1D	EPVYGAEADAENLIDTTDVIDRLPM	LVYVSREK	PGYDHNK	KAGAMNAL
os_42915 LOC_Os07g36610.1 cs1F	EPQLGMPASSGHPLDFSAVDVRLPI	LVYIAREK	PGYDHQK	KAGAMNAQ
AT1G55850.1 AT1G55850.1 cs1E	..E.....GNTIAIPT	LVYLSREK	RPQHNNF	KAGAMNAL
AT4G23990.1 AT4G23990.1 cs1G	MDD.....TKKYIMP	LIYVSREK	SKVSSHFF	KAGALNTL
AT2G32530.1 AT2G32530.1 cs1B	VG.....VENEVPH	FVYISREK	PNYLHHY	KAGAMNFL
os_25268 LOC_Os04g35020.1 cs1H	ERNH	PT.....	IIVKTRV

	190	200	210	220	230	240
AT5G22740.1 AT5G22740.1 cs1A	YVVFDA	DF.QPEP	DFLR	RSIP	FLMH....	NPNIAL
AT2G24630.1 AT2G24630.1 cs1C	FVAIFD	ADF.QPNS	DFLK	LTVP	HFKE....	KPELGL
AT2G21770.1 AT2G21770.1 cesA	YLLNV	DCDHY	INNS	KAIR	EMCF	MMDP
AT1G02730.1 AT1G02730.1 cs1D	FILNL	DCDHY	IYNS	MALR	EGMC	FMLD
os_42915 LOC_Os07g36610.1 cs1F	FIFNF	DGDHY	INNS	QAFR	AA	LCF
AT1G55850.1 AT1G55850.1 cs1E	IILNL	DCDMY	ANNS	SKSTR	RDAL	LCIL
AT4G23990.1 AT4G23990.1 cs1G	IILT	LD	CDMY	SNDP	ATPV	RALC
AT2G32530.1 AT2G32530.1 cs1B	YMLNV	DCDMY	ANEAD	VVR	QAMC	IFLQ
os_25268 LOC_Os04g35020.1 cs1H	IMLN	MD	CDME	VNNP	QAVL	HAMC

ExPASy: Expert Protein Analysis System at SIB

Collection of external/internal tools

This website collect and classify web links to hundreds of bioinfo tools

SIB Fellowship ▼

ExPASy
Bioinformatics Resource Portal

Query all databases ▾

Visual Guidance

Categories

- proteomics
- genomics
- structural bioinformatics
- systems biology
- phylogeny/evolution
- population genetics
- transcriptomics
- biophysics
- imaging
- IT infrastructure
- drug design

Resources A..Z

Links/Documentation

ExpASy is the **SIB Bioinformatics Resource Portal** which provides access to scientific databases and software tools (i.e., *resources* in different areas of life sciences including proteomics, genomics, phylogeny, systems biology, population genetics, transcriptomics and *Categories* in the left menu). On this portal you find resources from many different SIB groups as well as external institutions.

Featuring today

The Systems Biology Research Tool

Software package and API for systems biologists.
[\[details\]](#)

How to use this portal?

- Features and updates
- New to ExpASy
- Experienced ExpASy users: what is different

Click on genomics, then sequence alignment

This page lists tools for sequence alignment

Visual Guidance

Categories

proteomics

genomics

sequence alignment

similarity search

characterisation/annotation

structural bioinformatics

systems biology

phylogeny/evolution

population genetics

transcriptomics

biophysics


imaging


IT infrastructure

drug design


Resources A..Z


Links/Documentation


 SIB resources


 External resources - *(No support from the ExPASy Team)*


Tools


 [Alignment tools](#) • Four tools for multiple alignments • [\[more\]](#)


 [boxshade](#) • MSA pretty printer • [\[more\]](#)


 [ClustalW](#) • Multiple sequence alignment • [\[more\]](#)


 [ClustalW - PBIL](#) • Multiple sequence alignment program • [\[more\]](#)


 [ClustalW2](#) • Multiple sequence alignment program • [\[more\]](#)


 [Codon Suite](#) • codon-based sequence analysis • [\[more\]](#)


 [Decrease redundancy](#) • Sequence redundancy reduction • [\[more\]](#)

 [DIALIGN](#) • Local multiple sequence alignment • [\[more\]](#)


 [GENIO/logo](#) • RNA/DNA & Amino Acid Sequence Logos • [\[more\]](#)


 [Kalign - EBI](#) • Fast and accurate multiple sequence alignment • [\[more\]](#)


 [Kalign - SBC](#) • Fast and accurate multiple sequence alignment • [\[more\]](#)


 [LALIGN](#) • Pairwise alignment • [\[more\]](#)


 [MADAP](#) • clustering for genome annotation data • [\[more\]](#)

 [MAFFT - CBRC](#) • Multiple sequence alignment • [\[more\]](#)


 [MAFFT - EBI](#) • Multiple sequence alignment • [\[more\]](#)


 [MaxAlign](#) • Gap removal from alignments • [\[more\]](#)

 [Multialin](#) • Multiple sequence alignment • [\[more\]](#)


 [MUSCLE](#) • Multiple alignment server • [\[more\]](#)


 [Newick Utilities](#) • high-throughput phylogenetic tree processing • [\[more\]](#)


 [Phylogibbs](#) • regulatory sites discovery • [\[more\]](#)

 [SIBsim4](#) • spliced sequence alignment • [\[more\]](#)

 [T-Coffee](#) • sequence and structure multiple alignments • [\[more\]](#)

 [T-Coffee - EBI](#) • Multiple sequence alignment program • [\[more\]](#)

 [T-Coffee - WUR](#) • Multiple sequence alignment program • [\[more\]](#)

 [WebLogo](#) • Sequence logos • [\[more\]](#)

We're gonna try



http://weblogo.berkeley.edu/logo.cgi

Upload the file that we downloaded from <http://cys.bios.niu.edu/yyin/teach/PBB/cesa-pr.fa.aln>



[about](#) · [create](#) · [examples](#)

Multiple Sequence Alignment

Upload Sequence Data: No file chosen

Image Format & Size

Image Format: Logo Size per Line:

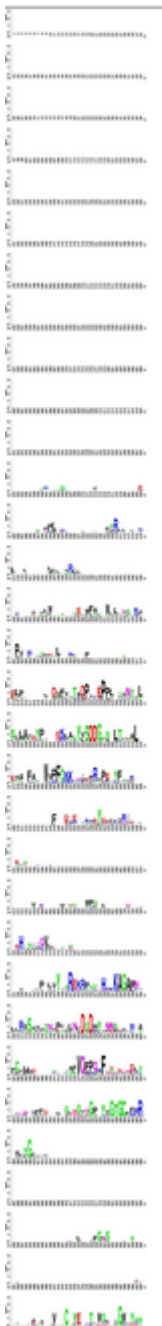
Advanced Logo Options

Sequence Type:	<input type="radio"/> amino acid <input type="radio"/> DNA / RNA <input checked="" type="radio"/> Automatic Detection	Logo Range: <input type="text"/>
First Position Number:	<input type="text" value="1"/>	Frequency Plot: <input type="checkbox"/>
Small Sample Correction:	<input checked="" type="checkbox"/>	
Multiline Logo (Symbols per Line):	<input checked="" type="checkbox"/> (<input type="text" value="32"/>)	

Advanced Image Options

Bitmap Resolution:	<input type="text" value="96"/> <input type="text" value="pixels/inch (dpi)"/>	Antialias Bitmaps: <input type="checkbox"/>
Title:	<input type="text"/>	Y-Axis Height: <input type="text"/>

Toggle this to allow logo shown in multiline

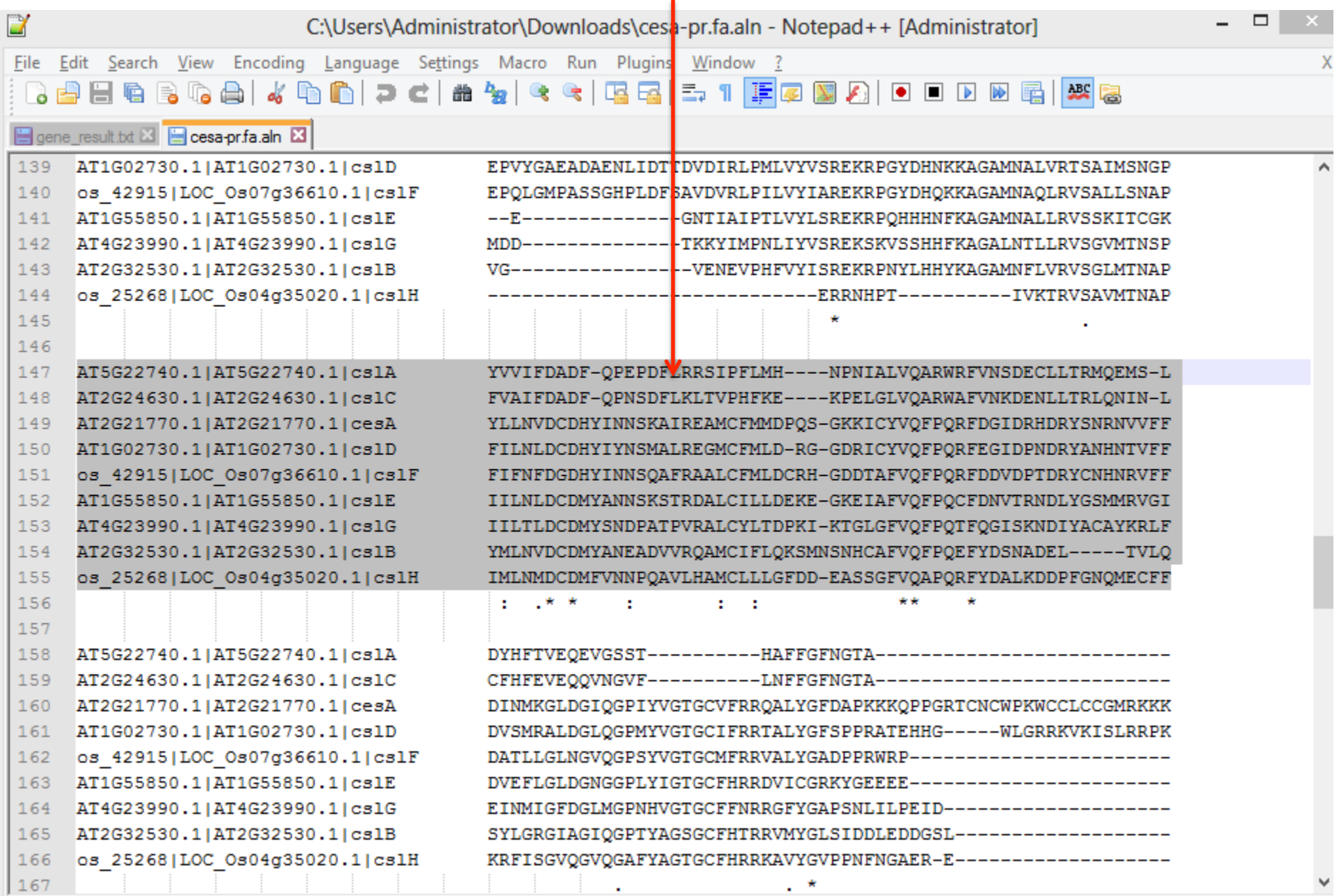


Click to increase



No need to use the entire alignment

You can also copy paste a segment of the alignment to weblogo



```
C:\Users\Administrator\Downloads\cesa-pr.fa.aln - Notepad++ [Administrator]
File Edit Search View Encoding Language Settings Macro Run Plugins Window ?
gene_result.txt cesa-pr.fa.aln
139 AT1G02730.1|AT1G02730.1|cs1D EPVYGAEDAENLIDT TDVDIRLPLMLVYVSREKRPGYDHNKKAGAMNALVRTSAIMSNP
140 os_42915|LOC_Os07g36610.1|cs1F EPQLGMPASSGHPLDF SAVDVRLPILVYIAREKRPGYDHQKKAGAMNAQLRVSALLSNAP
141 AT1G55850.1|AT1G55850.1|cs1E --E-----GNTIAIPTLVYLSREKRQHHHNFKAGAMNALLRVSSKITCGK
142 AT4G23990.1|AT4G23990.1|cs1G MDD-----TKKYIMP NLIYVSREKSKVSSHFFKAGALNTLLRVSGVMTNSP
143 AT2G32530.1|AT2G32530.1|cs1B VG-----VENEVPHFVYISREKRPNYLHHYKAGAMNFLVRVSGLMTNAP
144 os_25268|LOC_Os04g35020.1|cs1H -----ERRNHPT-----IVKTRVSAVMTNAP
145
146
147 AT5G22740.1|AT5G22740.1|cs1A YVVI FDADF-QPEPDFLRRSIPFLMH----NPNIALVQARWRVNSDECLLTRMQEMS-L
148 AT2G24630.1|AT2G24630.1|cs1C FVAIFDADF-QPNSDFLKLTVPHFKE----KPELGLVQARWAFVNKDENLLTRLQININ-L
149 AT2G21770.1|AT2G21770.1|cesA YLLNVDCDHYINNSKAIREAMCFMMDPQS-GKKICYVQFPQRFDGIDRHDYRNRRNVVFF
150 AT1G02730.1|AT1G02730.1|cs1D FILNLDCDHYIYNSMALREGMCFMLD-RG-GDRICYVQFPQRFEGIDPNDRYANHNTVFF
151 os_42915|LOC_Os07g36610.1|cs1F FIFNFDGDHYINNSQAFRAALCFMLDCRH-GDDTAFVQFPQRFDDVDPTDRYCNHNRVFF
152 AT1G55850.1|AT1G55850.1|cs1E IILNLDCDMYANNSKSTRDALCILLDEKE-GKEIAFVQFPQCFDNVTRNDLYGSMRVRGI
153 AT4G23990.1|AT4G23990.1|cs1G IILTLDCEMYSNDPATPVRALCYLTDPKI-KTGLGFVQFPQTFQGISKNDIYACAYKRLF
154 AT2G32530.1|AT2G32530.1|cs1B YMLNVDCDMYANEADVVRQAMCI FLQKSMNSNHCAFVQFPQEFYDSNADEL-----TVLQ
155 os_25268|LOC_Os04g35020.1|cs1H IMLNMDCEMFMVNNPQAVLHAMCLLLGFDD-EASSGFVQAPQRFYDALKDDPFGNQMECFE
156
157 : .* * : : : ** *
158 AT5G22740.1|AT5G22740.1|cs1A DYHFTVEQEVGSST-----HAFFGFNGTA-----
159 AT2G24630.1|AT2G24630.1|cs1C CFHFEVEQQVNGVF-----LNFFGFNGTA-----
160 AT2G21770.1|AT2G21770.1|cesA DINMKGLDGIQGPIYVGTGCVFRRQALYGFDA P K K K Q P P G R T C N C W P K W C C L C C G M R K K K
161 AT1G02730.1|AT1G02730.1|cs1D DVSMRALDGLQGPMYVGTGCI FRRTALYGFSPPRATEHHG-----WLGRRKVKISLRRPK
162 os_42915|LOC_Os07g36610.1|cs1F DATLLGLNGVQGPSYVGTGCMFRRVALY GADPPRWRP-----
163 AT1G55850.1|AT1G55850.1|cs1E DVEFLGLDGNNGPLYIGTGCFHRRDVICGRKYGEEEE-----
164 AT4G23990.1|AT4G23990.1|cs1G EINMIGFDGLMGP NHVGTGCFNRRGFY GAPS NLILPEID-----
165 AT2G32530.1|AT2G32530.1|cs1B SYLGRGIAGIQGPTYAGSGCFHTRRVMYGLSIDDLEDDGSL-----
166 os_25268|LOC_Os04g35020.1|cs1H KRFISGVQGVQGA FYAGTGCFHRRKAVYGVPPNFNGAER-E-----
167
. *
```

Paste the copied segment here



[about](#) · [create](#) · [examples](#)

WebLogo: a sequence logo generator

[GE Crooks](#), [G Hon](#), [JM Chandonia...](#) - [Genome ...](#), 2004 - [genome.cshlp.org](#)

Abstract **WebLogo** generates sequence logos, graphical representations of the patterns within a multiple sequence alignment. Sequence logos provide a richer and more precise description of sequence similarity than consensus sequences and can rapidly reveal ...

Cited by 3705 Related articles All 41 versions Web of Science: 2787 Cite Saved

Multiple Sequence Alignment

```

os_42915|LOC_Os07g36610.1|cs1F      FIFNFDGDHYINNSQAFRAALCFMLDCRH-
GDEAFVQFPQRFDDVDPTDRYCNHNRVFF
AT1G55850.1|AT1G55850.1|cs1E      IILNLDCDMYANNSKSTRDALCILLDEKE-
GKEIAFVQFPQCFDNVTRNDLYGSMRVGI
AT4G23990.1|AT4G23990.1|cs1G      IILTLDCDMYSNDPATFVRALCYLTDPKI-
KTGLGFVQFPQTFQGISKNDIYACAYKRLF
AT2G32530.1|AT2G32530.1|cs1B
YMLNVDCDMYANEADVVRQAMCIFLQKSMNSNHCAFVQFPQEFYDSNADEL-----TVLQ
os_25268|LOC_Os04g35020.1|cs1H      IMLNMDCFMVFNNPQAVLHAMCLLLGFDD-
EASSGFVQAPQRFYDALKDDPFNGQMECFE
        
```

Upload Sequence Data:
Choose File No file chosen

Image Format & Size

Image Format: PNG (bitmap)
Logo Size per Line:

Sequence Type:

First Position Number:

Small Sample Correction:

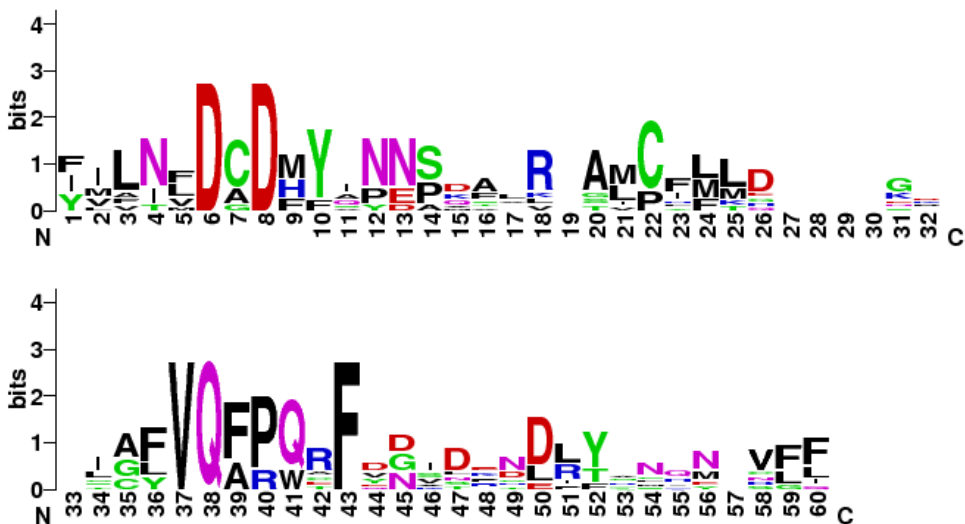
Multiline Logo (Symbols per Line):

Bitmap Resolution:

Title:

Show Y-Axis:

Show X-Axis:



With MSA you can build a phylogeny to describe the relatedness of seqs

Seqs

MSA

Phylogeny

Graph

Categories

- proteomics
- genomics
- structural bioinformatics
- systems biology
- phylogeny/evolution**
- population genetics
- transcriptomics
- biophysics
- imaging
- IT infrastructure
- drug design

Resources A..Z

Links/Documentation

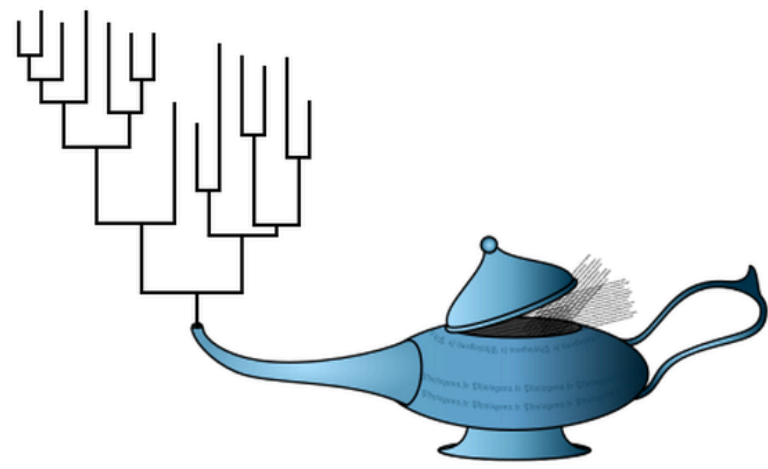
- algorithm • [more]
- Evolutionary Trace Server (TraceSuite II) • Maps evolutionary traces to structures • [more]
- fastsimcoal • coalescent simulation of genomic data • [more]
- Linear Classification • simple linear classification • [more]
- MLtree • maximum likelihood optimization • [more]
- MLTreeMap • phylogenetics and functionalities of metagenomes • [more]
- Newick Utilities • high-throughput phylogenetic tree processing • [more]
- PHYLIP • Package of programs for phylogenetic analysis • [more]
- Phylogenetic Tree • phylogenetic tree construction and printing • [more]
- Phylogeny.fr • Simple phylogenetic analysis • [more]**
- Phylogeny programs • Links to phylogeny programs • [more]
- RAxML • ML inference of large phylogenetic trees • [more]
- SuperTree • assemble phylogenetic trees • [more]

We are gonna try this website



- One Click
- Advanced
- A la Carte

Phylogeny.fr Robust Phylogenetic Analysis For The Non-Specialist



Three modes of phylogeny reconstruction

Try the one click mode

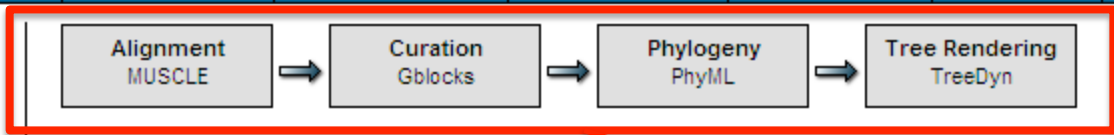
Phylogeny.fr is a free, simple to use web service dedicated to reconstructing and analysing phylogenetic relationships between molecular sequences.

Phylogeny.fr runs and connects various bioinformatics programs to reconstruct a robust phylogenetic tree from a set of sequences.

If you use this site, please cite:

Dereeper A.*, Guignon V.*, Blanc G., Audic S., Buffet S., Chevenet F., Dufayard J.F., Guindon S., Lefort V., Lescot M., Claverie J.M., Gascuel O. *Phylogeny.fr: robust phylogenetic analysis for the non-specialist*. Nucleic Acids Res. 2008 Jul 1;36(Wel Server issue):W465-9. Epub 2008 Apr 19. (PubMed) *: joint first authors

"One Click" Mode



1. Overview 2. Data & Settings

Name of the analysis (optional):

One click mode uses these tools

Upload your set of sequences in FASTA, EMBL or NEXUS format from a file.

No file chosen

Or paste it here [\(load example of sequences\)](#)

```
LLDWWRNQFYMIGATGVYLA AVLHIVLKRLLGLKGVRFKLTAKQLAGGARERFAELYDVHWSPLLAPT  
VVMVAVNVYAI GAAAGKAVVGGWTPAQVAGASAGLVFNWVWLVLLYPFALGIMGRWSKRPCALFALLVAAC  
AAVAAGFVAVHAVLAAGSAAPS WLGW SRGATAILPSSWRLKRGF  
>os_25268|LOC_Os04g35020.1|cslH  
MAVVAAAAATGSTTRSGGGGEGTRSGRKKPPPPPLQERVPLGRRAAWAWRLAGLAVLLLLLALLALRLL  
RHHGGAGGGDGGVWRVALVCEAWFAALCALNVS AKWSPVRFVTRPENLVAEGRTPSTTAAEY GELPAVDML  
VTTADPALEPPLVTVNTVLSLLALDYPRAGERLACYVSDDGCSPLTCHALREAAGFAAAWVVPFCRRYGVA  
VRAPFRYFSSSSSPESGGPADRKFLDDWT FMKDEYDKLVRRRIKNTIDERSLLRHGGGEFFAEFLNVERRNH  
PTIVKTRVSAVMTNAPIMLNMDCDMFVNNPQAVLHAMCLLLGFDDEASSGFVQAPQRFYDALKDDPFGNQ  
MECFKRFISGVQGVQAFYAGTGC FHRRKAVYGVPPNFNGAEREDTIGSSSYKELHTRFGNSEELNE SA  
RNI IWDLS SKPMVDISSRIEVAKAVSACNYDIGTCWGGQEVGWVYGLTEDILT GQRIHAMGWRSVLMVTE  
PPAFMGSAPIGGPACLTQFKRWATGQSEIIISRN NPILATMFKRLKFRQCLAYLIVLGWPLRAPFELCYG  
LLGPYCILT NQSF LPKASEDGF SVPLALFISYNTY NFM EYMACGLSARAWWNNHRMQRIISVSAWTLAFL  
TVLLKSLGLSETVFEVTGKDKSMSDDDDNTDGADPGRFTFDSL PVFI PVTALAMLNIVAVTVGACRVAFG  
TAEGVPCAPGIGEFMCCGWLVL CFFPFV RGI VWGKGSYGI PWSVKLKASLLVAMFVT FCKRN
```

Give this job a name

Maximum number of sequences is 200 for proteins and 200 for nucleic acids.
Maximum length of sequences is 2000 for proteins and 6000 for nucleic acids.

Use the Gblocks program to eliminate poorly aligned positions and divergent regions

Gblocks is a program automatically edit the alignment

To receive the results by e-mail, enter your address(es):

csi



- 1. Overview
- 2. Data & Settings
- 3. Alignment
- 4. Curation
- 5. Phylogeny
- 6. Tree Rendering

Tree Rendering results

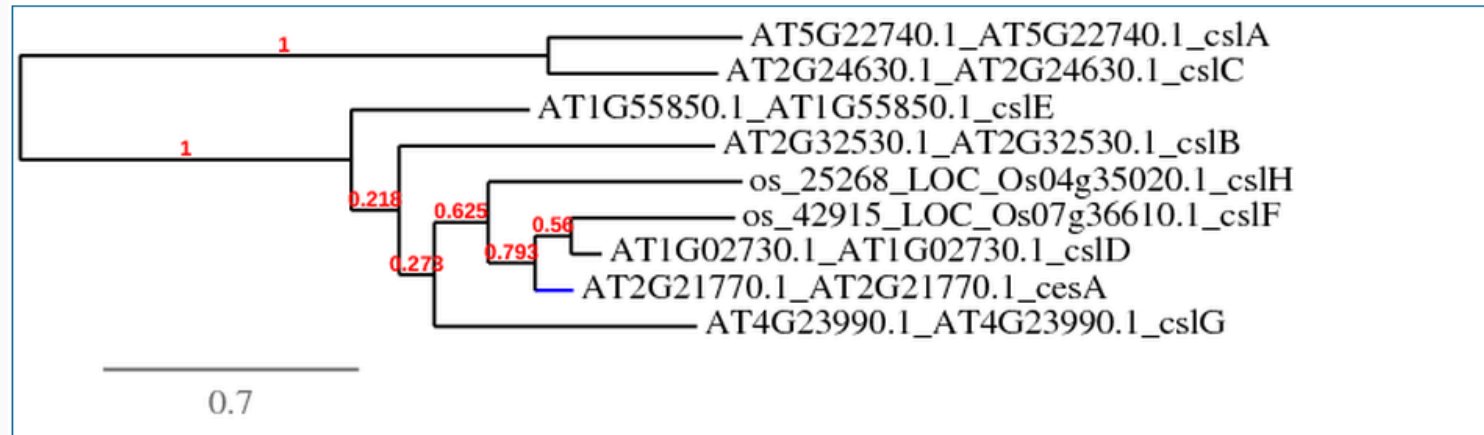


Figure 1: Phylogenetic tree (the branch length is proportional to the number of substitutions per site).

Dynamic Tree Edition

leaf
 branch

using color

and assign the group name

Add **annotations** using color

Reset to original tree

Reroot (**outgroup**)

Swap subtrees

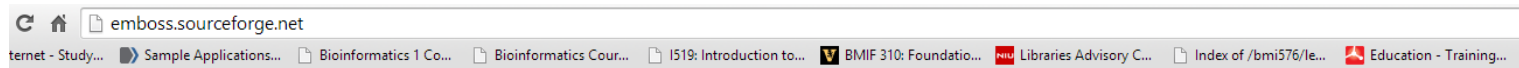
Reroot using **mid-point** rooting

Flip subtree

Change **leaf name**

EMBOSS: European Molecular Biology Open Software Suite

EMBOSS: The European Molecular Biology Open Software Suite (2000)
Rice,P. Longden,I. and Bleasby,A.
Trends in Genetics 16, (6) pp276--277



[About](#) • [Applications](#) • [GUIs](#) • [Servers](#) • [Downloads](#) • [Licence](#) • [User docs](#) • [Developer docs](#) • [Administrator docs](#) • [Get involved](#)

EMBOSS was most recently funded from May 2009 to Dec 2011 by BBSRC grant BBR/G02264X/1

Funded from May 2006 to April 2009 by BBSRC grant BB/D018358/1

About EMBOSS [Overview](#) • [Uses](#) • [FAQ](#) [Citing EMBOSS](#)

A high-quality package of free, Open Source software for molecular biology ... [more >](#)

Applications [EMBOSS](#) • [EMBASSY](#) • [Groups](#) [Proposed](#)

Hundreds of useful, well documented applications for molecular sequence and other analyses ... [more >](#)

GUIs [Jemboss](#) • [GUIs](#) • [Web](#) • [Others](#)

We support the Jemboss GUI but many others are available... [more >](#)

Servers [Portals](#) • [Servers](#) • [Mirrors](#) • [Misc](#)

Many EMBOSS portals, servers and mirrors are available ... [more >](#)

Downloads [Stable release](#) • [Developers \(CVS\) version](#) • [Getting started](#)

EMBOSS is open source software and is freely available to all ... [more >](#)

Licence [Licensing terms](#)

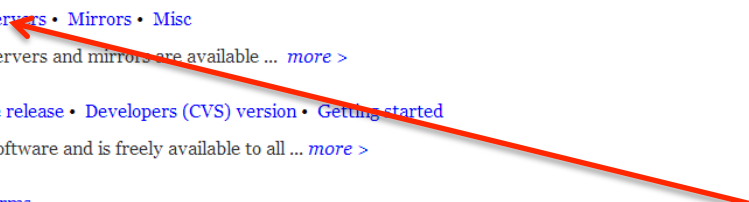
EMBOSS uses the General Public Licence (GPL) and Library GPL ... [more >](#)

User documents [FAQ](#) • [Tutorial](#) • [Running applications](#) • [Themes](#) • [Citing](#)

EMBOSS contain hundreds of computer programs written in C language for sequence analysis

The best way to use is to install it on a Linux computer

Here we're gonna try some public web servers that have EMBOSS package installed



EMBOSS servers (based on PISE)

The following sites are EMBOSS servers based on the [Pise](#) program.

[The MRC Clinical Sciences Centre, Imperial College, London.](#)
[Wellcome Trust Centre for Human Genetics, Oxford, UK](#)

EMBOSS servers (based on EMBOSS Explorer)

The following sites are EMBOSS servers based on the [EMBOSS Explorer](#) program.

[Wageningen Bioinformatics Webportal, Netherlands](#) ←
[The Centre for Genomics and Bioinformatics, Indiana University](#)
[Computer Centre, The University of Hong Kong](#)
[Cancer Vaccine Centre \(Bioinformatics\), Harvard University.](#)
[National Centre for High-Performance Computing, Taiwan.](#)
[Singapore Biomedical Computing Resource.](#)
[Robert Cedergren Center, Université de Montréal, Canada.](#)
[GSC, Japan](#)
[Centre for Comparative Genomics, Murdoch University, Australia](#)
[The University of Kansas Bioinformatics Core Facility](#)
[Southern Methodist University, Dallas, USA.](#)
[Purdue University, Indiana, USA](#)
[The Bioinformatics Center, National University of Singapore.](#)
[The University of Florida, USA.](#)
[The National Health Research Institute, Taiwan](#)
[Center for Genomics, Proteomics, and Bioinformatic, University of Hawaii at Manoa, Honolulu](#)
[Virginia Bioinformatics Institute, USA](#)
[Canadian Bioinformatics Resource](#)

Many others are not accessible, but this one is

350+ programs put into different groups

EMBOSS explorer

Welcome to EMBOSS explorer, a graphical user interface to the [EMBOSS](#) suite of bioinformatics tools.

To continue, select an application from the menu to the left. Move the mouse pointer over the name of an application in the menu to display a :

For more information about EMBOSS explorer, including how to download and install it locally, visit the [EMBOSS explorer](#) website.

Development of EMBOSS explorer has been supported by the [National Research Council of Canada](#) and [Genome Prairie](#).

ALIGNMENT
CONSENSUS
[cons](#)
[consambig](#)
[megamerger](#)
[merger](#)

ALIGNMENT
DIFFERENCES
[diffseq](#)

ALIGNMENT
DOT PLOTS
[dotmatcher](#)
[dotpath](#)
[dottup](#)
[polydot](#)

ALIGNMENT
GLOBAL
[esim4](#)
[est?genome](#)

This is called EMBOSS explorer, which is a web interface to support running EMBOSS programs through web

We will try a few programs in this package

The most basic one: translate a nucleotide seq to an amino acid seq (related to finding the open reading frames)

[vrnacofold](#)
[vrnacofoldconc](#)
[vrnacofoldpf](#)
[vrnadistance](#)
[vrnaduplex](#)
[vrnaeval](#)
[vrnaevalpair](#)
[vrnafold](#)
[vrnafoldpf](#)
[vrnaheat](#)
[vrnainverse](#)
[vrnafold](#)
[vrnaplot](#)
[vrnasubopt](#)

NUCLEIC TRANSCRIPTION

[jaspscan](#)
[tfscan](#)

NUCLEIC TRANSLATION

[backtranambig](#)
[backtranseq](#)
[coderet](#)
[plotorf](#)
[prettyseq](#)
[remap](#)
[showorf](#)
[showseq](#)
[sixpack](#)
[transeq](#)

PHYLOGENY CONSENSUS

[fconsense](#)

EMBOSS explorer

transla

1 of 1

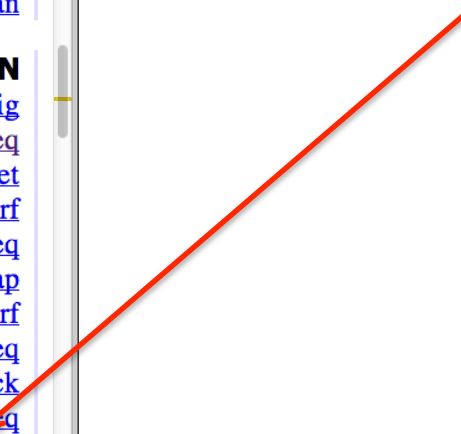
Welcome to EMBOSS explorer, a graphical user interface to the [EMBOSS](#) suite of bioinformatics tools.

To continue, select an application from the menu to the left. Move the mouse pointer over the name of an application in the menu to display a short description. To search for a particular application, use [wosname](#).

For more information about EMBOSS explorer, including how to download and install it locally, visit the [EMBOSS explorer](#) website.

Development of EMBOSS explorer has been supported by the [National Research Council of Canada](#) and [Genome Canada](#).

Find the program transeq in the nucleic translation group



Copy and paste the seq in <http://cys.bios.niu.edu/yyin/teach/PBB/nt-example.fa>
It's an assembled transcript from EST data of some algal species
We do not know if it indeed encode a protein and if yes where is the ORF
Remember mRNA contains untranslated region (UTR)

- [vrnaalifoldpf](#)
- [vrnacofold](#)
- [vrnacofoldconc](#)
- [vrnacofoldpf](#)
- [vrnadistance](#)
- [vrnaduplex](#)
- [vrnaeval](#)
- [vrnaevalpair](#)
- [vrnafold](#)
- [vrnafoldpf](#)
- [vrnaheat](#)
- [vrnainverse](#)
- [vrnalfold](#)
- [vrnaplot](#)
- [vrnasubopt](#)

transeq

Translate nucleic acid sequences ([read the manual](#))

Unshaded fields are optional and can safely be ignored. ([hide optional fields](#))

Input section

Select an input sequence. Use one of the following three fields:

1. To access a sequence from a database, enter the USA here:
2. To upload a sequence from your local computer, select it here: No file chosen

```
GTGCTACAAGCAGAGCATCGAGGCGGCATGCAGGCTGACTTG
GGAAGGCCGCATGCACGTCCAGGTCCTCGATGACTCAGACGA
CGAGGAGATCCAGGCGCTCATCAGAGACGAGGTCCGGCAG
GTGGCAGCAGAAGAACATGAGCATCGTATATCTCCAAGGGAC
AACCCGACTGGCCACAAGGCCGGCAACCTCGCGTTTGGGAT
GGCGGAGGCCGAGCCCCAGGGCTTCCAGTTTGTGGTGATCTT
TGACCGGACTTCATGCCCCAGCCCCGACTTCTCGCAGCG
CACCGTGCCGCTCTCCGCGGCCG
```

3. To enter the sequence data manually, type here:

Additional section

Frame(s) to translate

Choose all six frames

Code to use

Regions to translate (eg: 4-57,78-94)

nucleotide

NUCLEIC TRANSCRIPTION

- [jaspscan](#)
- [tfscan](#)

NUCLEIC TRANSLATION

- [backtranambig](#)
- [backtranseq](#)
- [coderet](#)
- [plotorf](#)
- [prettyseq](#)
- [remap](#)
- [showorf](#)
- [showseq](#)
- [sixpack](#)
- [transeq](#)

[vrnaalifoldpt](#)
[vrnacofold](#)
[vrnacofoldconc](#)
[vrnacofoldpf](#)
[vrnadistance](#)
[vrnaduplex](#)
[vrnaeval](#)
[vrnaevalpair](#)
[vrnafold](#)
[vrnafoldpf](#)
[vrnaheat](#)
[vrnainverse](#)
[vrnalfold](#)
[vrnaplot](#)
[vrnasubopt](#)

C TRANSCRIPTION

[jaspSCAN](#)
[tfscan](#)

GENE TRANSLATION

[backtranambig](#)
[backtranseq](#)
[coderet](#)
[plotorf](#)
[prettyseq](#)
[remap](#)
[showorf](#)
[showseq](#)
[sixpack](#)

OUTPUT FILE [outseq](#)

This is likely the right frame



```

>JO181643.1_1
RDKMHVGFPRSLGNKRIVGNGSTRFLGVTVKHWLHPGLATSFRTLCTVVTGSCSGSGGAI
PGTAGRQALPIHRASVDMGLQVSNACPRERDCAAGGSAVLGCRGGQAGGGVCPDPDPHV
QREGVLQAEHRGGMQADLGRPHARPGPR*LRRRGDPGAHQRRGRQVAEEHEHRISPPDN
PDWPQGRQPRVWDGGGRAPGLPVCGLD*RGLHAQPRLPAAHRAALPRPX
>JO181643.1_2
GIRCTWAFLDHWATSASLATVLLVCLALPGSMSSIQVLQPVFELCAQLLQALVRALEVLV
LAQLADKLFLLFTGRLWIWASKYRMPVLERETVPQEEDQLSLVAEEGKLEAESVLIQIPMC
NERECYKQSIEAACRLTWEGRMHVQVLDDSDDEEIQALIRDEVGRWQQKNMSIVYLHRTT
RTGHKAGNLAFGMAEAEPOGFQFVVIFDADFMPSPDFLQRTVPLFRGR
>JO181643.1_3
G*DARGLS*IIGQQAHRWQRFYSFVWRYLEACPPSRSCNQFSNFVHSCYRLLFGLWRCYS
WHSWPTSSSYSPGVCYGPSSIECLSSRERLCRRRRI SCPWLP RRASWRRSLS*SRSPCA
TRGSATSRASRRHAG*LGKAACTSRSSMTQTTRRSRRSSETR SAGGSRT*ASYISTGQP
GLATRPATSRLGWRRPSPRASSLW*SLTRTSCPAPTSCSAPCRSSAAA
>JO181643.1_4
RPRKSGTVRCRKSGLGMKSASKITTNWKPWGSASAI PNARLPALWPVRVVRWRYTMLMFF
CCHLPTSSLMSAWISSSESSRTWT CMRPSQVSLHAASMLCL*HSLSLHMGIWIRTD SAS
SLPSSATKDS*SSSCGTVSLSRTGIRYLEAHIHRRPVNRKLSLASCARNSTSRARTRACN
NCAQSSKTGCKTWMEDMLPGNAKQTSRTVANDALVAQ*SRKAHVHLIP
>JO181643.1_5
AAAEERHGALQEVGAGHEVRVKDHHKLEALGLGLRHPKREVAGLVASPGCPVEIYDAHVL
LLPPADLVSDERLDLLV*VIEDLDVHAAFPSQPACRLDALLVALPLVAHGDLQDRLRL
QLALLGNQGQLILLRHSLSLEDRHSILGGYPYQTPGE*EELVGQLCQE*HLQSPNKSL*
QLCTKFENWLQDLDDGGHASR*RQTNE*NRCQRACCPMI*ESPRASYPX
>JO181643.1_6
GRGRAARCAAGSRGWA*SPRQRSPQTGSPGARPPPSQTRGCRPCGQSGLSGGDIRCSCSS
AATCRPRL**APGSPRRLSHRGPGRACGLPKSACMPPRCSACSTPSRCTWGS GSGQTPPP
ACPPRQPRTADPPPAQSLSRGQAFDTWRPISTDAR*IGRACRPVAVPGIAPPEPEQEPVT
TVHKVRKLVARPGWRTCFQVTPNKRVEPLPTMRLLPNDLGKPTCILSR

```

If this is a correct result? You can take the nt seq to do blast at NCBI

NCBI/ BLAST/ blastx Translated BLAST: blastx

blastn blastp **blastx** tblastn tblastx

BLASTX search protein databases using a translated nucleotide query

Enter Query Sequence

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

JO181643.1 From
To

Or, upload file No file chosen

Genetic code

Job Title
Enter a descriptive title for your BLAST search

Align two or more sequences

Choose Search Set

Database

Organism Exclude
Optional
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude Models (XM/XP) Uncultured/environmental sample sequences
Optional

Entrez Query [YouTube](#) [Create custom database](#)
Optional
Enter an Entrez query to limit search

BLAST Search using **Blastx** (search protein databases using a translated nucleotide query)

Show results in a new window

Put the seq ID here because it is in GenBank already

Choose swiss-prot because it is smaller and high quality

Click formatting options
Choose plain text view
Click reformat

NCBI/ BLAST/ blastx/ Formatting Results - 2K239NB1014

[Edit and Resubmit](#) [Save Search Strategies](#) [Formatting options](#) [Download](#) [YouTube](#) [How to read this page](#) [Blast re](#)

Formatting options

Show Alignment as: **Plain text** Old View [Reset form to defaults](#)

Alignment View Pairwise

Display Graphical Overview Linkout Sequence Retrieval NCBI-gi

Masking Character: Lower Case Color: Grey

Limit results Descriptions: 100 Graphical overview: 100 Alignments: 100 Line length: 60

Organism Type common name, binomial, taxid, or group name. Only 20 top taxa will be shown.
 Exclude +

Entrez query:

Expect Min: Expect Max:

Percent Identity Min: Percent Identity Max:

[Reformat](#)

gb|JO181643.1| (685 letters)

RID [2K239NB1014](#) (Expires on 10-01 00:27 am)

Query ID [gi|341817447|gb|JO181643.1|](#)

Description TSA: Chaetosphaeridium globosum strain SAG 26.98 chglo_000001333313I17.F.ab1_c_s mRNA sequence

Molecule type rna

Database Name swissprot

Description Non-redundant UniProtKB/SwissProt sequences

Program BLASTX 2.2.30+ [Citation](#)

This is the alignment of our query with the best hit, the frame is +2, same as the transeq result

ALIGNMENTS

>sp|Q9SJA2.1|CSLC8_ARATH RecName: Full=Probable xyloglucan glycosyltransferase 8; AltName: Full=Cellulose synthase-like protein C8; Short=AtCslC8 [Arabidopsis thaliana] Length=690

Score = 166 bits (420), Expect = 5e-46, Method: Compositional matrix adjust.
Identities = 97/213 (46%), Positives = 127/213 (60%), Gaps = 17/213 (8%)
Frame = +2

```
Query 77 LALPGSMSSIQVLQPVF-----ELCAQLLQALVRALEVLFQAQLADKFLFTGRLW 229
      L +P S IQ L +F + A ++AL + VLFL Q D+L L G LW
Sbjct 140 LHIPTSKLEIQSLLHLFYVGWLSLRADYIAPPIKALSKFCIVLFLVQSVDRILILCLGCLW 199

Query 230 IWASKYRMPVLERETVTPQEEDQLSLVAEEGKLEAESVLIQIPMCNERECYKQSI EAACRL 409
      I K + P ++ E ++ E E VL+QIPMCNERE Y+QSI A C+L
Sbjct 200 IKFKKIK-PRIDEEHFRNDD-----FEGSGSEYPMVLVQIPMCNEREVYEQSISAVCQL 252

Query 410 TW-EGRMHVQVLDDSDDEEIQALIRDEVGRWQQKNMSIVYLHRTTRTGHKAGNLAFGMAE 586
      W + R+ VQVLDDSDDE IQ LIRDEV +W QK ++I+Y HR RTG+KAGNL M+
Sbjct 253 DWPKDRLLVQVLDDSDDESIQELIRDEVTKWSQKGVNIIYRHRLVRTGYKAGNLKSAMSC 312

Query 587 AEPQGFQFVVIFDADFMPSPDFLQRTVPLFRGR 685
      + ++FV IFDADF P+ DFL+ TVP F+ +
Sbjct 313 DYVEAYEFVAIFDADFQPNDFLKLTPHFKEK 345
```

[vrnacofold](#)
[vrnacofoldconc](#)
[vrnacofoldpf](#)
[vrnadistance](#)
[vrnaduplex](#)
[vrnaeval](#)
[vrnaevalpair](#)
[vrnafold](#)
[vrnafoldpf](#)
[vrnaheat](#)
[vrnainverse](#)
[vrnalfold](#)
[vrnaplot](#)
[vrnasubopt](#)

CLEIC TRANSCRIPTION

[jaspSCAN](#)
[tfscan](#)

NUCLEIC TRANSLATION

[backtranamig](#)
[backtranseq](#)
[coderet](#)
[plotorf](#)
[prettyseq](#)
[remap](#)
[showorf](#)
[showseq](#)
[sixpack](#)

plotorf

Plot potential open reading frames in a nucleotide sequence ([read the manual](#))

Unshaded fields are optional and can safely be ignored. ([hide optional fields](#))

Input section

Select an input sequence. Use one of the following three fields:

1. To access a sequence from a database, enter the USA here:
2. To upload a sequence from your local computer, select it here: No file chosen

```
>Chaetosphaeridium_globosum|gb|J0181643.1  
CGGGATAAGATGCACGTGGGCTTTCCTAGATCATTGGGCAAC  
AAGCGCATCGTTGGCAACGGTTCTACTCGTTTGTGGCGTTA  
CCTGGAAGCATGTCCTCCATCCAGGTCTTGAACCAAGTTTTCG  
AACTTTGTGCACAGTTGTTACAGGCTCTTGTTCCGGC  
TCTGGAGGTGCTATTCCTGGCACAGCTGGCCGACAAGCTCTT  
CCTATTCACCGGGCGTCTGTGGATATGGCCCTCCAAGTATCG  
AATGCCTGTCTCGAGAGAGAGACTGTGCCGAGGAGGAGGA
```

3. To enter the sequence data manually, type here:

Advanced section

Start codons

Stop codons

This is the longest ORF

[propnet](#)

IMAGE FILE [plotorf.1.png](#)

nucleotide

NUCLEIC REPEATS

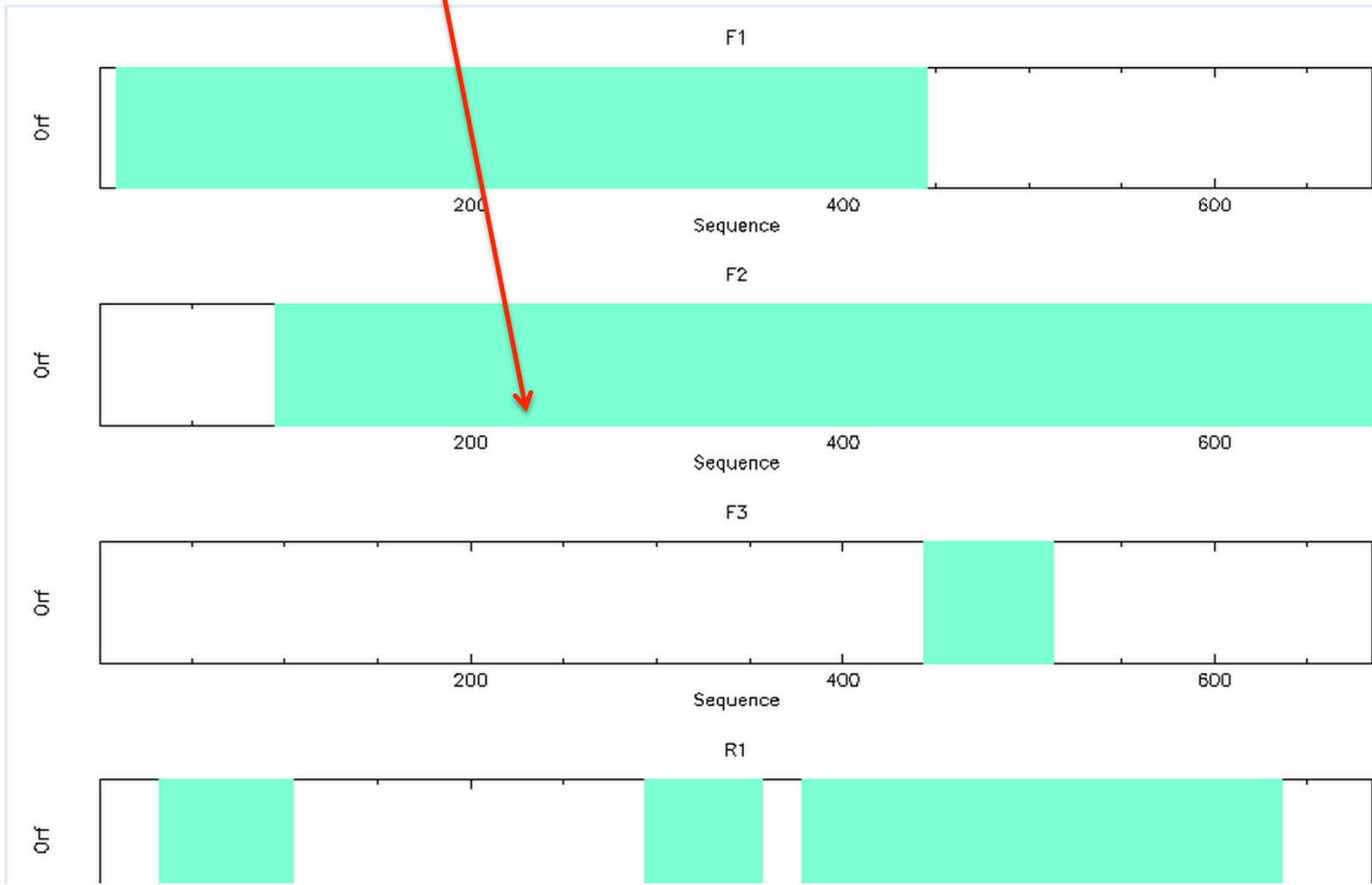
- [einvrted](#)
- [equicktandem](#)
- [etandem](#)
- [palindrome](#)

NUCLEIC RESTRICTION

- [recoder](#)
- [redata](#)
- [remap](#)
- [restover](#)
- [restrict](#)
- [showseq](#)
- [slient](#)

NUCLEIC RNA FOLDING

- [vrnaalifold](#)
- [vrnaalifoldpf](#)
- [vrnacofold](#)
- [vrnacofoldconc](#)
- [vrnacofoldpf](#)
- [vrnadistance](#)
- [vrnaduplex](#)
- [vrnaeval](#)
- [vrnaevalpair](#)
- [vrnafold](#)
- [vrnafoldpf](#)
- [vrnaheat](#)
- [vrnainverse](#)



<http://cys.bios.niu.edu/yyin/teach/PBB/nt-example.fa>

Complement
Reverse

ATGCGCTA
TACGCGAT
TAGCGCAT

- EDIT**
- [aligncopy](#)
- [aligncopypair](#)
- [biosed](#)
- [codcopy](#)
- [cutseq](#)
- [degapseq](#)
- [descseq](#)
- [entret](#)
- [extractalign](#)
- [extractfeat](#)
- [extractseq](#)
- [featcopy](#)
- [featreport](#)
- [listor](#)
- [makenucseq](#)
- [makeprotseq](#)
- [maskambignuc](#)
- [maskambigprot](#)
- [maskfeat](#)
- [maskseq](#)
- [newseq](#)
- [nohtml](#)
- [noreturn](#)
- [nospace](#)
- [notab](#)
- [notseq](#)
- [nthseq](#)
- [pasteseq](#)
- [revseq](#)
- [seqret](#)

revseq

Reverse and complement a nucleotide sequence ([read the manual](#))

Unshaded fields are optional and can safely be ignored. ([hide optional fields](#))

Input section

Select an input sequence. Use one of the following three fields:

1. To access a sequence from a database, enter the USA here:
2. To upload a sequence from your local computer, select it here: No file chosen

3. To enter the sequence data manually, type here:

Advanced section

Reverse sequence?

Complement sequence?

Output section

>JO181643.1

GCGGCCGCGGAAGAGCGGCACGGTGCCTGCAGGAAGTCGGGGCTGGGCATGAAGTCCGC
GTCAAAGATCACCAAACTGGAAGCCCTGGGGCTCGGCCTCCGCCATCCCAAACGCGAG
GTTGCCGGCCTTGTGGCCAGTCCGGGTTGTCCGGTGGAGATATACGATGCTCATGTTCTT
CTGCTGCCACCTGCCGACCTCGTCTCTGATGAGCGCCTGGATCTCCTCGTCGTCTGAGTC
ATCGAGGACCTGGACGTGCATGCGGCCTTCCCAAGTCAGCCTGCATGCCGCCTCGATGCT
CTGCTTGTAGCACTCCCTCTCGTTGCACATGGGGATCTGGATCAGGACAGACTCCGCCTC
CAGCTTGCCCTCCTCGGCAACCAAGGACAGCTGATCCTCCTCCTGCGGCACAGTCTCTCT
CTCGAGGACAGGCATTCGATACTTGGAGGCCATATCCACAGACGCCCGGTGAATAGGAA
GAGCTTGTCGGCCAGCTGTGCCAGGAATAGCACCTCCAGAGCCCGAACAAGAGCCTGTAA
CAACTGTGCACAAAGTTCGAAAACCTGGTTGCAAGACCTGGATGGAGGACATGCTTCCAGG
TAACGCCAAACAAACGAGTAGAACCGTTGCCAACGATGCGCTTGTGCCCCAATGATCTAG
GAAAGCCCACGTGCATCTTATCCCG

Region 2-7
↓
ATGCGCTA
↓
TGCGCT

- [showseq](#)
- [sixpack](#)
- [textsearch](#)

EDIT

- [aligncopy](#)
- [aligncopypair](#)
- [biosed](#)
- [codcopy](#)
- [cutseq](#)
- [degapseq](#)
- [descseq](#)
- [entret](#)
- [extractalign](#)
- [extractfeat](#)
- [extractseq](#)
- [featcopy](#)
- [featreport](#)
- [listor](#)
- [makenucseq](#)
- [makeprotseq](#)
- [maskambignuc](#)
- [maskambigprot](#)
- [maskfeat](#)
- [maskseq](#)

extractseq

Extract regions from a sequence ([read the manual](#))

Unshaded fields are optional and can safely be ignored. ([hide optional fields](#))

Input section

Select an input sequence. Use one of the following three fields:

1. To access a sequence from a database, enter the USA here:
2. To upload a sequence from your local computer, select it here: No file chosen

```
>Chaetosphaeridium_globosum|gb|J0181643.1
CGGGATAAGATGCACGTGGGCTTTCCTAGATCATTGGGCAAC
AAGCGCATCGTTGGCAACGGTTCTACTCGTTTGTTGGCGTTA
CCTGGAAGCATGTCCTCCATCCAGGTCTTGCAACCAGTTTTTCG
AACTTTGTGCACAGTTGTTACAGGCTCTTGTTCCGGGC
TCTGGAGGTGCTATTCTGGCACAGCTGGCCGACAAGCTCTT
CCTATTCACCGGGCGTCTGTGGATATGGGCCTCCAAGTATCG
AATGCCTGTCCTCGAGAGAGAGACTGTGCCGAGGAGGAGGA
```

3. To enter the sequence data manually, type here:

Required section

Regions to extract (eg: 4-57,78-94)

Required section

Region 57-400

[sixpack](#)
[textsearch](#)

EDIT

[aligncopy](#)
[aligncopypair](#)
[biosed](#)
[codcopy](#)
[cutseq](#)
[degapseq](#)
[descseq](#)
[entret](#)
[extractalign](#)
[extractalign](#)

OUTPUT FILE [outseq](#)

>JO181643.1

```
CAACGGTTCTACTCGTTTGTGGCGTTACCTGGAAGCATGTCCTCCATCCAGGTCTTGC  
AACCAGTTTTCGAACTTTGTGCACAGTTGTTACAGGCTCTTGTTTCGGGCTCTGGAGGTGC  
TATTCCTGGCACAGCTGGCCGACAAGCTCTTCCTATTCACCGGGCGTCTGTGGATATGGG  
CCTCCAAGTATCGAATGCCTGTCCTCGAGAGAGAGACTGTGCCGCAGGAGGAGGATCAGC  
TGTCTTGGTTGCCGAGGAGGGCAAGCTGGAGGCGGAGTCTGTCCTGATCCAGATCCCA  
TGTGCAACGAGAGGGAGTGCTACAAGCAGAGCATCGAGGCGGCA
```

ATGCGCTA

GC%=50%

freak

Generate residue/base frequency table or plot ([read the manual](#))

Unshaded fields are optional and can safely be ignored. ([hide optional fields](#))

Input section

Select an input sequence. Use one of the following three fields:

1. To access a sequence from a database, enter the USA here:
2. To upload a sequence from your local computer, select it here: No file chosen

```
>Chaetosphaeridium_globosum|gb|J0181643.1  
CGGGATAAGATGCACGTGGGCTTTCCTAGATCATTGGGCAAC  
AAGCGCATCGTTGGCAACGGTTCTACTCGTTTGTGGCGTTA  
CCTGGAAGCATGTCTCCATCCAGGTCTTGCAACCAGTTTTCC  
AACTTTGTGCACAGTTGTTACAGGCTCTTGTTCCGGGC  
TCTGGAGGTGCTATTCTGGCACAGCTGGCCGACAAGCTCTT  
CCTATTCACCCGGGCGTCTGTGGATATGGGCCTCCAAGTATCG  
AATGCCTGTCTCGAGAGAGAGACTGTGCCGAGGAGGAGGA
```

3. To enter the sequence data manually, type here:

Required section

Residue letters

Calculate GC content

S
t
r
e
r
k

Output section

Produce graphic?

Change to yes to get a pic

Output graphic format

[cusp](#)
[codcmp](#)
[cusp](#)
[syco](#)

NUCLEIC COMPOSITION

[banana](#)
[btwisted](#)
[chaos](#)
[compseq](#)
[dan](#)
[density](#)
[freak](#)
[isochore](#)
[sirna](#)
[wordcount](#)

NUCLEIC CPG ISLANDS

[cpgplot](#)
[cpgreport](#)
[geecce](#)
[newcpgreport](#)
[newcpgseek](#)

NUCLEIC GENE FINDING

[getorf](#)
[marscan](#)

[chips](#)
[codcmp](#)
[cusp](#)
[syco](#)

IMAGE FILE [freak.1.png](#)

POSITION

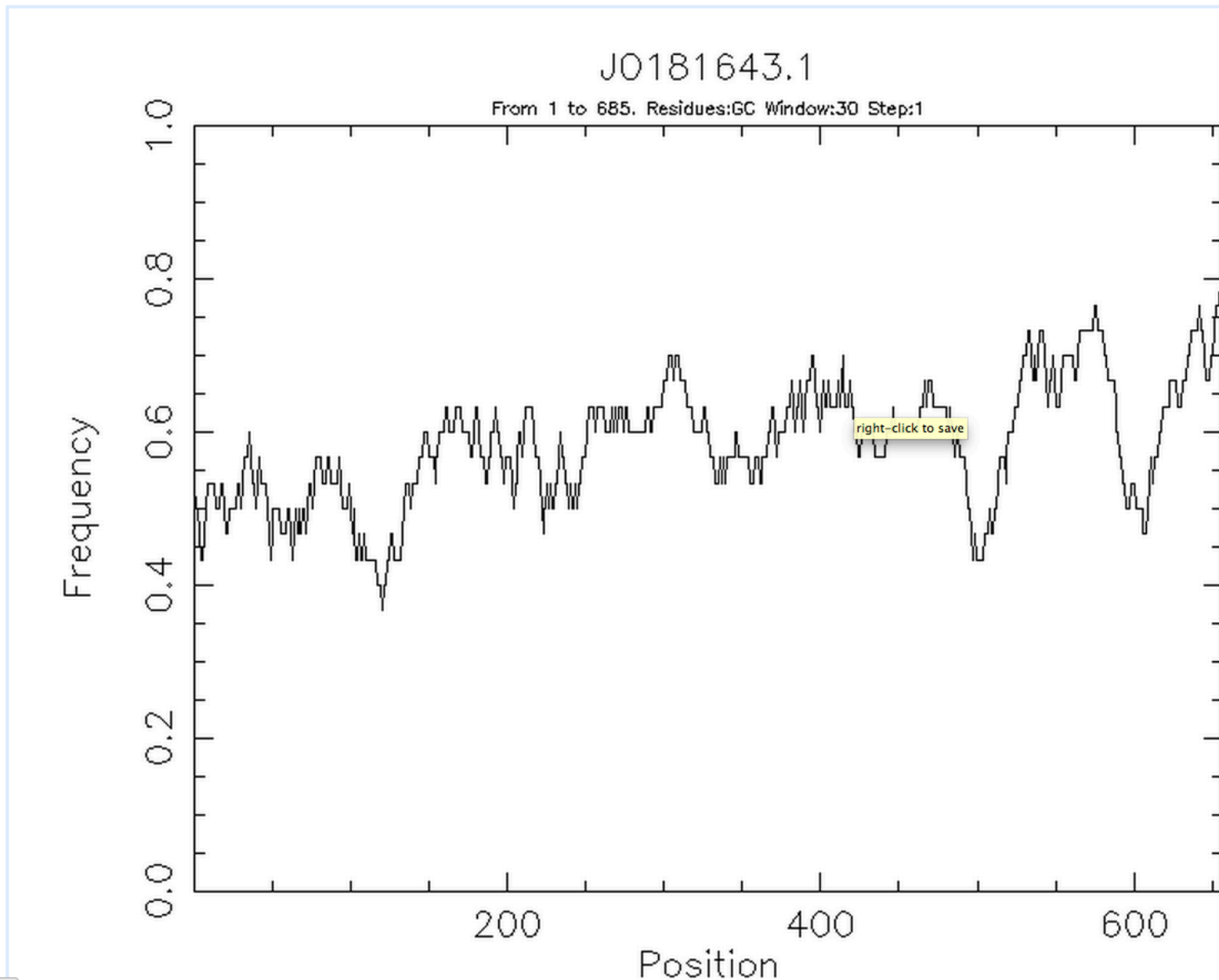
[banana](#)
[btwisted](#)
[chaos](#)
[compseq](#)
[dan](#)
[density](#)
[freak](#)
[isochore](#)
[sirna](#)
[wordcount](#)

CG ISLANDS

[cpgplot](#)
[cpgreport](#)
[geecee](#)
[newcpgreport](#)
[newcpgseek](#)

ORF FINDING

[getorf](#)
[marscan](#)
[plotorf](#)
[showorf](#)
[sixpack](#)
[syco](#)
[tcode](#)
[wobble](#)



ATGCGCTA

16 possible dinuc

64 possible trinuc

256 possible tetranuc

[siggenlig](#)
[sigplot](#)
[sigscan](#)
[sigscanlig](#)

compseq

Calculate the composition of unique words in sequences ([read the manual](#))

PROTEIN COMPOSITION

[backtranambig](#)
[backtranseq](#)
[charge](#)
[checktrans](#)
[compseq](#)
[emowse](#)
[freak](#)
[iep](#)
[mwcontam](#)
[mwfilter](#)
[octanol](#)
[pepinfo](#)
[pepstats](#)
[pepwindow](#)
[pepwindowall](#)
[wordcount](#)

Unshaded fields are optional and can safely be ignored. ([hide optional fields](#))

Input section

Select an input sequence. Use one of the following three fields:

1. To access a sequence from a database, enter the USA here:

2. To upload a sequence from your local computer, select it here: No file chosen

```
>Chaetosphaeridium_globosum|gb|J0181643.1  
CGGGATAAGATGCACGTGGGCTTTCCTAGATCATTGGGCAAC  
AAGCGCATCGTTGGCAACGGTCTACTCGTTTGTGGCGTTA  
CCTGGAAGCATGTCCTCCATCCAGGTCTTGCAACCAGTTTTTCG  
AACTTTGTGCACAGTTGTTACAGGCTCTTGTTCCGGGC  
TCTGGAGGTGCTATTCTGGCACAGCTGGCCGACAAGCTCTT  
CCTATTCACCGGGCGTCTGTGGATATGGGCCTCCAAGTATCG  
AATGCCTGTCTCGAGAGAGAGACTGTGCCGAGGAGGAGGA
```

3. To enter the sequence data manually, type here:

Program compseq output file (optional): No file chosen

Input section

Required section

Word size to consider (e.g. 2=dimer)

Default is dinuc

PROTEIN MOTIFS

[antigenic](#)
[digest](#)
[echlorop](#)
[eiprscan](#)
[elipr](#)

Application: scan genome to look for regions with abnormal compositions

PROTEIN COMPOSITION

- [backtranambig](#)
- [backtranseq](#)
- [charge](#)
- [checktrans](#)
- [compseq](#)
- [emowse](#)
- [freak](#)
- [iep](#)
- [mwcontam](#)
- [mwfilter](#)
- [octanol](#)
- [pepinfo](#)
- [pepstats](#)
- [pepwindow](#)
- [pepwindowall](#)
- [wordcount](#)

PROTEIN MOTIFS

- [antigenic](#)
- [digest](#)
- [echlorop](#)
- [eiprscan](#)
- [elipop](#)
- [smact](#)

OUTPUT FILE [outfile](#)

```
#
# Output from 'compseq'
#
# The Expected frequencies are calculated on the (false) assumption that every
# word has equal frequency.
#
# The input sequences are:
#      JO181643.1
```

```
Word size      2
Total count    684
```

Equal occurrence: 1/16



#	Word	Obs Count	Obs Frequency	Exp Frequency	Obs/Exp Frequency
#					
AA	19	0.0277778	0.0625000	0.4444444	
AC	32	0.0467836	0.0625000	0.7485380	
AG	55	0.0804094	0.0625000	1.2865497	
AT	30	0.0438596	0.0625000	0.7017544	
CA	53	0.0774854	0.0625000	1.2397661	
CC	48	0.0701754	0.0625000	1.1228070	
CG	45	0.0657895	0.0625000	1.0526316	
CT	44	0.0643275	0.0625000	1.0292398	
GA	52	0.0760234	0.0625000	1.2163743	
GC	63	0.0921053	0.0625000	1.4736842	
GG	63	0.0921053	0.0625000	1.4736842	
GT	35	0.0511696	0.0625000	0.8187135	
TA	12	0.0175439	0.0625000	0.2807018	
TC	47	0.0687135	0.0625000	1.0994152	
TG	50	0.0730994	0.0625000	1.1695906	
TT	36	0.0526316	0.0625000	0.8421053	

[seqword](#)
[siggen](#)
[siggenlig](#)
[sigplot](#)
[sigscan](#)
[sigscanlig](#)

PROTEIN COMPOSITION

[backtranambig](#)
[backtranseq](#)
[charge](#)
[checktrans](#)
[compseq](#)
[emowse](#)
[freak](#)
[iep](#)
[mwcontam](#)
[mwfilter](#)
[octanol](#)
[pepinfo](#)
[pepstats](#)
[pepwindow](#)
[pepwindowall](#)
[wordcount](#)

PROTEIN MOTIFS

[antigenic](#)
[digest](#)
[echlorop](#)
[eiprscan](#)
[elipop](#)

pepstats

Calculates statistics of protein properties ([read the manual](#))

Unshaded fields are optional and can safely be ignored. ([hide optional fields](#))

Input section

Select an input sequence. Use one of the following three fields:

1. To access a sequence from a database, enter the USA here:
2. To upload a sequence from your local computer, select it here: No file chosen

```
>AT2G21770.1|AT2G21770.1|cesa  
MNTGGRLIAGSHNRNEFVLINADDTARIRSAEELSGQTCKICRDEIEL  
TDNGEPFIACNECAFPTCRPCYEYERREGNQACPQCGTRYKRIKGS  
PRVEGDEEDDDIDDLHEFYGMDEPHVTEAALYMLRNTGRGTDE  
VSHLYSASPGSEVPLLYCDESDMYSDRHALIVPPSTGLGNRVHH  
VPFTDSFASIHTRPMVPQKDLTVYGYGSVAWKDRMEVWKKQIEK  
LQVVKNERVNDGGDGFIVDELDDPGLPMMDEGRQPLSRKLPIRR  
RINPYRMLIFCRLAILGLFFHYRILHPVNDAGLWLTVICEIWFVSW
```

3. To enter the sequence data manually, type here:

Amino acids properties and molecular weight data file. Use one of the following two fields:

1. To access a standard EMBOSS data file, enter the name here:
2. To upload a data file from your local computer, select it here: No file chosen

Molecular weight data file. Use one of the following two fields:

1. To access a standard EMBOSS data file, enter the name here:
2. To upload a data file from your local computer, select it here: No file chosen

[seqscan](#)
[seqsort](#)
[seqwords](#)
[siggen](#)
[siggenlig](#)
[sigplot](#)
[sigscan](#)
[sigscanlig](#)

PROTEIN COMPOSITION

[backtranambig](#)
[backtranseq](#)
[charge](#)
[checktrans](#)
[compseq](#)
[emowse](#)
[freak](#)
[iep](#)
[mwcontam](#)
[mwfilter](#)
[octanol](#)
[pepinfo](#)
[pepstats](#)
[pepwindow](#)
[pepwindowall](#)
[wordcount](#)

PROTEIN MOTIFS

[antigenic](#)
[digest](#)
[echlorop](#)
[eiprscan](#)
[elipop](#)
[emot](#)

OUTPUT FILE [outfile](#)

PEPSTATS of cesA from 1 to 1088

Molecular weight = 123446.87 Residues = 1088
 Average Residue Weight = 113.462 Charge = 5.5
 Isoelectric Point = 6.8610
 A280 Molar Extinction Coefficient = 211800
 A280 Extinction Coefficient 1mg/ml = 1.72
 Improbability of expression in inclusion bodies = 0.695

Residue	Number	Mole%	DayhoffStat
A = Ala	56	5.147	0.598
B = Asx	0	0.000	0.000
C = Cys	32	2.941	1.014
D = Asp	64	5.882	1.070
E = Glu	65	5.974	0.996
F = Phe	49	4.504	1.251
G = Gly	81	7.445	0.886
H = His	25	2.298	1.149
I = Ile	67	6.158	1.368
J = ---	0	0.000	0.000
K = Lys	63	5.790	0.877
L = Leu	102	9.375	1.267
M = Met	27	2.482	1.460
N = Asn	43	3.952	0.919
O = ---	0	0.000	0.000
P = Pro	63	5.790	1.114
Q = Gln	30	2.757	0.707
R = Arg	59	5.423	1.107
S = Ser	68	6.250	0.893
T = Thr	44	4.044	0.663
U = ---	0	0.000	0.000
V = Val	81	7.445	1.128
W = Trp	28	2.574	1.980
X = Xaa	0	0.000	0.000
Y = Tyr	41	3.768	1.108
Z = Glx	0	0.000	0.000

Popular tools developed at Technical University of Denmark

NUCLEOTIDE SEQUENCES

Whole genome visualization and analysis

« [GenomeAtlas](#)
DNA structural atlases for complete microbial Genomes

Gene finding and splice sites

« [EasyGene](#)
Genes in prokaryotes

« [EasyGene](#)
Genes in prokaryotes

« [HMMgene](#)
Genes in eukaryotes

[MetaRanker](#)
Identification of risk genes in complex phenotypes

[NetAspGene](#)
Intron splice sites in *Aspergillus* DNA

« [NetGene2](#)
Intron splice sites in human, *C. elegans* and *A. thaliana* DNA

[NetPlantGene](#)
Intron splice sites in *Arabidopsis thaliana* DNA

« [NetStart](#)
Translation start in vertebrate and *A. thaliana* DNA

[NetOTR](#)
Splice sites in 5' UTR regions of human genes

« [Promoter](#)
Transcription start sites in vertebrate DNA

« [RNAmmer](#)
Ribosomal RNA sub units

« [RNAmmer](#)
Ribosomal RNA sub units

Analysis of DNA microarray data

[GenePublisher](#)
Analysis of DNA microarray data

« [OligoWiz](#)
Design of oligonucleotides for DNA microarrays

SMALL MOLECULES

« [ChemProt](#)
Chemical-protein interactions

AMINO ACID SEQUENCES

Protein sorting

[ChloroP](#) »
Chloroplast transit peptides and their cleavage sites in plant proteins

[LipoP](#) »
Signal peptidase I & II cleavage sites in gram- bacteria

[NetNES](#) »
Leucine-rich nuclear export signals (NES) in eukaryotic proteins

[SecretomeP](#) »
Non-classical and leaderless secretion of proteins

[SignalP](#) »
Signal peptide and cleavage sites in gram+, gram- and eukaryotic amino acid sequences

[TargetP](#) »
Subcellular location of proteins: mitochondrial, chloroplastic, secretory pathway, or other

[TatP](#) »
Twin-arginine signal peptides

Post-translational modifications of proteins

[DictyOGlyc](#)
O-(alpha)-GlcNAc glycosylation sites (trained on *Dictyostelium discoideum* proteins)

[NetAcet](#)
N-terminal acetylation in eukaryotic proteins

[NetCGlyc](#) »
C-mannosylation sites in mammalian proteins

[NetCorona](#)
Coronavirus 3C-like proteinase cleavage sites in proteins

[NetGlycate](#) »
Glycation of ε amino groups of lysines in mammalian proteins

[NetNGlyc](#) »
N-linked glycosylation sites in human proteins

[NetNGlyc](#) »
N-linked glycosylation sites in human proteins

[NetOGlyc](#) »
O-GalNAc (mucin type) glycosylation sites in mammalian proteins

[NetOGlyc](#) »
O-GalNAc (mucin type) glycosylation sites in mammalian proteins

[NetPhorest](#)
Linear motif atlas for phosphorylation-dependent signaling

[NetPhos](#) »
Generic phosphorylation sites in eukaryotic proteins

[NetPhosBac](#)
Generic phosphorylation sites in bacterial proteins

[NetPhosK](#)
Kinase specific phosphorylation sites in eukaryotic proteins

[NetPhosYeast](#)
Serine and threonine phosphorylation sites in yeast proteins



Vertebrate

NetStart 1.0

Prediction Server

The NetStart server produces neural network predictions of translation start in vertebrate and *Arabidopsis thaliana* nucleotide sequences.

NetStart has been trained on cDNA-like sequences and will therefore presumably have better performance for cDNAs and ESTs. We have not tested the performance on genome data which may contain introns adjacent to the start codon.

[Instructions](#)

[Output format](#)

[About](#)

SUBMISSION

Paste a single sequence or several sequences in [FASTA](#) format into the field below:

Submit a file in [FASTA](#) format directly from your local disk:

No file chosen

Vertebrate **A. thaliana**

Restrictions: At most 50 sequences and 1,000,000 nucleotides per submission; each sequence not more than 500,000 nucleotides.

Confidentiality: The sequences are kept confidential and will be deleted after processing.

Translation start predictions for 1 dicot plant sequence

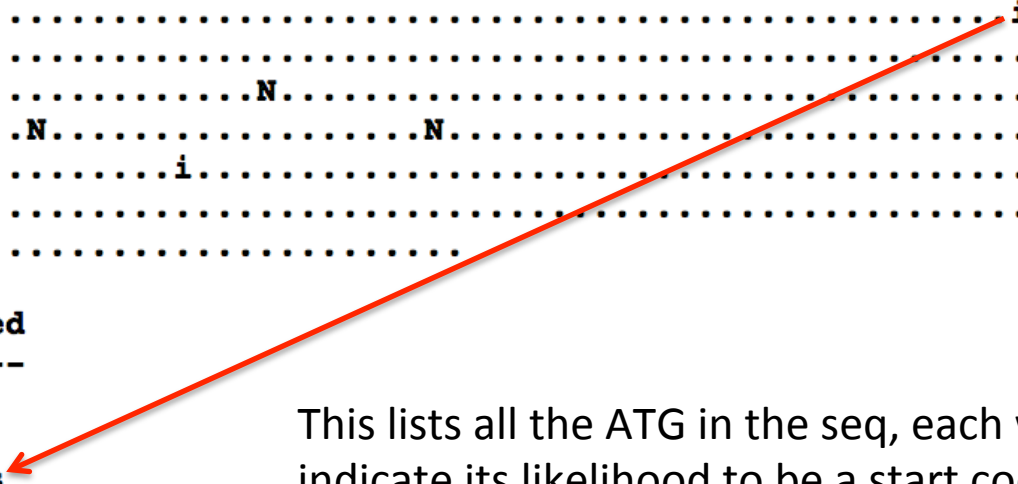
Name: Chaetosphaeridium_gl

CGGGATAAGATGCACGTGGGCTTTCCTAGATCATTGGGCAACAAGCGCATCGTTGGCAACGGTTCTACTCGTTTGTGG
 CGTTACCTGGAAGCATGTCCTCCATCCAGGTCTTGCAACCAGTTTTTCGAACTTTGTGCACAGTTGTTACAGGCTCTTGTT
 CGGGCTCTGGAGGTGCTATTCTGGCACAGCTGGCCGACAAGCTCTTCCTATTACCCGGGCGTCTGTGGATATGGGCCTC
 CAAGTATCGAATGCCTGTCCTCGAGAGAGAGACTGTGCCGCGAGGAGGAGATCAGCTGTCCTTGGTTGCCGAGGAGGGCA
 AGCTGGAGGCGGAGTCTGTCCTGATCCAGATCCCCATGTGCAACGAGAGGGAGTGCTACAAGCAGAGCATCGAGGCGGCA
 TGCAGGCTGACTTGGGAAGGCCGCATGCACGTCCAGGTCCTCGATGACTCAGACGACGAGGAGATCCAGGCGCTCATCAG
 AGACGAGGTCGGCAGGTGGCAGCAGAAGAACATGAGCATCGTATATCTCCACCGGACAACCCGGACTGGCCACAAGGCCG
 GCAACCTCGCGTTTGGGATGGCGGAGGCCGAGCCCCAGGGCTTCCAGTTTGTGGTGATCTTTGACGCGGACTTCATGCC
 AGCCCCGACTTCCTGCAGCGCACCGTGCCGCTCTTCCGCGGCCGC

.....N.....
N.....
i.....
i.....
N.....
N.....
i.....
i.....
N.....

Pos	Score	Pred
10	0.435	-
95	0.433	-
232	0.845	Yes
251	0.833	Yes
356	0.237	-
400	0.471	-
425	0.280	-
444	0.345	-
512	0.670	Yes
578	0.638	Yes
635	0.142	-

This lists all the ATG in the seq, each was scored to indicate its likelihood to be a start codon



NUCLEOTIDE SEQUENCES

Whole genome visualization and analysis

« [GenomeAtlas](#) »
DNA structural atlases for complete microbial Genomes

Gene finding and splice sites

« [EasyGene](#) »
Genes in prokaryotes

« [EasyGene](#) »
Genes in prokaryotes

« [HMMgene](#) »
Genes in eukaryotes

[MetaRanker](#)
Identification of risk genes in complex phenotypes

[NetAspGene](#)
Intron splice sites in Aspergillus DNA

« [NetGene2](#) »
Intron splice sites in human, C. elegans and A. thaliana DNA

[NetPlantGene](#)
Intron splice sites in Arabidopsis thaliana DNA

« [NetStart](#) »
Translation start in vertebrate and A. thaliana DNA

[NetUTR](#)
Splice sites in 5' UTR regions of human genes

« [Promoter](#) »
Transcription start sites in vertebrate DNA

« [RNAmer](#) »
Ribosomal RNA sub units

« [RNAmer](#) »
Ribosomal RNA sub units

Analysis of DNA microarray data

[GenePublisher](#)
Analysis of DNA microarray data

« [OligoWiz](#) »
Design of oligonucleotides for DNA microarrays

SMALL MOLECULES

« [ChemProt](#) »
Chemical-protein interactions

AMINO ACID SEQUENCES

Protein sorting

[ChloroP](#) »
Chloroplast transit peptides and their cleavage sites in plant proteins

[LipoP](#) »
Signal peptidase I & II cleavage sites in gram- bacteria

[NetNES](#) »
Leucine-rich nuclear export signals (NES) in eukaryotic proteins

[SecretomeP](#) »
Non-classical and leaderless secretion of proteins

[SignalP](#) »
Signal peptide and cleavage sites in gram+, gram- and eukaryotic amino acid sequences

[TargetP](#) »
Subcellular location of proteins: mitochondrial, chloroplastic, secretory pathway, or other

[TatP](#) »
Twin-arginine signal peptides

Post-translational modifications of proteins

[DictyOGlyc](#)
O-(alpha)-GlcNAc glycosylation sites (trained on *Dictyostelium discoideum* proteins)

[NetAcet](#)
N-terminal acetylation in eukaryotic proteins

[NetCGlyc](#) »
C-mannosylation sites in mammalian proteins

[NetCorona](#)
Coronavirus 3C-like proteinase cleavage sites in proteins

[NetGlycate](#) »
Glycation of ε amino groups of lysines in mammalian proteins

[NetNGlyc](#) »
N-linked glycosylation sites in human proteins

[NetNGlyc](#) »
N-linked glycosylation sites in human proteins

[NetOGlyc](#) »
O-GalNAc (mucin type) glycosylation sites in mammalian proteins

[NetOGlyc](#) »
O-GalNAc (mucin type) glycosylation sites in mammalian proteins

[NetPhorest](#)
Linear motif atlas for phosphorylation-dependent signaling

[NetPhos](#) »
Generic phosphorylation sites in eukaryotic proteins

[NetPhosBac](#)
Generic phosphorylation sites in bacterial proteins

[NetPhosK](#)
Kinase specific phosphorylation sites in eukaryotic proteins

[NetPhosYeast](#)
Serine and threonine phosphorylation sites in yeast proteins

SignalP 4.1 Server

SignalP 4.1 server predicts the presence and location of signal peptide cleavage sites in amino acid sequences from different organisms: Gram-positive prokaryotes, Gram-negative bacteria, and eukaryotes. It provides prediction of cleavage sites and a signal peptide/non-signal peptide prediction based on a combination of several artificial neural networks.

View the [version history](#) of this server. All the previous versions are available on line, for comparison and reference.

New: SignalP has been updated to version 4.1 with two new features:

- an option to choose a D-score cutoff that reproduces the sensitivity of SignalP 3.0 (this will make the false positive rate slightly higher, but still better than the default)
- a customizable minimum length of the predicted signal peptide (default 10).

Additionally, the documentation has been rewritten. The [Instructions](#) page is expanded, the [Output format](#) page has been clarified, and there are new [Performance](#) pages.

FAQ	Article abstracts	Instructions	Output format
---------------------	-----------------------------------	------------------------------	-------------------------------

SUBMISSION

Paste a single amino acid sequence or several sequences in **FASTA** format into the field below:

```
GIDDWWRNEQFWVIGGVSSHLFALFQGLLKVLGAVSTNFTVTSKAADDGEFSELYIFK
WTSLLIPPTLLIINIVGVIVGVSDAINNGYDSWGPLFGRLFFALWVIVHLYPFLKGLLGK
QDRVPTIILVWSILLASILTLLWVRVNPVSKDGPVLEICGLDCLK
```

<http://cys.bios.niu.edu/yyin/teach/PBB/cesa-pr.fa>, copy
paste the 1st seq

Submit a file in **FASTA** format directly from your local disk:

Choose File No file chosen

- Organism group** ([explain](#))
- Eukaryotes
 - Gram-negative bacteria
 - Gram-positive bacteria

- D-cutoff values** ([explain](#))
- Default (optimized for correlation)
 - Sensitive (reproduce SignalP 3.0's sensitivity)
 - User defined:
 - D-cutoff for SignalP-noTM networks
 - D-cutoff for SignalP-TM networks

- Output format** ([explain](#))
- Standard
 - Short (no graphics)
 - Long
 - All - SignalP-noTM and SignalP-TM output (no graphics)

- Method** ([explain](#))
- Input sequences may include TM regions
 - Input sequences do not include TM regions

- Gram**
- M
 - P
 - F

- Pos**
- -

antigen-specific binding of peptides to MHC class I alleles of known sequence
[NNAAlign](#)

Identifying sequence motifs in quantitative peptide data

[VDJsolver](#) »

Analysis of human immunoglobulin VDJ recombination

Protein function and structure

[ArchaeaFun](#)

Enzyme/non-enzyme and enzyme class (Archaea)

[CPHmodels](#)

Protein structure from sequence: distance constraints

[distanceP](#)

Protein distance constraints

[EPIPE](#) »

Functional differences of protein variants

[InterMap3D](#)

Co-evolving amino acids in proteins

[NetSurfP](#) »

Protein secondary structure and relative solvent accessibility

[NetTurnP](#)

β -turns and β -turn types in proteins

[ProtFun](#) »

Protein functional category and enzyme class (Eukarya)

[RedHom](#)

Reduction of sequence similarity in a data set

[TMHMM](#) »

Transmembrane helices in proteins

[VarDom](#)

Domains in the malaria antigen family PfEMP1

TMHMM Server v. 2.0

Prediction of transmembrane helices in proteins

NOTE: You can submit many proteins at once in one fasta file. Please limit each submission to at most 4000 proteins. Please tick the 'C

[Instructions](#)

SUBMISSION

Submission of a local file in **FASTA** format (HTML 3.0 or higher)

No file chosen

OR by pasting sequence(s) in **FASTA** format:

```
INLSDDLHQVLRWALGSVEIFLSRHCPIWYGYGGGLKWLERFSYINSVVYPWTSPLLVYCSLPAI
CLLTGKFIVPEISNYAGILFLMFMSIAVTGILEMQWGKIGIDDWWRNEQFWVIGGVSSHLFALF
QGLLKVLAVSTNFTVTSKAADDGEFSELYIFKWTSLIPPTLLIINIVGVIVGVSDAINNGYDS
WGPLFGRLFFALWVIVHLYPFLKGLLGKQDRVPTIILVWSILLASILLWVRVNPVSKDGPVLEI
CGLDCLK
```

Output format:

- Extensive, with graphics
- Extensive, no graphics
- One line per protein

Other options:

- Use old model (version 1)

PORTABLE VERSION

Next class: ClustalX and MEGA