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

Yanbin Yin

Spring 2013

Homework assignment 3

1. Download http://cys.bios.niu.edu/yyin/teach/PBB/purdue.cellwall.list.lignin.fa 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.

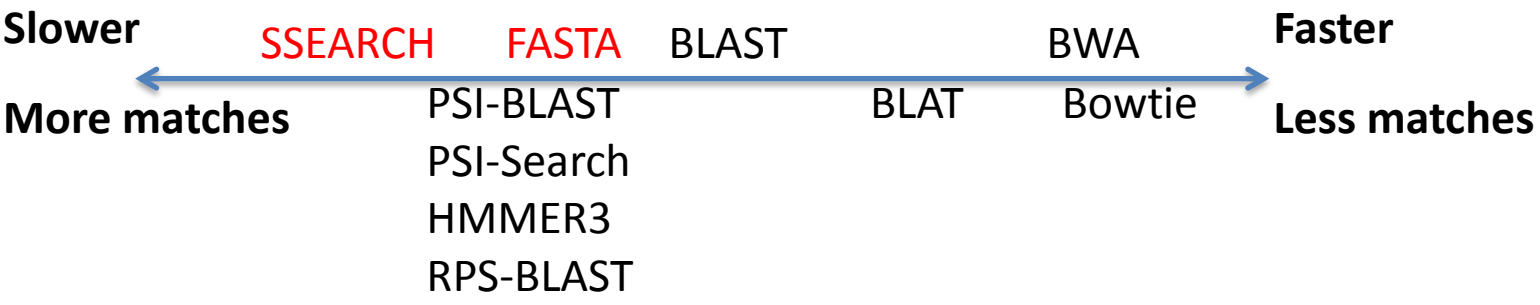
Due on Feb 21 (send by email)

Office hour:
Tue, Thu and Fri 2-4pm, MO325A
Or email: yyin@niu.edu

Outline

- Hands on Practice!

Pairwise alignment tools



EMBL-EBI Enter Text Here Find Terms of Use | Privacy | Cookies

Databases | Tools | Research | Training | Industry | About Us | Help | Site Index

- Tools Home
- Tools A-Z
- ID Mapping
- Literature
- Microarray Analysis
- Protein Functional Analysis
- Proteomic Services
- Sequence Analysis
- Similarity & Homology
- Structural Analysis
- Tools - Miscellaneous
- Web Services

- Databases
- Downloads

EBI > Tools

http://www.ebi.ac.uk/Tools/

Tools at the EBI

We provide a comprehensive range of bioinformatics tools.

This page shows a selection of those that are used most frequently. These include tools for the analysis and comparison of nucleotide and protein sequences, data from functional genomics experiments, text mining of the scientific literature and tools for determination and visualisation of macromolecular structures.

All these tools can be accessed over the web and most provide Web Services interfaces using SOAP or REST APIs.

Nucleotide and Protein sequence searching

Nucleotide sequence searches

The sequence databases that can be searched with the tools outlined below include EMBL-Bank, Coding Sequences, immunoglobulins and High throughput cDNA:

- ENA Search
- BLAST Nucleotide
- Fasta Genomes
- Ssearch Genomes

Protein sequence searches

The protein sequence databases available to search below include UniProtKB, sequences derived from macro molecular structures, immunoglobulins and sequences from patents:

- BLAST Protein
- PSI-Search
- Fasta Proteomes
- Ssearch Proteomes

Multiple Sequence Alignment

Alignment of three or more sequences to identify regions of conservation which may indicate functional constraints and infer evolutionary relationships.

- Clustal Omega

Pairwise Sequence Alignments

Alignment of two sequences to identify regions where the sequence is conserved and conversely regions where the sequence is not conserved.

- Needle
- LAlign

Sequence Similarity Searching

Sequence Similarity Searching is a method of searching sequence databases by using alignment to a query sequence. By statistically assessing how well database and query sequences match one can infer homology and transfer information to the query sequence.

The tools can be launched with different form pre-sets using the buttons - these can be changed on the tool page as well.

BLAST

NCBI BLAST NCBI BLAST (blastall) is the most commonly used sequence similarity search tool. It uses heuristics to perform fast **local** alignment searches.

Protein | Nucleotide | Vectors

WU-BLAST WU-BLAST is similar to NCBI BLAST but combines multiple parameter options into a simpler 'sensitivity' setting.

Protein | Nucleotide | Parasites

PSI-BLAST PSI-BLAST allows users to construct and perform a BLAST search with a custom, position-specific, scoring matrix which can help find distant evolutionary relationships. PHI-BLAST functionality is also available to restrict results using patterns.

Protein

FASTA

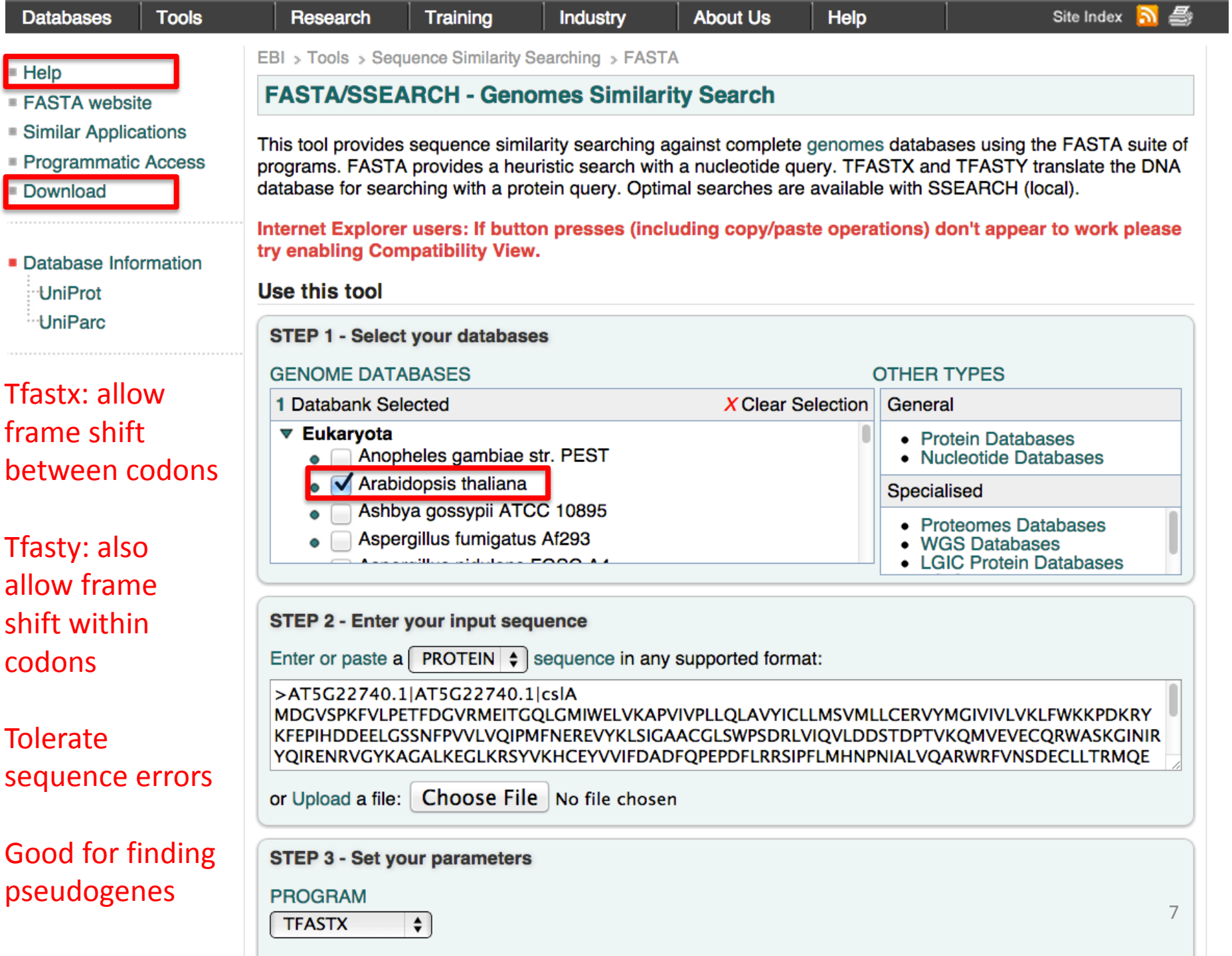
FASTA FASTA is another commonly used sequence similarity search tool which uses heuristics for fast **local** alignment searching.

Protein | Nucleotide | Proteomes | Genomes | Whole Genome Shotgun | LGIC Protein | LGIC Nucleotide

SSEARCH SSEARCH is an optimal (as opposed to heuristics-based) **local** alignment search tool using the Smith-Waterman algorithm. Optimal searches guarantee you find the best alignment score for your given parameters.

Protein | Nucleotide | Proteomes | Genomes | Whole Genome Shotgun | LGIC Protein | LGIC Nucleotide

PSI-Search PSI-Search combines the sensitivity of the Smith-Waterman search algorithm (SSEARCH)

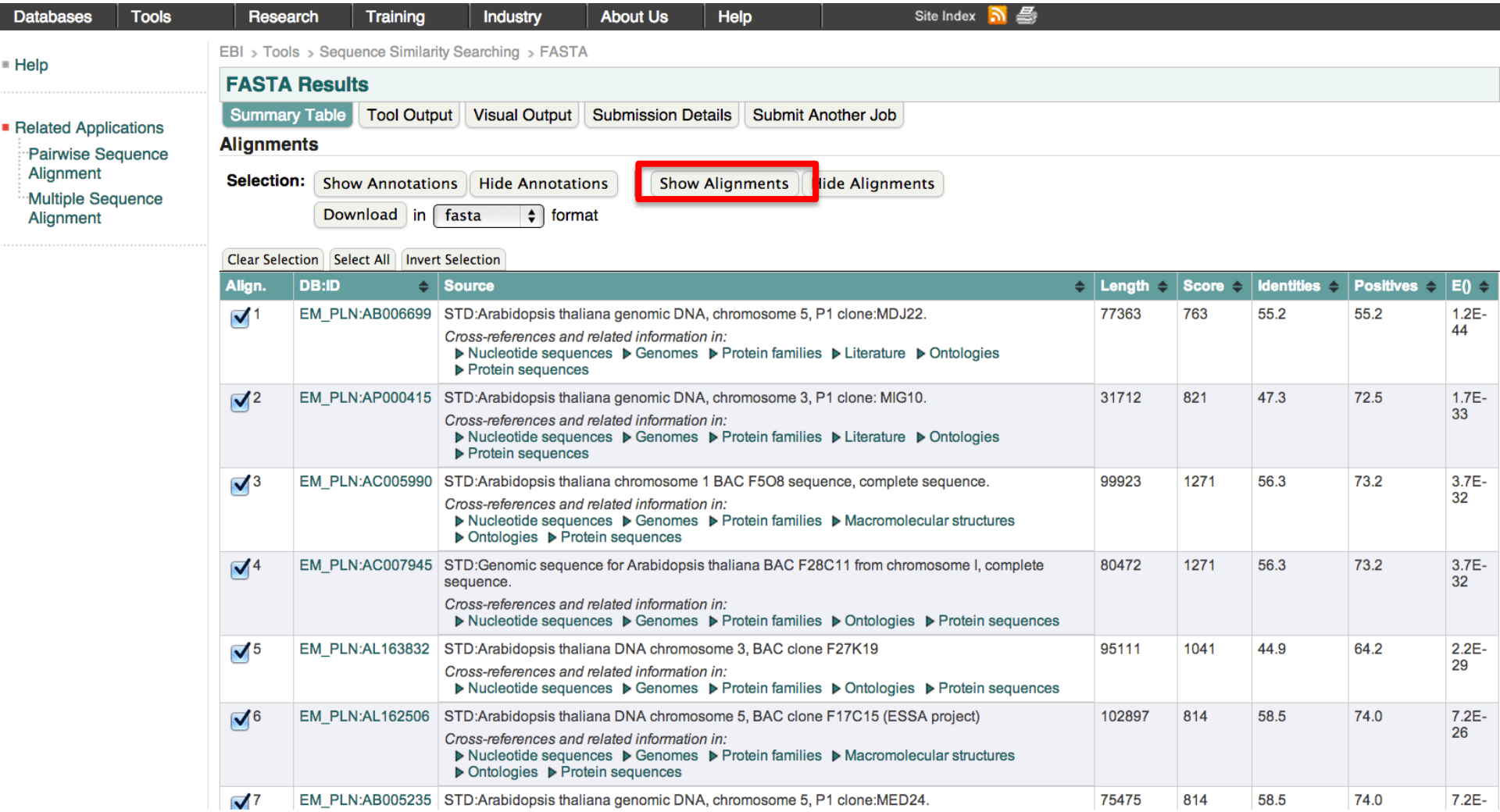


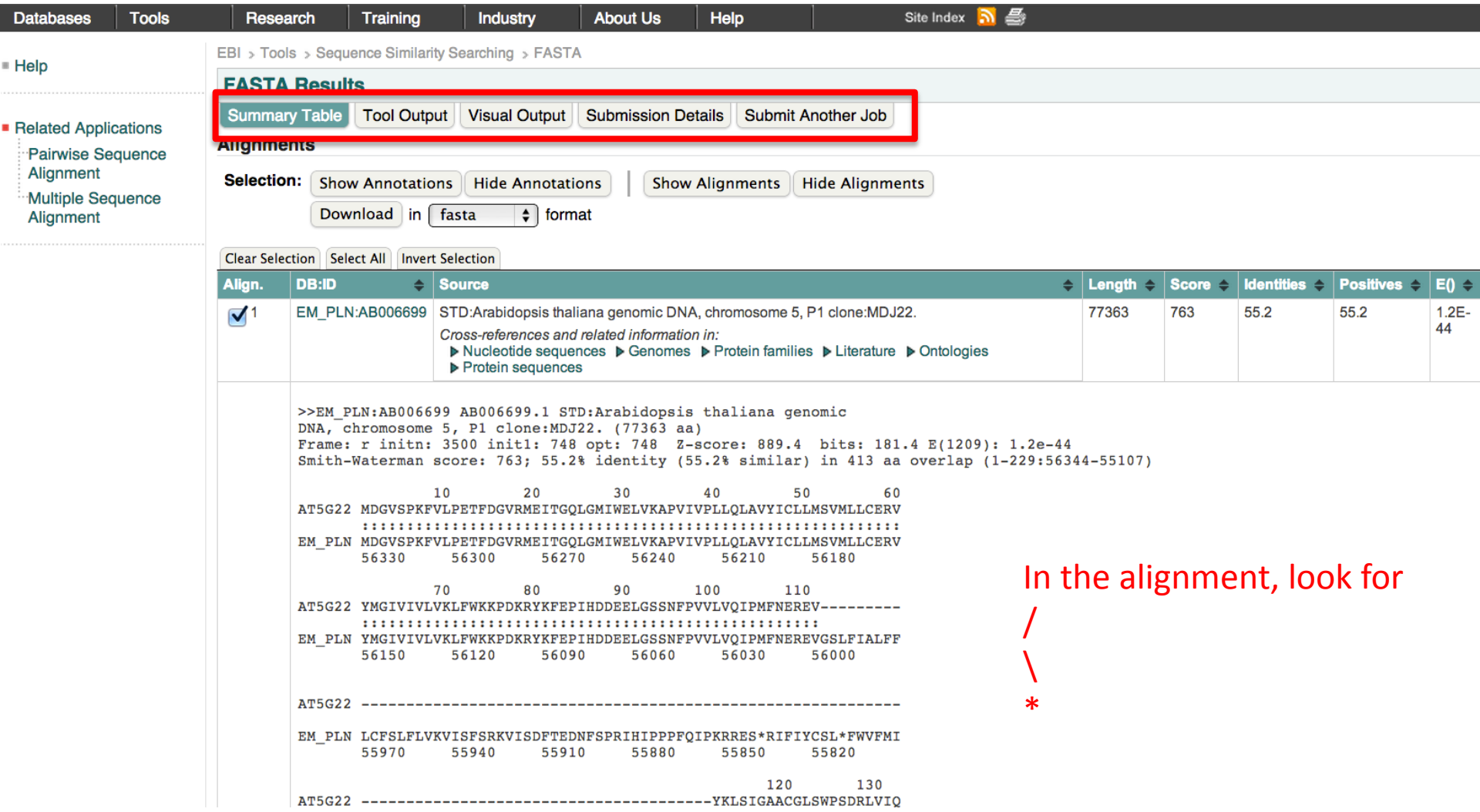
Tfastx: allow frame shift between codons

Tfasty: also allow frame shift within codons

Tolerate sequence errors

Good for finding pseudogenes


Databases
Tools
Research
Training
Industry
About Us
Help
Site Index
Help
Related Applications
Pairwise Sequence Alignment
Multiple Sequence Alignment
EBI › Tools › Sequence Similarity Searching › FASTA
FASTA Results
Summary Table
Tool Output
Visual Output
Submission Details
Submit Another Job
Alignments
Selection:
Show Annotations
Hide Annotations
Show Alignments
Hide Alignments
Download
in
fasta
format
Clear Selection
Select All
Invert Selection
Align.
DB:ID
Source
Length
Score
Identities
Positives
E()
1
EM_PLN:AB006699
STD:Arabidopsis thaliana genomic DNA, chromosome 5, P1 clone:MDJ22.
Cross-references and related information in:
Nucleotide sequences
Genomes
Protein families
Literature
Ontologies
Protein sequences
77363
763
55.2
55.2
1.2E-44
2
EM_PLN:AP000415
STD:Arabidopsis thaliana genomic DNA, chromosome 3, P1 clone: MIG10.
Cross-references and related information in:
Nucleotide sequences
Genomes
Protein families
Literature
Ontologies
Protein sequences
31712
821
47.3
72.5
1.7E-33
3
EM_PLN:AC005990
STD:Arabidopsis thaliana chromosome 1 BAC F5O8 sequence, complete sequence.
Cross-references and related information in:
Nucleotide sequences
Genomes
Protein families
Macromolecular structures
Ontologies
Protein sequences
99923
1271
56.3
73.2
3.7E-32
4
EM_PLN:AC007945
STD:Genomic sequence for Arabidopsis thaliana BAC F28C11 from chromosome I, complete sequence.
Cross-references and related information in:
Nucleotide sequences
Genomes
Protein families
Ontologies
Protein sequences
80472
1271
56.3
73.2
3.7E-32
5
EM_PLN:AL163832
STD:Arabidopsis thaliana DNA chromosome 3, BAC clone F27K19
Cross-references and related information in:
Nucleotide sequences
Genomes
Protein families
Ontologies
Protein sequences
95111
1041
44.9
64.2
2.2E-29
6
EM_PLN:AL162506
STD:Arabidopsis thaliana DNA chromosome 5, BAC clone F17C15 (ESSA project)
Cross-references and related information in:
Nucleotide sequences
Genomes
Protein families
Macromolecular structures
Ontologies
Protein sequences
102897
814
58.5
74.0
7.2E-26
7
EM_PLN:AB005235
STD:Arabidopsis thaliana genomic DNA, chromosome 5, P1 clone:MED24.
75475
814
58.5
74.0
7.2E-



In the alignment, look for

/

\

*

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

BLAST: shorter alignment

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      VNSDECLLTRMQEMSLDYHFTVEQEVGSSTHAFFGFNGTAGIWRIAAINEAGGWKDRTTV  287
                VN++ECL+TRMQEMSL+YHF  EQE GSS HAFFGFNGTAG+WRIAA+NEAGGWKDRTTV
Sbjct  8335055  VNANECLMTRMQEMSLNYHFVAEQESGSSIHAFFGFNGTAGVWRIAALNEAGGWKDRTTV  8335234

Query  288      EDMDLAVRASLRGWKFLYLGDL----------------------------QVKSELPSTF  319
                EDMDLAVRA L GWKF+Y+ D+                            QVK+ELPSTF
Sbjct  8335235  EDMDLAVRACLHGWKFVYVHDVEVLLL*PF*TQAKHNKISFKTLLSNLTFQVKNELPSTF  8335414

Query  320      RAFRFQQHRWSCGPANLFRKMVMEIVRNKKVRFWKKVYVIYSFFFVRKIIAHWVTFCFYC  379
                +A+RFQQHRWSCGPANL+RKM MEI++NKKV  WKK+Y+IY+FFF+RKI+ H  TF FYC
Sbjct  8335415  KAYRFQQHRWSCGPANLWRKMTMEILQNKKVSAWKKLYLIYNFFFIRKIVVHIFTFVFYC  8335594

Query  380      VVLPLTILVPEVKVPIWGSVYIPSIITILNSVGTPRSIHL  419
                ++LP T+L PE++VP W +VY P+ ITILN++ TPR  HL
Sbjct  8335595  LILPTTVLFPELQVPKWATVYFPTTITILNAIATPR*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
                +T DGV R  I G++  IW+  +  V +P+L+  V ICL+MS++L  ERVYM IV+V VK
Sbjct  8334112  DTTDGVVRSGIIGEIIYIWKQTRIFVFIPILKCLVTICLVMSLLLFIERVYMSIVVVFVK  8334291

Query  71       LFWKKPDKRYKFEPIHDDE-ELGSSNFPVVLVQIPMFNEREVY----------------  112
                L  + P+K +K+EPI+DD+ EL ++N+P+VL+QIPM+NE+EV
```

- Help
- FASTA website
- Similar Applications
- Programmatic Access
- Download

- Database Information
  - UniProt
  - UniParc

FASTA/SSEARCH - Proteome Similarity Search

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

Internet Explorer users: If button presses (including copy/paste operations) don't appear to work please try enabling Compatibility View.

Use this tool

STEP 1 - Select your databases

PROTEIN DATABASES

0 Databank Selected — X Clear Selection

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

OTHER TYPES

General

- Nucleotide Databases
- Protein Databases

Specialised

- Genomes Databases
- WGS Databases
- LGIC Protein Databases

STEP 2 - Enter your input sequence

Enter or paste a PROTEIN sequence in any supported format:

or upload a file: Choose File No file chosen

STEP 3 - Set your parameters

PROGRAM

SSEARCH

EMBL-EBI Enter Text Here Find Terms of Use | Privacy | Cookies

Databases | Tools | Research | Training | Industry | About Us | Help | Site Index

- Tools Home
- Tools A-Z
- ID Mapping
- Literature
- Microarray Analysis
- Protein Functional Analysis
- Proteomic Services
- Sequence Analysis
- Similarity & Homology
- Structural Analysis
- Tools - Miscellaneous
- Web Services

- Databases
- Downloads

EBI > Tools

http://www.ebi.ac.uk/Tools/

Tools at the EBI

We provide a comprehensive range of bioinformatics tools.

This page shows a selection of those that are used most frequently. These include tools for the analysis and comparison of nucleotide and protein sequences, data from functional genomics experiments, text mining of the scientific literature and tools for determination and visualisation of macromolecular structures.

All these tools can be accessed over the web and most provide Web Services interfaces using SOAP or REST APIs.

Nucleotide and Protein sequence searching

Nucleotide sequence searches

The sequence databases that can be searched with the tools outlined below include EMBL-Bank, Coding Sequences, immunoglobulins and High throughput cDNA:

- ENA Search
- BLAST Nucleotide
- Fasta Genomes
- Ssearch Genomes

Protein sequence searches

The protein sequence databases available to search below include UniProtKB, sequences derived from macro molecular structures, immunoglobulins and sequences from patents:

- BLAST Protein
- PSI-Search
- Fasta Proteomes
- Ssearch Proteomes

Multiple Sequence Alignment

Alignment of three or more sequences to identify regions of conservation which may indicate functional constraints and infer evolutionary relationships.

Pairwise Sequence Alignments

Alignment of two sequences to identify regions where the sequence is conserved and conversely regions where the sequence is not conserved.

- Needle

- Tools Home
- Tools A-Z
- Web Services
- Download

Pairwise Sequence Alignment

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

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

Global Alignment

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

Needle (EMBOSS) — EMBOSS Needle creates an optimal global alignment of two sequences using the Needleman-Wunsch algorithm.
Protein | Nucleotide

Needleman-wunsch algorithm

Stretcher (EMBOSS) — EMBOSS Stretcher uses a modification of the Needleman-Wunsch algorithm that allows larger sequences to be globally aligned.
Protein | Nucleotide

Local Alignment

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

Water (EMBOSS) — EMBOSS Water uses the Smith-Waterman algorithm (modified for speed enhancements) to calculate the local alignment of two sequences.
Protein | Nucleotide

Matcher (EMBOSS) — EMBOSS Matcher identifies local similarities between two sequences using a rigorous algorithm based on the LALIGN application.
Protein | Nucleotide

LALIGN — LALIGN finds internal duplications by calculating non-intersecting local alignments of protein or DNA sequences.
Protein | Nucleotide

Genomic Alignment

Help
EMBOSS website
Programmatic Access
Download

Nucleotide form
Similar Applications

CslA: 539 aa
CesA: 1089 aa

EMBOSS Needle - Pairwise Sequence Alignment (PROTEIN)

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

This is the form for protein sequences. Please go to the nucleotide form if you wish to align DNA or RNA sequences.

Internet Explorer users: If button presses (including copy/paste operations) don't appear to work please try enabling Compatibility View.

Use this tool

STEP 1 - Enter your protein sequences

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

```
>AT5G22740.1|AT5G22740.1|cslA
MDGVSPKFVLPETFDGVRMEITGQLGMIWELVKAPVIVPLLQLAVYICLLMSVMLLCERVYMGIVIVLVKLFWKKPDKRY
KFEPIHDDEELGSSNFPVVLVQIPMFNEREVYKLSIGAACGLSWPSDRLVIQVLDDSTDPTVKQMVEVECQRWASKGINIR
YQIRENRVGYKAGALKEGLKRSYVKHCEYVVIFDADFQPEPDFLRRSIPFLMHNPNIALVQARWRFVNSDECLLTRMQE
MSLDYHFTVEQEVGSSTHAFFGFNGTAGIWRIAAINEAGGWKDRTTVEDMDLAVRASLRGWKFLYLGDLQVKSELPST
FRAFRFQQHRWSCGPANLFRKMVMEIVRNKKVRFWKKVYVIYSFFFVRKIIAHWVTFCFYCVVLPLTILVPEVKVPIWGSV
YIPSIITILNSVGTPRSIHLLFYWILFENVMSLHRTKATLIGLFEAGRANEWVVTAKLGSGQSAKGNTKGIKRFPRIFKLPDRL
```

Or, upload a file: Choose File No file chosen

AND

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

```
>AT2G21770.1|AT2G21770.1|cesA
MNTGGRLIAGSHNRNEFVLINADDTARIRSAEELSGQTCKICRDEIELTDNGEPFIACNECAFPTCRPCYEYERREGNQA
CPQCGTRYKRIKGSPRVEGDEEDDDIDDLEHEFYGMDPEHVTEAALYYMRLNTGRGTDEVSHLYSASPGSEVPLLTYCD
EDSDMYSDRHALIVPPSTGLGNRVHHVPFTDSFASIHTRPMVPQKDLTVYGYGSVAWKDRMEVWKKQQIEKLQVVKN
ERVNDGDGDGFIVDELDDPGLPMMDEGRQPLSRKLPIRSSRINPYRMLIFCRLAILGLFFHYRILHPVNDAFGLWLTSVICE
IWFAVSWILDQFPKWYPIERETYLDRLSLRYEKEGKPSELAPVDVFVSTVDPLKEPPLITANTVLSILAVDYPVEKVACYVSD
DGAAMLTFEALSYTAEFARKWVPFCKKFSIEPRAPEWYFSQKMDYLKHKVDPAFVMERRAMKRDYEEFKVRINALVSVS
```

Or, upload a file: Choose File No file chosen

STEP 2 - Set your pairwise alignment options

```
#=======================================
#
# Aligned_sequences: 2
# 1: cslA
# 2: cesA
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 1162
# Identity:     125/1162 (10.8%)
# Similarity:   219/1162 (18.8%)
# Gaps:         700/1162 (60.2%)
# Score: 45.5
#
#
#=======================================

cslA               1 --------------------------------------------------      0

cesA               1 MNTGGRLIAGSHNRNEFVLINADDTARIRSAEELSGQTCKICRDEIELTD     50

cslA               1 --------------------------------------------------      0

cesA              51 NGEPFIACNECAFPTCRPCYEYERREGNQACPQCGTRYKRIKGSPRVEGD    100

cslA               1 --------------------------------------------------      0

cesA             101 EEDDDIDDLEHEFYGMDPEHVTEAALYYMRLNTGRGTDEVSHLYSASPGS    150

cslA               1 --------------------------------------------------      0

cesA             151 EVPLLTYCDEDSDMYSDRHALIVPPSTGLGNRVHHVPFTDSFASIHTRPM    200

cslA               1 --------------------------------------------------      0

cesA             201 VPQKDLTVYGYGSVAWKDRMEVWKKQQIEKLQVVKNERVNDGDGDGFIVD    250

cslA               1 --------------------------------------------------      0

cesA             251 ELDDPGLPMMDEGRQPLSRKLPIRSSRINPYRMLIFCRLAILGLFFHYRI    300

cslA               1 --------------------------------------------------      0

cesA             301 LHPVNDAFGLWLTSVICEIWFAVSWILDQFPKWYPIERETYLDRLSLRYE    350
```

Help
EMBOSS website
Programmatic Access
Download

Nucleotide form
Similar Applications

EMBOSS Water - Pairwise Sequence Alignment (PROTEIN)

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

This is the form for protein sequences. Please go to the nucleotide form if you wish to align DNA or RNA sequences.

Internet Explorer users: If button presses (including copy/paste operations) don't appear to work please try enabling Compatibility View.

Use this tool

STEP 1 - Enter your protein sequences

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

```
MDGVSPKFVLPETFDGVRMEITGQLGMIWELVKAPVIVPLLQLAVYICLLMSVMLLCERVYMGIVIVLVKLFWKKPDKRY
KFEPIHDDEELGSSNFPVVLVQIPMFNEREVYKLSIGAACGLSWPSDRLVIQVLDDSTDPTVKQMVEVECQRWASKGINIR
YQIRENRVGYKAGALKEGLKRSYVKHCEYVVIFDADFQPEPDFLRRSIPFLMHNPNIALVQARWRFVNSDECLLTRMQE
MSLDYHFTVEQEVGSSTHAFFGFNGTAGIWRIAAINEAGGWKDRTTVEDMDLAVRASLRGWKFLYLGDLQVKSELPST
FRAFRFQQHRWSCGPANLFRKMVMEIVRNKKVRFWKKVYVIYSFFFVRKIIAHWVTFCFYCVVLPLTILVPEVKVPIWGSV
YIPSIITILNSVGTPRSIHLLFYWILFENVMSLHRTKATLIGLFEAGRANEWVVTAKLGSGQSAKGNTKGIKRFPRIFKLPDRL
NTLELGFAAFLFVCGCYDFVHGKNNYFIYLFLQTMSFFISGLGWIGTYVPS*
```

Or, upload a file: Choose File No file chosen

AND

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

```
>AT2G21770.1|AT2G21770.1|cesA
MNTGGRLIAGSHNRNEFVLINADDTARIRSAEELSGQTCKICRDEIELTDNGEPFIACNECAFPTCRPCYEYERREGNQA
CPQCGTRYKRIKGSPRVEGDEEDDDIDDLEHEFYGMDPEHVTEAALYYMRLNTGRGTDEVSHLYSASPGSEVPLLTYCD
EDSDMYSDRHALIVPPSTGLGNRVHHVPFTDSFASIHTRPMVPQKDLTVYGYGSVAWKDRMEVWKKQQIEKLQVVKN
ERVNDGDGDGFIVDELDDPGLPMMDEGRQPLSRKLPIRSSRINPYRMLIFCRLAILGLFFHYRILHPVNDAFGLWLTSVICE
IWFAVSWILDQFPKWYPIERETYLDRLSLRYEKEGKPSELAPVDVFVSTVDPLKEPPLITANTVLSILAVDYPVEKVACYVSD
DGAAMLTFEALSYTAEFARKWVPFCKKFSIEPRAPEWYFSQKMDYLKHKVDPAFVMERRAMKRDYEEFKVKINALVSVS
```

Or, upload a file: Choose File No file chosen

STEP 2 - Set your pairwise alignment options

```
#=======================================
#
# Aligned_sequences: 2
# 1: cslA
# 2: cesA
# 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
#
#
#=======================================

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

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

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

cslA             407 I-LNSVGTPRSIHLLFYWILFENVMSLHRTKATLIGLFE------AGRAN    449
                     : ...:|.........:|::  ..:|.|     |..||:      ||.:.
cesA             918 MQWGKIGIDDWWRNEQFWVI--GGVSSH-----LFALFQGLLKVLAGVST    960

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


#---------------------------------------
#---------------------------------------
```

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

Align Sequences Protein BLAST

blastn | blastp | blastx | tblastn | tblastx

Enter Query Sequence

BLASTP programs search protein subjects using a protein query. more...

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

```
FNGTAGIWRIAAINEAGGWKDRTTVEDMDLAVRASLRGWKFLYLGDLQVKSELPSTFRAFRFQQ
HRWSCGPANLFRKMVMEIVRNKKVRFWKKVYVIYSFFFVRKIIAHWVTFCFYCVVLPLTILVPEVK
VPIWGSVYIPSIITILNSVGTPRSIHLLFYWILFENVMSLHRTKATLIGLFEAGRANEWVVTAKLGSG
QSAKGNTKGIKRFPRIFKLPDRLNTLELGFAAFLFVCGCYDFVHGKNNYFIYLFLQTMSFFISGLG
WIGTYVPS*
```

Query subrange

From

To

Or, upload file Choose File No file chosen

Job Title

Enter a descriptive title for your BLAST search

blast2seq

Align two or more sequences

Enter Subject Sequence

Enter accession number, gi, or FASTA sequence Clear

```
>AT2G21770.1|AT2G21770.1|cesA
MNTGGRLIAGSHNRNEFVLINADDTARIRSAEELSGQTCKICRDEIELTDNGEPFIACNECAFPTC
RPCYEYERREGNQACPQCGTRYKRIKGSPRVEGDEEDDDIDDLEHEFYGMDPEHVTEAALYYMR
LNTGRGTDEVSHLYSASPGSEVPLLTYCDEDSDMYSDRHALIVPPSTGLGNRVHHVPFTDSFASI
HTRPMVPQKDLTVYGYGSVAWKDRMEVWKKQQIEKLQVVKNERVNDGDGDGFIVDELDDPGL
```

Subject subrange

From

To

Or, upload file Choose File No file chosen

Program Selection

Algorithm

blastp (protein-protein BLAST)

Choose a BLAST algorithm

BLAST Search **protein sequence** using **Blastp (protein-protein BLAST)**

Download Graphics Sort by: E value

AT2G21770.1|AT2G21770.1|cesA

Sequence ID: lcl|34283 Length: 1089 Number of Matches: 4

Fragmented alignments

Range 1: 779 to 838 Graphics — Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps
23.5 bits(49)	0.044	Compositional matrix adjust.	14/60(23%)	22/60(36%)	2/60(3%)

```
Query  280  GWKDRTTVEDMDLAVRASLRGWKFLYLGDLQV--KSELPSTFRAFRFQQHRWSCGPANLF  337
            GW    +  ED+     +     GW+ +Y     +    K    P          Q  RW+ G    +F
Sbjct  779  GWIYGSVTEDILTGFKMHCHGWRSVYCMPKRAAFKGSAPINLSDRLHQVLRWALGSVEIF  838
```

Range 2: 414 to 426 Graphics — Next Match Previous Match First Match

Score	Expect	Method	Identities	Positives	Gaps
17.3 bits(33)	3.5	Compositional matrix adjust.	8/17(47%)	9/17(52%)	4/17(23%)

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

Range 3: 777 to 798 Graphics — Next Match Previous Match First Match

Score	Expect	Method	Identities	Positives	Gaps
16.5 bits(31)	5.7	Compositional matrix adjust.	5/22(23%)	11/22(50%)	0/22(0%)

```
Query  168  RVGYKAGALKEGLKRSYVKHCE  189
             +G+  G++ E +    +  HC
Sbjct  777  EIGWIYGSVTEDILTGFKMHCH  798
```

Range 4: 25 to 43 Graphics — Next Match Previous Match First Match

Score	Expect	Method	Identities	Positives	Gaps
15.8 bits(29)	9.2	Compositional matrix adjust.	7/19(37%)	10/19(52%)	0/19(0%)

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

- Help
- FAQ
- Programmatic Access
- Similar Applications

- Database Information
  - UniProt
  - UniParc

PSI-Search - Protein Similarity Search

PSI-Search combines the Smith-Waterman search algorithm with the PSI-BLAST profile construction strategy to find distantly related protein sequences.

N.B. PSI-SEARCH does not accept sequences containing bases containing translation STOP residues represented with the '*' (asterisk) character in the sequence.

Internet Explorer users: If button presses (including copy/paste operations) don't appear to work please try enabling Compatibility View.

Use this tool

STEP 1 - Select your database

PROTEIN DATABASES

UniProt Knowledgebase

STEP 2 - Enter your input sequence

Enter or paste a PROTEIN sequence in any supported format:

Upload a file: Choose File No file chosen

STEP 3 - Set your parameters

PSSM E-VALUE CUT-OFF

1.0e-3

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.)*

Multiple sequence alignment tools

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

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**.

COBALT@NCBI

Help
FAQ
Clustal website
Jalview
Programmatic Access
Download

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

ClustalW2 - Multiple Sequence Alignment

ClustalW2 is a general purpose multiple sequence alignment program for DNA or proteins.

Note: **ClustalW2 is no longer being maintained**. Please consider using the new version instead: Clustal Omega

Internet Explorer users: If button presses (including copy/paste operations) don't appear to work please try enabling Compatibility View.

Use this tool

STEP 1 - Enter your input sequences

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

```
>AT2G21770.1|AT2G21770.1|cesA
MNTGGRLIAGSHNRNEFVLINADDTARIRSAEELSGQTCKICRDEIELTDNGEPFIACNECAFPTCRPCYEYERREGNQA
CPQCGTRYKRIKGSPRVEGDEEDDDIDDLEHEFYGMDPEHVTEAALYYMRLNTGRGTDEVSHLYSASPGSEVPLLTYCD
EDSDMYSDRHALIVPPSTGLGNRVHHVPFTDSFASIHTRPMVPQKDLTVYGYGSVAWKDRMEVWKKQQIEKLQVVKN
ERVNDGDGDGFIVDELDDPGLPMMDEGRQPLSRKLPIRSSRINPYRMLIFCRLAILGLFFHYRILHPVNDAFGLWLTSVICE
IWFAVSWILDQFPKWYPIERETYLDRLSLRYEKEGKPSELAPVDVFVSTVDPLKEPPLITANTVLSILAVDYPVEKVACYVSD
DGAAMLTFEALSYTAEFARKWVPFCKKFSIEPRAPEWYFSQKMDYLKHKVDPAFVMERRAMKRDYEEFKVRINALVSVS
```

Or, upload a file: Choose File No file chosen

STEP 2 - Set your Pairwise Alignment Options

Alignment Type: (•) Slow ( ) Fast

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 - Set your Multiple Sequence Alignment Options

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 4 - Submit your job

ClustalW2 Results

Alignments | Result Summary | Guide Tree | Submission Details | Submit Another Job

Alignment

Download Alignment File | Hide Colors

```
CLUSTAL 2.1 multiple sequence alignment


AT1G02730.1|AT1G02730.1|cslD       -------MVKSAASQSPSPVTITVTPCKGSGDRSLGLTSPIPRASVITNQ 43
os_42915|LOC_Os07g36610.1|cslF     -------------------------------------------------- 
AT2G21770.1|AT2G21770.1|cesA       MNTGGRLIAGSHNRNEFVLINADDTARIRSAEELSGQTCKICRDEIELTD 50
AT2G32530.1|AT2G32530.1|cslB       -------------------------------------------------- 
os_25268|LOC_Os04g35020.1|cslH     -------------------------------------------------- 
AT1G55850.1|AT1G55850.1|cslE       -------------------------------------------------- 
AT4G23990.1|AT4G23990.1|cslG       -------------------------------------------------- 
AT5G22740.1|AT5G22740.1|cslA       -------------------------------------------------- 
AT2G24630.1|AT2G24630.1|cslC       -------------------------------------------------- 


AT1G02730.1|AT1G02730.1|cslD       N------SPLSSRATRRTSISSGNRRSNGDEGRYCSMSVEDLTAETTNSE 87
os_42915|LOC_Os07g36610.1|cslF     -------MALSPAAAGRT-----GRNNNNDAG------------------ 20
AT2G21770.1|AT2G21770.1|cesA       NGEPFIACNECAFPTCRPCYEYERREGNQACPQCG--------------- 85
AT2G32530.1|AT2G32530.1|cslB       -------------------------------------------------- 
os_25268|LOC_Os04g35020.1|cslH     -------------------------------------------------- 
AT1G55850.1|AT1G55850.1|cslE       -------------------------------------------------- 
AT4G23990.1|AT4G23990.1|cslG       -------------------------------------------------- 
AT5G22740.1|AT5G22740.1|cslA       -------------------------------------------------- 
AT2G24630.1|AT2G24630.1|cslC       -------------------------------------------------- 


AT1G02730.1|AT1G02730.1|cslD       CVLSYTVHIPPTPDHQTVFASQESEEDEMLKGNSNQKSFLSGTIFTGGFK 137
os_42915|LOC_Os07g36610.1|cslF     -------------------------------------------------- 
AT2G21770.1|AT2G21770.1|cesA       ---------------TRYKRIKGSPRVEGDEEDDDIDDLEHEFYGMDPE 119
AT2G32530.1|AT2G32530.1|cslB       -------------------------------------------------- 
os_25268|LOC_Os04g35020.1|cslH     -------------------------------------------------- 
AT1G55850.1|AT1G55850.1|cslE       -------------------------------------------------- 
```

Help
MAFFT website
Jalview
Programmatic Access
Download

Related Applications
Pairwise Sequence Alignment
Multiple Sequence Alignment
Phylogeny

MAFFT - Multiple Sequence Alignment

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

Internet Explorer users: If button presses (including copy/paste operations) don't appear to work please try enabling Compatibility View.

Use this tool

STEP 1 - Enter your input sequences

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

```
>AT2G21770.1|AT2G21770.1|cesA
MNTGGRLIAGSHNRNEFVLINADDTARIRSAEELSGQTCKICRDEIELTDNGEPFIACNECAFPTCRPCYEYERREGNQA
CPQCGTRYKRIKGSPRVEGDEEDDDIDDLEHEFYGMDPEHVTEAALYYMRLNTGRGTDEVSHLYSASPGSEVPLLTYCD
EDSDMYSDRHALIVPPSTGLGNRVHHVPFTDSFASIHTRPMVPQKDLTVYGYGSVAWKDRMEVWKKQQIEKLQVVKN
ERVNDGDGDGFIVDELDDPGLPMMDEGRQPLSRKLPIRSSRINPYRMLIFCRLAILGLFFHYRILHPVNDAFGLWLTSVICE
IWFAVSWILDQFPKWYPIERETYLDRLSLRYEKEGKPSELAPVDVFVSTVDPLKEPPLITANTVLSILAVDYPVEKVACYVSD
DGAAMLTFEALSYTAEFARKWVPFCKKFSIEPRAPEWYFSQKMDYLKHKVDPAFVMERRAMKRDYEEFKVKINALVSVS
```

Or upload a file: Choose File No file chosen

STEP 2 - Set your Parameters

OUTPUT FORMAT Pearson/FASTA

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

Help
Jalview

Related Applications
Multiple Sequence Alignment
Phylogeny

MAFFT Results

Alignments | Result Summary | Guide Tree | Submission Details | Submit Another Job

Alignment

Download Alignment File

```
>AT2G21770.1|AT2G21770.1|cesA
M---------------------------------------NTGGRLIAGSHNRNEFVLI
N-----------------------------------------------------------
----------------------ADDTARIRSAEELSGQTC--KICRDEIELTDNGEPF
IACNECAFPTCRPCYEYERREGNQACPQCGTRYKRIKGSPRVEGDEEDDDIDDLEHEFYG
MDPEHVTEAALYYMRLNTGRGTDEVSHLYSASPGSEVPLLTYCDEDSDMYSDRHALIVPP
STGLGNRVHHVPFTDSFASI-------HTRPMVPQKDLTVYGYGSVAW-KDRMEVWKKQQ
IEKLQVVKNERVNDGDGDGFIVDELDDPGLPMMDEG-RQP-----LSRKLPIRSSRINPY
RMLIFCRLAILGLFFHYRILHPVNDAF----GLWLTSVICEIWFAVSWILDQFPKWYPIE
RETYLDRL----------SLRYEKEG-----KPSELAPVDV-FVSTVDPLKEPPLITANT
VLSILAVDYP--VEKVACYVSDDGAAMLTFEALSYTAEFARKWVP-FCKKFSIEPRAPEW
YFSQKMD-YLKHKVDPAFVMERRAMKRDYEEFKVKINALVSVSQK---------------
----------------VPEDGWTMQDGTPWPG--------NNVRDHPGMIQVFLGHSGVC
DMDGNE--------------LPRLVYVSREKRPGFDHHKKAGAMNSLIRVSAVLSNAPY
LLNVDCDHYINNSKAIREAMCFMMDPQ-SGKKICYVQFPQRFDGIDRH--DRYSNRNVVF
FDINMKGLDGIQGPI--YVGTGCVFRRQALYGFDAPKKKQPPGRTCNCWPKWCCLCCGMR
KKKTGKVKDNQRKKPKETSKQIHALEHIEEGLQVTNAENNSETAQLKLEKKFGQSPVLVA
STLL----------LNG----------GVPSNVNPASLLRESIQVISCGYEEKTEWGKE
IGWIYGSVTEDILTGFKMHCHGWRSVYCMPKRAAFKGSAPINLSDRLHQVLRWALGSVEI
F--------LSRHCPIWYGYGGGLKWLERFSYINSVVYPWTSLPLLVYCSLPAICL-LTG
KFIVPEI-SNYAGILFLLMFMSIAVTGILEMQWGKIGIDDWWRNEQFWVIGGVSSHLFAL
FQGLLKVLAGVS-TNFTVT---------SKAAD----D-----GEFSELY---IFKW---
TSLL----------IPPTTLLIINIVGVIVGVSDAINNGY--DSWGPLFGRLFFALWVIV
HLY-----PFLKGLLGKQD--RVPTIILVW---------SILLASILTLL-------WV-
------RVNPFVSKDGPVLEICGLDCLK----*
>AT1G02730.1|AT1G02730.1|cslD
```

Help
MUSCLE website
Jalview
Programmatic Access
Download

Related Applications
Pairwise Sequence Alignment
Multiple Sequence Alignment
Phylogeny

MUSCLE - Multiple Sequence Alignment

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

Internet Explorer users: If button presses (including copy/paste operations) don't appear to work please try enabling Compatibility View.

Use this tool

STEP 1 - Enter your input sequences

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

```
DDWTFMKDEYDKLVRRIKNTDERSLLRHGGGEFFAEFLNVERRNHPTIVKTRVSAVMTNAPIMLNMDCDMFVNNPQA
VLHAMCLLLGFDDEASSGFVQAPQRFYDALKDDPFGNQMECFFKRFISGVQGVQGAFYAGTGCFHRRKAVYGVPPNF
NGAEREDTIGSSSYKELHTRFGNSEELNESARNIIWDLSSKPMVDISSRIEVAKAVSACNYDIGTCWGQEVGWVYGSLTED
ILTGQRIHAMGWRSVLMVTEPPAFMGSAPIGGPACLTQFKRWATGQSEIIISRNNPILATMFKRLKFRQCLAYLIVLGWPL
RAPFELCYGLLGPYCILTNQSFLPKASEDGFSVPLALFISYNTYNFMEYMACGLSARAWWNNHRMQRIISVSAWTLAFLT
VLLKSLGLSETVFEVTGKDKSMSDDDDNTDGADPGRFTFDSLPVFIPVTALAMLNIVAVTVGACRVAFGTAEGVPCAPG
IGEFMCCGWLVLCFFPFVRGIVWGKGSYGIPWSVKLKASLLVAMFVTFCKRN
```

Or upload a file: Choose File No file chosen

STEP 2 - Set your Parameters

OUTPUT FORMAT: 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

MUSCLE Results

Alignment

Download Alignment File | Show Colors

```
MUSCLE (3.8) multiple sequence alignment


AT5G22740.1|AT5G22740.1|cslA       ------------------------------------------------------------
AT2G24630.1|AT2G24630.1|cslC       ----------------------------------MAPRFDFSDLWAKETRRG-----
os_25268|LOC_Os04g35020.1|cslH     ---------------------------------------MAVVAAAAATGST-----
AT2G32530.1|AT2G32530.1|cslB       ------------------------------------------MADSSSSL-----
AT2G21770.1|AT2G21770.1|cesA       ------------------------MNTGGRLIAGSHNRNEFVLINADDTARI-----
AT1G02730.1|AT1G02730.1|cslD       MVKSAASQSPSPVTITVTPCKGSGDRSLGLTSPIPRASVITNQNSPLSSRATRRTSISSG
os_42915|LOC_Os07g36610.1|cslF     ---------------------------------------MALSPAAAGRTG----
AT4G23990.1|AT4G23990.1|cslG       ----------------------------------MYQVSLKQFVFLLKIKSTTM-----
AT1G55850.1|AT1G55850.1|cslE       ------------------------------------------MVNKDDRI-----


AT5G22740.1|AT5G22740.1|cslA       ------------------------------------------------------------
AT2G24630.1|AT2G24630.1|cslC       ------------------------------------------------------------
os_25268|LOC_Os04g35020.1|cslH     ------------------------------------------------------------
AT2G32530.1|AT2G32530.1|cslB       ------------------------------------------------------------
AT2G21770.1|AT2G21770.1|cesA       -RSAEELSGQTCKICRDEIELTDNGEPFIACNECAFPTC---RPCYEYERREGNQACPQC
AT1G02730.1|AT1G02730.1|cslD       NRRSNGDEGRYCSMSVEDLTAETTNSECVLSYTVHIPPTPDHQTVFASQESEEDEMLKGN
os_42915|LOC_Os07g36610.1|cslF     -RNNNNDAG---------------------------------------------------
AT4G23990.1|AT4G23990.1|cslG       ------------------------------------------------------------
AT1G55850.1|AT1G55850.1|cslE       ------------------------------------------------------------


AT5G22740.1|AT5G22740.1|cslA       ------------------------------------------------------------
AT2G24630.1|AT2G24630.1|cslC       ------------------------------------------------------------
os_25268|LOC_Os04g35020.1|cslH     ------------------------------------------------------------
AT2G32530.1|AT2G32530.1|cslB       ------------------------------------------------------------
AT2G21770.1|AT2G21770.1|cesA       GTRYKRIKGSPRVEGDEEDDDIDDLEHEFYGMDPEHVTEAALYY--------------M
AT1G02730.1|AT1G02730.1|cslD       SNQKSELSGTIETGGFKSVTRGHVIDCSMDRADPEKKSGQICWLKGCDEKVVHGRCFCGF
```

MAFFT > Clustal Omega > MUSCLE >> ClustalW

accuracy

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

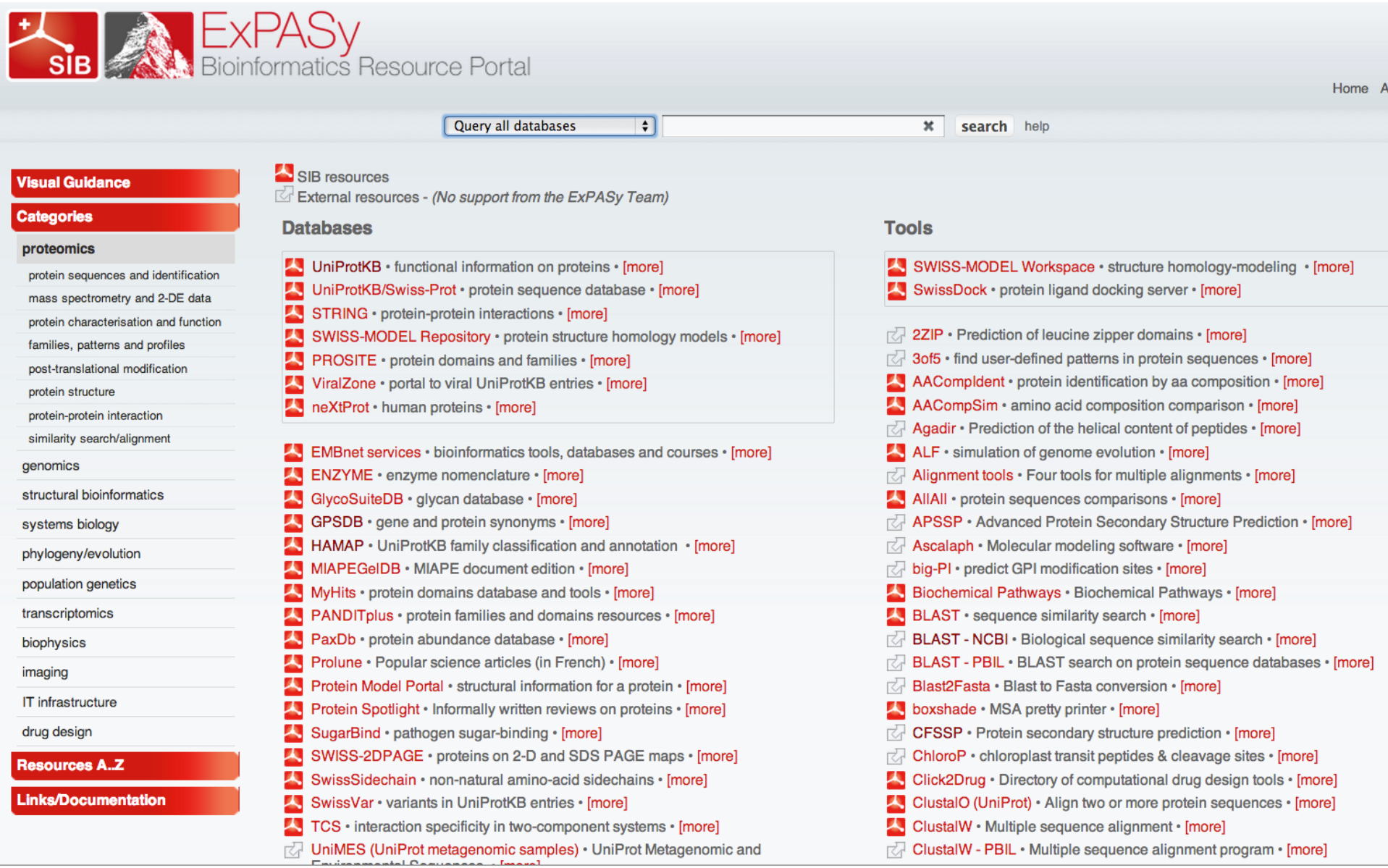
The figures are total column scores produced using bali score on core columns only. The average score over all families is given in the second column. The results for BAliBASE subgroupings are in columns 3–8. The total run time for all 218 families is given in the second last column. The last column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

Molecular Systems Biology 7:539, 2011

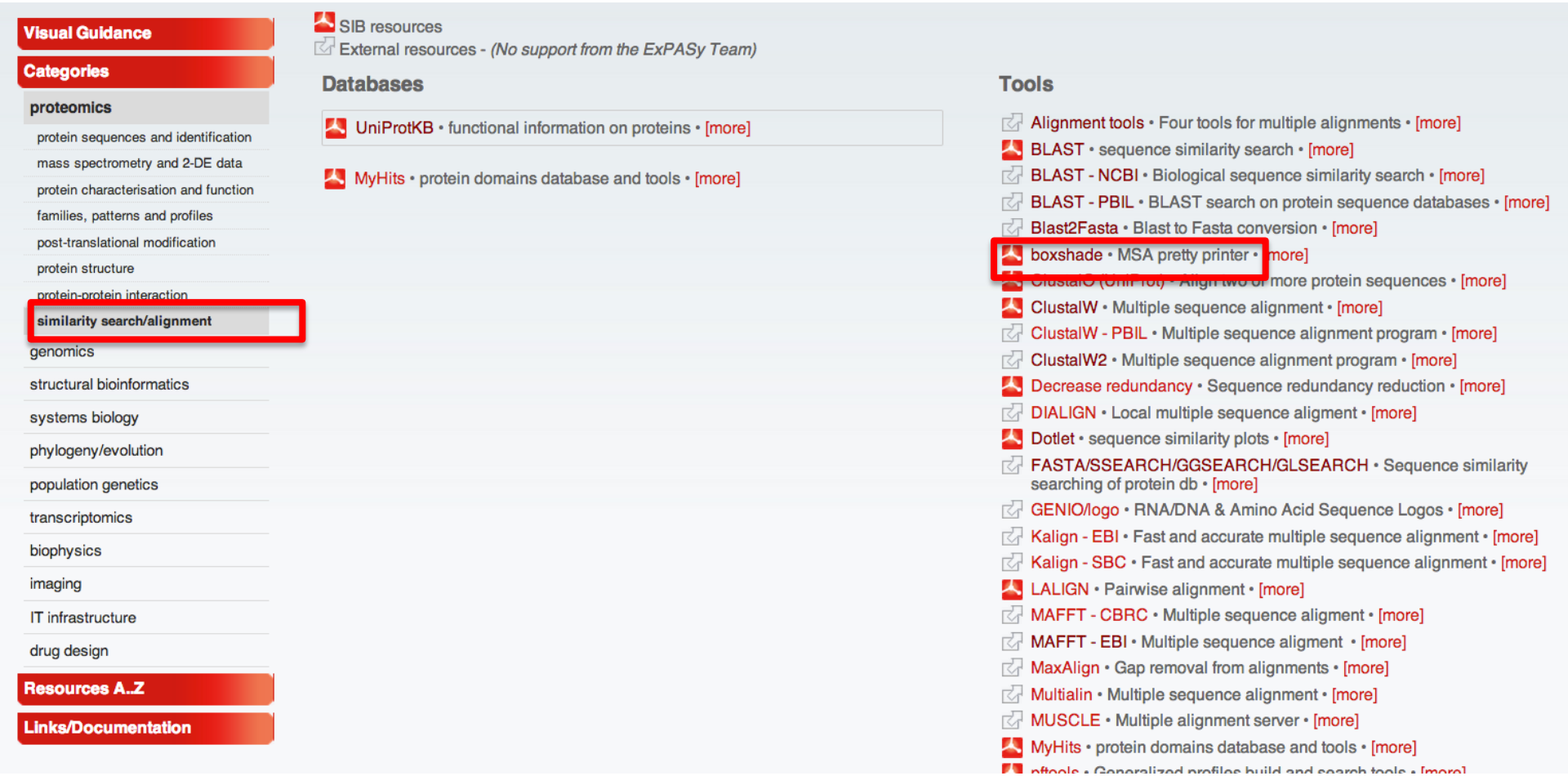
http://mafft.cbrc.jp/alignment/software/about.html

http://expasy.org/

Expert Protein Analysis System



boxshade


Visual Guidance
Categories
proteomics
protein sequences and identification
mass spectrometry and 2-DE data
protein characterisation and function
families, patterns and profiles
post-translational modification
protein structure
protein-protein interaction
similarity search/alignment
genomics
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)
Databases
UniProtKB • functional information on proteins • [more]
MyHits • protein domains database and tools • [more]
Tools
Alignment tools • Four tools for multiple alignments • [more]
BLAST • sequence similarity search • [more]
BLAST - NCBI • Biological sequence similarity search • [more]
BLAST - PBIL • BLAST search on protein sequence databases • [more]
Blast2Fasta • Blast to Fasta conversion • [more]
boxshade • MSA pretty printer • [more]
ClustalW • Multiple sequence alignment • [more]
ClustalW - PBIL • Multiple sequence alignment program • [more]
ClustalW2 • Multiple sequence alignment program • [more]
Decrease redundancy • Sequence redundancy reduction • [more]
DIALIGN • Local multiple sequence aligment • [more]
Dotlet • sequence similarity plots • [more]
FASTA/SSEARCH/GGSEARCH/GLSEARCH • Sequence similarity searching of protein db • [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]
MAFFT - CBRC • Multiple sequence aligment • [more]
MAFFT - EBI • Multiple sequence aligment • [more]
MaxAlign • Gap removal from alignments • [more]
Multialin • Multiple sequence alignment • [more]
MUSCLE • Multiple alignment server • [more]
MyHits • protein domains database and tools • [more]

Note: StartingJuly 17, version 3.21 is running on this server. The only changes are some improvements in the RTF output routine that enables RTF with shaded background for users of MS-Word 7.0
I have also compiled a list of List of frequently asked questions.
This server uses version 3.21 of BOXSHADE, written by K. Hofmann and M. Baron. BOXSHADE is in the public domain and available from Source Forge http://sourceforge.net/projects/boxshade/ The available version runs on PCs, VMS- and OSF1-machines and includes much more options than are implemented on this server.
This **server** takes a multiple-alignment file in either GCG's MSF-format or Clustals ALN-format.
Output can be created in the following formats:

- Postscript/EPS (using shaded background)
- RTF old (using colors)
- RTF new (using shaded background)
- XFIG-files (using shaded background)
- ASCII (showing similarities)
- ASCII (showing differences)
- HPGL (using colors)
- PICT (for later editing on MACs and PCs)

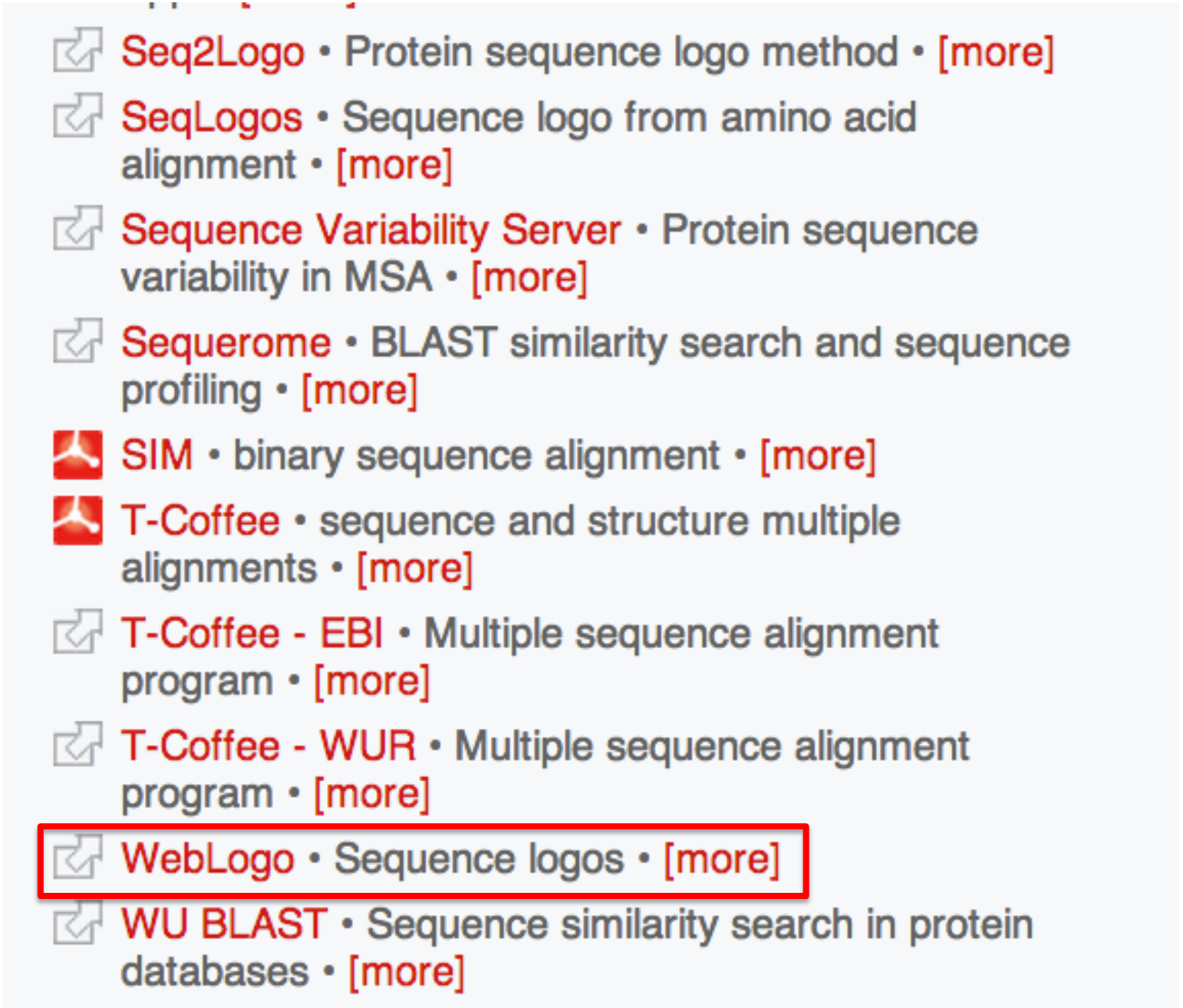
If you have problems using this server (like getting no result), read this and see the FAQ list.

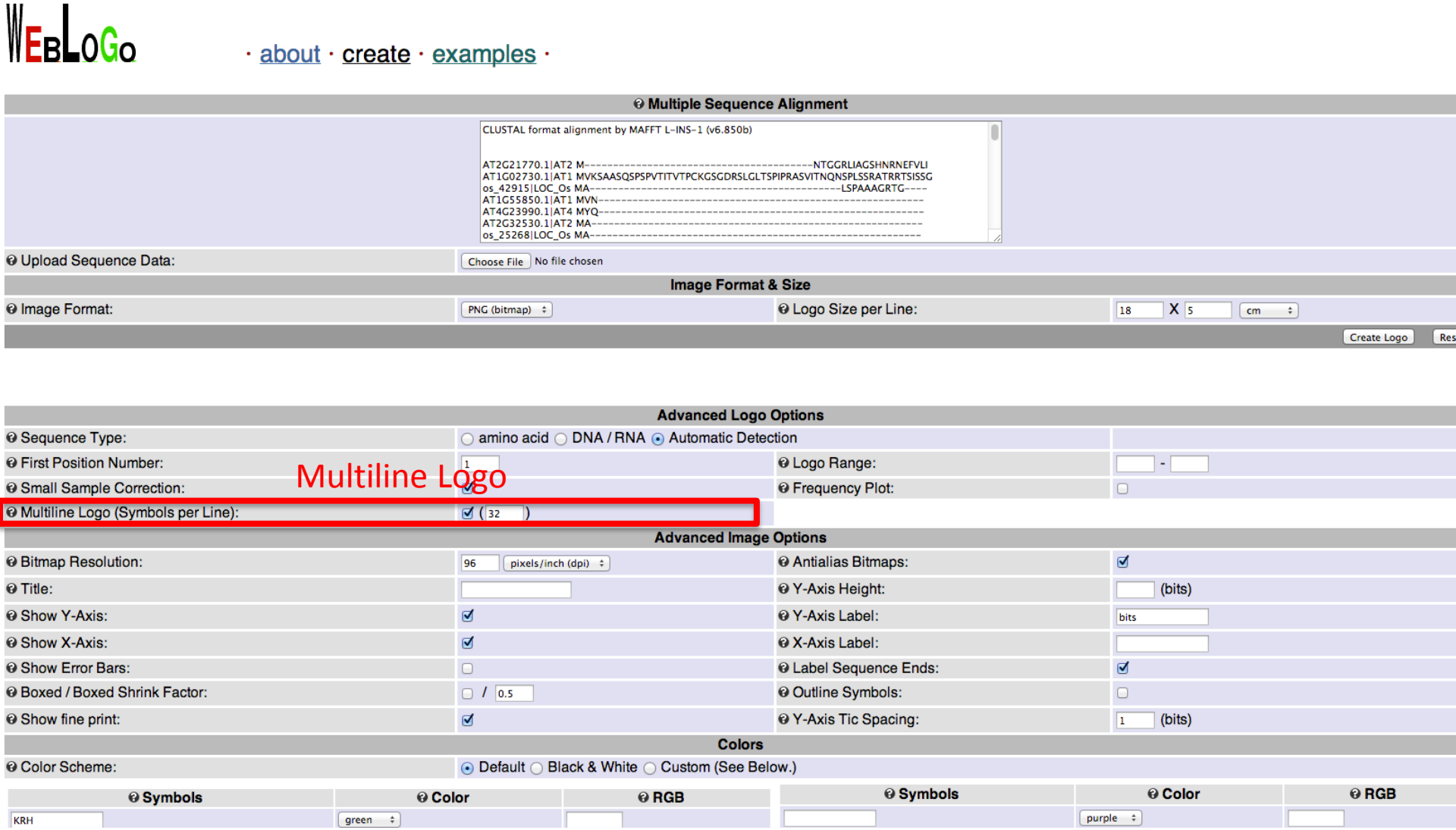
Output format	ASCII_similarities
Font Size	10
Consensus Line	consensus line with symbols
Fraction of sequences:	0.5 (that must agree for shading)
Enter sequence number:	9 only if 'consensus to a single sequence' is required
Query title (optional)	csl

When pasting MSF or ClustalW files, please make sure that the pasted text starts with the header line of the alignment and contains no extra blank lines at the bottom.

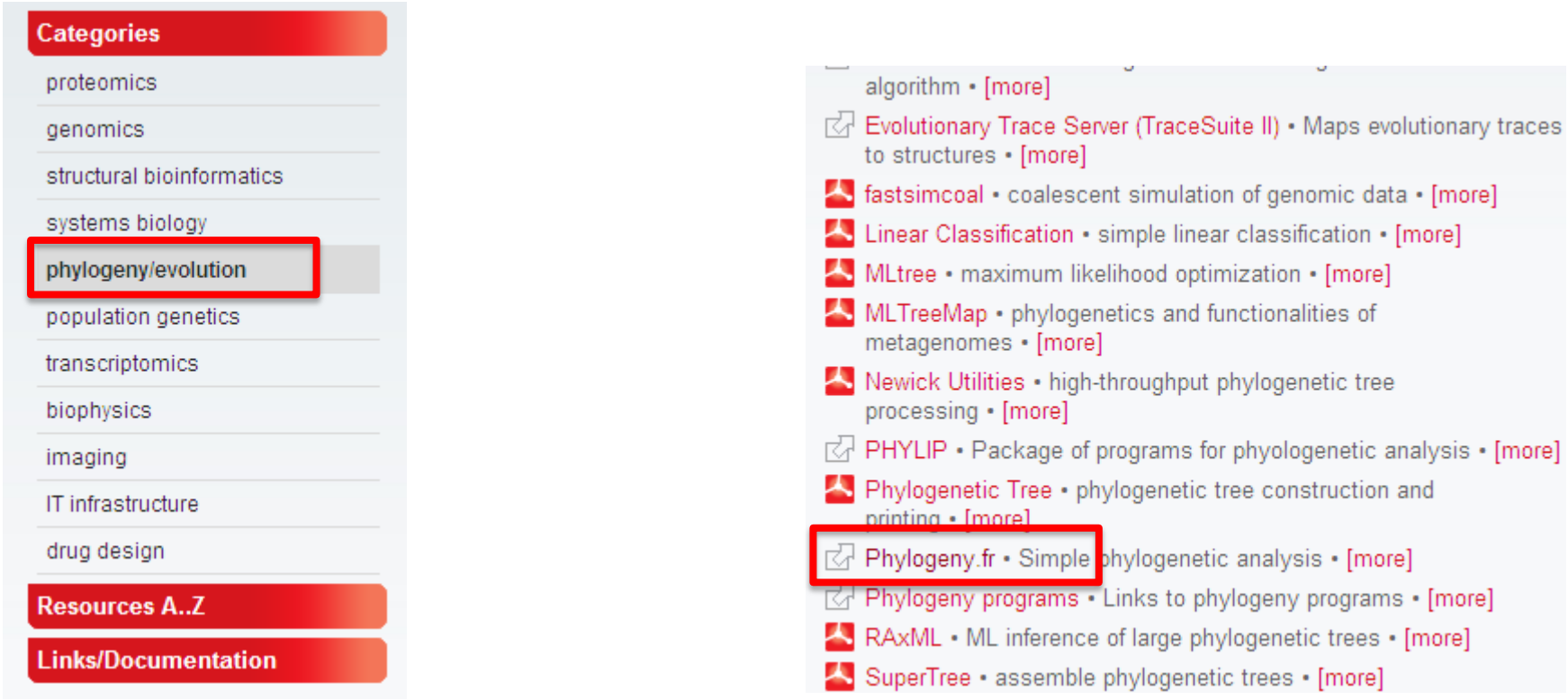
Input sequence format	ALN
Paste your multiple-alignment file (see above for valid formats)	CLUSTAL format alignment by MAFFT L-INS-1 (v6.850b) AT2G21770.1|AT2 M---------- NTGGRLIAGSHNRNEFVLI

Run BOXSHADE... Clear Input


Seq2Logo • Protein sequence logo method • [more]
SeqLogos • Sequence logo from amino acid alignment • [more]
Sequence Variability Server • Protein sequence variability in MSA • [more]
Sequerome • BLAST similarity search and sequence profiling • [more]
SIM • binary 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]
WU BLAST • Sequence similarity search in protein databases • [more]



Multiline Logo


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 phyologenetic 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]

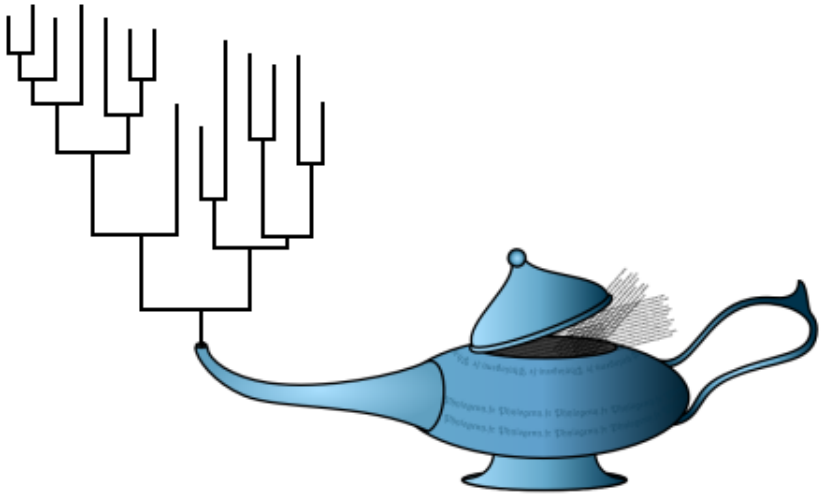
http://www.phylogeny.fr

LIRMM

Home | Phylogeny Analysis | Blast Explorer | Online Programs | Your Workspace | Documentation | Downloads | Contacts

- "One Click"
- "Advanced"
- "A la Carte"

Phylogeny.fr
Robust Phylogenetic Analysis For The Non-Specialist

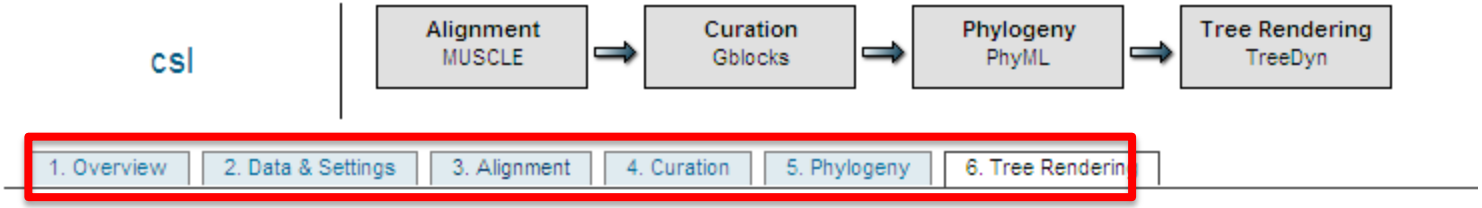


Phylogeny.fr is a free, simple to use web service dedicated to reconstructing and analysing phylogenetic relationships between molecular sequences.

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:


Home
Phylogeny Analysis
Blast Explorer
Online Programs
Your Workspace
Documentation
Downloads
Conta
"One Click" Mode
Alignment
MUSCLE
Curation
Gblocks
Phylogeny
PhyML
Tree Rendering
TreeDyn
1. Overview
2. Data & Settings
Name of the analysis (optional):
csl
Upload your set of sequences in FASTA, EMBL or NEXUS format from a file:
Choose File
No file chosen
Or paste it here (load example of sequences)
LLDWWRNEQFYMIGATGVYLAAVLHIVLKRLLGLKGVRFKLTAKQLAGGARERFAELYDVHWSPLLAPTV
VVMAVNVTAIGAAAGKAVVGGWTPAQVAGASAGLVFNVWVLVLLYPFALGIMGRWSKRPCALFALLVAAC
AAVAAGFVAVHAVLAAGSAAPSWLGWSRGATAILPSSWRLKRGF
>os_25268|LOC_Os04g35020.1|cslH
MAVVAAAAATGSTTRSGGGGGEGTRSGRKKPPPPPLQERVPLGRRAAWAWRLAGLAVLLLLALLALRLL
RHHGGAGGDGGVWRVALVCEAWFAALCALNVSAKWSPVRFVTRPENLVAEGRTPSTTAAEYGELPAVDML
VTTADPALEPPLVTVNTVLSLLALDYPRAGERLACYVSDDGCSPLTCHALREAAGFAAAWVPFCRRYGVA
VRAPFRYFSSSSSPESGGPADRKFLDDWTFMKDEYDKLVRRIKNTDERSLLRHGGGEFFAEFLNVERRNH
PTIVKTRVSAVMTNAPIMLNMDCDMFVNNPQAVLHAMCLLLGFDDEASSGFVQAPQRFYDALKDDPFGNQ
MECFFKRFISGVQGVQGAFYAGTGCFHRRKAVYGVPPNFNGAEREDTIGSSSYKELHTRFGNSEELNESA
RNIIWDLSSKPMVDISSRIEVAKAVSACNYDIGTCWGQEVGWVYGSLTEDILTGQRIHAMGWRSVLMVTE
PPAFMGSAPIGGPACLTQFKRWATGQSEIIISRNNPILATMFKRLKFRQCLAYLIVLGWPLRAPFELCYG
LLGPYCILTNQSFLPKASEDGFSVPLALFISYNTYNFMEYMACGLSARAWWNNHRMQRIISVSAWTLAFL
TVLLKSLGLSETVFEVTGKDKSMSDDDDNTDGADPGRFTFDSLPVFIPVTALAMLNIVAVTVGACRVAFG
TAEGVPCAPGIGEFMCCGWLVLCFFPFVRGIVWGKGSYGIPWSVKLKASLLVAMFVTFCKRN
Clear
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
To receive the results by e-mail, enter your address(es):



Tree Rendering results

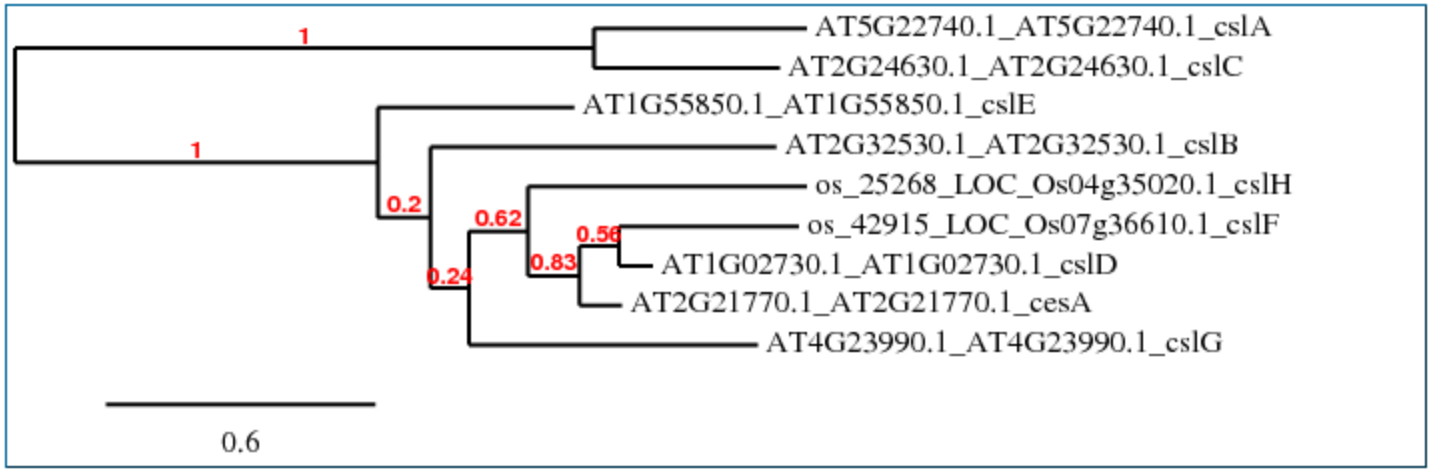


Figure 1: *Phylogenetic tree.*

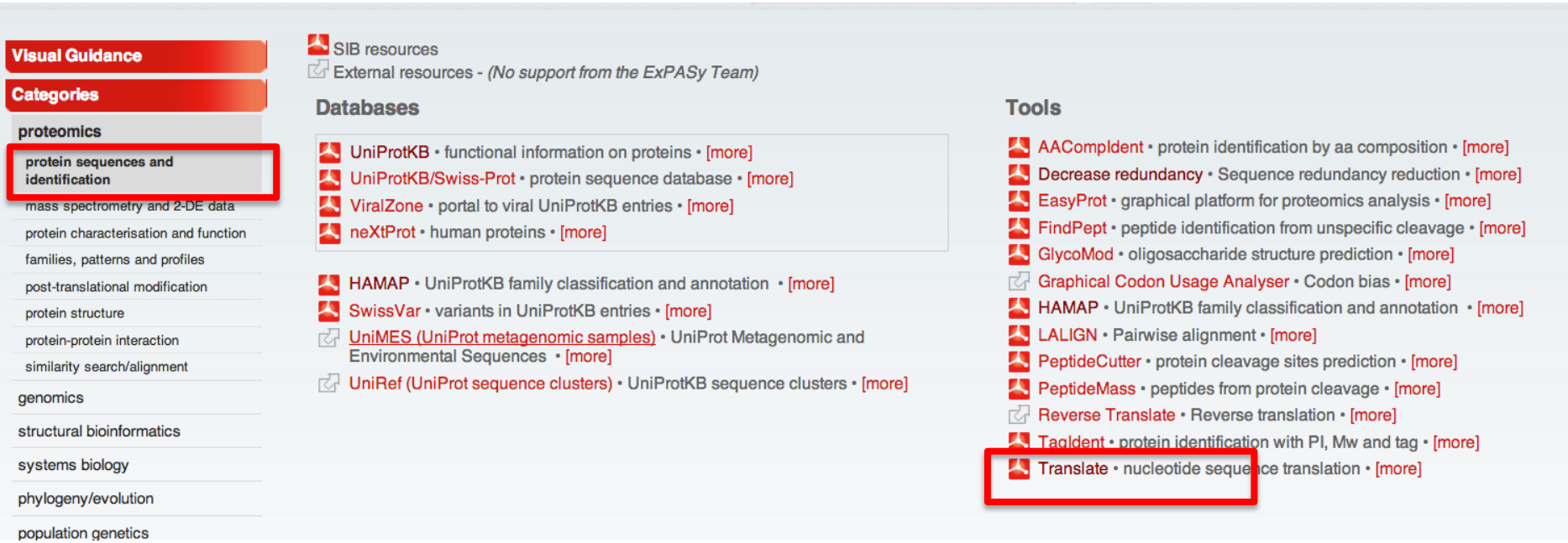
==> **Download the tree:** PNG - PDF - SVG - TGF (Treedyn format) - Newick - Text

Select an action:

Reset (cancel all changes)

Mid-point rooting

Use Genbank information to automatically rename leaves by: species and gi species only colorize


Visual Guidance
Categories
proteomics
protein sequences and identification
mass spectrometry and 2-DE data
protein characterisation and function
families, patterns and profiles
post-translational modification
protein structure
protein-protein interaction
similarity search/alignment
genomics
structural bioinformatics
systems biology
phylogeny/evolution
population genetics
SIB resources
External resources - (No support from the ExPASy Team)
Databases
UniProtKB • functional information on proteins • [more]
UniProtKB/Swiss-Prot • protein sequence database • [more]
ViralZone • portal to viral UniProtKB entries • [more]
neXtProt • human proteins • [more]
HAMAP • UniProtKB family classification and annotation • [more]
SwissVar • variants in UniProtKB entries • [more]
UniMES (UniProt metagenomic samples) • UniProt Metagenomic and Environmental Sequences • [more]
UniRef (UniProt sequence clusters) • UniProtKB sequence clusters • [more]
Tools
AACompIdent • protein identification by aa composition • [more]
Decrease redundancy • Sequence redundancy reduction • [more]
EasyProt • graphical platform for proteomics analysis • [more]
FindPept • peptide identification from unspecific cleavage • [more]
GlycoMod • oligosaccharide structure prediction • [more]
Graphical Codon Usage Analyser • Codon bias • [more]
HAMAP • UniProtKB family classification and annotation • [more]
LALIGN • Pairwise alignment • [more]
PeptideCutter • protein cleavage sites prediction • [more]
PeptideMass • peptides from protein cleavage • [more]
Reverse Translate • Reverse translation • [more]
TagIdent • protein identification with PI, Mw and tag • [more]
Translate • nucleotide sequence translation • [more]

Translate tool

Translate is a tool which allows the translation of a nucleotide (DNA/RNA) sequence to a protein sequence.

Please enter a DNA or RNA sequence in the box below (numbers and blanks are ignored).

```
>contig00008  gene=isogroup00001  length=862
tAAGATACCTCGAAATATTTTATTTGCATTTTATTTTGCATTGAGAATGGTAACTACATC
ACGCTGGCTTCCACTGAGCACATTCTCAGGGATGAAAAATGGACCGAATTCCAGCCTCTT
TTACAAAATGCCAGCCCCAAAATTTAGCAAGAGGACCAGTTGTCAGGATTCAAGGCCACC
TACACAGAAGCTATGTACTTGATAAGGTCCACCACACGGTTGCTGTAGCCCCACTCGTTA
TCATACCACGCGACAAACTTCATGAATTTTTCGTTTAACGCAATGCCGGCCTTTGCGTCG
AAGATACTCGACCTGCTATCACCAATCAGGTCTGAAGAGACGAGGTCCTCTTCTGTGTAT
CCAAGAATACCTTTCAGTGCTCCCTCTGATTCTTTCTTTATAACAGATTTTACTTCCTCG
TAAGAAGTTGCTTTCTCGAGCCTTACAGTCAGATCGACAACTGATACATCAGCAGTAGGC
ACACGGAATGCCATTCCTGTTAATTTCCCGTTTAATGCCGGAAGCACCTTCCCAACAGCC
TTTGCAGCTCCAGTTGAGCTTGGAATGATGTTGAATCCAGCACCTCGTCCACCCCTCCAG
TCTTTGGCAGAAGGTCCATCAACTGTTTTCTGCGTCGCTGTTACAGCGTGAACGGTAGTC
ATAAGGCCCTCAGTAATCCCAAAATTCTCATGTATAACCTTAGCAAGCGGGGCTAGACAA
TTAGTAGTGCAGCTTGCATTTGAGATTATGTCCAAGTCTGGAGTATAGTCCTTTTCATTG
ACACCTATAACGAACATAGGTGCATCCTTGCTGGGAGCGCTGATCACCACTTTCTTGGCT
```

Output format: Verbose ("Met", "Stop", spaces between residues)

Genetic code: Standard

Reset or TRANSLATE SEQUENCE

Better tool for open reading frame (ORF) prediction from EST data

The OrfPredictor (ORF-Predictor) server is designed for ORF prediction and translation of a batch of EST or cDNA sequences. See more details

Paste sequences below in FASTA format (Support multiple sequences in one FASTA file)

```
>contig00008 gene=isogroup00001 length=862
tAAGATACCTCGAAATATTTTATTTGCATTTTATTTTGCATTGAGAATGGTAACTACATC
ACGCTGGCTTCCACTGAGCACATTCTCAGGGATGAAAAATGGACCGAATTCCAGCCTCTT
TTACAAAATGCCAGCCCCAAAATTTAGCAAGAGGACCAGTTGTCAGGATTCAAGGCCACC
TACACAGAAGCTATGTACTTGATAAGGTCCACCACACGGTTGCTGTAGCCCCACTCGTTA
```

Or load from disk Choose File No file chosen

Clear

[Optional] **Paste BLASTX output** below in BLASTX output format

Or load from disk Choose File No file chosen

Clear

Select the strand for prediction both

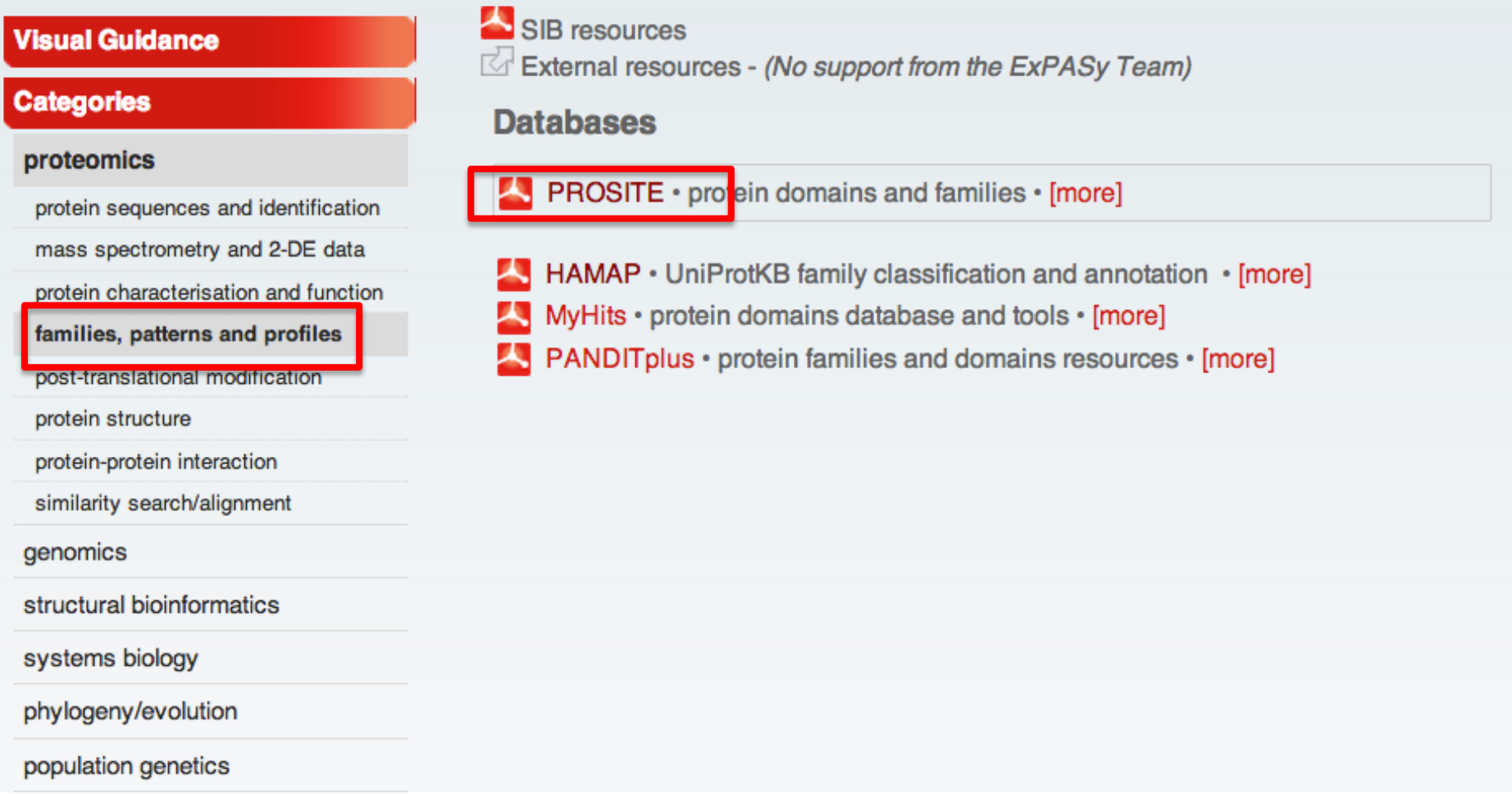
(Sequenced from 3' end, select '-'; from 5' end, select '+'; mixed or unknown, select 'both')

E-value in BLASTX 1e-5

E-mail results to: yanbin.yin@gmail.com

Submit Reset

http://proteomics.ysu.edu/tools/user_results/usr_WedFeb61910112013/


Visual Guidance
Categories
proteomics
protein sequences and identification
mass spectrometry and 2-DE data
protein characterisation and function
families, patterns and profiles
post-translational modification
protein structure
protein-protein interaction
similarity search/alignment
genomics
structural bioinformatics
systems biology
phylogeny/evolution
population genetics
SIB resources
External resources - (No support from the ExPASy Team)
Databases
PROSITE • protein domains and families • [more]
HAMAP • UniProtKB family classification and annotation • [more]
MyHits • protein domains database and tools • [more]
PANDITplus • protein families and domains resources • [more]

Home | ScanProsite | ProRule | Documents | Download

Database of protein domains, families and functional sites

PROSITE consists of documentation entries describing protein domains, families and functional sites as well as associated patterns and profiles to identify them [More... / References / PROSITE is complemented by **ProRule**, a collection of rules based on profiles and patterns, which increases the discriminatory power of profiles and patterns by providing additional in functionally and/or structurally critical amino acids [More...].

Release 20.89, of 31-Jan-2013 (1659 documentation entries, 1308 patterns, 1050 profiles and 1053 ProRule)

PROSITE access

e.g. PDOC00022, PS50089, SH3, zinc finger

Search add wildcard '*'

Browse:

- by documentation entry
- by ProRule description
- by taxonomic scope
- by number of positive hits

PROSITE tools

Scan a sequence against PROSITE patterns and profiles - quick scan

(Output includes graphical view and feature detection)

Enter your sequence(s) or (a) UniProtKB (Swiss-Prot or TrEMBL) AC or ID [Help]:

```
>AT5G22740.1|AT5G22740.1|cslA
MDGVSPKFVLPETFDGVRMEITGQLGMIWELVKAPVI
VPLLQLAVYICLLMSVMLLCERVYMGIVIVLVKLFWKK
PDKRYKFEPIHDDEELGSSNFPVVLVQIPMFNEREVYK
LSIGAACGLSWPSDRLVIQVLDDSTDPTVKQMVEVE
CQRWASKGINIRYQIRENRVGYKAGALKEGLKRSYVK
```

Scan Clear

exclude patterns with a high probability of occurrence

- **ScanProsite** - advanced scan
- **PRATT** - allows to interactively generate conserved pa
  of unaligned proteins.
- **MyDomains - Image Creato** - allows to generate cust

Google: cbs dtu

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
« 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

Instructions | Output format

SUBMISSION

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

```
>contig00008  gene=isogroup00001  length=862
tAAGATACCTCGAAATATTTTATTTGCATTTTATTTTGCATTGAGAATGGTAAC
TACATC
```

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

Choose File No file chosen

Vertebrate **A. thaliana** Submit Clear fields

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.*

CITATIONS

For publication of results, please cite:

Neural network prediction of translation initiation sites in eukaryotes: perspectives for EST and genome analysis.
A. G. Pedersen and H. Nielsen, ISMB: 5, 226-233, 1997.

PORTABLE VERSION

Would you prefer to run NetStart at your own site? NetStart 1.0 is available as a stand-alone software package, with the same functionality as the service abo
download page for academic users; other users are requested to contact CBS Software Package Manager at software@cbs.dtu.dk.

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
« 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

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 prok
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 tha
- 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 Performan

SUBMISSION

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

```
GIDDWWRNEQFWVIGGVSSHLFALFQGLLKVLAGVSTNFTVTSKAADDGEFSELYIFK
WTSLLIPPTTLLIINIVGVIVGVSDAINNGYDSWGPLFGRLFFALWVIVHLYPFLKGLLGK
QDRVPTIILVWSILLASILTLLWVRVNPFVSKDGPVLEICGLDCLK
```

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:
  - 0.45 *D-cutoff for SignalP-noTM networks*
  - 0.50 *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

Pan-specific binding of peptides to MHC class I alleles of known sequence

NNAlign
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)

Choose File No file chosen

OR by pasting sequence(s) in FASTA format:

```
INLSDRLHQVLRWALGSVEIFLSRHCPIWYGYGGGLKWLERFSYINSVVYPWTSLPLLVYCSLPAI
CLLTGKFIVPEISNYAGILFLLMFMSIAVTGILEMQWGKIGIDDWWRNEQFWVIGGVSSHLFALF
QGLLKVLAGVSTNFTVTSKAADDGEFSELYIFKWTSLLIPPTTLLIINIVGVIVGVSDAINNGYDS
WGPLFGRLFFALWVIVHLYPFLKGLLGKQDRVPTIILVWSILLASILTLLWVRVNPFVSKDGPVLEI
CGLDCLK
```

Output format:

- (•) Extensive, with graphics
- ( ) Extensive, no graphics
- ( ) One line per protein

Other options:

- [ ] Use old model (version 1)

Submit Clear

PORTABLE VERSION

Next class: JGI resource