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

Yanbin Yin

Homework assignment 4

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).

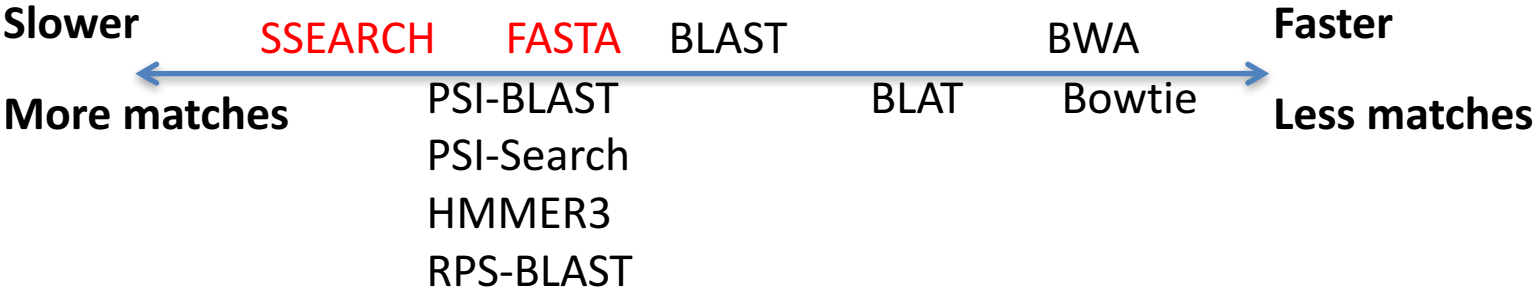
Write a report (in word or ppt) to include all the operations, screen shots and the final phylogenies from step 3 and 4.

Due on 10/10 (send by email, if there are 2+ files, put them in a zip file; include your last name in the file name)

Outline

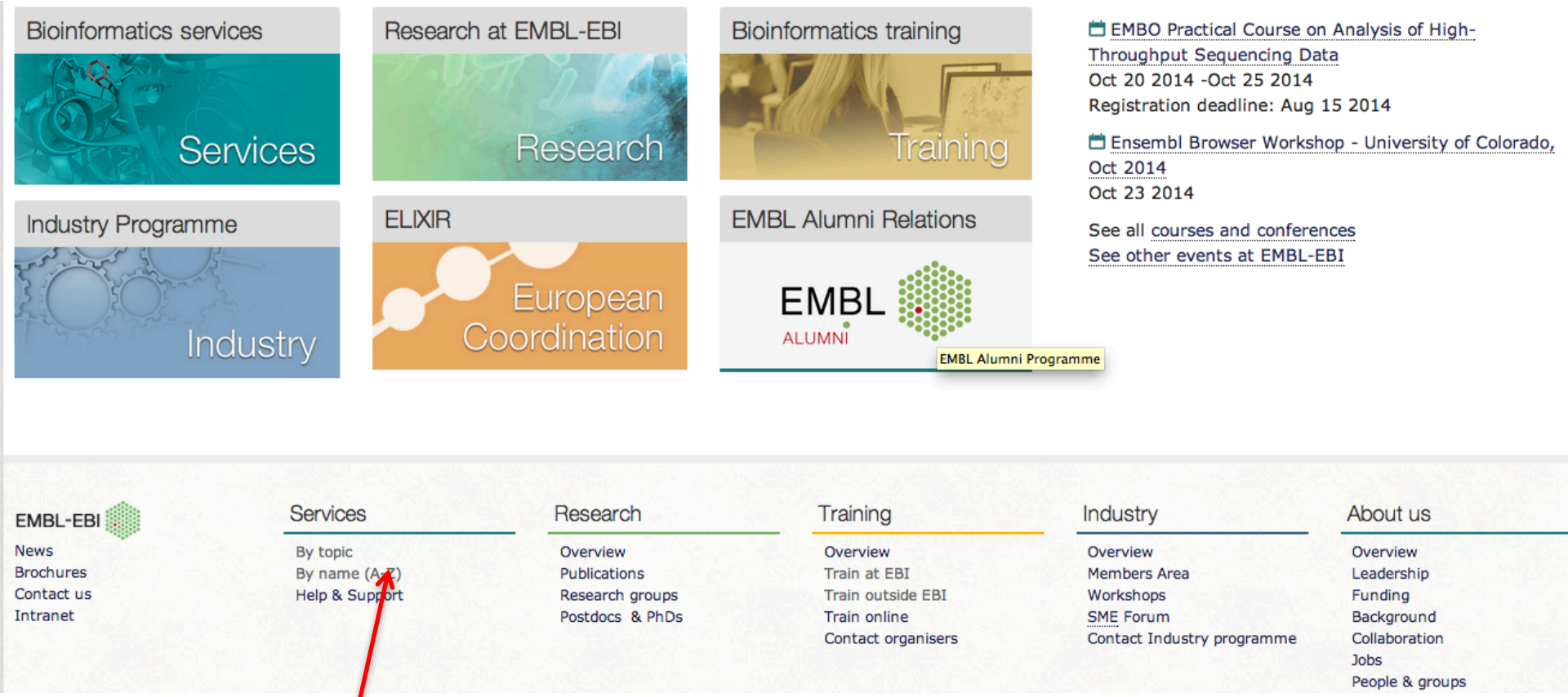
- Hands on exercises!

Pairwise alignment (including database search) tools



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

To the bottom of the page

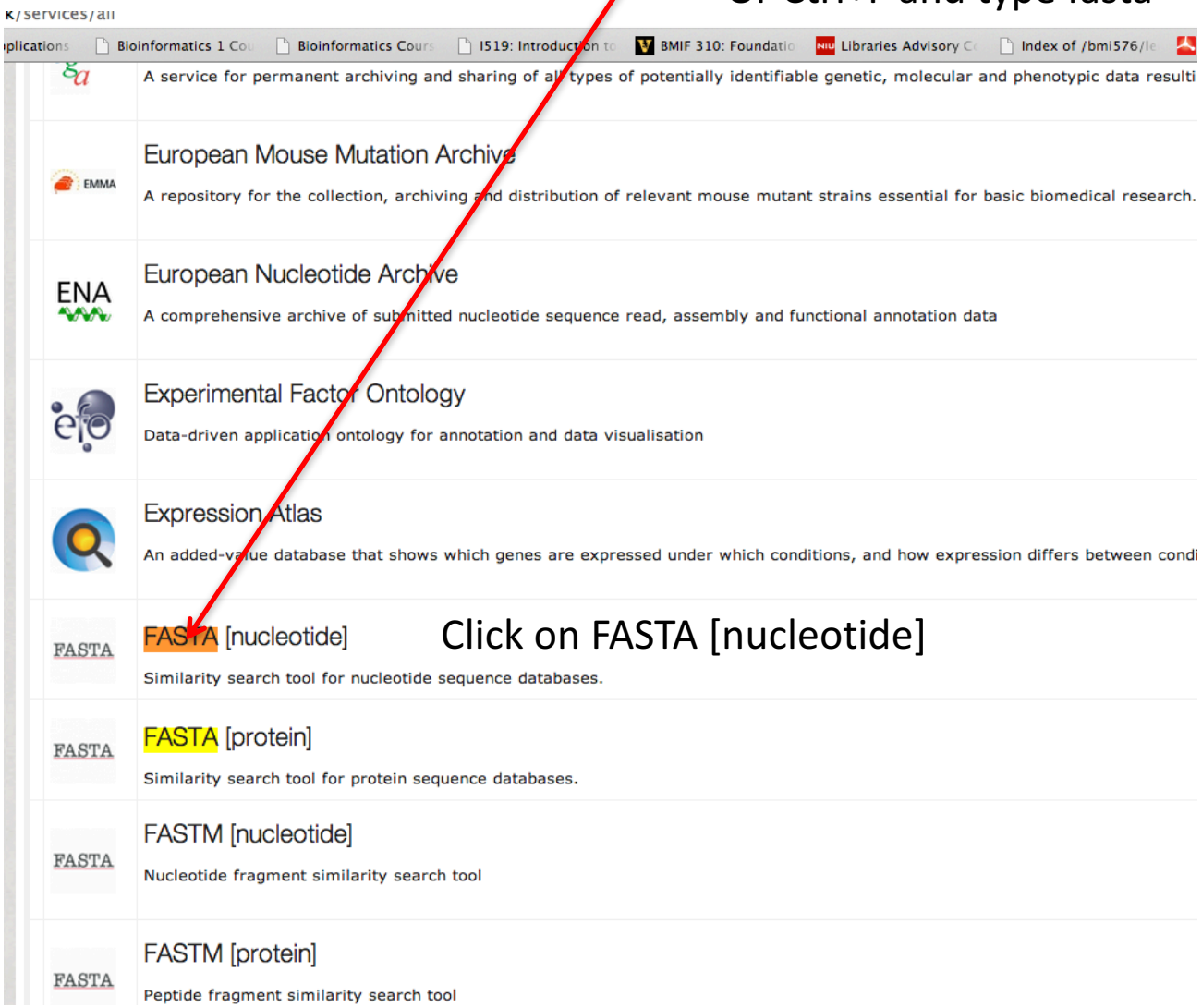


Click by names (A-Z)

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



Click on FASTA [nucleotide]

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 nuc
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 DNA sequence in any supported format:

or upload a file: Choose File No file chosen

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

Tools > Sequence Similarity Searching > FASTA

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

Click on this little arrow

Choose Arabidopsis

GENOME DATABASES

0 Databank Selected ✗ Clear Selection

- Eukaryota
- Archaea
- Bacteria
- Phage

Change here to protein

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

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.

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

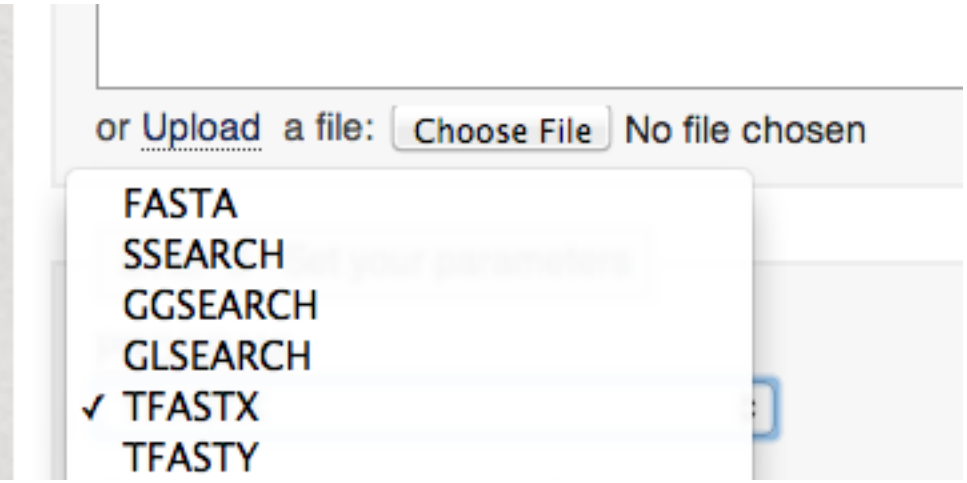
STEP 4 - Submit your job

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

Submit

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)

Protein | Nucleotide | Genomes | Proteomes | Whole Genome Shotgun | Web services | Help & Documentation | Share | Feedback

Tools > Sequence Similarity Searching > FASTA

Results for job fasta-I20140922-203900-0063-55691713-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 Clustal Omega

Align.	DB:ID	Source	Length	Score	Identities %	Positives %	E()
1	EM_PLN:AB016893	STD:Arabidopsis thaliana genomic DNA, chromosome 5, P1 clone:MYH9. *Cross-references and related information in:* Nucleotide sequences, Genomes, Protein families, Literature, Samples & ontologies, Protein sequences	82390	3862	59.4	67.9	2.4E-78
2	EM_PLN:AB025637	STD:Arabidopsis thaliana genomic DNA, chromosome 5, P1 clone:MVP7. *Cross-references and related information in:* Nucleotide sequences, Genomes, Protein families, Samples & ontologies, Protein sequences	39645	4258	64.2	70.7	1.4E-75
3	EM_PLN:AL161595	STD:Arabidopsis thaliana DNA chromosome 4, contig fragment No. 91 *Cross-references and related information in:* Nucleotide sequences, Genomes, Protein families, Macromolecular structures, Samples & ontologies, Protein sequences	198151	4660	70.0	74.7	7.3E-75
4	EM_PLN:AB006703	STD:Arabidopsis thaliana genomic DNA, chromosome 5, P1 clone:MRH10. *Cross-references and related information in:* Nucleotide sequences, Genomes, Protein families, Literature, Macromolecular structures, Samples & ontologies, Protein sequences	71522	2493	47.3	65.0	6.3E-64
5	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)

Tools > Sequence Similarity Searching > FASTA

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   ECAFPTCRPCYEYERREGNQACPQCGTRYKRIKG--------------------------
        :::::.::::::::::::::.:::: ::::::::
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-----EDSDMYSDRHALIVPPSTGLGNRVHHVPFTDS----
            :: . . ::.  ::.      : .. :: ::::: :: :  .:::.  : :
EM_PLN  WFIFSPITSTVLLSLFYCSFWLFF*DVEISSDSHALIVSPSPGHIHRVHQPHFPDPAGKF
    27040     27070     27100     27130     27160     27190

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

        230       240       250                           260
AT2G2   IEKLQVVKNERVND-GDGDGFIVDELD-DPGLPM-------------------MDEGRQP
         :: :::::..  .. ::::   .   : ::.  .                  :::::::
EM_PLN  NEKYQVVKHDGDSSLGDGDDADIPM*D\DPNTVFCHVSLLCFDIYEFCCSCFRMDEGRQP
    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)

Alignments

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      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

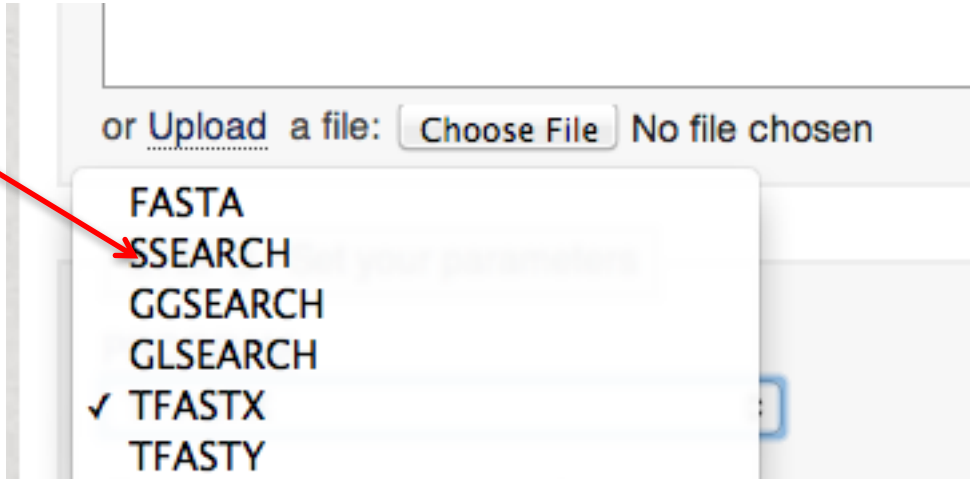
Score	Expect	Method	Identities	Positives	Gaps	Frame
137 bits(345)	3e-149	Compositional matrix adjust.	75/171(44%)	106/171(61%)	34/171(19%)	+1

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
```

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



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

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

Ctrl+F and type emboss

EMBOSS backtranambig
Back-translate a protein sequence to ambiguous nucleotide sequence
DNA back-translation

EMBOSS backtranseq
Back-translate a protein sequence to a nucleotide sequence, using codon frequency
DNA back-translation

EMBOSS cpgplot
Identification of potential CpG islands
CpG island and isochore detection

EMBOSS isochore
Plot potential isochore features
CpG island and isochore detection

EMBOSS matcher
Waterman-Eggert local alignment of two sequences
Pairwise sequence alignment | Local alignment

EMBOSS needle
Needleman-Wunsch global alignment of two sequences
Pairwise sequence alignment | Global alignment

Needleman-wunsch algorithm
Smith-Waterman algorithm

Equivalent to the bl2seq command of the BLAST package

EMBOSS contain hundreds of computer programs for sequence analysis

EMBOSS pepstats
Calculate statistics of protein properties
Protein property calculation

EMBOSS pepwindow
Draw a hydropathy plot for a protein sequence
Protein hydropathy calculation

EMBOSS seqret
Sequence format conversion tool
Sequence formatting

EMBOSS sixpack
Six frame nucleotide sequence translation, with ORF finding
DNA translation | Coding region prediction

EMBOSS stretcher
Improved version of the Needleman-Wunsch algorithm that allows rapid global alignment of two larger sequences
Pairwise sequence alignment | Global alignment

EMBOSS transeq
Nucleotide sequence translation in selected frames
DNA translation

EMBOSS water
Smith-Waterman local pairwise alignment of sequences

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

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: Choose File No file chosen

AND

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

Or, upload a file: Choose 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.

More options... *(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)*

Submit

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

This is needle output

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

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

cesA             351 KEGKPSELAPVDVFVSTVDPLKEPPLITANTVLSILAVDYPVEKVACYVS    400

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

cesA             401 DDGAAML-----TFEA--LSYTAEFARKW--------VP-------FCKK    428
                     .||.:..     ||:.  :..|.:....|        ||       .|..
cslA               1 MDGVSPKFVLPETFDGVRMEITGQLGMIWELVKAPVIVPLLQLAVYICLL     50
```

- gap
. Negative score
: positive score
| identical

This is different from what BLAST shows the alignment

This is water output

```
#=======================================
#
# 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
#
#
#=======================================

cesA             779 GWIYGSVTEDILTGFKMHCHGWRSVYCMPKRAAFKGSAPINLSDRLHQVL    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 MQWGKIGIDDWWRNEQFWVI--GGVSSH-----LFALFQGLLKVLAGVST    960
                     : ...:|.........:|::  ..:|.|     |..||:      ||.:.
cslA             407 I-LNSVGTPRSIHLLFYWILFENVMSLHRTKATLIGLFE------AGRAN    449

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


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

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

NIH U.S. National Library of Medicine | NCBI National Center for Biotechnology Information

BLAST® » blastp suite

Standard Protein BLAST

blastn | blastp | blastx | tblastn | tblastx

Enter Query Sequence

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

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

Query subrange

From

To

Or, upload file Choose File No file chosen

Job Title

Enter a descriptive title for your BLAST search

☑ **Align two or more sequences**

Select here to try blast 2 seq

Choose Search Set

Database Non-redundant protein sequences (nr)

Organism Optional

Enter organism name or id--completions will be suggested ☐ Exclude +

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

Exclude Optional

☐ Models (XM/XP) ☐ Uncultured/environmental sample sequences

Entrez Query Optional

YouTube Create custom database

Enter an Entrez query to limit search

Program Selection

Algorithm

- (●) blastp (protein-protein BLAST)
- ( ) PSI-BLAST (Position-Specific Iterated BLAST)
- ( ) PHI-BLAST (Pattern Hit Initiated BLAST)
- ( ) DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm

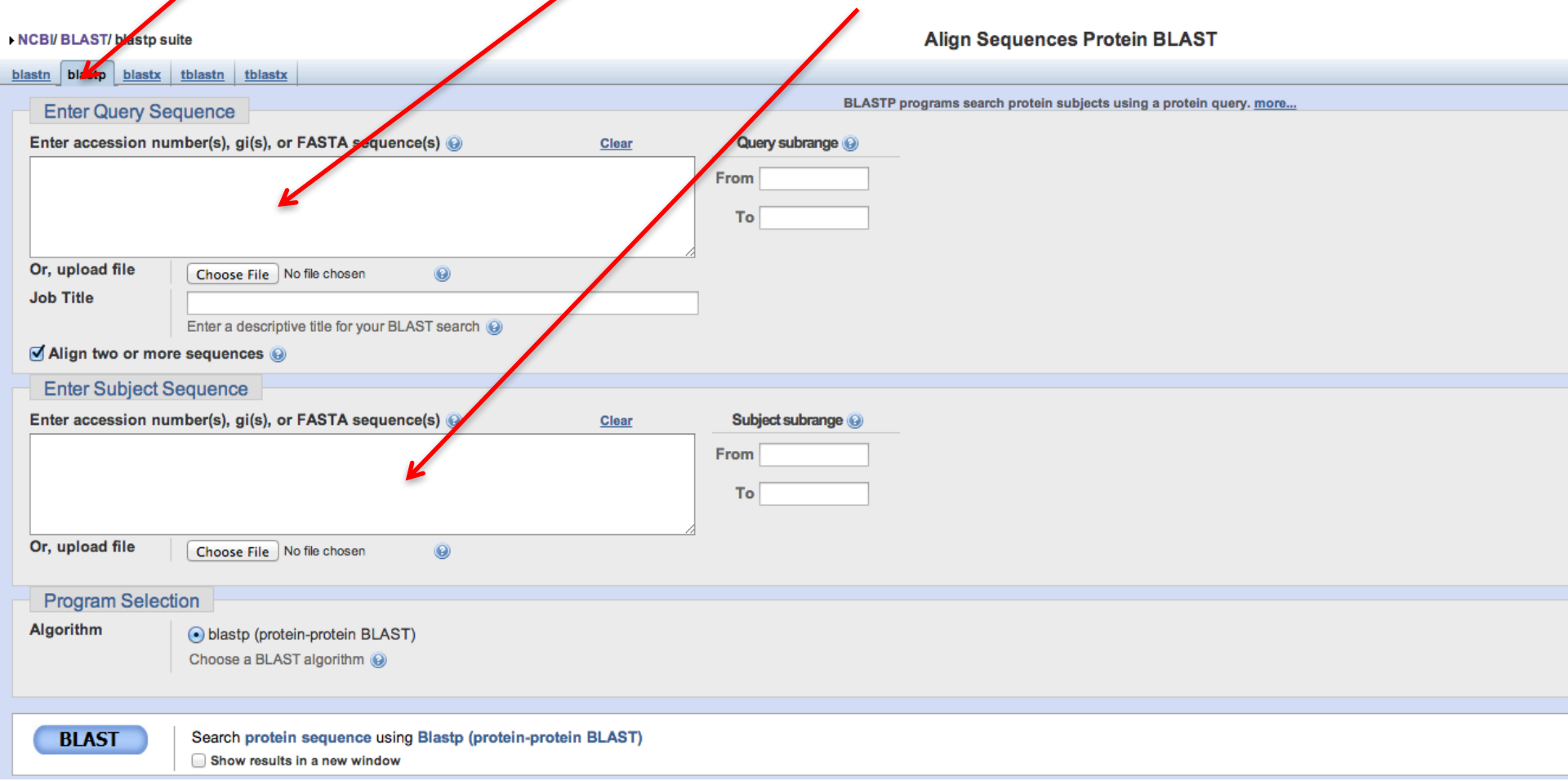
BLAST Search **protein sequence** using **Blastp (protein-protein BLAST)**

☐ **Show results in a new window**

+ Algorithm parameters

Choose blastp

Go to http://cys.bios.niu.edu/yyin/teach/PBB/cesa-pr.fa
Copy & paste CesA
Copy & paste CslA



This blast2seq output

Download Graphics Sort by: E value

Fragmented alignments

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  GWIYGSVTEDILTGFKMHCHGWRSVYCMPKRAAFKGSAPINLSDRLHQVLRWALGSVEIF  838
            GW   +  ED+    +    GW+ +Y    +   K   P        Q  RW+ G   +F
Sbjct  280  GWKDRTTVEDMDLAVRASLRGWKFLYLGDLQV--KSELPSTFRAFRFQQHRWSCGPANLF  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  YSFFFVRKIIAHWVTFC  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  EIGWIYGSVTEDILTGFKMHCH  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  TAKLGSGQSAKGNTKGIKR  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 t 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 sequenc

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 alig methods. Suitable for small alignments.

Launch T-Coffee

WebPRANK

The EBI has a new phylogeny-aware multiple sequence alignment program which mak 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 sequer sequence weighting, position-specific gap penalties and weight matri:
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 sequ **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 hi
RC Edgar - Nucleic acids research, 2004 - Oxford Univ Press
Abstract We describe **MUSCLE**, a new computer program for creating multiple alignment protein sequences. Elements of the algorithm include fast distance estimation using kme 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 f 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) Homo logous 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 alignme
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 progran
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
alignment 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

Tools > Multiple Sequence Alignment > Clustal Omega

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

Alignments | Result Summary | Phylogenetic Tree | Submission Details

Download Alignment File | Show Colors | Send to ClustalW2_Phylogeny

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|cslA        ------------------------------------------------------------
AT2G24630.1|AT2G24630.1|cslC        ------------------------------------------------------------
AT2G21770.1|AT2G21770.1|cesA        ------------------------------------------------------------
AT1G02730.1|AT1G02730.1|cslD        MVKSAASQSPSPVTITVTPCKGSGDRSLGLTSPIPRASVITNQNSPLSSRATRRTSISSG
os_42915|LOC_Os07g36610.1|cslF      ------------------------------------------------------------
AT1G55850.1|AT1G55850.1|cslE        ------------------------------------------------------------
AT4G23990.1|AT4G23990.1|cslG        ------------------------------------------------------------
AT2G32530.1|AT2G32530.1|cslB        ------------------------------------------------------------
os_25268|LOC_Os04g35020.1|cslH      ------------------------------------------------------------


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


AT5G22740.1|AT5G22740.1|cslA        ------------------------------------------------------------
AT2G24630.1|AT2G24630.1|cslC        ------------------------------------------------------------
AT2G21770.1|AT2G21770.1|cesA        HNRNEFVL----------INA----DDTARIRSAEELSGQTCKICRDEIELTDNGEPFIA
AT1G02730.1|AT1G02730.1|cslD        SNQKSFLSGTIFTGGFKSVTRGHVIDCSMDRADPEKKSGQICWLKGCD------EKVVHG
os_42915|LOC_Os07g36610.1|cslF      ------------------------------------------------------------
AT1G55850.1|AT1G55850.1|cslE        ------------------------------------------------------------
AT4G23990.1|AT4G23990.1|cslG        ------------------------------------------------------------
AT2G32530.1|AT2G32530.1|cslB        ------------------------------------------------------------
os_25268|LOC_Os04g35020.1|cslH      ------------------------------------------------------------
```

Tools > Multiple Sequence Alignment > Clustal Omega

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

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

```
(
(
(
AT5G22740.1|AT5G22740.1|cslA:0.28854,
AT2G24630.1|AT2G24630.1|cslC:0.29272)
:0.19734,
(
AT2G32530.1|AT2G32530.1|cslB:0.30390,
os_25268|LOC_Os04g35020.1|cslH:0.31325)
:0.03405)
:0.00339,
(
AT2G21770.1|AT2G21770.1|cesA:0.27403,
(
AT1G02730.1|AT1G02730.1|cslD:0.24931,
os_42915|LOC_Os07g36610.1|cslF:0.28135)
:0.03248)
:0.04783,
(
AT1G55850.1|AT1G55850.1|cslE:0.30074,
AT4G23990.1|AT4G23990.1|cslG:0.32201)
:0.03279);
```

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

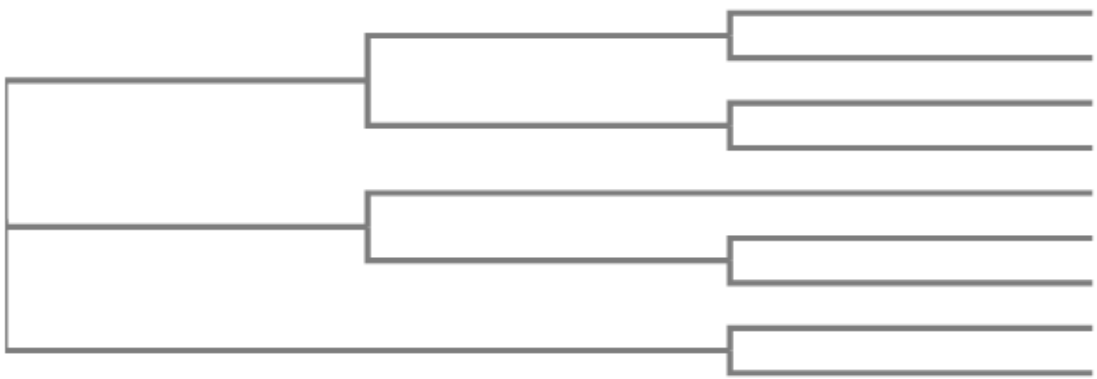
```
#
#
#  Percent Identity  Matrix - created by Clustal2.1
#
#
```

Matrix tells how similar each pair of seqs is

1: AT5G22740.1|AT5G22740.1|cslA	100.00	41.87	18.40	17.26	16.84	18.86	15.24	16.45	17.00
2: AT2G24630.1|AT2G24630.1|cslC	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|cslD	17.26	17.92	47.11	100.00	46.93	33.76	30.81	30.80	32.18
5: os_42915|LOC_Os07g36610.1|cslF	16.84	15.33	38.52	46.93	100.00	31.82	25.98	29.66	30.97
6: AT1G55850.1|AT1G55850.1|cslE	18.86	15.82	35.00	33.76	31.82	100.00	37.73	33.73	29.46
7: AT4G23990.1|AT4G23990.1|cslG	15.24	15.18	32.64	30.81	25.98	37.73	100.00	31.29	30.15
8: AT2G32530.1|AT2G32530.1|cslB	16.45	16.61	32.63	30.80	29.66	33.73	31.29	100.00	38.29
9: os_25268|LOC_Os04g35020.1|cslH	17.00	17.70	32.96	32.18	30.97	29.46	30.15	38.29	100.00

Phylogram

Branch length: Cladogram Real



You can always check the help page

This is MAFFT page

Input form | Web services | Help & Documentation

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

Tools > Multiple Sequence Alignment > MAFFT

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

IDs were truncated

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

AT2G21770.1|AT2 M-----------------------------------------NTGGRLIAGSHNRNEFVLI
AT1G02730.1|AT1 MVKSAASQSPSPVTITVTPCKGSGDRSLGLTSPIPRASVITNQNSPLSSRATRRTSISSG
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 MAPRFDFSDLWAKETRRGTPVVVKM---------------------------------
                *

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

Tools > Multiple Sequence Alignment > MUSCLE

This is MUSCLE page

Multiple Sequence Alignment

MUSCLE stands for **MU**ltiple **S**equence **C**omparison by **L**og- **E**xpectation. MUSCLE is claimed to achieve both better aver
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: Choose File No file chosen

STEP 2 - Set your Parameters

OUTPUT FORMAT: ClustalW

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

Input form | Web services | Help & Documentation

Tools > Multiple Sequence Alignment > MUSCLE

Results for job muscle-I20140924-181930-0252-89208085-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|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        ------------------------------------------------------------
```

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 t 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 sequenc

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 alig methods. Suitable for small alignments.

Launch T-Coffee

WebPRANK

The EBI has a new phylogeny-aware multiple sequence alignment program which mak 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 Protein sequence alignment in any supported format:

CLUSTAL O(1.2.1) multiple sequence alignment

AT5G22740.1|AT5G22740.1|cslA ----------
AT2G24630.1|AT2G24630.1|cslC ----------
AT2G21770.1|AT2G21770.1|cesA ----------

Upload a file: Choose File No file chosen

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

STEP 2 - Set input parameters

INPUT FORMAT
AUTOMATIC

STEP 3 - Set output parameters

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

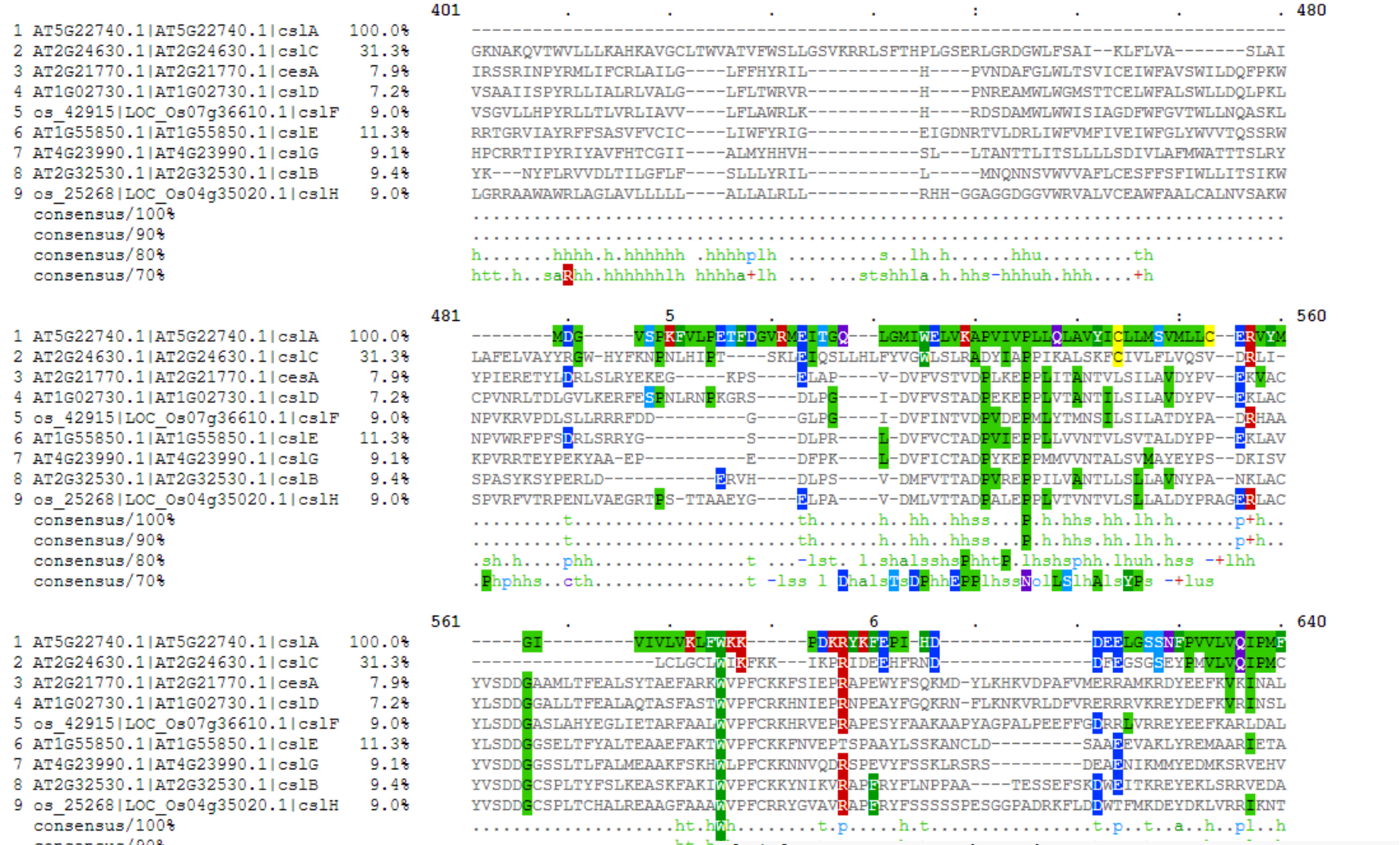
More options... *(Click here, if you want to view or change the default settings.)*

STEP 4 - Submit your job

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

Submit

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



Consensus letters explained at
http://bio-mview.sourceforge.net/manual/manual.html#ref-output-formats

Another MSA visualization tool: ESPript http://espript.ibcp.fr/ESPript/ESPript/

Click here to start

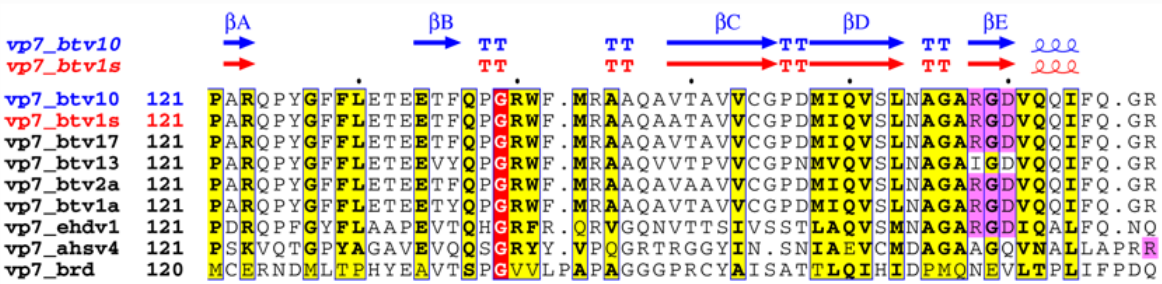
ESPript 3.0

• RUN ESPRIPT • | HELP | GALLERY | ABOUT | TRY ENDSCRIPT

NEWS
- September 9, 2014: if you want to relax, the ESPript 'Rebcross' game is online!
- July 1, 2014: ESPript3 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 ESPript
- December 23, 2013: we are pleased to announce the release of a major update: ESPript version 3.0!

What is ESPript?

- **ESPript**, *'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 ESPript figure (full size in PDF)

- **Key features:**
  - ESPript 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 ESPript figure shows:**
  - aligned sequences,
  - similarities,
  - consensus,
  - accessibility,
  - hydropathy,
  - secondary structures elements,
  - intermolecular contacts,
  - user-supplied markers.

ESPript 3.0

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

ENDscript / ESPript 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 Choose File No file chosen

Example file • Tutorial

Number sequences

After the file is uploaded

Secondary structure depiction

Top secondary structures | Parameters

Input file *Upload a file below OR click here:* PDB

Choose File No file chosen

Chain ID

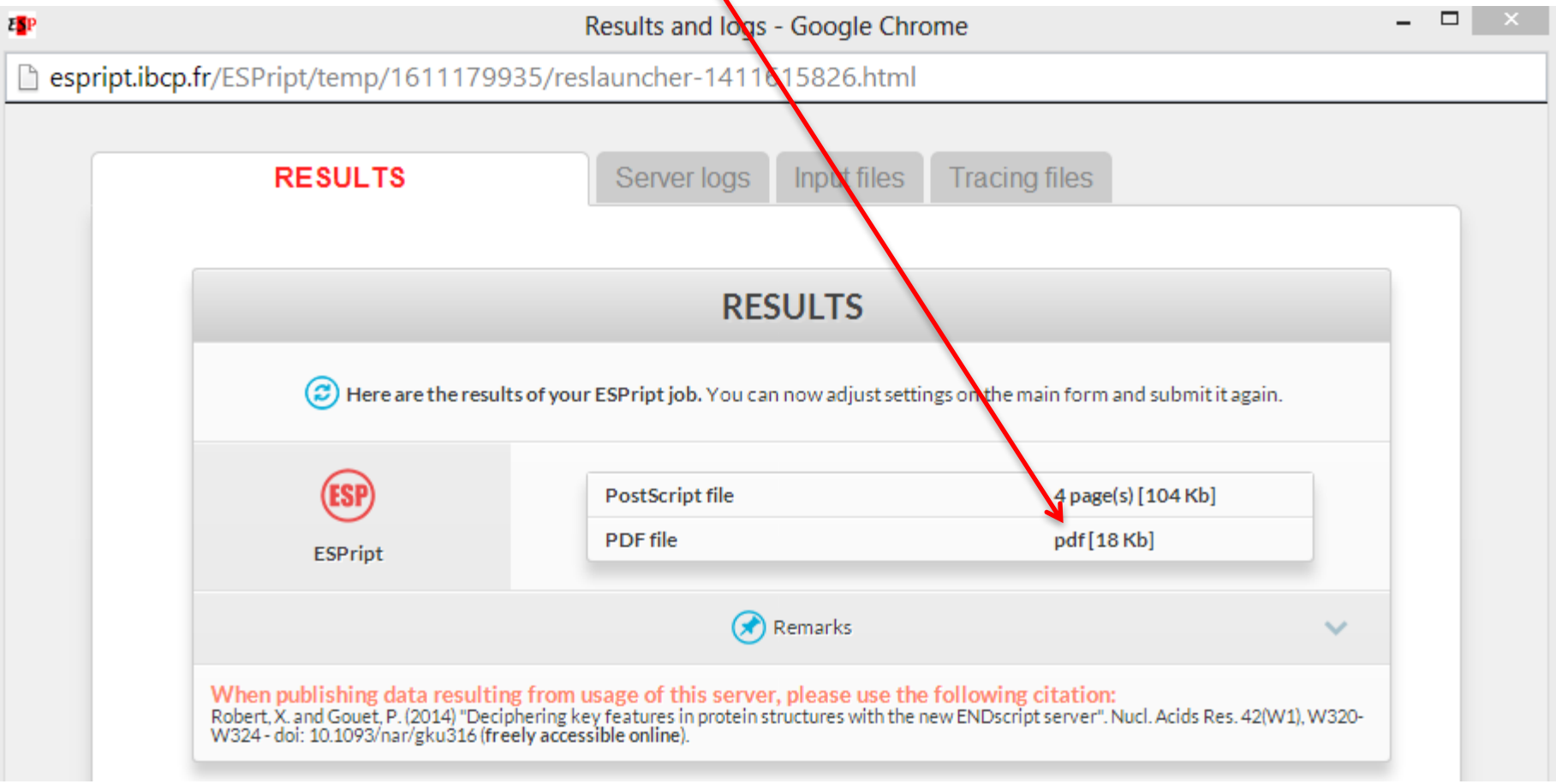
Relative accessibility

Sec. structure labels: α1,β1,α2,β2,...

Sec. structure labels: αA,βA,αB,βB,...

Sec. structure labels: α1,βA,α2,βB,...

A new window popped out, view in PDF



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://espript.ibcp.fr/ESPript/ESPript/esp_tutorial.php

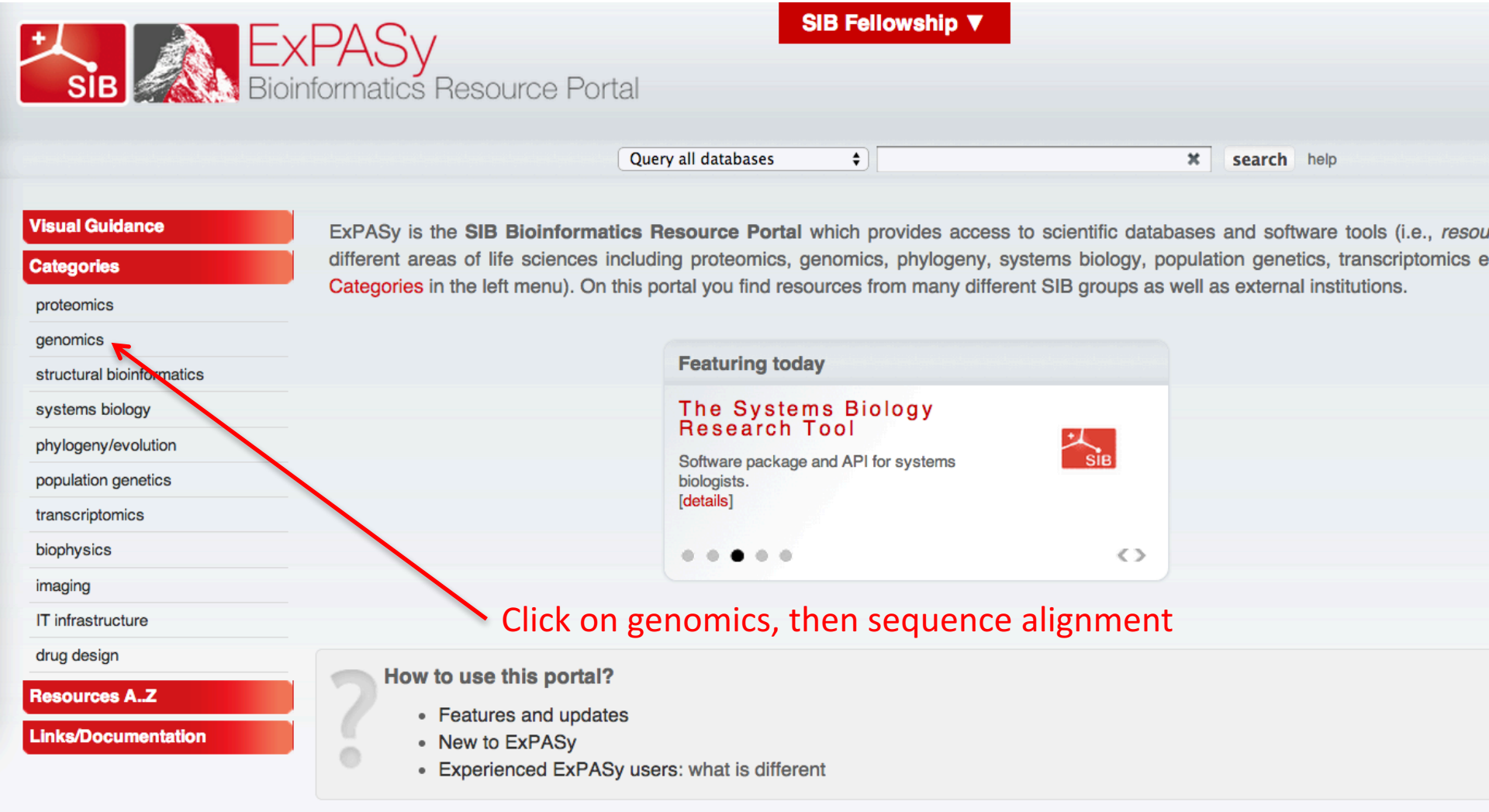
```
                                                150        160        170                180
AT5G22740.1|AT5G22740.1|cslA     TVK.QMVEV.....ECQRWASKGINIRYQIRENRVGY....KAGALKEGLKRSYVK.HCE
AT2G24630.1|AT2G24630.1|cslC     SIQ.ELIRD.....EVTKWSQKGVNIIYRHRLVRTGY....KAGNLKSAMSCDYVE.AYE
AT2G21770.1|AT2G21770.1|cesA     CDMD................GNELPRLVYVSREKRPGFDHHKKAGAMNSLIRVSAVLSNAP
AT1G02730.1|AT1G02730.1|cslD     EPVYGAEADAENLIDTTDVDIRLPMLVYVSREKRPGYDHNKKAGAMNALVRTSAIMSNGP
os_42915|LOC_Os07g36610.1|cslF   EPQLGMPASSGHPLDFSAVDVRLPILVYIAREKRPGYDHQKKAGAMNAQLRVSALLSNAP
AT1G55850.1|AT1G55850.1|cslE     ..E...............GNTIAIPTLVYLSREKRPQHHHNFKAGAMNALLRVSSKITCGK
AT4G23990.1|AT4G23990.1|cslG     MDD...............TKKYIMPNLIYVSREKSKVSSHHFKAGALNTLLRVSGVMTNSP
AT2G32530.1|AT2G32530.1|cslB     VG................VENEVPHFVYISREKRPNYLHHYKAGAMNFLVRVSGLMTNAP
os_25268|LOC_Os04g35020.1|cslH   ...........................ERRNHPT.....IVKTRVSAVMTNAP

                                190        200        210                220        230        240
AT5G22740.1|AT5G22740.1|cslA     YVVIFDADF.QPEPDFLRRSIPFLMH....NPNIALVQARWRFVNSDECLLTRMQEMS.L
AT2G24630.1|AT2G24630.1|cslC     FVAIFDADF.QPNSDFLKLTVPHFKE....KPELGLVQARWAFVNKDENLLTRLQNIN.L
AT2G21770.1|AT2G21770.1|cesA     YLLNVDCDHYINNSKAIREAMCFMMDPQS.GKKICYVQFPQRFDGIDRHDRYSNRNVVFF
AT1G02730.1|AT1G02730.1|cslD     FILNLDCDHYIYNSMALREGMCFMLD.RG.GDRICYVQFPQRFEGIDPNDRYANHNTVFF
os_42915|LOC_Os07g36610.1|cslF   FIFNFDGDHYINNSQAFRAALCFMLDCRH.GDDTAFVQFPQRFDDVDPTDRYCNHNRVFF
AT1G55850.1|AT1G55850.1|cslE     IILNLDCDMYANNSKSTRDALCILLDEKE.GKEIAFVQFPQCFDNVTRNDLYGSMMRVGI
AT4G23990.1|AT4G23990.1|cslG     IILTLDCDMYSNDPATPVRALCYLTDPKI.KTGLGFVQFPQTFQGISKNDIYACAYKRLF
AT2G32530.1|AT2G32530.1|cslB     YMLNVDCDMYANEADVVRQAMCIFLQKSMNSNHCAFVQFPQEFYDSNADEL.....TVLQ
os_25268|LOC_Os04g35020.1|cslH   IMLNMDCDMFVNNPQAVLHAMCLLLGFDD.EASSGFVQAPQRFYDALKDDPFGNQMECFF
```

ExPASy: Expert Protein Analysis System at SIB
Collection of external/internal tools

http://expasy.org/

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



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 aligment • [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 aligment • [more]
- MAFFT - EBI • Multiple sequence aligment • [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

WEBLOGO

· about · create · examples ·

Multiple Sequence Alignment

Upload Sequence Data: Choose File No file chosen

Image Format & Size

Image Format: PNG (bitmap)

Logo Size per Line:

Advanced Logo Options

Sequence Type: amino acid DNA / RNA Automatic Detection

First Position Number: 1

Logo Range:

Small Sample Correction:

Frequency Plot:

Multiline Logo (Symbols per Line): (32)

Advanced Image Options

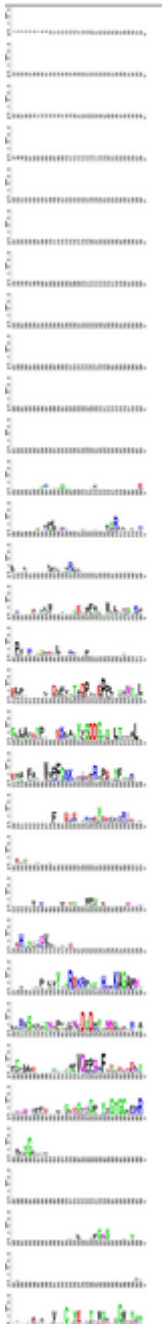
Bitmap Resolution: 96 pixels/inch (dpi)

Antialias Bitmaps:

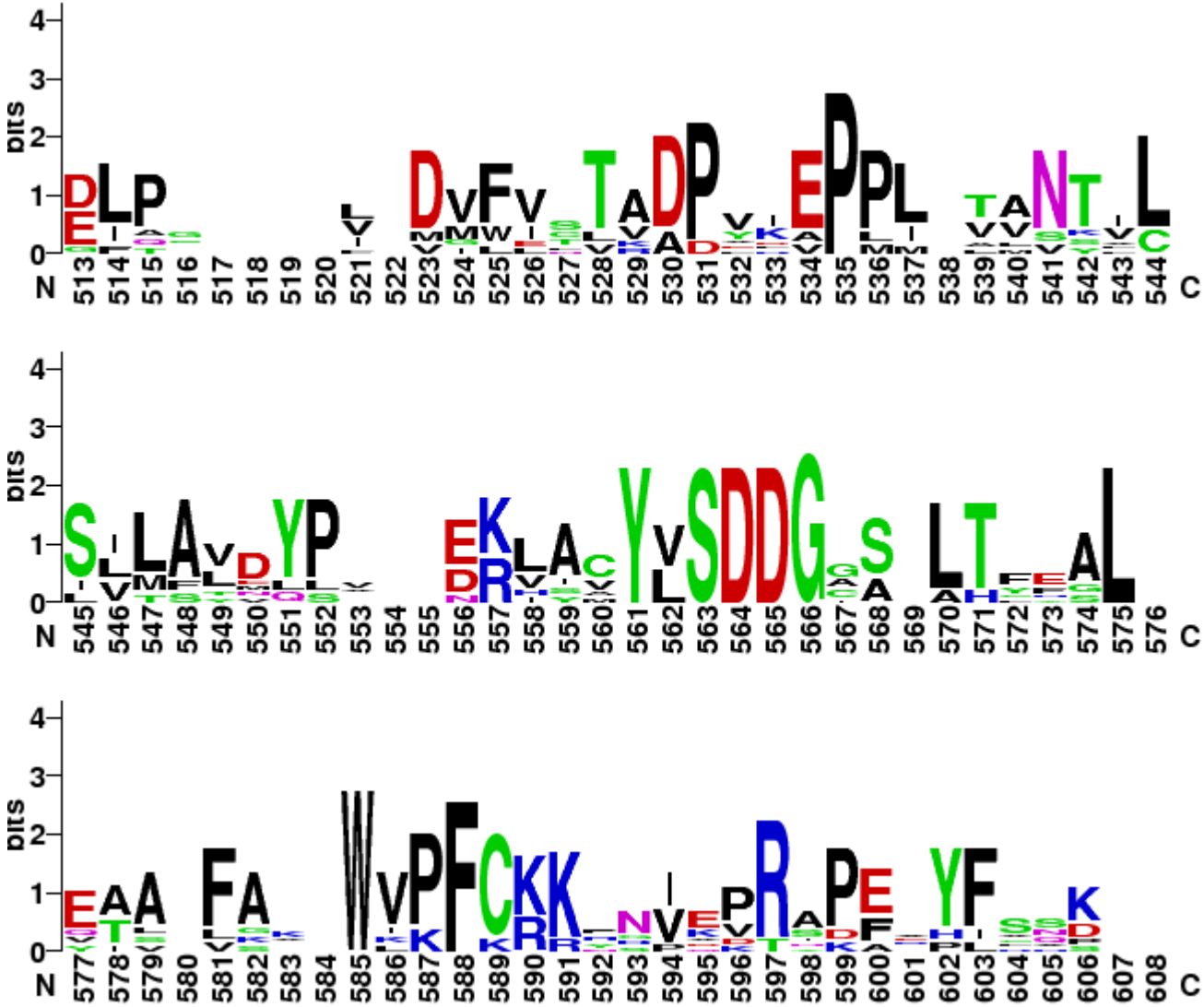
Title:

Y-Axis Height:

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|cslD        EPVYGAEADAENLIDTTDVDIRLPMLVYVSREKRPGYDHNKKAGAMNALVRTSAIMSNGP
140  os_42915|LOC_Os07g36610.1|cslF      EPQLGMPASSGHPLDFSAVDVRLPILVYIAREKRPGYDHQKKAGAMNAQLRVSALLSNAP
141  AT1G55850.1|AT1G55850.1|cslE        --E--------------GNTIAIPTLVYLSREKRPQHHHNFKAGAMNALLRVSSKITCGK
142  AT4G23990.1|AT4G23990.1|cslG        MDD--------------TKKYIMPNLIYVSREKSKVSSHHFKAGALNTLLRVSGVMTNSP
143  AT2G32530.1|AT2G32530.1|cslB        VG---------------VENEVPHFVYISREKRPNYLHHYKAGAMNFLVRVSGLMTNAP
144  os_25268|LOC_Os04g35020.1|cslH      ------------------------------ERRNHPT----------IVKTRVSAVMTNAP
145                                                                      *                   .
146
147  AT5G22740.1|AT5G22740.1|cslA        YVVIFDADF-QPEPDFLRRSIPFLMH----NPNIALVQARWRFVNSDECLLTRMQEMS-L
148  AT2G24630.1|AT2G24630.1|cslC        FVAIFDADF-QPNSDFLKLTVPHFKE----KPELGLVQARWAFVNKDENLLTRLQNIN-L
149  AT2G21770.1|AT2G21770.1|cesA        YLLNVDCDHYINNSKAIREAMCFMMDPQS-GKKICYVQFPQRFDGIDRHDRYSNRNVVFF
150  AT1G02730.1|AT1G02730.1|cslD        FILNLDCDHYIYNSMALREGMCFMLD-RG-GDRICYVQFPQRFEGIDPNDRYANHNTVFF
151  os_42915|LOC_Os07g36610.1|cslF      FIFNFDGDHYINNSQAFRAALCFMLDCRH-GDDTAFVQFPQRFDDVDPTDRYCNHNRVFF
152  AT1G55850.1|AT1G55850.1|cslE        IILNLDCDMYANNSKSTRDALCILLDEKE-GKEIAFVQFPQCFDNVTRNDLYGSMMRVGI
153  AT4G23990.1|AT4G23990.1|cslG        IILTLDCDMYSNDPATPVRALCYLTDPKI-KTGLGFVQFPQTFQGISKNDIYACAYKRLF
154  AT2G32530.1|AT2G32530.1|cslB        YMLNVDCDMYANEADVVRQAMCIFLQKSMNSNHCAFVQFPQEFYDSNADEL-----TVLQ
155  os_25268|LOC_Os04g35020.1|cslH      IMLNMDCDMFVNNPQAVLHAMCLLLGFDD-EASSGFVQAPQRFYDALKDDPFGNQMECFF
156                                       :  .* *    :       :  :           **    *
157
158  AT5G22740.1|AT5G22740.1|cslA        DYHFTVEQEVGSST----------HAFFGFNGTA--------------------------
159  AT2G24630.1|AT2G24630.1|cslC        CFHFEVEQQVNGVF----------LNFFGFNGTA--------------------------
160  AT2G21770.1|AT2G21770.1|cesA        DINMKGLDGIQGPIYVGTGCVFRRQALYGFDAPKKKQPPGRTCNCWPKWCCLCCGMRKKK
161  AT1G02730.1|AT1G02730.1|cslD        DVSMRALDGLQGPMYVGTGCIFRRTALYGFSPPRATEHHG-----WLGRRKVKISLRRPK
162  os_42915|LOC_Os07g36610.1|cslF      DATLLGLNGVQGPSYVGTGCMFRRVALYGADPPRWRP----------------------
163  AT1G55850.1|AT1G55850.1|cslE        DVEFLGLDGNGGPLYIGTGCFHRRDVICGRKYGEEEE----------------------
164  AT4G23990.1|AT4G23990.1|cslG        EINMIGFDGLMGPNHVGTGCFFNRRGFYGAPSNLILPEID--------------------
165  AT2G32530.1|AT2G32530.1|cslB        SYLGRGIAGIQGPTYAGSGCFHTRRVMYGLSIDDLEDDGSL------------------
166  os_25268|LOC_Os04g35020.1|cslH      KRFISGVQGVQGAFYAGTGCFHRRKAVYGVPPNFNGAER-E------------------
167                                                  .                 . *
```

Paste the copied segment here

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

WEBLOGO · about · create · examples

Multiple Sequence Alignment

```
os_42915|LOC_Os07g36610.1|cslF      FIFNFDGDHYINNSQAFRAALCFMLDCRH-
GDDVAFVQFPQRFDDVDPTDRYCNHNRVFF
AT1G55850.1|AT1G55850.1|cslE        IILNLDCDMYANNSKSTRDALCILLDEKE-
GKEIAFVQFPQCFDNVTRNDLYGSMMRVGI
AT4G23990.1|AT4G23990.1|cslG        IILTLDCDMYSNDPATPVRALCYLTDPKI-
KTGLGFVQFPQTFQGISKNDIYACAYKRLF
AT2G32530.1|AT2G32530.1|cslB
YMLNVDCDMYANEADVVRQAMCIFLQKSMNSNHCAFVQFPQEFYDSNADEL-----TVLQ
os_25268|LOC_Os04g35020.1|cslH      IMLNMDCDMFVNNPQAVLHAMCLLLGFDD-
EASSGFVQAPQRFYDALKDDPFGNQMECFF
```

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:

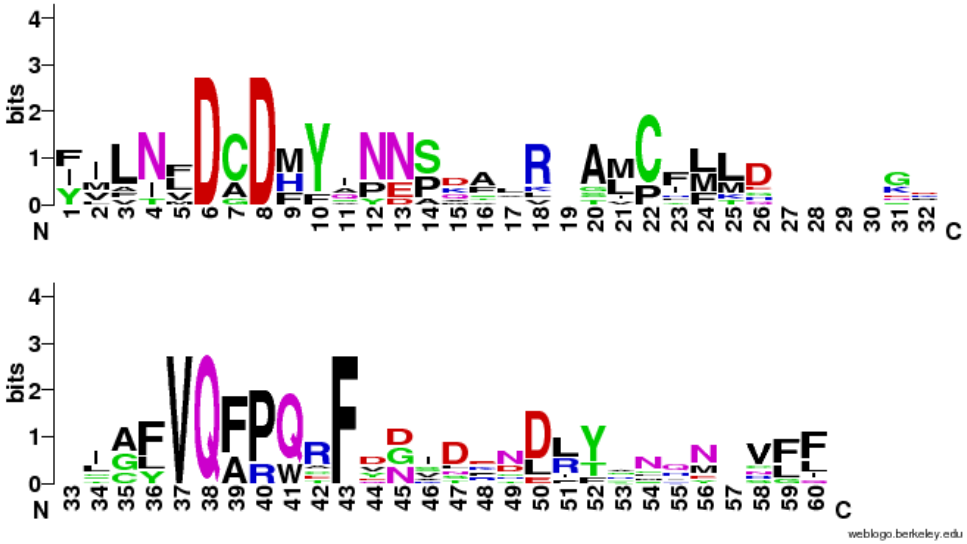
Multiline Logo (Symbols per Line):

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 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]

We are gonna try this website

http://phylogeny.lirmm.fr/phylo_cgi/index.cgi

phylogeny.lirmm.fr/phylo_cgi/index.cgi

CSR Internet - Study | Sample Applications | Bioinformatics 1 Cou | Bioinformatics Cours | I519: Introduction to | BMIF 310: Foundatio | Libraries Advisory C | Index of /bmi576/le | Education - Training

Méthodes et Algorithmes pour la Bio-informatique

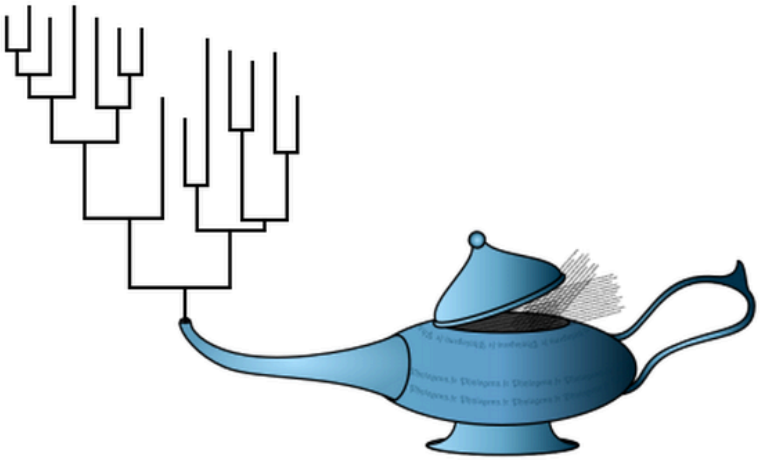
LIRMM

Home | Phylogeny | Blast: Sequence | Online Programs | Your Workspace | Documentation | Downloads | Contacts

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

Explorer

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:

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(Web Server issue):W465-9. Epub 2008 Apr 19. (PubMed)** *: joint first authors

Three modes of phylogeny reconstruction

Try the one click mode

http://cys.bios.niu.edu/yyin/teach/PBB/cesa-pr.fa

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

"One Click" Mode

Alignment MUSCLE → Curation Gblocks → Phylogeny PhyML → Tree Rendering TreeDyn

One click mode uses these tools

1. Overview | 2. Data & Settings

Name of the analysis (optional): csl

Give this job a name

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
VVMAVNVMAIGAAAGKAVVGGWTPAQVAGASAGLVFNVWVLVLLYPFALGIMGRWSKRPCALFALLVAAC
AAVAAGFVAVHAVLAAGSAAPSWLGWSRGATAILPSSWRLKRGF
>os_25268|LOC_Os04g35020.1|cslH
MAVVAAAAATGSTTRSGGGGGEGTRSGRKKPPPPPLQERVPLGRRAAWAWRLAGLAVLLLLLALLALRLL
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

Gblocks is a program automatically edit the alignment

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

csl

Alignment MUSCLE → Curation Gblocks → Phylogeny PhyML → Tree Rendering TreeDyn

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

Tree Rendering results

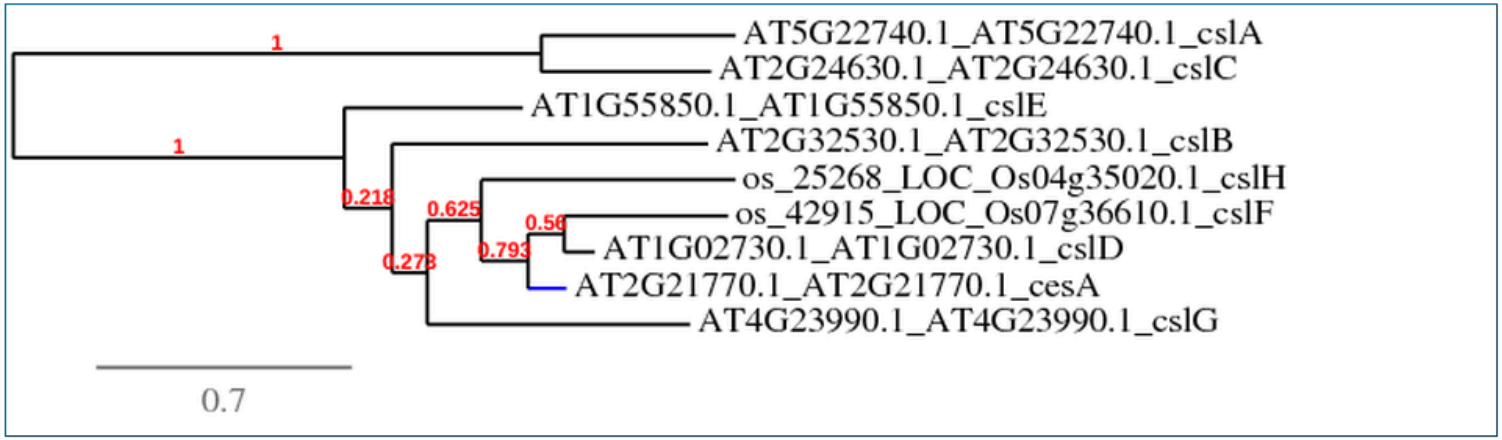


Figure 1: *Phylogenetic tree (the branch length is proportional to the number of substitutions per site).*

Dynamic Tree Edition

Color ☑ leaf ☑ branch using color black

and assign the group name

Add **annotations** using color red

Reset to original tree

Reroot (**outgroup**)

Swap subtrees

Reroot using **mid-point** rooting

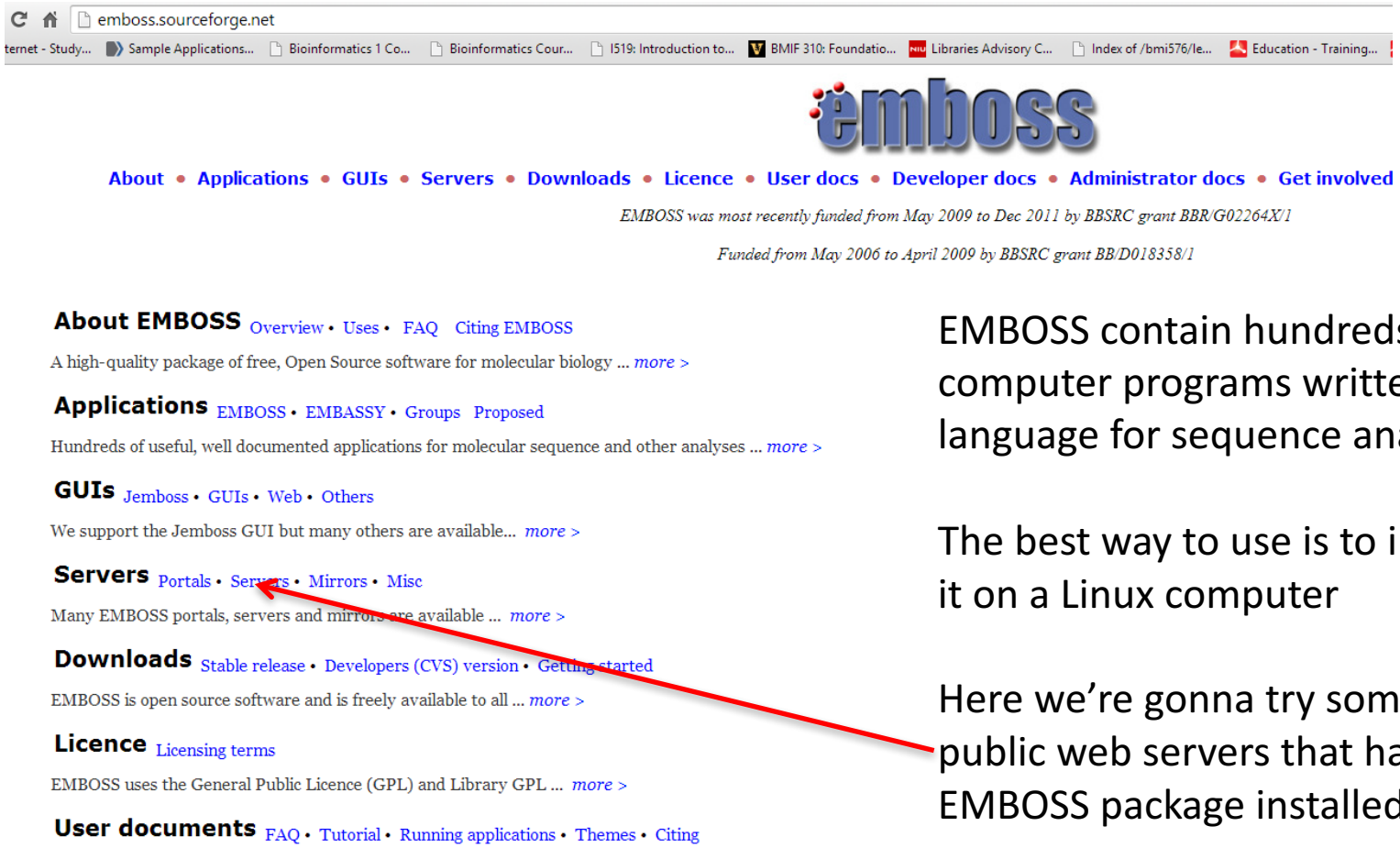
Flip subtree

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

http://emboss.sourceforge.net/



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.sourceforge.net/servers/#pise

rnet - Study... | Sample Applications... | Bioinformatics 1 Co... | Bioinformatics Cour... | I519: Introduction to... | BMIF 310: Foundatio

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

http://www.bioinformatics.nl/emboss-explorer/

350+ programs put into different groups

emboss.bioinformatics.nl

CSR Internet - Study... | Sample Applications... | Bioinformatics 1 Co... | Bioinformatics Cour... | I519: Introduction to... | BMIF 310: Foundatio... | Libraries Advisory C

[sort alphabetically]

ALIGNMENT CONSENSUS
cons
consambig
megamerger
merger

ALIGNMENT DIFFERENCES
diffseq

ALIGNMENT DOT PLOTS
dotmatcher
dotpath
dottup
polydot

ALIGNMENT GLOBAL
esim4
est2genome

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 s

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.

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
vrnalfold
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 applic the menu to display a short description. To search for a particular application, use wossname.

For more information about EMBOSS explorer, including how to download and install it locally, visit the EMB explorer website.

Development of EMBOSS explorer has been supported by the National Research Council of Canada and Genom Prairie.

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

NUCLEIC TRANSCRIPTION

jaspscan
tfscan

NUCLEIC TRANSLATION

backtranambig
backtranseq
coderet
plotorf
prettyseq
remap
showorf
showseq
sixpack
transeq

transeq

nucleotide

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: Choose File No file chosen
3. To enter the sequence data manually, type here:

```
GTGCTACAAGCAGAGCATCGAGGCGGCATGCAGGCTCACTTG
GGAAGGCCGCATGCACGTCCAGGTCCTCGATGACTCAGACGA
CGAGGAGATCCAGGCGCTCATCAGAGACGAGGTCGGCAG
GTGGCAGCAGAAGAACATGAGCATCGTATATCTCCACCGGAC
AACCCGGACTGGCCACAAGGCCGGCAACCTCGCGTTTGGGAT
GGCGGAGGCCGAGCCCCAGGGCTTCCAGTTTGTGGTGATCTT
TGACGCGGACTTCATGCCCAGCCCCGACTTCCTGCAGCG
CACCGTGCCGCTCTTCCGCGGCCGC
```

Additional section

Frame(s) to translate: -2, -3, Reverse three frames, All six frames

Choose all six frames

Code to use: Standard

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

vrnaalifoldpf
vrnacofold
vrnacofoldconc
vrnacofoldpf
vrnadistance
vrnaduplex
vrnaeval
vrnaevalpair
vrnafold
vrnafoldpf
vrnaheat
vrnainverse
vrnalfold
vrnaplot
vrnasubopt

C TRANSCRIPTION
jaspscan
tfscan

EIC TRANSLATION
backtranambig
backtranseq
coderet
plotorf
prettyseq
remap
showorf
showseq
sixpack

OUTPUT FILE outseq

This is likely the right frame

```
>JO181643.1_1
RDKMHVGFPRSLGNKRIVGNGSTRLFGVTWKHVLHPGLATSFRTLCTVVTGSCSGSGGAI
PGTAGRQALPIHRASVDMGLQVSNACPRERDCAAGGGSAVLGCRGGQAGGGVCPDPDPHV
QREGVLQAEHRGGMQADLGRPHARPGPR*LRRRGDPGAHQRRGRQVAAEEHEHRISPPDN
PDWPQGRQPRVWDGGGRAPGLPVCGDL*RGLHAQPRLPAAHRAALPRPX
>JO181643.1_2
GIRCTWAFLDHWATSASLATVLLVCLALPGSMSSIQVLQPVFELCAQLLQALVRALEVLF
LAQLADKLFLFTGRLWIWASKYRMPVLERETVPQEEDQLSLVAEEGKLEAESVLIQIPMC
NERECYKQSIEAACRLTWEGRMHVQVLDDSDDEEIQALIRDEVGRWQQKNMSIVYLHRTT
RTGHKAGNLAFGMAEAEPQGFQFVVIFDADFMPSPDFLQRTVPLFRGR
>JO181643.1_3
G*DARGLS*IIGQQAHRWQRFYSFVWRYLEACPPSRSCNQFSNFVHSCYRLLFGLWRCYS
WHSWPTSSSYSPGVCGYGPPSIECLSSRERLCRRRRISCPWLPRRASWRRSLS*SRSPCA
TRGSATSRASRRHAG*LGKAACTSRSSMTQTTRRSRRSSETRSAGGSRRT*ASYISTGQP
GLATRPATSRLGWRRPSPRASSLW*SLTRTSCPAPTSCSAPCRSSAAA
>JO181643.1_4
RPRKSGTVRCRKSGLGMKSASKITTNWKPWGSASAIPNARLPALWPVRVVRWRYTMLMFF
CCHLPTSSLMSAWISSSSESSRTWTCMRPSQVSLHAASMLCL*HSLSLHMGIWIRTDSAS
SLPSSATKDS*SSSCGTVSLSRTGIRYLEAHIHRRPVNRKSLSASCARNSTSRARTRACN
NCAQSSKTGCKTWMEDMLPGNAKQTSRTVANDALVAQ*SRKAHVHLIP
>JO181643.1_5
AAAEERHGALQEVGAGHEVRVKDHHKLEALGLGLRHPKREVAGLVASPGCPVEIYDAHVL
LLPPADLVSDERLDLLVV*VIEDLDVHAAFPSQPACRLDALLVALPLVAHGDLDQDRLRL
QLALLGNQGQLILLLRHSLSLEDRHSILGGPYPQTPGE*EELVGQLCQE*HLQSPNKSL*
QLCTKFENWLQDLDGGHASR*RQTNE*NRCQRCACCPMI*ESPRASYPX
>JO181643.1_6
GRGRAARCAAGSRGWA*SPRQRSPQTGSPGARPPPSQTRGCRPCGQSGLSGGDIRCSCSS
AATCRPRL**APGSPRRLSHRGPGRACGLPKSACMPPRCSACSTPSRCTWGSGSGQTPPP
ACPPRQPRTADPPPAAQSLSRGQAFDTWRPISTDAR*IGRACRPAVPGIAPPEPEQEPVT
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 q

Enter Query Sequence

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

JO181643.1

Put the seq ID here because it is in GenBank already

Query subrange

From

To

Or, upload file Choose File No file chosen

Genetic code Standard (1)

Job Title

Enter a descriptive title for your BLAST search

Align two or more sequences

Choose swiss-prot because it is smaller and high quality

Choose Search Set

Database UniProtKB/Swiss-Prot(swissprot)

Organism
Optional

Exclude +

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

Exclude
Optional

Models (XM/XP) Uncultured/environmental sample sequences

Entrez Query
Optional

Create custom database

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

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 Reformat

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:

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---------ELCAQLLQALVRALEVLFLAQLADKLFLFTGRLW  229
            L +P S   IQ L  +F         +  A  ++AL +   VLFL Q  D+L L  G LW
Sbjct  140  LHIPTSKLEIQSLLHLFYVGWLSLRADYIAPPIKALSKFCIVLFLVQSVDRLILCLGCLW  199

Query  230  IWASKYRMPVLERETVPQEEDQLSLVAEEGKLEAESVLIQIPMCNERECYKQSIEAACRL  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  DYVEAYEFVAIFDADFQPNSDFLKLTVPHFKEK  345
```

vrnacofold
vrnacofoldconc
vrnacofoldpf
vrnadistance
vrnaduplex
vrnaeval
vrnaevalpair
vrnafold
vrnafoldpf
vrnaheat
vrnainverse
vrnalfold
vrnaplot
vrnasubopt

CLEIC TRANSCRIPTION

jaspscan
tfscan

NUCLEIC TRANSLATION

backtranambig
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: Choose File No file chosen
3. To enter the sequence data manually, type here:

```
>Chaetosphaeridium_globosum|gb|JO181643.1
CGGGATAAGATGCACGTGGGCTTTCCTAGATCATTGGGCAAC
AAGCGCATCGTTGGCAACGGTTCTACTCGTTTGTTTGGCGTTA
CCTGGAAGCATGTCCTCCATCCAGGTCTTGCAACCAGTTTTCG
AACTTTGTGCACAGTTGTTACAGGCTCTTGTTCGGGC
TCTGGAGGTGCTATTCCTGGCACAGCTGGCCGACAAGCTCTT
CCTATTCACCGGGCGTCTGTGGATATGGGCCTCCAAGTATCG
AATGCCTGTCCTCGAGAGAGAGACTGTGCCGCAGGAGGAGGA
```

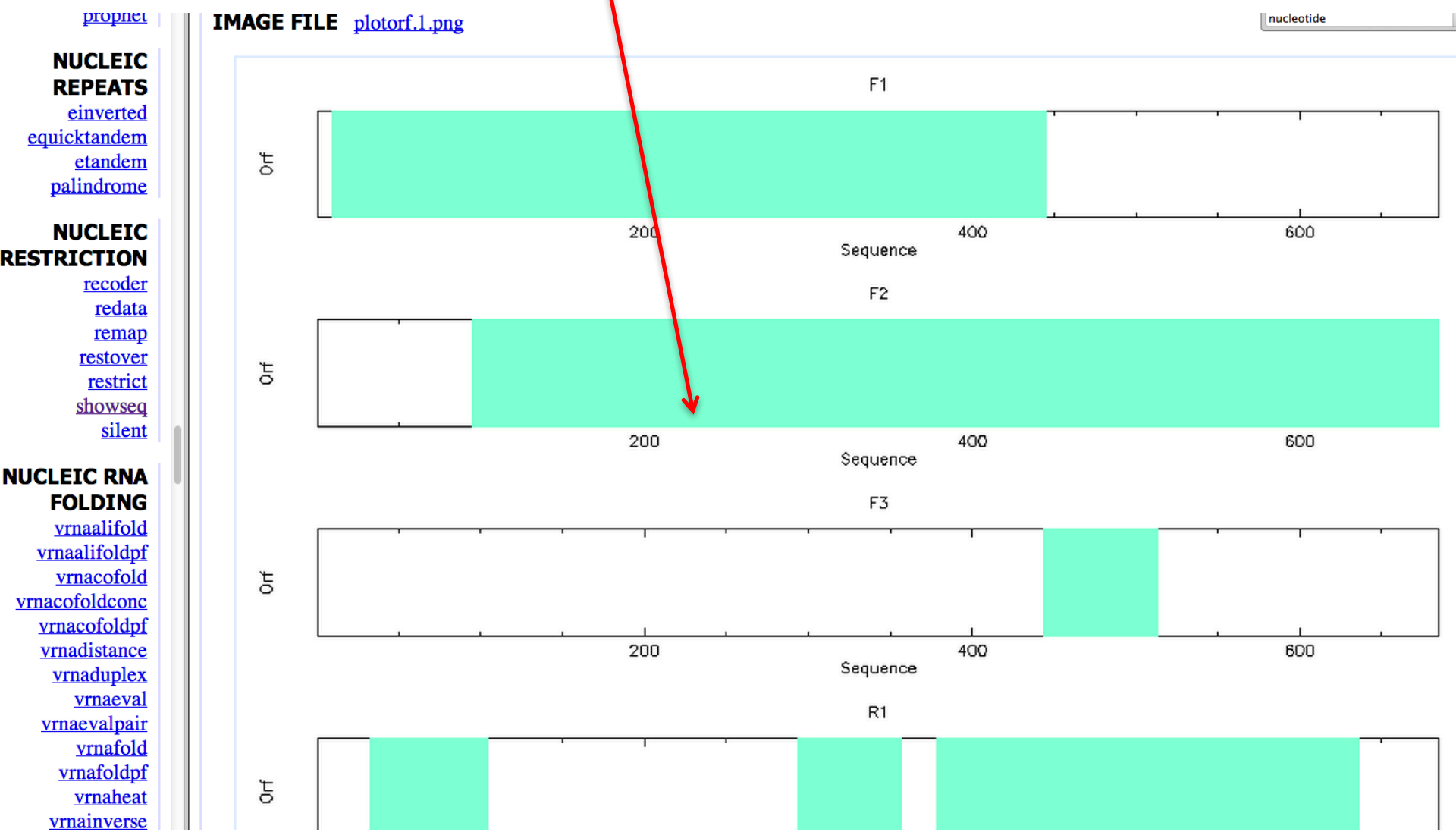
Input section

Advanced section

Start codons ATG

Stop codons TAA,TAG,TGA

This is the longest ORF



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

ATGCGCTA

Complement

TACGCGAT

Reverse

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: Choose File No file chosen
3. To enter the sequence data manually, type here:

Advanced section

Reverse sequence? Yes

Complement sequence? Yes

Output section

```
>JO181643.1
GCGGCCGCGGAAGAGCGGCACGGTGCGCTGCAGGAAGTCGGGGCTGGGCATGAAGTCCGC
GTCAAAGATCACCACAAACTGGAAGCCCTGGGGCTCGGCCTCCGCCATCCCAAACGCGAG
GTTGCCGGCCTTGTGGCCAGTCCGGGTTGTCCGGTGGAGATATACGATGCTCATGTTCTT
CTGCTGCCACCTGCCGACCTCGTCTCTGATGAGCGCCTGGATCTCCTCGTCGTCTGAGTC
ATCGAGGACCTGGACGTGCATGCGGCCTTCCCAAGTCAGCCTGCATGCCGCCTCGATGCT
CTGCTTGTAGCACTCCCTCTCGTTGCACATGGGGATCTGGATCAGGACAGACTCCGCCTC
CAGCTTGCCCTCCTCGGCAACCAAGGACAGCTGATCCTCCTCCTGCGGCACAGTCTCTCT
CTCGAGGACAGGCATTCGATACTTGGAGGCCCATATCCACAGACGCCCGGTGAATAGGAA
GAGCTTGTCGGCCAGCTGTGCCAGGAATAGCACCTCCAGAGCCCGAACAAGAGCCTGTAA
CAACTGTGCACAAAGTTCGAAAACTGGTTGCAAGACCTGGATGGAGGACATGCTTCCAGG
TAACGCCAAACAAACGAGTAGAACCGTTGCCAACGATGCGCTTGTTGCCCAATGATCTAG
GAAAGCCCACGTGCATCTTATCCCG
```

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

ATGCGCTA

Region 2-7

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: Choose File No file chosen
3. To enter the sequence data manually, type here:

```
>Chaetosphaeridium_globosum|gb|JO181643.1
CGGGATAAGATGCACGTGGGCTTTCCTAGATCATTGGGCAAC
AAGCGCATCGTTGGCAACGGTTCTACTCGTTTGTTTGGCGTTA
CCTGGAAGCATGTCCTCCATCCAGGTCTTGCAACCAGTTTTCG
AACTTTGTGCACAGTTGTTACAGGCTCTTGTTCGGGC
TCTGGAGGTGCTATTCCTGGCACAGCTGGCCGACAAGCTCTT
CCTATTCACCGGGCGTCTGTGGATATGGGCCTCCAAGTATCG
AATGCCTGTCCTCGAGAGAGAGACTGTGCCGCAGGAGGAGGA
```

Required section

Required section

Regions to extract (eg: 4-57,78-94) 57-400

Region 57-400

sixpack
textsearch

EDIT
aligncopy
aligncopypair
biosed
codcopy
cutseq
degapseq
descseq
entret
extractalign

OUTPUT FILE outseq

```
>JO181643.1
CAACGGTTCTACTCGTTTGTTTGGCGTTACCTGGAAGCATGTCCTCCATCCAGGTCTTGC
AACCAGTTTTCGAACTTTGTGCACAGTTGTTACAGGCTCTTGTTCGGGCTCTGGAGGTGC
TATTCCTGGCACAGCTGGCCGACAAGCTCTTCCTATTCACCGGGCGTCTGTGGATATGGG
CCTCCAAGTATCGAATGCCTGTCCTCGAGAGAGAGACTGTGCCGCAGGAGGAGGATCAGC
TGTCCTTGGTTGCCGAGGAGGGCAAGCTGGAGGCGGAGTCTGTCCTGATCCAGATCCCCA
TGTGCAACGAGAGGGAGTGCTACAAGCAGAGCATCGAGGCGGCA
```

ATGCGCTA

GC%=50%

codcmp
cusp
syco

NUCLEIC COMPOSITION

banana
btwisted
chaos
compseq
dan
density
freak
isochore
sirna
wordcount

NUCLEIC CPG ISLANDS

cpgplot
cpgreport
geecee
newcpgreport
newcpgseek

NUCLEIC GENE FINDING

getorf
marscan

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: Choose File No file chosen
3. To enter the sequence data manually, type here:

```
>Chaetosphaeridium_globosum|gb|JO181643.1
CGGGATAAGATGCACGTGGGCTTTCCTAGATCATTGGGCAAC
AAGCGCATCGTTGGCAACGGTTCTACTCGTTTGTTTGGCGTTA
CCTGGAAGCATGTCCTCCATCCAGGTCTTGCAACCAGTTTTCG
AACTTTGTGCACAGTTGTTACAGGCTCTTGTTCGGGC
TCTGGAGGTGCTATTCCTGGCACAGCTGGCCGACAAGCTCTT
CCTATTCACCGGGCGTCTGTGGATATGGGCCTCCAAGTATCG
AATGCCTGTCCTCGAGAGAGAGACTGTGCCGCAGGAGGAGGA
```

Required section

Residue letters gc

Calculate GC content

Output section

Produce graphic? Yes

Change to yes to get a pic

Output graphic format PNG

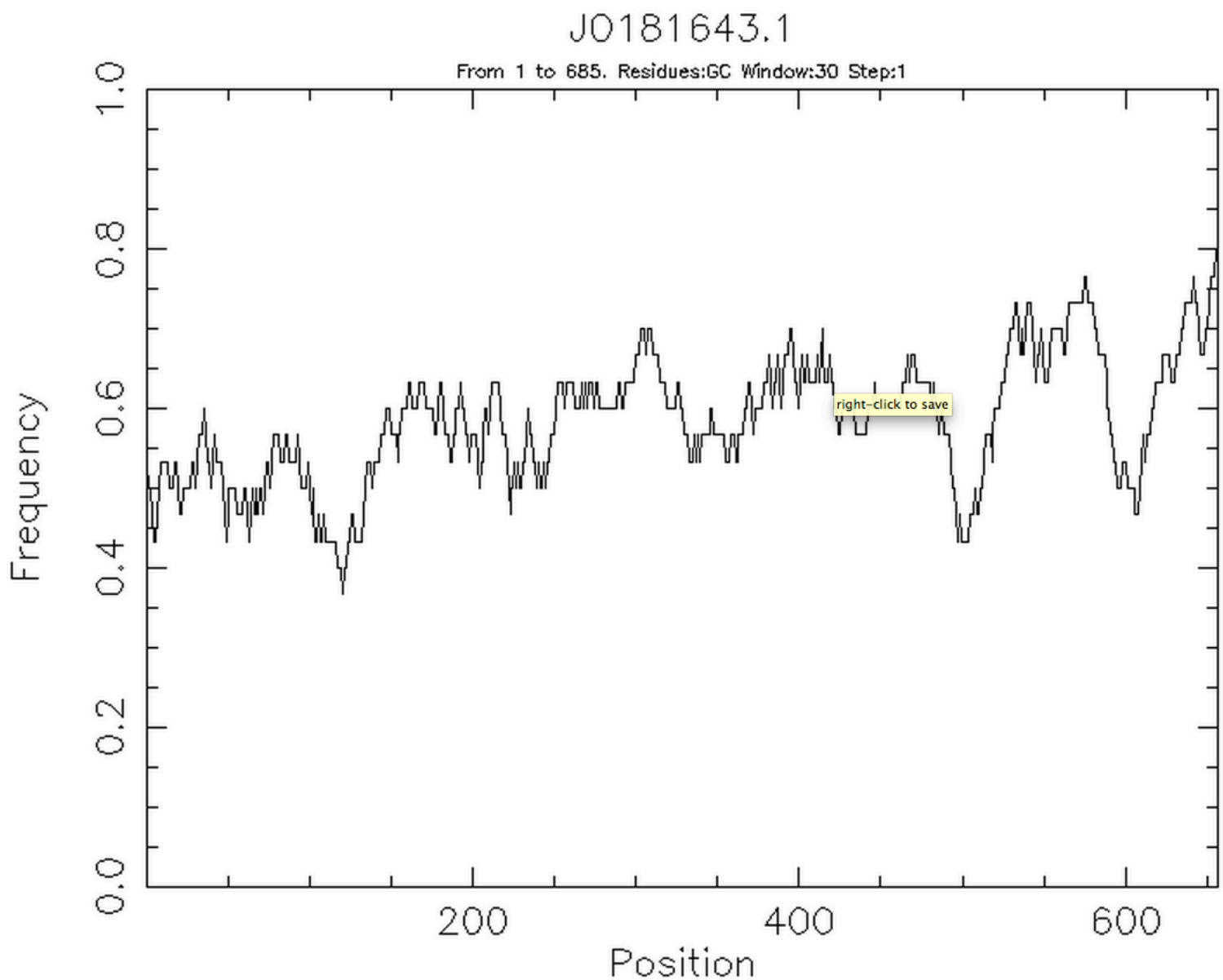
chips
codcmp
cusp
syco

POSITION
banana
btwisted
chaos
compseq
dan
density
freak
isochore
sirna
wordcount

G ISLANDS
cpgplot
cpgreport
geecee
newcpgreport
newcpgseek

E FINDING
getorf
marscan
plotorf
showorf
sixpack
syco
tcode
wobble

IMAGE FILE freak.1.png



ATGCGCTA

16 possible dinuc
64 possible trinuc
256 possible tetranuc

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

compseq

Calculate the composition of unique words in 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: Choose File No file chosen
3. To enter the sequence data manually, type here:

```
>Chaetosphaeridium_globosum|gb|JO181643.1
CGGGATAAGATGCACGTGGGCTTTCCTAGATCATTGGGCAAC
AAGCGCATCGTTGGCAACGGTTCTACTCGTTTGTTTGGCGTTA
CCTGGAAGCATGTCCTCCATCCAGGTCTTGCAACCAGTTTTCG
AACTTTGTGCACAGTTGTTACAGGCTCTTGTTCGGGC
TCTGGAGGTGCTATTCCTGGCACAGCTGGCCGACAAGCTCTT
CCTATTCACCGGGCGTCTGTGGATATGGGCCTCCAAGTATCG
AATGCCTGTCCTCGAGAGAGAGACTGTGCCGCAGGAGGAGGA
```

Input section

Program compseq output file (optional): Choose File No file chosen

Required section

Default is dinuc

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

Application: scan genome to look for regions with abnormal compositions

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

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

#
# 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
```

Equal occurrence: 1/16

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

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

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: Choose File No file chosen
3. To enter the sequence data manually, type here:

```
>AT2G21770.1|AT2G21770.1|cesA
MNTGGRLIAGSHNRNEFVLINADDTARIRSAEELSGQTCKICRDEIEL
TDNGEPFIACNECAFPTCRPCYEYERREGNQACPQCGTRYKRIKGS
PRVEGDEEDDDIDDLEHEFYGMDPEHVTEAALYYMRLNTGRGTDE
VSHLYSASPGSEVPLLTYCDEDSDMYSDRHALIVPPSTGLGNRVHH
VPFTDSFASIHTRPMVPQKDLTVYGYGSVAWKDRMEVWKKQQIEK
LQVVKNERVNDGDGDGFIVDELDDPGLPMMDEGRQPLSRKLPIRSS
RINPYRMLIFCRLAILGLFFHYRILHPVNDAFGLWLTSVICEIWFAVSW
```

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: Eamino.dat
2. To upload a data file from your local computer, select it here: Choose File 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: Emolwt.dat
2. To upload a data file from your local computer, select it here: Choose File No file chosen

seqsearch
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

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

http://www.cbs.dtu.dk/services/

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

http://www.cbs.dtu.dk/services/

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

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

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.

Vert

Instructions | Output format | Ab

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:

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.*

```
Translation start predictions for 1 dicot plant sequence


  Name: Chaetosphaeridium_gl
CGGGATAAGATGCACGTGGGCTTTCCTAGATCATTGGGCAACAAGCGCATCGTTGGCAACGGTTCTACTCGTTTGTTTGG
CGTTACCTGGAAGCATGTCCTCCATCCAGGTCTTGCAACCAGTTTTCGAACTTTGTGCACAGTTGTTACAGGCTCTTGTT
CGGGCTCTGGAGGTGCTATTCCTGGCACAGCTGGCCGACAAGCTCTTCCTATTCACCGGGCGTCTGTGGATATGGGCCTC
CAAGTATCGAATGCCTGTCCTCGAGAGAGAGACTGTGCCGCAGGAGGAGGATCAGCTGTCCTTGGTTGCCGAGGAGGGCA
AGCTGGAGGCGGAGTCTGTCCTGATCCAGATCCCCATGTGCAACGAGAGGGAGTGCTACAAGCAGAGCATCGAGGCGGCA
TGCAGGCTGACTTGGGAAGGCCGCATGCACGTCCAGGTCCTCGATGACTCAGACGACGAGGAGATCCAGGCGCTCATCAG
AGACGAGGTCGGCAGGTGGCAGCAGAAGAACATGAGCATCGTATATCTCCACCGGACAACCCGGACTGGCCACAAGGCCG
GCAACCTCGCGTTTGGGATGGCGGAGGCCGAGCCCCAGGGCTTCCAGTTTGTGGTGATCTTTGACGCGGACTTCATGCCC
AGCCCCGACTTCCTGCAGCGCACCGTGCCGCTCTTCCGCGGCCGC
.........N.............................................................................
..............N.........................................................................
.......................................................................i...............
..........i..............................................................................
.................................N.......................................................N
.......................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
« 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

FAQ	Article abstracts	Instructions	Output format

SUBMISSION

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

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*

Gra
- ( ) N
- (•) P
- ( ) P

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

Pos

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: ClustalX and MEGA