Multiple sequence alignment

Making a multiple sequence alignment with ClustalW

Making and comparing multiple sequence alignments with Tcoffee

Comparing sequences you cannot align

Many criteria for building a multiple sequence alignment

Criterion	Meaning
Structural similarity	Amino acids that play the same role in each structure are in the same column. Structure superposition programs are the only ones that use this criterion.
Evolutionary similarity	Amino acids or nucleotide related to the same amino acid (or nucleotide) in the common ancestor of all the sequences are put in the same column. No automatic program explicitly uses this criterion, but they all try to deliver an alignment that respects it.
Functional similarity	Amino acids or nucleotides with the same function are in the same column. No automatic program explicitly uses this criterion, but if the information is available, you can force some programs to respect it or you can edit your alignment manually
Sequence similarity	Amino acids in the same column are those that yield an alignment with maximum similarity. Most programs use sequence similarity because it is the easiest criterion. When the sequences are closely related, structural, evolutionary, and functional similarities are equivalent to sequence similarity

Main applications of multiple sequence alignments

Application	Procedure
Extrapolation	A good multiple alignment can help convince you that an uncharacterized sequence is really a member of a protein family
Phylogenetic analysis	If you carefully choose the sequences to include in your multiple alignment, you can reconstruct the history of these proteins
Pattern identification	By discovering very conserved positions, you can identify a region that is characteristic of a function (in proteins or in nucleic acid sequences)
Domain identification	It is possible to turn a multiple sequence alignment into a profile that describes a protein family or a protein domain. You can use this profile to scan databases for new members of the family.
DNA regulatory elements	You can turn a DNA multiple alignment of a binding site into a weight matrix and scan other DNA sequences for potential similar binding sites
Structure prediction	A good multiple alignment can give you an almost perfect prediction of your protein secondary structure for both proteins and RNA. Sometimes it can also help in the building of a 3-D model
PCR analysis	A multiple alignment can help you identify the less degenerated portions of a protein family, in order to fish out new members by PCR. If this is what you want to do, you can use the following site: blocks.fhcrc.org/codehop.html.

Evolutionary rules

Important amino acids (or nucleotides) are not allowed to mutate

Less important residues change more easily, sometimes randomly, and sometimes in order to adapt a function

⬇

Conserved & not conserved = important and less important

A few guidelines for selecting sequences

Problem	diagnostics
Proteins or DNA	Use proteins whenever possible
Many sequences	Start with 10-15 sequences and avoid aligning more than 50
Very different sequences	Sequences that are less than 30% identical with more than half of the other sequences in the set cause troubles
Identical sequences	They never help. Avoid using sequences of more than 90% identity
Partial sequences	Multiple alignment programs prefer sequences that are roughly the same length
Repeated sequences	Mostly cause problems

BLAST servers integrating multiple alignment methods

Address

- http://www.expasy.org/tools/blast/
- npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_blast.html
- srs.ebi.ac.uk

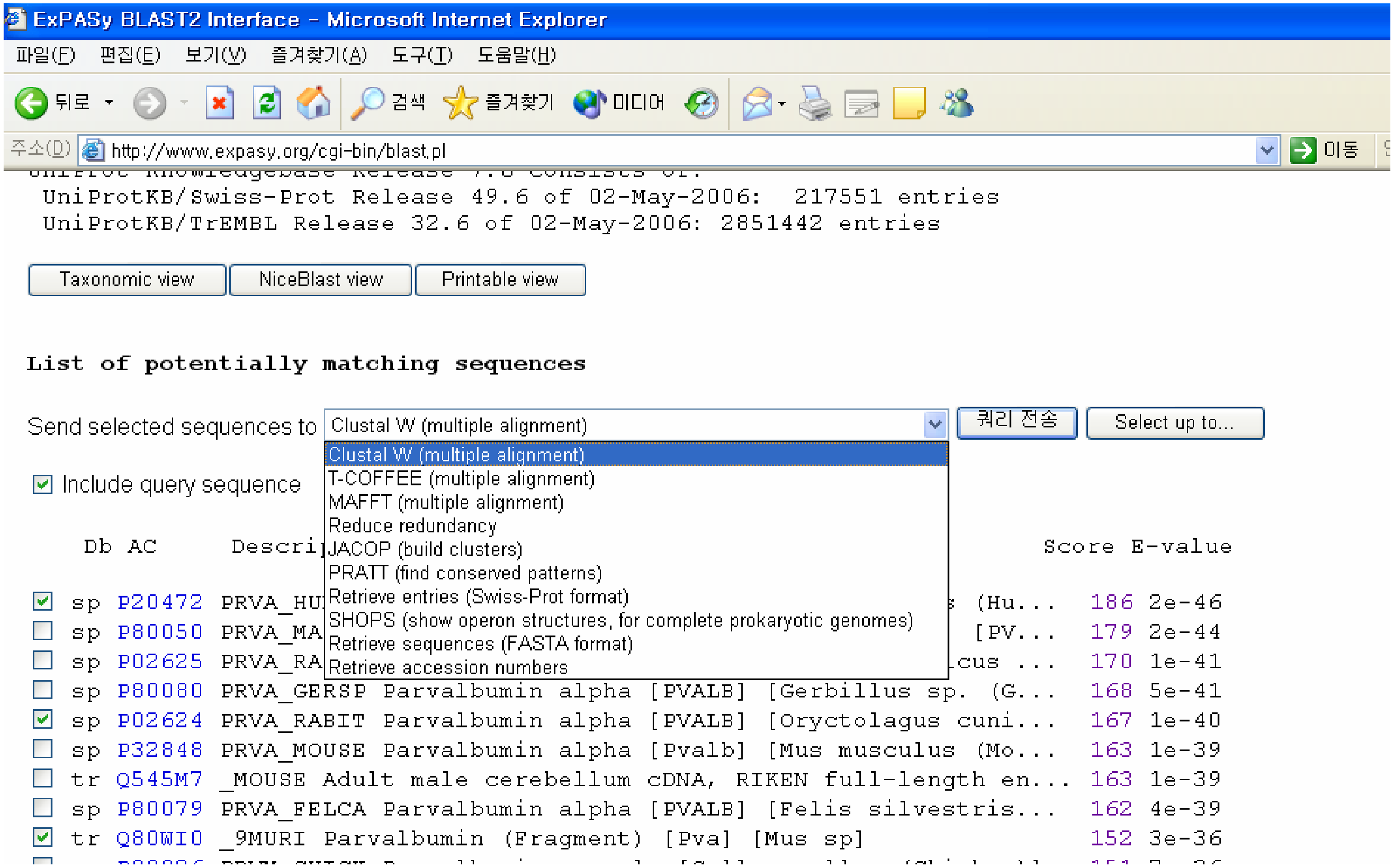
What you can do there

- Extract entire sequences
- Extract sequences in FASTA
- Submit sequences to ClustalW
- Submit sequences to Tcoffee

Multiple alignment at ExPASy server

Go to http://www.expasy.org/tools/blast/

Enter the sequence Accession Number or paste your sequence


ExPASy BLAST2 Interface - Microsoft Internet Explorer
파일(F) 편집(E) 보기(V) 즐겨찾기(A) 도구(T) 도움말(H)
뒤로
검색
즐겨찾기
미디어
주소(D)
http://www.expasy.org/cgi-bin/blast.pl
이동
Taxonomic view
NiceBlast view
Printable view
쿼리 전송
Select up to...

```
UniProt Knowledgebase Release 7.6 consists of:
 UniProtKB/Swiss-Prot Release 49.6 of 02-May-2006:  217551 entries
 UniProtKB/TrEMBL Release 32.6 of 02-May-2006: 2851442 entries
```

List of potentially matching sequences

Send selected sequences to: Clustal W (multiple alignment)

- Clustal W (multiple alignment)
- T-COFFEE (multiple alignment)
- MAFFT (multiple alignment)
- Reduce redundancy
- JACOP (build clusters)
- PRATT (find conserved patterns)
- Retrieve entries (Swiss-Prot format)
- SHOPS (show operon structures, for complete prokaryotic genomes)
- Retrieve sequences (FASTA format)
- Retrieve accession numbers

☑ Include query sequence

```
    Db  AC      Descri                                              Score E-value
☑  sp  P20472  PRVA_HU                                      s (Hu...   186  2e-46
☐  sp  P80050  PRVA_MA                                       [PV...    179  2e-44
☐  sp  P02625  PRVA_RA                                      cus ...    170  1e-41
☐  sp  P80080  PRVA_GERSP Parvalbumin alpha [PVALB] [Gerbillus sp. (G...  168  5e-41
☑  sp  P02624  PRVA_RABIT Parvalbumin alpha [PVALB] [Oryctolagus cuni...  167  1e-40
☐  sp  P32848  PRVA_MOUSE Parvalbumin alpha [Pvalb] [Mus musculus (Mo...  163  1e-39
☐  tr  Q545M7  _MOUSE Adult male cerebellum cDNA, RIKEN full-length en...  163  1e-39
☐  sp  P80079  PRVA_FELCA Parvalbumin alpha [PVALB] [Felis silvestris...  162  4e-39
☑  tr  Q80WI0  _9MURI Parvalbumin (Fragment) [Pva] [Mus sp]              152  3e-36
```

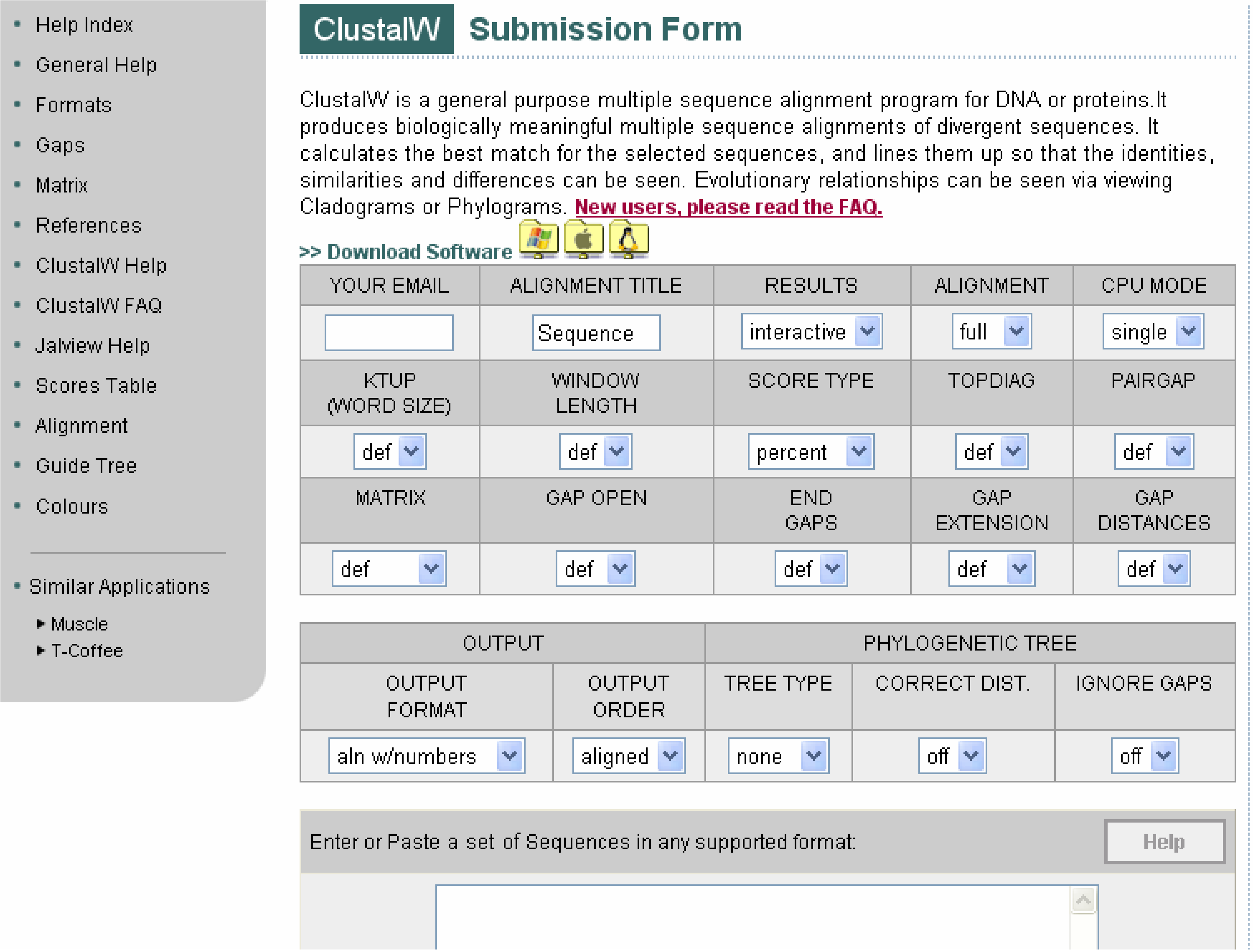
Choosing the right multiple sequence alignment method

Using ClustalW

The most commonly used program for multiple alignments

Pairwise alignment program: It uses a progressive algorithm,
building the alignment progressively

ClustalW at http://www.ebi.ac.uk/clustalw/index.html



ClustalW Submission Form

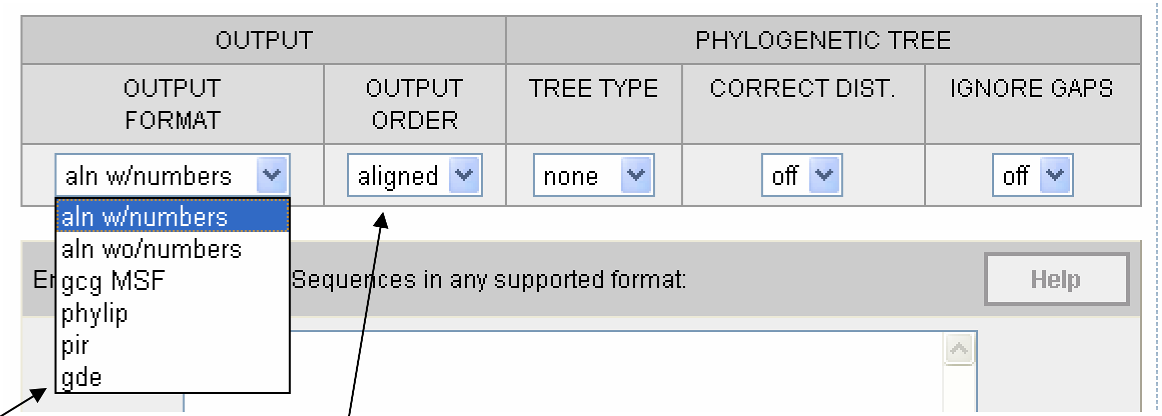
ClustalW is a general purpose multiple sequence alignment program for DNA or proteins.It produces biologically meaningful multiple sequence alignments of divergent sequences. It calculates the best match for the selected sequences, and lines them up so that the identities, similarities and differences can be seen. Evolutionary relationships can be seen via viewing Cladograms or Phylograms. **New users, please read the FAQ.**

>> Download Software

YOUR EMAIL	ALIGNMENT TITLE	RESULTS	ALIGNMENT	CPU MODE
	Sequence	interactive	full	single
KTUP (WORD SIZE)	WINDOW LENGTH	SCORE TYPE	TOPDIAG	PAIRGAP
def	def	percent	def	def
MATRIX	GAP OPEN	END GAPS	GAP EXTENSION	GAP DISTANCES
def	def	def	def	def

OUTPUT		PHYLOGENETIC TREE		
OUTPUT FORMAT	OUTPUT ORDER	TREE TYPE	CORRECT DIST.	IGNORE GAPS
aln w/numbers	aligned	none	off	off

Enter or Paste a set of Sequences in any supported format:



OUTPUT		PHYLOGENETIC TREE		
OUTPUT FORMAT	OUTPUT ORDER	TREE TYPE	CORRECT DIST.	IGNORE GAPS
aln w/numbers	aligned	none	off	off

Er Sequences in any supported format:

You can always change a format (www.bimcore.emory.edu/Pise/)

Aligned sequence order

The results come in three sections

Pairwise scores: pairwise comparison
The multiple alignments:
The guide tree: contains the tree that ClustalW used to guide its progressive alignment strategy
there are several options if you click mouse button on the graph
you can also see a true phylogenetic tree

Changing ClustalW parameters

Substitution matrix: no effect, if your sequences are closely related
Gap-opening penalty: the higher the value, the more difficult it is to insert a gap
Gap-extension penalty: controls the size of the gaps

It is better not to change parameters in order to force ClustalW to produce an alignment that you know is right

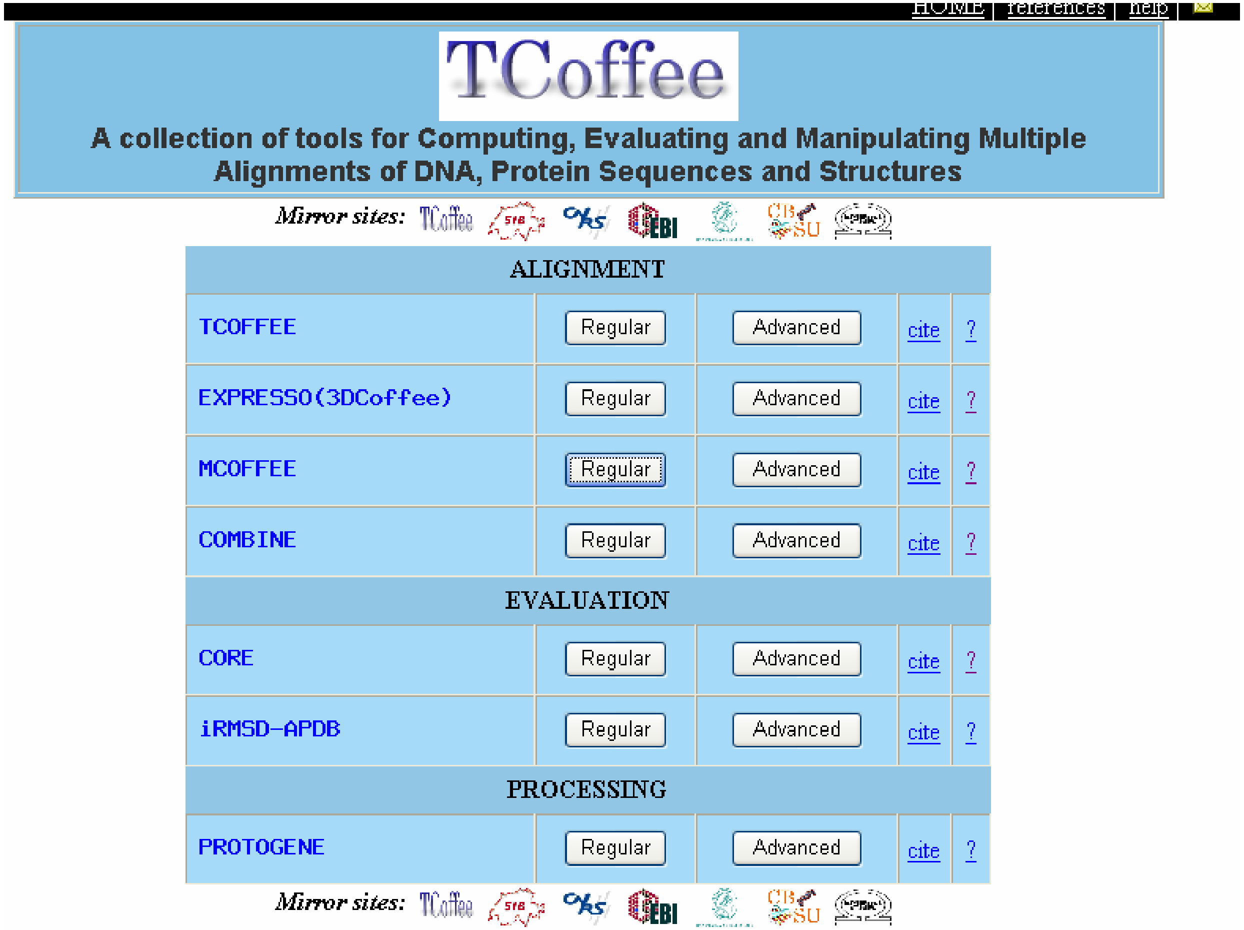
Making & evaluating alignments with Tcoffee

One of the most recently developed methods
More accurate alignments at the cost of a slightly longer running time
It compares segments across the entire sequence set

Usage	Description
Multiple alignment	
Evaluation	Evaluate the reliability of an existing multiple alignment
using structures	If some of your sequence have a known structure, Tcoffee can use them to help the alignment.
Combining alignments	If you have several alignments of the same sequences produced with different methods (ClustalW and Tcoffee), you can use Tcoffee to combine these alignments into a single one. Tcoffee also shows you the regions where your alignments agree most

Point your browser to the Tcoffee server

http://tcoffee.vital-it.ch/cgi-bin/Tcoffee/tcoffee_cgi/index.cgi



TCoffee

A collection of tools for Computing, Evaluating and Manipulating Multiple Alignments of DNA, Protein Sequences and Structures

Mirror sites:

ALIGNMENT				
TCOFFEE	Regular	Advanced	cite	?
EXPRESSO(3DCoffee)	Regular	Advanced	cite	?
MCOFFEE	Regular	Advanced	cite	?
COMBINE	Regular	Advanced	cite	?
EVALUATION				
CORE	Regular	Advanced	cite	?
iRMSD-APDB	Regular	Advanced	cite	?
PROCESSING				
PROTOGENE	Regular	Advanced	cite	?

Mirror sites:

Interpreting your multiple sequence alignment

Surface loops that evolve rapidly: gap-rich blocks
Core regions inside the protein that evolve less rapidly: gap-free blocks

The last line contains signs such as (*), (:), or (.)
(*) A star indicates an entirely conserved column
(:) A colon indicates columns where all the residues have roughly the same size and the same hydropathy
(.) A period indicates columns where the size or the hydropathy has been preserved in the course of evolution

Good block: a unit at least 10-30 amino acids long exhibiting at least 1-3 stars, 5-7 colons, and a few periods

Important amino acids for evaluating conserved columns

Amino acids	characteristics
W, Y, F	conserved tryptophan is common
G, P	loops
C	disulfide bridges
H, S	catalytic sites
K, R, D, E	ligand binding and salt bridge
L	rarely conserved except leucine zipper

Advanced multiple alignments

motif-finding methods available online

Gibbs sampler: http://bioweb.pasteur.fr/seqanal/interfaces/gibbs-simple.html

local alignments

scrambles your sequences, aligns them randomly until a good solution appears

Other sites

Pratt
eMotif
MEME
TEIRESIAS
Bioprospector
Improbizer
BLOCK-Maker

Multiple alignment in the right format

A classification of multiple sequence alignment formats

Name	Type	Usage
Post-script, pdf, html	Graphic	terminal formats suitable for printing only
FASTA	text	easy to manipulate
PIR	text	similar to FASTA
MSF	text	most standard multiple alignment format
Selex	text	extended version of MSF
ALN	text	simplified version of MSF default output of ClustalW supported by many programs
Phylip	text	variant of ALN useful for doing phylogenetic analysis supported by most phylogenetic packages

Converting format

Pasteur Institute: http://bioweb.pasteur.fr/seqanal/interfaces/fmtseq.html

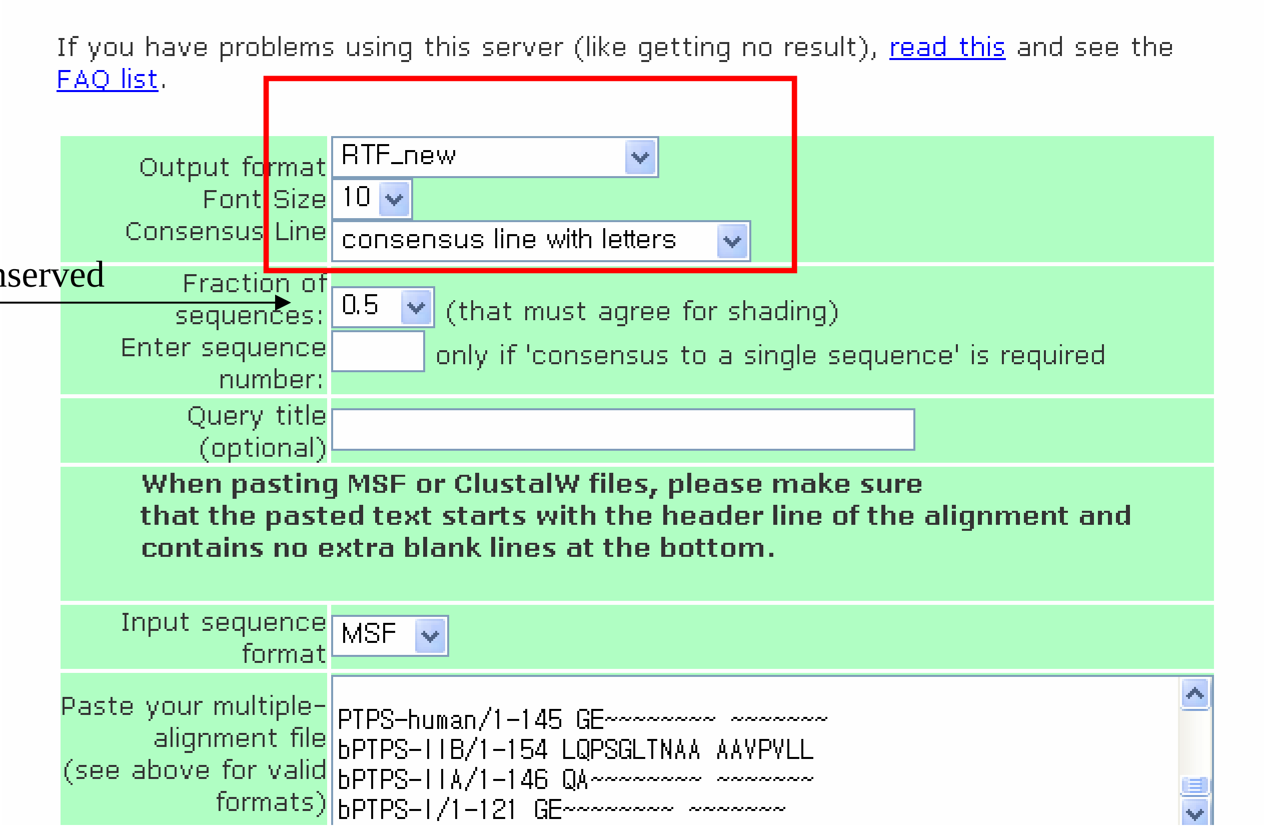
Others on the Web

- FMTSEQ
- READSEQ
- SEQCHECK

Multiple alignment for publication

1. Boxshade

www.ch.embnet.org/software/BOX_form.html



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

Output format: RTF_new

Font Size: 10

Consensus Line: consensus line with letters

Fraction of sequences: 0.5 (that must agree for shading)

Enter sequence number: only if 'consensus to a single sequence' is required

Query title (optional):

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

Input sequence format: MSF

Paste your multiple-alignment file (see above for valid formats)

```
PTPS-human/1-145 GE~~~~~~~~ ~~~~~~~~
bPTPS-IIB/1-154 LQPSGLTNAA AAVPVLL
bPTPS-IIA/1-146 QA~~~~~~~~ ~~~~~~~~
bPTPS-I/1-121 GE~~~~~~~~ ~~~~~~~~
```

Half of the amino acids to be conserved for some shading occur

Black: identical

Grey: similar

2. Logos

www.cbs.dtu.dk/~gorodkin/appl/plogo.html

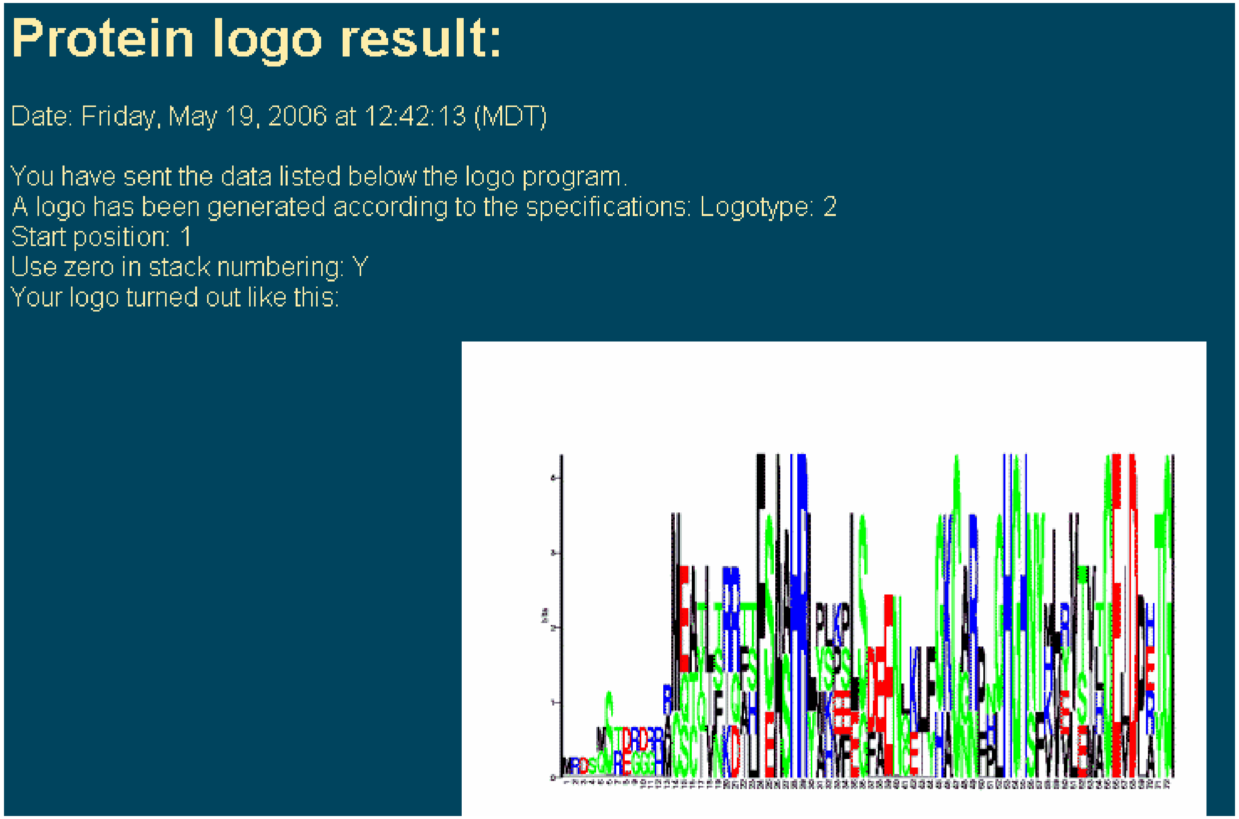
Get your alignment in FASTA

Copy & paste the FASTA alignment into a word processing program

Replace the name with a space for each sequence

> MSTEGGGRRCQAQVS

space



Protein logo result:

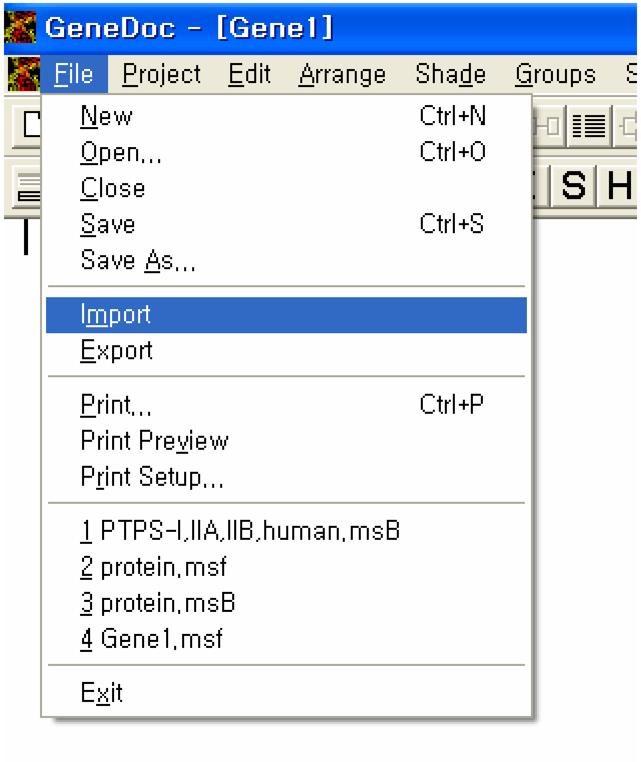
Date: Friday, May 19, 2006 at 12:42:13 (MDT)

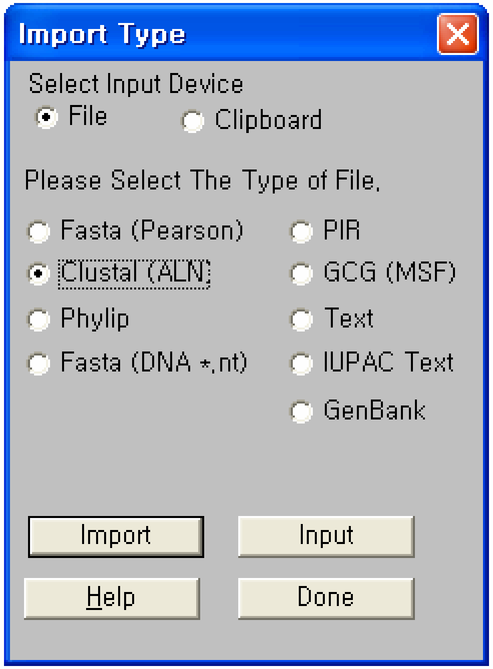
You have sent the data listed below the logo program.
A logo has been generated according to the specifications: Logotype: 2
Start position: 1
Use zero in stack numbering: Y
Your logo turned out like this:

3. GENEDOC

http://www.psc.edu/biomed/genedoc/

Install the program in your PC

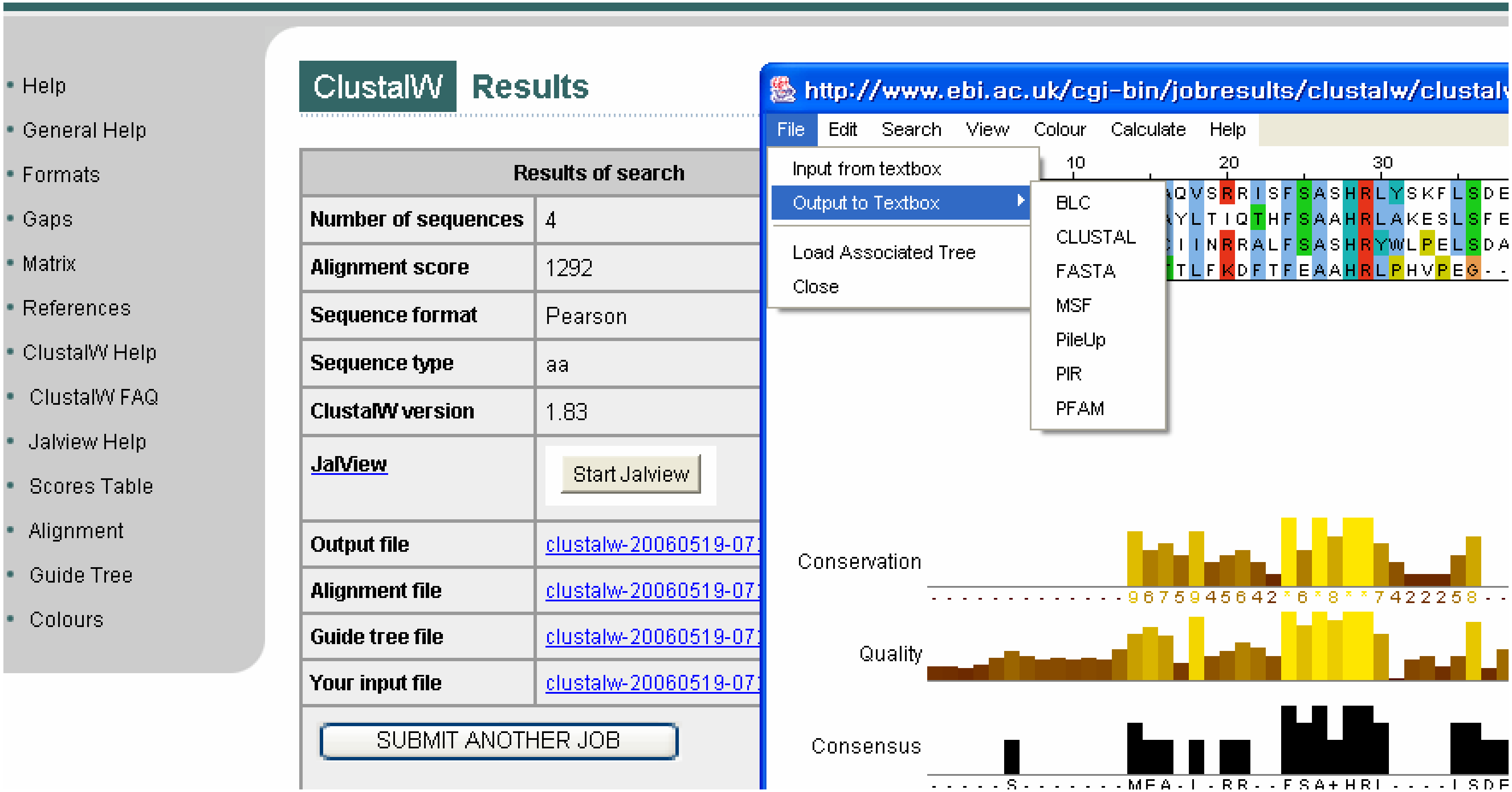




Select Input Device
● File ○ Clipboard

Please Select The Type of File.
○ Fasta (Pearson) ○ PIR
● Clustal (ALN) ○ GCG (MSF)
○ Phylip ○ Text
○ Fasta (DNA *,nt) ○ IUPAC Text
○ GenBank

Jalview http://www.ebi.ac.uk/clustalw/index.html



ClustalW Results

Results of search	
Number of sequences	4
Alignment score	1292
Sequence format	Pearson
Sequence type	aa
ClustalW version	1.83
JalView	Start Jalview
Output file	clustalw-20060519-071
Alignment file	clustalw-20060519-071
Guide tree file	clustalw-20060519-071
Your input file	clustalw-20060519-071

Assignment

Multiple alignment with Tcoffee:

- Tcoffee
- Expresso
- Mcoffee
- Core

Compare two groups of sequences:

- Combine