wEMBOSS interface to EMBOSS

EMBnet Course: Introduction to Bioinformatics

Geneva, 2 March 2006

Lorenza Bordoli

Swiss Institute of Bioinformatics

Outline

- What is EMBOSS?
- Major programs
- The wEMBOSS package

Why EMBOSS ?

History:

• Wisconsin (sequence analysis) package, **GCG** (Genetics Computer Group) founded in 1982 as a service of the Department of Genetics at the University of Wisconsin;

• Widely used, sources available for inspection (programs could be algorithmically verified and adapted to needs);

• Since 1998 EGCG (extended GCG) developed academic add-on, started as a small collection of programs to support EMBL's research activities, in particular the development of automated DNA sequencing;

• GCG became a private company in 1990, now belongs to Accelerys;

• Currently sources not freely available anymore, no longer possible to distribute academic software source code which uses the GCG libraries!

•1999 - EGCG split from GCG to become EMBOSS;

October 2005: version 3.0.0

What is EMBOSS ?

• http://emboss.sourceforge.net/

• EMBOSS, The *E*uropean *M*olecular *B*iology *O*pen Source *S*oftware *S*uite, is a package of high-quality FREE Open Source software for sequence analysis;

• EMBOSS includes hundreds of applications (+150). They share a similar interface, but each comes with its own documentation:

- Many sequence analysis & display programs.
- Protein 3D structure prediction being developed.
- Other assorted programs, eg: enzyme kinetics.
- Extensible (with some C programming knowledge)!

• Complete list of the programs in the currently release: http://emboss.sourceforge.net/apps/#Overview

• Grouped applications: http://emboss.sourceforge.net/apps/groups.html

EMBOSS !

- Free Open Source (for most Unix platforms, including MacOSX)
- GCG successor (compatible with GCG file format)
- Public domain (GNU Public License)
- Written by HGMP/Sanger/EBI/Denmark … etc
- Easy to install locally:
  but no interface, requires local databases
  Unix command-line only
- Interfaces:
  Jemboss, www2gcg, w2h, wEMBOSS… (with account)
  Pise, EMBOSS-GUI, SRS (no account)
  Staden, Kaptain, CoLiMate, Jemboss (local)

History *(cont.)*:

- The UK Medical Research Council is to close the Research and Bioinformatics Divisions of the RFCGR (the current home of EMBOSS) in the middle of 2005. The MRC Press Office has stated:
"All MRC can say at this stage is that Council have made a decision to close the Research and Bioinformatics Divisions. However, the Director has been asked to draw up a closing down plan for consideration by Council in July."
- "This action will more than halve the current core development team and will therefore adversely affect the development and support of EMBOSS. We hope that alternative sources of funding can be found."
- EMBOSS has moved to SourceForge.net (http://sourceforge.net/);

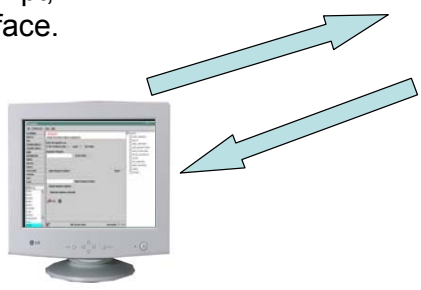
EMBOSS: Introduction

• The EMBOSS package consists of a large number of separate programs that have a specific function.

• They usually take a (number of) input file(s) and some parameters that are important to the function and produce output in the form of files, plots, web pages or simple text output.

Running EMBOSS Programs

EMBOSS programs are run by:
• typing them at the UNIX prompt,
• or by using a graphical interface.



Remote server:
ludwig-sun1.unil.ch

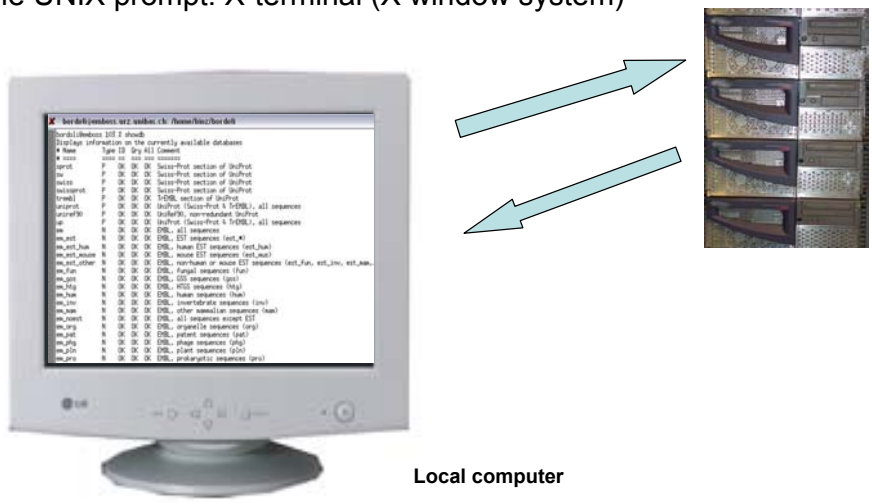
Remote server:
you personal account

Local computer:
your PC in the lab, in the course room,…

Running EMBOSS Programs

EMBOSS programs are run by:
• typing them at the UNIX prompt: X-terminal (X window system)

Remote server:
ludwig-sun1.unil.ch

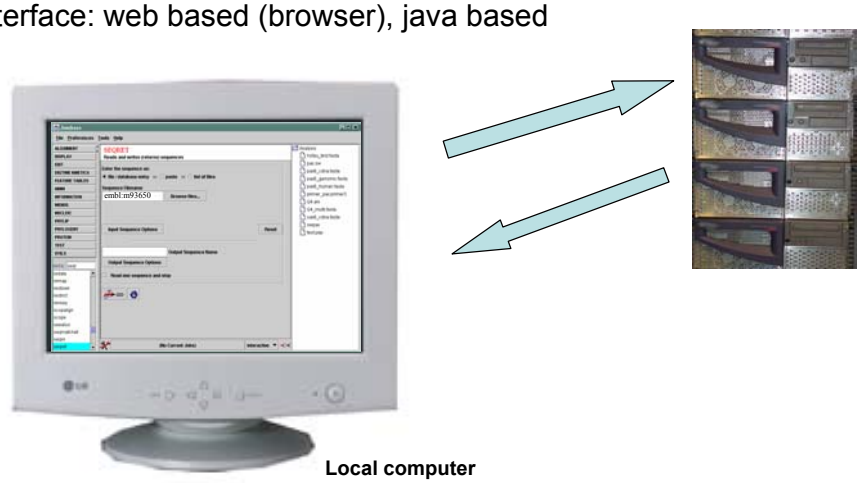


Local computer

Running EMBOSS Programs

EMBOSS programs are run by:
• or by using an interface: web based (browser), java based

Remote server:
ludwig-sun1.unil.ch



Local computer

Graphical interfaces to EMBOSS

- wEMBOSS: web interface to EMBOSS
- Jemboss: java based interface to EMBOSS
- others : http://emboss.sourceforge.net/

Some major programs:

- **General:**

wossname	lists all **EMBOSS** programs
showdb	shows the available databases

- **Sequence retrieval:**

seqret	retrieves and/or changes format of a sequence
seqretset	retrieve and or change formats of a number
seqretall	of sequences at once
transeq	translate a DNA sequence to protein
backtranseq	translate a protein sequence to DNA
extractseq	extract regions from a sequence
cutseq	remove a region from a sequence
pasteseq	inserts a sequence into another sequence
infoseq	display information about a sequence
splitter	split a sequence into smaller sequences

Some major programs *(cont.)*:

• **Sequence comparison**

needle	Needleman-Wusch sequence alignment
water	Smith-Waterman sequence alignment
stretcher	Myers and Miller global alignment
matcher	Huang and Miller local alignment
dottup	
dotmatcher	dotplot comparisons of two sequences.
prettyplot	plots multiple sequence alignments
polydot	
supermatcher	dotplot comparisons of multiple sequences
emma	ClustalW program (***clustal***, wEMBOSS 1.4.0: new wrapper)

• **Sequence parameters**

cusp	generates a codon usage table
syco	synonymous codon usage plot
dan	calculates DNA/RNA melting temperature
compseq	sequence composition tables

Some major programs *(cont.)*:

• **DNA Sequence features**

remap	restriction map of the sequence
remap	
cpgplot	
cpgreport	CpG island detection
etandem	
einverted	finds tandem and inverted repeats
plotorf	plots potential ORFs
showorf	pretty display of potential ORFs
fuzznuc	DNA pattern search
tfscan	scans sequence for TF binding sites

Some major programs *(cont.)*:

- **Protein Sequence features**

ief	Isoelectric point calculation
antigenic	Finds potential antigenic sites
digest	protein digestion map
findkm	Vmax and Km calculations
fuzzpro	protein pattern search
garnier	protein 2D structure prediction
helixturnhelix	finds nucleic acid binding motifs
octanol	
pepwindow	displays protein hydropathy
patmatdb	
patmatmotifs	searching with motifs vs protein sequences
pepcoil	predicts coiled coil regions
pepinfo	
pepstats	Protein information
Hammer package	***ehmmpfam***, ***ehmmsearch***, ***ehmmbuild***, …
Phylip package	***efitch***, ***edolpenny***, ***edollop***, …

Working with sequences :

- EMBOSS reads sequences from files or databases.
- It automatically recognizes the input sequence format.
- You can easily specify many output formats.

Uniform Sequence Address (USA)

- = a standard way of specifying a **sequence** to be read into a program in EMBOSS
- Sequences can be in **databases** or in **files**
- It has the following syntax:

```
format::database:entry
  format::file:entry
```

In general, a USA specifies
- what sequence format to expect
- what file or database to open
- what entry to look for

Uniform Sequence Address (USA)

```
format::database:entry
```

- Of these only the "`file`" or "`database`" are necessary;
- If format is omitted: EMBOSS will check and recognizes the format (occasionally needs a hint) * ;
- if the "`entry` " part is omitted, all of the entries in the file or database are read in;

* EMBOSS recognizes: GCG, FASTA, ClustalW, MSF, EMBL, GenBank, DNAStrider, Phylip, PIR, PAUP,ASN.1, NBRF, Fitch, IntelliGenetics

Uniform Sequence Address (USA)

The most common ways of specifying a sequence are:

• to type the name of the file that the sequence is in: `myfile.seq`

• or type "`db:entry`", where "`db`" is the name of the database and "`entry`" is either the ID or the accession number (AC) of the sequence in the database

Ex.:

database:accession	`embl:X65923`
database:ID	`swissprot:opsd_xenla`
file name	`myfile.seq`

ACs and IDs …

• An entry in a database: uniquely identified in that database. Most sequence databases have two such identifiers for each sequence - an ID name and an Accession number.

• Why are there two such identifiers?

  •The ID name: a human-readable name that had some indication of the function of its sequence: OPSD_HUMAN in Swiss-Prot
  !! ID names are not guaranteed to remain the same between different versions of a database.

  • Accession numbers: unique alphanumeric identifiers that are guaranteed to remain with that sequence through the rest of the life of the database: **P08100.** If two sequences are merged into one, then the new sequence will get a new Accession number and the Accession numbers of the merged sequences will be retained as 'secondary' Accession numbers.

Databases

You can easily find out what are the database name in your EMBOSS installation by running the ***showdb*** program:

```
Displays information on the currently available databases
#Name         Type ID  Qry All Comment
#====         ==== ==  === === =======
sw             P   OK  OK  OK  Swiss-Prot section of UniProt
swiss          P   OK  OK  OK  Swiss-Prot section of UniProt
swiss-prot     P   OK  OK  OK  Swiss-Prot section of UniProt
trembl         P   OK  OK  OK  TrEMBL section of UniProt
uniprot        P   OK  OK  OK  UniProt (Swiss-Prot & TrEMBL),
…
```

Databases

```
 #Name         Type ID  Qry All Comment
#====         ==== ==  === === =======
sw             P   OK  OK  OK  Swiss-Prot section of UniProt
```

• `ID` allows programs to extract a single explicitly named entry from the database: `embl:x13776` ;

• `Query` indicates that programs can extract a set of matching wildcard entry names: `sw:opsd_*` ;

• `All` allows programs to analyze all entries in the database sequentially: `embl:*` ;

Uniform Sequence Address (USA)

- you may also use:

`filename`	all sequences in a file
`filename:entry`	an entry in a file
`dbname`	all sequences in a database (not recommended)
`dbname:entry`	a sequence in a database
`@listfile`	a list file
`list::listfile`	a list file

Specifying a List File

• Instead of containing the sequences themselves, a `listefile` contains "references" to sequences using any valid USA.

• Example of a ListFile:

`opsd_abyko.fasta`	: the name of a sequence file;
`sw:opsd_xenla`	: a specific sequence in the Swiss-Prot database;
`@anotherlist`	: the name of a second list file;

• Blank lines and lines starting with a '#' character are ignored in List Files: a way to add your comments: this won't be read by the programs.

The full USA syntax

`filename`	: a file containing one or more sequences
`filename:entry` **`mysequences:opsd_xenla`**	: a given sequence in the file. The 'entry' is the ID or AC of the sequence in that file
`filename:entry[start:end]` **`mysequences:opsd_xenla[1:20]`** **`mysequences:opsd_xenla[-1:-20]`** **`mysequences:[1:20:r]`**	: a part of the sequence can be specified by the range : the last 20 residues/nucleotides : reverse-complemented (nucleotide sequences)

Specifying Search Fields

- Beside ID names or AC numbers there are other ways to specify sequences.
- A typical sequence entry in EMBL format is:

```
ID HSFAU standard; DNA; UNC; 518 BP.
AC X65923;
SV X65923.1
DE H.sapiens fau mRNA
KW fau gene.
OS Homo sapiens (human)
OC Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo. SQ Sequence 518
BP; 125 A; 139 C; 148 G; 106 T; 0 other;
```

- It is also useful to find sequences that contain words occurring in their description filed ("`DE`" line), their Keyword field ("`KE`" line), …

Specifying Search Fields

• You can specify which field you are searching by using one of the following Search Field Names:

Name	Searches for
acc	Accession number
des	Description
id	ID name
key	Keyword
org	Organism Name
sv	Sequence Version/GI Number

• Examples:

`embl-des:fau` : database

`embl-des:h*emoglobin` : database

`myclones.seq:des:fau` : file

Sequence Formats

• Sequences can be read and written in a variety of formats;

• Sequences are stored in databases or in files as simple text (ASCII text);

• Microsoft WORD format is not a sequence format (save the file as text *.txt file!!!)

• The default sequence file format is fasta:

```
>xyz some other comment
ttcctctttctcgactccatcttcgcggtagctgggaccgccgttcagtcgccaatatgc
agctctttgtccgcgcccaggagctacacaccttcgaggtgaccggccaggaaacggtcg
cccagatcaaggctcatgtagcctcactggagggcatt
```

xyz: ID name

• Sequence Database Format: EMBL, GenBank, SwissProt, PIR;

• Sequence File: Files can hold sequences in standard recognized formats;

Sequence Formats

• Currently input/output supported formats (more than 42):
http://emboss.sourceforge.net/docs/themes/SequenceFormats.html

• Input Sequence Formats: fasta, EMBL (embl/em), Swiss-Prot (swissprot/swiss/sw), GCG (gcg), MSF (msf), Genbank (genbank), raw,…

• Output Sequence Formats: embl, gcg, swiss, CLUSTALW (clustal, aln), genbank, …

Multiple sequences, single file

- EMBOSS writes many sequences to a single file. Most sequence formats can deal with this:
  Fasta, EMBL, PIR, MSF, Clustal, Phylip, etc.
  BUT NOT: Plain, Staden and GCG
- EMBOSS reads many sequences from a single file.
  Use `filename:entryname` if you wish to specify a single sequence.
  If there is only one sequence, or you wish to read all entries, use just the filename.
- The program ***seqretsplit*** will split an existing multiple sequence file into many files.

Alignment output Formats

- Several formats have been written or adopted for EMBOSS output: http://emboss.sourceforge.net/docs/themes/AlignFormats.html
  - Multiple sequence alignment: fasta, msf,..
  - Pair-wise alignment: pair, score,…
- Each program that writes an alignment has a default alignment format defined for that program. However you can change the output formats from the output file format menu

Feature Formats

• A feature is a region of interest in a specified nucleic or protein sequence. It has a specified start and end position. It has a name describing what type of thing it is:
Ex: Swiss-Prot Feature table

```
FT DISULFID 3 40
FT DISULFID 4 32
FT DISULFID 16 26
FT VARIANT 22 22 P -> S (IN ISOFORM SI).
FT VARIANT 25 25 L -> I (IN ISOFORM SI).
```

• When reading or writing features associated with a sequence, there are a standard set of formats that are used: UFO (Universal Feature Object) e.g. Swiss-Prot (swissprot), EMBL (embl), PIR (pir),…
http://emboss.sourceforge.net/docs/themes/FeatureFormats.html

• ***showfeat*** useful for displaying features.
• ***extractfeat*** useful for extracting the sequences of features.

Feature: Example

Example: PAX4_HUMAN

```
FT   CHAIN         1     350       Paired box protein Pax-4.
FT                                 /FTId=PRO_0000050180.
FT   DOMAIN        5     131       Paired.
FT   DNA_BIND    170     229       Homeobox.
FT   REGION      278     350       Transcription repression.
FT   VARSPLIC    239     257       Missing (in isoform Pax4V).
FT                                 /FTId=VSP_002359.
FT   VARSPLIC    258     350       QSPGSVPTAALPALEPLGPSCYQLCWATAPERCLSDTPPKA
FT                                 CLKPCWDCGSFLLPVIAPSCVDVAWPCLDASLAHHLIGGAG
FT                                 KATPTHFSHWP -> AVPWQCAHSSPACPGTTGSLLLSAVL
FT                                 GNSTRKVSE (in isoform Pax4V).
FT                                 /FTId=VSP_002360.
FT   VARSPLIC    305     350       DCGSFLLPVIAPSCVDVAWPCLDASLAHHLIGGAGKATPTH
FT                                 FSHWP -> GHLPPQPNSLDSGLLCLPCPSSHCPLASLSGS
FT                                 QALLWPGCPLLYGLE (in isoform 3).
FT                                 /FTId=VSP_012925.
```

Feature: Example

wSHOWFEAT Output

Output file:
/import/bc2/home2/schwede/bordoli/wProjects/Alignment/.showfeat.06.02.16:16.07.44/pax4_human.showfeat [right click to save]

```
PAX4_HUMAN
Paired box protein Pax-4.
|=========================================================| 350
|---------------------------------------------------------| chain
|--------------------|                                      domain
                          |---------|                       dna_bind
                                    |---|                   varsplic
                                        |---------------|   varsplic
                                           |------------|   site
                                                 |-------|  varsplic
```

Report Formats

- There are many ways in which the results of an analysis can be reported: http://emboss.sourceforge.net/docs/themes/ReportFormats.html
- ***garnier*** predicts protein secondary structure.

```
########################################
# Program: garnier
# Rundate: Thu Feb 16 2006 16:53:12
# Report_format: tagseq
# Report_file: pax4_human.garnier
########################################

#=======================================
#
# Sequence: PAX4_HUMAN     from: 1    to: 350
# HitCount: 114
#
# DCH = 0, DCS = 0
#
#  Please cite:
#  Garnier, Osguthorpe and Robson (1978) J. Mol. Biol. 120:97-120
#
#
#=======================================

          .   10    .   20    .   30    .   40    .   50
      MHQDGISSMNQLGGLFVNGRPLPLDTRQQIVRLAVSGMRPCDISRILKVS
helix HH
sheet         EE   EEEE         EEEEEEEE   EE   EEEEEE
turns    TT       T    T     TTT        TTT  TTT      TT
 coil   C  CCC  CC      CCCCC
```

Report Formats

- Many EMBOSS programs are now able to output their results in a standard report format - you can change the report format used from the report format output menu

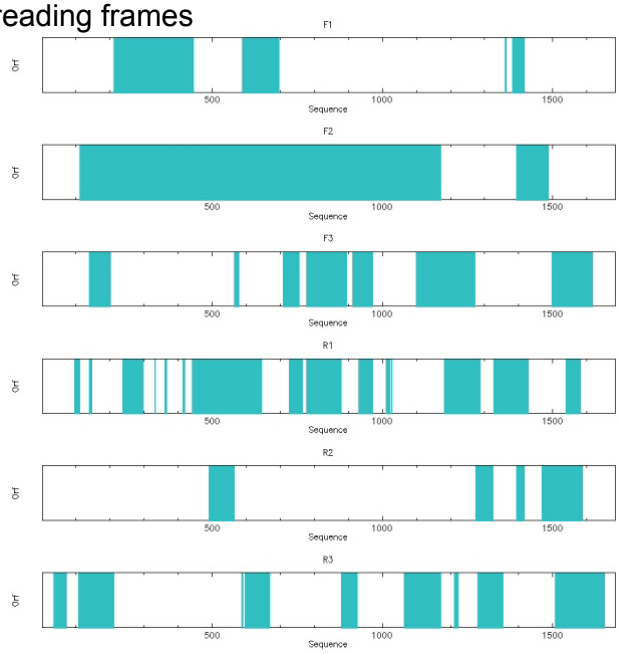
- examples:

embl	Writes a report in EMBL feature table format
pir	Writes a report in PIR feature table format
swiss	Writes a report in SwissProt feature table format
excel	This is a TAB-delimited table format suitable for reading into spread-sheet programs such as Excel.
seqtable	A simple table format that includes the feature sequence

```
 Start        End          [tagnames]    Sequence
[start]       [end]        [tagvalues] [sequence]
```

Graphic Formats

- Graphic format: PNG, ps (postscript)
- ***plotorf*** plot potential open reading frames

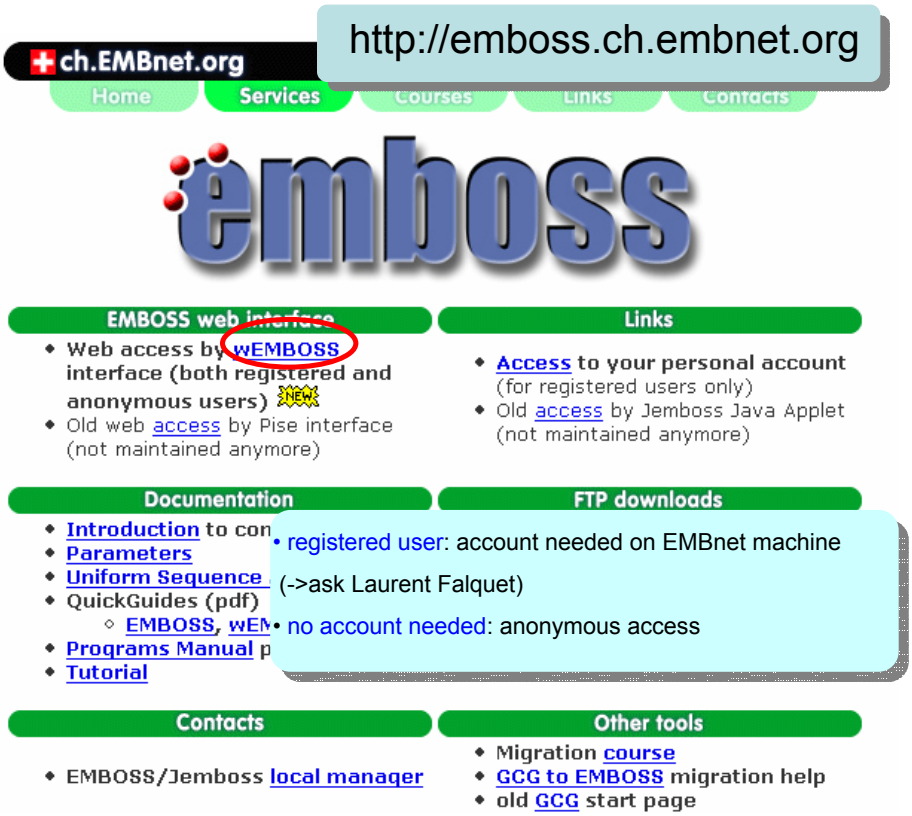


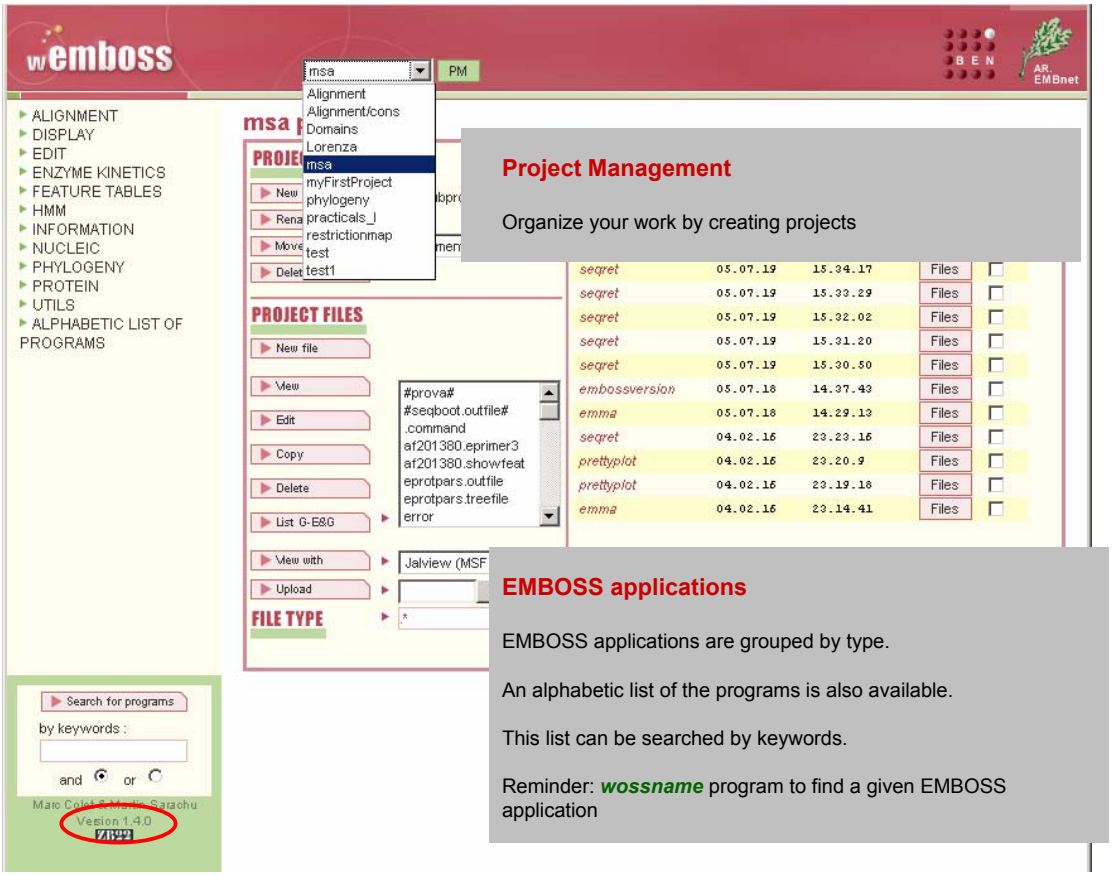
Some useful programs to start with:

- ***wossname***: Finds programs by keywords in their one-line documentation;
- ***showdb***: Displays information on the currently available databases;
- ***seqret***: Reads and writes (returns) sequences: retrieve and reformat sequences;

wEMBOSS

- http://www.wemboss.org/
- web interface to EMBOSS (current version: 1.5.0)
- Each user has a separate and private workspace.
- Organize your work by creating projects and subprojects.
- Results saved for easy recover & review.
- Authors: Marc Colet, Martin Sarachu
- wEMBOSS is a joint effort between Argentinian EMBnet Node and the Belgian EMBnet node





Project Management

Organize your work by creating projects

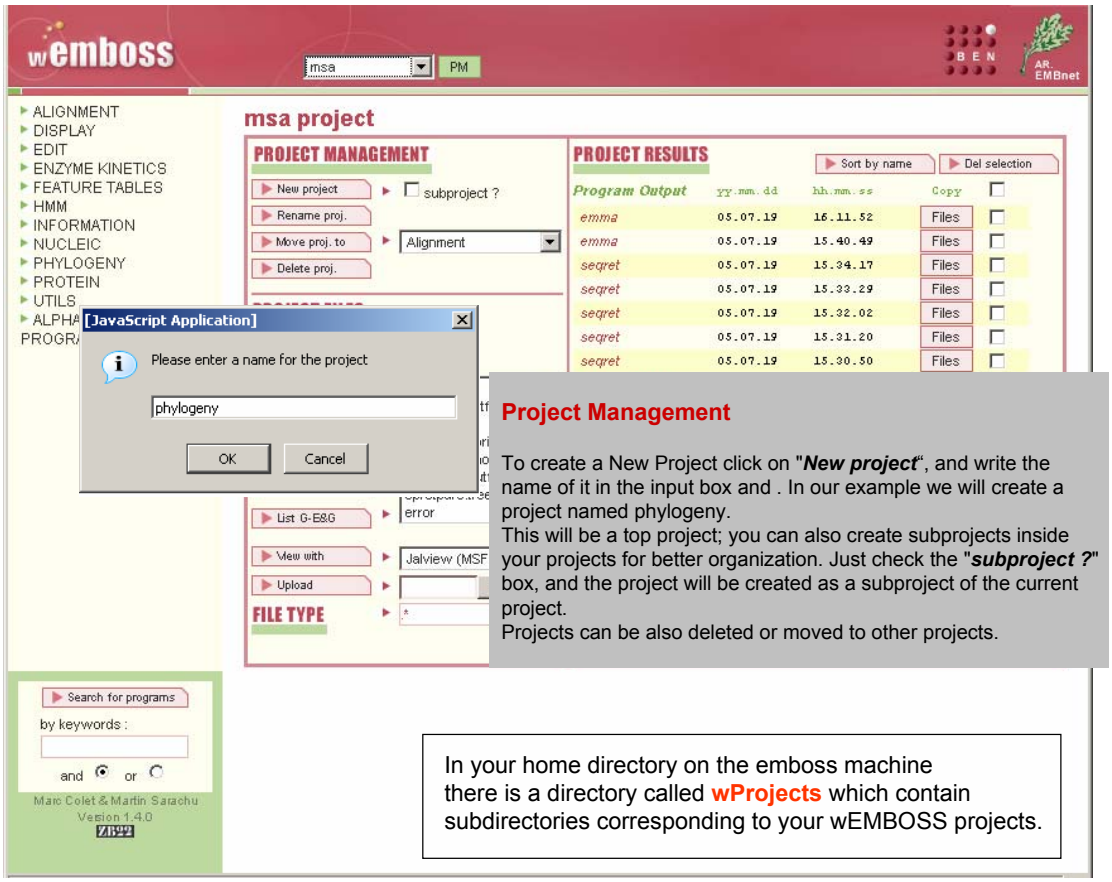
EMBOSS applications

EMBOSS applications are grouped by type.

An alphabetic list of the programs is also available.

This list can be searched by keywords.

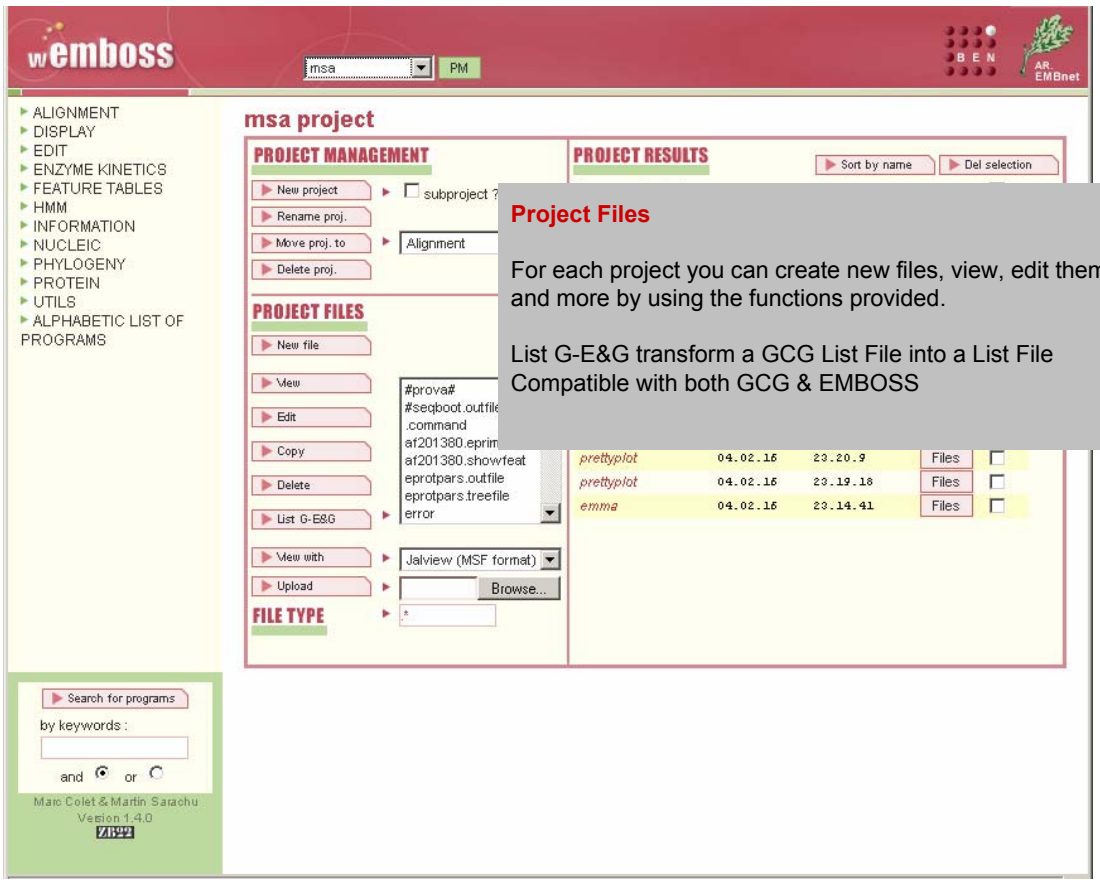
Reminder: ***wossname*** program to find a given EMBOSS application



Project Management

To create a New Project click on "***New project***", and write the name of it in the input box and . In our example we will create a project named phylogeny.
This will be a top project; you can also create subprojects inside your projects for better organization. Just check the "***subproject ?***" box, and the project will be created as a subproject of the current project.
Projects can be also deleted or moved to other projects.

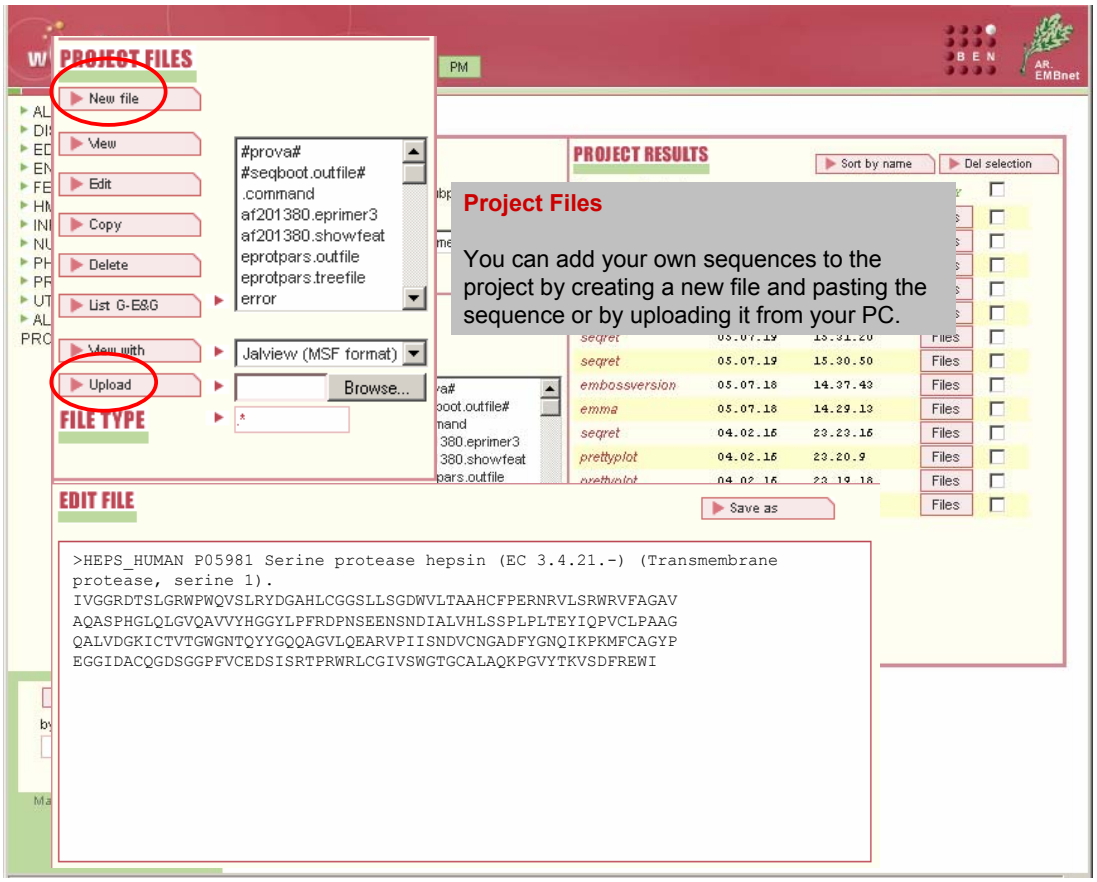
In your home directory on the emboss machine there is a directory called **wProjects** which contain subdirectories corresponding to your wEMBOSS projects.



Project Files

For each project you can create new files, view, edit them, and more by using the functions provided.

List G-E&G transform a GCG List File into a List File Compatible with both GCG & EMBOSS



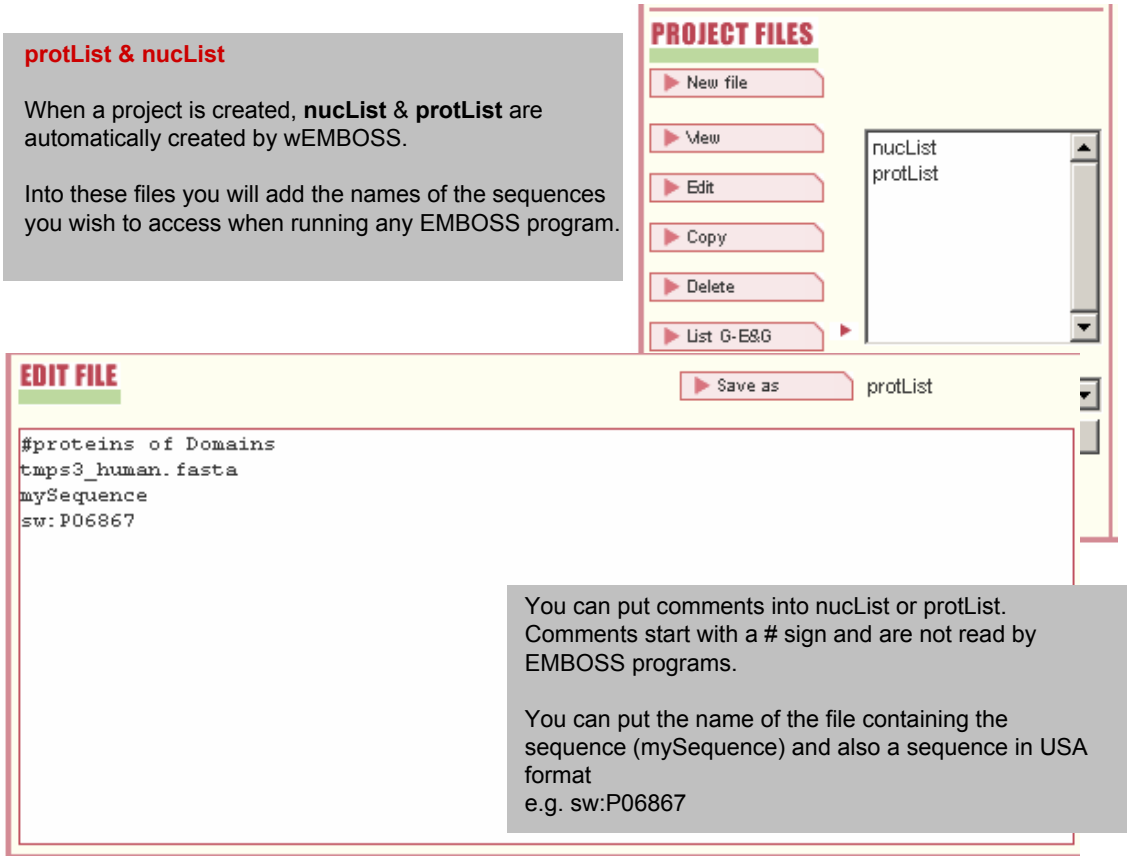
Project Files

You can add your own sequences to the project by creating a new file and pasting the sequence or by uploading it from your PC.

protList & nucList

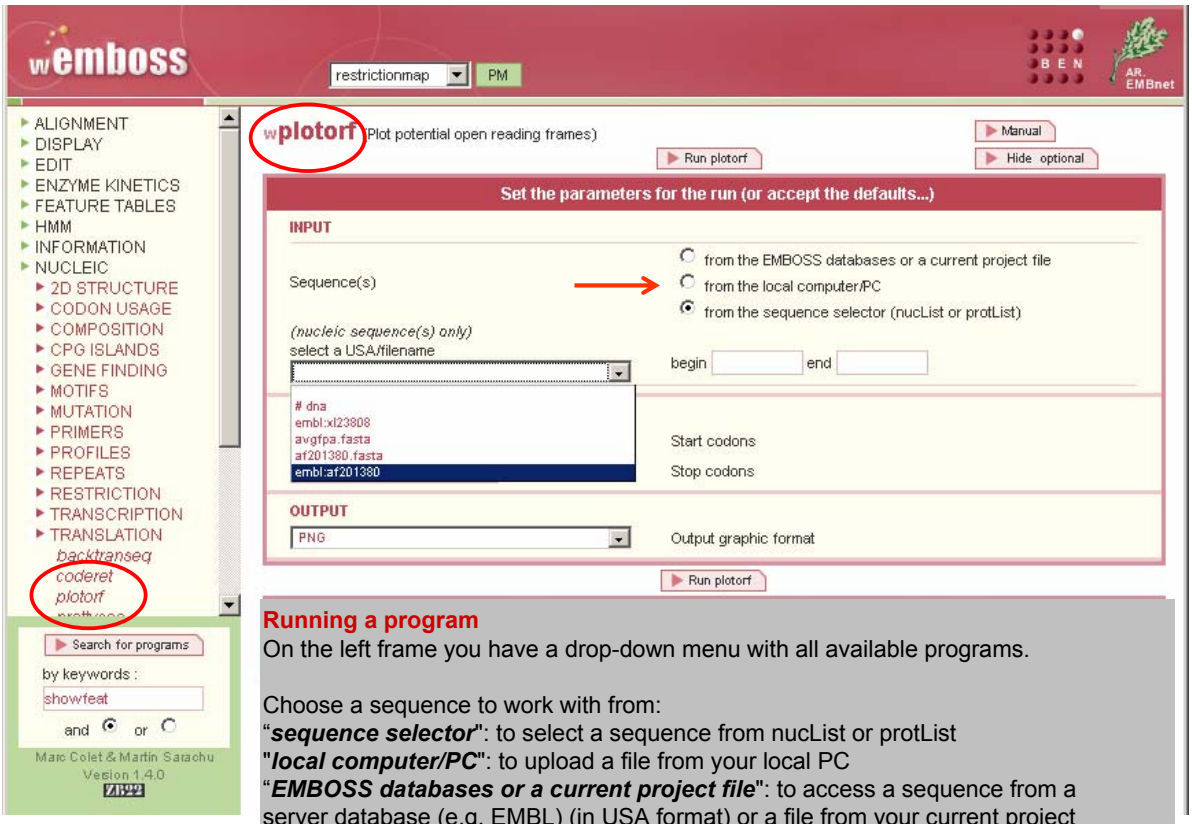
When a project is created, **nucList** & **protList** are automatically created by wEMBOSS.

Into these files you will add the names of the sequences you wish to access when running any EMBOSS program.



You can put comments into nucList or protList. Comments start with a # sign and are not read by EMBOSS programs.

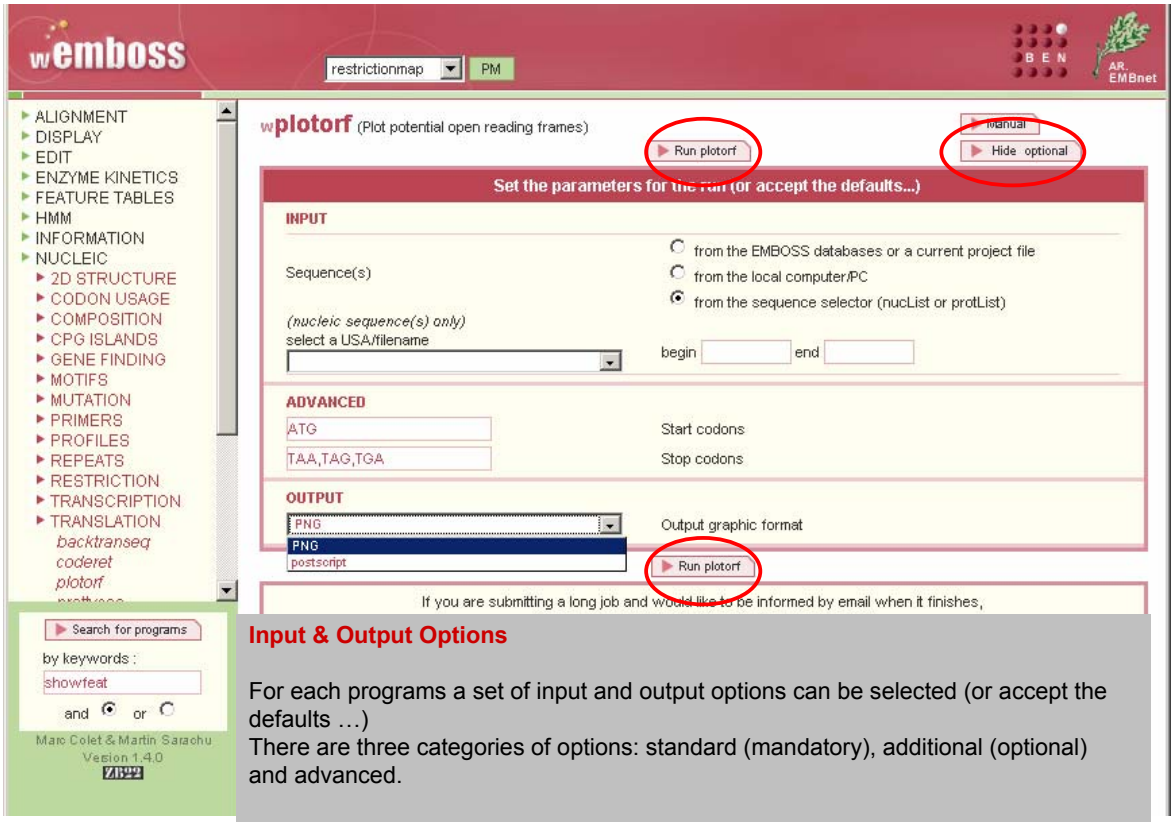
You can put the name of the file containing the sequence (mySequence) and also a sequence in USA format
e.g. sw:P06867



Running a program

On the left frame you have a drop-down menu with all available programs.

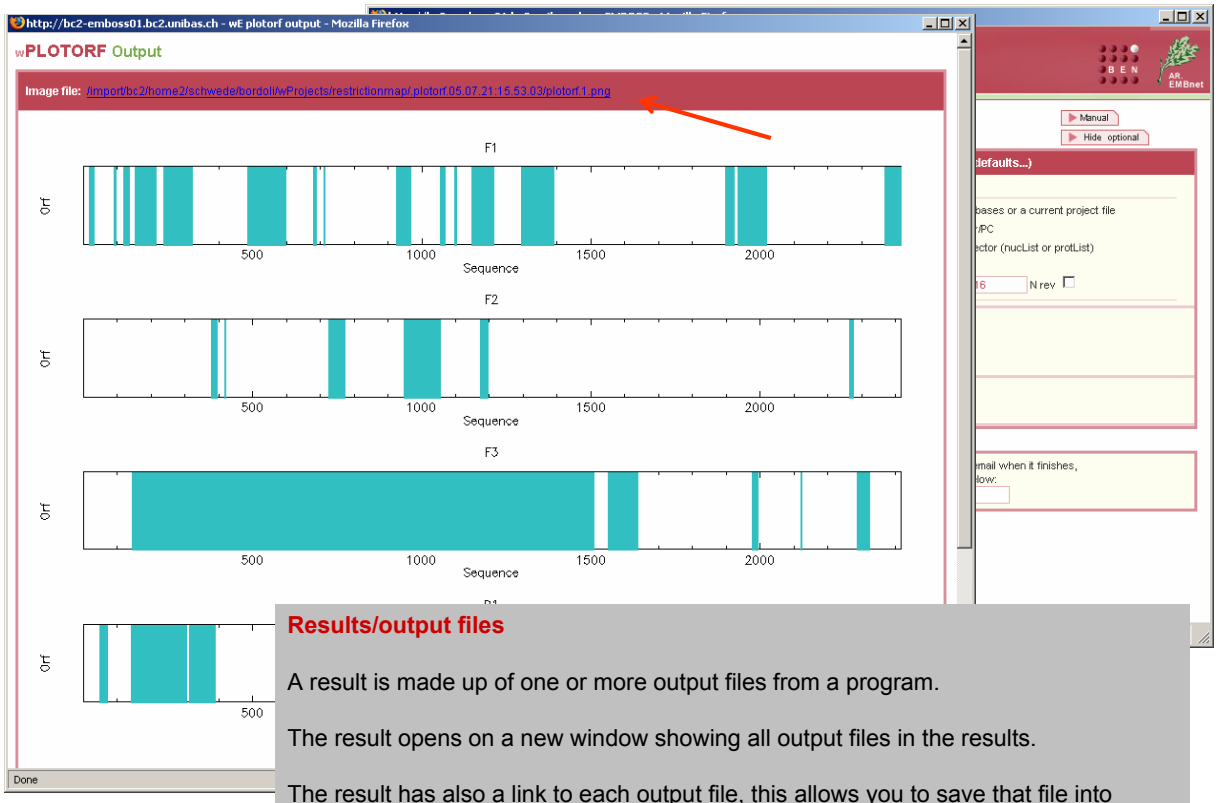
Choose a sequence to work with from:
"***sequence selector***": to select a sequence from nucList or protList
"***local computer/PC***": to upload a file from your local PC
"***EMBOSS databases or a current project file***": to access a sequence from a server database (e.g. EMBL) (in USA format) or a file from your current project



Input & Output Options

For each programs a set of input and output options can be selected (or accept the defaults …)
There are three categories of options: standard (mandatory), additional (optional) and advanced.

And then you can run the program

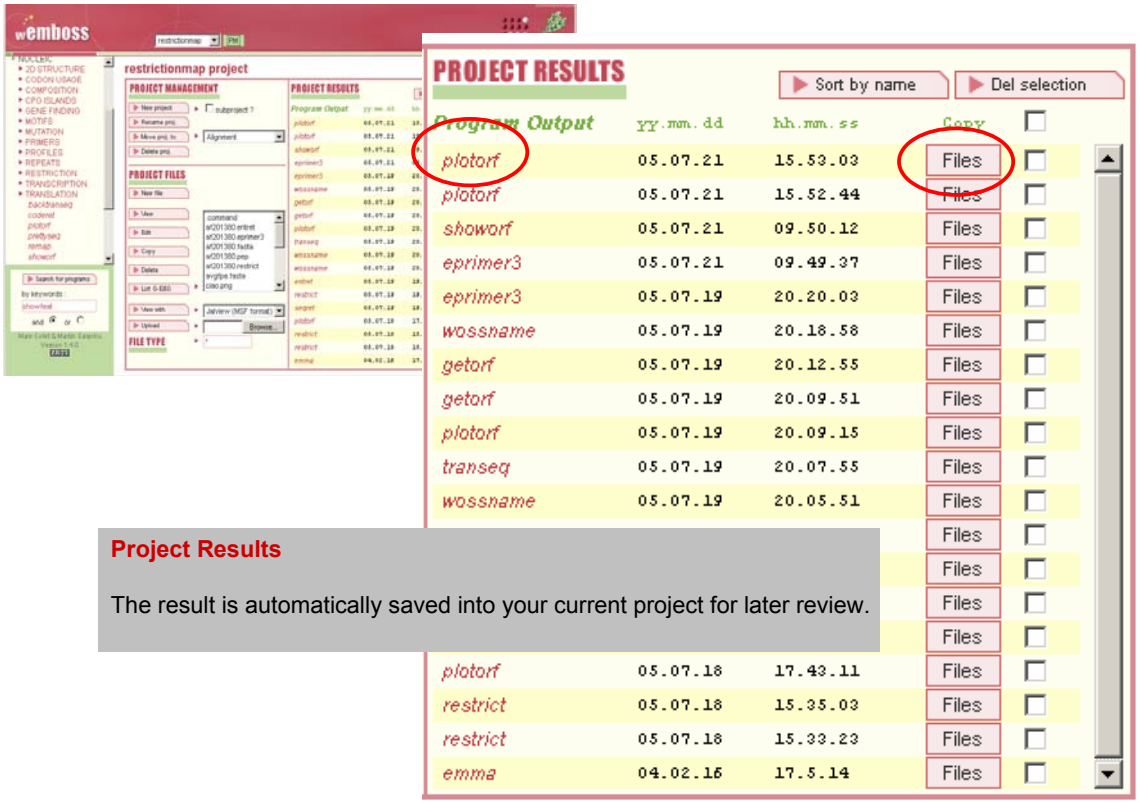


Results/output files

A result is made up of one or more output files from a program.

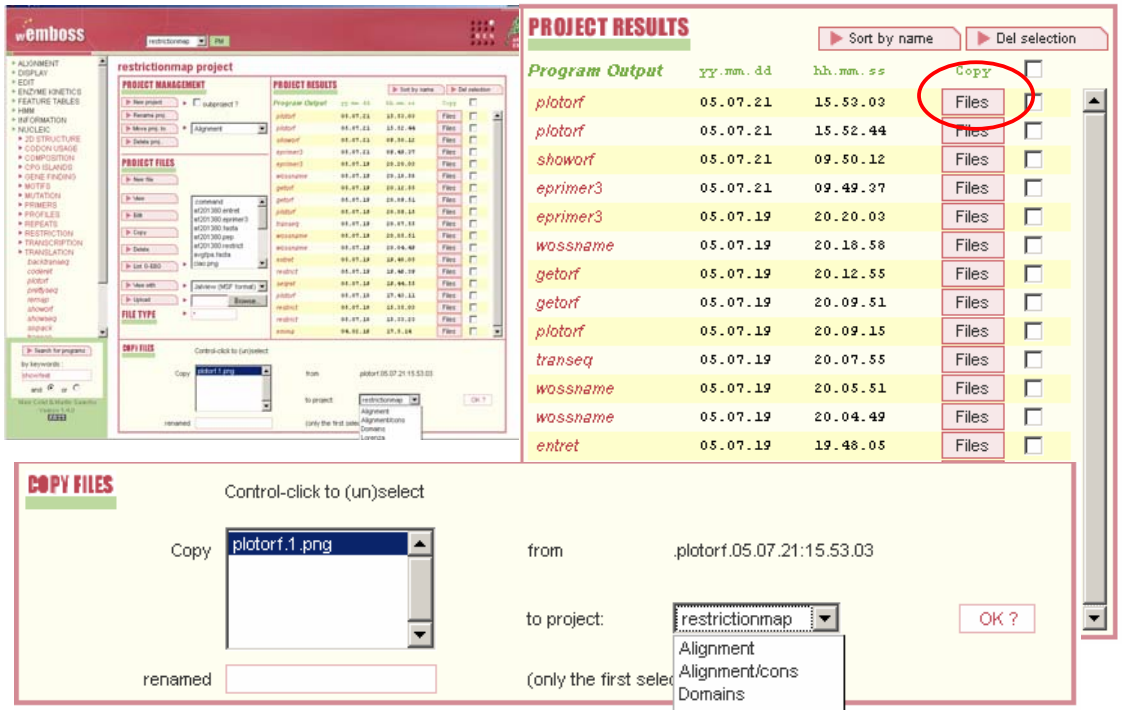
The result opens on a new window showing all output files in the results.

The result has also a link to each output file, this allows you to save that file into your local computer.



Project Results

The result is automatically saved into your current project for later review.

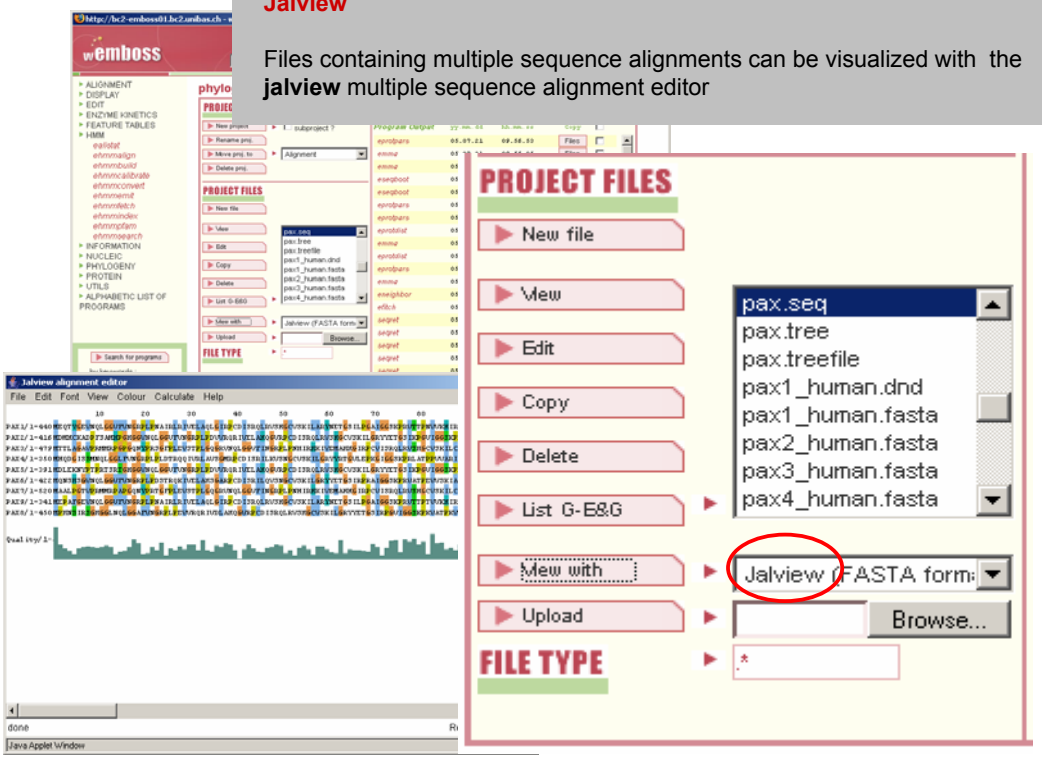


Project Results

The result file(s) can be copied into the list of files of the current or of other Projects

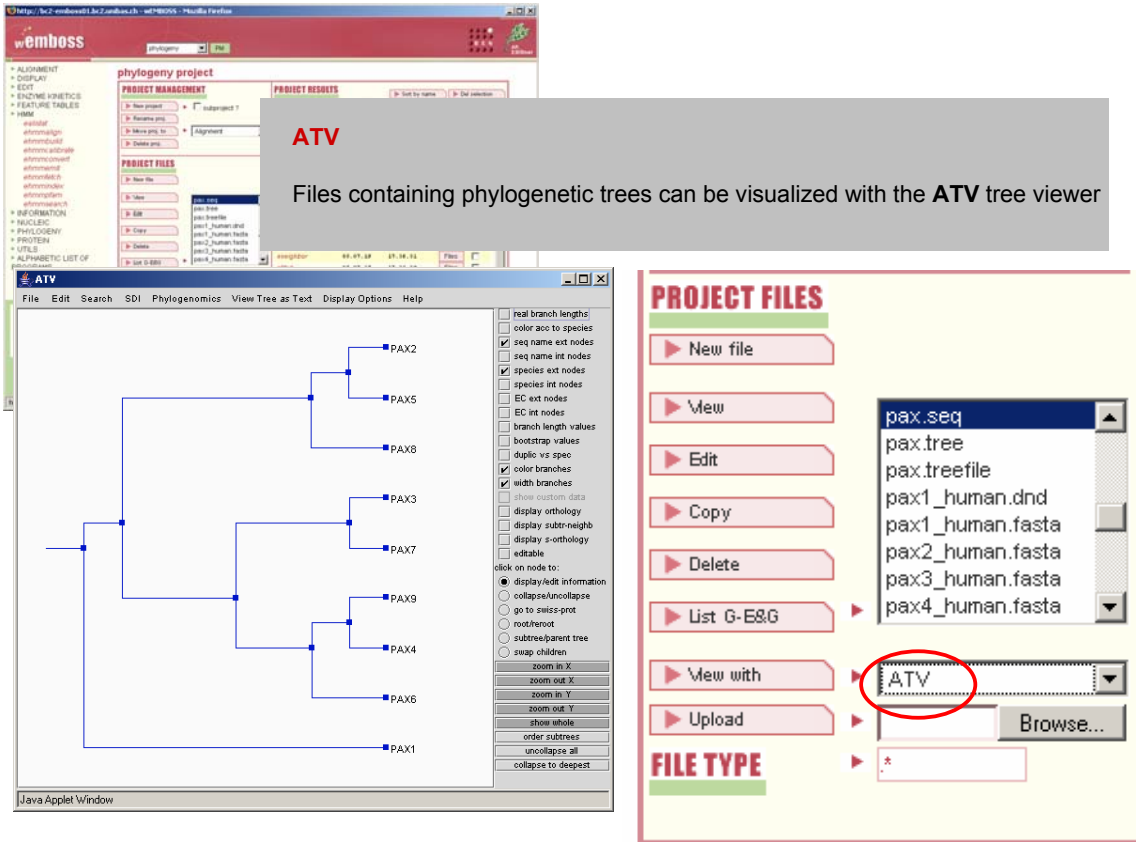
Jalview

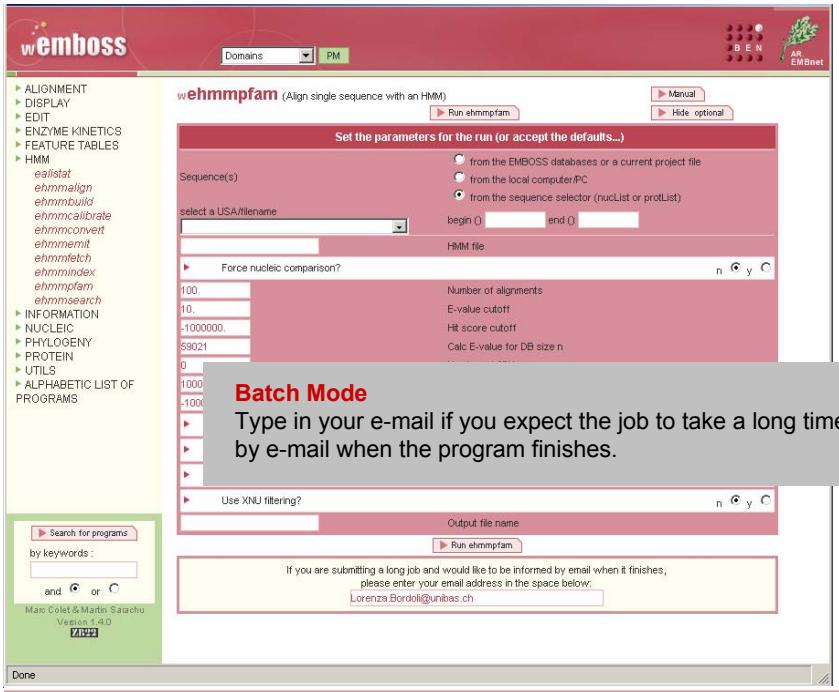
Files containing multiple sequence alignments can be visualized with the **jalview** multiple sequence alignment editor



ATV

Files containing phylogenetic trees can be visualized with the **ATV** tree viewer

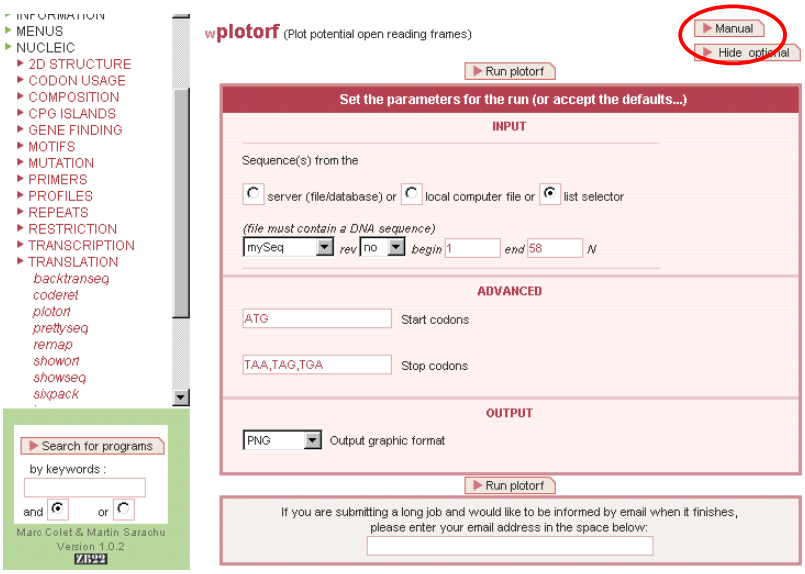




Batch Mode

Type in your e-mail if you expect the job to take a long time. You will be warned by e-mail when the program finishes.

Help on a program!



What EMBOSS does NOT

• The major deficiencies in the EMBOSS package are:
BLAST, FASTA, ASSEMBLY (GelMerge, GelEnter,…) , PAUP, sequence editor

You should use the publicly available software:

- Blast - NCBI, HGMP, many other sites
- Fasta – HGMP
- Assembly - Staden package
- PHRED, PHRAP
- (PAUP, package not free)
- sequence editor: pico, emacs, vi

What EMBOSS does NOT

• The major deficiencies in the EMBOSS package are:
BLAST, FASTA, ASSEMBLY (GelMerge, GelEnter,…) , PAUP, sequence editor

• Graphical Interface:

- BLAST:
  - SIB: http://www.expasy.org/tools/blast/
  - Swiss EMBnet: http://www.ch.embnet.org/software/BottomBLAST.html?
  - NCBI: http://www.ncbi.nlm.nih.gov/BLAST/
- FASTA:
  EBI: http://www.ebi.ac.uk/fasta33/
- ClustalW:
  - Swiss EMBnet: http://www.ch.embnet.org/software/ClustalW.html
- PAUP no graphical interface, use Phylip instead (part of EMBOSS)

References

- http://emboss.sourceforge.net/
- UK HGMP Resource Centre, Userguide, 2002
- wEMBOSS: http://www.wemboss.org/