



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 Accelrys;
- 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 *European Molecular Biology Open Source Software Suite*, 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.



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

Remote server:
ludwig-sun1.unil.ch



Remote server:
you personal
account



Running EMBOSS Programs

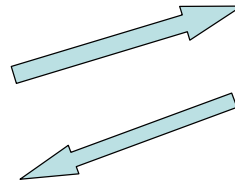
EMBOSS programs are run by:

- typing them at the UNIX prompt: X-terminal (X window system)

Remote server:
ludwig-sun1.unil.ch



Local computer



Running EMBOSS Programs

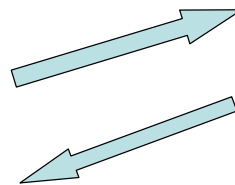
EMBOSS programs are run by:

- 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:**

wosname 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

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

• Sequence parameters

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



Some major programs (cont.):

• DNA Sequence features

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



Some major programs (*cont.*):

- **Protein Sequence features**

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



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	<code>embl:X65923</code>
database:ID	<code>swissprot:opsd_xenla</code>
file name	<code>myfile.seq</code>



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:

<code>filename</code>	all sequences in a file
<code>filename:entry</code>	an entry in a file
<code>dbname</code>	all sequences in a database (not recommended)
<code>dbname:entry</code>	a sequence in a database
<code>@listfile</code>	a list file
<code>list::listfile</code>	a list file



Specifying a List File

- Instead of containing the sequences themselves, a `listfile` 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

<code>filename</code>	: a file containing one or more sequences
<code>filename:entry</code> <code>mysequences:opsd_xenla</code>	: a given sequence in the file. The 'entry' is the ID or AC of the sequence in that file
<code>filename:entry[start:end]</code> <code>mysequences:opsd_xenla[1:20]</code> <code>mysequences:opsd_xenla[-1:-20]</code>	: a part of the sequence can be specified by the range
<code>mysequences:[1:20:r]</code>	: 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
ttcctctttctcgactccatcttcgcggtagctgggaccgcccgttcagtcgccaatatgc
agctctttgtccgcccagagctacacaccttcgaggtgaccggccaggaaacggtcg
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 *secretsplit* 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                                     KATPTHFSHP -> AVPWQCAHSSPACPGTTGSLLSAVL
FT                                     GNSTRKVSE (in isoform Pax4V).
FT                                     /FTid=VSP_002360.
FT   VARSPLIC       305  350   DCGSFLLPVIAPSCVDVAWPCLDASLAHHLIGGAGKATPTH
FT                                     FSHWP -> GHLPPQPNLSDSGLLCLPCSSHCLASLSGS
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://embooss.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  EEEEE
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:

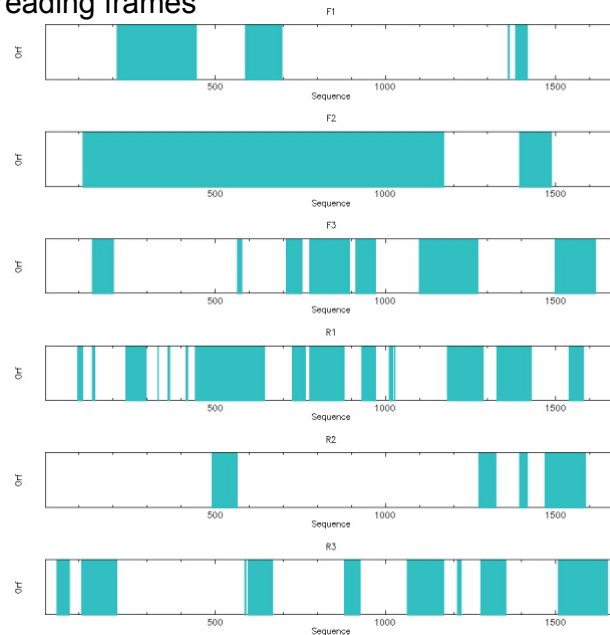
- [embl](#) Writes a report in EMBL feature table format
- [pir](#) Writes a report in PIR feature table format
- [swiss](#) Writes a report in SwissProt feature table format
- [excel](#) This is a TAB-delimited table format suitable for reading into spread-sheet programs such as Excel.
- [seqtable](#) 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

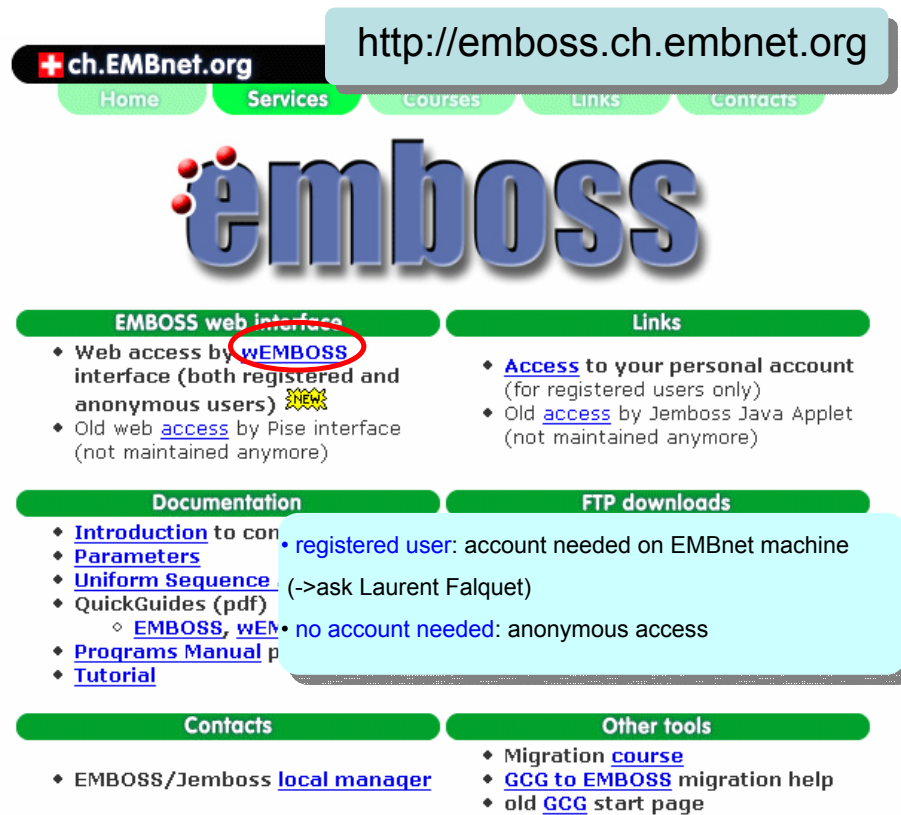


Some useful programs to start with:

- *wosname*: Finds programs by keywords in their one-line documentation;
- *showdb*: Displays information on the currently available databases;
- *secret*: 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



The screenshot shows the website <http://emboss.ch.embnet.org>. The navigation bar includes Home, Services, Courses, Links, and Contacts. The main heading is "emboss".

EMBOSS web interface

- ♦ Web access by **wEMBOSS** interface (both registered and anonymous users) 
- ♦ Old web [access](#) by Pise interface (not maintained anymore)

Links

- ♦ [Access](#) to your personal account (for registered users only)
- ♦ Old [access](#) by Jemboss Java Applet (not maintained anymore)

Documentation

- ♦ [Introduction](#) to con
- ♦ [Parameters](#)
- ♦ [Uniform Sequence QuickGuides](#) (pdf) (->ask Laurent Falquet)
- ♦ [EMBOSS, wEMBOSS, Jemboss](#)
- ♦ [Programs Manual](#) p
- ♦ [Tutorial](#)

FTP downloads

- ♦ **registered user:** account needed on EMBnet machine
- ♦ **no account needed:** anonymous access

Contacts

- ♦ EMBOSS/Jemboss [local manager](#)

Other tools

- ♦ Migration [course](#)
- ♦ [GCG to EMBOSS](#) migration help
- ♦ old [GCG](#) start page

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: **wosname** program to find a given EMBOSS application

Program	yy.mm.dd	hh.mm.ss	Files	Del selection
secret	05.07.19	15.34.17	Files	<input type="checkbox"/>
secret	05.07.19	15.33.29	Files	<input type="checkbox"/>
secret	05.07.19	15.32.02	Files	<input type="checkbox"/>
secret	05.07.19	15.31.20	Files	<input type="checkbox"/>
secret	05.07.19	15.30.50	Files	<input type="checkbox"/>
embossversion	05.07.18	14.37.43	Files	<input type="checkbox"/>
emma	05.07.18	14.29.13	Files	<input type="checkbox"/>
secret	04.02.16	23.23.16	Files	<input type="checkbox"/>
prettyplot	04.02.16	23.20.9	Files	<input type="checkbox"/>
prettyplot	04.02.16	23.19.18	Files	<input type="checkbox"/>
emma	04.02.16	23.14.41	Files	<input type="checkbox"/>

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.

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

wemboss B E N AR. EMBnet

msa PM

- ▶ ALIGNMENT
- ▶ DISPLAY
- ▶ EDIT
- ▶ ENZYME KINETICS
- ▶ FEATURE TABLES
- ▶ HMM
- ▶ INFORMATION
- ▶ NUCLEIC
- ▶ PHYLOGENY
- ▶ PROTEIN
- ▶ UTILS
- ▶ ALPHABETIC LIST OF PROGRAMS

msa project

PROJECT MANAGEMENT

▶ New project subproject?

▶ Rename proj.

▶ Move proj. to Alignment

▶ Delete proj.

PROJECT RESULTS

Sort by name Del selection

PROJECT FILES

▶ New file

▶ View

▶ Edit

▶ Copy

▶ Delete

▶ List G-E&G

▶ View with Jalview (MSF format)

▶ Upload Browse...

FILE TYPE

*.

Search for programs

by keywords:

and or

Maio Colet & Martin Sarachu
Version 1.4.0
2002

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 & EMBOS

prettyplot	04.02.16	23.20.9	Files	<input type="checkbox"/>
prettyplot	04.02.16	23.19.18	Files	<input type="checkbox"/>
emma	04.02.16	23.14.41	Files	<input type="checkbox"/>

wemboss B E N AR. EMBnet

PM

- ▶ ALIGNMENT
- ▶ DISPLAY
- ▶ EDIT
- ▶ ENZYME KINETICS
- ▶ FEATURE TABLES
- ▶ HMM
- ▶ INFORMATION
- ▶ NUCLEIC
- ▶ PHYLOGENY
- ▶ PROTEIN
- ▶ UTILS
- ▶ ALPHABETIC LIST OF PROGRAMS

PROJECT FILES

▶ New file

▶ View

▶ Edit

▶ Copy

▶ Delete

▶ List G-E&G

▶ View with Jalview (MSF format)

▶ Upload Browse...

PROJECT RESULTS

Sort by name Del selection

FILE TYPE

*.

Search for programs

by keywords:

and or

Maio Colet & Martin Sarachu
Version 1.4.0
2002

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.

EDIT FILE

Save as

```

>HEPS_HUMAN P05981 Serine protease hepsin (EC 3.4.21.-) (Transmembrane
protease, serine 1).
IVGGRDTSLGRWPQVSLRYDGAHLCCGSLLSGDWVLTAAHCPPERNRVLSRWRVFAGAV
AQASPHGLQLGVQAVVYHGGYLPFRDPNSEENSNDIALVHLSSPLPLTEYIQPVCLPAAG
QALVDGKICTVTGWGNTQYYGQAGVLQEARVPIISNDVCNGADFYGNQIKPKMFCAGYP
EGGIDACQGDSSGPFVCEDSISRTPRWRLCGIVSWGTCALAKPKGVYTKVSDFREWI

```

seqret	05.07.19	15.21.20	Files	<input type="checkbox"/>
seqret	05.07.19	15.20.50	Files	<input type="checkbox"/>
embossversion	05.07.18	14.37.43	Files	<input type="checkbox"/>
emma	05.07.18	14.29.13	Files	<input type="checkbox"/>
seqret	04.02.16	23.23.16	Files	<input type="checkbox"/>
prettyplot	04.02.16	23.20.9	Files	<input type="checkbox"/>
prettyplot	04.02.16	23.19.18	Files	<input type="checkbox"/>
seqret	04.02.16	23.19.18	Files	<input type="checkbox"/>

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.

PROJECT FILES

- New file
- New
- Edit
- Copy
- Delete
- List G-E&G

```
nuclist
protList
```

EDIT FILE

Save as protList

```
#proteins of Domains
tmps3_human.fasta
mySequence
sw:P06867
```

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

The screenshot shows the wEMBOSS web interface. The top navigation bar includes the wEMBOSS logo, a 'restrictionmap' dropdown, a 'PM' button, and a 'B E N' logo with 'AR. EMBnet' text. A left sidebar menu lists various programs, with 'wplotorf' circled in red. Below the menu is a search box and a 'Search for programs' button. The main content area is titled 'wplotorf Plot potential open reading frames)'. It features a 'Run plotorf' button and a 'Manual' button. Below this is a form titled 'Set the parameters for the run (or accept the defaults...)'. The 'INPUT' section has three radio buttons: 'from the EMBOSS databases or a current project file', 'from the local computer/PC', and 'from the sequence selector (nuclist or protList)'. An orange arrow points to the third option. Below the radio buttons is a text input field for 'select a USA filename' with a dropdown arrow, and 'begin' and 'end' input fields. The 'OUTPUT' section has a dropdown menu set to 'PNG' and an 'Output graphic format' label. At the bottom of the form is another 'Run plotorf' button. A grey box at the bottom of the screenshot contains the text: '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'

wplotorf (Plot potential open reading frames)

Set the parameters for the run (or accept the defaults...)

INPUT

Sequence(s) from the EMBOSS databases or a current project file
 from the local computer/PC
 from the sequence selector (nuclist or protList)

(nucleic sequence(s) only)
 select a USA/filename begin end

ADVANCED

ATG Start codons
 TAA,TAG,TGA Stop codons

OUTPUT

Output graphic format

If you are submitting a long job and would like to be informed by email when it finishes,

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

wPLOTORF Output

Image file: http://imgprofs2/home2/schwede/hordollw/Projects/restrictionmap/plotorf/05.07.21.15.53.03/plotorf_1.png

F1
 0rf
 500 1000 1500 2000
 Sequence

F2
 0rf
 500 1000 1500 2000
 Sequence

F3
 0rf
 500 1000 1500 2000
 Sequence

F4
 0rf
 500
 Sequence

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

Sort by name Del selection

Program Output	yy.mm.dd	hh.mm.ss	Copy
plotorf	05.07.21	15.53.03	Files
plotorf	05.07.21	15.52.44	Files
showorf	05.07.21	09.50.12	Files
eprimer3	05.07.21	09.49.37	Files
eprimer3	05.07.19	20.20.03	Files
wosname	05.07.19	20.18.58	Files
getorf	05.07.19	20.12.55	Files
getorf	05.07.19	20.09.51	Files
plotorf	05.07.19	20.09.15	Files
transeq	05.07.19	20.07.55	Files
wosname	05.07.19	20.05.51	Files
plotorf	05.07.18	17.43.11	Files
restrict	05.07.18	15.35.03	Files
restrict	05.07.18	15.33.23	Files
emma	04.02.16	17.5.14	Files

Project Results

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

PROJECT RESULTS

Sort by name Del selection

Program Output	yy.mm.dd	hh.mm.ss	Copy
plotorf	05.07.21	15.53.03	Files
plotorf	05.07.21	15.52.44	Files
showorf	05.07.21	09.50.12	Files
eprimer3	05.07.21	09.49.37	Files
eprimer3	05.07.19	20.20.03	Files
wosname	05.07.19	20.18.58	Files
getorf	05.07.19	20.12.55	Files
getorf	05.07.19	20.09.51	Files
plotorf	05.07.19	20.09.15	Files
transeq	05.07.19	20.07.55	Files
wosname	05.07.19	20.05.51	Files
wosname	05.07.19	20.04.49	Files
entret	05.07.19	19.48.05	Files

COPY FILES

Control-click to (un)select

Copy plotorf.1.png from plotorf.05.07.21:15.53.03

renamed [] to project: restrictionmap (only the first selected)

Alignment Alignment/cons Domains

OK ?

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

The screenshot shows the weMBOSS interface with the Jalview alignment editor open. The editor displays a multiple sequence alignment of PAX genes from various species, with a color scale at the bottom. To the right, a 'PROJECT FILES' panel lists files like pax.seq, pax.tree, pax.treefile, and pax1_human.fasta. The 'New with' button is highlighted with a red circle, and its dropdown menu is open, showing 'Jalview (FASTA form:)' selected.

ATV

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

The screenshot shows the weMBOSS interface with the ATV tree viewer open. The viewer displays a phylogenetic tree with nodes labeled PAX1 through PAX9. To the right, a 'PROJECT FILES' panel lists files like pax.seq, pax.tree, pax.treefile, and pax1_human.fasta. The 'View with' button is highlighted with a red circle, and its dropdown menu is open, showing 'ATV' selected.

If you are submitting a long job and would like to be informed by email when it finishes, please enter your email address in the space below:

Lorenza.Bordoli@unibas.ch



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/>