



# Introduction to EMBOSS

Etienne de Villiers

# What is EMBOSS!



- A suite of bioinformatics programs.
- Open source software – freely available.
- Public domain (GNU Public Licence).
- Written by HGMP/Sanger/EBI/Norway ... etc.

# What it aims to do



- A useful, integrated set of programs.
- They share a common look and feel.
- Incorporates many small and large programs.
- Easy to run from the command line.
- Easy to call from other programs (e.g. perl).
- Easy to set up behind GUIs and Web interfaces.

# Scope of applications



- There are many EMBOSS programs (150+).

- See:

<http://www.uk.embnet.org/Software/EMBOSS/Apps/>

- Many sequence analysis & display programs.
- Protein 3D structure prediction being developed.
- Other assorted programs, eg: enzyme kinetics.

# Running EMBOSS Programs

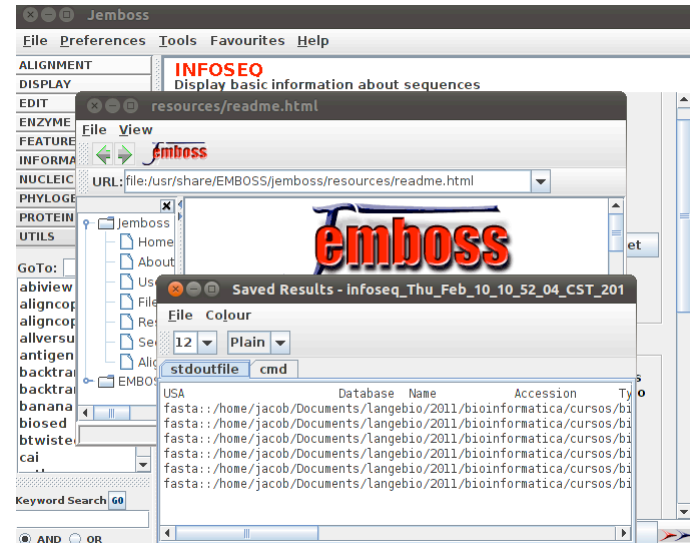
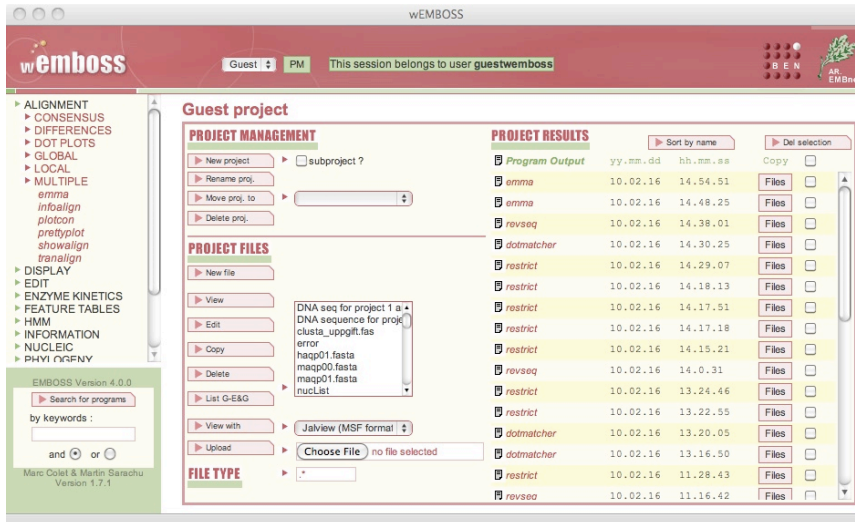


- EMBOSS programs are run by:
  - ◆ Typing them at the UNIX prompt.
  - ◆ Or by using a graphical interface.



# Graphical interfaces to EMBOSS

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



# Some major programmes



- General:
  - ◆ wosname list EMBOSS programs
  - ◆ showdb shows the available databases
- Sequence retrieval
  - ◆ seqret retrieve and/or changes format of sequence
  - ◆ transeq translate DNA sequence to protein
  - ◆ backtranseq translate protein sequence to DNA
  - ◆ extractseq extract region from a sequence
  - ◆ cutseq remove regions from sequence
  - ◆ splitter split a sequence into smaller sequences

# Some major programs (cont)



- Sequence comparison:
  - ◆ needle Needleman-Wunsch sequence alignment
  - ◆ water Smith-Waterman sequence alignment
  - ◆ dotmatcher dotplot comparison of two sequences
  - ◆ prettyplot plots multiple sequence alignments
  - ◆ emma ClustalW program
- Protein sequence features
  - ◆ fuzzpro protein pattern search
  - ◆ helixturnhelix finds nucleic acid binding motifs
  - ◆ pepcoil predict coiled coil regions
  - ◆ pepstats Protein information



# Working with sequences



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

# 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 files as text \*.txt file!).
- The default sequence file format is **fasta**

```
>seq_name
```

```
acgcatgctagcagcagctagctagggcgcgatcgatcatctagctagctcg  
atcgatcgatcgatcgacgatcgatcgatcgatcgatgcatgcatcgatgc  
atgctacgatgatcga
```

- Sequence database formats are **EMBL**, **GenBank**, **SwissProt**, **PIR**

# An example EMBOSS program

- It is easy to forget the name of a program.
- To find EMBOSS programs, use **wosname**.
- **wosname** finds programs by looking for keywords in the description or the name of the program.

# Running at the command-line



- Type **wosname** at the Unix % prompt  
`unix % wosname`
- Displays one-line description.
- Prompts you for information:

```
Finds programs by keywords in their one-line documentation
Keyword to search for: restrict
```

```
SEARCH FOR 'RESTRICT'
```

```
recode           Remove restriction sites but maintain the
                  same translation
```

```
remap           Display a sequence with restriction cut
                  sites, translation
```

```
etc.....
```

# wEMBOSS



- <http://www.wemboss.org>
- Is a web interface to EMBOSS.
- Each user has a separate and private workspace.
- Organize your work by creating projects and subprojects.
- Results are save for easy recovery and review.

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

msa

Alignment

AlignmentIcons

Domains

Lorenza

msa

myFirstProject

phylogeny

practicals\_1

restrictionmap

test

test1

### PROJECT FILES

▶ New file

▶ View

▶ Edit

▶ Copy

▶ Delete

▶ List G-ESG

▶ View with

▶ Upload

### FILE TYPE

#prova#

#seqboot.outfile#

.command

af201380.eprimer3

af201380.showtest

eprotpars.outfile

eprotpars.treefile

error

Jalview (MSF

\*

## Project Management

Organize your work by creating projects

secret	05.07.19	15.04.17	Files	<input type="checkbox"/>
secret	05.07.19	15.03.29	Files	<input type="checkbox"/>
secret	05.07.19	15.02.02	Files	<input type="checkbox"/>
secret	05.07.19	15.01.20	Files	<input type="checkbox"/>
secret	05.07.19	15.00.50	Files	<input type="checkbox"/>
embossversion	05.07.18	14.07.40	Files	<input type="checkbox"/>
emma	05.07.18	14.09.10	Files	<input type="checkbox"/>
secret	04.02.16	20.03.16	Files	<input type="checkbox"/>
prettyplot	04.02.16	20.02.09	Files	<input type="checkbox"/>
prettyplot	04.02.16	20.01.18	Files	<input type="checkbox"/>
emma	04.02.16	20.04.41	Files	<input type="checkbox"/>

## 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

▶ Search for programs

by keywords :

and ☐ or ☐

Mao Collet & Maria Sarachu

Version 1.4.0

17/09/19

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

## msa project

## PROJECT MANAGEMENT

- ▶ New project ▶ ☐ subproject ?
- ▶ Rename proj.
- ▶ Move proj. to ▶ Alignment
- ▶ Delete proj.

## PROJECT RESULTS

▶ Sort by name ▶ Del selection

Program Output	yy.mm.dd	hh.mm.ss	Copy	
emma	05.07.19	16.11.52	Files	<input type="checkbox"/>
emma	05.07.19	15.40.49	Files	<input type="checkbox"/>
seqlot	05.07.19	15.34.17	Files	<input type="checkbox"/>
seqlot	05.07.19	15.33.29	Files	<input type="checkbox"/>
seqlot	05.07.19	15.32.02	Files	<input type="checkbox"/>
seqlot	05.07.19	15.31.20	Files	<input type="checkbox"/>
seqlot	05.07.19	15.30.50	Files	<input type="checkbox"/>

[JavaScript Application]

Please enter a name for the project

phylogeny

OK Cancel

- ▶ Use G-ESS ▶ error
  - ▶ View with ▶ Jalview (MSF)
  - ▶ Upload ▶
- FILE TYPE
- ▶

## 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.

▶ Search for programs

by keywords :

 and ☒ or ☐

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

- ▶ 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
- ▶ Delete proj.

### PROJECT FILES

- ▶ New file
- ▶ View
- ▶ Edit
- ▶ Copy
- ▶ Delete
- ▶ List G-ESG
- ▶ View with
- ▶ Upload
- ▶
- ▶

### FILE TYPE

### PROJECT RESULTS

▶ Sort by name

▶ Del selection

## 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

<i>prettyplot</i>	04.02.16	22.20.9	Files	<input type="checkbox"/>
<i>prettyplot</i>	04.02.16	22.19.18	Files	<input type="checkbox"/>
<i>emma</i>	04.02.16	22.14.41	Files	<input type="checkbox"/>

▶ Search for programs

by keywords :

and ☐ or ☐

Mao Colet & Martin Sarrachu  
Version 1.4.0

2019



## PROJECT FILES

New file

New

Edit

Copy

Delete

List G-ESG

New with

Upload

## FILE TYPE

```
#prova#
#seqboot.outfile#
.command
af201380.eprimer3
af201380.showfeat
eprotpars.outfile
eprotpars.treefile
error
```

Jalview (MSF format)

Browse...

## PROJECT RESULTS

Sort by name

Del selection

### 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.

seqret	05.07.19	15.01.20	Files	<input type="checkbox"/>
seqret	05.07.19	15.00.50	Files	<input type="checkbox"/>
embossversion	05.07.18	14.07.40	Files	<input type="checkbox"/>
emma	05.07.18	14.29.13	Files	<input type="checkbox"/>
seqret	04.02.16	22.22.16	Files	<input type="checkbox"/>
prettyplot	04.02.16	22.20.9	Files	<input type="checkbox"/>
prettyplot	04.02.16	22.19.18	Files	<input type="checkbox"/>
			Files	<input type="checkbox"/>

Save as

## EDIT FILE

```
>HEPS_HUMAN P05981 Serine protease hepsin (EC 3.4.21.-) (Transmembrane
protease, serine 1).
IVGGRDTSLGRWPQVSLRYDGAHLCGGSLLSGDWVLTAAHCFFPERNRVLSRWVVFAGAV
AQASPHGLQLGVQAVVYHGGYLPFRDPNSEENSNDIALVHLSSPLPLTEYIQQPVCLPAAG
QALVDGKICTVTGWGNTQYYGQQAGVLQEARVPIISNDVCNGADFYGNQIKPKMFCAGYP
EGGIDACQGDSCGGPFVCEDSISRTPRWRLCGIVSWGTCALAQKPGVYTKVSDFREWI
```

## 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

▶ View

▶ Edit

▶ Copy

▶ Delete

▶ List G-E&G

nucList  
protList

▶ Save as

protList

## EDIT FILE

```
#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

- ▶ ALIGNMENT
- ▶ DISPLAY
- ▶ EDIT
- ▶ ENZYME KINETICS
- ▶ FEATURE TABLES
- ▶ HMM
- ▶ INFORMATION
- ▶ NUCLEIC
  - ▶ 2D STRUCTURE
  - ▶ CODON USAGE
  - ▶ COMPOSITION
  - ▶ CPG ISLANDS
  - ▶ GENE FINDING
  - ▶ MOTIFS
  - ▶ MUTATION
  - ▶ PRIMERS
  - ▶ PROFILES
  - ▶ REPEATS
  - ▶ RESTRICTION
  - ▶ TRANSCRIPTION
  - ▶ TRANSLATION
- backtranseq
- codonrat
- plotorf
- prottrans

**plotorf**

(Plot potential open reading frames)

Manual

Run plotorf

Hide optional

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

### INPUT

Sequence(s)

- ☐ from the EMBOS 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

# dna  
embl:xl23808  
avgfpa.fasta  
af201380.fasta  
embl:af201380

begin end

Start codons

Stop codons

### OUTPUT

PNG

Output graphic format

Run plotorf

## 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

Search for programs

by keywords :

showfeat

and or

Marc Colet & Martin Sarachu

Version 1.4.0

2002

- ▶ ALIGNMENT
- ▶ DISPLAY
- ▶ EDIT
- ▶ ENZYME KINETICS
- ▶ FEATURE TABLES
- ▶ HMM
- ▶ INFORMATION
- ▶ NUCLEIC
  - ▶ 2D STRUCTURE
  - ▶ CODON USAGE
  - ▶ COMPOSITION
  - ▶ CPG ISLANDS
  - ▶ GENE FINDING
  - ▶ MOTIFS
  - ▶ MUTATION
  - ▶ PRIMERS
  - ▶ PROFILES
  - ▶ REPEATS
  - ▶ RESTRICTION
  - ▶ TRANSCRIPTION
  - ▶ TRANSLATION
- backtranseq
- coderet
- plotorf
- protseq

▶ Search for programs

by keywords :

showfeat

and ☒ or ☐

Mac Coklet & Martin Sarrachu

Version 1.4.0

2002

wplotorf (Plot potential open reading frames)

▶ Run plotorf

▶ Manual

▶ Hide optional

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

PNG

Output graphic format

PNG

postscript

▶ Run plotorf

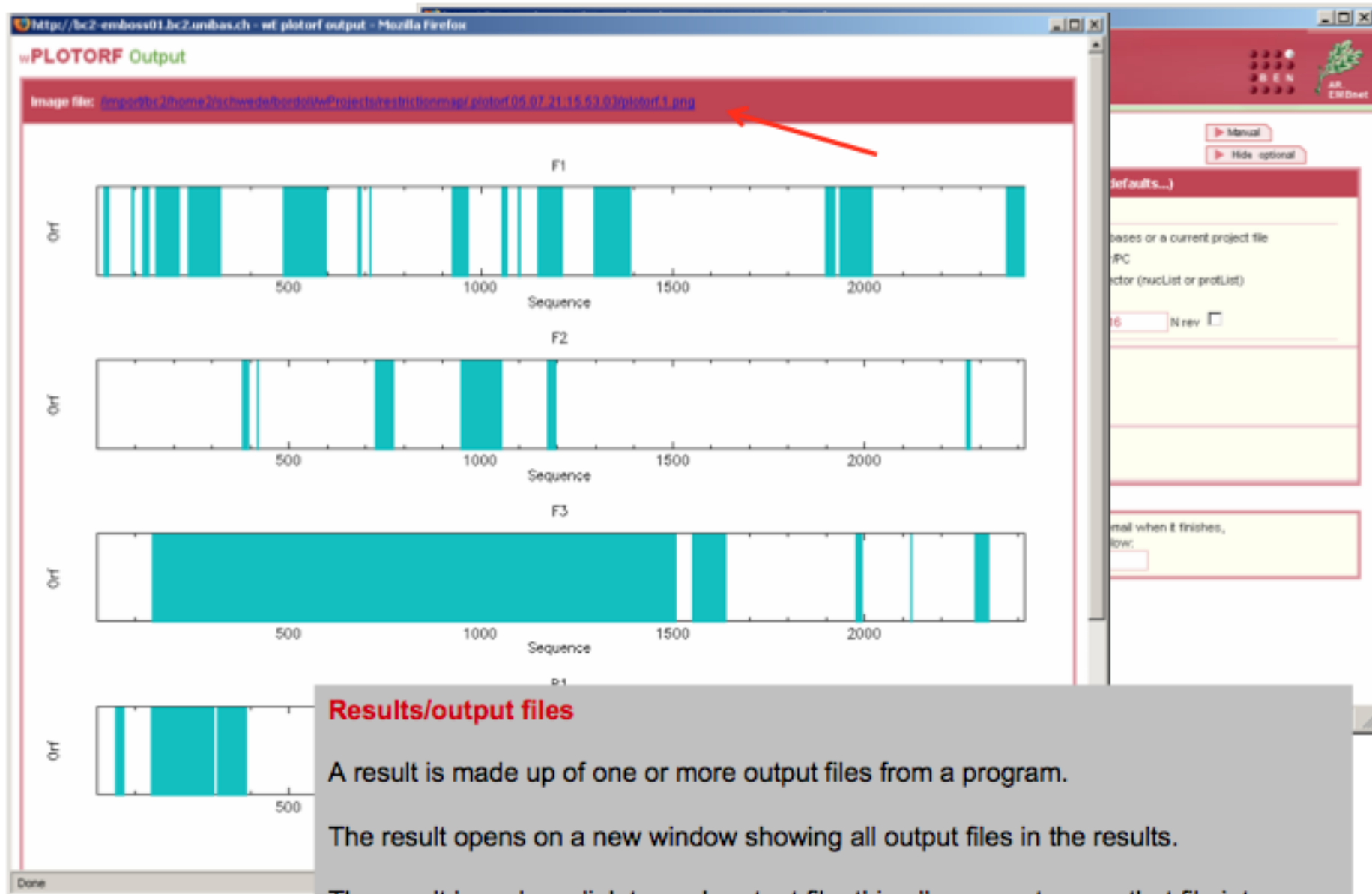
If you are submitting a long job and would like to be informed by email when it finishes,

## 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



### 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.



- 2D STRUCTURE
- COXON USAGE
- COMPOSITION
- CPO CLANES
- GENE FINDING
- WOTFS
- MUTATION
- PROBES
- PROFILES
- REPEATS
- RESTRICTION
- TRANSCRIPTION
- TRANSLATION
- Background
- colored
- plot
- preseq
- rmmap
- showoff

restrictionmap project

PROJECT MANAGEMENT

PROJECT FILES

PROJECT RESULTS

Program Output	yy.mm.dd	hh.mm.ss	Copy
<i>plotorf</i>	05.07.21	15.53.03	<input type="button" value="Files"/>
<i>plotorf</i>	05.07.21	15.52.44	<input type="button" value="Files"/>
<i>showorf</i>	05.07.21	09.50.12	<input type="button" value="Files"/>
<i>eprimer3</i>	05.07.21	09.49.37	<input type="button" value="Files"/>
<i>eprimer3</i>	05.07.19	20.20.03	<input type="button" value="Files"/>
<i>wossname</i>	05.07.19	20.18.58	<input type="button" value="Files"/>
<i>getorf</i>	05.07.19	20.12.55	<input type="button" value="Files"/>
<i>getorf</i>	05.07.19	20.09.51	<input type="button" value="Files"/>
<i>plotorf</i>	05.07.19	20.09.15	<input type="button" value="Files"/>
<i>transeq</i>	05.07.19	20.07.55	<input type="button" value="Files"/>
<i>wossname</i>	05.07.19	20.05.51	<input type="button" value="Files"/>
<i>plotorf</i>	05.07.18	17.43.11	<input type="button" value="Files"/>
<i>restrict</i>	05.07.18	15.35.03	<input type="button" value="Files"/>
<i>restrict</i>	05.07.18	15.33.23	<input type="button" value="Files"/>
<i>emma</i>	04.02.16	17.5.14	<input type="button" value="Files"/>

Project Results

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

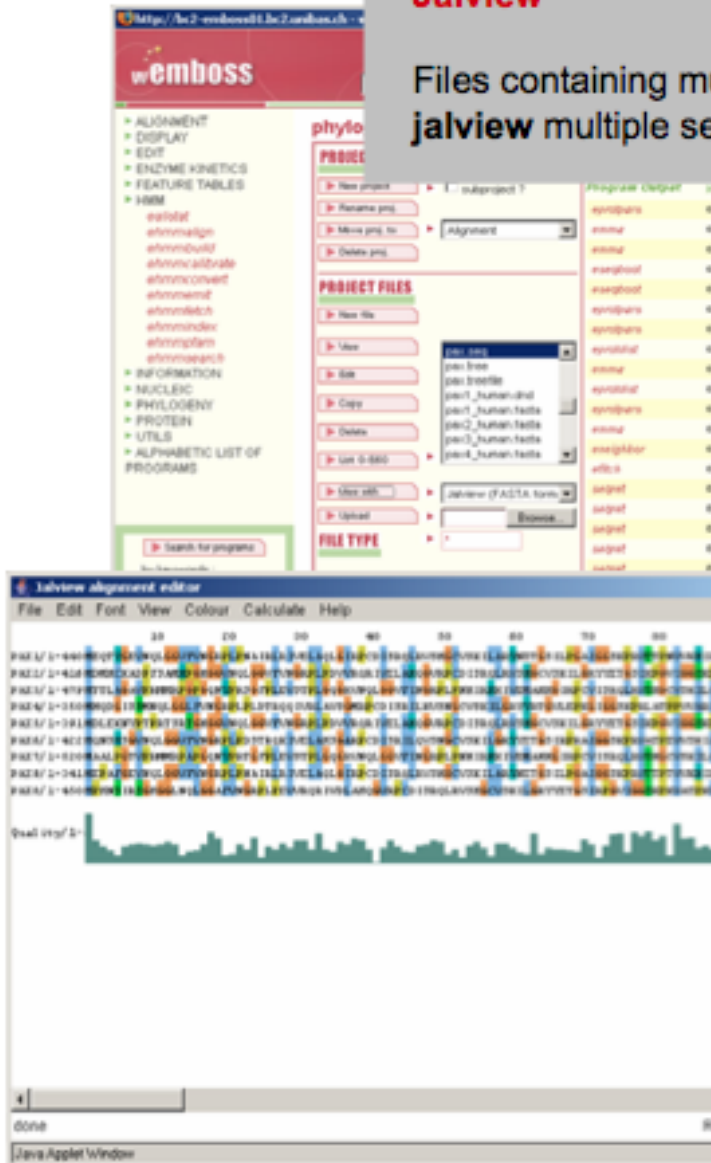


**Jalview**

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

**Jalview**

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



## PROJECT FILES

► New file

▶ View

Edit

Copy

▶ Delete

► List G-E&G

▶ **View with**

▶ Upload

## FILE TYPE

pax.seq  
pax.tree  
pax.treefile  
pax1\_human.dnd  
pax1\_human.fasta  
pax2\_human.fasta  
pax3\_human.fasta  
pax4\_human.fasta

pax.tree  
pax.treefile  
pax1\_human.dnd  
pax1\_human.fasta  
pax2\_human.fasta  
pax3\_human.fasta  
pax4\_human.fasta

```
pax.treefile
pax1_human.dnd
pax1_human.fasta
pax2_human.fasta
pax3_human.fasta
pax4_human.fasta
```

pax1\_human.dnd  
pax1\_human.fasta  
pax2\_human.fasta  
pax3\_human.fasta  
pax4\_human.fasta

pax1\_human.fasta  
pax2\_human.fasta  
pax3\_human.fasta  
pax4\_human.fasta

pax2\_human.fasta  
pax3\_human.fasta  
pax4\_human.fasta

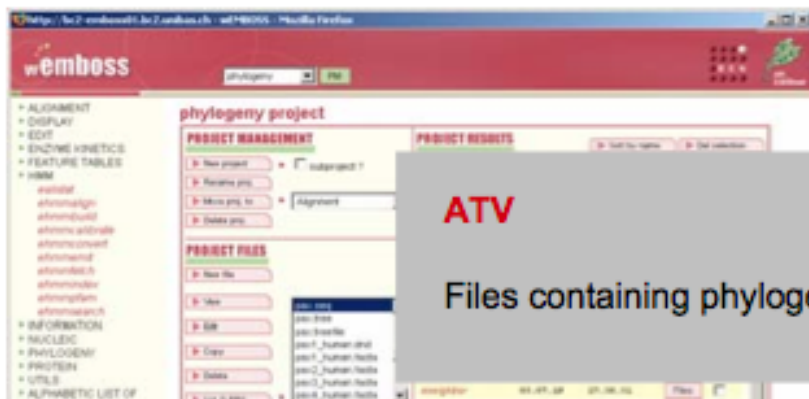
pax3\_human.fasta  
pax4\_human.fasta

pax4\_human.fasta

Jalview (FASTA form: 

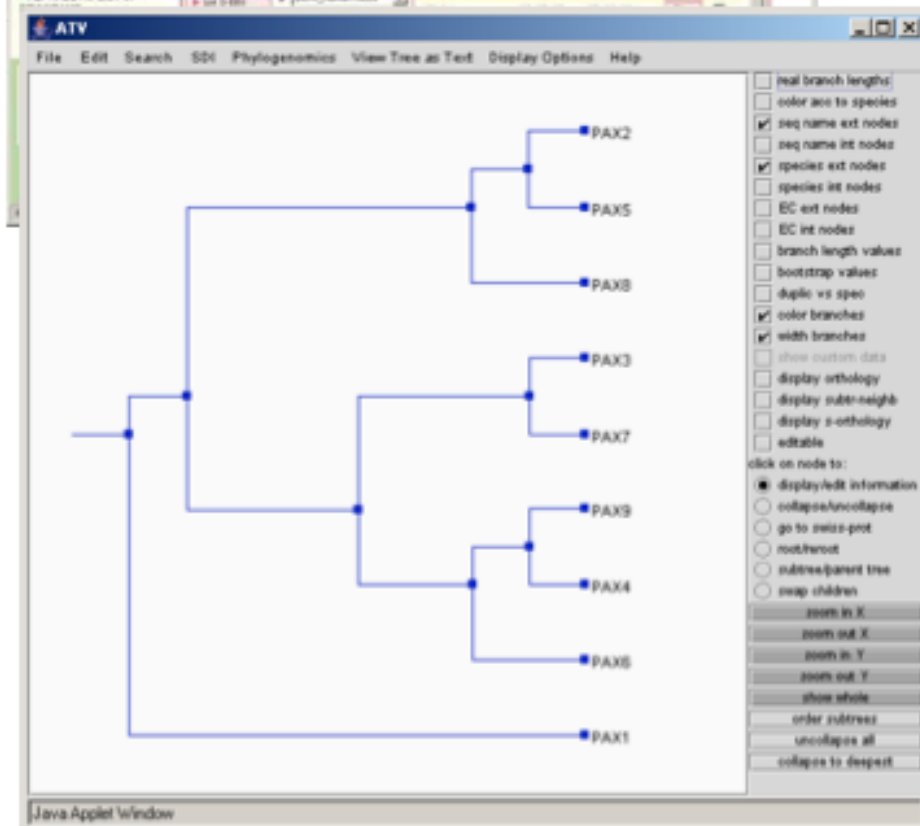
☆





## ATV

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



## PROJECT FILES

▶ New file

▶ View

▶ Edit

▶ Copy

▶ Delete

▶ List G-E&G

▶ View with

▶ Upload

pax.seq  
pax.tree  
pax.treefile  
pax1\_human.dnd  
pax1\_human.fasta  
pax2\_human.fasta  
pax3\_human.fasta  
pax4\_human.fasta

▶ ATV

Browse...

## FILE TYPE

▶ \*

# Help on a program



- INFORMATION
- MENUS
- NUCLEIC
  - 2D STRUCTURE
  - CODON USAGE
  - COMPOSITION
  - CPG ISLANDS
  - GENE FINDING
  - MOTIFS
  - MUTATION
  - PRIMERS
  - PROFILES
  - REPEATS
  - RESTRICTION
  - TRANSCRIPTION
  - TRANSLATION
    - backtranseq*
    - coderet*
    - plotorf*
    - prettyseq*
    - remap*
    - showorf*
    - showseq*
    - sixpack*

► Search for programs

by keywords :

and ☒ or ☐

Marc Colet & Martin Sarachu

Version 1.0.2

ZIP22

**wplotorf** (Plot potential open reading frames)

► Manual

► Hide optional

► Run plotorf

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

## INPUT

Sequence(s) from the

☒ server (file/database) or ☐ local computer file or ☒ list selector

(file must contain a DNA sequence)

mySeq  rev  no  begin 1 end 58 N

## ADVANCED

ATG Start codons

TAA,TAG,TGA Stop codons

## OUTPUT

PNG  Output graphic format

► Run plotorf

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:

**THE END**

