

- **E** uropean
- **M** olecular
- **B** iology
- **O** pen
- **S** oftware
- **S** uite

<http://emboss.sourceforge.net/>
and CVS at Open Bio

EMBOSS: History

- EMBOSS started in March 1996
- First requirements based on problems in existing commercial software, and needs for public source code
- First ajax library written August 1996
- 30 potential sites identified November 1996
- Wellcome Trust proposal February 1997 (Sanger, HGMP and EBI)
- Accepted August 1997
- Project started November 1997.
- EMBOSS 1.0.0 released on 15th July 2000.
- EMBOSS 2.0.0 released on 15th July 2002.
- EMBOSS 2.9.0 released on 15th July 2004.

EMBOSS: Current status

- 200+ applications in EMBOSS
- 100+ applications in EMBASSY
- 40+ known interfaces
- 21,000 unique downloads
- 500,000 runs / month at RFCGR

Hinxton Campus



EMBOSS: Original Aims (1996)

- Developing new tools for sequence analysis
- Replacing popular but obsolete EGCG applications
- Integrating with public packages: SRS and ACEDB
- Integrating with popular user interface packages
- Integrating with other publicly available packages and tools
- Encouraging developers to use the EMBOSS libraries

Original Target Users

Each of the following groups had their own special needs which EMBOSS aimed to satisfy:

- Sanger Centre genomic sequencing and analysis groups
- RFCGR/HGMP registered academic users (about 10,000)
- EMBnet service providers in 30 other countries with over 30,000 users
- Academic users everywhere
- Pharmaceutical and biotechnology industry
- Bioinformatics developers

EMBOSS: Standards

- GPL/LGPL licensing
- All code in ANSI standard C
- Support for all common Unix platforms
- Choice of sequence formats
- Choice of database formats
- Site specific customising
- User interface definitions

Software libraries

- AJAX
 - Strings, sequences, command-line, graphics API
- NUCLEUS
 - Algorithms
- Third party
 - Graphics
 - Xml

Code Documentation

We have developed our own source code documentation standard

- Based on JavaDoc with extensions
- Automatic conversion to HTML
- Validation and correction daily
- Indexed in SRS srs.rfcgr.mrc.ac.uk and srs.ebi.ac.uk

[Reset](#)[Previous Entry](#) Entry 5 of 9 from [Query 4](#) [Next Entry](#)

Entry Information

Entry from: [EFUNC](#)

Entry Options

Link to related information:

[Link](#)

Save entry:

[Save](#)

View:

[Printer Friendly](#)

Name

ajSeqReverseStr

Source file: [ajax/ajseq.c](#)

Description

Reverses and complements a nucleotide sequence provided as a string.



Input

Type	Name	R/W	Comment
AjPStr*	pthis	Update	Sequence as a string.

Returns

Type	Comment
void	

Prototype

void  ajSeqReverseStr  (AjPStr* pthis);

Body

```
{
    char *cp;
    char *cq;
    char tmp;

    cp = ajStrStrMod(pthis);
    cq = cp + ajStrLen(*pthis) - 1;

    while (cp < cq)
    {
        tmp = ajSeqBaseComp(*cp);
        *cp = ajSeqBaseComp(*cq);
        *cq = tmp;
        cp++;
        cq--;
    }

    if (cp == cq)
        *cp = ajSeqBaseComp(*cp);

    return;
}
```

EMBOSS: ACD Files

- AJAX Command Definition (ACD) files control all EMBOSS applications.
- ACD files are a complete user interface definition
- Simple syntax with dependencies
- Flexible command line syntax
- Automatic processing
- Information provided at startup
- Conversion for Web, GUI and other interfaces
- Developer tools: `acdc`, `acdtrace`, `entrails`
- QA testing tools: `achtable`, `acdvalid`

Example ACD file

This test file shows how dependencies are handled in ACD:

```
application: ajtest [ documentation: "Testing"  
                    group: "test" ]  
sequence: sequence [ parameter: "y" ]  
boolean: test [ default: "y" ]  
float: fval [ standard: "y" default: "2.5"  
             maximum: "100.0" ]  
integer: aval [ default: "10"  
              maximum: "$(sequence.length)" ]  
integer: bval [ default: "$(aval)"  
              maximum: "@($aval.maximum)+1" ]  
outfile: out [ parameter: "y" ]
```

ACD Processing

A single call (`embInit`) handles:

- Parsing the ACD file
- Parsing the command line
- Prompting the user
- Validation
- Opening files
- Reading sequences etc.

All values are treated as strings until they are used.

All sequences are read, and passed as sequence objects.

No further interaction with the user is expected.

EMBOSS: Sequences

Uniform Sequence Address (USA): URL-style naming

database : entryname

- embl : ecoompa
- swissprot-id : opsd_bovin
- embl-acc : x13776

format :: filename

- fasta :: /nfs/users/pmr/paamir.fa
- gcg :: ecoompa.em_ba

format :: filename : entryname

- fasta :: unfinished : AH6.1

Also **@listfile** and **asis::gctgactgactgatg**

To be extended to LSIDs **urn:lsid:emboss.org:dbname:entryname**

Seqret

Seqret is a very simple application

- It reads a sequence USA (in any format, from anywhere)
- It writes a sequence USA (in any format)

If you tell it the sequence has feature annotation:

- It reads the features (in any format)
- It writes the features (in any format)

Seqret has 13 lines of code

The source code seqret.c

```
#include "emboss.h"

int main(int argc, char **argv) {
    AjPSeqall  seqall;
    AjPSeqout  outseq;
    AjPSeq seq = NULL;
    embInit("seqret", argc, argv);
    seqall = ajAcdGetSeqall ("sequence");
    outseq = ajAcdGetSeqout ("seqout");
    while (ajSeqallNext (seqall, &seq))
        ajSeqWrite (outseq, seq);
    ajSeqWriteClose (outseq);
    ajExit();
}
```


EMBOSS Quality Control

- Nightly build with no compiler warnings
- 2,000 test runs (including expected fail conditions)
- 200 valgrind memory leak tests
- Code documentation validation and indexing
- ACD file validation
- ACD documentation completeness
- Program documentation: description, command line qualifiers, example run(s) and input/output files
- Web site updates

The Applications (programs)

Contents

- [Introduction](#)
- [Groups of applications](#)
- [List of applications](#)
- [EMBASSY programs](#)
- [Proposed new applications](#)

Introduction

The programs are listed in alphabetical order, divided into four sections. Look at the individual applications or go to the [GROUPS](#) page.

Applications in the [current release](#)

If you have any comments about any of the programs, mail the EMBOSS open discussion list embooss@embnet.org.
Bug reports should be sent to the EMBOSS Support Team embooss-bug@embnet.org

Program name	Author(s)	Description
aaindexextract	RFCGR	Extract data from AAINDEX
abiview	RFCGR	Reads ABI file and display the trace
acdc	Sanger	Tests definition files for any EMBOSS application.
antigenic	RFCGR	Finds antigenic sites in proteins
backtranseq	RFCGR	Back translate a protein sequence
banana	Sanger	Bending and Curvature Plot in B-DNA
biosed	RFCGR	Replace or delete sequence sections
btwisted	RFCGR	Calculates the twisting in a B-DNA sequence
cai	RFCGR	CAI codon usage statistic
chaos	Sanger	Create a chaos plot for a sequence.

PROTEIN MOTIFS

Program name	Description
antigenic	Finds antigenic sites in proteins
digest	Protein proteolytic enzyme or reagent cleavage digest
epestfind	Finds PEST motifs as potential proteolytic cleavage sites
fuzzpro	Protein pattern search
fuzztran	Protein pattern search after translation
helixturnhelix	Report nucleic acid binding motifs
oddcomp	Finds protein sequence regions with a biased composition
patmatdb	Search a protein sequence with a motif
patmatmotifs	Search a PROSITE motif database with a protein sequence
pepcoil	Predicts coiled coil regions
preg	Regular expression search of a protein sequence
pscan	Scans proteins using PRINTS
sigcleave	Reports protein signal cleavage sites
meme	Motif detection

emboss helixturnhelix

Function

Report nucleic acid binding motifs

Description

helixturnhelix uses the method of Dodd and Egan and finds helix-turn-helix nucleic acid binding motifs in proteins.

The helix-turn-helix motif was originally identified as the DNA-binding domain of phage repressors. One alpha-helix lies in the wide groove of DNA; the other lies at an angle across DNA.

Usage

Here is a sample session with **helixturnhelix**

```
% helixturnhelix
Report nucleic acid binding motifs
Input sequence(s): tsw:laci_ecoli
Output report [laci_ecoli.hth]:
```

[Go to the input files for this example](#)

[Go to the output files for this example](#)

Command line arguments

```
Standard (Mandatory) qualifiers:
[-sequence]      seqall      Sequence database USA
[-outfile]       report       Output report file name

Additional (Optional) qualifiers:
-mean            float        Mean value
```

Standard (Mandatory) qualifiers		Allowed values	Default
[-sequence] (Parameter 1)	Sequence database USA	Readable sequence(s)	Required
[-outfile] (Parameter 2)	Output report file name	Report output file	
Additional (Optional) qualifiers		Allowed values	Default
-mean	Mean value	Number from 1.000 to 10000.000	238.71
-sd	Standard Deviation value	Number from 1.000 to 10000.000	293.61
-minsd	Minimum SD	Number from 0.000 to 100.000	2.5
-eightyseven	Use the old (1987) weight data	Boolean value Yes/No	No
Advanced (Unprompted) qualifiers		Allowed values	Default
(none)			

Input file format

helixturnhelix reads one or more protein sequence USAs.

Input files for usage example

'tsw:laci_ecoli' is a sequence entry in the example protein database 'tsw'

Database entry: tsw:laci_ecoli

```
ID  LACI_ECOLI      STANDARD;      PRT;   360 AA.
AC  P03023; P71309; Q47338; O09196;
DT  21-JUL-1986 (Rel. 01, Created)
DT  01-NOV-1997 (Rel. 35, Last sequence update)
DT  15-DEC-1998 (Rel. 37, Last annotation update)
DE  LACTOSE OPERON REPRESSOR.
GN  LACI.
OS  Escherichia coli.
OC  Bacteria; Proteobacteria; gamma subdivision; Enterobacteriaceae;
OC  Escherichia.
```

Output file format

The output is a standard EMBOSS report file.

The results can be output in one of several styles by using the command-line qualifier **-rformat xxx**, where 'xxx' is replaced by the name of the required format. The available format names are: embl, genbank, gff, pir, swiss, trace, listfile, dbmotif, diffseq, excel, featable, motif, regions, seqtable, simple, srs, table, tagseq

See: <http://emboss.sourceforge.net/docs/themes/ReportFormats.html> for further information on report formats.

By default **helixturnhelix** writes a 'motif' report file.

Output files for usage example

File: laci_ecoli.hth

```
#####
# Program: helixturnhelix
# Rundate: Tue Apr 06 14:29:48 2004
# Report_format: motif
# Report_file: laci_ecoli.hth
#####

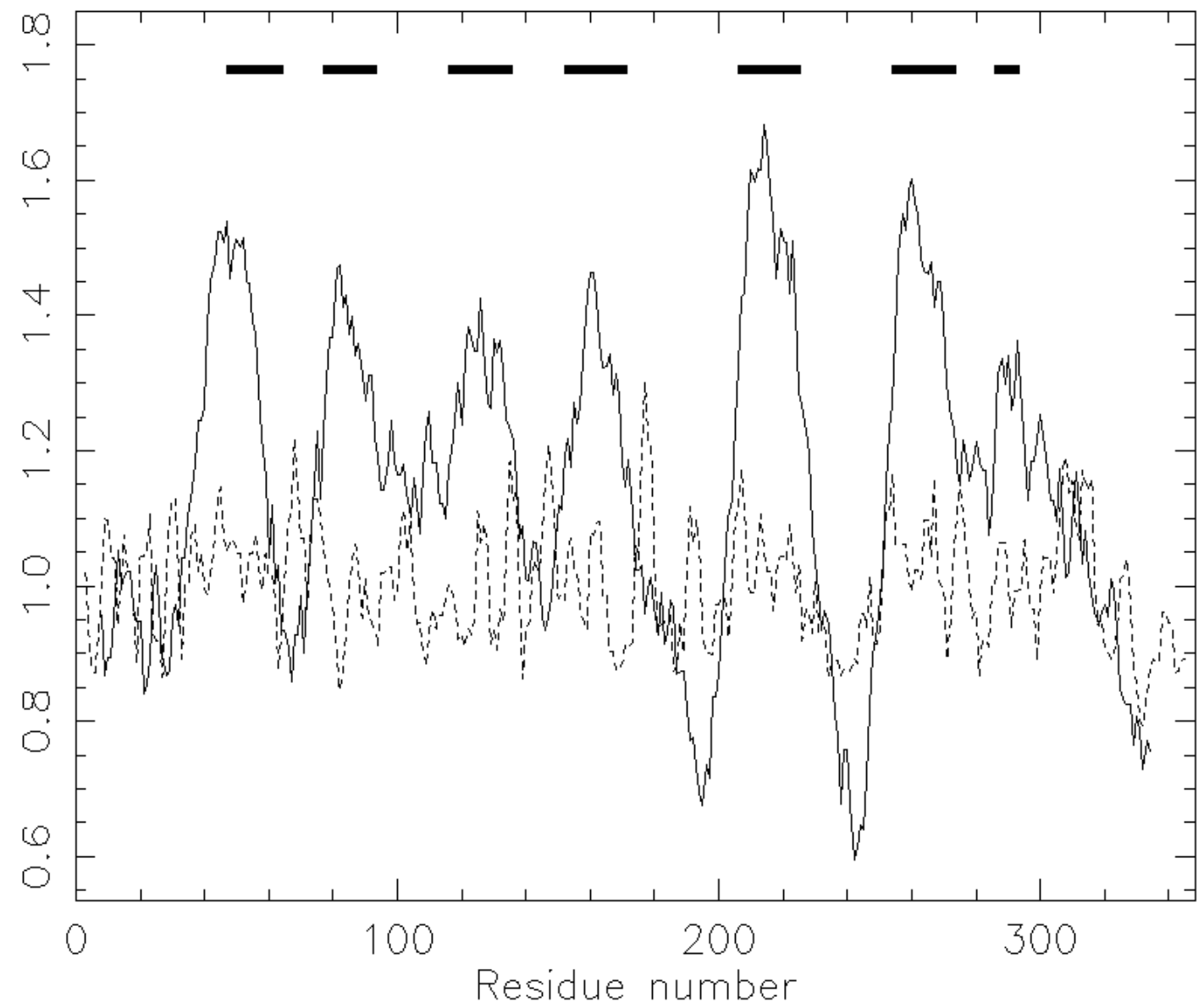
#=====
#
# Sequence: LACI_ECOLI      from: 1    to: 360
# HitCount: 1
#
# Hits above +2.50 SD (972.73)
#
#=====

Maximum_score_at at "*"

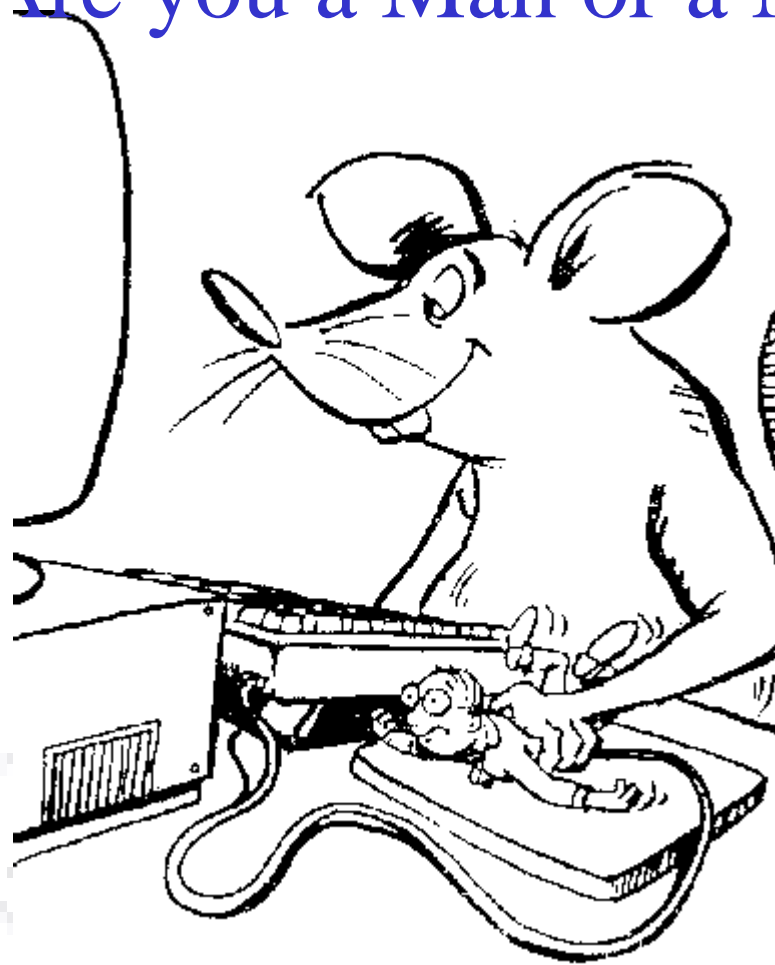
(1) Score 2160.000 length 22 at residues 4->25
      *
Sequence: VTLYDVAEYAGVSYQTVSRVVN
          |                |
          4                25
Standard_deviations: 6.54

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

tmap (14/12/99)



Users: Are you a Man or a Mouse?



Command Line

EMBOSS has many possible command lines:

- Unix style

```
% seqret
```

```
  What sequence []: embl:paamir
```

```
  Output file [paamir.fasta]:
```

```
% seqret embl:paamir -send 100 -auto
```

```
% seqret embl:paamir -se 100 -auto
```

```
% seqret -se 100 embl:paamir -auto
```

- GCG style

```
% seqret embl:paamir -send=100 -auto
```

- VMS style

```
$ seqret /SEQUENCE=EMBL:PAAMIR /SEND=100 /AUTO
```


Web Interfaces

We would like to fully automate the generation of "external application definitions" for the Web, GUI and other interfaces, including these collaborators:

- Pise (Catherine Letondal, Pasteur)
- AppLab (Martin Senger, EBI)
- W2H (DKFZ Germany)
- wEMBOSS (EMBnet Belgium+Argentina)
- SRS applications (LION Bioscience)
- BioNavigator (Entigen Australia)
- GCG SeqWeb (in-house pharma)

Web interface (W2H)

Go to Main Page Go to Results Page Go to Files Page Advanced Mode Exit ?

 **Application antigenic** [Help on antigenic](#)

Antigenic looks for potential antigenic regions using the method of Kolaskar.

Run: just now daily weekly monthly

Begin **End** **List item**

Nucl Circular
 Prot Reverse

Command Line (*parameters/options can only be set in the section(s) below*):

Output filename

Antigenicity data file Using:

Optional parameters

3 Minimum length to report

Web interface (SRS)

The screenshot shows a web browser window titled "Launch - Tiscali 10.0". The address bar contains the URL: `http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?page+Launch+id+1g5Zt1NEFs1+-appl+antigenic+-launchFrom+top`. The browser's menu bar includes "File", "Edit", "View", "Favorites", "Tools", and "Help". The address bar also features navigation buttons (Back, Forward, Home, Stop, Reload) and a search field.

The web page content includes a navigation menu with items: "Quick Search", "Library Page", "Query Form", "Tools", "Results", "Projects", "Views", "Databanks", and "HELP". The main content area is titled "Antigenic" and features a "More Info..." link. Below the title, there is a "Job name:" field containing the text "temp" and a "Launch" button. A text area is provided for "Enter 1 or more FASTA sequences :", with a "Browse..." button below it. An alternative "or enter file name" field with a "Browse..." button is also present. A "Required Options" section contains a "Minimum length" field set to "6".

On the left side, there are two panels: "Job Options" with a "Reset" button and a note about the LSF batch queuing system, and "Parameter set options" with a "Save current parameter set" button and an "as:" field. The footer of the page reads: "SRS Release 7.1.1 Copyright © 1997-2003 LION bioscience AG. All Rights Reserved. [Terms of Use](#) [Feedback](#)".

The Windows taskbar at the bottom shows the "start" button and several open applications: "Google Search: ...", "GNU General Pu...", "EMBOSS - Tiscali...", "Launch - Tiscali ...", "Newc-04", and "TurkCourse". The system tray on the right shows "Internet" and the time "15:57".

GUI Interfaces

- Jemboss (RFCGR/HGMP Hinxton + Sanger Institute)
- Taverna (EBI)
- Luke McCarthy (PBI, Canada)
- Kaptain (Thomas Siegmund)
- Staden package (Kathryn Beal, LMB)
- CINEMA (EMBnet Manchester)

GUI Interfaces: Luke McCarthy

The screenshot shows a web browser window titled "EMBOSS GUI - Tiscali 10.0". The address bar shows the URL "http://bioinfo.pbi.nrc.ca:8090/EMBOSS/". The browser's toolbar includes navigation buttons (Back, Forward, Home, Stop, Reload), a search bar, and various utility icons. The main content area displays the EMBOSS web interface, which features a navigation menu on the left and a central form area.

EMBOSS version 2.3.1

PROTEIN 3D STRUCTURE

- contacts
- dichet
- interface
- psiblasts
- scopalign
- seqsort
- siggen
- sigscan

PROTEIN COMPOSITION

- backtranseq
- charge
- checktrans
- compseq
- emowse
- freak
- iep
- mwfilter
- octanol
- pepinfo
- pepstats
- pepwindow
- pepwindowall

PROTEIN MOTIFS

- antigenic
- digest
- fuzzpro
- fuzztran
- helixturnhelix
- odcomp
- patmatdb
- patmatmotifs
- pepcoil
- preg
- pscan
- sigcleave

PROTEIN MUTATION

(Finds antigenic sites in proteins)

Help ?

Fields with a coloured background are optional and can safely be ignored...
[Hide optional fields]

1. SET THE PARAMETERS FOR THE RUN (OR ACCEPT THE DEFAULTS...)

input section

Select a set of sequences.

Use one of the following three fields: *(file must contain protein sequences)*

- To access a sequence from a database, enter the USA path here: *(dbname:entry)*
- Or, upload a sequence file from your local computer here:
- Or enter the sequence data manually here:

required section

Minimum length (integer)

output section

Report format:

Done

start | Google Search: ... | GNU General Pu... | EMBOSS - Tiscali... | EMBOSS GUI - T... | Newc-04 | TurkCourse | Internet | 15:59

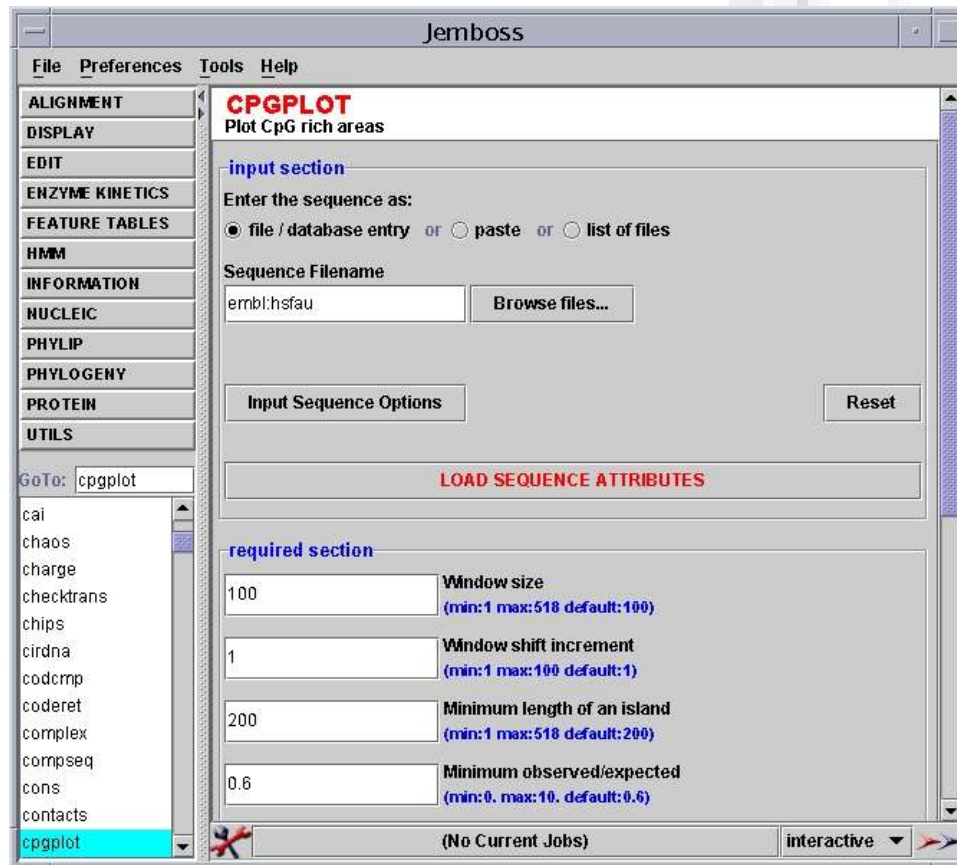
GUI Interfaces: Kaptain

The screenshot displays the KEMBOSS GUI interface with several windows open:

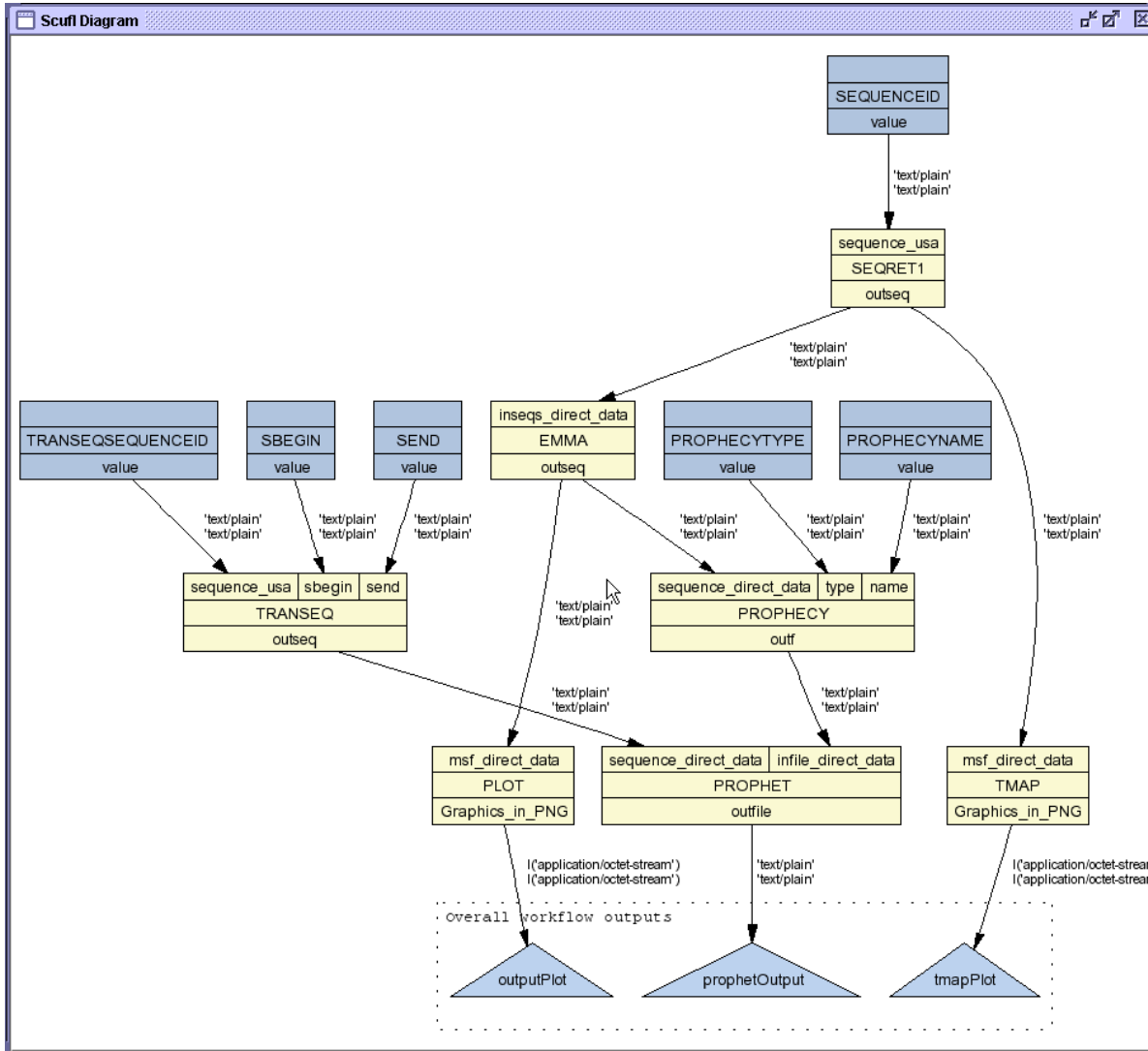
- KEMBOSS**: The main application window with a menu bar (File, Edit, Search, Preferences, Shell, Macro, Windows, Help) and a sidebar containing various tools like Alignment consensus, Alignment fit, Alignment multiple, etc.
- Untitled (seq01.txt)**: A text editor window showing a CLUSTAL W(1.4) multiple sequence alignment of protein sequences.
- Multiple alignment program - interface to ClustalW**: A dialog box for configuring the ClustalW alignment parameters. It includes fields for Gap opening penalty (100), Gap extension penalty (50), Substitution matrix (BLOSUM), and DNA substitution table. It also has checkboxes for "Use Transition weight" and "Reduce gap penalty for hydrophilic residues".
- Seq01.txt**: A window showing a circular plot of protein sequences, with residues numbered 1 to 24 around the perimeter.
- Shows protein sequences as helices**: A window for plotting protein sequences as helices, with parameters for Steps (100), Turns (10), Squares, Stars, and Octagons.
- List available databases**: A window showing a table of available databases.

Name	Type	DB	Seq	Alt	Comment
swiss	P	OK	-	-	EMBL
genbank	N	OK	-	-	Genbank
cd protein P	OK	OK	OK	OK	NCBI
swiss	P	OK	-	-	SWISS
trnswiss	P	OK	OK	OK	SWISS

GUI Interfaces: Jemboss



GUI Interfaces: Taverna



Other Interfaces

- SoapLab/Taverna Web Services (Martin Senger, EBI)
- BioMoby-S Web Services (Martin Senger, EBI)
- SRS 7.x APIs: Perl, C++, Java (Lion Bioscience)
- Other companies with EMBOSS integrated: Turboworx, MetaLife, InforSense, ...
- AppLab CORBA (Martin Senger, EBI)
- BioPerl
- Staden SPIN and SPIN2
- Other academic projects ...

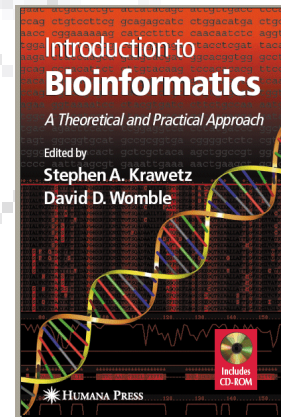
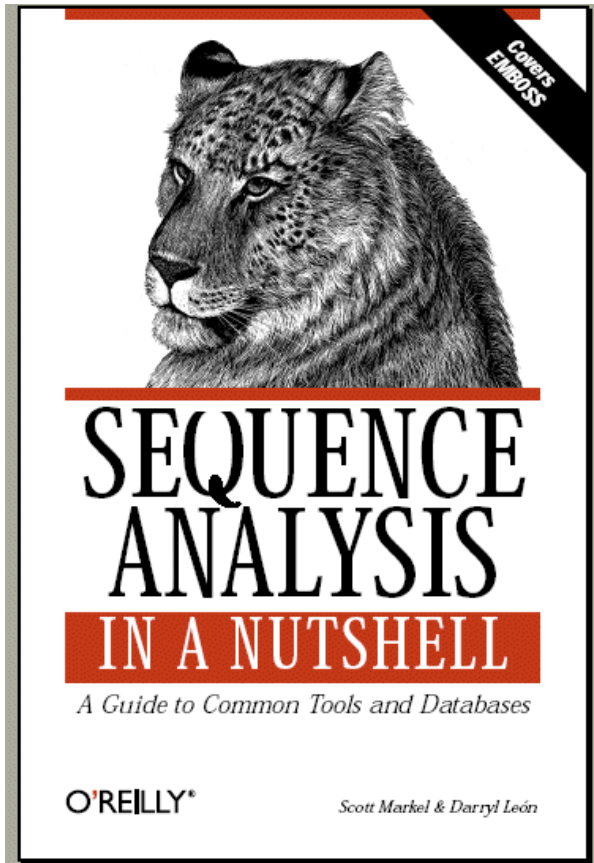
Adding to EMBOSS

- All contributions are welcomed
- Applications must be linked to the EMBOSS libraries
- For now, code should be in ANSI C or C++
- Preferably under a GPL licence, but ...
- The "EMBOSS Associated Applications Directory" covers software as separate packages, or contributed with non-GPL licences
 - MSE, TOPO, EMNU, PHYLIP
 - MEME, ESIM4, HMMER
 - DOMAINATRIX (Jon Ison)
 - PHYLIPNEW (phylip 3.6)
 - APPENDIXD
- (Even commercial software could be included)
- (But companies may prefer to use the GPL licence)

Technical Support

- All EMBOSS code is fully tested – program examples etc.
- Memory leaks are tested and fixed before each release
- All EMBOSS code is fully and automatically documented.
- Emboss-bug mailing list support group with expertise in software installation and maintenance.
- Emboss-dev developers mailing list
- Emboss users mailing list
- RFCGR/HGMP and EMBnet bioinformatics training courses.
- Program documentation generated and checked.
- Series of books in preparation (for June 2005).
- Sourceforge trackers for bugs and feature requests

EMBOSS Books



24 Analyzing Sequences Using the Staden Package and EMBOSS

Rodger Staden, David P. Judge, and James K. Bonfield

Disaster proof software licences



Disaster proof software licences

- 1996 EMBOSS started by Peter Rice (Sanger) and Alan Bleasby (SEQNET Daresbury), in collaboration with Thure Etzold (EBI)
- 1997 funding approved by the Wellcome Trust
- 1998 SEQNET relocated to Hinxton (HGMP)
- 1999 Thure goes to LION Bioscience
- 2000 Peter leaves Sanger – EMBOSS goes to Alan at HGMP
- 2001 LION (Peter) adds EMBOSS to SRS and updates EMBOSS
 - CCP11 funding for EMBOSS development
- 2002 Peter leaves LION
- 2003 Peter joins EBI – integrating EMBOSS in myGrid services
 - Medical Research Council terminates funding for Rodger Staden
 - HGMP is renamed after Rosalind Franklin

Disaster proof software licences

- 1996 EMBOSS started by Peter Rice (Sanger) and Alan Bleasby (SEQNET Daresbury), in collaboration with Thure Etzold (EBI)
- 1997 funding approved by the Wellcome Trust
- 1998 SEQNET relocated to Hinxton (HGMP)
- 1999 Thure goes to LION Bioscience
- 2000 Peter leaves Sanger – EMBOSS goes to Alan at HGMP
- 2001 LION (Peter) adds EMBOSS to SRS and updates EMBOSS
 - CCP11 funding for EMBOSS development
- 2002 Peter leaves LION
- 2003 Peter joins EBI – integrating EMBOSS in myGrid services
 - Medical Research Council terminates funding for Rodger Staden
 - HGMP is renamed after Rosalind Franklin
- 2004 April 1st: RFCGR to be closed by MRC within 15 months

- All the code is still licensed to everyone under GPL.

EMBOSS new home page



[news](#) • [what?](#) • [apps](#) • [docs](#) • [frontends](#) • [get](#) • [admin](#)
[developers](#) • [support](#) • [co-ordination](#) • [licence](#) • [credits](#)

hosted by 

latest stable release: [2.9.0](#)

The EMBOSS project gratefully acknowledges the support of [CCP11](#).

Latest news about EMBOSS

[News page](#)

Important

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. We will keep you informed of further developments.

EMBOSS Survey 2004

It would help the EMBOSS project tremendously if you complete [this survey](#).

This will help us to both gain support for EMBOSS and to steer the future development of the project down lines suggested by you.

We will keep you informed of further developments.

EMBOSS Survey

- First responses:
 - 405/471 do use EMBOSS
 - Of these, 403/405 users find it valuable
 - 397/471 may attend a users course
 - 196/471 " " a developers course
 - 422/471 want EMBOSS for free
 - But 49/471 were "not sure"
 - 415/471 preferred public funding for EMBOSS
 - 56/471 preferred part industry funding
 - Now we will ask industry ...

EMBOSS Phase Three

- Now on emboss.sourceforge.net and open-bio
- Coordinated by EBI et al. from end July 2005
- Various funding options under consideration
- Books in preparation by the EMBOSS Team (summer 2005)
- New applications
- Non-sequence input and output datatypes (MicroArray, ChemBoss)
- New graphics libraries
- New database access methods (B+ tree indexing)
- Life Science identifiers
- Ontologies: The ACD Ontology "TAO" and RDF templates in ACD
- New user interfaces
- Closer integration with workflows and web/grid services
- ... And much, much more ...
- And possibly commercial sponsorship

EMBOSS in Industry



EMBOSS in Industry

- Large companies (e.g. big pharma) with in-house bioinformatics groups
- Small biotechs with no in-house expertise
- Bioinformatics solution providers
- Bioinformatics service providers
- Workflow/pipeline software providers
- Hardware manufacturers
- Providers of operating systems and office software

EMBOSS in Academia

- Current EMBOSS user community
- Former RFCGR service users
- UK university bioinformatics services

- Needs for software, support, training, custom developments, user forums, developer workshops, etc. ...

Acknowledgements

- HGMP: Alan Bleasby, Jon Ison, Tim Carver, Hugh Morgan, Claude Beesley, Lisa Mullan, Damian Counsell, Gary Williams, Val Curwen, Mark Faller, Sinead O’Leary, Thon deBoer, Martin Bishop
- EBI: Peter Rice, Martin Senger, Tom Oinn, Jaina Mistry , Rodrigo Lopez, Sharmilla Pillai
- LION: Thomas Laurent, Bijay Jassal, Thure Etzold
- Sanger Centre: Ian Longden, Richard Bruskiwich, Simon Kelley
- National bioinformatics service providers in: Norway, Spain, Italy, Netherlands, Germany, Belgium, Russia, China, Canada, Australia, and Manchester
- Others: Catherine Letondal, Don Gilbert, Rodger Staden, Bill Pearson,