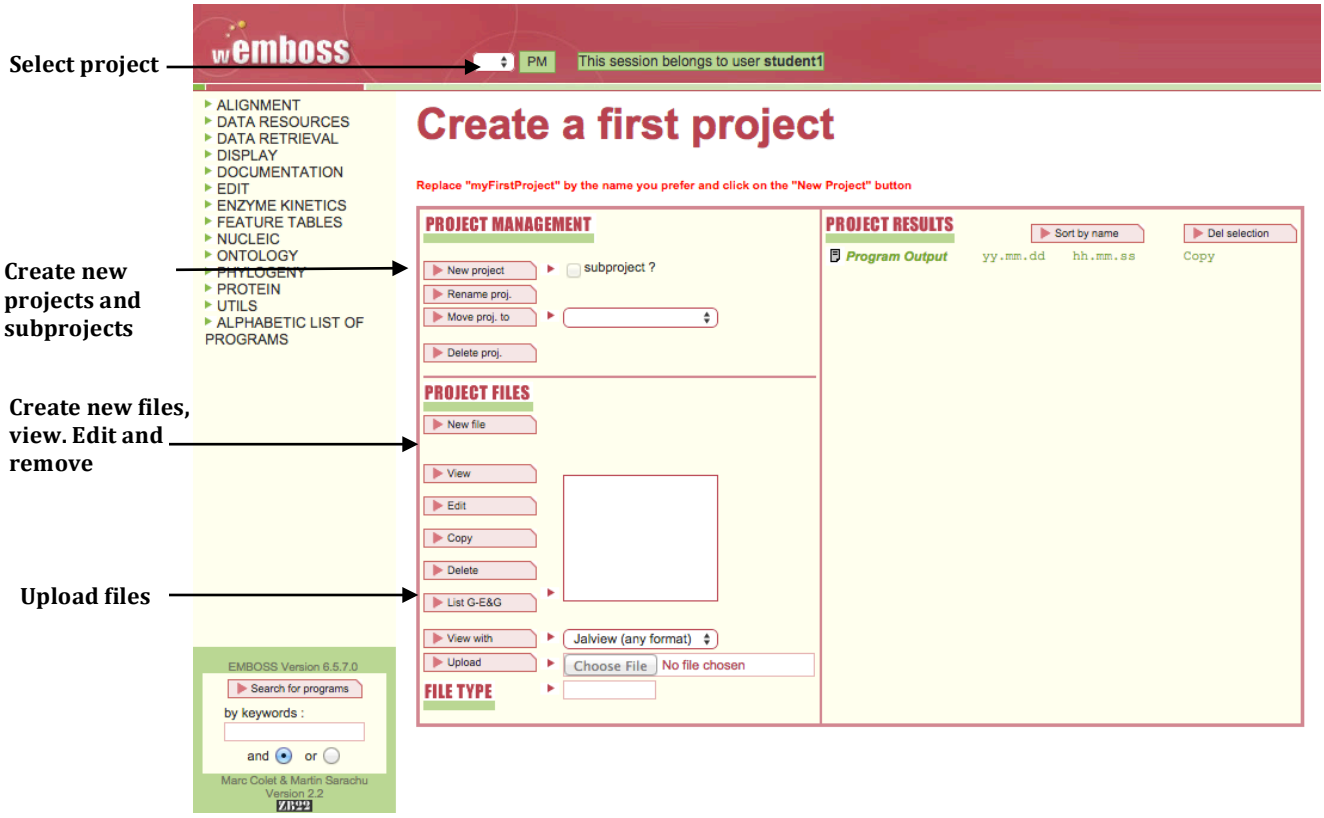
**wEMBOSS PRACTICAL**

From the eBioKit interface launch wEMBOSS and put in your login details it should give you the wEMBOSS interface. (that looks as in the diagram below)




## 1. Create a new project

To start of you need to create a new project | *click on create new project and give it a name*

Throughout this tutorial, we're going to look at members of the rhodopsin family of G-protein coupled receptors. The general principles are, of course, applicable to any sequences you would like to analyze. We will be working with sequences retrieved from EMBL but you can also use EMBOSS with sequences in text files.

We will begin with the EMBL sequence for *Xenopus laevis* rhodopsin whose identifiers are L07770.1
You need to tell EMBOSS where to read the sequence(s) you want to analyse. EMBOSS can read sequences either from text files or directly from a sequence database.

## 2. Retrieving sequences from databases

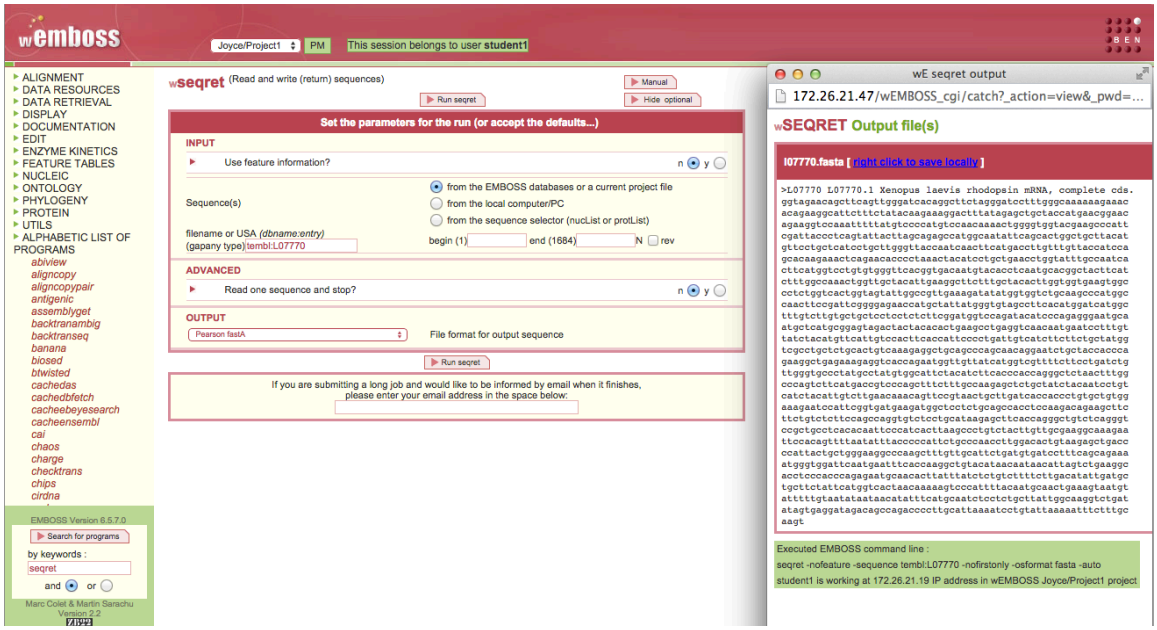
The EMBOSS program can read sequences from various databases, to enable you to do this is a tool called seqret.

In the search for programs by key word section, search for the program

**seqret**

Seqret reads in a sequence, and writes it out. The program should give an interface as shown below that allows you to get sequences from a database.

- Select the option | *from the EMBOSS databases or a current project file*
- Type in the filename as *embl:L07770*
- Click run seqret
- A page appears with the output as shown below




- You can change the output format that you get from the database.
- Change output to GCG and run seqret see what you get.
- Change output to EMBL new and run seqret see what you get
- Change output to Genebank and run seqret see what you get.
- You are now able to read sequences from the various databases available in EMBOSS

To know which databases are available to EMBOSS you can use the program showdb In the search for programs by key word section search for the program

**showdb**

Showdb displays information on configured databases

- Click run showdb
- A page appears that lists the available databases.

## 3. Sequence annotation

Sequence databases do not just contain sequences; they also contain a great deal of associated information (annotation) about the sequence entries. By default EMBOSS does not return all this information when you run seqret.To retrieve the full entry for a sequence in it's original database form you can use the utility entret.

In the search for programs by key word section search for the program

**entret**

Entret reads and writes (returns) flat-file entries

- Select the option | *from the EMBOSS databases or a current project file*
- Type in the filename as *embl:L07770*
- Click run entret
- A page appears with the output of annotations

There is a lot of information here. Near the bottom, just above the sequence itself is a list of features associated with the sequence. A feature is any defined region of the sequence that has a particular description associated with it. We can view a simple graphical overview of the sequence features using the utility showfeat:

In the search for programs by key word section search for the program

**showfeat**

Showfeat utility shows features of a sequence.

- Select the option | *from the EMBOSS databases or a current project file*
- Type in the filename as *embl:L07770*
- Click run showfeat
- A page appears with the output

## 4. Dot-plots

The most intuitive representation of the comparison between two sequences uses dot-plots. One sequence is represented on each axis and significant matching regions are distributed along diagonals in the matrix.

In the search for programs by key word section search for the program

**dottup**

Dottup generates DNA sequence dot plot

- Select the option | *from the EMBOSS databases or a current project file*
- Type in the filename as *embl:L07770*
- The second sequences filename *embl:L04692*
- Wordsize : 10
- Output graphic format: PNG
- Click run dottup
- A page appears with the output

## 5. Global alignment

A global alignment is one that compares the two sequences over their entire lengths, and is appropriate for comparing sequences that are expected to share similarity over the whole length. The alignment maximizes regions of similarity and minimizes gaps using the scoring matrices and gap parameters provided to the program. The EMBOSS program needle is an implementation of the Needleman-Wunsch [] algorithm for global alignment; the computation is rigorous and needle can be time consuming to run if the sequences are long.

In the search for programs by key word section search for the program

**needle**

Needle uses the Needleman-Wunsch global alignment of two sequences

- Select the option | *from the EMBOSS databases or a current project file*
- Type in the filename as *embl:L07770*
- The second sequences filename *embl:L04692*
- Gap opening penalty [10.0]:
- Gap extension penalty [0.5]
- Click run needle
- A page appears with the output alignment

## 6. Local alignment

A second comparison method, local alignment, searches for regions of local similarity and need not include the entire length of the sequences. Local alignment methods are very useful for scanning databases or when you do not know that the sequences are similar over their entire lengths. The EMBOSS program water is a rigorous implementation of the Smith Waterman algorithm for local alignments.

In the search for programs by key word section search for the program

**water**

Water program uses the Smith-Waterman local alignment of two sequences

- Select the option | *from the EMBOSS databases or a current project file*
- Type in the filename as *embl:L07770*
- The second sequences filename *embl:L04692*
- Gap opening penalty [10.0]:
- Gap extension penalty [0.5]
- Click run needle
- A page appears with the output alignment

## Protein analysis

EMBOSS applications that can be used to analyze protein sequences. Obviously, the pairwise sequence comparison methods illustrated in the previous chapter with nucleic acid sequences can also be used with protein sequences.

## 7. Identifying the ORF

In this section we'll show you some simple EMBOSS applications for translating your cDNA sequence into protein. First, we need to identify our open reading frame. We can get a rapid visual overview of the distribution of ORFs in the six frames of our sequence using the EMBOSS program plotorf.

In the search for programs by key word section search for the program

**plotorf**

Plotorf plots potential open reading frames

- Select the option | *from the EMBOSS databases or a current project file*
- Type in the filename as *embl:L07770*

- Click run plotorf
- You will see a graphical output that shows the potential open reading frames (ORF) in all six frames open reading frames (ORF) in all six frames

The longest ORF is in frame 2 from around position 100 to 1200. We will now identify the exact start and end points for our translation. To do this, we can use the EMBOSS program getorf.

In the search for programs by key word section search for the program

**getorf**

Getorf program finds and extracts open reading frames

- Select the option | *from the EMBOSS databases or a current project file*
- Type in the filename as *embl:L07770*
- Click run getorf
- You will various translations and the start and end point

## 8. Translating DNA to amino acid sequences

These allow you to specify a sub-region of your sequence; in this case we will ask transeq to translate only the part of *L07770* that we have identified as the coding region.

In the search for programs by key word section search for the program

**transeq**

Transeq program translates nucleic acid sequences

- Select the option | *from the EMBOSS databases or a current project file*
- Type in the filename as *embl:L07770*
- Begin 2
- End 1171
- Click run transeq
- The output should be the translated sequence
- On the result window, save the protein sequence by *right click to save locally.*
- Copy the sequence and paste it in a notepad page save the sequence in your computer by the name *L07770.pep*

## 9. Motifs

Motifs are functional units of proteins, these can be searched in protein sequences using the emboss program patmatmotifs.

In the search for programs by key word section search for the program

**patmatmotifs**

Patmatmotifs program search a motif database with a protein sequence

- Select the option | *from the local computer/PC*
- Choose the file you saved in the last step from your computer *L07770.pep*
- Click run patmatmotif
- Which motifs are present in your sequence..?
- *#Hint you should have at least three motifs*

## 10. Protein fingerprints

PRINTS is a database that defines functional protein families, identifying each domain by a number of short, particularly well conserved sequences. A full match to one of these "fingerprints" will match all the relevant short sequences in the correct order. The PRINTS database can be searched using the pscan program which is available within EMBOSS.

In the search for programs by key word section search for the program

**pscan**

Pscan scans protein sequence with fingerprints from the PRINTS database

- Select the option | *from the local computer/PC*
- Choose the file you saved in the last step from your computer *L07770.pep*
- Minimum number of elements per fingerprint [2]
- Maximum number of elements per fingerprint [20]
- Click run pscan
- Which protein families are present in your sequence..?
- *#Hint you should have at least four protein families*