BASIC **L**OCAL **A**LIGNMENT **S**EARCH **T**OOL (**BLAST**)

Using BLAST from Roseobase

Roseobase is a source of genome sequence information and analysis tool for marine Roseobacters, for more information please see: http://roseobase.org/

1. Go to www.roseobase.org/blast/APBio.html
2. Choose program (blastn, blastp, blastx, tblastn, tblastx) and database (organism) to search.

BLAST Programs

blastn: (Nucleotide BLAST) This program searches a nucleotide database using a nucleotide query.

blastp: (Protein BLAST) This program searches a protein database using a protein query.

blastx: Search protein database using a translated nucleotide query.

tblastn: Search translated nucleotide database using a protein query.

tblastx: search translated nucleotide database using a translated nucleotide query.

During this exercise we will be using **blastp.**

3. Enter your data as sequence in FASTA format (please see FASTA format description handout).
4. Paste sequence (or upload a file in FASTA format)

```
>RecA_E. coli K12_Accession P0A7G6
maidenkqkalaaalgqiekqfgkgsimrlgedrsmdvetistgslsldialgagglpmg
riveiygpessgkttltlqviaaaqregktcafidaehaldpiyarklgvdidnllcsqp
dtgeqaleicdalarsgavdvivvdsvaaltpkaeiegeigdshmglaarmmsqamrkla
gnlkqsntllifinqirmkigvmfgnpetttggnalkfyasvrldirrigavkegenvvg
setrvkvvknkiaapfkqaefqilygeginfygelvdlgvkekliekagawysykgekig
qgkanatawlkdnpetakeiekkvrelllsnpnstpdfsvddsegvaetnedf
```

5. Click on search

Checking the Results

You will be checking the E value and the % identities. An excellent match will have an E value = 0.0 When going from protein to protein (blastp) the closer the E value is to 0.0 the more likely the protein has similar structure and maybe the same function. E values =< e-6 and % identities >= 30% are typically considered as a homology (which means the same or closely related functional proteins).

Finding the name of the protein

To find the name of the protein we will be using the NCBI (National Center for Biotechnology Information) website www.ncbi.nlm.nih.gov .

1. On the NCBI home page click on the pull down menu next to search and choose protein.
1. Enter the tag number (e.g., ISM_08395) on the window next to the pull down menu.
2. Click **GO**

   The name of the protein will be the name that appears under the blue number (gene ID).

FASTA format description

A sequence in FASTA format consists of a single-line description, followed by lines of sequence data. The first character of the description line is a greater-than (">") symbol in the first column. All lines should be shorter than 80 charcters. An example sequence in FASTA format is:

```
>Name of the sequence
ctgcgagNcgcgcgatgatagMMM-
NNNnnnnncgcggcgagcatgtagcatgctagctgtcgcgagcactUUUURRRrrrrrrr
cggccgagatcaggcgatgcatgcgcagggagcagcgagcgacgagcacagcatgctagctagatgcatgctaVvv
vcgtaggcagc
cgccgagagacgatggagctgc
```

Sequences have to be represented in the standard IUB/IUPAC amino acid and nucleic acid codes, with these exceptions: lower-case letters are accepted and are mapped into upper-case; a single hyphen or dash can be used to represent a gap of indeterminate length; and in amino acid sequences, U and * are acceptable letters (see below). Before submitting a request, any numerical digits in the query sequence should either be removed or replaced by appropriate letter codes (e.g., N for unknown nucleic acid residue or X for unknown amino acid residue).

The nucleic acid codes supported are:

A --> adenosine	M --> A C (amino)
C --> cytidine	S --> G C (strong)
G --> guanine	W --> A T (weak)
T --> thymidine	B --> G T C
U --> uridine	D --> G A T
R --> G A (purine)	H --> A C T
Y --> T C (pyrimidine)	V --> G C A
K --> G T (keto)	N --> A G C T (any)
	- gap of indeterminate length

For those programs that use amino acid query sequences (BLASTP and TBLASTN), the accepted amino acid codes are:

A alanine	P praline
B aspartate or asparagine	Q glutamine
C cystine	R arginine
D aspartate	S serine
E glutamate	T threonine
F phenylalanine	U selenocysteine
G glycine	V valine
H histidine	W tryptophan
I isoleucine	Y tyrosine
K lysine	Z glutamate or glutamine
L leucine	X any
M methionine	* translation stop
N asparagine	- gap of indeterminate length