Microbial Bioinformatics

Introduction

These exercises are for you to learn how to use bioinformatics' tools to explore bacterial genomes, and complements the 'Microbiology of Safe Food' 2nd Edition (Wiley-Blackwell).

Table of Contents

If you have real problems with all this simply contact me at stephen.forsythe@ntu.ac.uk.

Exercise 1: Phylogenetic tree construction

Part 1 : ClustalW

Introduction

This exercise introduces you to multiple sequence alignments, whereby your unknown sequence is compared to a number of preselected sequences (retrieved using BLAST; Exercise 2), and drawing a relationship tree, which is more commonly known as a ‘phylogenetic tree’. The main program you will be using is ClustalW.

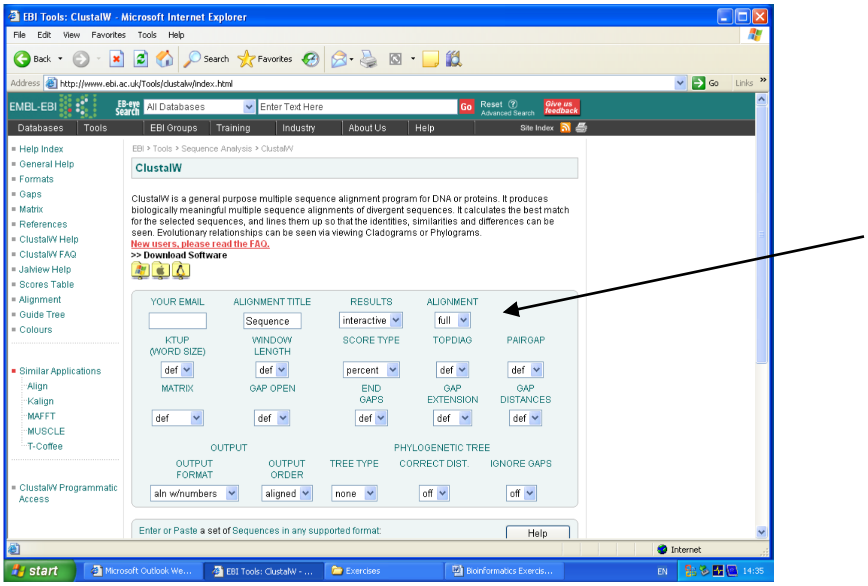
My suggestion is to open Internet Explore and have the ClustalW site and this guide open in different windows. You can do this by clicking on this link now and reducing it to the bottom of the screen until you need it.

ClustalW at http://www.ebi.ac.uk/clustalw/index.html.

This part shows you how to construct a relationship tree based on sequence similarity. You’ll use the online multiple sequence alignment program ClustalW to determine where your organism is on the ‘tree of life’.

Web address: http://www.ebi.ac.uk/clustalw/index.html.

Here is a screen print of the window that you will see.



Address: http://www.ebi.ac.uk/Tools/clustalw/index.html

ClustalW

ClustalW is a general purpose multiple sequence alignment program for DNA or proteins. It produces biologically meaningful multiple sequence alignments of divergent sequences. It calculates the best match for the selected sequences, and lines them up so that the identities, similarities and differences can be seen. Evolutionary relationships can be seen via viewing Cladograms or Phylograms.

New users, please read the FAQ.

>> Download Software

YOUR EMAIL	ALIGNMENT TITLE	RESULTS	ALIGNMENT	
	Sequence	interactive	full	
KTUP (WORD SIZE)	WINDOW LENGTH	SCORE TYPE	TOPDIAG	PAIRGAP
def	def	percent	def	def
MATRIX	GAP OPEN	END GAPS	GAP EXTENSION	GAP DISTANCES
def	def	def	def	def

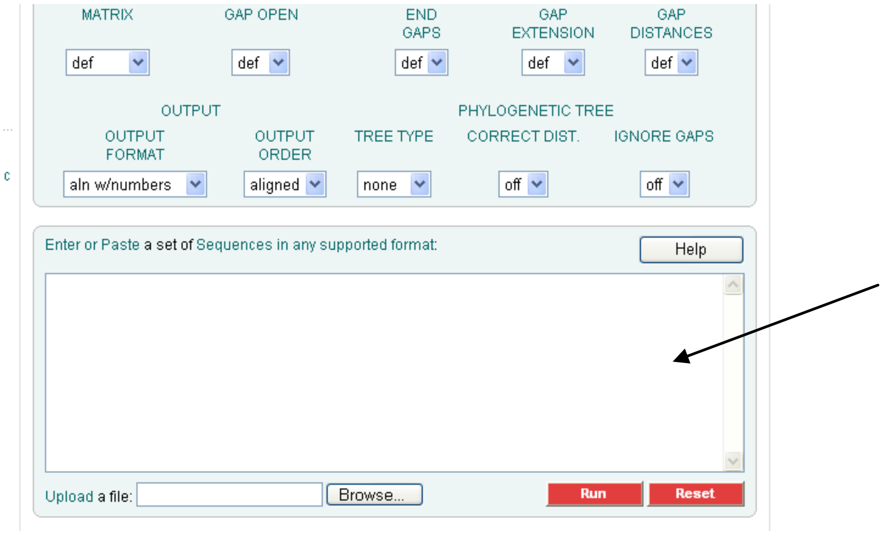
OUTPUT		PHYLOGENETIC TREE		
OUTPUT FORMAT	OUTPUT ORDER	TREE TYPE	CORRECT DIST.	IGNORE GAPS
aln w/numbers	aligned	none	off	off

Enter or Paste a set of Sequences in any supported format:

Ignore, this is an option – not a requirement

Leave all these dropdown menu options at the default settings.

You can enter your sequence data by cutting and pasting or upload a file.



MATRIX	GAP OPEN	END GAPS	GAP EXTENSION	GAP DISTANCES
def	def	def	def	def

OUTPUT		PHYLOGENETIC TREE		
OUTPUT FORMAT	OUTPUT ORDER	TREE TYPE	CORRECT DIST.	IGNORE GAPS
aln w/numbers	aligned	none	off	off

Enter or Paste a set of Sequences in any supported format:

Upload a file:

For this exercise we will use the cut and paste option.

You have three sets of sequences to analyse.

a) 16S ribosomal DNA sequences for a range of bacteria
b) 16S rDNA sequences for the lactic acid bacteria
c) Amino acid sequences for staphylococcal and streptococcal toxins.

Together you can make similar figures to Fig. 2.1, 2.20, and 2.23 in the book.

If you want to learn how to find similar sequences go to Exercise 2 for 16S rDNA sequences and Exercise 3 for toxin sequences.

Part 2 : 16S-18S rDNA sequence alignment, 'Tree of Life'

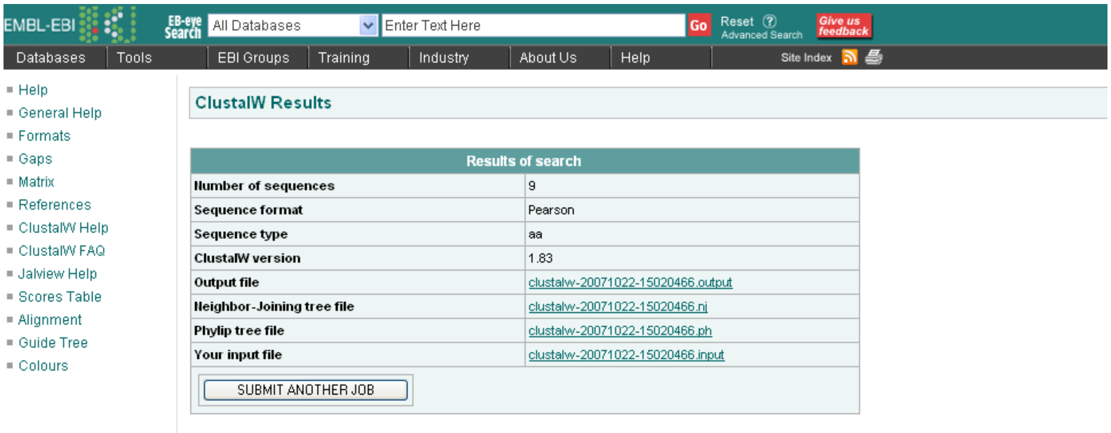
'All' you need to do initially is to cut and paste each set of sequences together into the ClustalW (http://www.ebi.ac.uk/clustalw/index.html) window.

Here is the LINK (http://www.foodmicrobe.com/Fig1.txt) for the 16S sequence files for this exercise. It is named 'Fig 1 16S seq file'.

If you wish you can also turn on the 'color' (yeap American) alignment option, but it's not important, especially if you're colo(u)r blind like me!

After a few minutes you'll get a new screen. There are two formats depending upon how many sequences you have compare.

If it looks like this:



ClustalW Results

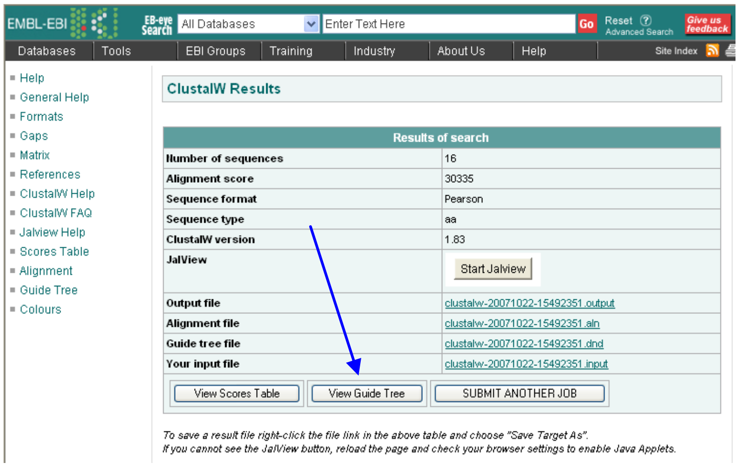
Results of search	
Number of sequences	9
Sequence format	Pearson
Sequence type	aa
ClustalW version	1.83
Output file	clustalw-20071022-15020466.output
Neighbor-Joining tree file	clustalw-20071022-15020466.nj
Phylip tree file	clustalw-20071022-15020466.ph
Your input file	clustalw-20071022-15020466.input

Scroll down and you will see that the sequences have been aligned. You will also find at the bottom of the page a ‘tree’ showing the relative similarities of the sequences. However please note that they were in fact different lengths and for more precise comparisons you will need to trim the sequences to comparable lengths.

Click on ‘Show as Phylogram tree’ for a more understandable diagram.

Background comment, from ClustalW index:
A phylogram is a branching diagram (tree) that is assumed to be an estimate of a phylogeny. The branch lengths are proportional to the amount of inferred evolutionary change. A cladogram is a branching diagram (tree) assumed to be an estimate of a phylogeny where the branches are of equal length. Therefore, cladograms show common ancestry, but do not indicate the amount of evolutionary "time" separating taxa. It is possible to see the tree distances by clicking on the diagram to get a menu of options. The options available allow you to do things like changing the colours of lines and fonts and showing the distances.

Alternatively, you may get a screen which does not have the tree at the bottom. This occurs when you have compared a large number of sequences. Instead click on ‘View Guide Tree’ and continue as per the example above. Scroll down, and click on ‘Show as Phylogram tree’.



ClustalW Results

Results of search	
Number of sequences	16
Alignment score	30335
Sequence format	Pearson
Sequence type	aa
ClustalW version	1.83
JalView	Start Jalview
Output file	clustalw-20071022-15492351.output
Alignment file	clustalw-20071022-15492351.aln
Guide tree file	clustalw-20071022-15492351.dnd
Your input file	clustalw-20071022-15492351.input

To save a result file right-click the file link in the above table and choose "Save Target As".
If you cannot see the JalView button, reload the page and check your browser settings to enable Java Applets.

Note how the label in the FASTA files is used for the phylogenetic tree. Also it only uses the letters after the '>' symbol until there is a space.

Does the tree make sense? You need to look up the organisms to answer this question.

Using this first 16S rDNA sequence file you will be constructing a 'Tree of Life' which is focused on micro-organisms associated with food, In order to better follow the organisation of the tree you need to know what the different organisms are.

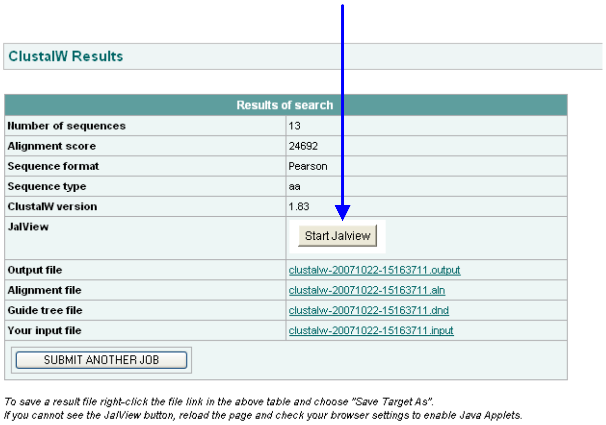
Bacillus cereus – Gram positive eubacterium, aerobic spore-former
Bacteroides fragilis – Gram negative eubacterium
Campylobacter jejuni – Gram negative eubacterium
Clostridium perfringens – Gram positive eubacterium, strict anaerobe, spore-former
Clostridum botulinum - Gram positive eubacterium, strict anaerobe, spore-former
Escherichia coli – Gram negative eubacterium
Lactobacillus acidophilus – Gram positive eubacterium
Listeria monocytogenes – Gram positive eubacterium
Micrococcus luteus – Gram positive eubacterium
Pseudomonas aeruginosa – Gram negative eubacterium
Pyrococcus furiosus – thermophilic archaea organism
Salmonella Enterica – Gram negative eubacterium
Staphylococcus aureus – Gram positive eubacterium

What does it tell you about the organisms you are not familiar with?

Traditionally cellular life is divided into prokaryotic and eukaryotic. However, what does the 'Tree of life' show you about 'bacteria'?

How can you find the differences between the sequences which have lead to the branches?

Looking at the alignments, can you regions of insertion or deletions? It may help you to click on ‘Start Jalview’



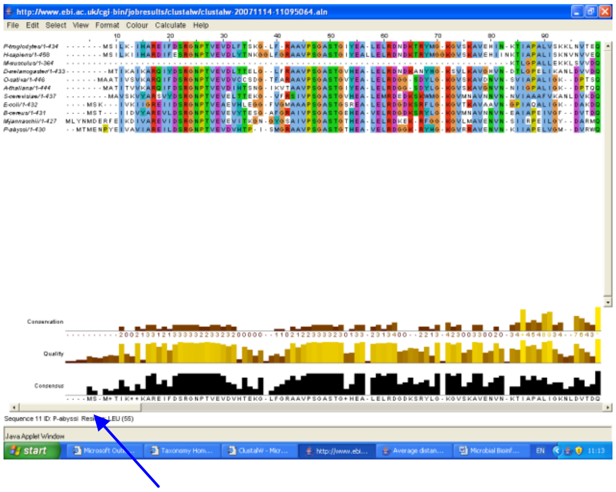
ClustalW Results

Results of search	
Number of sequences	13
Alignment score	24692
Sequence format	Pearson
Sequence type	aa
ClustalW version	1.83
JalView	Start Jalview
Output file	clustalw-20071022-15163711.output
Alignment file	clustalw-20071022-15163711.aln
Guide tree file	clustalw-20071022-15163711.dnd
Your input file	clustalw-20071022-15163711.input

SUBMIT ANOTHER JOB

To save a result file right-click the file link in the above table and choose "Save Target As".
If you cannot see the JalView button, reload the page and check your browser settings to enable Java Applets.

You’ll get a very colourful window:



Use the slider to scroll along, and you will see the regions of similarity, as well as insertions/deletions.

Printing the tree from ClustalW is not easy (please let me know if you find a simpler solution than mine). I use the ‘PrtScrn’button and then paste into Word™. For the book I used text boxes in Word™ to add the addition labels.

Part 3: Lactic acid bacteria phylogenetic tree

Just as before you need to cut and paste the set of sequences into the ClustalW window at http://www.ebi.ac.uk/clustalw/index.html.

Here is the LINK for the lactic acid bacteria 16S rDNA sequence files; http://www.foodmicrobe.com/Fig2.txt .

Hopefully by reading through the book (Section 3.7.1), and referring to the figure you've generated you'll see that this collection of organisms have been named according to their ability to produce lactic acid, but otherwise they are fairly divergent. In fact the *Bifidobacterium* genus is a considerable distance from all the others. It may look similar to *Lactobacillus* down the microscope but its genetic history is very different.

Part 4: Amino acid sequence alignment

Just as there is a single letter code for the nucleotides in DNA, so there is also one for amino acids. This is great as the sequence of amino acids in a protein can be represented as a string of single letters and compared to another just as we've already done for DNA sequences. The single letter codes are given in the Appendix.

Just as before 'all' you need to do initially is to cut and paste the set of sequences into the ClustalW window; http://www.ebi.ac.uk/Tools/clustalw2/index.html . Here is the LINK for the staphylococcal-streptococcal toxins for this exercise; http://www.foodmicrobe.com/Fig3.txt .

You'll see although the sequences this time are amino acids, the results is again a phylogenetic tree based on the similarities between each file.

Confession: I did not know about the link between staphylococcal and streptococcal toxins when I first made the tree on p99. As I was preparing the chapter I wanted to put a variety of topics in the bioinformatics/phylogeny section as tasters of what can be done. I already had two examples based on ribosomal DNA (rDNA) sequences ('Tree of Life' and 'lactic acid bacteria') and wanted a protein example to complement them. The staphylococcal toxin (SEA etc) group came to mind and not being a specialist in staphylococcal toxins I thought it would be interesting to collate their sequences and see what the result was. First I used the terms 'staphylococcal' and 'SEA' PubMed search for the protein sequences (Exercise 2), and having made a core of sequences I did BLAST searching (Exercise 3) to see if there were any staphylococcal sequences I'd overlooked, and that's when I came across the matches with streptococcal toxins (!). So I collated those as well and then made the phylogenetic tree with ClustalW. The result was Figure 2.21 in the book. Pure serendipity.

What you should know by the end of Exercise 1:

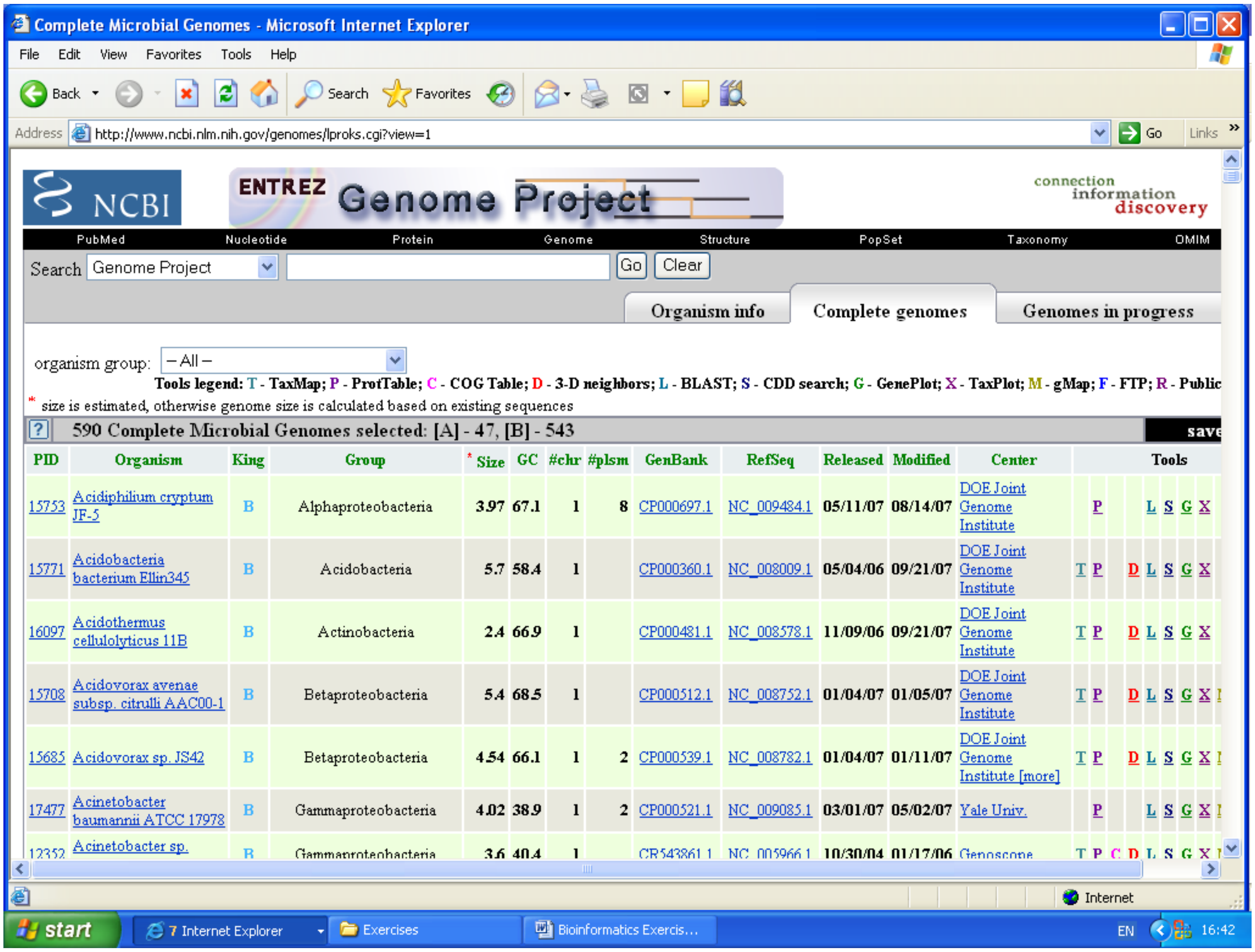
Difference between BLAST and ClustalW
How to construct a phylogenetic tree, not cladogram
How to look for sequence differences
How to identify regions of interest that may relate to taxonomy.

Exercise 2: Finding 16S rRNA sequences

Part 1: Using NCBI for organisms which have been sequenced

Introduction

(a) Go to the NCBI Entrez genome site at Complete Microbial Genomes;
http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi


Complete Microbial Genomes - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Address
Go
Links
NCBI
ENTREZ Genome Project
connection information discovery
PubMed Nucleotide Protein Genome Structure PopSet Taxonomy OMIM
Search
Go
Clear
Organism info
Complete genomes
Genomes in progress
save
Internet
start
7 Internet Explorer
Exercises
Bioinformatics Exercis...
EN
16:42

Address: http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi?view=1

Search: Genome Project

organism group: – All –

Tools legend: T - TaxMap; P - ProtTable; C - COG Table; D - 3-D neighbors; L - BLAST; S - CDD search; G - GenePlot; X - TaxPlot; M - gMap; F - FTP; R - Public

* size is estimated, otherwise genome size is calculated based on existing sequences

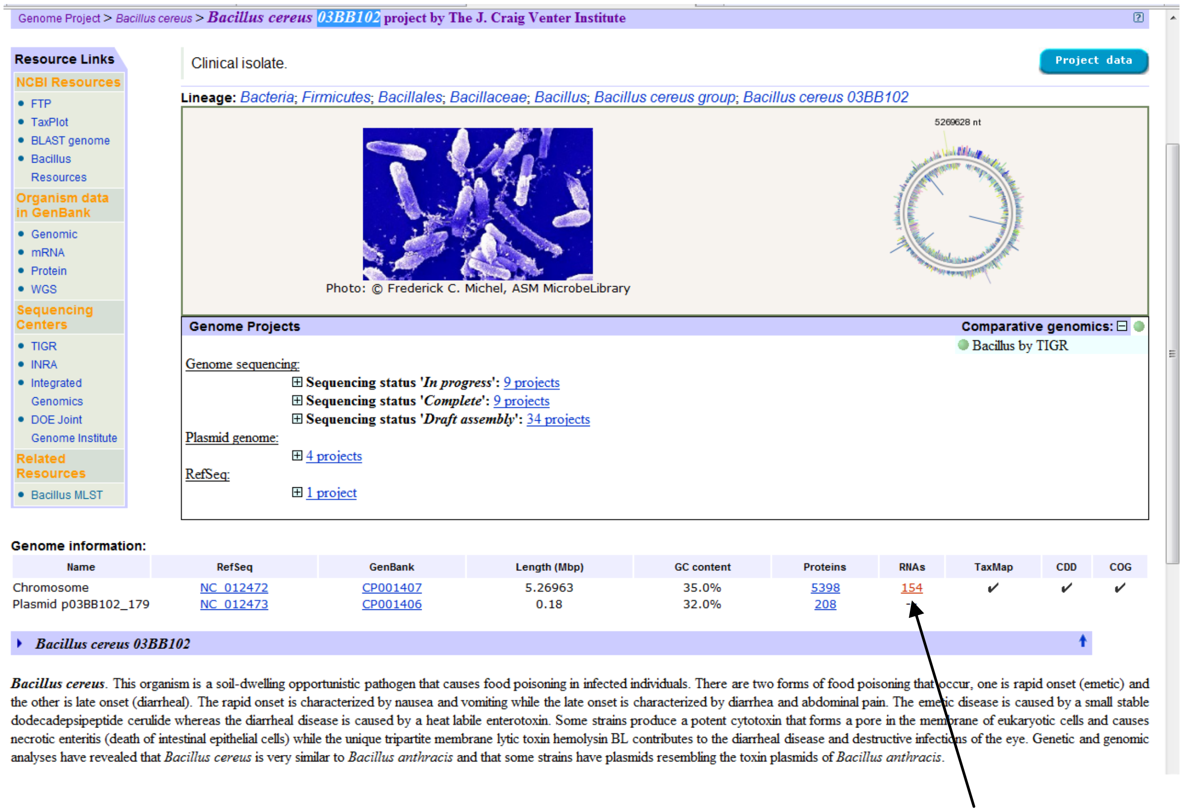
590 Complete Microbial Genomes selected: [A] - 47, [B] - 543

PID	Organism	King	Group	* Size	GC	#chr	#plsm	GenBank	RefSeq	Released	Modified	Center	Tools
15753	Acidiphilium cryptum JF-5	B	Alphaproteobacteria	3.97	67.1	1	8	CP000697.1	NC_009484.1	05/11/07	08/14/07	DOE Joint Genome Institute	P L S G X
15771	Acidobacteria bacterium Ellin345	B	Acidobacteria	5.7	58.4	1		CP000360.1	NC_008009.1	05/04/06	09/21/07	DOE Joint Genome Institute	T P D L S G X
16097	Acidothermus cellulolyticus 11B	B	Actinobacteria	2.4	66.9	1		CP000481.1	NC_008578.1	11/09/06	09/21/07	DOE Joint Genome Institute	T P D L S G X
15708	Acidovorax avenae subsp. citrulli AAC00-1	B	Betaproteobacteria	5.4	68.5	1		CP000512.1	NC_008752.1	01/04/07	01/05/07	DOE Joint Genome Institute	T P D L S G X
15685	Acidovorax sp. JS42	B	Betaproteobacteria	4.54	66.1	1	2	CP000539.1	NC_008782.1	01/04/07	01/11/07	DOE Joint Genome Institute [more]	T P D L S G X
17477	Acinetobacter baumannii ATCC 17978	B	Gammaproteobacteria	4.02	38.9	1	2	CP000521.1	NC_009085.1	03/01/07	05/02/07	Yale Univ.	P L S G X
12352	Acinetobacter sp.	B	Gammaproteobacteria	3.6	40.4	1		CR543861.1	NC_005966.1	10/30/04	01/17/06	Genoscope	T P C D L S G X

(b) Scroll down to the organism of choice, in this case *Bacillus cereus* 03BB102.

Click on the organism name.

You will get a long list of *B. cereus* genome projects, and towards the bottom:



Genome Project > *Bacillus cereus* > ***Bacillus cereus* 03BB102 project by The J. Craig Venter Institute**

Clinical isolate.

Lineage: *Bacteria*; *Firmicutes*; *Bacillales*; *Bacillaceae*; *Bacillus*; *Bacillus cereus group*; *Bacillus cereus 03BB102*

Photo: © Frederick C. Michel, ASM MicrobeLibrary

Genome Projects — **Comparative genomics:** ⊟ ● Bacillus by TIGR

Genome sequencing:

⊞ **Sequencing status '*In progress*':** 9 projects

⊞ **Sequencing status '*Complete*':** 9 projects

⊞ **Sequencing status '*Draft assembly*':** 34 projects

Plasmid genome:

⊞ 4 projects

RefSeq:

⊞ 1 project

Genome information:

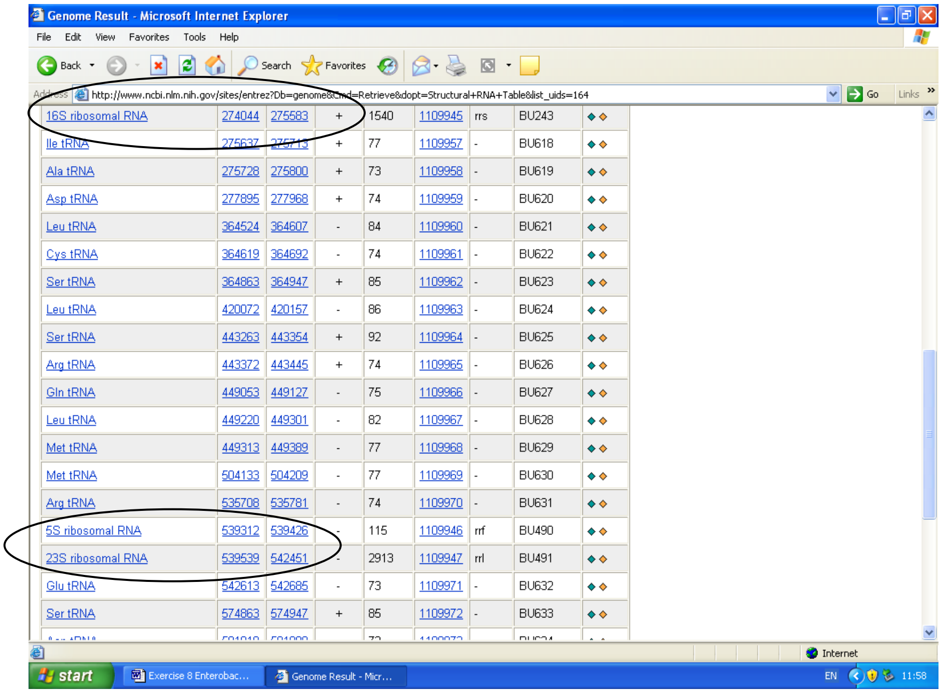
Name	RefSeq	GenBank	Length (Mbp)	GC content	Proteins	RNAs	TaxMap	CDD	COG
Chromosome	NC_012472	CP001407	5.26963	35.0%	5398	154	✓	✓	✓
Plasmid p03BB102_179	NC_012473	CP001406	0.18	32.0%	208	-			

Bacillus cereus 03BB102

Bacillus cereus. This organism is a soil-dwelling opportunistic pathogen that causes food poisoning in infected individuals. There are two forms of food poisoning that occur, one is rapid onset (emetic) and the other is late onset (diarrheal). The rapid onset is characterized by nausea and vomiting while the late onset is characterized by diarrhea and abdominal pain. The emetic disease is caused by a small stable dodecadepsipeptide cerulide whereas the diarrheal disease is caused by a heat labile enterotoxin. Some strains produce a potent cytotoxin that forms a pore in the membrane of eukaryotic cells and causes necrotic enteritis (death of intestinal epithelial cells) while the unique tripartite membrane lytic toxin hemolysin BL contributes to the diarrheal disease and destructive infections of the eye. Genetic and genomic analyses have revealed that *Bacillus cereus* is very similar to *Bacillus anthracis* and that some strains have plasmids resembling the toxin plasmids of *Bacillus anthracis*.

To find the ribosomal genes (23S, 16S, 5S) encoded on a genome, click on the RNAs link.

Another window will open up and you can see a list of the relevant genes to download.



Genome Result - Microsoft Internet Explorer

http://www.ncbi.nlm.nih.gov/sites/entrez?Db=genome&Cmd=Retrieve&dopt=Structural+RNA+Table&list_uids=164

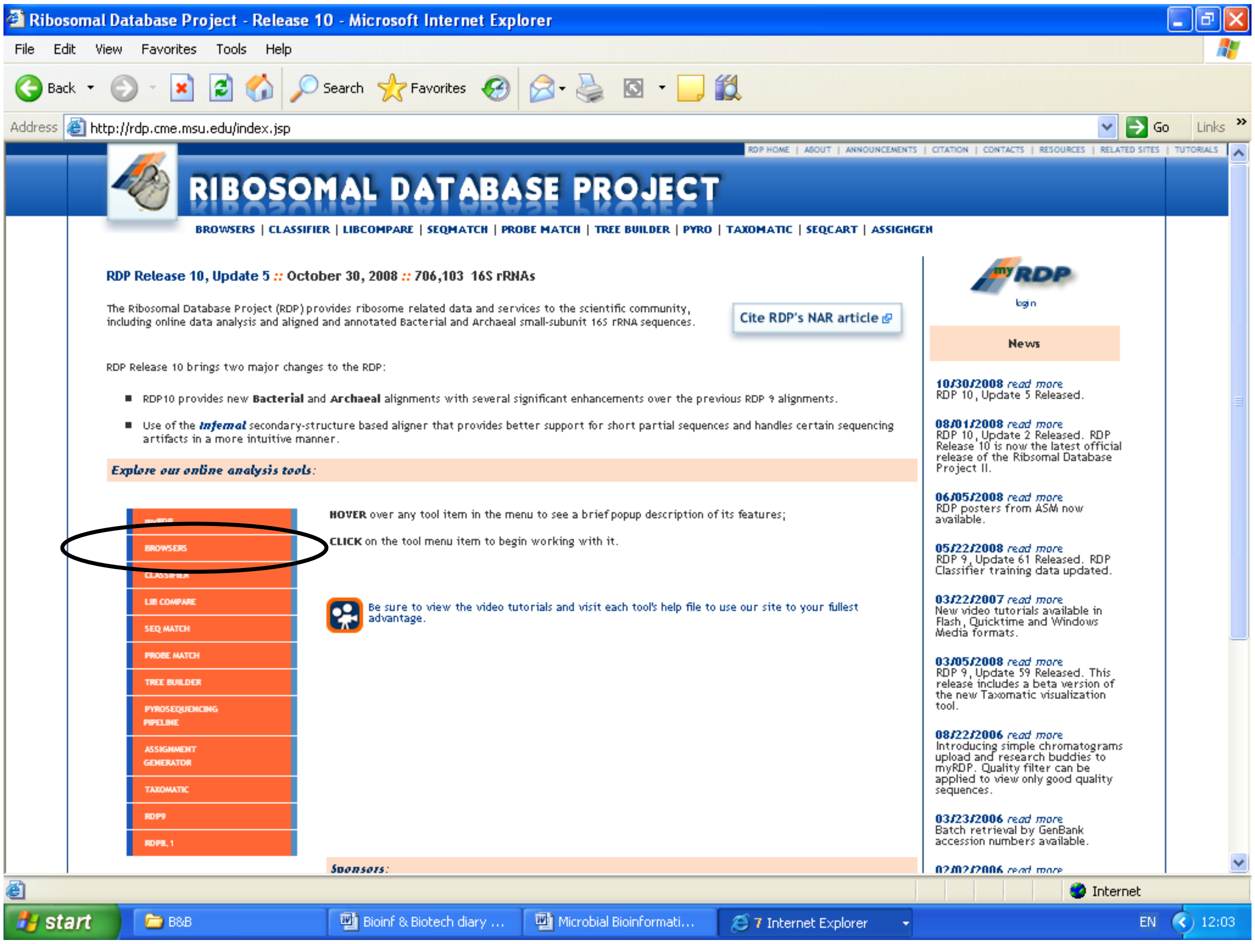
16S ribosomal RNA	274044	275583	+	1540	1109945	rrs	BU243	◆◇
Ile tRNA	275637	275713	+	77	1109957	-	BU618	◆◇
Ala tRNA	275728	275800	+	73	1109958	-	BU619	◆◇
Asp tRNA	277895	277968	+	74	1109959	-	BU620	◆◇
Leu tRNA	364524	364607	-	84	1109960	-	BU621	◆◇
Cys tRNA	364619	364692	-	74	1109961	-	BU622	◆◇
Ser tRNA	364863	364947	+	85	1109962	-	BU623	◆◇
Leu tRNA	420072	420157	-	86	1109963	-	BU624	◆◇
Ser tRNA	443263	443354	+	92	1109964	-	BU625	◆◇
Arg tRNA	443372	443445	+	74	1109965	-	BU626	◆◇
Gln tRNA	449053	449127	-	75	1109966	-	BU627	◆◇
Leu tRNA	449220	449301	-	82	1109967	-	BU628	◆◇
Met tRNA	449313	449389	-	77	1109968	-	BU629	◆◇
Met tRNA	504133	504209	-	77	1109969	-	BU630	◆◇
Arg tRNA	535708	535781	-	74	1109970	-	BU631	◆◇
5S ribosomal RNA	539312	539426	-	115	1109946	rrf	BU490	◆◇
23S ribosomal RNA	539539	542451	-	2913	1109947	rrl	BU491	◆◇
Glu tRNA	542613	542685	-	73	1109971	-	BU632	◆◇
Ser tRNA	574863	574947	+	85	1109972	-	BU633	◆◇

Tip: Not all sequences deposited are in the same orientation. You may find you have a collection of 16S sequences which do not make a sensible tree, yet if you BLAST them you always get the right organism. It is possible that you have downloaded the reverse complement which will mess things up when using ClustalW as it compares exactly what you have submitted, whereas BLAST will compare in the three reading frames of the top strand and the three reading frames of the bottom strand. Use the program at http://arep.med.harvard.edu/labgc/adnan/projects/Utilities/revcomp.html to generate reverse complement sequence and retry making the phylogram. This can occur with other DNA sequence files.

Part 2: Using RDPII

Go to the Ribosomal Database Project II at http://rdp.cme.msu.edu/index.jsp .

Below is a screen shot of what you should get:



Ribosomal Database Project - Release 10 - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://rdp.cme.msu.edu/index.jsp Go Links
RDP HOME | ABOUT | ANNOUNCEMENTS | CITATION | CONTACTS | RESOURCES | RELATED SITES | TUTORIALS
RIBOSOMAL DATABASE PROJECT
BROWSERS | CLASSIFIER | LIBCOMPARE | SEQMATCH | PROBE MATCH | TREE BUILDER | PYRO | TAXOMATIC | SEQCART | ASSIGNGEN
Cite RDP's NAR article
myRDP
login
myRDP
BROWSERS
CLASSIFIER
LIB COMPARE
SEQ MATCH
PROBE MATCH
TREE BUILDER
PYROSEQUENCING PIPELINE
ASSIGNMENT GENERATOR
TAXOMATIC
RDP9
RDPII. 1
Sponsors:
Internet
start
B&B
Bioinf & Biotech diary ...
Microbial Bioinformati...
7 Internet Explorer
EN
12:03

RDP Release 10, Update 5 :: October 30, 2008 :: 706,103 16S rRNAs

The Ribosomal Database Project (RDP) provides ribosome related data and services to the scientific community, including online data analysis and aligned and annotated Bacterial and Archaeal small-subunit 16S rRNA sequences.

RDP Release 10 brings two major changes to the RDP:

- RDP10 provides new **Bacterial** and **Archaeal** alignments with several significant enhancements over the previous RDP 9 alignments.
- Use of the ***Infernal*** secondary-structure based aligner that provides better support for short partial sequences and handles certain sequencing artifacts in a more intuitive manner.

Explore our online analysis tools:

HOVER over any tool item in the menu to see a brief popup description of its features;

CLICK on the tool menu item to begin working with it.

Be sure to view the video tutorials and visit each tool's help file to use our site to your fullest advantage.

News

10/30/2008 *read more*
RDP 10, Update 5 Released.

08/01/2008 *read more*
RDP 10, Update 2 Released. RDP Release 10 is now the latest official release of the Ribsomal Database Project II.

06/05/2008 *read more*
RDP posters from ASM now available.

05/22/2008 *read more*
RDP 9, Update 61 Released. RDP Classifier training data updated.

03/22/2007 *read more*
New video tutorials available in Flash, Quicktime and Windows Media formats.

03/05/2008 *read more*
RDP 9, Update 59 Released. This release includes a beta version of the new Taxomatic visualization tool.

08/22/2006 *read more*
Introducing simple chromatograms upload and research buddies to myRDP. Quality filter can be applied to view only good quality sequences.

03/23/2006 *read more*
Batch retrieval by GenBank accession numbers available.

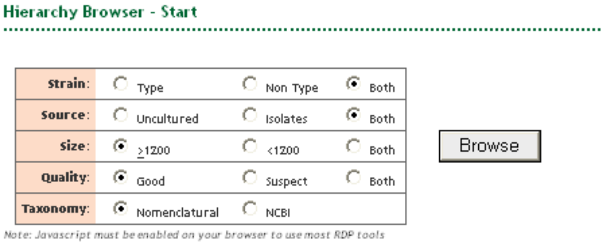
02/02/2006 *read more*

There's lots, and lots of tools you can use here, but for now we will keep to retrieving 16S rRNA sequences for specific organisms.

This example will look for some key bacteria which cause foodborne illness:
Salmonella Enterica and *S.* Typhimurium
Escherichia coli O157
Staphylococcus aureus
Bacillus cereus
Clostridium perfringens and *Clostridium botulinum*
Campylobacter jejuni

Click on the ‘Browser’ (circled above) to get the following page:

Choose the following:



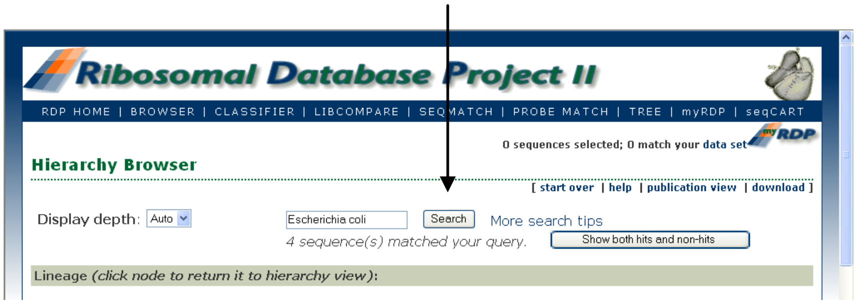
Hierarchy Browser - Start

Strain:	○ Type	○ Non Type	● Both
Source:	○ Uncultured	○ Isolates	● Both
Size:	● ≥1200	○ <1200	○ Both
Quality:	● Good	○ Suspect	○ Both
Taxonomy:	● Nomenclatural	○ NCBI	

Note: Javascript must be enabled on your browser to use most RDP tools

and then click on ‘Browse’

Enter the organism of choice in the search box.



Ribosomal Database Project II
RDP HOME | BROWSER | CLASSIFIER | LIBCOMPARE | SEQMATCH | PROBE MATCH | TREE | myRDP | seqCART
myRDP
[start over | help | publication view | download]
Search
More search tips
Show both hits and non-hits

0 sequences selected; 0 match your data set

Hierarchy Browser

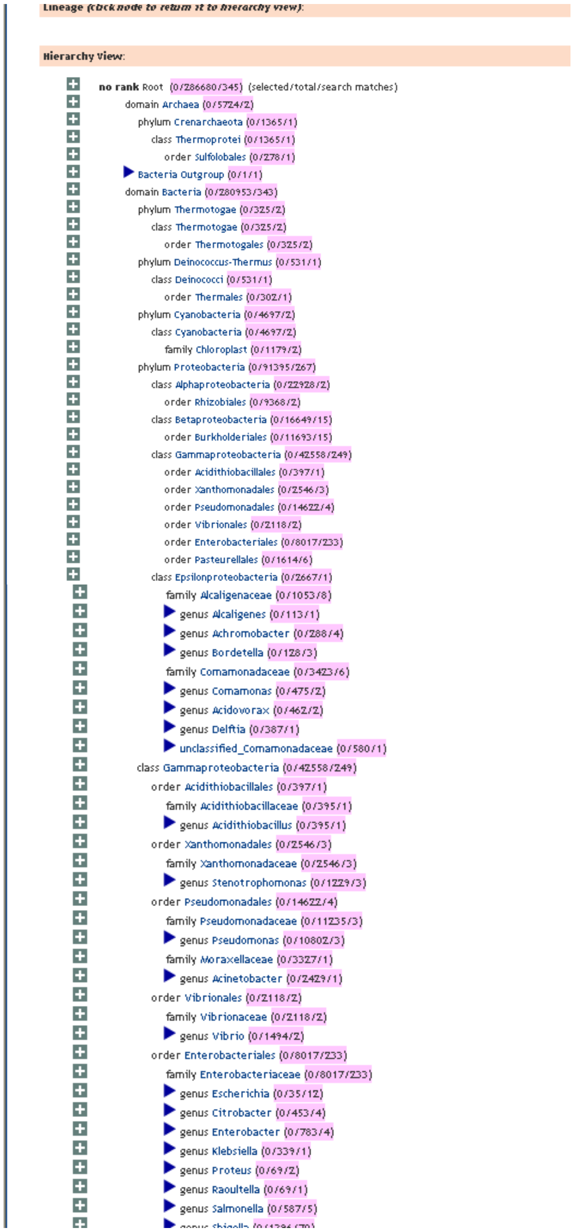
Display depth: Auto

Escherichia coli

4 sequence(s) matched your query.

Lineage (*click node to return it to hierarchy view*):

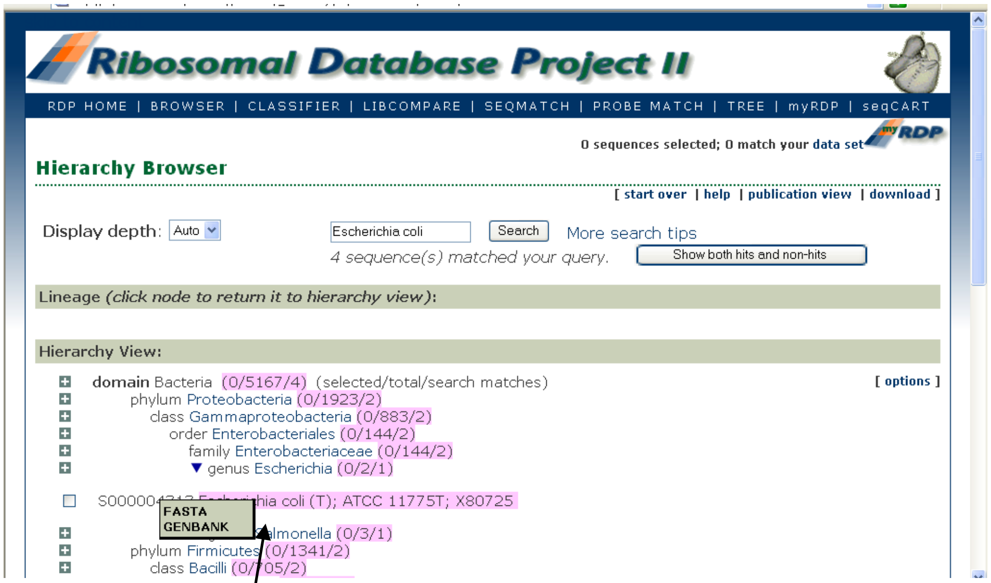
You will need to ‘drill’ down via the ‘Proteobacteria’, and then ‘*Enterobacteriaceae’* to find *Escherichia coli*.



Lineage (click node to return it to hierarchy view):

Hierarchy View:

```
no rank Root (0/286680/345) (selected/total/search matches)
    domain Archaea (0/5724/2)
      phylum Crenarchaeota (0/1365/1)
        class Thermoprotei (0/1365/1)
          order Sulfolobales (0/278/1)
    ▶ Bacteria Outgroup (0/1/1)
    domain Bacteria (0/280953/343)
      phylum Thermotogae (0/325/2)
        class Thermotogae (0/325/2)
          order Thermotogales (0/325/2)
      phylum Deinococcus-Thermus (0/531/1)
        class Deinococci (0/531/1)
          order Thermales (0/302/1)
      phylum Cyanobacteria (0/4697/2)
        class Cyanobacteria (0/4697/2)
          family Chloroplast (0/1179/2)
      phylum Proteobacteria (0/91395/267)
        class Alphaproteobacteria (0/22928/2)
          order Rhizobiales (0/9368/2)
        class Betaproteobacteria (0/16649/15)
          order Burkholderiales (0/11693/15)
        class Gammaproteobacteria (0/42558/249)
          order Acidithiobacillales (0/397/1)
          order Xanthomonadales (0/2546/3)
          order Pseudomonadales (0/14622/4)
          order Vibrionales (0/2118/2)
          order Enterobacteriales (0/8017/233)
          order Pasteurellales (0/1614/6)
        class Epsilonproteobacteria (0/2667/1)
          family Alcaligenaceae (0/1053/8)
          ▶ genus Alcaligenes (0/113/1)
          ▶ genus Achromobacter (0/288/4)
          ▶ genus Bordetella (0/128/3)
          family Comamonadaceae (0/3423/6)
          ▶ genus Comamonas (0/475/2)
          ▶ genus Acidovorax (0/462/2)
          ▶ genus Delftia (0/387/1)
          ▶ unclassified_Comamonadaceae (0/580/1)
      class Gammaproteobacteria (0/42558/249)
        order Acidithiobacillales (0/397/1)
          family Acidithiobacillaceae (0/395/1)
          ▶ genus Acidithiobacillus (0/395/1)
        order Xanthomonadales (0/2546/3)
          family Xanthomonadaceae (0/2546/3)
          ▶ genus Stenotrophomonas (0/1229/3)
        order Pseudomonadales (0/14622/4)
          family Pseudomonadaceae (0/11235/3)
          ▶ genus Pseudomonas (0/10802/3)
          family Moraxellaceae (0/3327/1)
          ▶ genus Acinetobacter (0/2429/1)
        order Vibrionales (0/2118/2)
          family Vibrionaceae (0/2118/2)
          ▶ genus Vibrio (0/1494/2)
        order Enterobacteriales (0/8017/233)
          family Enterobacteriaceae (0/8017/233)
          ▶ genus Escherichia (0/35/12)
          ▶ genus Citrobacter (0/453/4)
          ▶ genus Enterobacter (0/783/4)
          ▶ genus Klebsiella (0/339/1)
          ▶ genus Proteus (0/69/2)
          ▶ genus Raoultella (0/69/1)
          ▶ genus Salmonella (0/587/5)
          ▶ genus Shigella (0/1296/79)
```


Ribosomal Database Project II
RDP HOME | BROWSER | CLASSIFIER | LIBCOMPARE | SEQMATCH | PROBE MATCH | TREE | myRDP | seqCART
myRDP
[start over | help | publication view | download]
Search
More search tips
Show both hits and non-hits
[options]
FASTA
GENBANK

0 sequences selected; 0 match your data set

Hierarchy Browser

Display depth: Auto

Escherichia coli

4 sequence(s) matched your query.

Lineage (*click node to return it to hierarchy view*):

Hierarchy View:

domain Bacteria (0/5167/4) (selected/total/search matches)
phylum Proteobacteria (0/1923/2)
class Gammaproteobacteria (0/883/2)
order Enterobacteriales (0/144/2)
family Enterobacteriaceae (0/144/2)
genus Escherichia (0/2/1)

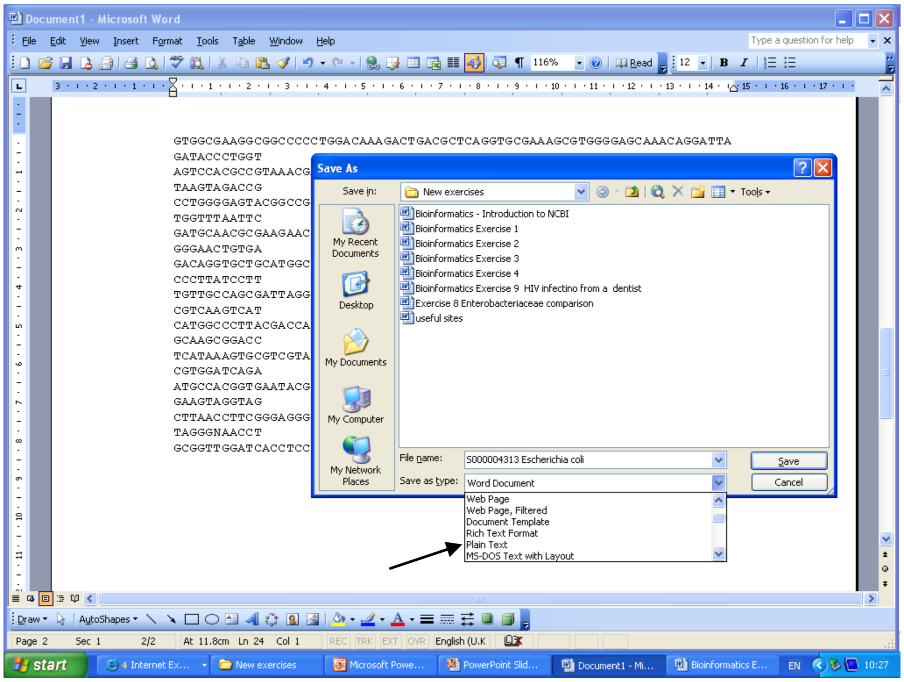
□ S000004313 Escherichia coli (T); ATCC 11775T; X80725

Salmonella (0/3/1)
phylum Firmicutes (0/1341/2)
class Bacilli (0/705/2)

Move your cursor over the selected organism number on the left-hand side, and a temporary window will open arrowed above. This is a bit tricky as it is not apparent that these embedded links exist.

Click on the FASTA option as this will be more appropriate for you.

Another window will open with the sequence in FASTA format. Copy and paste to a Word file (use font size 8 to save space), and save in PLAIN TEXT format using the ‘Save as’ and pull down options.



```
GTGGCGAAGGCGGCCCCCTGGACAAAGACTGACGCTCAGGTGCGAAAGCGTGGGGAGCAAACAGGATTA
GATACCCTGGT
AGTCCACGCCGTAAACG
TAAGTAGACCG
CCTGGGGAGTACGGCCG
TGGTTTAATTC
GATGCAACGCGAAGAAC
GGGAACTGTGA
GACAGGTGCTGCATGGC
CCCTTATCCTT
TGTTGCCAGCGATTAGG
CGTCAAGTCAT
CATGGCCCTTACGACCA
GCAAGCGGACC
TCATAAAGTGCGTCGTA
CGTGGATCAGA
ATGCCACGGTGAATACG
GAAGTAGGTAG
CTTAACCTTCGGGAGGG
TAGGGNAACCT
GCGGTTGGATCACCTCC
```

Bioinformatics - Introduction to NCBI
Bioinformatics Exercise 1
Bioinformatics Exercise 2
Bioinformatics Exercise 3
Bioinformatics Exercise 4
Bioinformatics Exercise 9 HIV infectino from a dentist
Exercise 8 Enterobacteriaceae comparison
useful sites

File name: S000004313 Escherichia coli

Save as type: Word Document

Web Page
Web Page, Filtered
Document Template
Rich Text Format
Plain Text
MS-DOS Text with Layout

Page 2 Sec 1 2/2 At 11.8cm Ln 24 Col 1 REC TRK EXT OVR English (U.K

You need to ensure the label line is useful.

For example:
>Escherichia-coli (type strain)
ACTACATTCGA…etc

This is okay as the 'Escherichia-coli' will appear as the label in the resulting tree when you do the phylogenetic analysis. The '(type strain)' will not appear as part of the label as there is a space after 'coli'.

I **strongly** recommend that you put a list together of the organism name, strain and sequence accession number. If you get into this habit, it will save you lots of wasted time repeating searches in the future as you can just enter the accession number into the NCBI website. This is a lot easier than a general search for each organism.

If you're not sure what the accession number is, look at the line below which is the first line for the *E. coli* FASTA file:

S000004313 Escherichia coli (T); ATCC 11775T; X80725

Background comment: The culture collection code ATCC 11775, which refers tot eh American Type Culture Collection (www.atcc.org). 'T' means it is the type strain, and has therefore been well studied.

The only numbers left are S000004313, and X80725.

Go to NCBI web site (http://www.ncbi.nlm.nih.gov/) opt for the 'CoreNucleotide' in the 'Search box, and enter each value in the search facility and see what you get. One makes sense, the other doesn't.
The alternative method is to go back to the Heirarchy Browser and click on Genbank instead of FASTA.

Also remember that these exercises are to introduce you to the basic tools in bioinformatics, they are applicable to eukaryotic as well as prokaryotic systems. You may also find them of use in other projects. All the best!

What you should know by the end of Exercise 7, Part 2:

How to use ClustalW for multisequence alignment analysis
How to construct labelled phylogenetic trees
How to interpret features of an 'unknown' organism, from similarity comparison.

Exercise 3: BLAST searching

Introduction

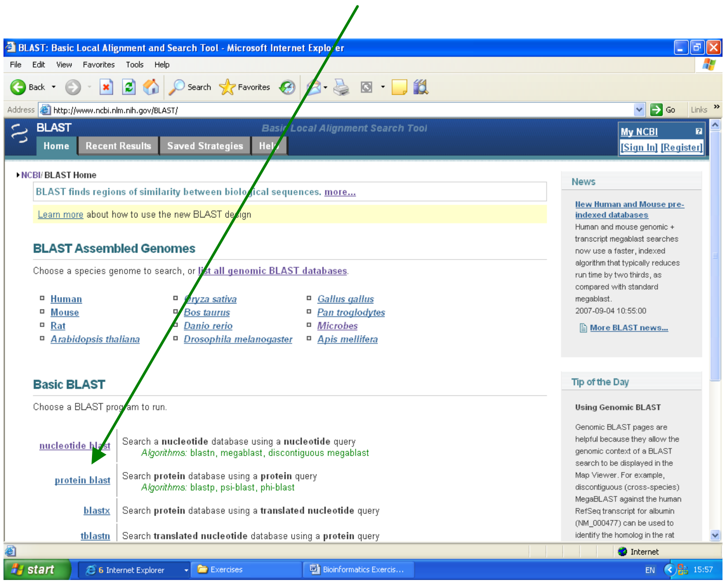
The aim of this exercise is to identify the amino acid sequence for an unknown protein using BLASTP:

```
MNRRDFIKNTAIASAASVAGLSVPSSMLGAQEEDWKWDKAVCRFCGTGCGIMIARKDGKIVATKGDPAAP
VNRGLNCIKGYFNAKIMYGEDRLVMPLLRMNEKGEFDKKGKFQQVSWQRAFDEMEKQFKKAYNELGVTGI
GIFGSGQYTIQEGYAALKLAKAGFRTNNIDPNARHCMASAVVGFMQTFGVDEPSGCYDDIELTDTIITWG
ANMAEMHPILWSRVSDRKLSNLDKVKVVNLSTFSNRTSNIADIEIIFKPNTDLAIWNYIAREIVYNHPEA
MDMKFIKDHCVFATGYADIGYGMRNNPNHPKFKESEKDTVEKENVITLDDEEATSLSYLGVKAGDKFEMK
HQGVADKNWEISFDEFKKGLAPYTLEYTARVAKGDDNESLEDFKKKLQELANLYIEKNRKVVSFWTMGFN
QHTRGSWVNEQAYMVHFLLGKQAKPGSGAFSLTGQPSACGTAREVGTFSHRLPADMVVANPKHREISEKI
WKVPAKTINPKPGSPYLNIMRDLEDGKIKFAWVQVNNPWQNTANANHWIAAAREMDNFIVVSDCYPGISA
KVADLILPSAMIYEKWGAYGNAERRTQHWKQQVLPVGAAMSDTWQILEFAKRFKLKEVWKEQKVDNKLTL
PSVLEEAKAMGYSEDDTLFDVLFANKEAKSFNPNDAIAKGFDNTDVKGDERKIQGSDGKEFTGYGFFVQK
YLWEEYRKFGLGHGHDLADFDTYHKVRGLRWPVVNGKETQWRFNTKFDYYAKKAAPNSDFAFYGDFNKML
TNGDLIAPKDEKEHSIKNKAKIFFRPFMKAPERPSKEYPFWLATGRVLEHWHSGTMTMRVPELYRAVPEA
LCYMSEKDGEKLGLNQGDLVWVESRRGKVKARVDMRGRNKPPVGLVYVPWFDENVYINKVTLDATCPLSK
QTDFKKCAVKIYKA
```

Tip: Look up what the differences are between the 6 BLAST programs; http://blast.ncbi.nlm.nih.gov/Blast.cgi.

You will find the Appendix at the end of the document useful for looking up the single letter codes for the amino acids.

Use the NCBI BLAST site at http://www.ncbi.nlm.nih.gov/BLAST/
Click on the Protein-protein BLAST (blastp)


BLAST: Basic Local Alignment and Search Tool - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Address
Go
Links
BLAST
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
My NCBI
[Sign In] [Register]
Internet
start
6 Internet Explorer
Exercises
Bioinformatics Exercis...
EN
15:57

http://www.ncbi.nlm.nih.gov/BLAST/

NCBI/ BLAST Home

BLAST finds regions of similarity between biological sequences. more...

Learn more about how to use the new BLAST design

BLAST Assembled Genomes

Choose a species genome to search, or list all genomic BLAST databases.

- Human
- Mouse
- Rat
- *Arabidopsis thaliana*
- *Oryza sativa*
- *Bos taurus*
- *Danio rerio*
- *Drosophila melanogaster*
- *Gallus gallus*
- *Pan troglodytes*
- *Microbes*
- *Apis mellifera*

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast
protein blast	Search **protein** database using a **protein** query *Algorithms:* blastp, psi-blast, phi-blast
blastx	Search **protein** database using a **translated nucleotide** query
tblastn	Search **translated nucleotide** database using a **protein** query

News

New Human and Mouse pre-indexed databases
Human and mouse genomic + transcript megablast searches now use a faster, indexed algorithm that typically reduces run time by two thirds, as compared with standard megablast.
2007-09-04 10:55:00

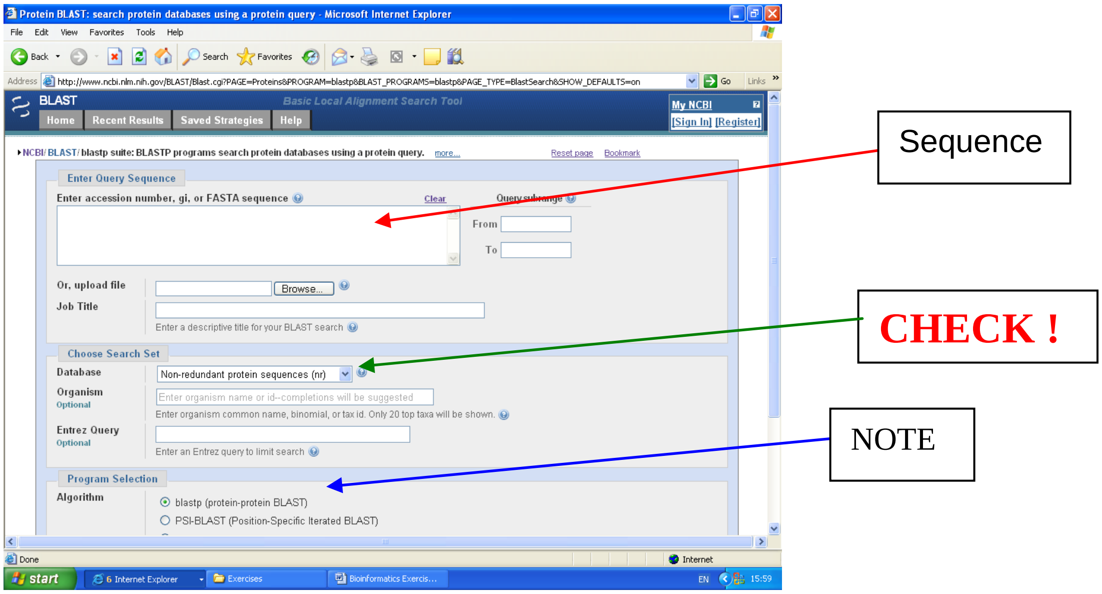
More BLAST news...

Tip of the Day

Using Genomic BLAST

Genomic BLAST pages are helpful because they allow the genomic context of a BLAST search to be displayed in the Map Viewer. For example, discontiguous (cross-species) MegaBLAST against the human RefSeq transcript for albumin (NM_000477) can be used to identify the homolog in the rat

Because of variation in codon usage, in this example use the 'BLASTP' search. The top of the screen should look like this:



Protein BLAST: search protein databases using a protein query - Microsoft Internet Explorer

Address: http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi?PAGE=Proteins&PROGRAM=blastp&BLAST_PROGRAMS=blastp&PAGE_TYPE=BlastSearch&SHOW_DEFAULTS=on

NCBI/ BLAST/ blastp suite: BLASTP programs search protein databases using a protein query.

Enter Query Sequence

Enter accession number, gi, or FASTA sequence

Database: Non-redundant protein sequences (nr)

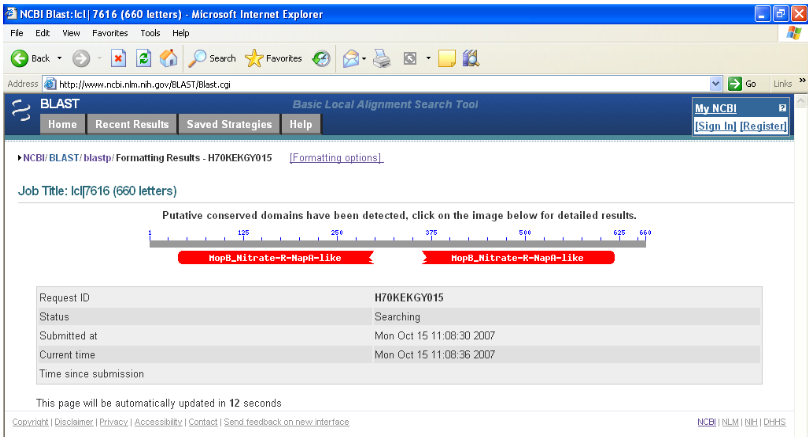
Organism (Optional): Enter organism name or id--completions will be suggested

Algorithm:
- ● blastp (protein-protein BLAST)
- ○ PSI-BLAST (Position-Specific Iterated BLAST)

Paste the amino acid sequence on the previous page into the ‘Query Sequence’ box.

Be **VERY** careful to check the other pull down menus (‘Check’ above). Unfortunately frequently the BLAST pages default to searching the human genome, which is not quite what we wanted to do for 5 minutes….!

Press ‘BLAST’ and then another window will open.
Bottom of Form



NCBI Blast:lcl|7616 (660 letters) - Microsoft Internet Explorer

Address: http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi

NCBI/ BLAST/ blastp/ Formatting Results - H70KEKGY015 [Formatting options]

Job Title: lcl|7616 (660 letters)

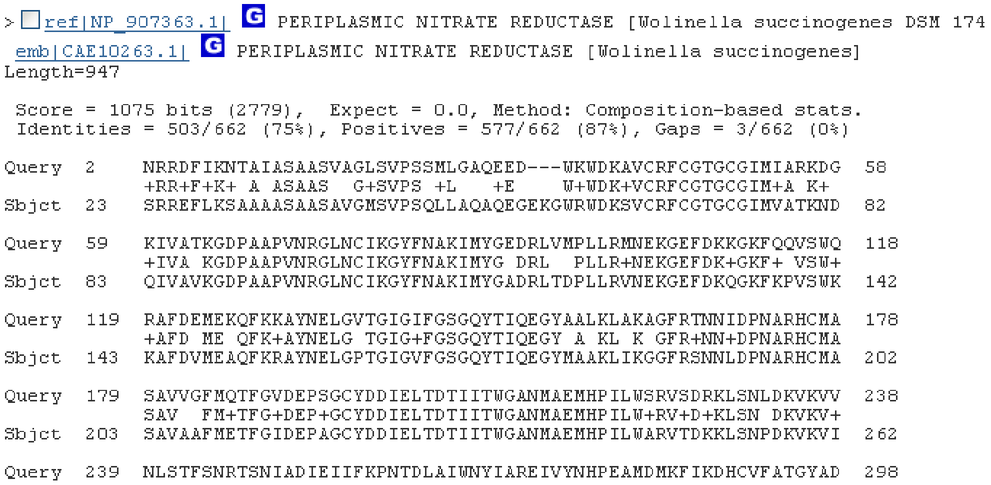
Putative conserved domains have been detected, click on the image below for detailed results.

Request ID	H70KEKGY015
Status	Searching
Submitted at	Mon Oct 15 11:08:30 2007
Current time	Mon Oct 15 11:08:36 2007
Time since submission	

This page will be automatically updated in 12 seconds

The window will update, and you can scroll down a little to a list of organisms in which your amino acid sequence has already been found in. The nearest match will be first. Scroll further down and you can see the regions where the similar occurs (simple consensus sequence).

The example below shows regions of homology, insert/deletion (‘indel’), and where equivalent amino acids are found in the same position. Try to find these for yourself, before asking for help.



```
> □ref|NP_907363.1| G PERIPLASMIC NITRATE REDUCTASE [Wolinella succinogenes DSM 174
 emb|CAE10263.1| G PERIPLASMIC NITRATE REDUCTASE [Wolinella succinogenes]
Length=947

 Score = 1075 bits (2779),  Expect = 0.0, Method: Composition-based stats.
 Identities = 503/662 (75%), Positives = 577/662 (87%), Gaps = 3/662 (0%)

Query  2    NRRDFIKNTAIASAASVAGLSVPSSMLGAQEED---WKWDKAVCRFCGTGCGIMIARKDG  58
            +RR+F+K+ A ASAAS  G+SVPS +L   +E    W+WDK+VCRFCGTGCGIM+A K+
Sbjct  23   SRREFLKSAAAASAASAVGMSVPSQLLAQAQEGEKGWRWDKSVCRFCGTGCGIMVATKND  82

Query  59   KIVATKGDPAAPVNRGLNCIKGYFNAKIMYGEDRLVMPLLRMNEKGEFDKKGKFQQVSWQ  118
            +IVA KGDPAAPVNRGLNCIKGYFNAKIMYG DRL  PLLR+NEKGEFDK+GKF+ VSW+
Sbjct  83   QIVAVKGDPAAPVNRGLNCIKGYFNAKIMYGADRLTDPLLRVNEKGEFDKQGKFKPVSWK  142

Query  119  RAFDEMEKQFKKAYNELGVTGIGIFGSGQYTIQEGYAALKLAKAGFRTNNIDPNARHCMA  178
            +AFD ME QFK+AYNELG TGIG+FGSGQYTIQEGY A KL K GFR+NN+DPNARHCMA
Sbjct  143  KAFDVMEAQFKRAYNELGPTGIGVFGSGQYTIQEGYMAAKLIKGGFRSNNLDPNARHCMA  202

Query  179  SAVVGFMQTFGVDEPSGCYDDIELTDTIITWGANMAEMHPILWSRVSDRKLSNLDKVKVV  238
            SAV  FM+TFG+DEP+GCYDDIELTDTIITWGANMAEMHPILW+RV+D+KLSN DKVKV+
Sbjct  203  SAVAAFMETFGIDEPAGCYDDIELTDTIITWGANMAEMHPILWARVTDKKLSNPDKVKVI  262

Query  239  NLSTFSNRTSNIADIEIIFKPNTDLAIWNYIAREIVYNHPEAMDMKFIKDHCVFATGYAD  298
```

The line between the two sequences gives the matching amino acid, or ‘+’ for when the amino acids are structurally equivalent, and therefore would not lead to a major change in the 3D structure of the protein.

‘Identities 503/662’ means that identical amino acids were found in 503 positions out of 662.

‘Positives = 577/662’ mean that identical plus structurally similar amino acids were found in 577 out of 662 positions.

‘Gaps= 3/662’ refers to the stretching of the sequence by 3 positions to improve the fit.

Go back to the BLASTP page and enter an Accession number instead of the sequence in the Query sequence box, for example ZP_01071654. You’ll see another reason why noting an Accession number can be useful.

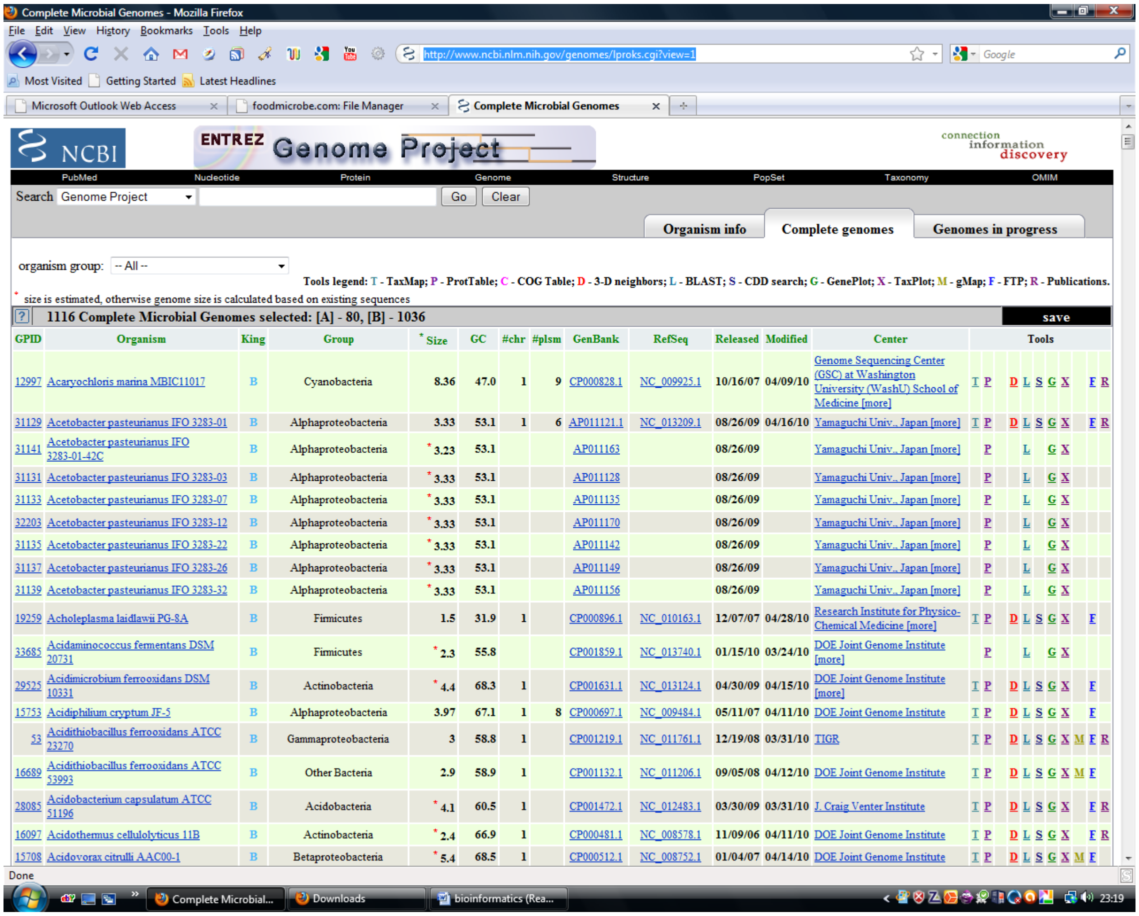
What you should know at the end of this Exercise:

Difference between DNA and amino acid sequences
Differences between BLASTN and BLASTP
Why two DNA sequences for the same enzyme can differ?
What do ‘Sbjct’ and ‘Query’ refer to when you do a BLAST search?
What is an ‘Indel’?
What do the terms ‘Identities’, ‘Positives’ and ‘Gaps’ refer to?

Exercise 4: Searching a bacterial genome

Introduction

(a) Go to the NCBI Entrez genome site at the NCBI ‘Complete Microbial Genomes’:
(http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi?view=1


Complete Microbial Genomes - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi?view=1
Google
Most Visited
Getting Started
Latest Headlines
Microsoft Outlook Web Access
foodmicrobe.com: File Manager
Complete Microbial Genomes
NCBI
ENTREZ Genome Project
connection information discovery
PubMed Nucleotide Protein Genome Structure PopSet Taxonomy OMIM
Search Genome Project
Go
Clear
Organism info
Complete genomes
Genomes in progress
organism group: -- All --
save
Done
Complete Microbial...
Downloads
bioinformatics (Rea...
23:19

Tools legend: T - TaxMap; P - ProtTable; C - COG Table; D - 3-D neighbors; L - BLAST; S - CDD search; G - GenePlot; X - TaxPlot; M - gMap; F - FTP; R - Publications.

* size is estimated, otherwise genome size is calculated based on existing sequences

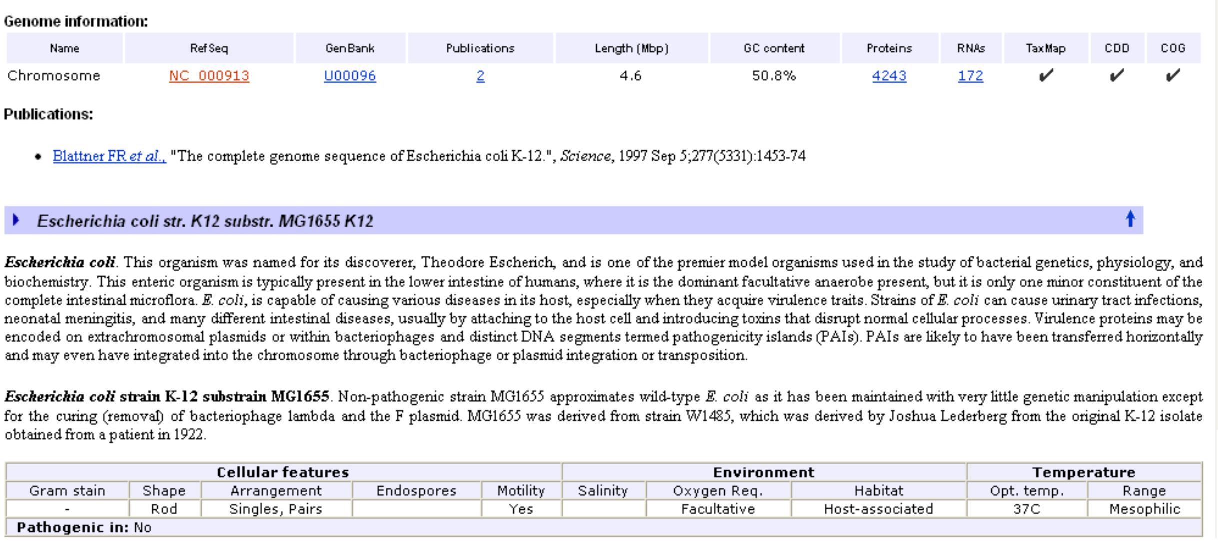
1116 Complete Microbial Genomes selected: [A] - 80, [B] - 1036

GPID	Organism	King	Group	* Size	GC	#chr	#plsm	GenBank	RefSeq	Released	Modified	Center	Tools
12997	Acaryochloris marina MBIC11017	B	Cyanobacteria	8.36	47.0	1	9	CP000828.1	NC_009925.1	10/16/07	04/09/10	Genome Sequencing Center (GSC) at Washington University (WashU) School of Medicine [more]	T P D L S G X F R
31129	Acetobacter pasteurianus IFO 3283-01	B	Alphaproteobacteria	3.33	53.1	1	6	AP011121.1	NC_013209.1	08/26/09	04/16/10	Yamaguchi Univ., Japan [more]	T P D L S G X F R
31141	Acetobacter pasteurianus IFO 3283-01-42C	B	Alphaproteobacteria	* 3.23	53.1			AP011163		08/26/09		Yamaguchi Univ., Japan [more]	P L G X
31131	Acetobacter pasteurianus IFO 3283-03	B	Alphaproteobacteria	* 3.33	53.1			AP011128		08/26/09		Yamaguchi Univ., Japan [more]	P L G X
31133	Acetobacter pasteurianus IFO 3283-07	B	Alphaproteobacteria	* 3.33	53.1			AP011135		08/26/09		Yamaguchi Univ., Japan [more]	P L G X
32203	Acetobacter pasteurianus IFO 3283-12	B	Alphaproteobacteria	* 3.33	53.1			AP011170		08/26/09		Yamaguchi Univ., Japan [more]	P L G X
31135	Acetobacter pasteurianus IFO 3283-22	B	Alphaproteobacteria	* 3.33	53.1			AP011142		08/26/09		Yamaguchi Univ., Japan [more]	P L G X
31137	Acetobacter pasteurianus IFO 3283-26	B	Alphaproteobacteria	* 3.33	53.1			AP011149		08/26/09		Yamaguchi Univ., Japan [more]	P L G X
31139	Acetobacter pasteurianus IFO 3283-32	B	Alphaproteobacteria	* 3.33	53.1			AP011156		08/26/09		Yamaguchi Univ., Japan [more]	P L G X
19259	Acholeplasma laidlawii PG-8A	B	Firmicutes	1.5	31.9	1		CP000896.1	NC_010163.1	12/07/07	04/28/10	Research Institute for Physico-Chemical Medicine [more]	T P D L S G X F
33685	Acidaminococcus fermentans DSM 20731	B	Firmicutes	* 2.3	55.8			CP001859.1	NC_013740.1	01/15/10	03/24/10	DOE Joint Genome Institute [more]	P L G X
29525	Acidimicrobium ferrooxidans DSM 10331	B	Actinobacteria	* 4.4	68.3	1		CP001631.1	NC_013124.1	04/30/09	04/15/10	DOE Joint Genome Institute [more]	T P D L S G X F
15753	Acidiphilium cryptum JF-5	B	Alphaproteobacteria	3.97	67.1	1	8	CP000697.1	NC_009484.1	05/11/07	04/11/10	DOE Joint Genome Institute	T P D L S G X F
53	Acidithiobacillus ferrooxidans ATCC 23270	B	Gammaproteobacteria	3	58.8	1		CP001219.1	NC_011761.1	12/19/08	03/31/10	TIGR	T P D L S G X M F R
16689	Acidithiobacillus ferrooxidans ATCC 53993	B	Other Bacteria	2.9	58.9	1		CP001132.1	NC_011206.1	09/05/08	04/12/10	DOE Joint Genome Institute	T P D L S G X M F
28085	Acidobacterium capsulatum ATCC 51196	B	Acidobacteria	* 4.1	60.5	1		CP001472.1	NC_012483.1	03/30/09	03/31/10	J. Craig Venter Institute	T P D L S G X F R
16097	Acidothermus cellulolyticus 11B	B	Actinobacteria	* 2.4	66.9	1		CP000481.1	NC_008578.1	11/09/06	04/11/10	DOE Joint Genome Institute	T P D L S G X F R
15708	Acidovorax citrulli AAC00-1	B	Betaproteobacteria	* 5.4	68.5	1		CP000512.1	NC_008752.1	01/04/07	04/14/10	DOE Joint Genome Institute	T P D L S G X M F

(b) Scroll down to the organism of choice, in this case *Escherichia coli* K12 MG1655.

Click on the Accession number 225 (NOT the organism name).

You will get a long list of *E. coli* genome projects, and towards the bottom:



Genome information:

Name	RefSeq	GenBank	Publications	Length (Mbp)	GC content	Proteins	RNAs	TaxMap	CDD	COG
Chromosome	NC_000913	U00096	2	4.6	50.8%	4243	172	✔	✔	✔

Publications:

- Blattner FR *et al.*, "The complete genome sequence of Escherichia coli K-12.", *Science*, 1997 Sep 5;277(5331):1453-74

Escherichia coli str. K12 substr. MG1655 K12

Escherichia coli. This organism was named for its discoverer, Theodore Escherich, and is one of the premier model organisms used in the study of bacterial genetics, physiology, and biochemistry. This enteric organism is typically present in the lower intestine of humans, where it is the dominant facultative anaerobe present, but it is only one minor constituent of the complete intestinal microflora. *E. coli*, is capable of causing various diseases in its host, especially when they acquire virulence traits. Strains of *E. coli* can cause urinary tract infections, neonatal meningitis, and many different intestinal diseases, usually by attaching to the host cell and introducing toxins that disrupt normal cellular processes. Virulence proteins may be encoded on extrachromosomal plasmids or within bacteriophages and distinct DNA segments termed pathogenicity islands (PAIs). PAIs are likely to have been transferred horizontally and may even have integrated into the chromosome through bacteriophage or plasmid integration or transposition.

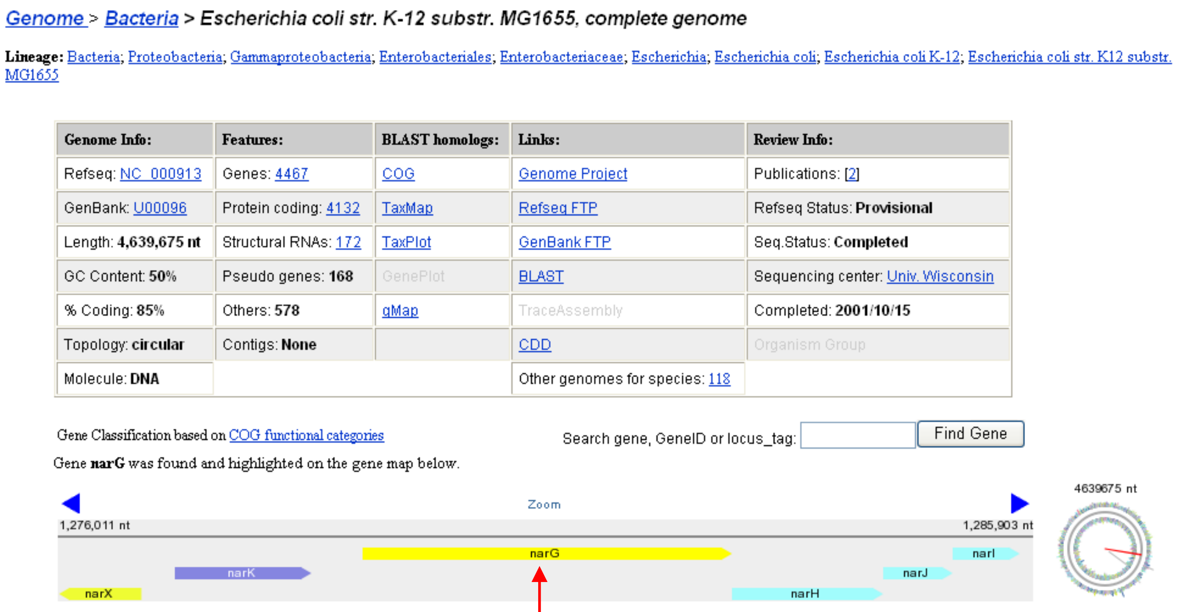
***Escherichia coli* strain K-12 substrain MG1655**. Non-pathogenic strain MG1655 approximates wild-type *E. coli* as it has been maintained with very little genetic manipulation except for the curing (removal) of bacteriophage lambda and the F plasmid. MG1655 was derived from strain W1485, which was derived by Joshua Lederberg from the original K-12 isolate obtained from a patient in 1922.

Cellular features					Environment			Temperature	
Gram stain	Shape	Arrangement	Endospores	Motility	Salinity	Oxygen Req.	Habitat	Opt. temp.	Range
-	Rod	Singles, Pairs		Yes		Facultative	Host-associated	37C	Mesophilic
Pathogenic in: No									

The bottom portion has useful background information on *E. coli* and the particular strain MG1655. You'll see that it is a non-pathogenic strain. Therefore we can use it to compare with the genome of pathogenic strains such as *E. coli* O157:H7.

(c) About half way down the page enter the gene name narG and click on 'Find Gene'. This will then give you two maps of the genome, the linear one has the gene in the middle (which you can click on for further information), and a tiny (!) circular map indicating the general position.

The linear map is also of use to see if any other genes of interest are nearby. Find out what some of them are? Decide whether the order of genes make sense.



Genome > Bacteria > *Escherichia coli str. K-12 substr. MG1655, complete genome*

Lineage: Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales; Enterobacteriaceae; Escherichia; Escherichia coli; Escherichia coli K-12; Escherichia coli str. K12 substr. MG1655

Genome Info:	Features:	BLAST homologs:	Links:	Review Info:
Refseq: NC_000913	Genes: 4467	COG	Genome Project	Publications: [2]
GenBank: U00096	Protein coding: 4132	TaxMap	Refseq FTP	Refseq Status: **Provisional**
Length: **4,639,675 nt**	Structural RNAs: 172	TaxPlot	GenBank FTP	Seq.Status: **Completed**
GC Content: **50%**	Pseudo genes: **168**	GenePlot	BLAST	Sequencing center: Univ. Wisconsin
% Coding: **85%**	Others: **578**	gMap	TraceAssembly	Completed: **2001/10/15**
Topology: **circular**	Contigs: **None**		CDD	Organism Group
Molecule: **DNA**			Other genomes for species: 118	

Gene Classification based on COG functional categories

Search gene, GeneID or locus_tag: ▭ Find Gene

Gene **narG** was found and highlighted on the gene map below.

Look up the colour coding; remember the Appendix section?

What is the function of narG? Click here to get the link to the Genbank entry, and convert to FASTA format. You will now have gone full circle with Exercises 2-4.

Background comment: The colour coding is a common feature of genomes as it helps to visualise the organisation of the genes, but the code can change between Web sites.

The example used above does require knowledge of the relevant enzymes. In this case it should be remembered that nitrate reductase is not a single protein, but a complex involving electron carriers and membrane proteins some of which are not catalytic but anchor the catalytic subunit into the membrane. In a future Exercise you will come across the KEGG site(http://www.genome.jp/kegg/)which gives excellent background to metabolic pathways.

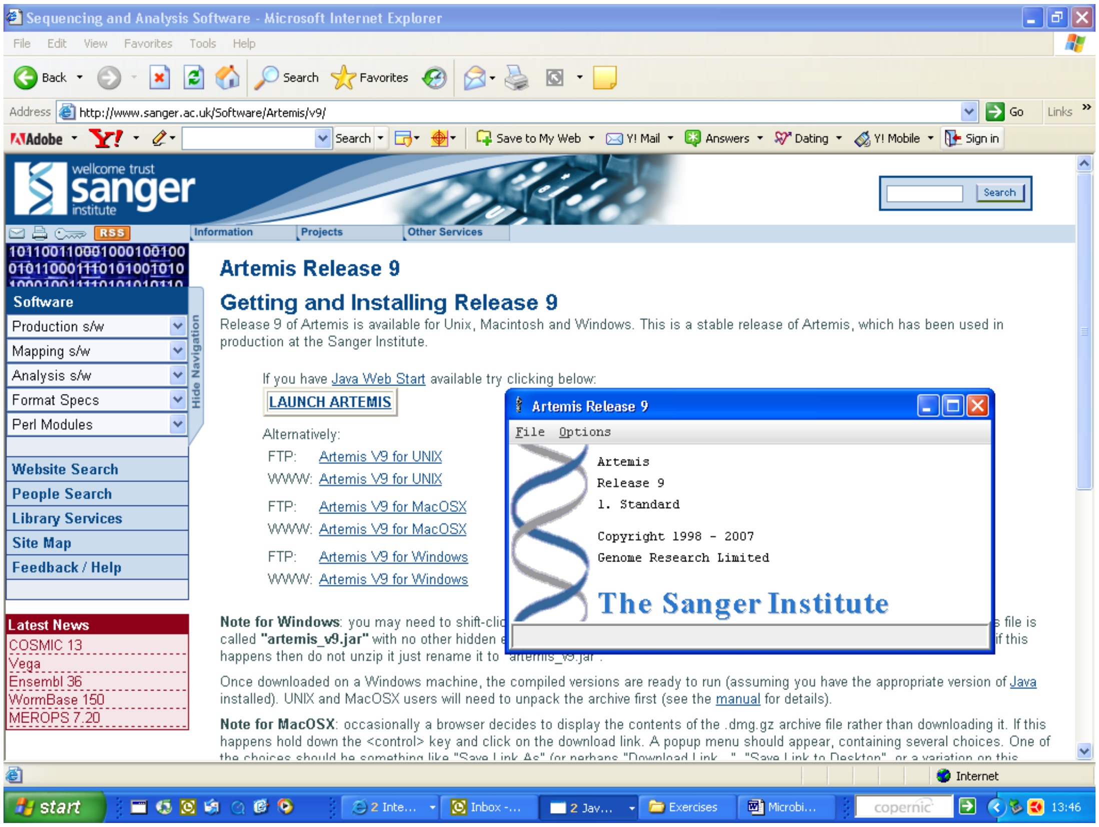
Comment on the location of the genes. Are they clustered together? Is there evidence for operons? You need to look at the neighbouring genes and decide if their function is related to your gene.

Try repeating the exercise using other genes of interest in other modules.

Other parts of the Entrez site to explore are the Entrez Taxonomy at http://www.ncbi.nlm.nih.gov/sites/entrez?db=taxonomy

Exercise 5: Investigating a bacterial genome using Artemis

Click on this link [URL](http://www.sanger.ac.uk/resources/software/artemis/) to the Sanger Centre and access Artemis release 9: http://www.sanger.ac.uk/resources/software/artemis/



Artemis Release 9

Getting and Installing Release 9

Release 9 of Artemis is available for Unix, Macintosh and Windows. This is a stable release of Artemis, which has been used in production at the Sanger Institute.

If you have Java Web Start available try clicking below:

LAUNCH ARTEMIS

Alternatively:

FTP: Artemis V9 for UNIX
WWW: Artemis V9 for UNIX

FTP: Artemis V9 for MacOSX
WWW: Artemis V9 for MacOSX

FTP: Artemis V9 for Windows
WWW: Artemis V9 for Windows

Note for Windows: you may need to shift-clic … s file is called **"artemis_v9.jar"** with no other hidden e … if this happens then do not unzip it just rename it to "artemis_v9.jar".

Once downloaded on a Windows machine, the compiled versions are ready to run (assuming you have the appropriate version of Java installed). UNIX and MacOSX users will need to unpack the archive first (see the manual for details).

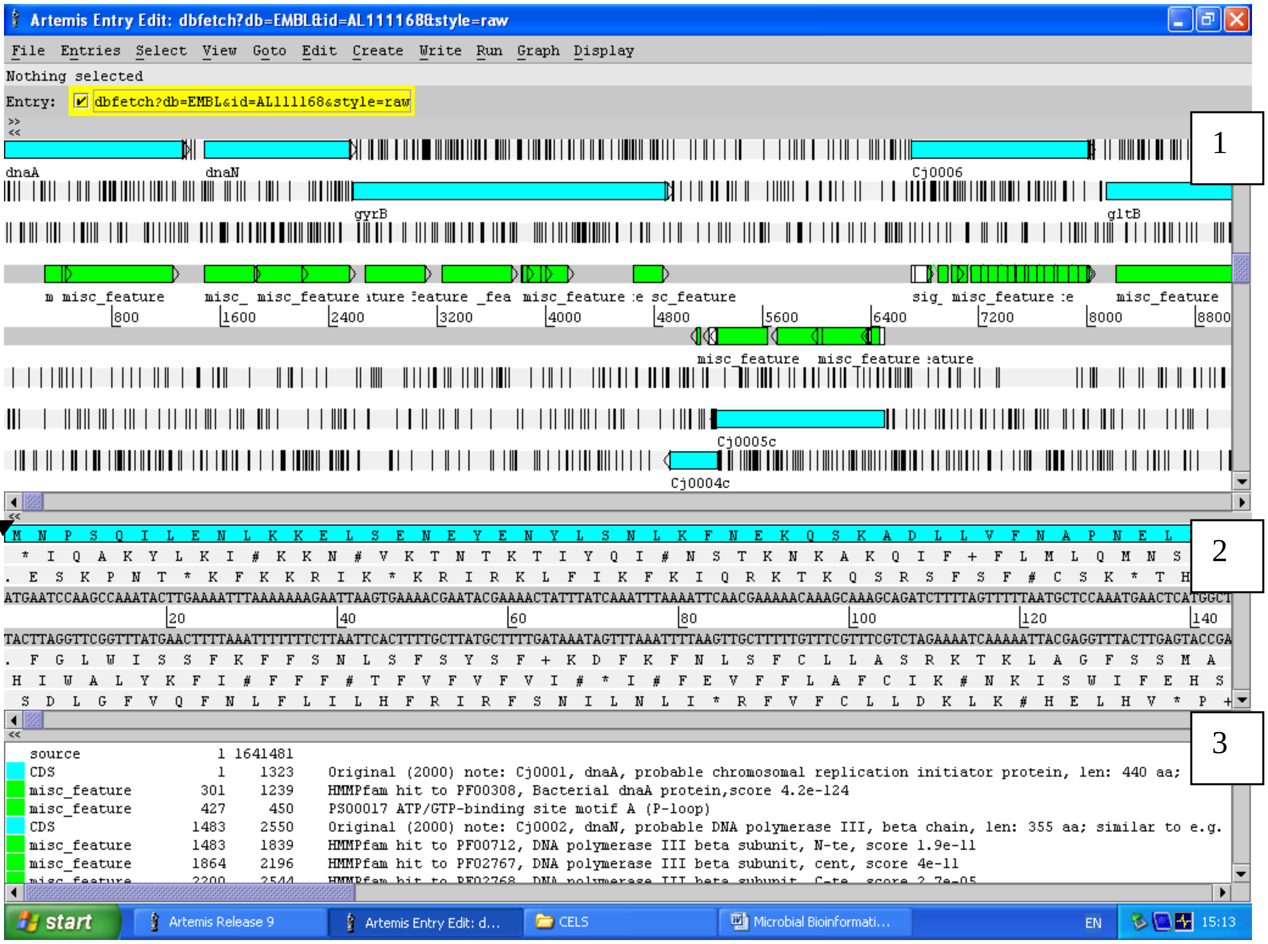
Note for MacOSX: occasionally a browser decides to display the contents of the .dmg.gz archive file rather than downloading it. If this happens hold down the <control> key and click on the download link. A popup menu should appear, containing several choices. One of the choices should be something like "Save Link As" (or perhaps "Download Link..." "Save Link to Desktop", or a variation on this

```
Artemis
Release 9
1. Standard

Copyright 1998 - 2007
Genome Research Limited
```

The Sanger Institute

When you get this small Window, click on ‘File’ then ‘Open from EBI –Dbfetch’ and enter the accession number U00096. This is the genome for *E. coli* strain MG1655.

After about 30-45 seconds, and clicking ‘No’ to viewing any warning, you will get a complicated looking screen!


Artemis Entry Edit: dbfetch?db=EMBL&id=AL111168&style=raw
File Entries Select View Goto Edit Create Write Run Graph Display
1
dnaA
dnaN
Cj0006
gyrB
gltB
m misc_feature
misc_ misc_feature ature feature _fea misc_feature e sc_feature
sig_ misc_feature e
misc_feature
800
1600
2400
3200
4000
4800
5600
6400
7200
8000
8800
misc_feature misc_feature ature
Cj0005c
Cj0004c
2
3
start
Artemis Release 9
Artemis Entry Edit: d...
CELS
Microbial Bioinformati...
EN
15:13

Nothing selected

Entry: ☑ dbfetch?db=EMBL&id=AL111168&style=raw

```
M N P S Q I L E N L K K E L S E N E Y E N Y L S N L K F N E K Q S K A D L L V F N A P N E L
 * I Q A K Y L K I # K K N # V K T N T K T I Y Q I # N S T K N K A K Q I F + F L M L Q M N S
. E S K P N T * K F K K R I K * K R I R K L F I K F K I Q R K T K Q S R S F S F # C S K * T H
ATGAATCCAAGCCAAATACTTGAAAATTTAAAAAAAGAATTAAGTGAAAACGAATACGAAAACTATTTATCAAATTTAAAATTCAACGAAAAACAAAGCAAAGCAGATCTTTTAGTTTTTAATGCTCCAAATGAACTCATGGCT
               |20               |40               |60               |80               |100              |120              |140
TACTTAGGTTCGGTTTATGAACTTTTAAATTTTTTTCTTAATTCACTTTTGCTTATGCTTTTGATAAATAGTTTAAATTTTAAGTTGCTTTTTGTTTCGTTTCGTCTAGAAAATCAAAAATTACGAGGTTTACTTGAGTACCGA
. F G L W I S S F K F F S N L S F S Y S F + K D F K F N L S F C L L A S R K T K L A G F S S M A
 H I W A L Y K F I # F F F # T F V F V F V I # * I # F E V F F L A F C I K # N K I S W I F E H S
  S D L G F V Q F N L F L I L H F R I R F S N I L N L I * R F V F C L L D K L K # H E L H V * P +
```

```
  source             1 1641481
🟦 CDS                1    1323   Original (2000) note: Cj0001, dnaA, probable chromosomal replication initiator protein, len: 440 aa;
🟩 misc_feature     301    1239   HMMPfam hit to PF00308, Bacterial dnaA protein,score 4.2e-124
🟩 misc_feature     427     450   PS00017 ATP/GTP-binding site motif A (P-loop)
🟦 CDS             1483    2550   Original (2000) note: Cj0002, dnaN, probable DNA polymerase III, beta chain, len: 355 aa; similar to e.g.
🟩 misc_feature    1483    1839   HMMPfam hit to PF00712, DNA polymerase III beta subunit, N-te, score 1.9e-11
🟩 misc_feature    1864    2196   HMMPfam hit to PF02767, DNA polymerase III beta subunit, cent, score 4e-11
🟩 misc_feature    2200    2544   HMMPfam hit to PF02768, DNA polymerase III beta subunit, C-te, score 2.7e-05
```

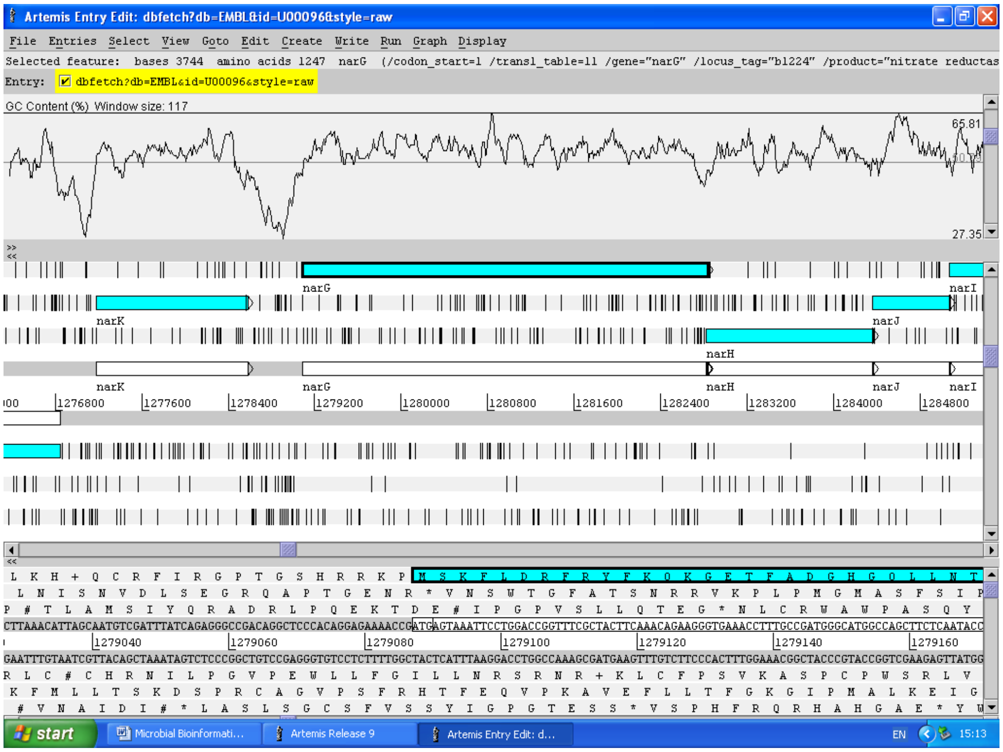
These three areas refer to:
Main sequence view panel. The central grey lines represent the forward and reverse DNA strands at the top and bottom. Above and below these lies are the three forward and three reverse reading frames. The stop codons are marked as black vertical bars. The coloured boxes are genes and any other features.
The nucleotide and amino acid sequences correspond with the ‘Main sequence view panel’. This is a zoomed in version of the main panel to show. To show this double click on a coloured box, and the sequences will shift and highlight the region which you selected.
This bottom area lists the various features on the DNA. If you select a gene, then it will be highlighted in this area. You can also scroll through the list.

Note that both panels can be scrolled left and right, as well as zoomed in and out. The consequence of this is that it is easy to make both panels look alike – which might not be very useful.

ADVICE: Don’t use the zoom slide bars for now!!!

Scroll the middle slider (arrowed) to 1279200, you should find a gene that you have already looked at. Alternatively use to Goto then Navigator and enter narG in the window for ‘Goto Feature With Gene Name’.

Now on the top toolbar select Graph, and then click % GC content.



Artemis Entry Edit: dbfetch?db=EMBL&id=U00096&style=raw

```
Selected feature:  bases 3744  amino acids 1247  narG  (/codon_start=1 /transl_table=11 /gene="narG" /locus_tag="b1224" /product="nitrate reductas
Entry:  ☑ dbfetch?db=EMBL&id=U00096&style=raw
```

The GC content will now be displayed (you can fine tune this with the right-hand side slide bar), and you can see that *nar*G is flanked on the left by the *nar*K gene, and this is flanked by low GC regions on both sides.

Take your time browsing the genome, you can click on genes and see what they are, and look for regions of %CG variation and whether there’s anything of particular interest there – such as a virulence gene (from horizontal gene transfer) or phage DNA.

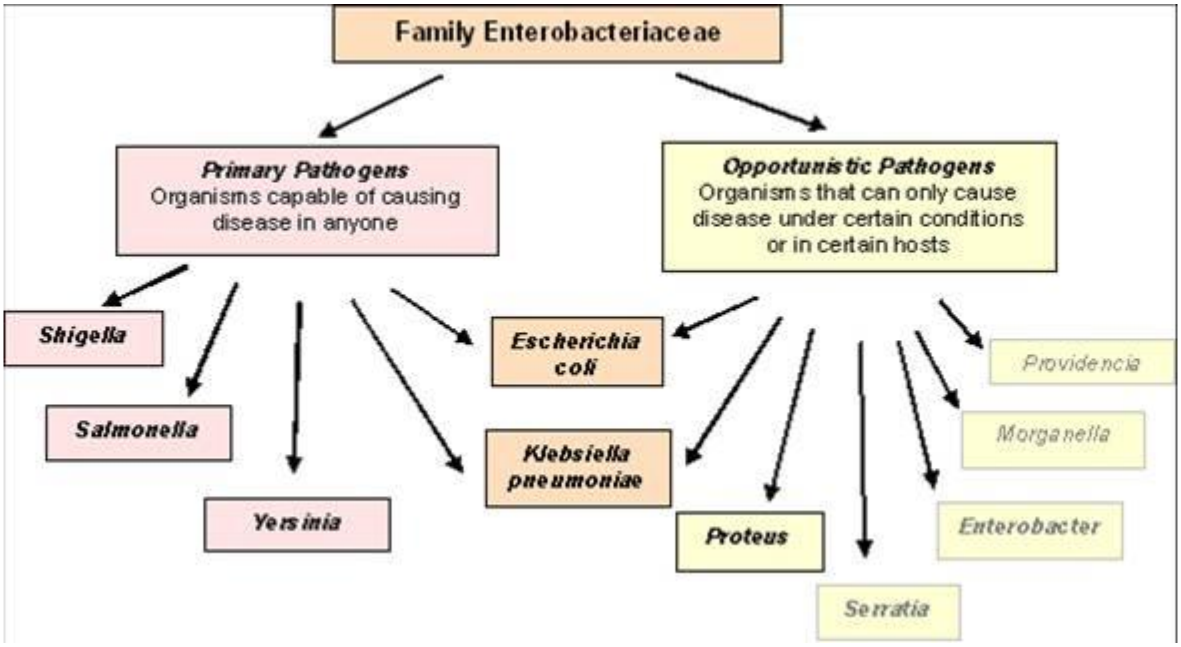
Note: A useful way of using Artemis to look at separate genomes is to have the NCBI Microbial Genome website (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi?view=1) open in a separate window and then you can select the Genbank (not RefSeq) accession number quickly. If you want to look at the genome of the organism you are using in your final year project – then have a go.

Here’s a few Accession numbers for whole genomes:

E. coli O157 AE005174	*Yersinia pestis* AE009952
Salmonella Typhi AL513382	*Salmonella* Typhimurium AE006468

Exercise 6: Comparison of whole genomes in *Enterobacteriaceae* using WebAct.

Below is a diagram from a book of the *Enterobacteriaceae* Family. The most well known members are *Escherichia coli*, and *Salmonella*, but there are others; *Enterobacter*, *Shigella*, *Edwardsiella*, *Klebsiella*, *Citrobacter*, *Proteus*, *Providencia*, *Yersinia, Hafnia*, *Morganella*, *Erwinia*, *Buttiauxella* and *Serratia*.


Family Enterobacteriaceae
Primary Pathogens
Organisms capable of causing disease in anyone
Opportunistic Pathogens
Organisms that can only cause disease under certain conditions or in certain hosts
Shigella
Salmonella
Yersinia
Escherichia coli
Klebsiella pneumoniae
Proteus
Serratia
Enterobacter
Morganella
Providencia

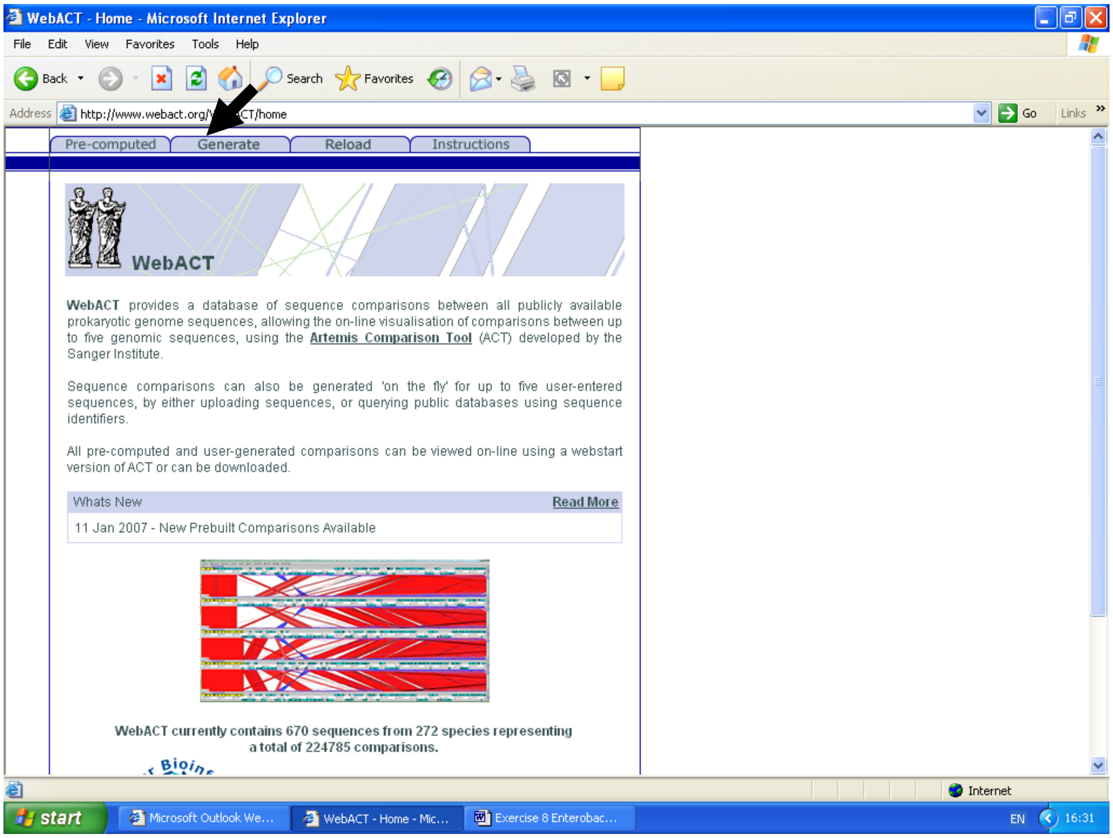
But this diagram is very simplistic and does not give any indication of relatedness, sharing of pathogenicity genes, or evolutionary development. Ask yourself ‘Why are *E. coli* and *K. pneumoniae* in the middle?’

This exercise is designed to give you the opportunity to investigate the *Enterobacteriaceae* Family. Questions to ask are ‘What is in common between the organisms, and hence why they are in the Family’. Related to this you can consider ‘What are the differences between strains within a bacterial species’. This leads on to questions such as ‘Why are most of the *E. coli* in your intestines harmless, yet *E. coli* O157:H7 so pathogenic?’.

Why is this useful? One answer is that by knowing the differences on the bacterial surface one can work towards producing specific vaccines, or detection methods by generating antibodies against unique proteins without cross-reaction with very similar, yet ‘harmless’ organisms.

Go to ‘WebACT Prebuilt Comparison: Select Sequences ‘ at
http:// www.webact.org/WebACT .

Click on the ‘Pre-computed’ tab.



WebACT provides a database of sequence comparisons between all publicly available prokaryotic genome sequences, allowing the on-line visualisation of comparisons between up to five genomic sequences, using the **Artemis Comparison Tool** (ACT) developed by the Sanger Institute.

Sequence comparisons can also be generated 'on the fly' for up to five user-entered sequences, by either uploading sequences, or querying public databases using sequence identifiers.

All pre-computed and user-generated comparisons can be viewed on-line using a webstart version of ACT or can be downloaded.

11 Jan 2007 - New Prebuilt Comparisons Available

WebACT currently contains 670 sequences from 272 species representing a total of 224785 comparisons.

This web site enables you to align whole bacterial genomes.

Some useful RefSeq numbers are:

Organism	RefSeq number
E. coli K12	U00096
E. coli O157	NC_002655
Yersinia pestis	NC_010159
Salmonella Typhi	NC_003198
Salmonella Typhimurium	NC_003197

Warning: It can get very tiresome zooming in and out and rescaling for the screen the alignments from the program. So breathe deeply, and be patient. It is a very powerful program and is used routinely by the professionals…..!

We will start with a ‘simple’ genome to genome comparison, and then increase the number of comparisons. The software default is to compare 2 sequences.

1) Compare two strains of the same species; *E. coli* K12 and *E. coli* O157.

You may already know that *E. coli* K12 has been used for many decades in laboratories for studies into the biochemistry and genetics of bacteria. *E. coli* O157 is a pathogenic variety of *E. coli* that can causes a number of serious infections, such as haemorrhagic colitis, and haemorrhagic uraemic syndrome.

Select the two genomes from the two windows. I'll assume *E. coli* K12 is in the upper window for the example below.
Follow the menus, and click 'Next' twice, then 'Start ACT'.
Click 'No' for warnings.
A small diagram of the two genomes will appear and you can see if they are similar lengths or not.
Do not download the files, but click 'Show'.
Move the cursor to the middle section, and right click on a white area – turn off 'Locked'.
Move up to the sequences and right click, turn off 'stop codons' for both sequences.

You will find a lot of red lines, indicating near identical DNA sequence matches. The blue lines are where there are good matches, but the sequences are in reverse orientation.

You can slide the sequences past each other.

What do the gaps represent?

A region which has a red block can be centralised by double clicking on it.

Use the sliders for each sequence to zoom in and out.

The middle slider acts as a filter, the default is low resolution.

Go to 242400 on the bottom slider. You will find a large region which is not joined to any corresponding region in *E. coli* K12 above. Consider what this means, and remember both are the same species.

Go to the tool bar at the top of the window.
Click on 'Graph'
Two identical menus appear, one for the upper sequence, and the other for the lower sequence.
Click on 'GC content %' for the upper sequence.
A new upper diagram will appear.

If you scroll along, and change the zoom, you will notice regions which are significantly higher or lower than others. These can be regions which have special function, or may be 'foreign' DNA which has been incorporated ie. lysogenic bacteriophages, antibiotic resistance, and pathogenicity islands.

For *E. coli* look at positions: 293000. You will see the %GC content dips for a long region, and you will find the *yag* genes (*yag*L, etc). This is the coding region for a prophage.

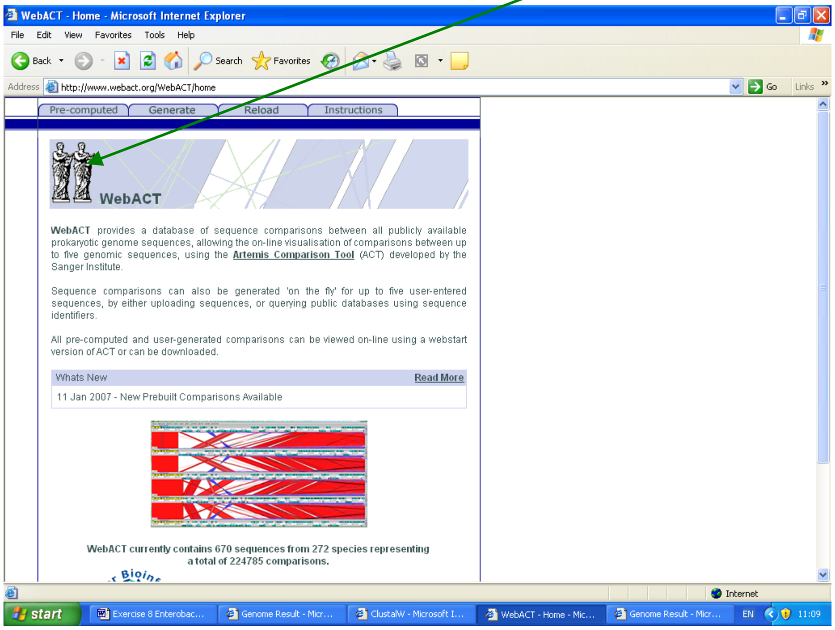
Go to position 353600 on the *E. coli* O157 sequence (bottom sequence) and you will see that eaeH is different lengths in the two genomes, and only the N-terminal region matches. They then match up again with *ykg*A.

Move the arrow back and highlight the *eae*H gene for *E. coli* O157. Then right click for 'View' menu, and then click on 'View Selected Features'. You will then find a description of the gene function, and the amino acid sequence. Could the *eae*H gene of *E. coli* O157 be involved in virulence?

A longer route is to use the link to the RefSeq files **at (blue)** NC_000913.2 and NC_002695.1 at via NCBI (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi) to look up genes from *E. coli* K12.

Can you find other genes which are unique to *E. coli* O157? Hint : look for gaps, ie. positions 18463-24969, 241522- 277462 and 300075-330124, 501537-503093. Look at the %GC pattern, notice how often it suddenly drops below the average value.

Are these possibly related to virulence?
To try another comparison you need to click on the two statues, and not on the more obvious 'Pre-computed'. This will take you back to the beginning.


WebACT - Home - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Address http://www.webact.org/WebACT/home
Go
Links
Pre-computed
Generate
Reload
Instructions
WebACT

WebACT provides a database of sequence comparisons between all publicly available prokaryotic genome sequences, allowing the on-line visualisation of comparisons between up to five genomic sequences, using the **Artemis Comparison Tool** (ACT) developed by the Sanger Institute.

Sequence comparisons can also be generated 'on the fly' for up to five user-entered sequences, by either uploading sequences, or querying public databases using sequence identifiers.

All pre-computed and user-generated comparisons can be viewed on-line using a webstart version of ACT or can be downloaded.

Whats New **Read More**

11 Jan 2007 - New Prebuilt Comparisons Available

WebACT currently contains 670 sequences from 272 species representing a total of 224785 comparisons.

Other comparisons to consider. Note – you might well see a mass of lines, initially. So switch off the stop codons for both sequences, zoom out, and move the middle slide bar until the profile makes ‘sense’.

Compare *E. coli* K12 (NC_000913) with *Salmonella* Typhimurium (NC_003197)
Salmonella Typhimurium (NC_003198) with *Salmonella* Typhi (NC_003197)
E. coli K12 with *Buchnera aphidicola* (NC_002528)
E. coli K12 with *Pseudomonas aeruginosa* (NC_009656)

Using the basics of whether the bacteria have the same genus name, and the 16S tree from Part 2, ask yourself if the comparisons make sense.

Now to go further, change the number of sequences to 3!
Compare *Shigella dysenteriae* (CP000034), *E. coli* K12 (NC_000913) and *Salmonella* Typhimurium
Compare *E. coli* K12 with *Salmonella* Typhi, and *S.* Typhimurium.

Notes:
1. The Accession number is specific to that sequence. There will be hundreds of *E. coli* 16S sequences in the database, and you would have a problem locating the same sequence if you searched twice, but a few weeks apart. By noting the accession number for sequence files you will make life MUCH easier for yourself.

2. Bacterial names are normally in italics as they are Genus and Species name, ie. *Escherichia* [genus] *coli* [species], however you’ll notice that I have not done this for the *Salmonella* names above. The reason is that they are an exception. We now recognise that the 2400+ varieties of *Salmonella* are in fact in only two species. This is totally unhelpful, and so the second name refers to the serovar. So the name is *Salmonella* [genus] and Typhi [serovar].
The story is a little more complicated, but it’s not important enough to cover in any more depth here.

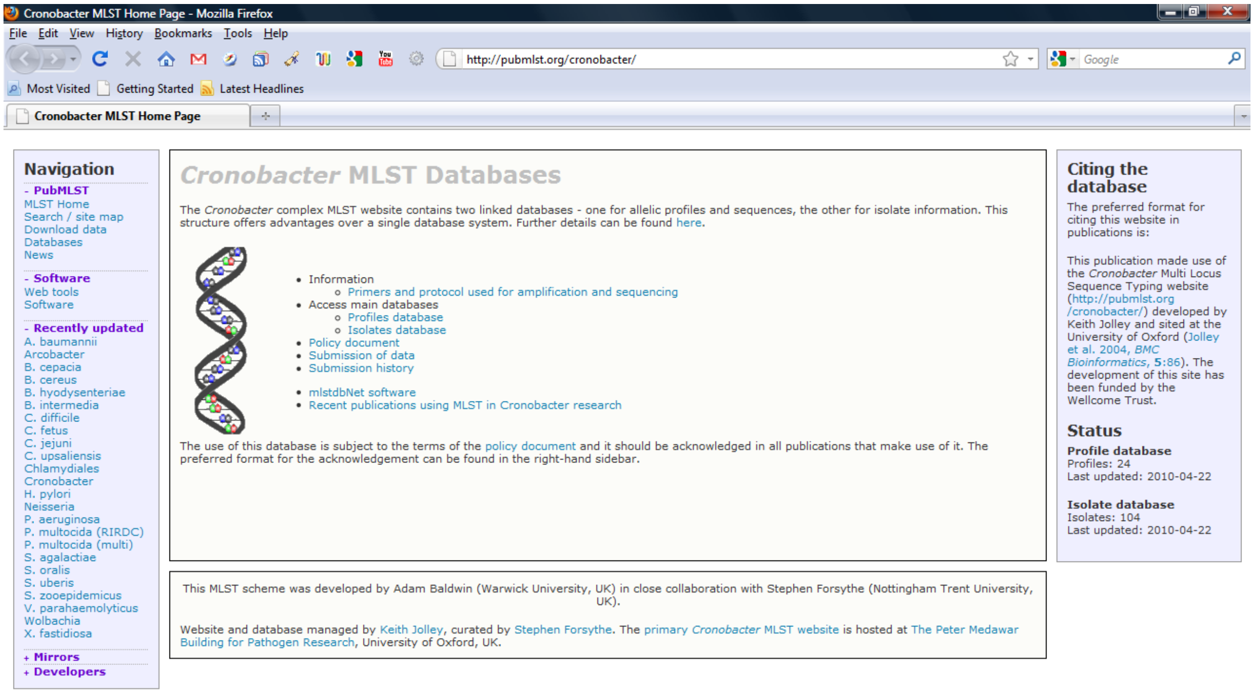
Exercise 7: Multilocus sequence typing analysis (MLST)

Part 1: Introduction to online MLST database analysis

Chapter 5 (p249) refers to MLST as a means of analysing portions of genes and generating profiles. The example below uses the PubMLST web site (http://pubmlst.org/) which is run by the University of Oxford, UK. Not only can the user upload their own sequence data and compare with the online database, but you can also download sequences for you own analysis. The example used here is for *Cronobacter* spp. (Section 4.8.2)

The *Cronobacter* MLST scheme, in keeping with other MLST schemes, is based on 7 loci which are independent of each other. Each PCR product is about 400-500 bases in length and can be chained together to give a pseudogene of ~3500 bases. This is considerably bigger than the partial 16S rDNA sequence which is commonly used, and which also suffers from constraints of limited variability.

Go to http://pubMLST.org/cronobacter to get the homepage below.


Cronobacter MLST Home Page - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://pubmlst.org/cronobacter/
Google
Most Visited Getting Started Latest Headlines
Cronobacter MLST Home Page

Navigation

- PubMLST
MLST Home
Search / site map
Download data
Databases
News

- Software
Web tools
Software

- Recently updated
A. baumannii
Arcobacter
B. cepacia
B. cereus
B. hyodysenteriae
B. intermedia
C. difficile
C. fetus
C. jejuni
C. upsaliensis
Chlamydiales
Cronobacter
H. pylori
Neisseria
P. aeruginosa
P. multocida (RIRDC)
P. multocida (multi)
S. agalactiae
S. oralis
S. uberis
S. zooepidemicus
V. parahaemolyticus
Wolbachia
X. fastidiosa

+ Mirrors
+ Developers

***Cronobacter* MLST Databases**

The *Cronobacter* complex MLST website contains two linked databases - one for allelic profiles and sequences, the other for isolate information. This structure offers advantages over a single database system. Further details can be found here.

- Information
  - Primers and protocol used for amplification and sequencing
- Access main databases
  - Profiles database
  - Isolates database
- Policy document
- Submission of data
- Submission history

- mlstdbNet software
- Recent publications using MLST in Cronobacter research

The use of this database is subject to the terms of the policy document and it should be acknowledged in all publications that make use of it. The preferred format for the acknowledgement can be found in the right-hand sidebar.

This MLST scheme was developed by Adam Baldwin (Warwick University, UK) in close collaboration with Stephen Forsythe (Nottingham Trent University, UK).

Website and database managed by Keith Jolley, curated by Stephen Forsythe. The primary *Cronobacter* MLST website is hosted at The Peter Medawar Building for Pathogen Research, University of Oxford, UK.

Citing the database

The preferred format for citing this website in publications is:

This publication made use of the *Cronobacter* Multi Locus Sequence Typing website (http://pubmlst.org/cronobacter/) developed by Keith Jolley and sited at the University of Oxford (Jolley et al. 2004, *BMC Bioinformatics*, **5**:86). The development of this site has been funded by the Wellcome Trust.

Status

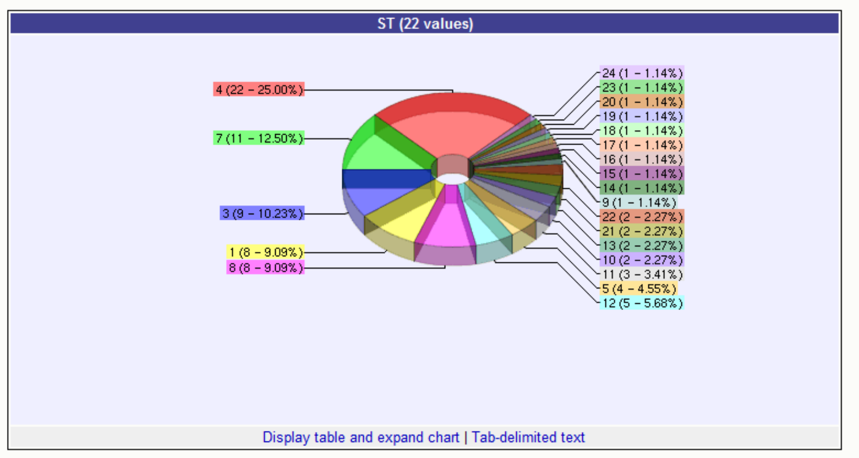
Profile database
Profiles: 24
Last updated: 2010-04-22

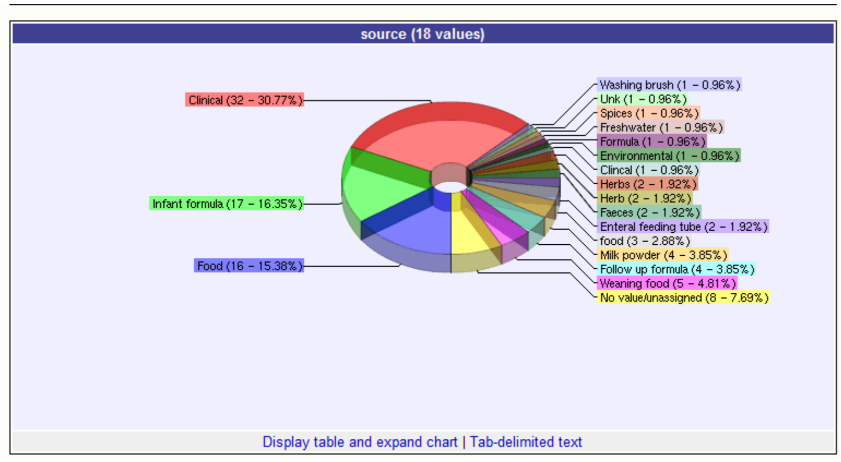
Isolate database
Isolates: 104
Last updated: 2010-04-22

From here you can access the protocols, and the databases. The original paper describing this site was by Baldwin *et al.* (2009) (BMC Microbiology;

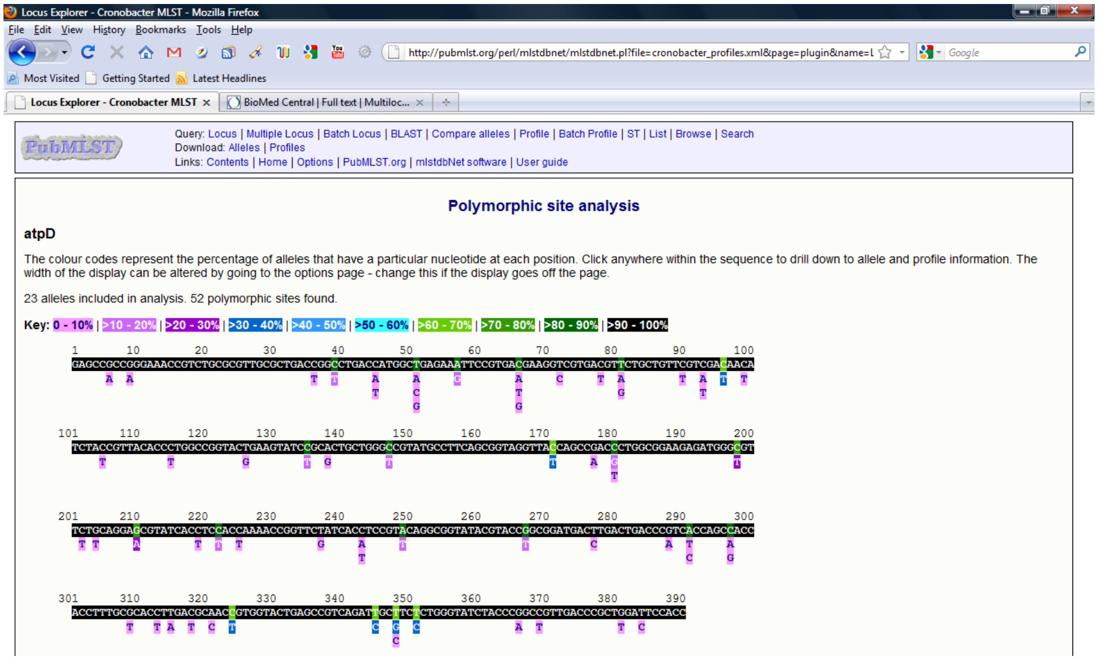
http://www.biomedcentral.com/1471-2180/9/223) for *C. sakazakii* and *C. malonaticus* but has been extended since then to cover all *Cronobacter* species.

Click on 'Isolates database' and the 'Detailed statistics' under the 'Database statistics' heading. You've now obtain a series of pie charts indicating the sources of the strains, sequence types and allele frequencies.





To follow allele frequencies, it is possibly easier to look at the variations. So go back to the homepage, then click on 'Profiles database' and then 'Locus explorer'. For polymorphic site analysis there are pull down menu to select one of the seven alleles in the scheme. For now use the default of *atp*D and click 'submit' and you'll get the following screen. Remember that just one nucleotide difference means it is a different allele. These are numbered according to the order they were found, and do not relate to the number of differences. So allele sequences for *atp*D 1 and 26 can be only one nucleotide different, whereas *atp*D 1 and 2 could have 18 differences between them.


Locus Explorer - Cronobacter MLST - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://pubmlst.org/perl/mlstdbnet/mlstdbnet.pl?file=cronobacter_profiles.xml&page=plugin&name=L
Google
Most Visited
Getting Started
Latest Headlines
Locus Explorer - Cronobacter MLST
BioMed Central | Full text | Multiloc...
PubMLST

Query: Locus | Multiple Locus | Batch Locus | BLAST | Compare alleles | Profile | Batch Profile | ST | List | Browse | Search
Download: Alleles | Profiles
Links: Contents | Home | Options | PubMLST.org | mlstdbNet software | User guide

Polymorphic site analysis

atpD

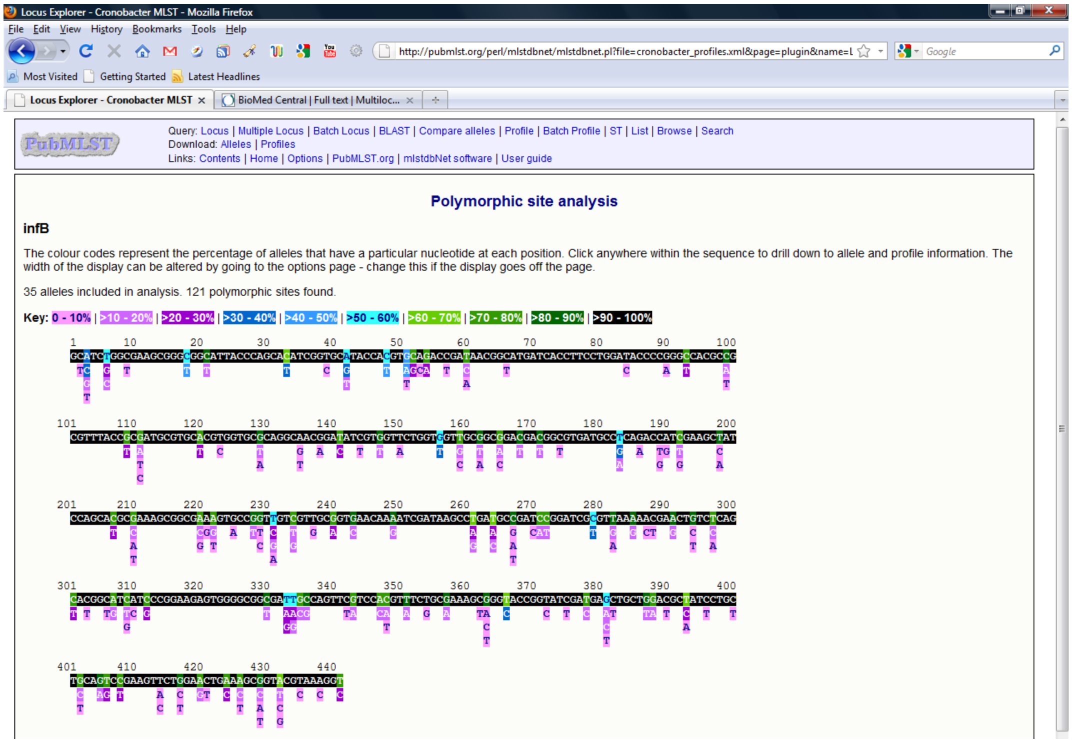
The colour codes represent the percentage of alleles that have a particular nucleotide at each position. Click anywhere within the sequence to drill down to allele and profile information. The width of the display can be altered by going to the options page - change this if the display goes off the page.

23 alleles included in analysis. 52 polymorphic sites found.

Key: 0 - 10% | >10 - 20% | >20 - 30% | >30 - 40% | >40 - 50% | >50 - 60% | >60 - 70% | >70 - 80% | >80 - 90% | >90 - 100%

```
1        10        20        30        40        50        60        70        80        90       100
GAGCCGCCGGGAAACCGTCTGCGCGTTGCGCTGACCGGCCTGACCATGGCTGAGAAATTCCGTGACGAAGGTCGTGACGTTCTGCTGTTCGTCGACAACA
A A T T A A G A C T A T A T T
T C T G T
G G

101      110       120       130       140       150       160       170       180       190       200
TCTACCGTTACACCCTGGCCGGTACTGAAGTATCCGCACTGCTGGGCCGTATGCCTTCAGCGGTAGGTTACCAGCCGACCCTGGCGGAAGAGATGGGCGT
T T G T G T T A G T
T

201      210       220       230       240       250       260       270       280       290       300
TCTGCAGGAGCGTATCACCTCCACCAAAACCGGTTCTATCACCTCCGTACAGGCGGTATACGTACCGGCGGATGACTTGACTGACCCGTCACCAGCCACC
T T A T T T G A T T C A T A
T C G

301      310       320       330       340       350       360       370       380       390
ACCTTTGCGCACCTTGACGCAACCGTGGTACTGAGCCGTCAGATTGCTTCTCTGGGTATCTACCCGGCCGTTGACCCGCTGGATTCCACC
T T A T C T C G C A T T C
C
```

You can see that the variation is in a few locations. The colour coding shows the frequency with which the nucleotide changes at each location. If you instead choose *inf*B you get the following:


Locus Explorer - Cronobacter MLST - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://pubmlst.org/perl/mlstdbnet/mlstdbnet.pl?file=cronobacter_profiles.xml&page=plugin&name=L
Google
Most Visited
Getting Started
Latest Headlines
Locus Explorer - Cronobacter MLST
BioMed Central | Full text | Multiloc...
PubMLST

Query: Locus | Multiple Locus | Batch Locus | BLAST | Compare alleles | Profile | Batch Profile | ST | List | Browse | Search
Download: Alleles | Profiles
Links: Contents | Home | Options | PubMLST.org | mlstdbNet software | User guide

Polymorphic site analysis

infB

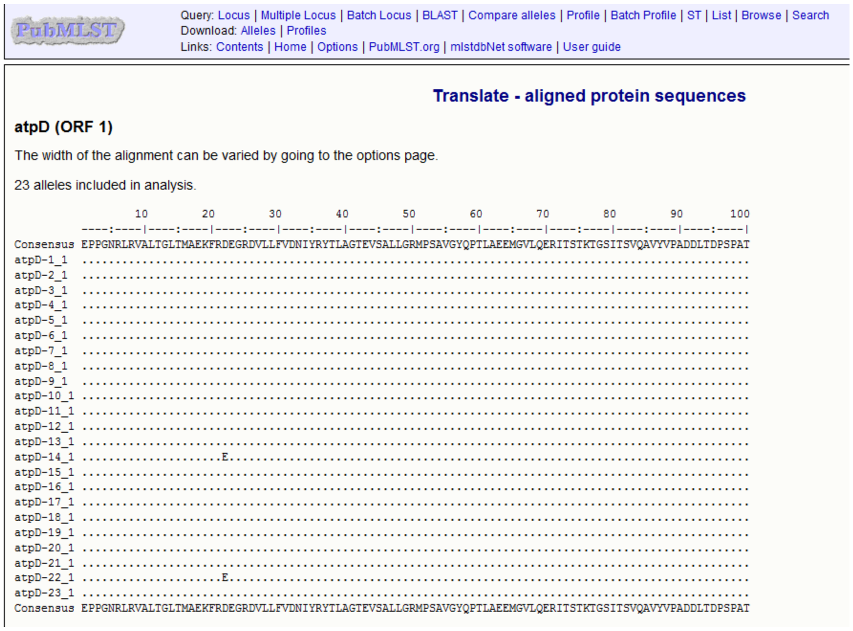
The colour codes represent the percentage of alleles that have a particular nucleotide at each position. Click anywhere within the sequence to drill down to allele and profile information. The width of the display can be altered by going to the options page - change this if the display goes off the page.

35 alleles included in analysis. 121 polymorphic sites found.

Key: 0 - 10% | >10 - 20% | >20 - 30% | >30 - 40% | >40 - 50% | >50 - 60% | >60 - 70% | >70 - 80% | >80 - 90% | >90 - 100%

```
1        10        20        30        40        50        60        70        80        90       100
GCATCTGGCGAAGCGGGCGGCATTACCCAGCACATCGGTGCATACCACGTGCAGACCGATAACGGCATGATCACCTTCCTGGATACCCCGGGCCACGCCG
TC G T T T T C G T AGCA T C T C A T A
G C T T A T
T

101      110       120       130       140       150       160       170       180       190       200
CGTTTACCGCGATGCGTGCACGTGGTGCGCAGGCAACGGATATCGTGGTTCTGGTGGTTGCGGCGGACGACGGCGTGATGCCTCAGACCATCGAAGCTAT
T A T C T G A C T T A T G T A T T T G A TG T C
T A T C A C A G G A
C

201      210       220       230       240       250       260       270       280       290       300
CCAGCACGCGAAAGCGGCGAAAGTGCCGGTTGTCGTTGCGGTGAACAAAATCGATAAGCCTGATGCCGATCCGGATCGCGTTAAAAACGAACTGTCTCAG
T C CGG A TT C T G A C G A A G CAT T G G CT G C C
A G T C G G G C A A T A
T A T

301      310       320       330       340       350       360       370       380       390       400
CACGGCATCATCCCGGAAGAGTGGGGCGGCGATTGCCAGTTCGTCCACGTTTCTGCGAAAGCGGGTACCGGTATCGATGAGCTGCTGGACGCTATCCTGC
T T TG TC G T AACG TA CA A G A TA C C T C AT TA T C T T
G GG T C C A
T T

401      410       420       430       440
TGCAGTCCGAAGTTCTGGAACTGAAAGCGGTACGTAAAGGT
C AG T A C GT C C C T C C C
T C T T A C
T G
```

What you are looking at are variations in the DNA sequence for some housekeeping genes. Do these variations affect the protein sequence? To find out instead of using the ‘Polymorphic site analysis’ option use the ‘Translate’ instead. When you repeat for *atp*D and *inf*B you get the following:



Query: Locus | Multiple Locus | Batch Locus | BLAST | Compare alleles | Profile | Batch Profile | ST | List | Browse | Search
Download: Alleles | Profiles
Links: Contents | Home | Options | PubMLST.org | mlstdbNet software | User guide

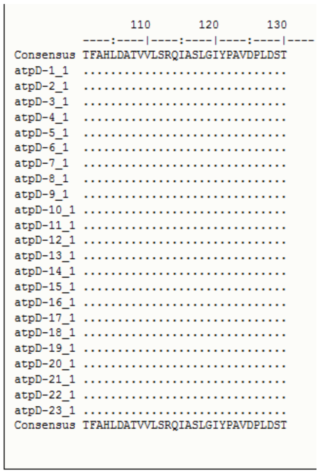
Translate - aligned protein sequences

atpD (ORF 1)

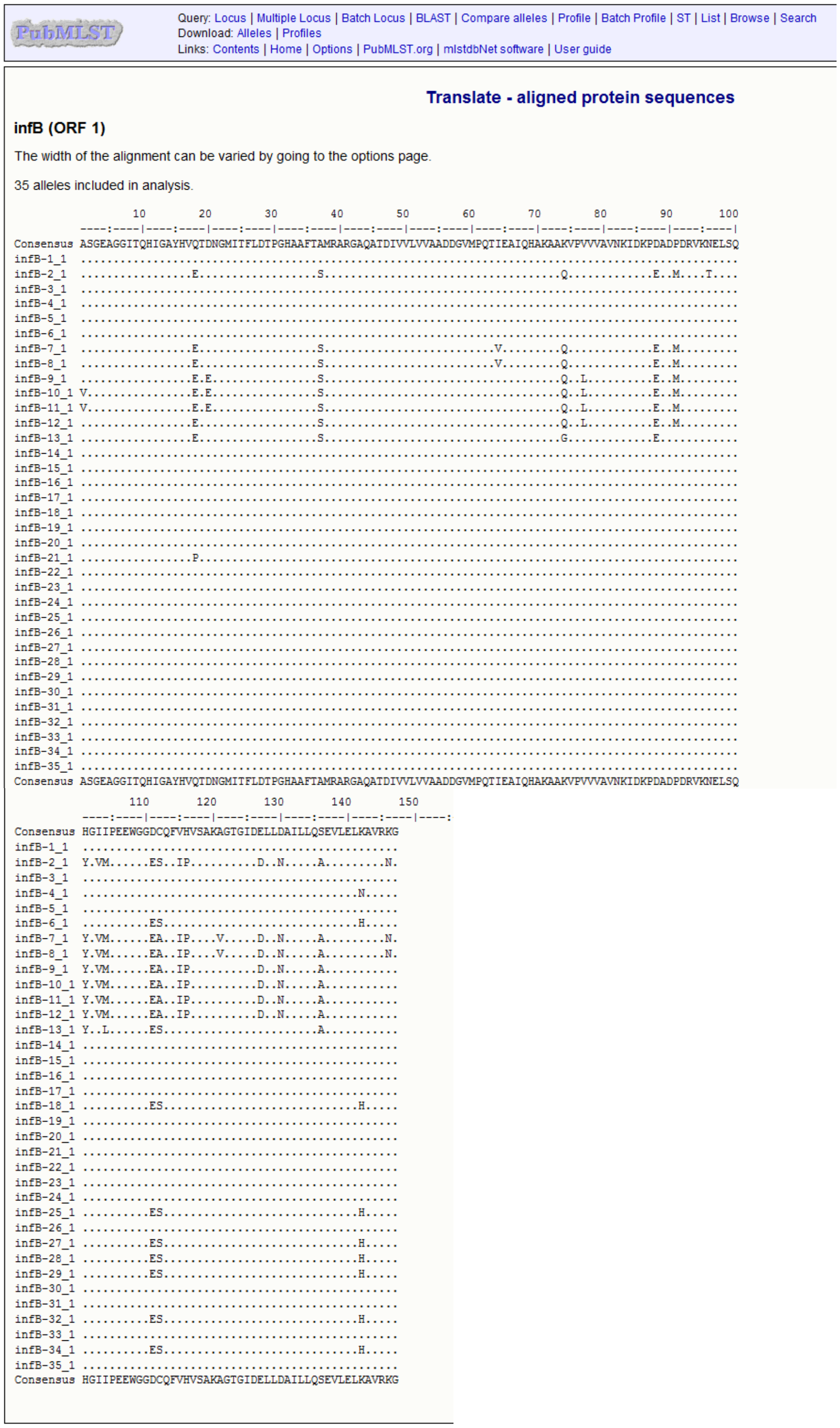
The width of the alignment can be varied by going to the options page.

23 alleles included in analysis.

```
                  10        20        30        40        50        60        70        80        90       100
          ----:----|----:----|----:----|----:----|----:----|----:----|----:----|----:----|----:----|----:----|
Consensus EPPGNRLRVALTGLTMAEKFRDEGRDVLLFVDNIYRYTLAGTEVSALLGRMPSAVGYQPTLAEEMGVLQERITSTKTGSITSVQAVYVPADDLTDPSPAT
atpD-1_1  ....................................................................................................
atpD-2_1  ....................................................................................................
atpD-3_1  ....................................................................................................
atpD-4_1  ....................................................................................................
atpD-5_1  ....................................................................................................
atpD-6_1  ....................................................................................................
atpD-7_1  ....................................................................................................
atpD-8_1  ....................................................................................................
atpD-9_1  ....................................................................................................
atpD-10_1 ....................................................................................................
atpD-11_1 ....................................................................................................
atpD-12_1 ....................................................................................................
atpD-13_1 ....................................................................................................
atpD-14_1 .....................E..............................................................................
atpD-15_1 ....................................................................................................
atpD-16_1 ....................................................................................................
atpD-17_1 ....................................................................................................
atpD-18_1 ....................................................................................................
atpD-19_1 ....................................................................................................
atpD-20_1 ....................................................................................................
atpD-21_1 ....................................................................................................
atpD-22_1 .....................E..............................................................................
atpD-23_1 ....................................................................................................
Consensus EPPGNRLRVALTGLTMAEKFRDEGRDVLLFVDNIYRYTLAGTEVSALLGRMPSAVGYQPTLAEEMGVLQERITSTKTGSITSVQAVYVPADDLTDPSPAT
```



```
                 110       120       130
          ----:----|----:----|----:----|----
Consensus TFAHLDATVVLSRQIASLGIYPAVDPLDST
atpD-1_1  ..............................
atpD-2_1  ..............................
atpD-3_1  ..............................
atpD-4_1  ..............................
atpD-5_1  ..............................
atpD-6_1  ..............................
atpD-7_1  ..............................
atpD-8_1  ..............................
atpD-9_1  ..............................
atpD-10_1 ..............................
atpD-11_1 ..............................
atpD-12_1 ..............................
atpD-13_1 ..............................
atpD-14_1 ..............................
atpD-15_1 ..............................
atpD-16_1 ..............................
atpD-17_1 ..............................
atpD-18_1 ..............................
atpD-19_1 ..............................
atpD-20_1 ..............................
atpD-21_1 ..............................
atpD-22_1 ..............................
atpD-23_1 ..............................
Consensus TFAHLDATVVLSRQIASLGIYPAVDPLDST
```



Query: Locus | Multiple Locus | Batch Locus | BLAST | Compare alleles | Profile | Batch Profile | ST | List | Browse | Search
Download: Alleles | Profiles
Links: Contents | Home | Options | PubMLST.org | mlstdbNet software | User guide

Translate - aligned protein sequences

infB (ORF 1)

The width of the alignment can be varied by going to the options page.

35 alleles included in analysis.

```
                  10        20        30        40        50        60        70        80        90       100
          ----:----|----:----|----:----|----:----|----:----|----:----|----:----|----:----|----:----|----:----|
Consensus ASGEAGGITQHIGAYHVQTDNGMITFLDTPGHAAFTAMRARGAQATDIVVLVVAADDGVMPQTIEAIQHAKAAKVPVVVAVNKIDKPDADPDRVKNELSQ
infB-1_1  ....................................................................................................
infB-2_1  .................E..................S....................................Q.............E..M....T....
infB-3_1  ....................................................................................................
infB-4_1  ....................................................................................................
infB-5_1  ....................................................................................................
infB-6_1  ....................................................................................................
infB-7_1  .................E..................S..........................V.........Q.............E..M.........
infB-8_1  .................E..................S..........................V.........Q.............E..M.........
infB-9_1  .................E.E................S....................................Q..L..........E..M.........
infB-10_1 V................E.E................S....................................Q..L..........E..M.........
infB-11_1 V................E.E................S....................................Q..L..........E..M.........
infB-12_1 .................E..................S....................................Q..L..........E..M.........
infB-13_1 .................E..................S....................................G.............E............
infB-14_1 ....................................................................................................
infB-15_1 ....................................................................................................
infB-16_1 ....................................................................................................
infB-17_1 ....................................................................................................
infB-18_1 ....................................................................................................
infB-19_1 ....................................................................................................
infB-20_1 ....................................................................................................
infB-21_1 .................P..................................................................................
infB-22_1 ....................................................................................................
infB-23_1 ....................................................................................................
infB-24_1 ....................................................................................................
infB-25_1 ....................................................................................................
infB-26_1 ....................................................................................................
infB-27_1 ....................................................................................................
infB-28_1 ....................................................................................................
infB-29_1 ....................................................................................................
infB-30_1 ....................................................................................................
infB-31_1 ....................................................................................................
infB-32_1 ....................................................................................................
infB-33_1 ....................................................................................................
infB-34_1 ....................................................................................................
infB-35_1 ....................................................................................................
Consensus ASGEAGGITQHIGAYHVQTDNGMITFLDTPGHAAFTAMRARGAQATDIVVLVVAADDGVMPQTIEAIQHAKAAKVPVVVAVNKIDKPDADPDRVKNELSQ

                 110       120       130       140       150
          ----:----|----:----|----:----|----:----|----:----|----:
Consensus HGIIPEEWGGDCQFVHVSAKAGTGIDELLDAILLQSEVLELKAVRKG
infB-1_1  ...............................................
infB-2_1  Y.VM......ES..IP..........D..N.....A.........N.
infB-3_1  ...............................................
infB-4_1  .........................................N.....
infB-5_1  ...............................................
infB-6_1  ..........ES.............................H.....
infB-7_1  Y.VM......EA..IP....V.....D..N.....A.........N.
infB-8_1  Y.VM......EA..IP....V.....D..N.....A.........N.
infB-9_1  Y.VM......EA..IP..........D..N.....A...........
infB-10_1 Y.VM......EA..IP..........D..N.....A...........
infB-11_1 Y.VM......EA..IP..........D..N.....A...........
infB-12_1 Y.VM......EA..IP..........D..N.....A...........
infB-13_1 Y..L......ES.......................A...........
infB-14_1 ...............................................
infB-15_1 ...............................................
infB-16_1 ...............................................
infB-17_1 ...............................................
infB-18_1 ..........ES.............................H.....
infB-19_1 ...............................................
infB-20_1 ...............................................
infB-21_1 ...............................................
infB-22_1 ...............................................
infB-23_1 ...............................................
infB-24_1 ...............................................
infB-25_1 ..........ES.............................H.....
infB-26_1 ...............................................
infB-27_1 ..........ES.............................H.....
infB-28_1 ..........ES.............................H.....
infB-29_1 ..........ES.............................H.....
infB-30_1 ...............................................
infB-31_1 ...............................................
infB-32_1 ..........ES.............................H.....
infB-33_1 ...............................................
infB-34_1 ..........ES.............................H.....
infB-35_1 ...............................................
Consensus HGIIPEEWGGDCQFVHVSAKAGTGIDELLDAILLQSEVLELKAVRKG
```

So the variations in *atp*D did not result in much variation in the amino acid sequence; position 23 D and E.

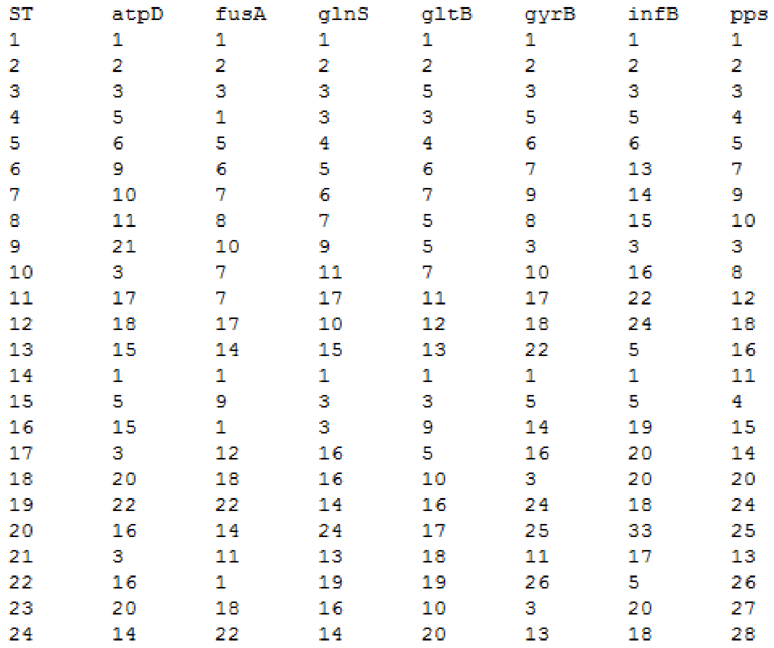
Why did it not cause much change? Think about the genetic code and its redundancy and the third base variation.

There appears to be more amino acid sequence change for *inf*B. For example, at position 18 it is Q or E. Use the Appendix (b) ‘Amino acid codes’ to look up what these amino acids are, and consider how significant these variations in the amino acid sequences are. For this example it is glutamine (Q) and glutamate (E). You may be surprised how informative this analysis can be. Remember you are using the online analysis facility. Next we’ll look at how easy the sequences can be downloaded for you own analysis.

Part 2: Data download from MLST database

Go to back to the *Cronobacter* MLST home page (http://pubmlst.org/cronobacter/), and select the ‘Profiles database’.
At the top right hand corner you’ll see ‘Downloads’ where you can either download the allele sequences, or allelic profiles.

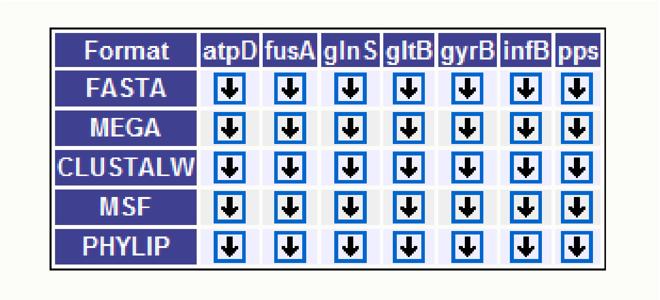
The allelic profiles look like this:



ST	atpD	fusA	glnS	gltB	gyrB	infB	pps
1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2
3	3	3	3	5	3	3	3
4	5	1	3	3	5	5	4
5	6	5	4	4	6	6	5
6	9	6	5	6	7	13	7
7	10	7	6	7	9	14	9
8	11	8	7	5	8	15	10
9	21	10	9	5	3	3	3
10	3	7	11	7	10	16	8
11	17	7	17	11	17	22	12
12	18	17	10	12	18	24	18
13	15	14	15	13	22	5	16
14	1	1	1	1	1	1	11
15	5	9	3	3	5	5	4
16	15	1	3	9	14	19	15
17	3	12	16	5	16	20	14
18	20	18	16	10	3	20	20
19	22	22	14	16	24	18	24
20	16	14	24	17	25	33	25
21	3	11	13	18	11	17	13
22	16	1	19	19	26	5	26
23	20	18	16	10	3	20	27
24	14	22	14	20	13	18	28

Which gives the allele numbers making up each Sequence Type (ST)?

What is more useful is downloading the allele sequences:



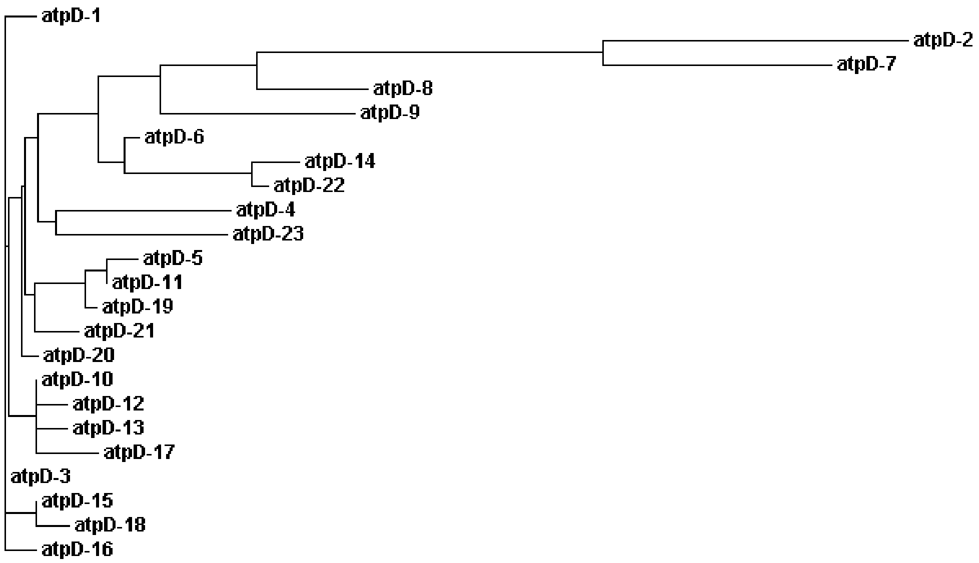
Format	atpD	fusA	glnS	gltB	gyrB	infB	pps
FASTA	⬇	⬇	⬇	⬇	⬇	⬇	⬇
MEGA	⬇	⬇	⬇	⬇	⬇	⬇	⬇
CLUSTALW	⬇	⬇	⬇	⬇	⬇	⬇	⬇
MSF	⬇	⬇	⬇	⬇	⬇	⬇	⬇
PHYLIP	⬇	⬇	⬇	⬇	⬇	⬇	⬇

By clicking on the FASTA arrow for atpD you get all the atpD sequences in the database. I'm sure you'll agree this is a lot faster than using Genbank and NCBI!

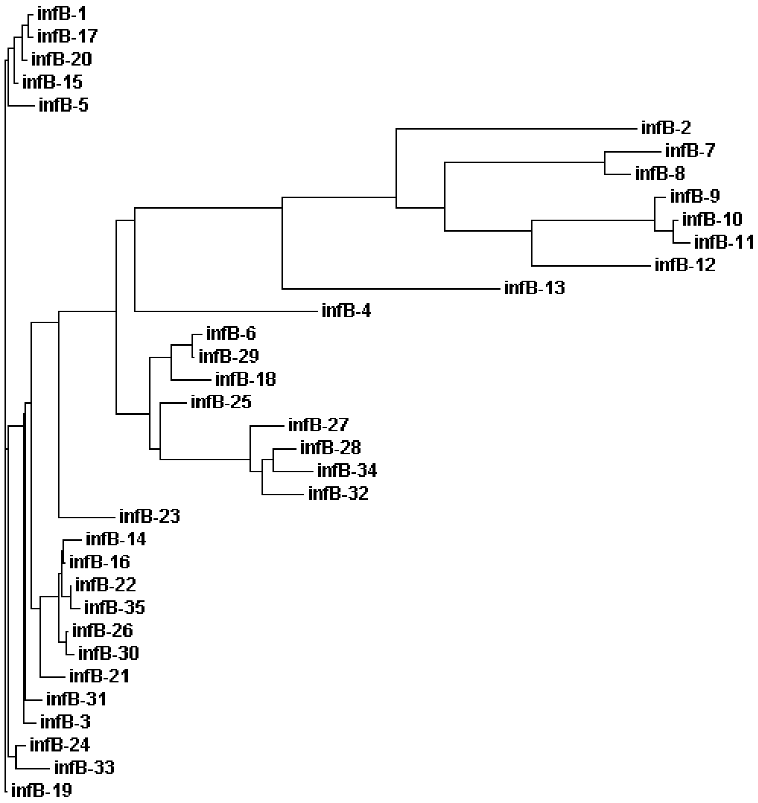
Let's complete the circle and go back to Exercise 1 to make phylogenetic trees. Download the *atp*D sequences and save in Notepad or another simple word processor, and then download the *inf*B sequences.

Use the ClustalW program at http://www.ebi.ac.uk/clustalw/index.html to make two phylogenetic trees.

*atp*D phylogram



*inf*B phylogram



You can then use text boxes to more fully annotate the trees with respect to whether certain branches are food isolates, or a particular species within a genus, etc.

Appendix : Sequence codes

a) DNA codes

Sequences are expected 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 or C (amino)
C	→ cytidine	S	→ G or C (strong)
G	→ guanine	W	→ A T (weak)
T	→ thymidine	B	→ G, T or C
U	→ uridine	D	→ G, A or T
R	→ G A (purine)	H	→ A, C or T
Y	→ T C (pyrimidine)	V	→ G, C or A
K	→ G T (keto)	N	→ A, G, C or T (any)
		-	gap of indeterminate length

b) Amino acid code

These codes are used by BLASTP and TBLASTN.

A	alanine	P	proline
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

c) Colour Code.

Be aware that this coding can change between Web sites.

Amino acid biosynthesis

Purines, pyrimidines, nucleosides, and nucleotides

Fatty acid, phospholipid and sterol metabolism

Biosynthesis of cofactors, prosthetic groups, and carriers

Central intermediary metabolism

Energy metabolism

Transport and binding proteins

DNA replication, restriction, modification, recombination, and repair

Transcription

Translation

Regulatory functions

Cell envelope

Cellular processes

Other categories

Hypothetical

Below is the black and white version of the above:

Can you see it?