PART 1: GENOME BROWSING WITH ARTEMIS

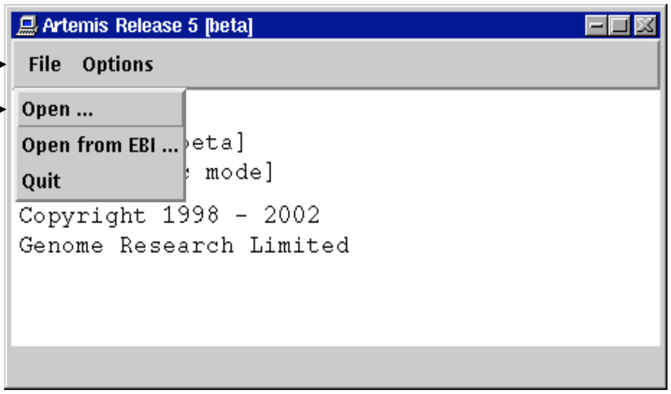
1. Starting up the Artemis software

In the Unix window type

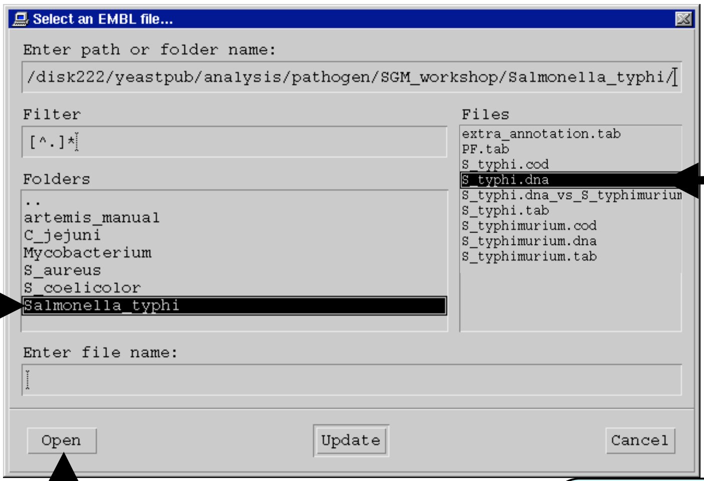
```
artemis <ret>
```

A small start-up window will appear (see below).
Now follow the sequence of numbers to load up the *Salmonella typhi* chromosome sequence.
Ask a demonstrator for help if you have any problems.

1
2
Click 'File' then 'Open



```
eta]
: mode]
Copyright 1998 - 2002
Genome Research Limited
```



```
Enter path or folder name:
/disk222/yeastpub/analysis/pathogen/SGM_workshop/Salmonella_typhi/
Filter
[^.]*
Folders
..
artemis_manual
C_jejuni
Mycobacterium
S_aureus
S_coelicolor
Salmonella_typhi
Files
extra_annotation.tab
PF.tab
S_typhi.cod
S_typhi.dna
S_typhi.dna_vs_S_typhimurium
S_typhi.tab
S_typhimurium.cod
S_typhimurium.dna
S_typhimurium.tab
Enter file name:
```

3
Double click to select Salmonella typhi directory

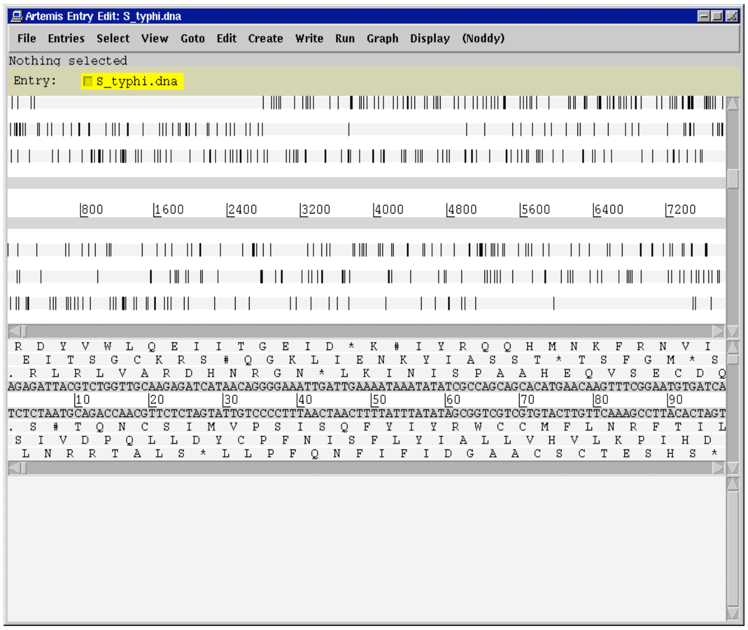
4
Single click to select DNA file

5
Single click to open file in Artemis then wait

DNA sequence files will have the suffix '.dna'. Annotation files end with '.tab'.

2. Loading annotation files (entries) into Artemis

Hopefully you will now have an Artemis window like this! If not, ask a demonstrator for assistance.

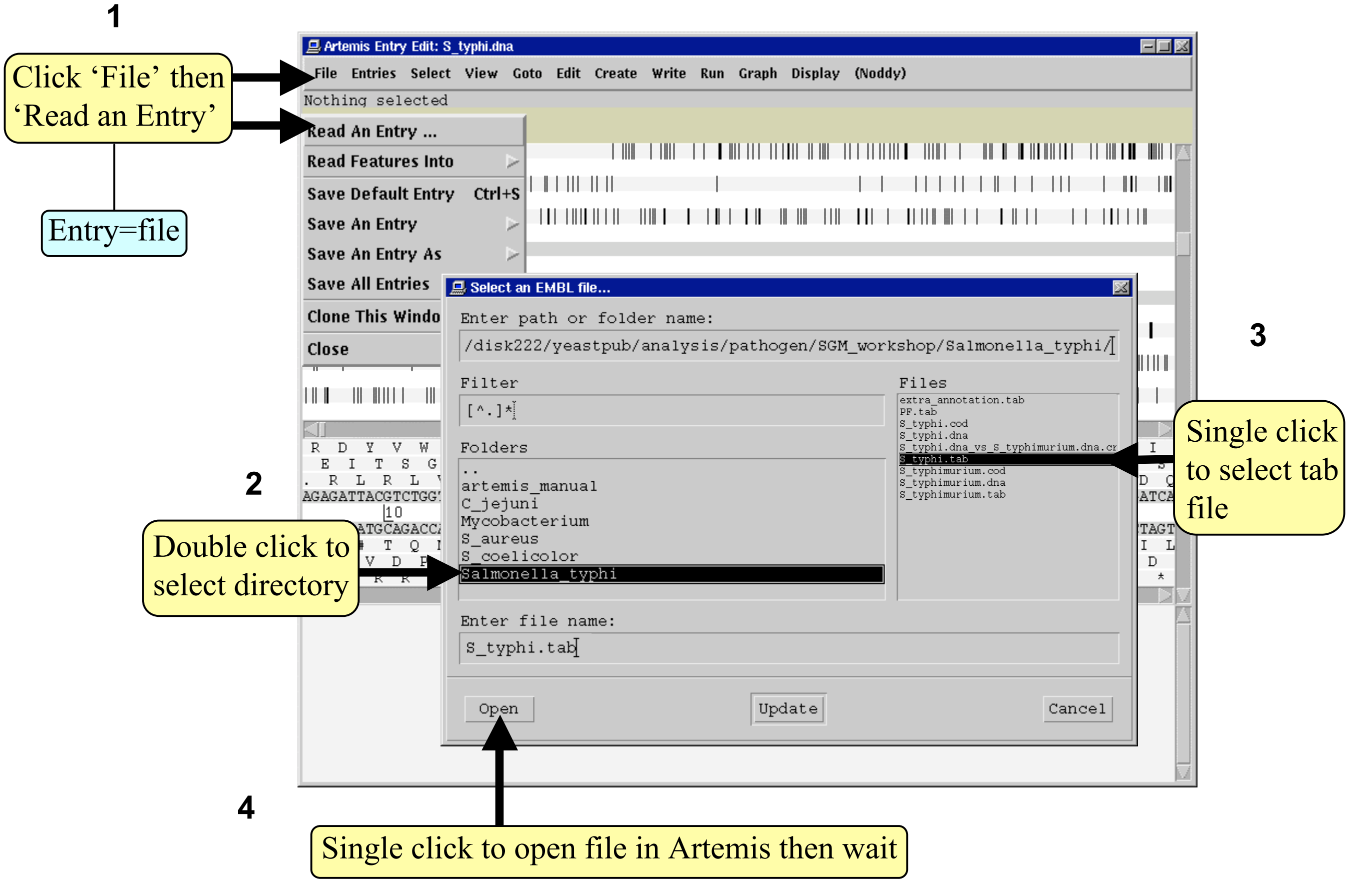

Artemis Entry Edit: S_typhi.dna
File Entries Select View Goto Edit Create Write Run Graph Display (Noddy)
|800 |1600 |2400 |3200 |4000 |4800 |5600 |6400 |7200

Nothing selected

Entry: S_typhi.dna

```
 R  D  Y  V  W  L  Q  E  I  I  T  G  E  I  D  *  K  #  I  Y  R  Q  Q  H  M  N  K  F  R  N  V  I
  E  I  T  S  G  C  K  R  S  #  Q  G  K  L  I  E  N  K  Y  I  A  S  S  T  *  T  S  F  G  M  *  S
.  R  L  R  L  V  A  R  D  H  N  R  G  N  *  L  K  I  N  I  S  P  A  A  H  E  Q  V  S  E  C  D  Q
AGAGATTACGTCTGGTTGCAAGAGATCATAACAGGGGAAATTGATTGAAAATAAATATATCGCCAGCAGCACATGAACAAGTTTCGGAATGTGATCA
         |10       |20       |30       |40       |50       |60       |70       |80       |90
TCTCTAATGCAGACCAACGTTCTCTAGTATTGTCCCCTTTAACTAACTTTTATTTATATAGCGGTCGTCGTGTACTTGTTCAAAGCCTTACACTAGT
.  S  #  T  Q  N  C  S  I  M  V  P  S  I  S  Q  F  Y  I  Y  R  W  C  C  M  F  L  N  R  F  T  I  L
 S  I  V  D  P  Q  L  L  D  Y  C  P  F  N  I  S  F  L  Y  I  A  L  L  V  H  V  L  K  P  I  H  D
  L  N  R  R  T  A  L  S  *  L  L  P  F  Q  N  F  I  F  I  D  G  A  A  C  S  C  T  E  S  H  S  *
```

Now follow the numbers to load up the annotation file for the *Salmonella typhi* chromosome.



1 Click 'File' then 'Read an Entry'

Entry=file

Nothing selected

Enter path or folder name:

/disk222/yeastpub/analysis/pathogen/SGM_workshop/Salmonella_typhi/

Filter

[^.]*

Folders

```
..
artemis_manual
C_jejuni
Mycobacterium
S_aureus
S_coelicolor
Salmonella_typhi
```

Files

```
extra_annotation.tab
PF.tab
S_typhi.cod
S_typhi.dna
S_typhi.dna_vs_S_typhimurium.dna.cr
S_typhi.tab
S_typhimurium.cod
S_typhimurium.dna
S_typhimurium.tab
```

Enter file name:

S_typhi.tab

2 Double click to select directory

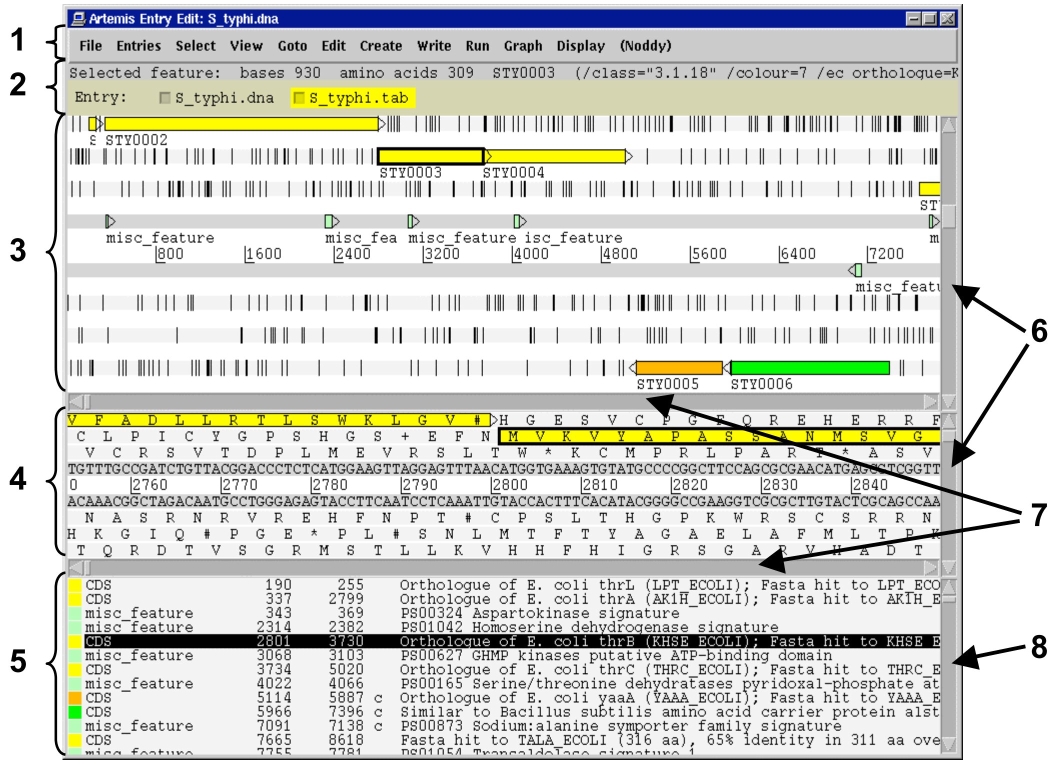
3 Single click to select tab file

4 Single click to open file in Artemis then wait

What's an "Entry"? It's a file of DNA or amino acid features which can be overlaid onto the DNA in the main Artemis view panel.

3. The basics of Artemis

Now you have an Artemis window open let's look at what's in there.


Artemis Entry Edit: S_typhi.dna
File Entries Select View Goto Edit Create Write Run Graph Display (Noddy)
Entry: S_typhi.dna S_typhi.tab
STY0002
STY0003
STY0004
STY0005
STY0006
misc_feature
misc_fea
misc_feature
isc_feature
misc_featu
800
1600
2400
3200
4000
4800
5600
6400
7200
1
2
3
4
5
6
7
8

```
Selected feature:  bases 930  amino acids 309  STY0003  (/class="3.1.18" /colour=7 /ec orthologue=K
```

```
V  F  A  D  L  L  R  T  L  S  W  K  L  G  V  # H  G  E  S  V  C  P  G  F  Q  R  E  H  E  R  R  F
 C  L  P  I  C  Y  G  P  S  H  G  S  +  E  F  N  M  V  K  V  Y  A  P  A  S  S  A  N  M  S  V  G
  V  C  R  S  V  T  D  P  L  M  E  V  R  S  L  T  W  *  K  C  M  P  R  L  P  A  R  T  *  A  S  V
TGTTTGCCGATCTGTTACGGACCCTCTCATGGAAGTTAGGAGTTTAACATGGTGAAAGTGTATGCCCCGGCTTCCAGCGCGAACATGAGCGTCGGTT
0     |2760     |2770     |2780     |2790     |2800     |2810     |2820     |2830     |2840
ACAAACGGCTAGACAATGCCTGGGAGAGTACCTTCAATCCTCAAATTGTACCACTTTCACATACGGGGCCGAAGGTCGCGCTTGTACTCGCAGCCAA
  N  A  S  R  N  R  V  R  E  H  F  N  P  T  #  C  P  S  L  T  H  G  P  K  W  R  S  C  S  R  R  N
H  K  G  I  Q  #  P  G  E  *  P  L  #  S  N  L  M  T  F  T  Y  A  G  A  E  L  A  F  M  L  T  P  K
 T  Q  R  D  T  V  S  G  R  M  S  T  L  L  K  V  H  H  F  H  I  G  R  S  G  A  R  V  H  A  D  T
```

```
CDS              190     255     Orthologue of E. coli thrL (LPT_ECOLI); Fasta hit to LPT_ECO
CDS              337    2799     Orthologue of E. coli thrA (AK1H_ECOLI); Fasta hit to AK1H_E
misc_feature     343     369     PS00324 Aspartokinase signature
misc_feature    2314    2382     PS01042 Homoserine dehydrogenase signature
CDS             2801    3730     Orthologue of E. coli thrB (KHSE_ECOLI); Fasta hit to KHSE E
misc_feature    3068    3103     PS00627 GHMP kinases putative ATP-binding domain
CDS             3734    5020     Orthologue of E. coli thrC (THRC_ECOLI); Fasta hit to THRC_E
misc_feature    4022    4066     PS00165 Serine/threonine dehydratases pyridoxal-phosphate at
CDS             5114    5887 c   Orthologue of E. coli yaaA (YAAA_ECOLI); Fasta hit to YAAA_E
CDS             5966    7396 c   Similar to Bacillus subtilis amino acid carrier protein alst
misc_feature    7091    7138 c   PS00873 Sodium:alanine symporter family signature
CDS             7665    8618     Fasta hit to TALA_ECOLI (316 aa), 65% identity in 311 aa ove
misc_feature    7755    7781     PS01054 Transaldolase signature 1
```

1. Drop-down menus. There's lots in there so don't worry about them right now.
2. Shows what entries are currently loaded (bottom line) and gives details regarding the feature selected in the window below; in this case gene STY0003 (top line).
3. This is the Main Sequence View panel showing. The central 2 grey lines represent the forward (top) and reverse (bottom) DNA strands. Above and below those are the 3 forward and 3 reverse reading frames. Stop codons are marked as black vertical bars. Genes and other features (eg. Pfam and Prosite matches) are displayed as coloured boxes.
4. This panel has a similar layout to the main panel but is zoomed in to show nucleotides and amino acids. Double click on a gene in the main view to see the zoomed view of the start of that gene. Note that both this and the main panel can be scrolled left and right (7, below) zoomed in and out (6, below).
5. This panel lists the various features in the order that they occur on the DNA with the selected gene highlighted. The list can be scrolled (8, below).
6. Sliders for zooming view panels.
7. Sliders for scrolling along the DNA.
8. Slider for scrolling feature list.

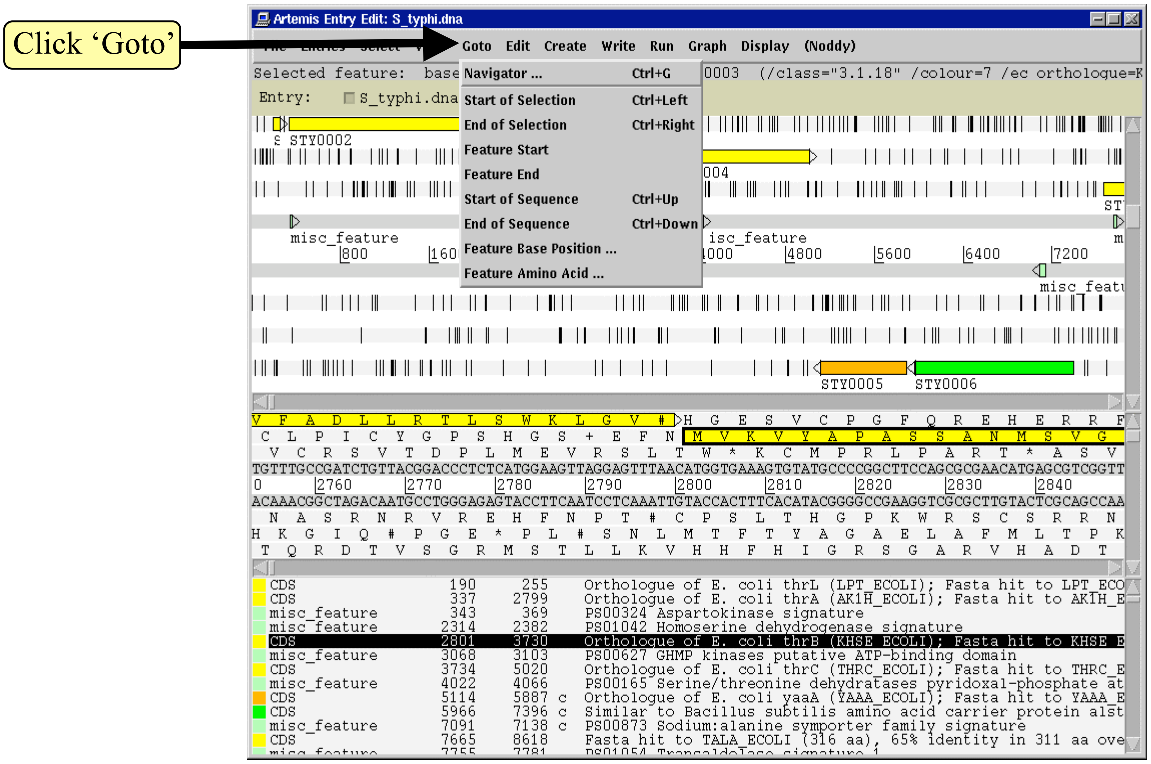
For 6, 7 and 8 the sliders work like those in Windows. It is possible to zoom out to view a whole genome! Try it out but be careful, zooming out quickly with lots of features displayed may temporarily lock up the display. Switch off feature files using the buttons in 2 and switch off stop codons by right mouse botton clicking in the main view panel. If you have

4. Getting around in Artemis

The 3 main ways of getting to where you want to be in Artemis are the Goto dropdown menu, the Navigator and the Feature Selector. The best method depends on what you're trying to do and knowing which one to use comes with practice.

4.1 The 'Goto'menu

The functions on this menu (ignore the Navigator for now) are shortcuts for getting to locations within a selected feature or for jumping to the start or end of the DNA sequence. This one's really intuitive so give it a try!


Click 'Goto'

Artemis Entry Edit: S_typhi.dna
File Entries Select Goto Edit Create Write Run Graph Display (Noddy)
Selected feature: base ... 0003 (/class="3.1.18" /colour=7 /ec ortholoque=K
Entry: S_typhi.dna
Navigator ... Ctrl+G
Start of Selection Ctrl+Left
End of Selection Ctrl+Right
Feature Start
Feature End
Start of Sequence Ctrl+Up
End of Sequence Ctrl+Down
Feature Base Position ...
Feature Amino Acid ...
STY0002
004
misc_feature
isc_feature
800
160
4000
4800
5600
6400
7200
misc_featu
STY0005
STY0006

```
V  F  A  D  L  L  R  T  L  S  W  K  L  G  V  # H  G  E  S  V  C  P  G  F  Q  R  E  H  E  R  R  F
 C  L  P  I  C  Y  G  P  S  H  G  S  +  E  F  N  M  V  K  V  Y  A  P  A  S  S  A  N  M  S  V  G
  V  C  R  S  V  T  D  P  L  M  E  V  R  S  L  T  W  *  K  C  M  P  R  L  P  A  R  T  *  A  S  V
TGTTTGCCGATCTGTTACGGACCCTCTCATGGAAGTTAGGAGTTTAACATGGTGAAAGTGTATGCCCCGGCTTCCAGCGCGAACATGAGCGTCGGTT
0    |2760     |2770     |2780     |2790     |2800     |2810     |2820     |2830     |2840
ACAAACGGCTAGACAATGCCTGGGAGAGTACCTTCAATCCTCAAATTGTACCACTTTCACATACGGGGCCGAAGGTCGCGCTTGTACTCGCAGCCAA
  N  A  S  R  N  R  V  R  E  H  F  N  P  T  #  C  P  S  L  T  H  G  P  K  W  R  S  C  S  R  R  N
H  K  G  I  Q  #  P  G  E  *  P  L  #  S  N  L  M  T  F  T  Y  A  G  A  E  L  A  F  M  L  T  P  K
 T  Q  R  D  T  V  S  G  R  M  S  T  L  L  K  V  H  H  F  H  I  G  R  S  G  A  R  V  H  A  D  T
```

```
CDS                190    255    Orthologue of E. coli thrL (LPT_ECOLI); Fasta hit to LPT_ECO
CDS                337   2799    Orthologue of E. coli thrA (AK1H_ECOLI); Fasta hit to AK1H_E
misc_feature       343    369    PS00324 Aspartokinase signature
misc_feature      2314   2382    PS01042 Homoserine dehydrogenase signature
CDS               2801   3730    Orthologue of E. coli thrB (KHSE_ECOLI); Fasta hit to KHSE_E
misc_feature      3068   3103    PS00627 GHMP kinases putative ATP-binding domain
CDS               3734   5020    Orthologue of E. coli thrC (THRC_ECOLI); Fasta hit to THRC_E
misc_feature      4022   4066    PS00165 Serine/threonine dehydratases pyridoxal-phosphate at
CDS               5114   5887 c  Orthologue of E. coli yaaA (YAAA_ECOLI); Fasta hit to YAAA_E
CDS               5966   7396 c  Similar to Bacillus subtilis amino acid carrier protein alst
misc_feature      7091   7138 c  PS00873 Sodium:alanine symporter family signature
CDS               7665   8618    Fasta hit to TALA_ECOLI (316 aa), 65% identity in 311 aa ove
misc_feature      7755   7781    PS01054 Transaldolase signature 1
```

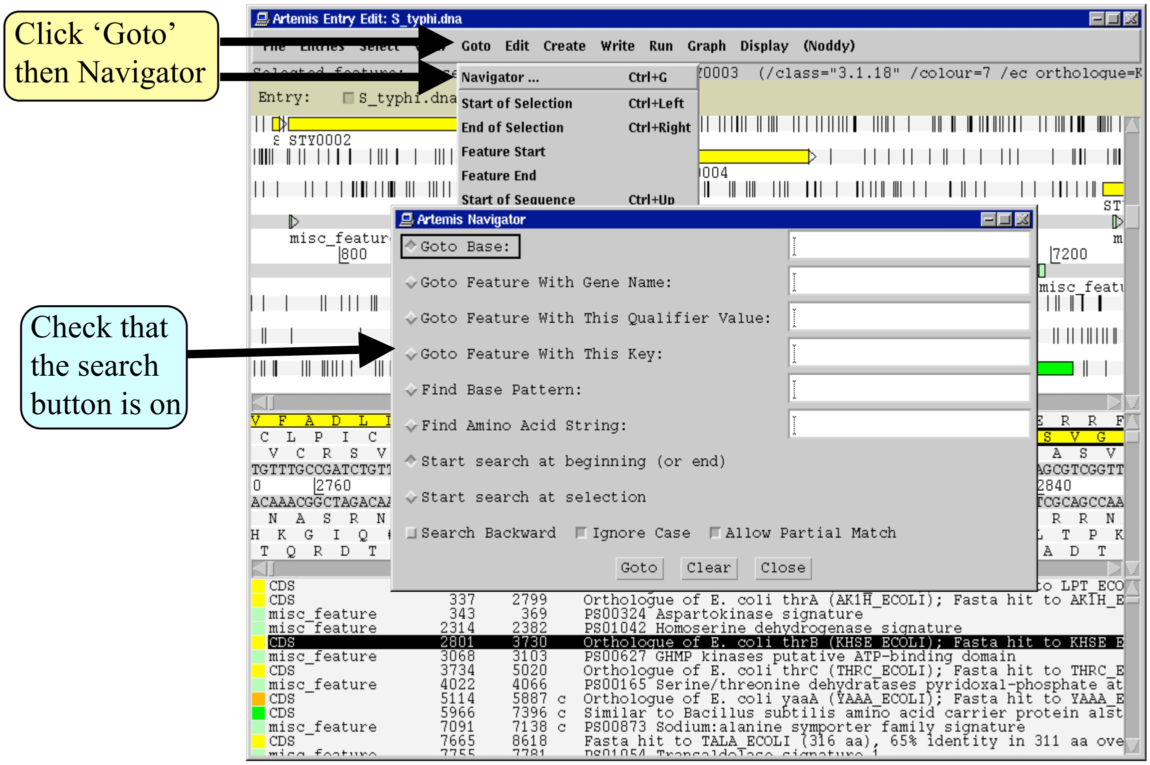
It may seem that 'Goto' 'Start of Selection' and 'Goto' 'Feature Start' do the same thing. Well they do if you have a feature selected but 'Goto' 'Start of Selection' will also work for a region which you have highlighted by click-dragging in the main window. So yes, give it a try!

Suggested tasks:

1. Zoom out, highlight a large region by clicking the left hand button and dragging the cursor then go to the start and end of the highlighted region.
2. Select a gene then go to the start and end.
3. Go to the start and end of the genome sequence.
4. Select a gene. Within it, go to a base (nucleotide) and/or amino acid of your choice.

4.2 Navigator

The Navigator panel is fairly intuitive so open it up and give it a try.


Artemis Entry Edit: S_typhi.dna
File Entries Select View Goto Edit Create Write Run Graph Display (Noddy)
Navigator ... Ctrl+G
Start of Selection Ctrl+Left
End of Selection Ctrl+Right
Feature Start
Feature End
Start of Sequence Ctrl+Up
Artemis Navigator
Goto Base:
Goto Feature With Gene Name:
Goto Feature With This Qualifier Value:
Goto Feature With This Key:
Find Base Pattern:
Find Amino Acid String:
Start search at beginning (or end)
Start search at selection
Search Backward
Ignore Case
Allow Partial Match
Goto
Clear
Close

Click ‘Goto’ then Navigator

Check that the search button is on

```
CDS
CDS                  337    2799     Orthologue of E. coli thrA (AK1H_ECOLI); Fasta hit to AK1H_E
misc_feature         343     369     PS00324 Aspartokinase signature
misc_feature        2314    2382     PS01042 Homoserine dehydrogenase signature
CDS                 2801    3730     Orthologue of E. coli thrB (KHSE_ECOLI); Fasta hit to KHSE E
misc_feature        3068    3103     PS00627 GHMP kinases putative ATP-binding domain
CDS                 3734    5020     Orthologue of E. coli thrC (THRC_ECOLI); Fasta hit to THRC_E
misc_feature        4022    4066     PS00165 Serine/threonine dehydratases pyridoxal-phosphate at
CDS                 5114    5887 c   Orthologue of E. coli yaaA (YAAA_ECOLI); Fasta hit to YAAA_E
CDS                 5966    7396 c   Similar to Bacillus subtilis amino acid carrier protein alst
misc_feature        7091    7138 c   PS00873 Sodium:alanine symporter family signature
CDS                 7665    8618     Fasta hit to TALA_ECOLI (316 aa), 65% identity in 311 aa ove
```

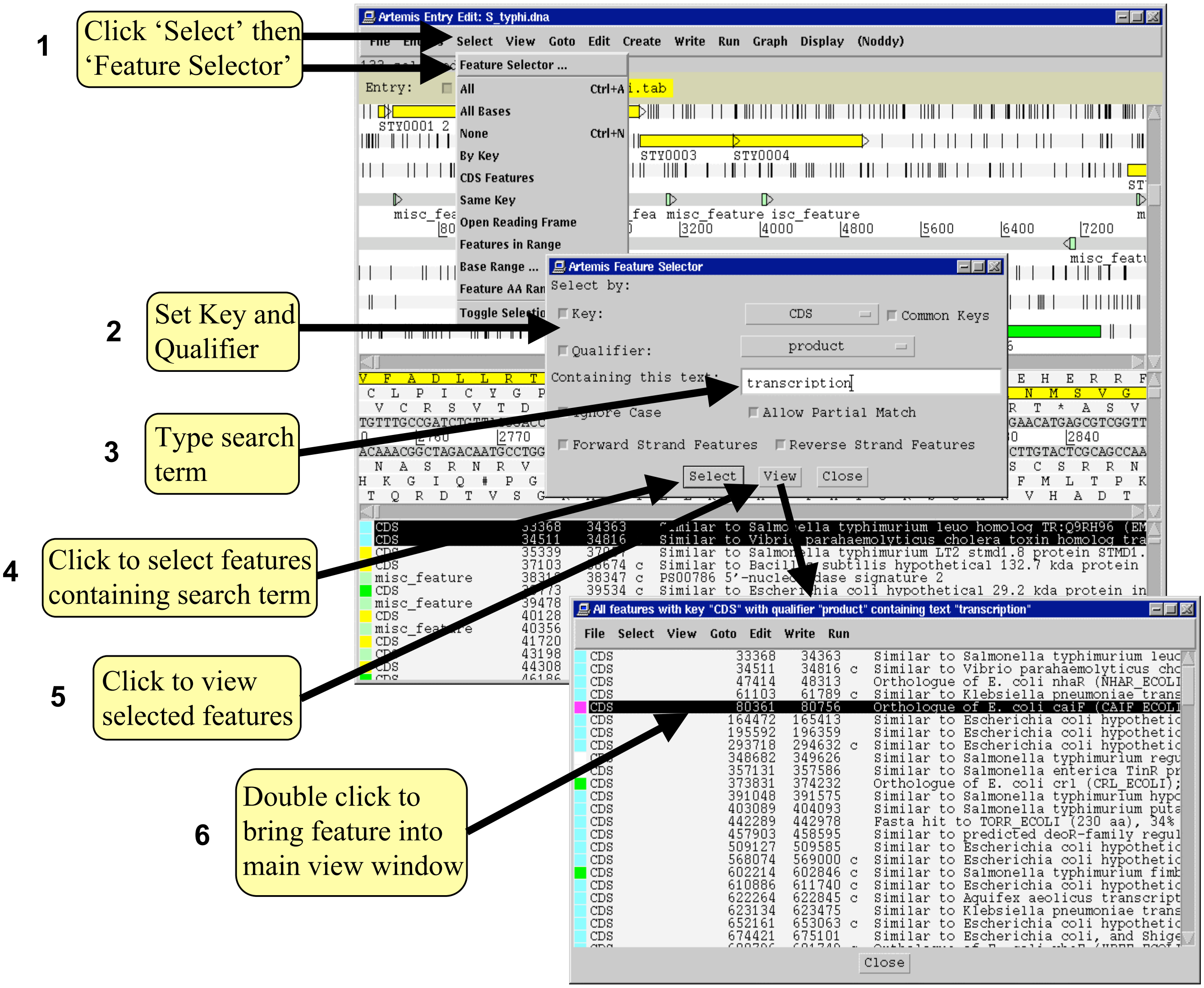
Suggestions of where to go:

1. Think of a number between 1 and 4809037 and go to that base (notice how the cursors on the horizontal sliders move with you).
2. Your favourite gene name (it may not be there so you could try ‘fts’).
3. Use ‘Goto Feature With This Qualifier value’ to search the contents of all qualifiers for a particular term. For example using the word ‘pseudogene’ will take you to the next feature with the word ‘pseudogene’ in any of its qualifiers. Note how repeated clicking of the ‘Goto’ button takes you through the pseudogenes as they occur on the chromsome.
4. tRNA genes. Type ‘tRNA’ in the ‘Goto Feature With This Key’.
5. Regulator-binding DNA consensus sequence (real or made up!). Note that degenerate base values can be used (Appendix I).
6. Amino acid consensus sequences (real or made up!). You can use ‘X’s. Note that it searches all six reading frames regardless of whether the amino acids are encoded or not.

What are Keys and Qualifiers? See Appendix II

4.3 Feature Selector

This is a great tool for selecting/extracting a subset of features from the genome for viewing as a set or individually. Again the panel is fairly intuitive so open it up and give it a try.



1 Click ‘Select’ then ‘Feature Selector’

2 Set Key and Qualifier

3 Type search term

4 Click to select features containing search term

5 Click to view selected features

6 Double click to bring feature into main view window

☐ Key: CDS ☐ Common Keys
☐ Qualifier: product
Containing this text: transcription
☐ Ignore Case ☐ Allow Partial Match
☐ Forward Strand Features ☐ Reverse Strand Features

All features with key "CDS" with qualifier "product" containing text "transcription"

```
CDS        33368   34363     Similar to Salmonella typhimurium leuc
CDS        34511   34816 c   Similar to Vibrio parahaemolyticus chc
CDS        47414   48313     Orthologue of E. coli nhaR (NHAR_ECOLI
CDS        61103   61789 c   Similar to Klebsiella pneumoniae trans
CDS        80361   80756     Orthologue of E. coli caiF (CAIF ECOLI
CDS       164472  165413     Similar to Escherichia coli hypothetic
CDS       195592  196359     Similar to Escherichia coli hypothetic
CDS       293718  294632 c   Similar to Escherichia coli hypothetic
CDS       348682  349626     Similar to Salmonella typhimurium regu
CDS       357131  357586     Similar to Salmonella enterica TinR pr
CDS       373831  374232     Orthologue of E. coli crl (CRL_ECOLI);
CDS       391048  391575     Similar to Salmonella typhimurium hypc
CDS       403089  404093     Similar to Salmonella typhimurium puta
CDS       442289  442978     Fasta hit to TORR_ECOLI (230 aa), 34%
CDS       457903  458595     Similar to predicted deoR-family regul
CDS       509127  509585     Similar to Escherichia coli hypothetic
CDS       568074  569000 c   Similar to Escherichia coli hypothetic
CDS       602214  602846 c   Similar to Salmonella typhimurium fimb
CDS       610886  611740 c   Similar to Escherichia coli hypothetic
CDS       622264  622845 c   Similar to Aquifex aeolicus transcript
CDS       623134  623475     Similar to Klebsiella pneumoniae trans
CDS       652161  653063 c   Similar to Escherichia coli hypothetic
CDS       674421  675101     Similar to Escherichia coli, and Shige
```

Suggested missions:

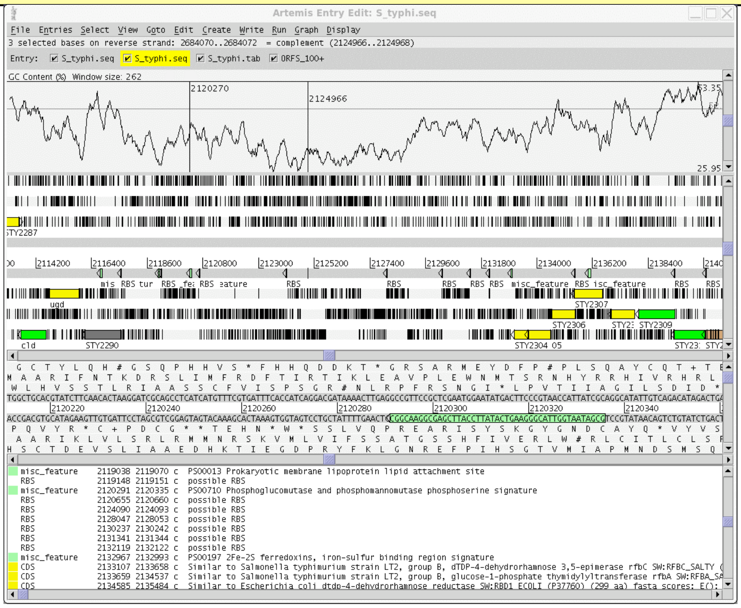
1. Select and view all ‘tRNAs’
2. Select and view all ‘repeat_regions’
3. Select and view all CDSs with ‘fts’ in the gene name
4. Select and view all CDSs with ‘fl’ in the gene name
5. View the genome distribution of a subset of genes. To do this you will need to create a new entry (click ‘Create’ then ‘New Entry’), then copy the selected subset of genes to this New Entry (click ‘Edit’ then ‘Copy selected Features To’ then ‘no name’). Now inactivate the ‘S_typhi.tab’ file (click the button next to the file name just below the menu bar). Remove all stop codons (right button click in main view panel then click ‘Stop Codons’. Now zoom out!

PART 2: GENOME ANNOTATION WITH ARTEMIS

Clearly there are many more features of Artemis which we will not have time to explain in detail. Before getting on with this next section it might be worth browsing the menu features. Hopefully you will find most of them easy to understand.

This exercise uses the files and data you already have loaded into Artemis from the previous section. By a method of your choice go to the region located between bases 2117712 to 2133101 on the DNA sequence - it is bordered by the *ugd* gene which codes for UDP-glucose-6-dehydrogenase. You can use either the Navigator, Feature Selector or Goto functions, discussed previously, to get there. The region you arrive at should look like the window below.

The completed annotation has been removed from this region. Select the region containing the empty ORFs by clicking and dragging on the display then goto "CREATE" and "MARK ORFS IN RANGE" . Click OK in the next window and the ORFS should fill up with blue boxes



```
3 selected bases on reverse strand: 2684070..2684072  = complement (2124966..2124968)
```

```
misc_feature   2119038 2119070 c  PS00013 Prokaryotic membrane lipoprotein lipid attachment site
RBS            2119148 2119151 c  possible RBS
misc_feature   2120291 2120335 c  PS00710 Phosphoglucomutase and phosphomannomutase phosphoserine signature
RBS            2120655 2120660 c  possible RBS
RBS            2124090 2124093 c  possible RBS
RBS            2128047 2128053 c  possible RBS
RBS            2130237 2130242 c  possible RBS
RBS            2131341 2131344 c  possible RBS
RBS            2132119 2132122 c  possible RBS
misc_feature   2132967 2132993 c  PS00197 2Fe-2S ferredoxins, iron-sulfur binding region signature
CDS            2133107 2133658 c  Similar to Salmonella typhimurium strain LT2, group B, dTDP-4-dehydrorhamnose 3,5-epimerase rfbC SW:RFBC_SALTY (
CDS            2133659 2134537 c  Similar to Salmonella typhimurium strain LT2, group B, glucose-1-phosphate thymidylyltransferase rfbA SW:RFBA_SA
CDS            2134585 2135484 c  Similar to Escherichia coli dtdp-4-dehydrorhamnose reductase SW:RBD1_ECOLI (P37760) (299 aa) fasta scores: E():
```

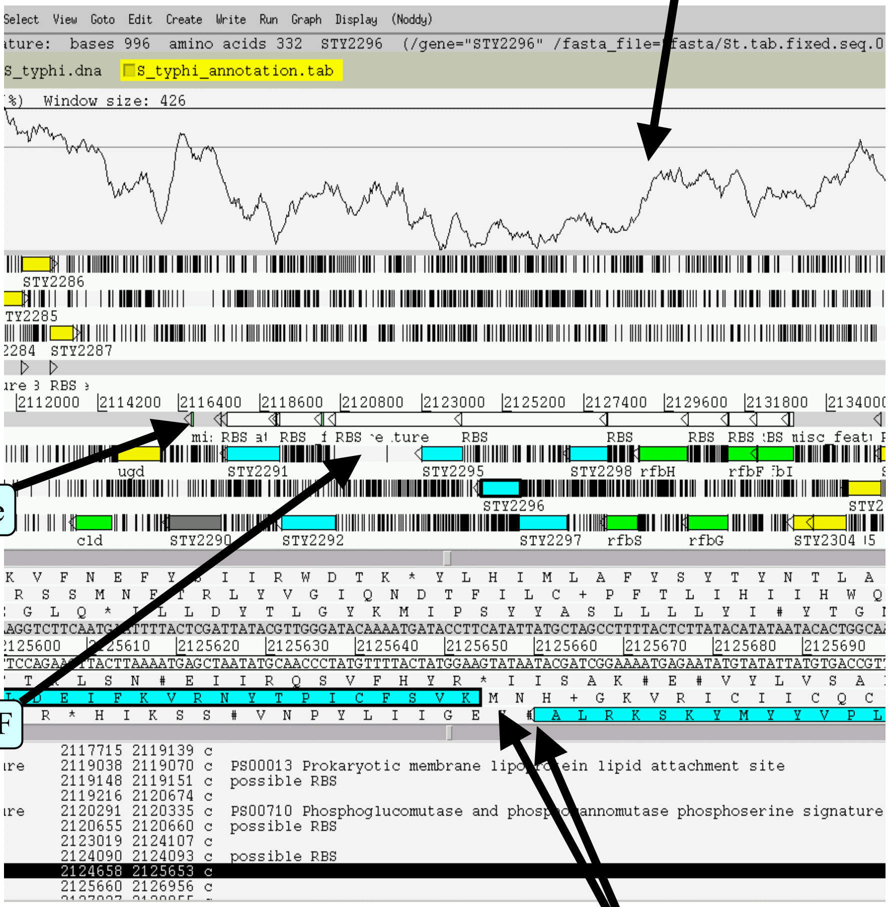
Things to do:

•Check the translational start and stop site. Does the start codon seem sensible? Does it have a reasonable upstream ribosome-binding site? Does the CDS end on a stop codon? There are several methods for adjusting the CDS coordinates:

> •Click 'Edit' then 'Trim Selected Features to……..'. This will automatically trim start codons to either ATG or other allowable start codon.
>
> •Click 'Edit' then 'Fix Stop Codons'
>
> •Click 'Edit' then 'Edit Selected Features and manually alter the coordinates in the 'location' box.
>
> •Goto the Artemis start-up window, click 'Options' then 'Enable Direct Editing'. This now allows you to drag the start/stop location by clicking and holding down the left hand mouse button.

- Has the genefinder missed any CDSs? There are several methods for creating a new feature (don’t forget to check the start/stop of the CDS you create):
    - Double click middle button within empty ORF then click ‘Create’ then ‘Create Feature From Base Range’
    - Left button click and drag to highlight a region then click ‘Create’ then ‘Mark ORFs In Range’
    - Click ‘Create’ then ‘New Feature’ and type in the desired coordinates in the ‘location’ box
- Compare your predicted CDSs with others in database. To run BLASTP searches select a CDS (hold down shift if you want to select more than one), click ‘Run’ then ‘Run BLASTP on selected features. When the search is done, a banner will appear saying ‘fasta process completed’. To view the search results click ‘View’ then ‘Search Results’ then ‘fasta results’. The results will appear in a scrollable window. How does you predicted start/stop compare with similar proteins in the database?

Things to notice



```
ature:  bases 996  amino acids 332  STY2296  (/gene="STY2296" /fasta_file=fasta/St.tab.fixed.seq.0
%)  Window size: 426
```

```
     2117715 2119139 c
ure  2119038 2119070 c  PS00013 Prokaryotic membrane lipoprotein lipid attachment site
     2119148 2119151 c  possible RBS
     2119216 2120674 c
ure  2120291 2120335 c  PS00710 Phosphoglucomutase and phosphomannomutase phosphoserine signature
     2120655 2120660 c  possible RBS
     2123019 2124107 c
     2124090 2124093 c  possible RBS
     2124658 2125653 c
     2125660 2126956 c
```

GC content graph

Prosite feature

Empty ORF

e.g’s of start and stop codons to fix

Other information to view:

Plots/Graphs

•DNA plots. These are listed in the 'Graph' menu.

Click 'Graph' then 'GC Content (%)'. Notice how the GC content is lower in the region you are annotating. You may need to zoom out a little to fully appreciate the variation. This could indicate laterally acquired DNA.

•Feature plots. These can be displayed by selecting a feature then click 'View' and 'Show Feature Plots'. The window which appears shows plots predicting hydrophobicity, hydrophilicity and coiled-coil regions for the protein product of the selected CDS.

Load additional files

The results from Prosite searches should already be on display as pale-green boxes on the grey DNA lines.The results from the Pfam protein motif searches can be viewed loading the appropriate file. Click on 'File' then 'Read an Entry' and select the appropriate file (PF.tab). Each Pfam match will appear as a feature in the main display panel on the grey DNA lines. To see the details click the feature then click 'View' then 'View Selection' or click 'Edit' then Edit Selected Features'. Please ask if you are unsure about Prosite and Pfam.

Further information on specific Prosite or Pfam entries can be found on the web at

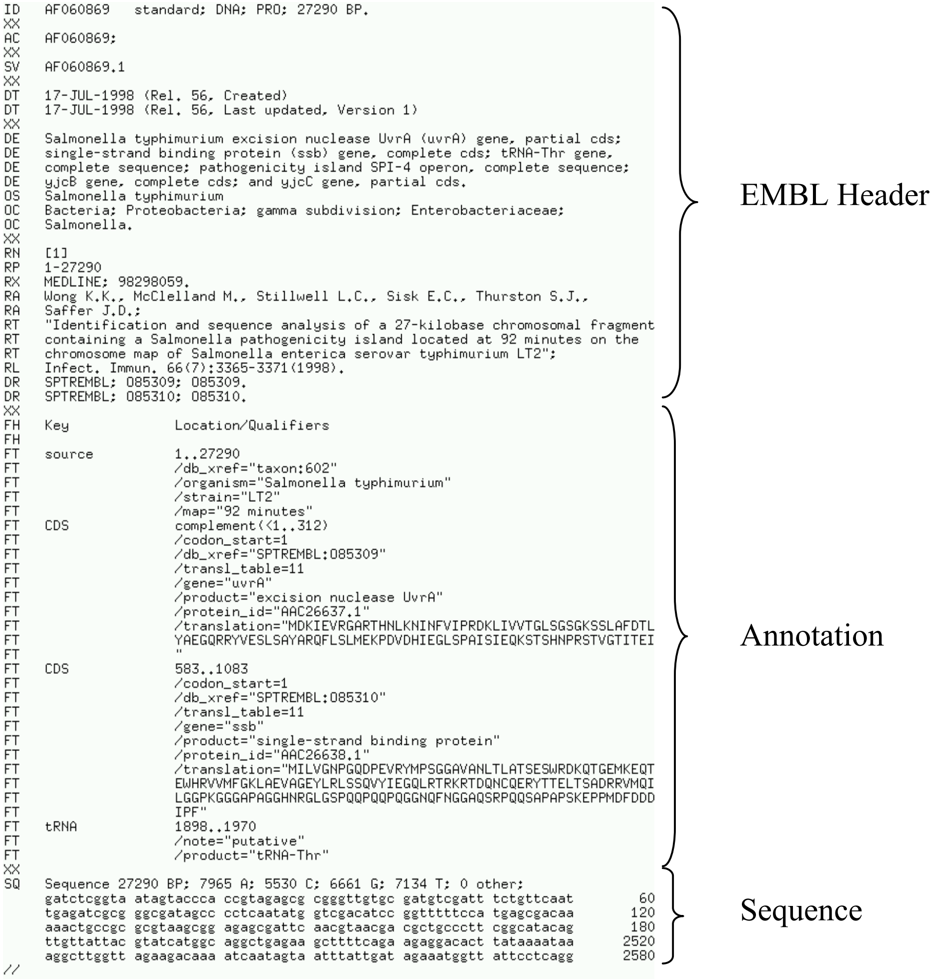
http://www.expasy.ch/prosite and http://www.sanger.ac.uk/software/Pfam/tsearch.shtml

Bearing all the information in mind, you can now create your own annotation. Highlight a CDS then click 'Edit' then 'Edit Selected Features'. In this window you can 'Add Qualifiers' (top right). Aim to include '/note', '/gene' and '/product' as a minimum. The '/note' qualifier allows you to describe the gene in free text and gives you the opportunity to explain your '/product' prediction.

To see how this region was originally annotated load in the file 'extra_annotation.tab'

PART 3: OTHER EXERCISES

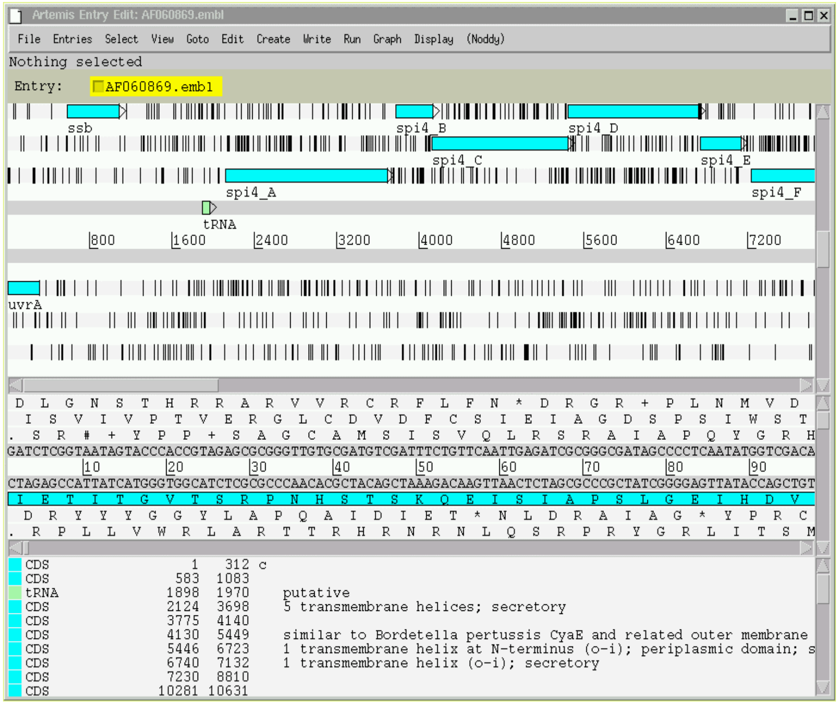
Exercise One:
Perhaps one of the most routinely useful tasks that Artemis can be used to perform is the presentation of data from the EMBL, Genbank or DDBJ databases. Having performed a random transposon mutagenesis you may have obtained some sequence for the gene into which a transposon has inserted. You can then use this sequence to search the public databases and identify similar genes. The complete, original database submission for such a gene can be downloaded from the database and viewed in Artemis.
In the embl_files directory there is just such a file called AF060869.embl. This is simply a text file as shown below:



```
ID   AF060869   standard; DNA; PRO; 27290 BP.
XX
AC   AF060869;
XX
SV   AF060869.1
XX
DT   17-JUL-1998 (Rel. 56, Created)
DT   17-JUL-1998 (Rel. 56, Last updated, Version 1)
XX
DE   Salmonella typhimurium excision nuclease UvrA (uvrA) gene, partial cds;
DE   single-strand binding protein (ssb) gene, complete cds; tRNA-Thr gene,
DE   complete sequence; pathogenicity island SPI-4 operon, complete sequence;
DE   yjcB gene, complete cds; and yjcC gene, partial cds.
OS   Salmonella typhimurium
OC   Bacteria; Proteobacteria; gamma subdivision; Enterobacteriaceae;
OC   Salmonella.
XX
RN   [1]
RP   1-27290
RX   MEDLINE; 98298059.
RA   Wong K.K., McClelland M., Stillwell L.C., Sisk E.C., Thurston S.J.,
RA   Saffer J.D.;
RT   "Identification and sequence analysis of a 27-kilobase chromosomal fragment
RT   containing a Salmonella pathogenicity island located at 92 minutes on the
RT   chromosome map of Salmonella enterica serovar typhimurium LT2";
RL   Infect. Immun. 66(7):3365-3371(1998).
DR   SPTREMBL; O85309; O85309.
DR   SPTREMBL; O85310; O85310.
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..27290
FT                   /db_xref="taxon:602"
FT                   /organism="Salmonella typhimurium"
FT                   /strain="LT2"
FT                   /map="92 minutes"
FT   CDS             complement(<1..312)
FT                   /codon_start=1
FT                   /db_xref="SPTREMBL:O85309"
FT                   /transl_table=11
FT                   /gene="uvrA"
FT                   /product="excision nuclease UvrA"
FT                   /protein_id="AAC26637.1"
FT                   /translation="MDKIEVRGARTHNLKNINFVIPRDKLIVVTGLSGSGKSSLAFDTL
FT                   YAEGQRRYVESLSAYARQFLSLMEKPDVDHIEGLSPAISIEQKSTSHNPRSTVGTITEI
FT                   "
FT   CDS             583..1083
FT                   /codon_start=1
FT                   /db_xref="SPTREMBL:O85310"
FT                   /transl_table=11
FT                   /gene="ssb"
FT                   /product="single-strand binding protein"
FT                   /protein_id="AAC26638.1"
FT                   /translation="MILVGNPGQDPEVRYMPSGGAVANLTLATSESWRDKQTGEMKEQT
FT                   EWHRVVMFGKLAEVAGEYLRLSSQVYIEGQLRTRKRTDQNCQERYTTELTSADRRVMQI
FT                   LGGPKGGGAPAGGHNRGLGSPQQPQQPQGGNQFNGGAQSRPQQSAPAPSKEPPMDFDDD
FT                   IPF"
FT   tRNA            1898..1970
FT                   /note="putative"
FT                   /product="tRNA-Thr"
XX
SQ   Sequence 27290 BP; 7965 A; 5530 C; 6661 G; 7134 T; 0 other;
     gatctcggta atagtaccca ccgtagagcg cgggttgtgc gatgtcgatt tctgttcaat        60
     tgagatcgcg ggcgatagcc cctcaatatg gtcgacatcc ggtttttcca tgagcgacaa       120
     aaactgccgc gcgtaagcgg agagcgattc aacgtaacga cgctgccctt cggcatacag       180
     ttgttattac gtatcatggc aggctgagaa gcttttcaga agaggacact tataaaataa      2520
     aggcttggtt agaagacaaa atcaatagta atttattgat agaaatggtt attcctcagg      2580
//
```

Artemis can read, understand and display the EMBL/Genbank/DDBJ file graphically. This allows you to view the published results more clearly and, perhaps more importantly, allows you to carry out further analyses.

This is what you will see if you ‘load up’ the EMBL file into Artemis. So clicking on a gene feature (blue box) and by clicking on the Edit menu and selecting Edit Selected Feature you can view all the annotation that is associated with that gene, taken directly from the EMBL text file.



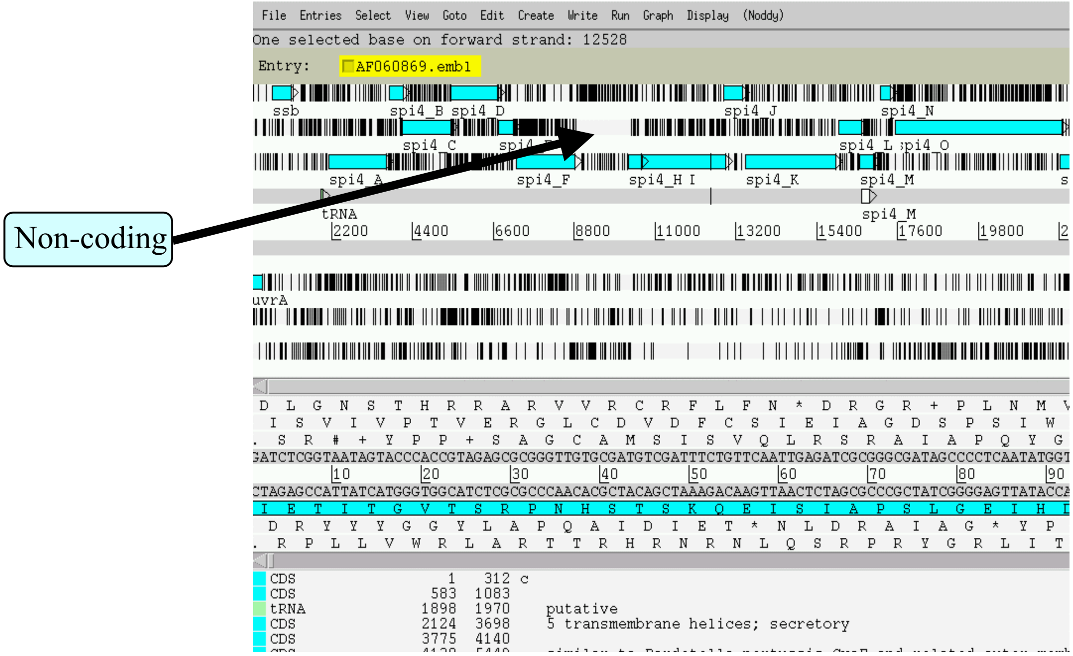
```
Nothing selected
Entry:  AF060869.embl

 D  L  G  N  S  T  H  R  R  A  R  V  V  R  C  R  F  L  F  N  *  D  R  G  R  +  P  L  N  M  V  D
  I  S  V  I  V  P  T  V  E  R  G  L  C  D  V  D  F  C  S  I  E  I  A  G  D  S  P  S  I  W  S  T
.  S  R  #  +  Y  P  P  +  S  A  G  C  A  M  S  I  S  V  Q  L  R  S  R  A  I  A  P  Q  Y  G  R  H
GATCTCGGTAATAGTACCCACCGTAGAGCGCGGGTTGTGCGATGTCGATTTCTGTTCAATTGAGATCGCGGGCGATAGCCCCTCAATATGGTCGACA
         |10       |20       |30       |40       |50       |60       |70       |80       |90
CTAGAGCCATTATCATGGGTGGCATCTCGCGCCCAACACGCTACAGCTAAAGACAAGTTAACTCTAGCGCCCGCTATCGGGGAGTTATACCAGCTGT
 I  E  T  I  T  G  V  T  S  R  P  N  H  S  T  S  K  Q  E  I  S  I  A  P  S  L  G  E  I  H  D  V
  D  R  Y  Y  Y  G  G  Y  L  A  P  Q  A  I  D  I  E  T  *  N  L  D  R  A  I  A  G  *  Y  P  R  C
.  R  P  L  L  V  W  R  L  A  R  T  T  R  H  R  N  R  N  L  Q  S  R  P  R  Y  G  R  L  I  T  S  M

CDS                 1   312  c
CDS               583  1083
tRNA             1898  1970    putative
CDS              2124  3698    5 transmembrane helices; secretory
CDS              3775  4140
CDS              4130  5449    similar to Bordetella pertussis CyaE and related outer membrane
CDS              5446  6723    1 transmembrane helix at N-terminus (o-i); periplasmic domain; s
CDS              6740  7132    1 transmembrane helix (o-i); secretory
CDS              7230  8810
CDS             10281 10631
```

You may now perform new database searches on all these genes within Artemis. Remember this EMBL file and many like it were submitted to the database several years ago and in general the annotation is not updated to take account of new data.

You may notice a region shown below for which there is no predicted gene. This type of non-coding region in enterics (this is *Salmonella*) is very unusual. See next page.



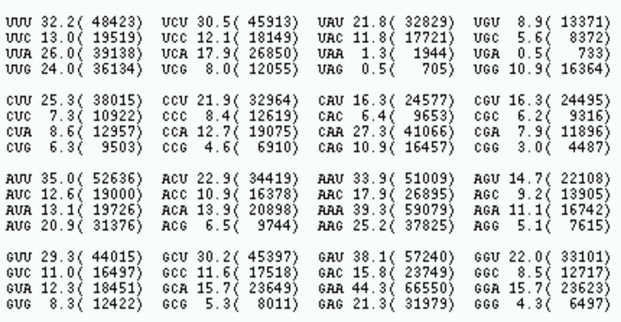
```
One selected base on forward strand: 12528
Entry:  AF060869.embl

 D  L  G  N  S  T  H  R  R  A  R  V  V  R  C  R  F  L  F  N  *  D  R  G  R  +  P  L  N  M  V
  I  S  V  I  V  P  T  V  E  R  G  L  C  D  V  D  F  C  S  I  E  I  A  G  D  S  P  S  I  W
.  S  R  #  +  Y  P  P  +  S  A  G  C  A  M  S  I  S  V  Q  L  R  S  R  A  I  A  P  Q  Y  G
GATCTCGGTAATAGTACCCACCGTAGAGCGCGGGTTGTGCGATGTCGATTTCTGTTCAATTGAGATCGCGGGCGATAGCCCCTCAATATGGT
         |10       |20       |30       |40       |50       |60       |70       |80       |90
CTAGAGCCATTATCATGGGTGGCATCTCGCGCCCAACACGCTACAGCTAAAGACAAGTTAACTCTAGCGCCCGCTATCGGGGAGTTATACCA
 I  E  T  I  T  G  V  T  S  R  P  N  H  S  T  S  K  Q  E  I  S  I  A  P  S  L  G  E  I  H  D
  D  R  Y  Y  Y  G  G  Y  L  A  P  Q  A  I  D  I  E  T  *  N  L  D  R  A  I  A  G  *  Y  P
.  R  P  L  L  V  W  R  L  A  R  T  T  R  H  R  N  R  N  L  Q  S  R  P  R  Y  G  R  L  I  T

CDS                 1   312  c
CDS               583  1083
tRNA             1898  1970    putative
CDS              2124  3698    5 transmembrane helices; secretory
CDS              3775  4140
```

To determine whether this non-coding region is truly as published, load into Artemis the codon usage information for *Salmonella.* Click on the Graph menu, then select ‘add usage plots..’. A window will appear asking you to select a file. Select the file ‘S_typhi.cod. This file contains codon usage information taken from a public web site (see below). Please ask for an explanation of the graphs which appear in your Artemis window.

Codon usage table taken from:
http://www.kazusa.or.jp/codon/

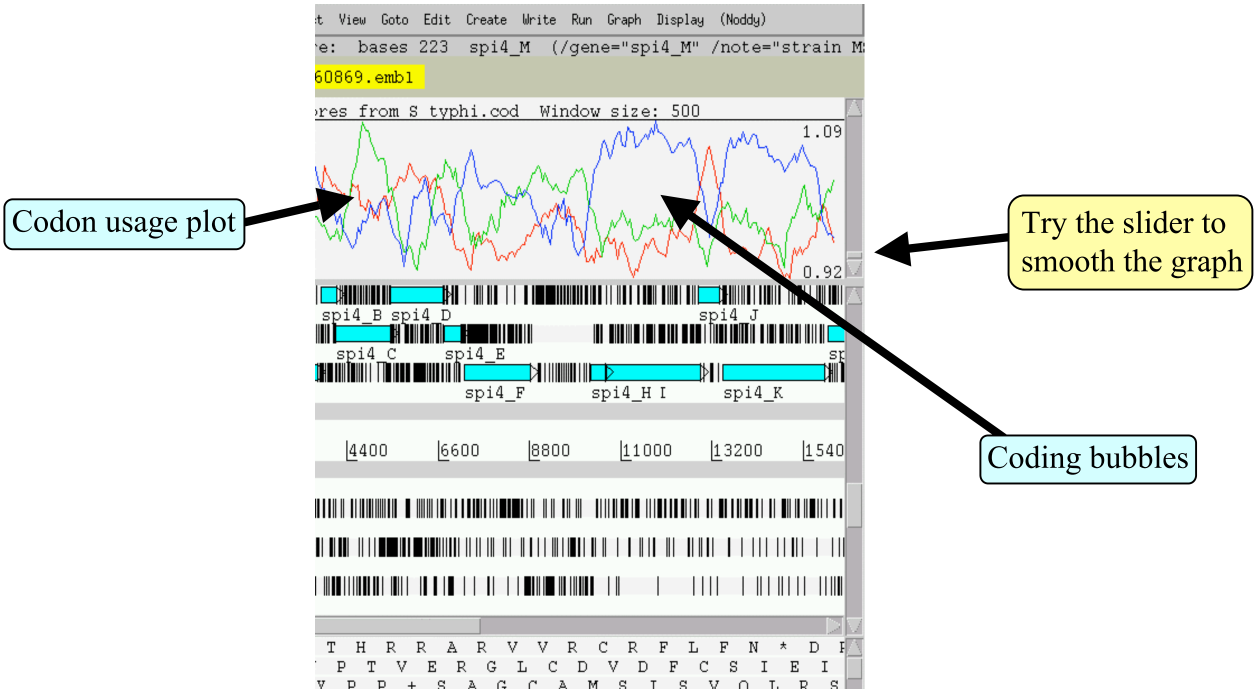


```
UUU 32.2( 48423)  UCU 30.5( 45913)  UAU 21.8( 32829)  UGU  8.9( 13371)
UUC 13.0( 19519)  UCC 12.1( 18149)  UAC 11.8( 17721)  UGC  5.6(  8372)
UUA 26.0( 39138)  UCA 17.9( 26850)  UAA  1.3(  1944)  UGA  0.5(   733)
UUG 24.0( 36134)  UCG  8.0( 12055)  UAG  0.5(   705)  UGG 10.9( 16364)

CUU 25.3( 38015)  CCU 21.9( 32964)  CAU 16.3( 24577)  CGU 16.3( 24495)
CUC  7.3( 10922)  CCC  8.4( 12619)  CAC  6.4(  9653)  CGC  6.2(  9316)
CUA  8.6( 12957)  CCA 12.7( 19075)  CAA 27.3( 41066)  CGA  7.9( 11896)
CUG  6.3(  9503)  CCG  4.6(  6910)  CAG 10.9( 16457)  CGG  3.0(  4487)

AUU 35.0( 52636)  ACU 22.9( 34419)  AAU 33.9( 51009)  AGU 14.7( 22108)
AUC 12.6( 19000)  ACC 10.9( 16378)  AAC 17.9( 26895)  AGC  9.2( 13905)
AUA 13.1( 19726)  ACA 13.9( 20898)  AAA 39.3( 59079)  AGA 11.1( 16742)
AUG 20.9( 31376)  ACG  6.5(  9744)  AAG 25.2( 37825)  AGG  5.1(  7615)

GUU 29.3( 44015)  GCU 30.2( 45397)  GAU 38.1( 57240)  GGU 22.0( 33101)
GUC 11.0( 16497)  GCC 11.6( 17518)  GAC 15.8( 23749)  GGC  8.5( 12717)
GUA 12.3( 18451)  GCA 15.7( 23649)  GAA 44.3( 66550)  GGA 15.7( 23623)
GUG  8.3( 12422)  GCG  5.3(  8011)  GAG 21.3( 31979)  GGG  4.3(  6497)
```


t View Goto Edit Create Write Run Graph Display (Noddy)
60869.embl
ores from S typhi.cod Window size: 500
1.09
0.92
spi4_B spi4_D
spi4_J
spi4_C
spi4_E
sp
spi4_F
spi4_H I
spi4_K
4400
6600
8800
11000
13200
1540
T H R R A R V V R C R F L F N * D F
P T V E R G L C D V D F C S I E I
V P P + S A G C A M S I S V Q L R S

e: bases 223 spi4_M (/gene="spi4_M" /note="strain M

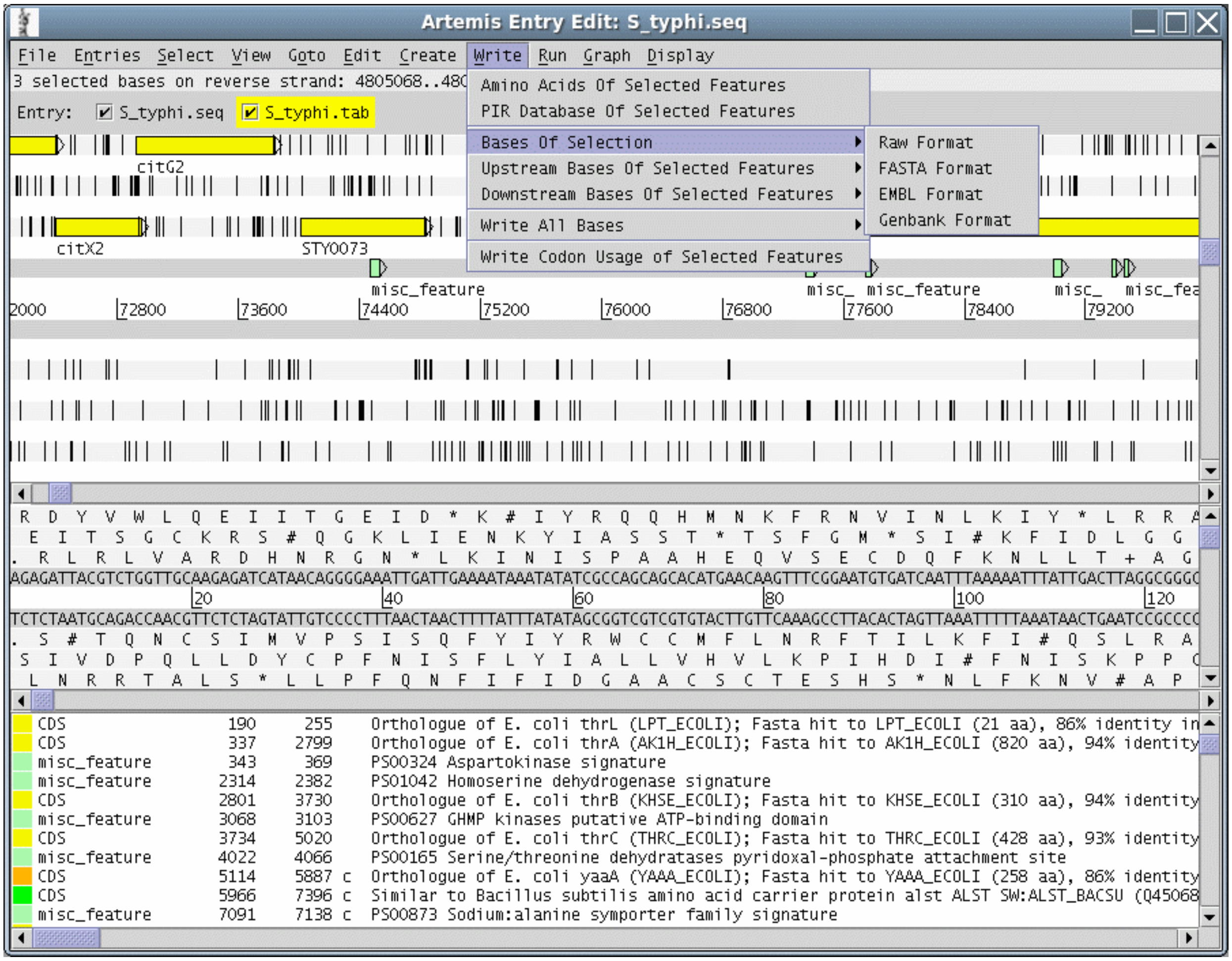
Codon usage plot

Try the slider to smooth the graph

Coding bubbles

When you first load the codon usage table into Artemis the graphs calculated for both upper and lower strands will be displayed (not shown). To add/remove one of these to/ from view click on the Graph menu and check the box alongside the option ‘codon usage from S_typhi.cod’ (the reverse plot is also represented in this list).

Based on the codon usage table Artemis calculates for each triplet in succession a score based on how well it matches the commonly used codons for that organism. The three lines shown above represent the scores for each reading frame. If the codons for a particular frame match those of the calculated codon usage table a high score is given. Practically speaking this manifests itself as a ‘coding bubble’ where a gap opens up in the plot indicating that this region is likely to be coding (see above). The plot suggests that this empty region actually encodes a product.


Artemis Entry Edit: S_typhi.seq
File Entries Select View Goto Edit Create Write Run Graph Display
Entry: ☑ S_typhi.seq ☑ S_typhi.tab
Amino Acids Of Selected Features
PIR Database Of Selected Features
Bases Of Selection
Upstream Bases Of Selected Features
Downstream Bases Of Selected Features
Write All Bases
Write Codon Usage of Selected Features
Raw Format
FASTA Format
EMBL Format
Genbank Format
citG2
citX2
STY0073
misc_feature
misc_
misc_feature
misc_
misc_fea
2000
72800
73600
74400
75200
76000
76800
77600
78400
79200

3 selected bases on reverse strand: 4805068..480

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

```
🟨 CDS              190    255    Orthologue of E. coli thrL (LPT_ECOLI); Fasta hit to LPT_ECOLI (21 aa), 86% identity in
🟨 CDS              337   2799    Orthologue of E. coli thrA (AK1H_ECOLI); Fasta hit to AK1H_ECOLI (820 aa), 94% identity
🟩 misc_feature     343    369    PS00324 Aspartokinase signature
🟩 misc_feature    2314   2382    PS01042 Homoserine dehydrogenase signature
🟨 CDS             2801   3730    Orthologue of E. coli thrB (KHSE_ECOLI); Fasta hit to KHSE_ECOLI (310 aa), 94% identity
🟩 misc_feature    3068   3103    PS00627 GHMP kinases putative ATP-binding domain
🟨 CDS             3734   5020    Orthologue of E. coli thrC (THRC_ECOLI); Fasta hit to THRC_ECOLI (428 aa), 93% identity
🟩 misc_feature    4022   4066    PS00165 Serine/threonine dehydratases pyridoxal-phosphate attachment site
🟧 CDS             5114   5887 c  Orthologue of E. coli yaaA (YAAA_ECOLI); Fasta hit to YAAA_ECOLI (258 aa), 86% identity
🟩 CDS             5966   7396 c  Similar to Bacillus subtilis amino acid carrier protein alst ALST SW:ALST_BACSU (Q45068
🟩 misc_feature    7091   7138 c  PS00873 Sodium:alanine symporter family signature
```

It is worth pointing out that you can export data from artemis in a number of different formats. Try this on the region that you annotated. Simply select the genes you made, you may then give them names by going to EDIT and AUTOMATICALLY CREATE GENE NAMES. Then go to the write menu and you can create a file of the protein sequences, nucleotide sequences, or the upstream and downstream sequences. Try this and look at the files you create.

Comparative Genomics

Introduction

The Artemis Comparison Tool (ACT), also written by Kim Rutherford, was designed to extract the additional information that can only be gained by comparing the growing number of genomes from closely related organisms.
ACT is based on Artemis, and so you will already be familiar with many of its core functions. ACT, is essentially composed of three layers or windows. The top and bottom layers are mini Artemis windows (with their inherited functionality), showing the linear representations of the genomes with their associated features. The middle window shows red blocks, which span this middle layer and link conserved regions within the two genomes, above and below. Consequently, if you were comparing two identical genome sequences you would see a solid red block extending over the length of the two sequences in this middle layer. If insertions were present in either of the genomes, they would show up as breaks between the solid red conserved regions. Data used to draw these red blocks and link conserved regions is generated by running pairwise BlastN or tBlastX comparisons of the genomes (details of how this is done are outlined in Appendix II and will be covered in detail in Module 6).

Aims

The aim of this Module is for you to become familiar with the basic functioning of ACT by using a series of worked examples. Some of these examples will touch on exercises that were used in previous Modules, this is intentional. Hopefully, as well as introducing you to the basics of ACT this Module will also show you how ACT can be used for not only looking at genome evolution but also to backup, or question, gene models and so on.

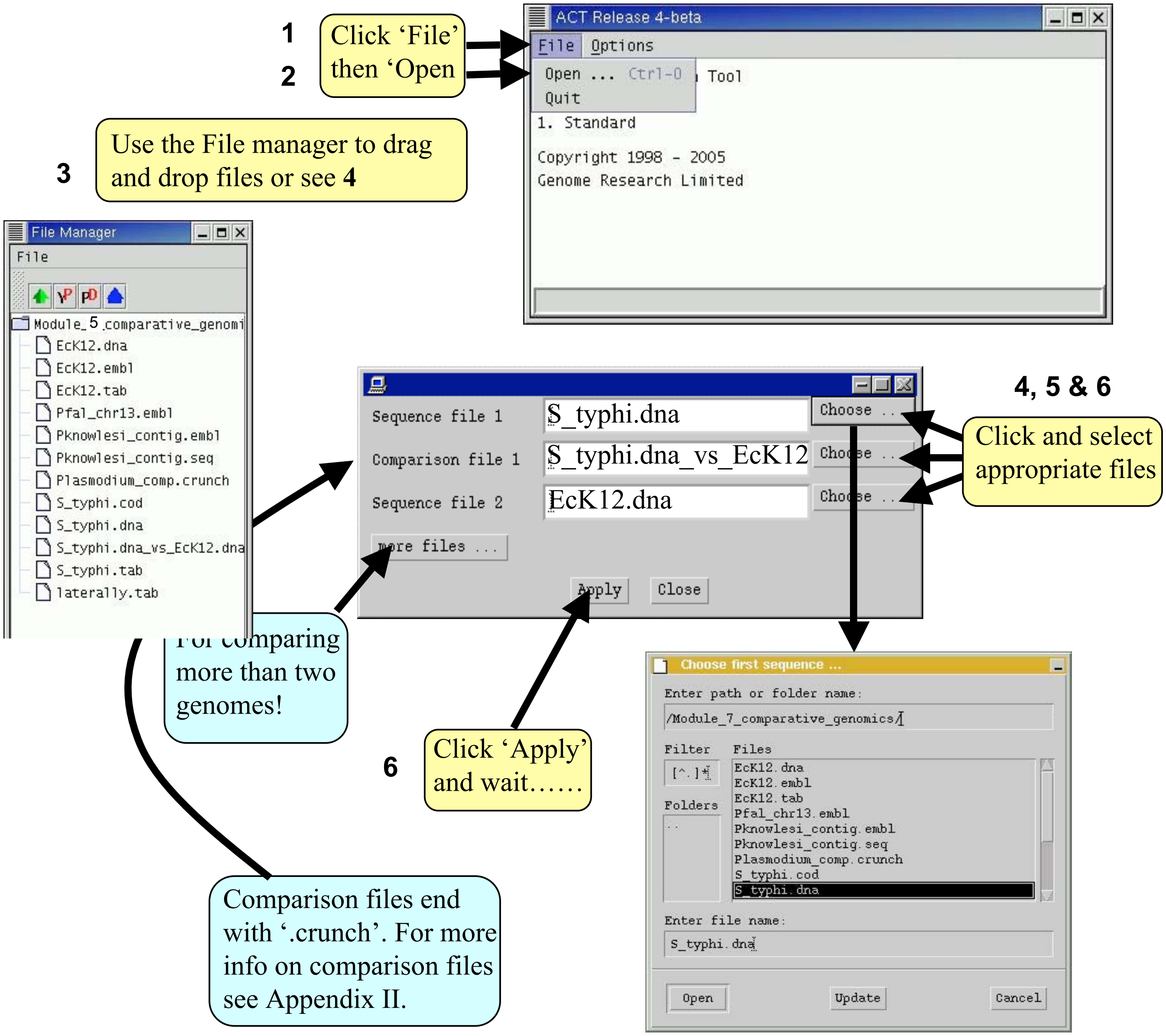
1. Starting up the ACT software

Make sure you’re in the **ACT/S_TYPHI-ECOLI** directory.
Then type

```
act & [return]
```

A small start up window will appear.
Now let’s load up a *S. typhi* versus *Escherichia coli* comparison.

The files you will need for this exercise are: *S_typhi.dna*
S_typhi.dna_vs_EcK12.dna.crunch
EcK12.dna



1 Click ‘File’
2 then ‘Open

3 Use the File manager to drag and drop files or see **4**

Sequence file 1: S_typhi.dna
Comparison file 1: S_typhi.dna_vs_EcK12
Sequence file 2: EcK12.dna

4, 5 & 6 Click and select appropriate files

For comparing more than two genomes!

Choose first sequence ...

Enter path or folder name: /Module_7_comparative_genomics/

Filter: [^.]*

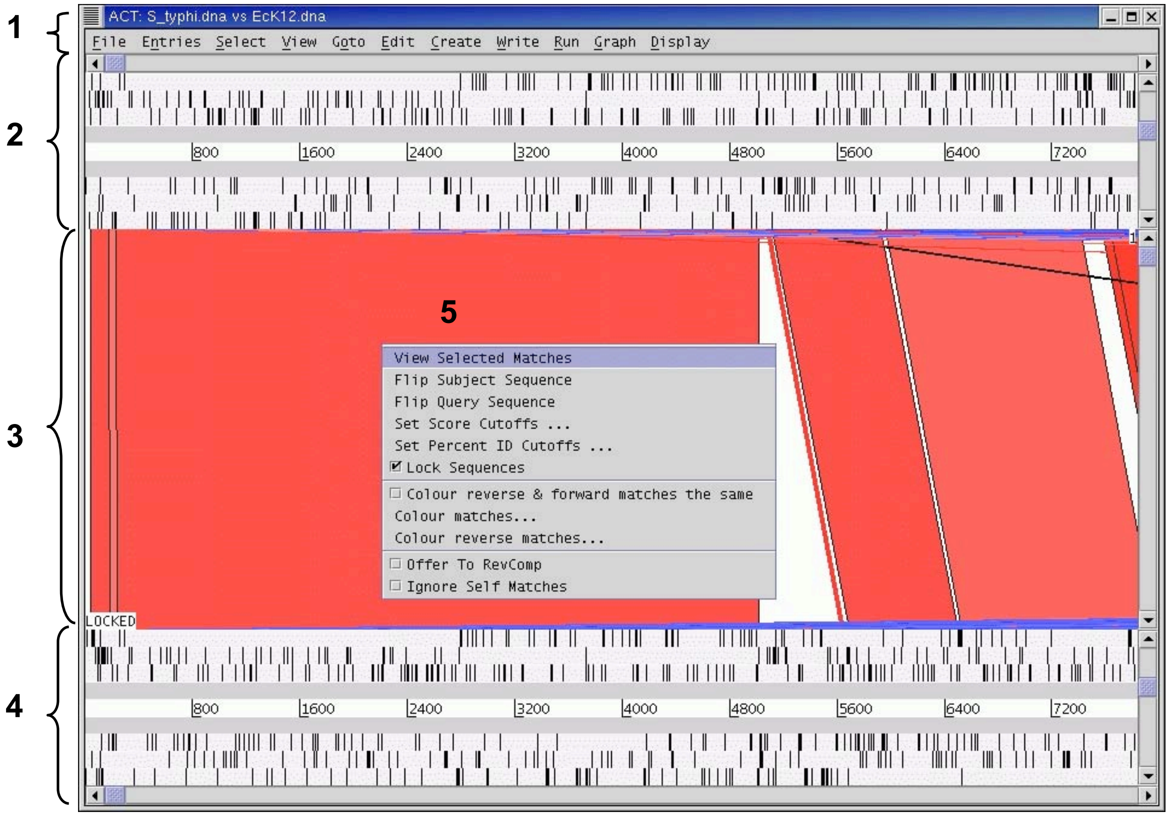
Enter file name: S_typhi.dna

6 Click ‘Apply’ and wait……

Comparison files end with ‘.crunch’. For more info on comparison files see Appendix II.

2. The basics of ACT

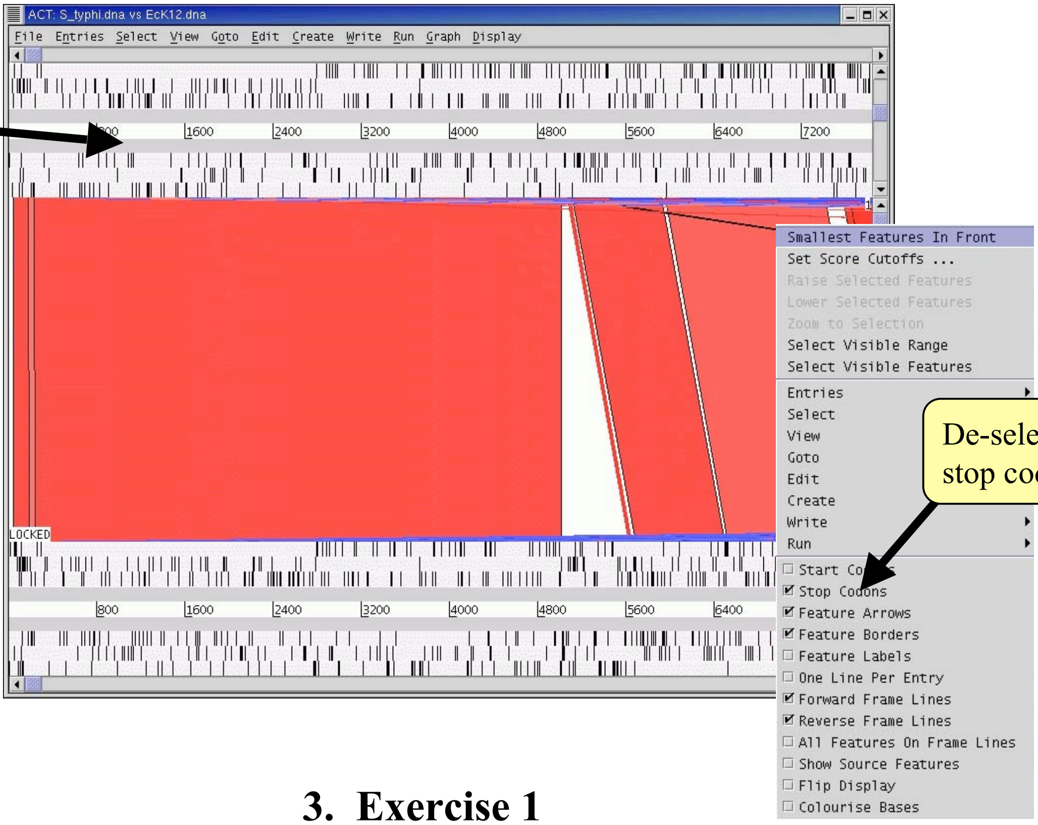
You should now have a window like this so let's see what's there.



1. Drop-down menus. These are mostly the same as in Artemis. The major difference you'll find is that after clicking on a menu header you will then need to select a DNA sequence before going to the full drop-down menu.
2. This is the Sequence view panel for 'Sequence file 1' (Subject Sequence) you selected earlier. It's a slightly compressed version of the Artemis main view panel. The panel retains the sliders for scrolling along the genome and for zooming in and out.
3. The Comparison View. This panel displays the regions of similarity between two sequences. Red blocks link similar regions of DNA with the intensity of red colour directly proportional to the level of similarity. Double clicking on a red block will centralise it. Blue blocks link regions that are inverted with respect to each other.
4. Artemis-style Sequence View panel for 'Sequence file 2' (Query Sequence).
5. Right button click in the Comparison View panel brings up this important ACT-specific menu which we will use later.

1

Right button click here



☐ Start Codons
☑ Stop Codons
☑ Feature Arrows
☑ Feature Borders
☐ Feature Labels
☐ One Line Per Entry
☑ Forward Frame Lines
☑ Reverse Frame Lines
☐ All Features On Frame Lines
☐ Show Source Features
☐ Flip Display
☐ Colourise Bases

2

De-select stop codons

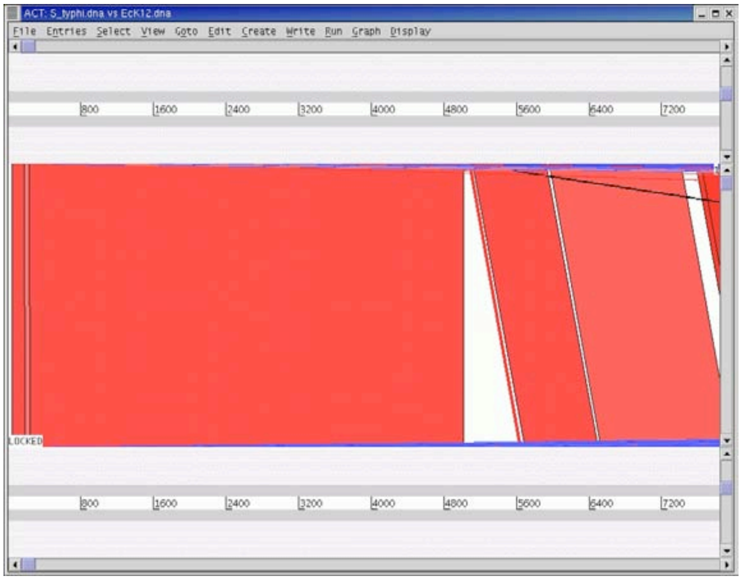
3. Exercise 1

Introduction & Aims

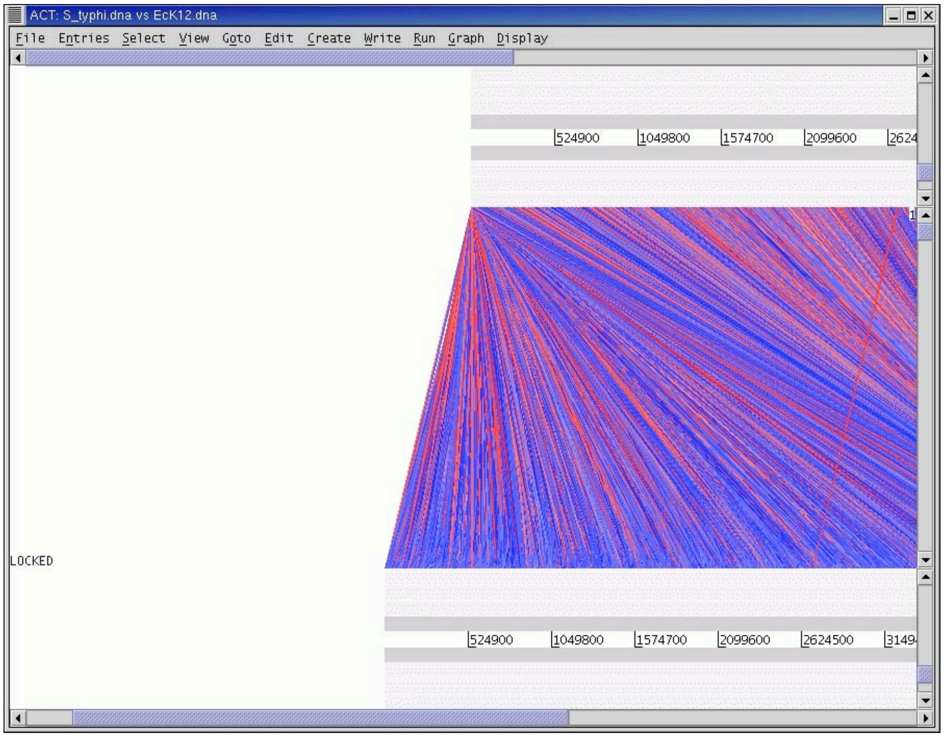
In this first exercise we are going to explore the basic features of ACT. Using the ACT session you have just opened we firstly are going to zoom outwards until we can see the entire *S. typhi* genome compared against the entire *E. coli* K12 genome. As for the Artemis exercises we should turn off the stop codons to clear the view and speed up the process of zooming out.

The only difference between ACT and Artemis when applying changes to the sequence views is that in ACT you must click the right mouse button over the specific sequence that you wish to change, as shown above.

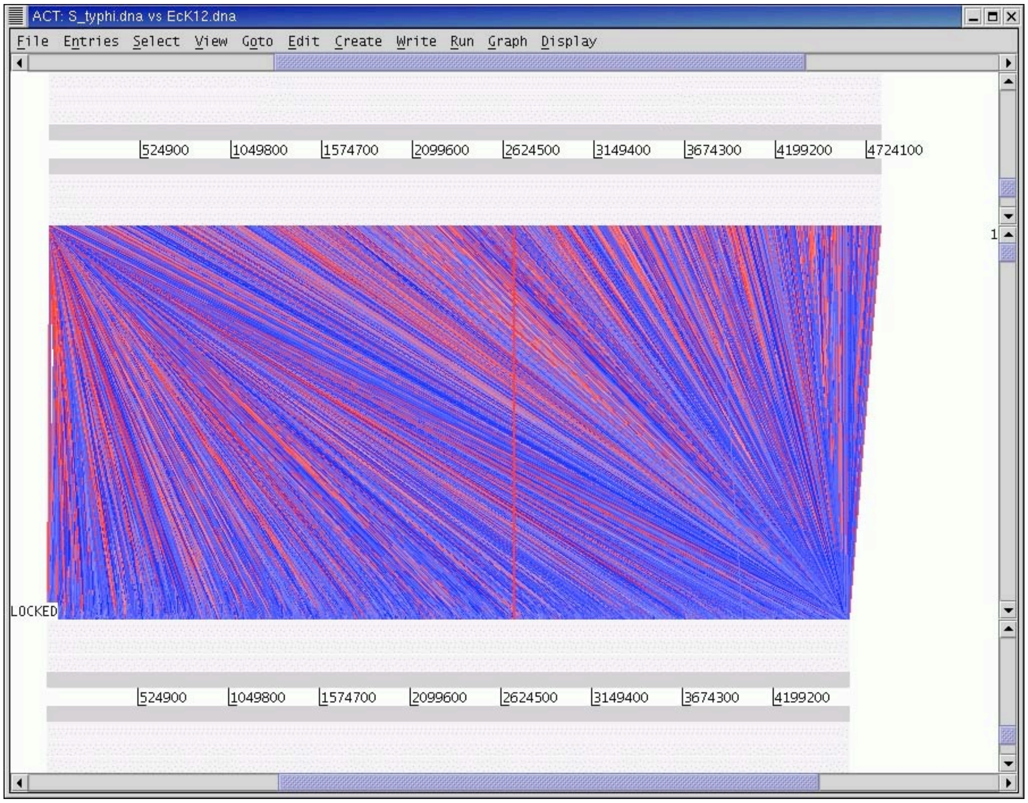
Now turn the stop codons off in the other sequence too. Your ACT window should look something like the one below:


ACT: S_typhi.dna vs EcK12.dna
File Entries Select View Goto Edit Create Write Run Graph Display
800
1600
2400
3200
4000
4800
5600
6400
7200
LOCKED

Use the vertical sliders to zoom out. Drag or click the slider downwards from one of the genomes. The other genome will stay in synch.

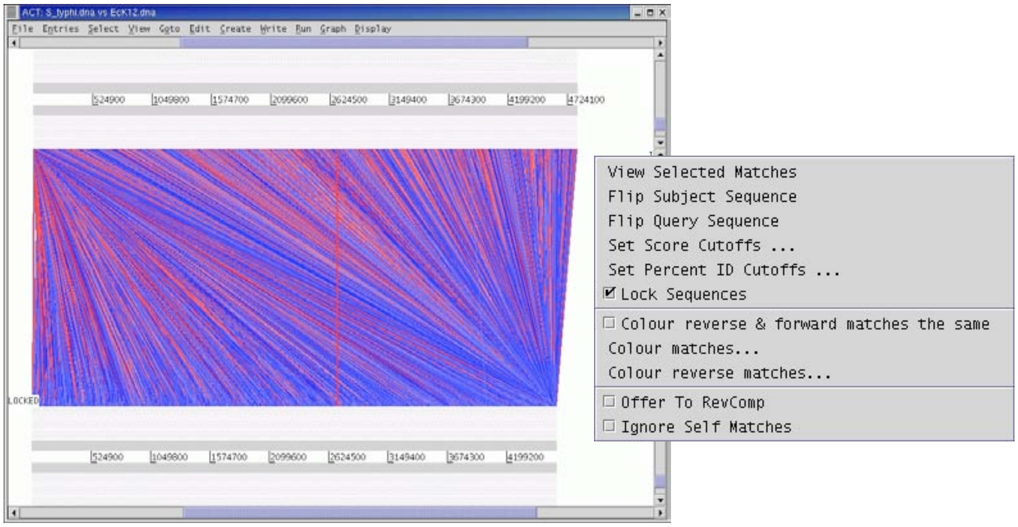

ACT: S_typhi.dna vs EcK12.dna
File Entries Select View Goto Edit Create Write Run Graph Display
524900
1049800
1574700
2099600
2624
LOCKED
524900
1049800
1574700
2099600
2624500
3149

Once zoomed out your ACT window should look similar to the one shown above. If the genomes in view fall out of view to the right of the screen, use the horizontal sliders to scroll the image and bring the whole sequence into view, as shown below. You may have to play around with the level of zoom to get the whole genomes shown in the same screen as shown below.



Notice that when you scroll along with either slide both genomes move together. This is because they are 'locked' together. Right click over the middle comparison view panel. A small menu will appear, select Unlock sequences and then scroll one of the horizontal sliders. Notice that 'LOCKED' has disappeared from the comparison view panel and the genomes will now move independently

LOCKED



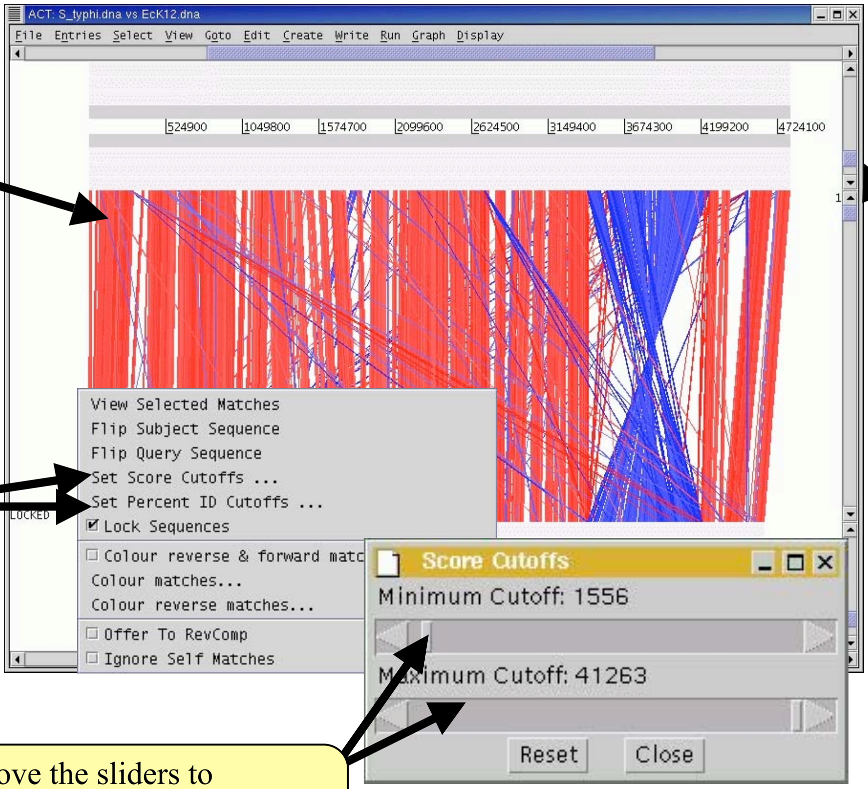
You can optimise your image by either removing 'low scoring' (or percentage ID) hits from view, as shown below **1-3** or by using the slider on the the comparison view panel (**4**). The slider allows you to filter the regions of similarity based on the length of sequence over which the similarity occurs, sometimes described as the "footprint".

1

Right button click in the Comparison View panel

2

Select either Set Score Cutoffs or Set Percent ID Cutoffs



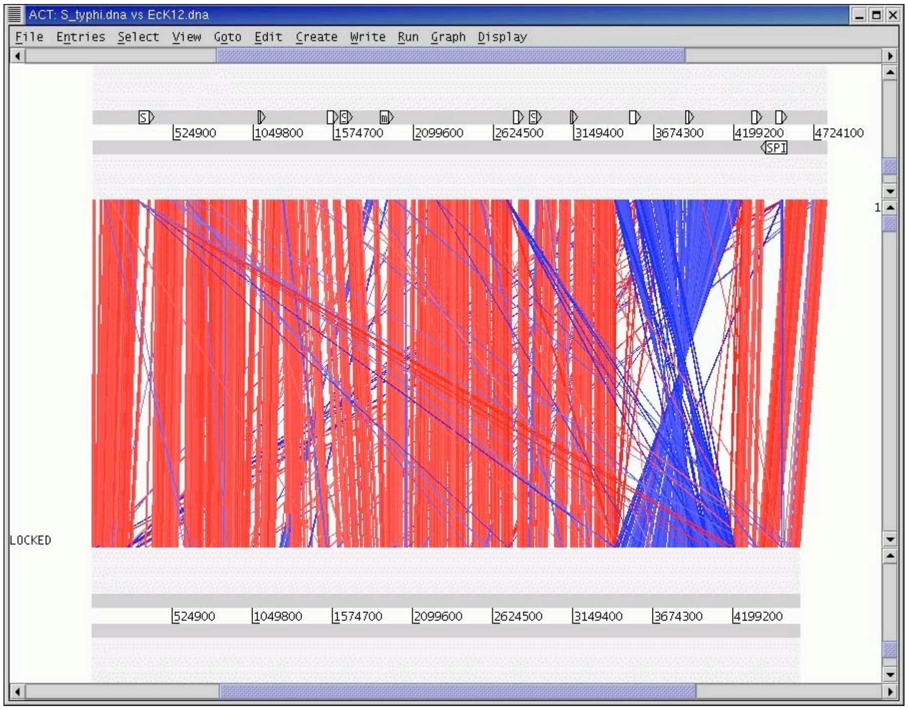
Minimum Cutoff: 1556

Maximum Cutoff: 41263

4

3

Move the sliders to manipulate the comparison view image



4. Things to try out in ACT

Load into the top sequence (*S.typhi*) a '.tab' file called 'laterally.tab'. You will need to use the 'File' menu and select the correct genome sequence ('*S.typhi*.dna') before you can read in an entry. If you are zoomed out and looking at the whole of both genomes you should see the above. The small white boxes are the regions of atypical DNA covering regions that we looked at in the first Artemis exercise. It is apparent that there is a backbone sequence shared with *E. coli* K12. Into this various chunks of DNA, specific the *S. typhi* (with respect to *E. coli* K12) have been inserted.

5. More things to try out in ACT

1. Double click red boxes to centralise them.
2. Zoom right in to view the base pairs and amino acids of each sequence.
3. Load annotation files into the sequence view panels.
4. You could load in the appropriate '.tab' files for each genome (S_typhi.tab and EcK12.tab) and view the annotation of a particular region. Also try using some of the other Artemis features eg. graphs etc.
5. Find an inversion in one genome relative to the other then flip one of the sequences.

Once you have finished this exercise remember to close this ACT session down completely before starting the next exercise

OTHER EXERSISES IN ACT:

There are three other directories containing comparisons between other species that you can look at

A_FUM: a comparison between three fungi. *Aspergillus fumigatus, Aspergilllus nidulans* and *Podospora anserina*. This is a comparison of a rearrangement in a gene cluster.

Lmajor_Tbrucei: a comparison of a strand switch region between in T. brucei and L. major.

PFAL:PKNOW: a comparison between Plasmodium knowlesi and Plasmodium falciparum.

MAKING A COMPARISON FILE.

Using the program blastall you can make a comparison file. Try this on one of the smaller sequences.

1. First you must generate a database from one of the sequences.

```
> formatdb –i myseq1 –p F
```

2. For a BLASTN comparison type:

```
>blastall –i myseq2 –d myseq1 –o blast.out –p blastn –m8
```

(The –m8 flag generates the format that ACT can read.)

3. For a TBLASTX comparison type

```
>blastall –i myseq2 –d myseq1 –o blast.out –p tblastx –m8
```

Look at these comparisons in ACT and compare the output using blastn and tblastx

Appendix I: List of degenerate nucleotide value/IUB Base Codes.

R = A or G

S = G or C

B = C, G or T

Y = C or T

W = A or T

D = A, G or T

K = G or T

N = A, C, G or T

H = A, C or T

M = A or C

V = A, C or G

Appendix II: Keys and Qualifiers – a brief explanation of what they are and a sample of the one's we use.

1 – Keys: They describe features with DNA coordinates and once marked, they all appear in the Artemis main window. The ones we use are:

➔ CDS: Marks the extent of the gene.
➔ RBS: Ribosomal binding site
➔ misc_feature: Miscellaneous feature in the DNA
➔ rRNA: Ribosomal RNA
➔ repeat_region
➔ repeat_unit
➔ stem_loop
➔ tRNA: Transfer RNA

2 – Qualifiers: They describe features with protein coordinates. Once marked they appear in the lower part of the Artemis window. They describe the gene whose coordinates appear in the 'location' part of the editing window. The ones we commonly use for annotation at the Sanger Institute are:

➔ Class: Classification scheme we use "in-house" developed from Monica Riley's MultiFun assignments (see Appendix V).
➔ Colour: Also used in-house in order to differentiate between different types of genes and other features.
➔ Gene: This qualifier either gives the gene a name or a systematic gene number.
➔ Label: Allows you to label a gene/feature in the main view panel.
➔ Note: This qualifier allows for the inclusion of free text. This could be a description of the evidence supporting the functional prediction or other notable features/information which cannot be described using other qualifiers.
➔Partial: When a region in the DNA hits a protein in the database but lacks start and/or stop codons and the match does not include the whole length of the protein, it can be considered as a partial gene.
➔ Product: The assigned possible function for the protein goes here.
➔ Pseudo: Matches in different frames to consecutive segments of the same protein in the databases can be linked or joined as one and edited in one window. They are marked as pseudogenes. They are normally not functional and are considered gene remnants.

Appendix III:Useful Web addresses

Major Public Sequence Repositories

DNA Data Bank of Japan (DDBJ)	http://www.ddbj.nig.ac.jp
EMBL Nucleotide Sequence Database	http://www.ebi.ac.uk/embl.html
Genomes at the EBI	http://www.ebi.ac.uk/genomes/
GenBank	http://www.ncbi.nlm.nih.gov/

Microbial Genome Databases Resources

Sanger Microbial Genomes	http://www.sanger.ac.uk/Projects/Microbes/
TIGR Microbial Database	http://www.tigr.org/tdb/mdb/mdbcomplete.html
Institute Pasteur GenoList databases *Including: SubtiList, Colbri, TubercuList, Leproma, PyloriGene, MypuList, ListiList, CandidaDB,*	http://genolist.pasteur.fr
Pseudomonas Genome Database	http://www.pseudomonas.com/
Clusters of Orthologous Groups of proteins (COGs)	http://www.ncbi.nlm.nih.gov/COG/
SCODBII (*S. coelicolor* database)	http://www.jiio16.jic.bbsrc.ac.uk/S.coelicolor

Protein Motif Databases

Prosite	http://www.expasy.ch/prosite/
Pfam	http://www.expasy.ch/prosite/
BLOCKS	http://blocks.fhcrc.org
InterPro	http://www.ebi.ac.uk/interpro/
PRINTS	http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/
SMART	http://smart.embl-heidelberg.de
InterPro	http://www.ebi.ac.uk/interpro/index.html

Protein feature prediction tools

TMHMM Prediction of transmembrane helices in proteins	http://www.cbs.dtu.dk/services/TMHMM-2.0/
SignalP Prediction Server	http://www.cbs.dtu.dk/services/SignalP/
PSORT protein prediction	http://psort.ims.u-tokyo.ac.jp/form.html

Metabolic Pathways and Cellular Regulation

EcoCyc	http://ecocyc.org/
ENZYME	http://www.expasy.ch/enzyme/
Kyoto Encyclopedia of Genes and Genomes (KEGG)	http://www.genome.ad.jp/kegg
MetaCyc	http://ecocyc.org/

Miscellaneous sites

NCBI BLAST website	http://www.ncbi.nlm.nih.gov/BLAST/
The tmRNA website	http://www.indiana.edu/~tmrna/
tRNAscan-SE Search Server	http://www.genetics.wustl.edu/eddy/tRNAscan-SE/
Codon usage database	http://www.kazusa.or.jp/codon/
RNAgenie RNA gene prediction	http://rnagene.lbl.gov/
GO Gene Ontology Consortium	http://www.geneontology.org/
Artemis homepage	http://www.sanger.ac.uk/Software/Artemis/
ACT homepage	http://www.sanger.ac.uk/Software/ACT/