

Module 1

Artemis

Introduction

Artemis is a DNA viewer and annotation tool, free to download and use, written by Kim Rutherford from the Sanger Institute (Rutherford *et al.*, 2000). The program allows the user to view a range of files, from simple sequence files (e.g. fasta format) to EMBL/Genbank entries, as well as the results of sequence analyses, in a highly interactive and intuitive graphical format. Artemis is routinely used by the Pathogen Genomics group for annotation and analysis of both prokaryotic and eukaryotic genomes, and can also be used to visualise mapped data from next generation sequencing. Several types/sets of information can be viewed simultaneously within different contexts. For example, Artemis gives you the two views of the same genome region, so you can zoom in to inspect detailed DNA sequence motifs, and also zoom out to view local gene architecture (e.g. operons), or even an entire chromosome or genome, all within one screen. It is also possible to perform analyses within Artemis and save the output for future reference.

Aims

The aim of this Module is for you to become familiar with the basic functions of Artemis using a series of worked examples. These examples are designed to take you through the most immediately useful functions. However, there will be time, and encouragement, for you to explore other menus; features of Artemis that are not described in the exercises in this manual, but which may be of particular interest to some users. Like all the Modules in this workshop, please remember:

IF YOU DON' T UNDERSTAND, PLEASE ASK!

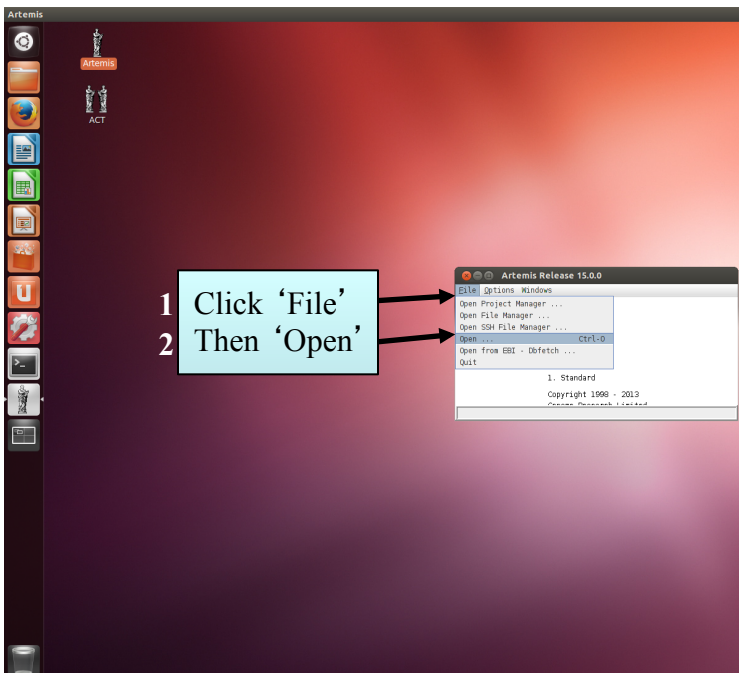
Artemis Exercise 1

1. Starting up the Artemis software

Double click the Artemis icon on the desktop.

A small start-up window will appear (see below). The directory **Module_1_Artemis** contains all files you will need for this module.

Now follow the sequence of numbers to load up the *Salmonella* Typhi chromosome sequence. Ask a demonstrator for help if you have any problems.

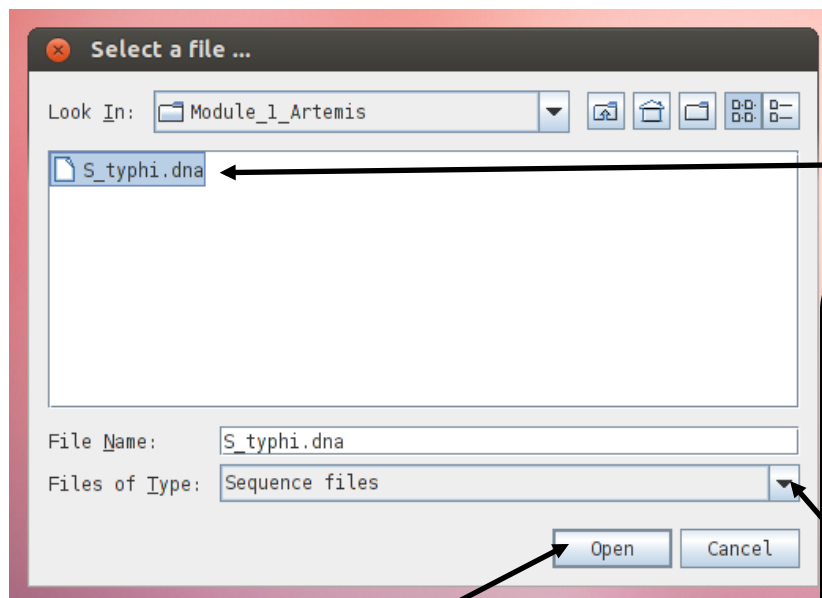


1 Click 'File'
2 Then 'Open'

In the 'Options' menu you can switch between prokaryotic and eukaryotic mode.

You can also start Artemis from the terminal window by typing 'art'.

For simplicity it is a good idea to open a new start up window for each Artemis session and close down any sessions once you have finished an exercise.



3 Single click to select file *S_typhi.dna*

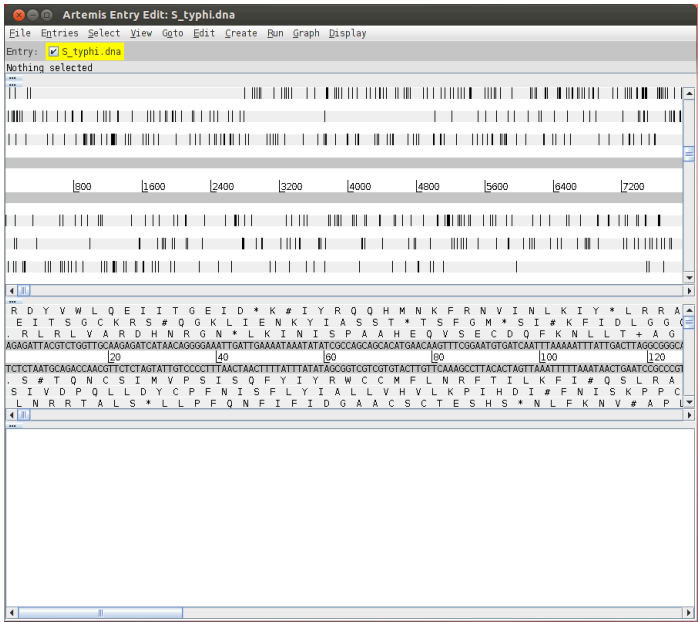
4 Single click to open file in Artemis then wait

Change to 'All Files' if you want to display all the files in the directory.

Use this feature to choose the type of file to be displayed in this panel. DNA sequence files will have the suffix '.dna'. Annotation files end with '.tab'. You can also open '.embl' files.

2. Loading an annotation file (entry) into Artemis

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



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

1

Click 'File' then 'Read an Entry'

Entry = file

What's an "Entry"? It's a file of DNA and/or features which can be overlaid onto the sequence information displayed in the main Artemis view panel.

2

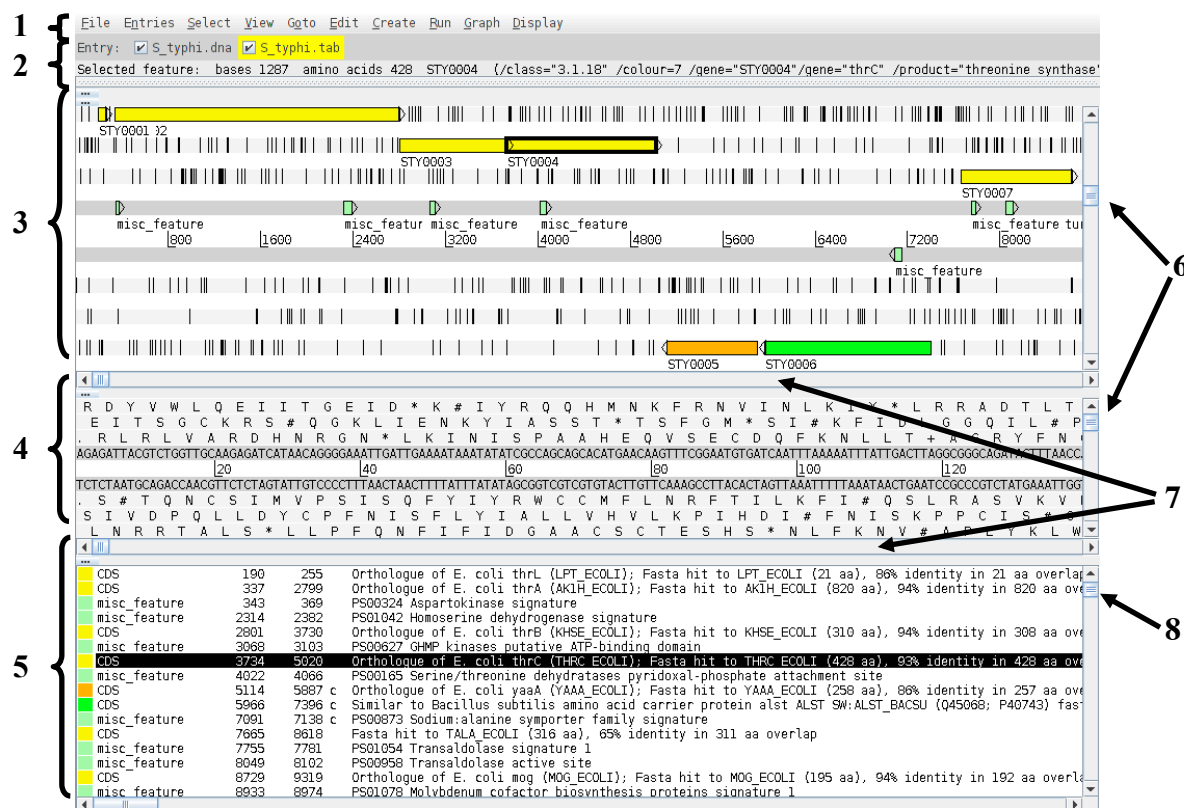
Single click to select file S_typhi.tab

3

Single click to open file in Artemis then wait (click 'no' if an error window pops up)

3. The basics of Artemis

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



1. **Drop-down menus:** There's lots in there so don't worry about all the details right now.
2. **Entry (top line):** shows which entries are currently loaded with the default entry highlighted in yellow (this is the entry into which newly created features are created). Selected feature: the details of a selected feature are shown here; in this case gene STY0004 (yellow box surrounded by thick black line).
3. This is the main **sequence view panel**. 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 on the reading frames as black vertical bars. Genes and other annotated features (eg. Pfam and Prosite matches) are displayed as coloured boxes. We often refer to predicted genes as coding sequences or CDSs.
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 CDS in the main view to see the zoomed view of the start of that CDS. Note that both this and the main panel can be scrolled left and right (7, below) zoomed in and out (6, below).
5. **Feature panel:** This panel contains details of the various features, listed in the order that they occur on the DNA. Any selected features are 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.

4. Getting around in Artemis

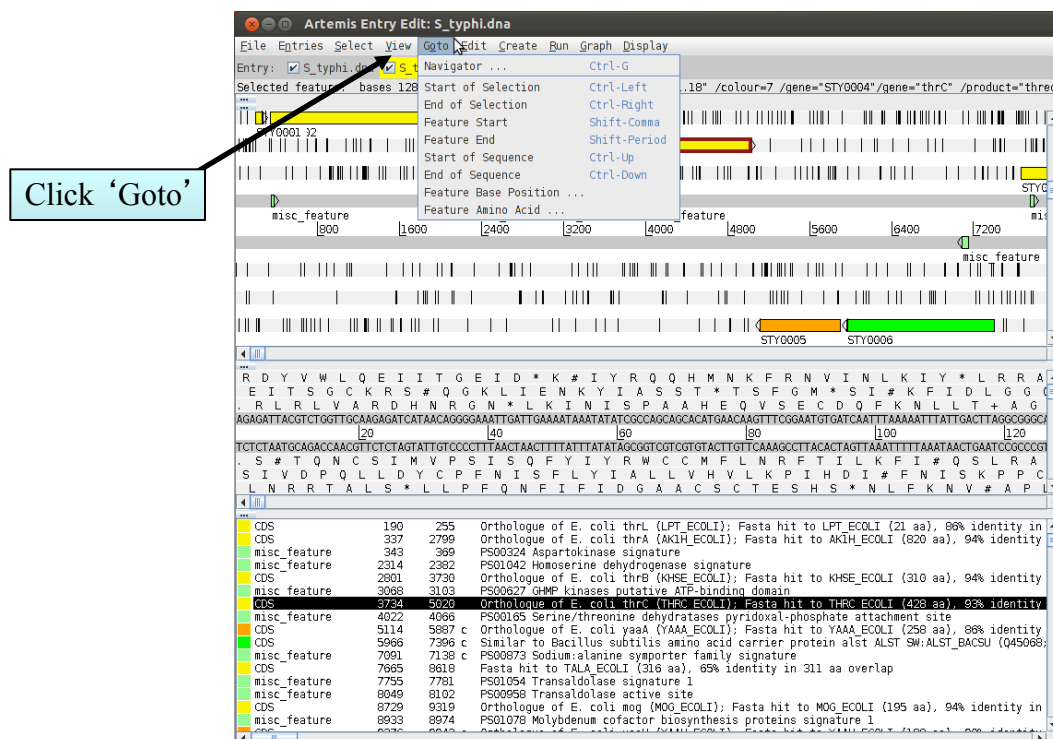
There are three main ways of getting to a particular DNA region in Artemis:

- the Goto drop-down menu
- the Navigator and
- the Feature Selector (which we will use in Part IV)

The best method depends on what you're trying to do. Knowing which one to use comes with practice.

4.1 The 'Goto' menu

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



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 selected by click-dragging in the main window.

So yes, give it a try!

Suggested tasks:

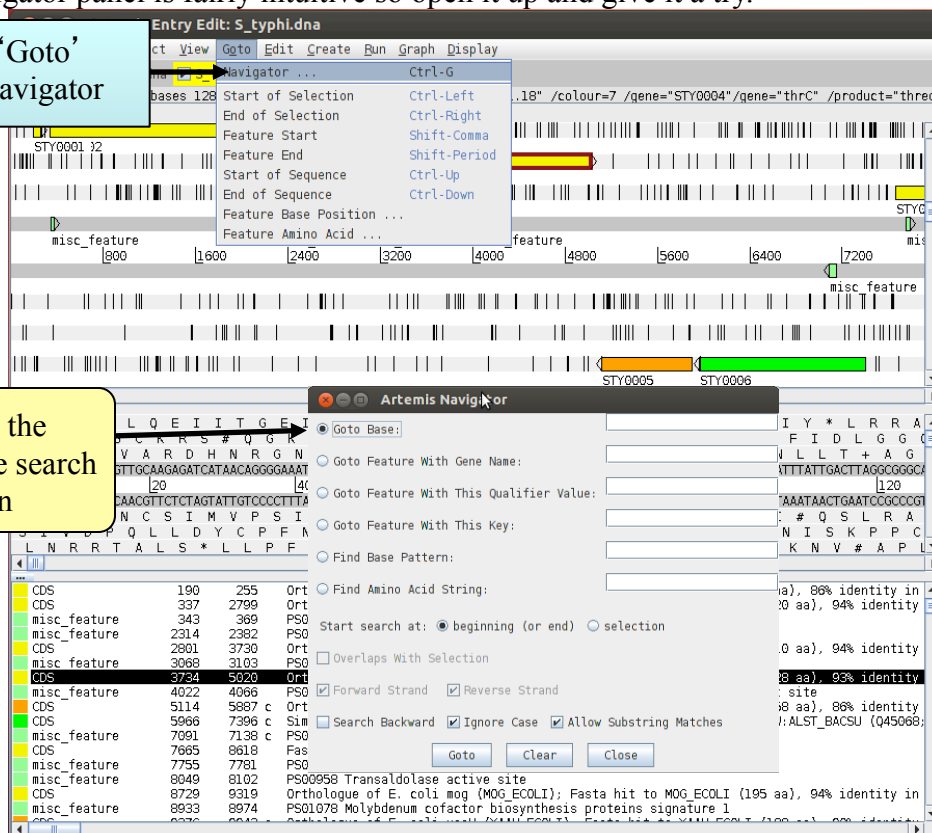
1. Zoom out, select / highlight a large region of sequence by clicking the left hand button and dragging the cursor then go to the start and end of this selected region.
2. Select a CDS then go to the start and end.
3. Go to the start and end of the genome sequence.
4. Select a CDS. Within it, go to a base (nucleotide) and/or amino acid of your choice.
5. Highlight a region then, from the right click menu, select 'Zoom to Selection'.

4.2 Navigator

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

Click 'Goto'
then Navigator

Check that the
appropriate search
button is on



Suggestions about 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 to the following pseudogene in the order that they occur on the chromosome.
4. Look at **Appendix VIII** which is a functional classification scheme used for the annotation of *S. Typhi*. Each CDS has a class qualifier best describing its function. Use the '**Goto Feature With This Qualifier value**' search to look for CDSs belonging to a class of interest by searching with the appropriate class values.
5. tRNA genes. Type 'tRNA' in the '**Goto Feature With This Key**'.
6. Regulator-binding DNA consensus sequence (real or made up!). Note that degenerate base values can be used (**Appendix Xa**).
7. 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 IV**

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 menus. Hopefully you will find most of them easy to understand.

Artemis Exercise 2

This part of the exercise uses the files and data you already have loaded into Artemis from Part I. By a method of your choice go to the region from bases 2188349 to 2199512 on the DNA sequence. This region is bordered by the *fbaB* gene which codes for fructose-bisphosphate aldolase. You can use the Navigator function discussed previously to get there. The region you arrive at should look similar to that shown below (maybe you have to use the zoom sliders).

The screenshot shows the Artemis genome browser interface. The top menu bar includes File, Entries, Select, View, Goto, Edit, Create, Run, Graph, and Display. Below the menu bar, the entry list shows S_typhi.dna and S_typhi.tab. The main display area shows a DNA sequence with various features. A yellow callout box labeled 'CDS features' points to a green bar, and another yellow callout box labeled 'Misc features' points to an orange bar. The bottom of the screen shows a list of features with their coordinates and descriptions.

Feature Type	Start	End	Description
CDS	190	255	Orthologue of E. coli thrL (LPT_ECOLI); Fasta hit to LPT_ECOLI (21 aa), 86% identity in 21 aa overlap
CDS	337	2799	Orthologue of E. coli thrA (AKIH_ECOLI); Fasta hit to AKIH_ECOLI (820 aa), 94% identity in 820 aa overlap
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 in 308 aa overlap
misc_feature	3068	3103	PS00627 GMP kinases putative ATP-binding domain
CDS	3734	5020	Orthologue of E. coli thrC (THRC_ECOLI); Fasta hit to THRC_ECOLI (428 aa), 93% identity in 428 aa overlap
misc_feature	4022	4066	PS00165 Serine/threonine dehydratases pyridoxal-phosphate attachment site
CDS	5114	5887	Orthologue of E. coli yaaA (YAAA_ECOLI); Fasta hit to YAAA_ECOLI (258 aa), 86% identity in 257 aa overlap
misc_feature	5966	7396	Similar to Bacillus subtilis amino acid carrier protein alaT ALST SW:ALST_BACSU (Q40506; P40743) fast
CDS	7091	7138	PS00873 Sodium:alanine symporter family signature
CDS	7665	8618	Fasta hit to TALA_ECOLI (316 aa), 65% identity in 311 aa overlap
misc_feature	7755	7781	PS01054 Transaldolase signature 1
misc_feature	8049	8102	PS00958 Transaldolase active site
CDS	8729	9319	Orthologue of E. coli mog (MOG_ECOLI); Fasta hit to MOG_ECOLI (195 aa), 94% identity in 192 aa overlap
misc_feature	8933	8974	PS01078 Molybdenum cofactor biosynthesis proteins signature 1

Once you have found this region have a look at some of the information available:

Information to view:

Annotation

If you click on a particular feature you can view the annotation associated with it: select a CDS feature (or any other feature) and click on the 'Edit' menu and select 'Selected Feature in Editor'. A window will appear containing all the annotation that is associated with that CDS. The format for this information is constrained by that which can be submitted to the EMBL database.

Viewing amino acid or protein sequence

Click on the 'View' menu and you will see various options for viewing the bases or amino acids of the feature you have selected, in two formats i.e. EMBL (view -> selection) or fasta (view -> bases or view -> amino acids). This can be very useful when using other programs that are not integrated into Artemis e.g. those available on the Web that require you to cut and paste sequence into them.

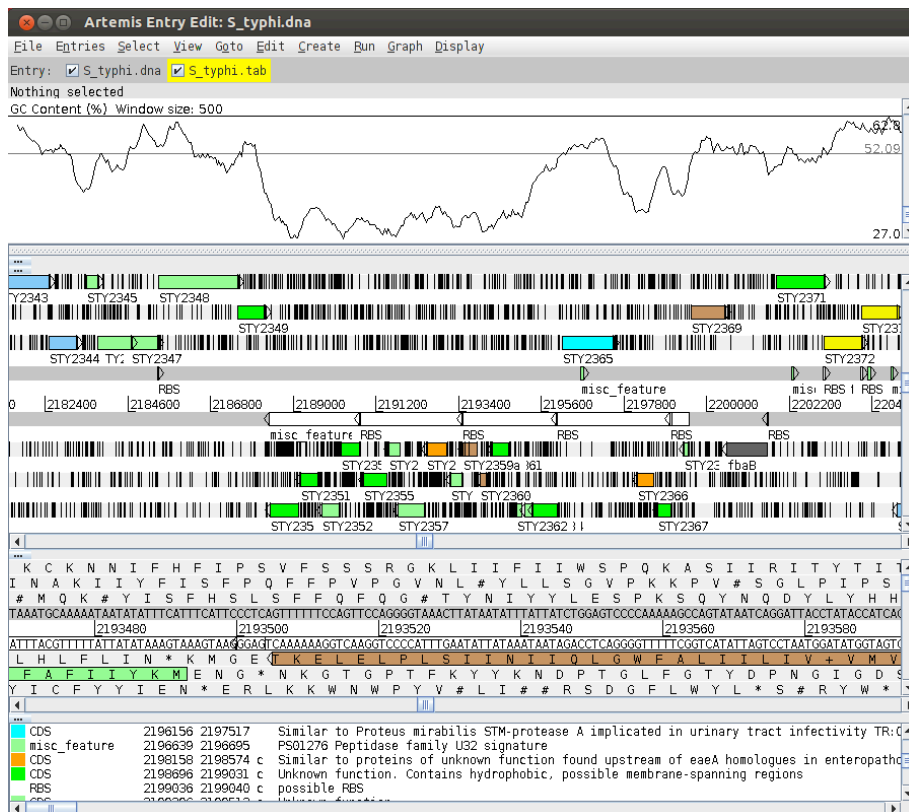
Plots/Graphs

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

In addition to looking at the fine detail of the annotated features it is also possible to look at the characteristics of the DNA covering the region displayed. This can be done by adding various plots to the display, showing different characteristics of the DNA. Some of the plots can be used to look at the protein coding potential of translation frames within the DNA, and others can be used to search for horizontally acquired DNA (such as GC frame plot).

To view the graphs:

Click on the 'Graph' menu to see all those available and then tick the box for 'GC Content (%)'. To adjust the smoothing of the graph you change the window size over which the points on the graph are calculated, using the slider shown below.



DNA plot

Slider for smoothing

Notice how the plot shows a marked deviation around the region you are currently looking at. To fully appreciate how anomalous this region is, move the genome view by scrolling to the left and right of this region. The apparent unusual nucleotide content of this region is indicative of laterally acquired DNA that has inserted into the genome.

As well as looking at the characteristics of small regions of the genome, it is possible to zoom out and look at the characteristics of the genome as a whole. To view the entire genome, you can use the sliders indicated above. However, be careful zooming out quickly with all the features being displayed, as this may temporarily lock up the computer. Read further to see how to zoom out.

1. To make this process faster and clearer, **switch off stop codons** by clicking with the right mouse button in the main view panel. A menu will appear with an option to de-select 'Stop Codons' (see below).
2. You will also need to temporarily **remove all of the annotated features** from the Artemis display window. In fact if you leave them on, which you can, they would be too small to see when you zoomed out to display the entire genome. To remove the annotation click on the S_typhi.tab entry button on the grey entry line of the Artemis window shown above.

2 To de-select the annotation click here.

Artemis Entry Edit: S_typhi.dna

File Entries Select View Goto Edit Create Run Graph Display

Entry: ☒ S_typhi.dna ☒ S_typhi.tab

Nothing selected

GC Content (%) Window size: 500

Smallest Features In Front
Set Score Cutoffs ...
Raise Selected Features
Lower Selected Features
Zoom to Selection
Select Visible Range
Select Visible Features
Frame Line Features ...

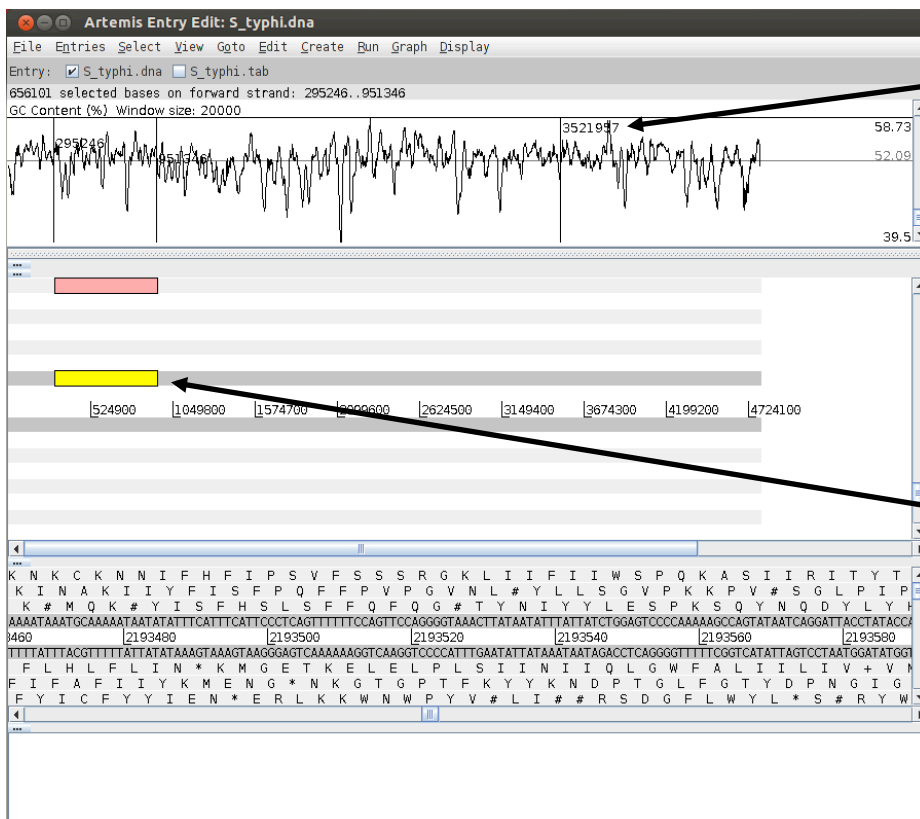
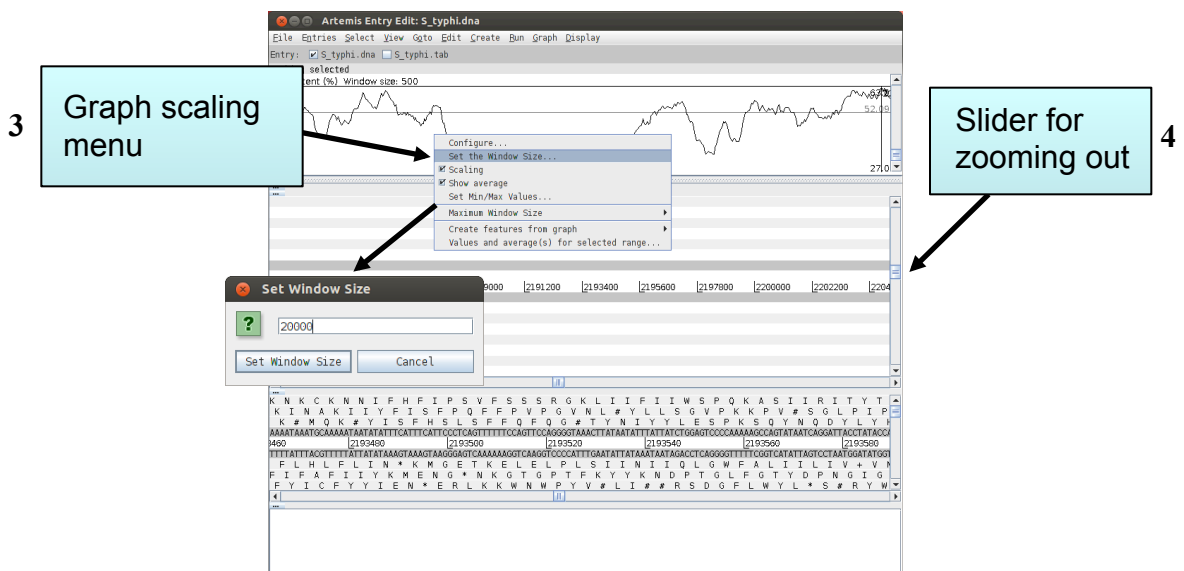
Entries
Select
View
Goto
Edit
Create
Write
Run

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

No stop codons shown on frame lines

Menu item for de-selecting stop codons

- One final tip is to **adjust the scaling** for each graph displayed before zooming out. This increases the maximum window size over which a single point for each plot is calculated. To adjust the scaling click with the right mouse button over a particular graph window. A menu will appear with an option "Set the Window size" (see above), set the window size to '20000'. You should do this for each graph displayed (if you get an error message press continue).
- You are now ready to zoom out by dragging or clicking the slider indicated below. Once you have zoomed out fully to see the entire genome you will need to adjust the smoothing of the graphs using the vertical graph sliders as before, to have a similar view to that shown below.



Click with the left mouse button in a graph window. A line and a number will appear. The number is the relative position within the genome (bps).

Click and drag to highlight a region on the main DNA line. Notice that the boundaries of this region are now marked in the graph windows that you previously clicked in.

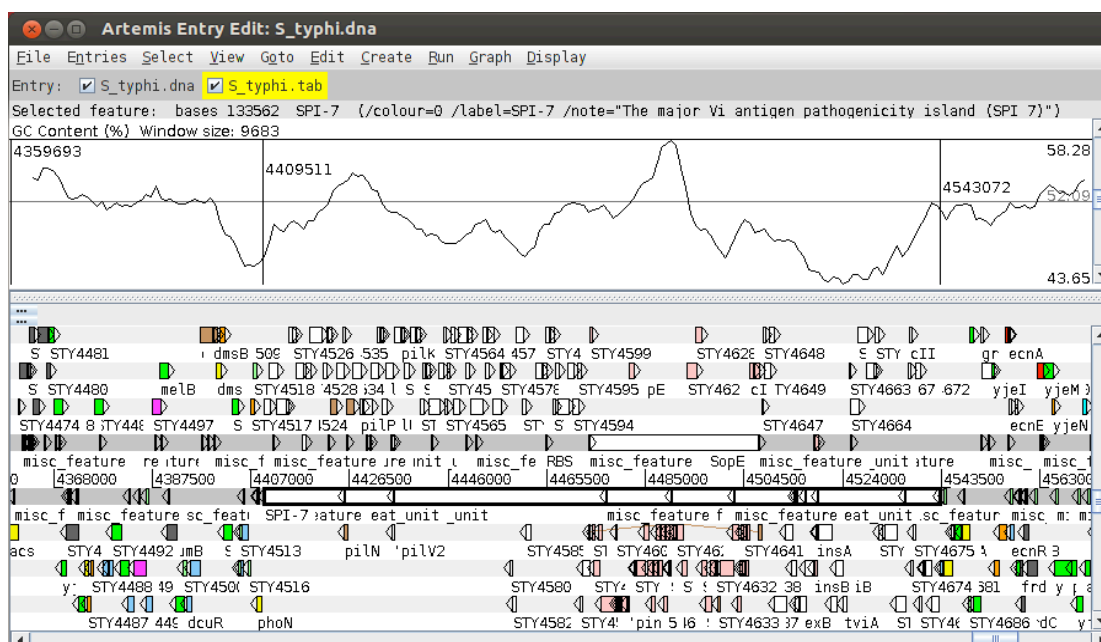
Artemis Exercise 3

Now go to position 4409511. The next region we are looking at is defined as a *Salmonella* pathogenicity island (SPI). SPI-7, or the major Vi pathogenicity island, is ~134 kb in length and contains ~30 kb of integrated bacteriophage.

The region you should be looking at is shown below and is a classical example of a *Salmonella* pathogenicity island (SPI). The definitions of what constitutes a pathogenicity island are quite diverse. However, below is a list of characteristics which are commonly seen within these regions, as described by Hacker *et al.*, 1997.

1. Often inserted alongside stable RNAs
2. Atypical G+C contents.
3. Carry virulence-related functions
4. Often carry genes encoding transposase or integrase-like proteins
5. Unstable and self-mobilisable
6. Of limited phylogenetic distribution

Have a look in and around this region and look for some of these features.



We are going to extract this region from the whole genome sequence and perform some more detailed analysis on it. We will aim to write and save new EMBL format files which will include just the annotations and DNA for this region.

Follow the numbers on the next page to complete the task.

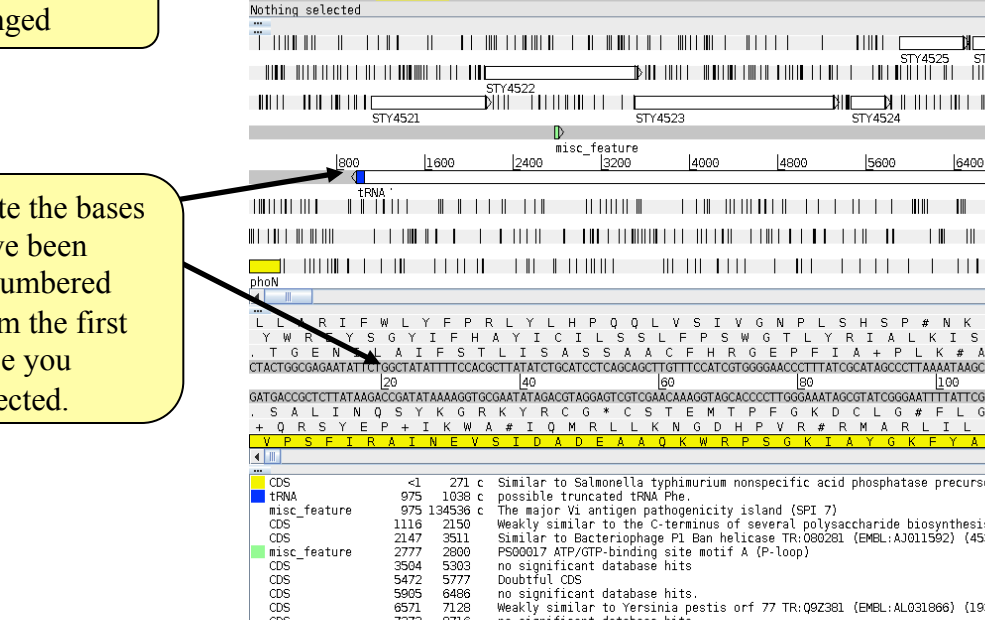
The screenshot shows the Artemis software interface for editing a DNA entry named 'S_typhi.dna'. The 'Edit' menu is open, displaying options such as 'Undo', 'Redo', 'Selected Features in Editor', 'Subsequence (and Features)', 'Find/Replace Qualifier Text ...', 'Qualifier of Selected Feature(s)', 'Selected Feature(s)', 'Move Selected Features To', 'Copy Selected Features To', 'Trim Selected Features', 'Extend Selected Features', 'Fix Stop Codons', 'Automatically Create Gene Names', 'Fix Gene Names', 'Bases', 'Contig Reordering', and 'Header Of Default Entry'. Annotations include:

- A blue box labeled 'Click 'Edit'' with an arrow pointing to the 'Edit' menu item.
- A blue box labeled 'Click 'Edit subsequence (and features)'' with an arrow pointing to the 'Subsequence (and Features)' menu option.
- A blue box labeled 'Select region by clicking with the left mouse button & dragging' with an arrow pointing to a selected region in the genomic track.

A new Artemis window will appear displaying only the region that you highlighted

Note the entry names have changed

Note the bases have been renumbered from the first base you selected.



Artemis Entry Edit

File Entries Select View Goto Edit Create Run Graph Display

Entry ☒ no name ☒ no name

Nothing selected

STY4521 STY4522 STY4523 STY4524 STY4525

misc_feature

tRNA

phoN

LLLRIRIFWLYFPRLLYHPQQLVLSIVGNPLSHSP#NKKPPYHSGK
YWRYSYSGYIFHAYICILSSSLFSPSWGTLRYIALKISLLIIIAES
TGTGENLAIFSTLISASSAACFHRGEPFIA+PLK#ASLS+RK
CTACTGGGAGGAATATTTCGCTATATTTCACAGCTTATATCTGCATCTCCAGACGCTGTTTCCATCTGTGGGAAACCTTTATCGCATAGCCCTTAAATAAGCCCTTCCTTATCATAGGGAAGAA
GATGACCGCTCTTATGAAGATAAAAGTGGGAATATAGAGCTAGGAGTCTGGACAAAGTAGCACCCTTGGGAATAGCTATCGGGAATTTTATTCGGAGAAATAGTATTCGCGCTTC
SAILNQSYKCGKCGCSTGMTTFGKDDLLSG*
+QRSYEP+IKWA*IQMRLCLKNGDHPVR#RMARLLTLRRIMASL
VPSFIRAINEVSIDADEAAQKWRPSPGKIAYYGFYAEKDYRF

CDS <1 271 c Similar to Salmonella typhimurium nonspecific acid phosphatase precursor phoN SW:PHON.SALT

tRNA 975 1038 c possible truncated tRNA Phe

misc_feature 975 134536 c The major Vi antigen pathogenicity island (SPI 7)

CDS 1116 2150 Weakly similar to the C-terminus of several polysaccharide biosynthesis proteins e.g. Str

CDS 2147 3511 Similar to Bacteriophage P1 Ban helicase TR:080281 (EMBL:AJ011592) (453 aa) fasta scores:

misc_feature 2777 2800 P500017 ATP/GTP-binding site motif A (P-loop)

CDS 3504 5303 no significant database hits

CDS 5472 5777 Doubtful CDS

CDS 5905 6496 no significant database hits

CDS 6571 7128 Weakly similar to Yersinia pestis orf 77 TR:Q9Z381 (EMBL:AL031806) (193 aa) fasta scores:

CDS 7373 8716 no significant database hits

misc_feature 8718 9154 Low G+C region containing repeat region with 10xTGTG(A/-)(T/C)AAAA(A/G)T.

CDS 9302 10081 no significant database hits

CDS 10192 12166 Previously sequenced Salmonella typhi topoisomerase B TopB TR:Q9RPF5 (EMBL:AF090001) (664

CDS 12654 13303 no significant database hits

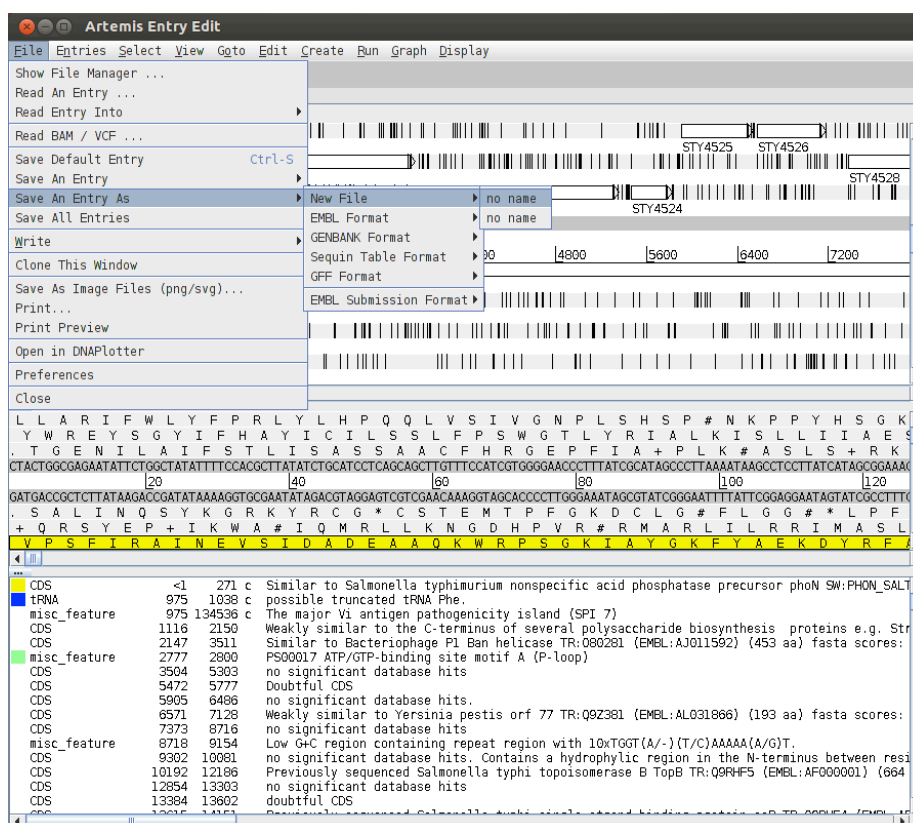
CDS 13384 13602 doubtful CDS

CDS 13635 14163 Previously sequenced Salmonella typhi topoisomerase B TopB TR:Q9RPF5 (EMBL:AF090001) (664

Note that the two entries on the grey 'Entry' line are now denoted 'no name'. They represent the same information in the same order as the original Artemis window but simply have no assigned 'Entry' names. As the sub-sequence is now viewed in a new Artemis session, this prevents the original files (S_typhi.dna and S_typhi.tab) from being over-written.

We will save the new files with relevant names to avoid confusion. So click on the 'File' menu then 'Save An Entry As' and then 'New File'. Another menu will ask you to choose one of the entries listed. At this point they will both be called 'no name'. Left click on the top entry in the list. A window will appear asking you to give this file a name. Save this file as spi7.dna

Do the same again for the second unnamed entry and save it as spi7.tab



We are going to look at this region in more detail and to attempt to define the limits of the bacteriophage that lies within this region. Luckily for us all the phage-related genes within this region have been given a colour code number 12 (pink; for a list of the other numerical values that Artemis will display as colours for features see **Appendix IX**). We are going to use this information to select all the relevant phage genes using the Feature selector as shown below and then define the limits of the bacteriophage.

First we need to create a new entry (click 'Create' then 'New Entry'). Another entry will appear on the entry line called, you guessed it, 'no name'. We will eventually copy all our phage-related genes into here.

1 Click 'Select' then 'Feature Selector'

2 Set Key to 'CDS' and Qualifier to 'colour'

3 Type search term

4 Click to select features containing search term

5 Click to view selected features in a list

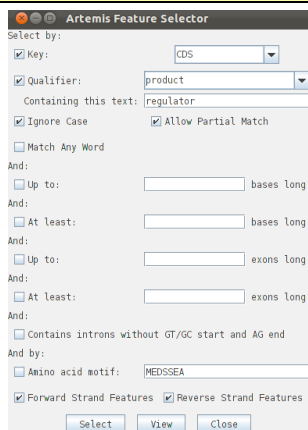
6 feature list

The genes listed in **(6)** are only those fitting your selection criteria. They can be copied or cut / moved in to a new entry so we can view them in isolation from the rest of the information within spi7.tab.

Firstly in window **(6)** select all of the CDSs shown by clicking on the ‘Select’ menu and then selecting ‘All’. All the features listed in window **(6)** should now be highlighted. To copy them to another entry (file) click ‘Edit’ then ‘Copy Selected Features To’ then ‘no name’. Close the two smaller feature selector windows and return to the SPI-7 Artemis window. You could rename the ‘no name’ entry as phage.tab, as you did before (if you can’t remember how to do it have a look at page 14). Temporarily remove the features contained in ‘spi7.tab’ file by left clicking on the entry button on the grey entry line. Only the phage genes should remain.

Additional methods for selecting/extracting features using the Feature Selector

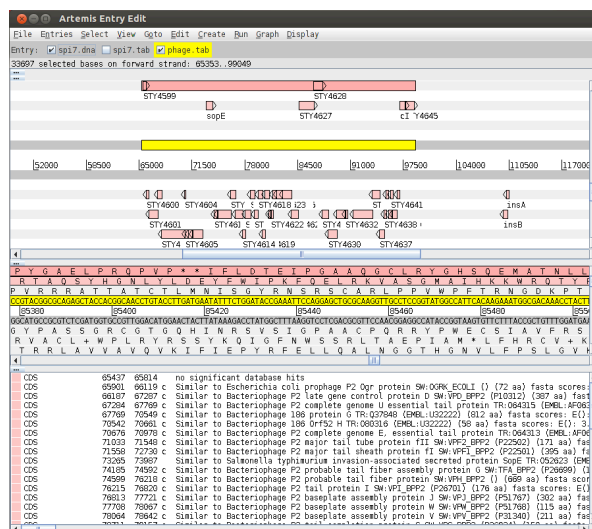
It is worth noting that the Feature Selector can be used in many other ways to select and extract subsets of features from the genome, using eg text or amino acid searches.



Space for a search term or amino acid motif

Defining the extent of the prophage

Even from this preliminary analysis it is clear that the prophage occupies a fairly discrete region within SPI-7 (see below). It is often useful to create a new DNA feature to define the limits of this type of genome landmark. To do this use the left mouse button to click and drag over the region that you think defines the prophage.



While the region is highlighted, click on the 'Create' menu and select 'Create feature from base range'. A feature edit window will appear. The default 'Key' value given by Artemis when creating a new feature is 'CDS'. With this 'Key' the newly created feature would automatically be put on the translation line. However, if we change this to 'misc_feature' (an option in the 'Key' drop down menu in the top left hand corner of the Edit window), Artemis will place this feature on the DNA line. This is perhaps more appropriate and is easier to visualise. You can also add a qualifier, such as '/label': select 'label' from the 'Add Qualifier' list and click 'Add Qualifier', '/label=' will appear in the text window; add text of your choice, then click 'OK'. That text will be used as a feature label to be displayed in the main sequence view panel.

To see how well you have done, tick the little box to turn on spi7.tab.

Your final task is to write out the spi7 files in EMBL submission format, and create a merged annotation and sequence file in EMBL submission format. In Artemis you are going to copy the annotation features from the '.tab' file into the '.dna' file, and then save this entry in EMBL format. Don't worry about error messages popping up. This is because not all entries are accepted by the EMBL database.

1 Click 'Select' then 'All'

2 Click 'Edit', then 'Copy Selected Features To'

3 Select 'spi7.dna'

The screenshot shows the Artemis Entry Edit window. The 'Edit' menu is open, and the 'Copy Selected Features To' option is highlighted. The 'spi7.dna' file is selected in the list of target files. The background shows a genomic map with various features and a sequence viewer at the bottom.

4 Click 'File' then 'Save An Entry As'

5 'EMBL Submission Format'

6 Select 'spi7.dna'

7 Save file as spi7.embl

The screenshot shows the Artemis Entry Edit window. The 'File' menu is open, and the 'Save An Entry As' option is highlighted. The 'EMBL Submission Format' is selected in the list of file formats. The 'spi7.dna' file is selected in the list of target files. The background shows a genomic map with various features and a sequence viewer at the bottom.

Now open the EMBL format file that you have just created in Artemis.

Artemis Entry Edit: spi7.embl

File Entries Select View Goto Edit Create Run Graph Display

Entry: ☒ spi7.embl

Nothing selected

STY4521 STY4523 STY4524 STY4525 STY4526 STY4528

misc_feature

tRNA feature

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

Feature	Start	End	Description
tRNA	1	64	c possible truncated tRNA Phe.
misc_feature	1	133562	c The major Vi antigen pathogenicity island (SPI 7)
CDS	142	1176	Weakly similar to the C-terminus of several polysaccharide biosynthesis proteins e.g. Str
CDS	1173	2537	Similar to Bacteriophage P1 Ban helicase TR:080281 (EMBL:AJ011592) (453 aa) fasta scores:
misc_feature	1803	1826	PS00017 ATP/GTP-binding site motif A (P-loop)
CDS	2530	4329	no significant database hits
CDS	4498	4803	Doubtful CDS
CDS	4931	5512	no significant database hits.
CDS	5597	6154	Weakly similar to Yersinia pestis orf 77 TR:Q9Z381 (EMBL:AL031866) (193 aa) fasta scores:
CDS	6399	7742	no significant database hits
misc_feature	7744	8180	Low G+C region containing repeat region with 10xTGGT(A/-)(T/C)AAAA(A/G)T.
CDS	8328	9107	no significant database hits. Contains a hydrophilic region in the N-terminus between resi
CDS	9218	11212	Previously sequenced Salmonella typhi topoisomerase B TopB TR:Q9RHF5 (EMBL:AF000001) (664
CDS	11880	12329	no significant database hits
CDS	12410	12628	doubtful CDS
CDS	12641	13177	Previously sequenced Salmonella typhi single strand binding protein ssB TR:Q9RHE4 (EMBL:AF

You will see that the colours of the features have now changed. This is because not all the qualifiers in the previous entry are accepted by the EMBL database, so some have not been saved in this format. This includes the '/colour' qualifier, so Artemis displays the features with default colours.

When you download sequence files from EMBL and visualize them in Artemis you will notice that they are displayed using default colours. You can customize your own annotation files with the '/colour' qualifier and chosen number (**Appendix IX**), to differentiate features. To do this you can use the Feature Selector to select certain features and annotate them all using the 'Edit', 'Change Qualifiers of Selected' function.

Artemis Exercise 4

This exercise will introduce you to database searches and will give you a first insight in the annotation of genes.

The gene you will work on is *hpcC* (STY1136). Go to this gene by using one the different methods you have learned so far.

As you can see the gene is full with stop codons indicating that we are looking at a pseudogene. To correct the annotation we are going to use database search. Follow now the numbers in the figure below to start a database search. The search may take a couple of minutes to run; a banner will pop up to tell you when its complete (3).

1 Select CDS

2 Start fasta

3 OK

Artemis Entry Edit: S_typhi.dna

Select View Goto Edit Create Run Graph Display

phi.dna ☒ S_typhi.tab

Feature: bases 1466 amino acids 488

hpcC hpcB

misc_feature

1099200 1100000 1100800

V V V E V E G V G R L V N R I V

L S L K + K A W V A W L T E S S

C R * S R R R G S P G # P N R C

TTGTGCTTGAAGTAGAAGCGGTGGGTGGCTGGTTAACCGAATCGTCA

1099540 1099560

AACAGCAACTTCATCTTCGCGACCCGCGGACCAATTGGCTTAGCAGT

T T T S T S P T P R R T L R I T L S S L A F H L F Y I M P N V A F V N Q T V V V E L C

N D N F Y F A H T A Q N V S D D T L L A C F S S F L D N S # R C F R Q L Y R S S G S

Q R Q L L L R P D G P # G F R * H P S R L I F F I F * Q I L P L F T A P L S + K W V

Run fasta on selected features against

Run blastp on selected features against

Run clustalx (PROTEIN) on selected features

Run jalview (PROTEIN) on selected features

Run tblastx on selected features against

Run blastn on selected features against

Run blastx on selected features against

Run fastx (%uniprot) on selected features

Run clustalx (DNA) on selected features

Run jalview on selected features

Set fasta options

Set blastp options

Set clustalx options

Set jalview options

Set tblastx options

Set blastn options

Set blastx options

Set fastx options

Set clustalx options

fasta process completed

OK

Feature	Start	End	Description
CDS	1099596	1101061	Pseudogene. Similar to Escherichia coli 5-carboxymethyl-2-hydroxyisocaproate semialdehyde dehydrogenase
misc_feature	1100321	1100344	PS00687 Aldehyde dehydrogenases glutamic acid active site
misc_feature	1100405	1100440	PS00070 Aldehyde dehydrogenases cysteine active site
CDS	1101063	1101914	Similar to Escherichia coli 3,4-dihydroxyphenylacetate 2,3-dioxygenase hpcB SW:HPCB ECOL
CDS	1101924	1102304	Similar to Escherichia coli 5-carboxymethyl-2-hydroxyisocaproate delta-isomerase hpcD SW:HF
CDS	1102448	1103251	Similar to Escherichia coli 2-oxo-hepta-3-ene-1,7-dioic acid hydratase hpcG SW:HPCG ECOL
CDS	1103262	1104053	Similar to Escherichia coli 2,4-dihydroxyhept-2-ene-1,7-dioic acid aldolase hpcH or hpaI
CDS	1104125	1105501	Similar to Escherichia coli putative 4-hydroxyphenylacetate permease hpaX TR:Q46984 (EME
CDS	1105511	1106407	Similar to Escherichia coli 4-hydroxyphenylacetate 3-monooxygenase operon regulatory pro
misc_feature	1106237	1106395	PS00041 Bacterial regulatory proteins, araC family signature
CDS	1106421	1107359	No significant database matches
CDS	1106783	1109088	Orthologue of E. coli yccD (YCCD_ECOLI); Fasta hit to YCCD_ECOLI (101 aa), 74% identity
CDS	1109088	1109088	Fasta hit to DNAI_ECOLI (375 aa), 35% identity in 353 aa overlap
CDS	1109814	1109873	PS00636 Nt-dnaI domain signature
misc_feature	1110244	1110606	Similar to Salmonella typhimurium suppressor for copper-sensitivity a ScaA TR:Q33917 (EM
CDS	1110655	1112541	Similar to Salmonella typhimurium suppressor for copper-sensitivity B precursor ScaB TR:
CDS	1112550	1113201	Similar to Salmonella typhimurium suppressor for copper-sensitivity C precursor ScaC TR:

To view the search results click 'View', then 'Search Results', then 'fasta results'. The results will appear in a scrollable window. Scroll down to the first sequence comparison and you should see the results as shown in the next figure.

Gene in database

To correct the annotation we have to edit the CDS now. Left click on the right amino acid continuing the amino acid sequence on the second frame (have a look in the fasta results and look at the sequence of the gene in the database when you are not sure) and drag till the end of the gene. Then click 'Create' 'Feature from base range' and 'OK'. A new blue CDS feature will appear on the appropriate frame line.

As the original gene annotation is too long we have to shorten it. Click on the original *hpcC* CDS, 'Edit' 'Selected features in Editor'. A window will pop up and you can change the end position in 'location' (the end position is the last base of the stop codon).

[illegible]

Artemis Feature Edit: hpcC

Key: CDS Add Qualifier: note

Location: 1099596..1099896

Complement Grab Range Remove Range Goto Feature Select Feature Tidy TAT Object Edit

```

/EC_number="EC 1.2.1.."
/blast_file="./old_whole_genome/blastp/St.tab.seq.0136.out"
/class="3.4.3"
/colour="11"
/gene="hpcC"
/gene="STY1136"
/rhth_file="./old_whole_genome/rhth/COFBA-St.tab.seq.0125.out"
/name="Pseudogene. Similar to Escherichia coli 5-carboxymethyl-2-hydroxyumconate
senialdehyde dehydrogenase hpcC SM:hPC_ECOLI (P42269) fasta scores: E(1): 0.96; 2% id in
366 aa. This CDS contains a frameshift after codon 97. The sequence has been checked and
is believed to be correct"
/product="5-carboxymethyl-2-hydroxyumconate senialdehyde dehydrogenase (pseudogene)"
/pseudo
/fasta_file="fasta_S_typhi.tab.seq.0132.out"
/fasta_file="kumpirot_bacteria.fasta_S_typhi.tab.seq.00009.out"

```

OK Cancel Apply

The new CDS feature can then be merged with the original gene as shown below (1-3).

A small window will appear asking you whether you are sure you want to merge these features. Another window will then ask you if you want to 'delete old features'. If you click 'yes' the CDS features you have just merged will disappear leaving the single merged CDS. If you select 'no' all of the three CDS features (the two CDSs you started with plus the merged feature) will be retained.

2

Click 'Edit'

3

'Selected Features' 'Merge'

1

Select both the original gene-model and the new CDS feature, which is to be merged with it to form a new gene

The screenshot shows the Artemis Entry Edit: S_typhi.dna window. The 'Edit' menu is open, and the 'Merge' option is selected. The 'Selected Features' window is also open, showing the 'Merge' button. The main display area shows a gene model with features hpcC and hpcD. The sequence view below shows the DNA sequence. The 'result' window at the bottom right shows the merged feature details, including the feature name, coordinates, and sequence.

Tip: To select more than one feature (of any type) you must hold the shift key down.

Artemis Exercise 4 - Second part

In the first part of the exercise you have learned how to correct a gene annotation. But what if you think a gene is missing?

Remember that there are loads of genomes that were submitted to the databases several years ago and in general the annotation is not updated to take into account new data. Sometimes it is worth checking regions which look strange to you.

Go to position 2,248,400 by using one of the different methods you have learned so far. If you look carefully you will notice a region shown below which there is no predicted gene. This type of non-coding region in *Salmonella* is very unusual (this is also true for other bacteria). To determine if this non-coding region is truly as published, load the codon usage information for *Salmonella* into Artemis by following the figure below.

The file 'S_typhi.cod' contains codon usage information taken from a public website (see below).

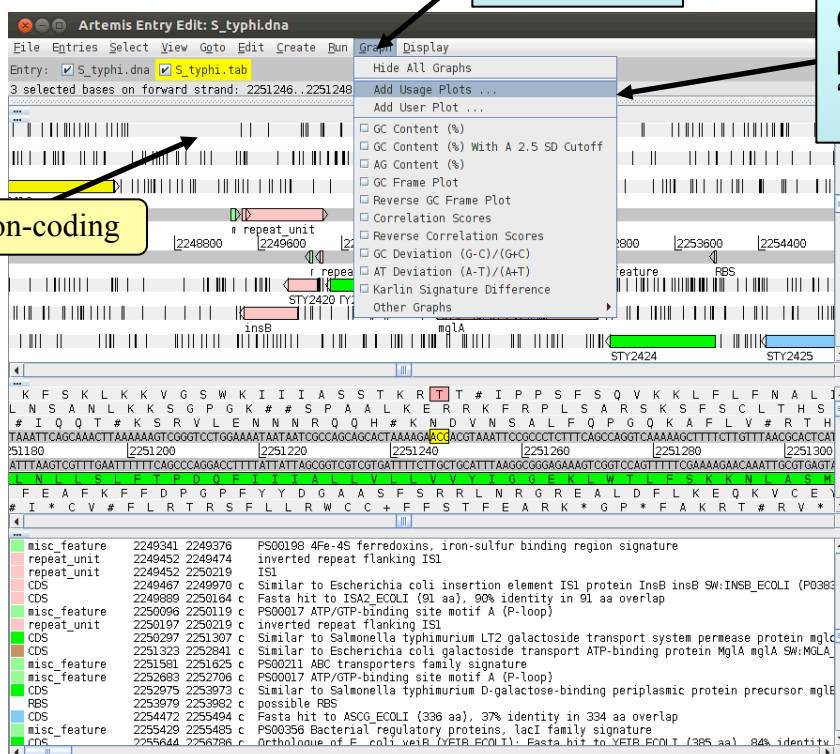
1

Click Graph

2

Click 'Add usage plots' and select 'S_typhi.cod'

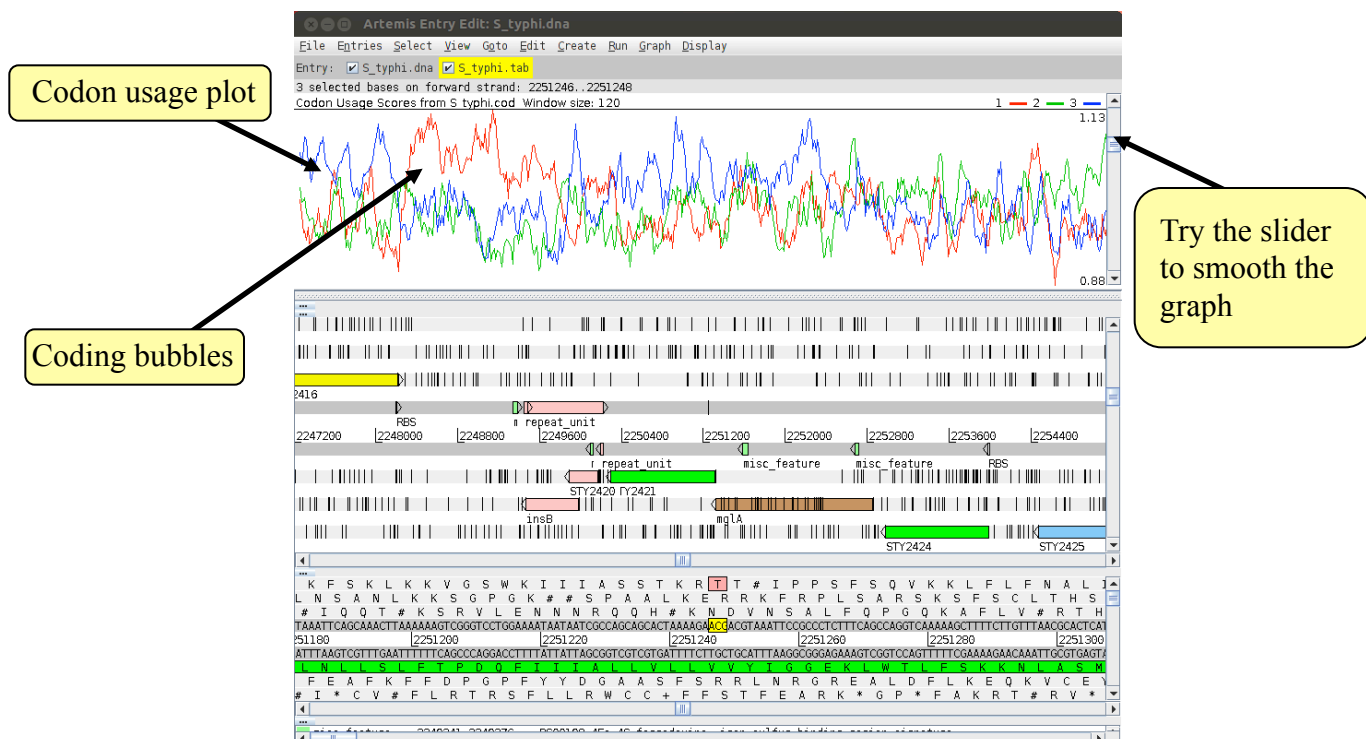
Non-coding



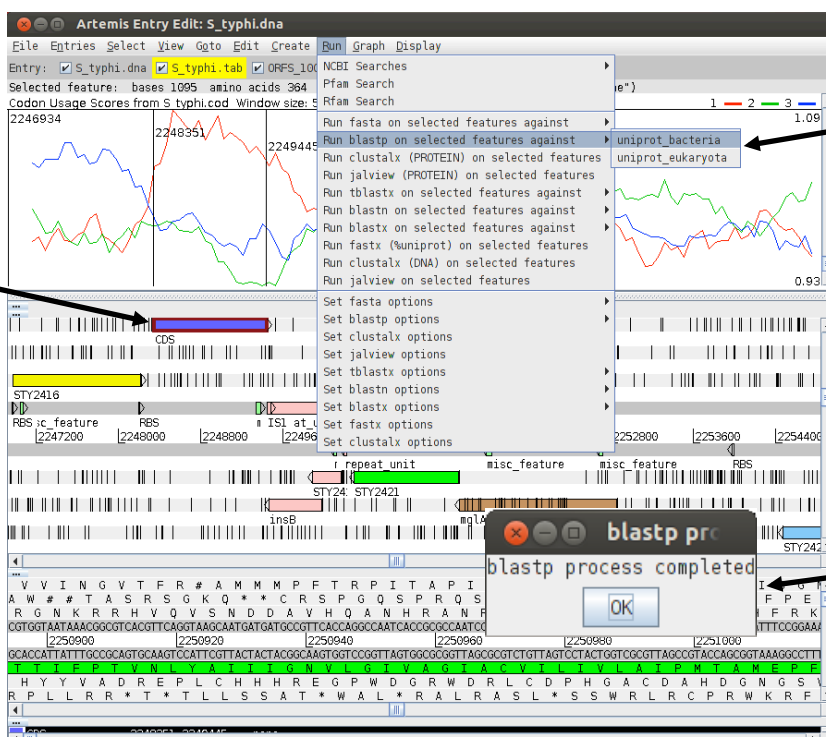
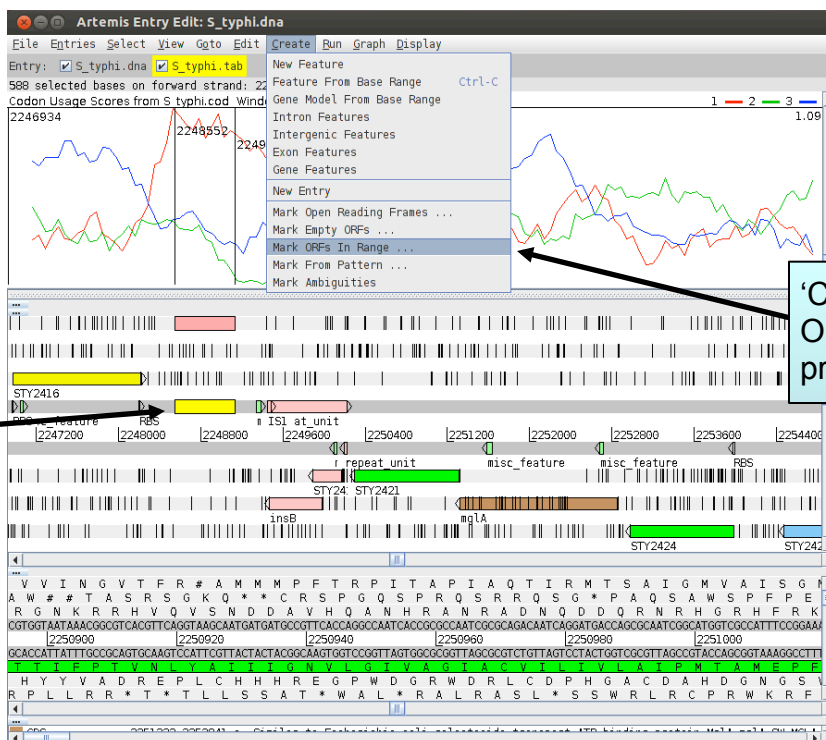
UUU 24.4(2445)	UCU 12.1(1212)	UAU 18.6(1864)	UGU 6.5(651)
UUC 15.6(1563)	UCC 18.5(1854)	UAC 13.3(1335)	UGC 5.8(585)
UUA 15.8(1582)	UCA 15.3(1534)	UAA 1.6(158)	UGA 1.4(145)
UUG 12.2(1222)	UCG 9.5(952)	UAG 0.5(47)	UGG 13.3(1326)
CUU 16.1(1614)	CCU 18.5(1855)	CAU 11.4(1143)	CGU 14.3(1426)
CUC 11.2(1120)	CCC 6.8(681)	CAC 7.2(725)	CGC 11.7(1170)
CUA 6.6(656)	CCA 9.3(938)	CAA 13.8(1381)	CGA 6.7(666)
CUG 34.1(3411)	CCG 14.1(1415)	CAG 27.2(2716)	CGG 8.2(822)
AUU 27.3(2729)	ACU 14.8(1476)	AUA 27.8(2699)	AGU 13.2(1319)
AUC 28.4(2837)	AAC 28.1(2812)	AAG 22.7(2266)	AGC 16.4(1639)
AUA 9.9(988)	ACA 14.4(1445)	AAA 35.4(3546)	AGA 6.8(599)
AUG 26.8(2597)	ACG 15.5(1552)	AAG 17.5(1752)	AGG 4.7(472)
GUU 28.3(2833)	GCU 17.7(1778)	GAA 33.9(3394)	GGU 19.5(1949)
GUC 15.4(1541)	GCC 21.6(2168)	GAC 28.8(1998)	GGC 21.8(2098)
GUA 12.8(1281)	GCA 21.8(2185)	GAA 34.8(3481)	GGA 12.6(1259)
GUG 19.4(1941)	GCG 19.3(1931)	GAG 28.8(2088)	GGG 13.6(1359)

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

When you first load the codon 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 the view click on the Graph menu and check the box alongside the option 'codon usage scores 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 in 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. So now we have to create the open reading frame (ORF), blast the amino acid sequence and add the annotation. Follow the instruction on the next page to do this.



To view the search results click 'View', then 'Search Results', then 'blastp results'. The results will appear in a scrollable window. You see that the product of the gene is "NAD-dependent dihydropyrimidine dehydrogenase subunit PreA). To add the product to the annotation follow the instruction in the next page.

The image shows two screenshots from the Artemis software interface. The top screenshot shows the 'Artemis Entry Edit: S_typhi.dna' window with the 'Edit' menu open. A callout box labeled '1' points to a feature on the track, and another callout box labeled '2' points to the 'Edit' menu. The bottom screenshot shows the 'Artemis Feature Edit: CDS' dialog box. Callout boxes labeled '3', '4', and '5' point to the 'product' dropdown, the 'Add Qualifier' button, and the text area respectively.

1 Click on ORF

2 'Edit', 'Selected feature in editor'

3 Select 'product' from dropdown list

4 Click 'Add qualifier'

5 Add result from blastp search

The annotation of the ORF is now complete. You can add as much information as you want. Have a look at the other qualifiers if some time is left. The last thing you have to do is copy the annotated feature to *S_typhi.tab*. To do that select the feature and go to 'Edit', 'Copy selected features to' and click '*S_typhi.tab*'. Don't forget to save the tab file.