

Module 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 visualize 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.

Artemis is not the best viewer for large genomes, so that later in the course you will be introduced to IGV....

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

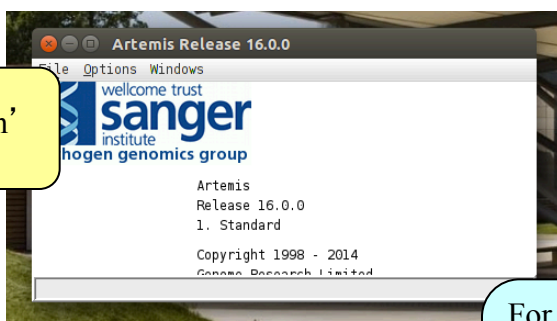
1. Starting up the Artemis software

In the terminal type
art &

A small start-up window will appear (see below). The directory is `/export/projects/bioinfo3/to16r/BioinfoWorkshop/Data/Module_Viewer/` and contains all files you will need for this module. (You might have copied the data already somewhere else).

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

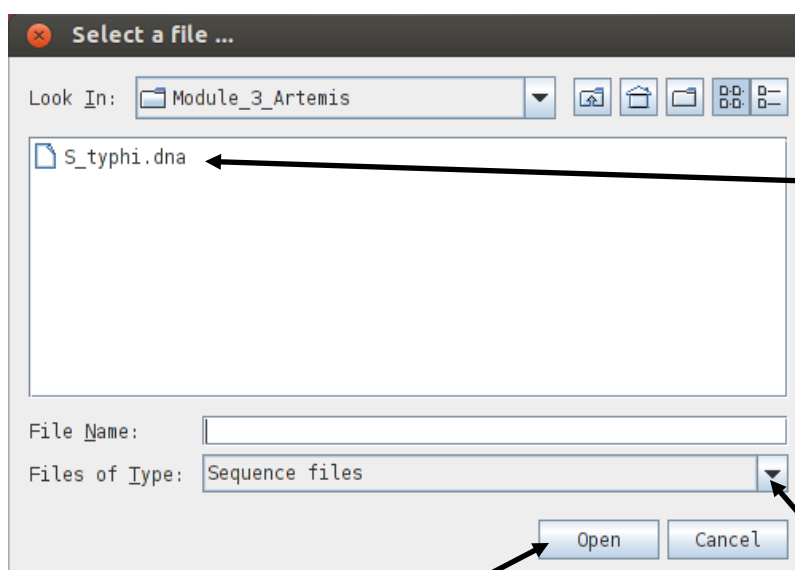
1. Click 'File'
2. Then 'Open'



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

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.

- 4 Single click to open file in Artemis then wait

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)

The annotated screenshot shows the 'File' menu open with 'Read An Entry ...' selected. A 'Select a file ...' dialog box is open, showing the file list with 'S_typhi.tab' selected. Arrows point from the numbered instructions to the corresponding UI elements: 'File' and 'Read An Entry ...' in the menu, 'S_typhi.tab' in the file list, and the 'Open' button in the dialog box.

3. The basics of Artemis

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

The screenshot shows the Artemis genome browser interface. At the top, there is a menu bar and a status bar. Below that, a selected feature is shown: bases 1287, amino acids 428, STY0004. The main panel displays a DNA sequence with various features represented by colored boxes. A feature list is shown at the bottom, listing features such as CDS, misc_feature, and other annotations. Numbered callouts 1-8 point to specific UI elements:

- 1. File menu
- 2. Entry selection (S_typhi.dna, S_typhi.tab)
- 3. Selected feature details (bases 1287, amino acids 428, STY0004)
- 4. DNA sequence view (nucleotides and amino acids)
- 5. Feature list (CDS, misc_feature, etc.)
- 6. Sliders for zooming view panels
- 7. Sliders for scrolling along the DNA
- 8. Slider for scrolling feature list

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 Exercise 4)

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!

Click 'Goto'

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 in
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
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
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
CDS	5966	7386	Similar to Bacillus subtilis amino acid carrier protein alst ALST SW:ALST_BACSU (Q45068);
misc_feature	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
misc_feature	8933	8974	PS01078 Molybdenum cofactor biosynthesis proteins 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 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.

The image shows the Artemis software interface. The top window is titled 'Artemis Entry Edit: S_typhi.dna'. A yellow callout box with a black border points to the 'Goto' button in the menu bar, containing the text 'Click 'Goto' then Navigator'. Below this, the Navigator panel is visible, showing a list of features with their coordinates. A second yellow callout box points to the 'Goto' button in the Navigator panel, containing the text 'Check that the appropriate search button is on'. In the foreground, a search dialog box titled 'Artemis Navigator' is open. It has several search options: 'Goto Base:', 'Goto Feature With Gene Name:', 'Goto Feature With This Qualifier Value:', 'Goto Feature With This Key:', 'Find Base Pattern:', and 'Find Amino Acid String:'. The 'Goto Feature With This Qualifier Value:' option is selected. The dialog also has checkboxes for 'Overlaps With Selection', 'Forward Strand', 'Reverse Strand', 'Search Backward', 'Ignore Case', and 'Allow Substring Matches'. There are 'Goto', 'Clear', and 'Close' buttons at the bottom of the dialog. The background shows a genomic map with various features like CDS, misc_feature, and STY0001-32.

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 '*fis*').
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 VI** 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 VIII**).
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 Optional 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 *fbxB* 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.

The screenshot shows the Artemis genome browser interface. The main display area shows a DNA sequence with various features represented by colored bars and labels. The features are organized into tracks, including CDS (Coding DNA Sequence) features and Misc features. The DNA sequence is displayed in a monospaced font, with a green highlight under the sequence. The features are labeled with STY IDs and coordinates. Two callout boxes on the right side of the image point to specific features: 'CDS features' points to a green bar labeled 'STY2368' and 'Misc features' points to a yellow bar labeled 'misc_feature'.

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	PS1042 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 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 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	c Similar to Bacillus subtilis amino acid carrier protein alaT ALST SW:ALST_BACSU (Q45068; P40743) fasta hit
CDS	7091	7138	c PS00873 Sodium:alanine symporter family signature
misc_feature	7665	8618	Fasta hit to TALA_ECOLI (316 aa), 65% identity in 311 aa overlap
misc_feature	7755	7781	PS1054 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	PS10178 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 or fasta. 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.

Load additional files

You should be able to see the results from Prosite searches, run on the translation of each CDS, as pale-green boxes on the grey DNA lines. The results from the Pfam protein motif searches are not yet shown, but can be viewed by loading the appropriate file. Click on 'File' then 'Read an Entry' and select the file PF.tab. Each Pfam match will appear as a coloured blue feature in the main display panel on the grey DNA lines. To see the details click the feature then click 'View' then 'Selection' or click 'Edit' then 'Selected Features in Editor'. Please ask if you are unsure about Prosite and Pfam.

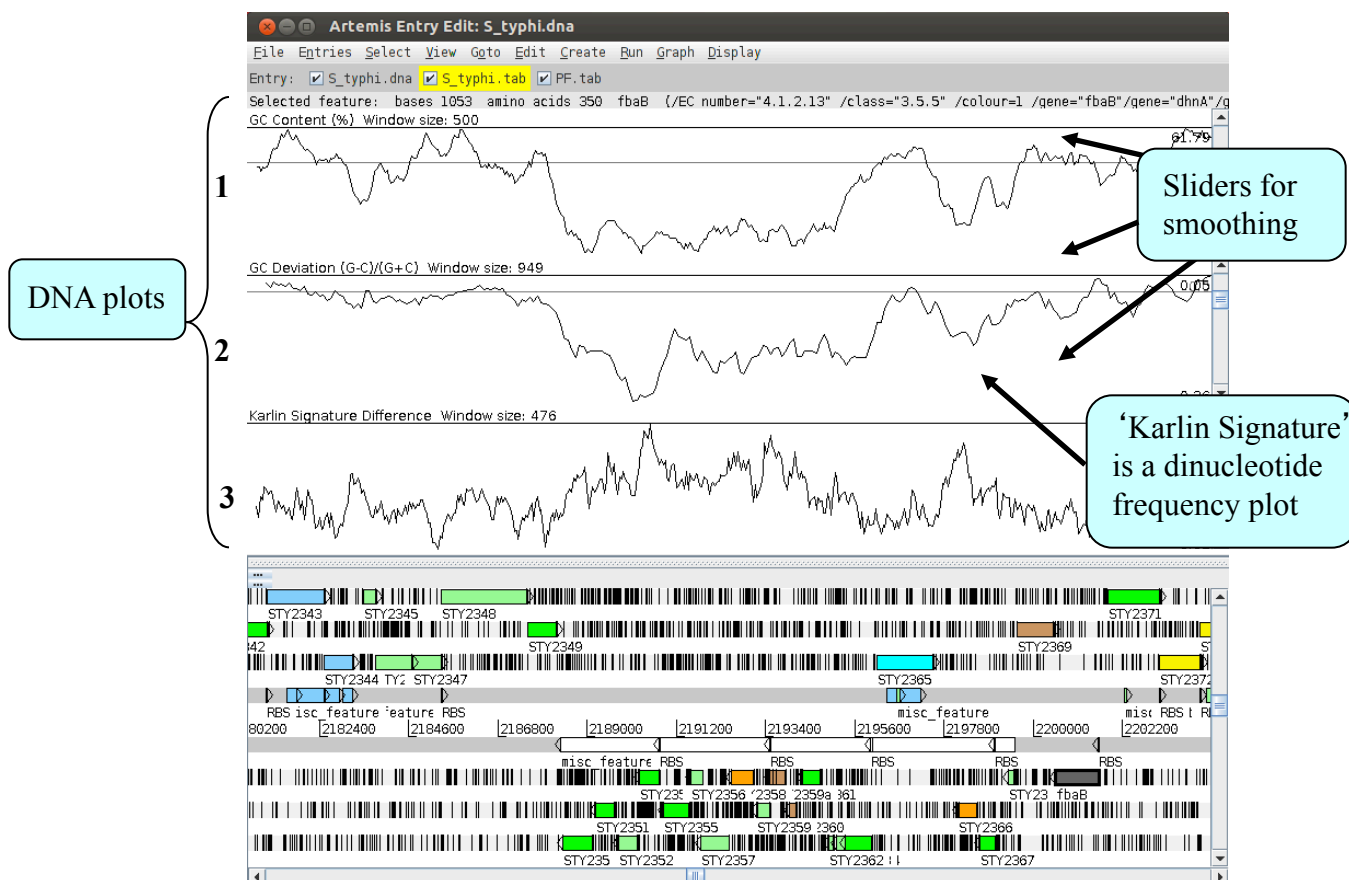
Further information on specific Prosite or Pfam entries can be found on the web at:
<http://ca.expasy.org/prosite> and <http://pfam.sanger.ac.uk/>

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, such as GC frame plot, and others can be used to search for horizontally acquired DNA.

The plot information is generated dynamically by Artemis and although this is a relatively speedy exercise for a small region of DNA, on a whole genome view (we will move onto this later) this may take a little time, so be patient.

To view the graphs:

Click on the 'Graph' menu to see all those available. Perhaps some of the most useful plots are the (1) 'GC Content (%)', (2) 'GC Deviation' and (3) 'Karlin Signature Difference' as shown below. To adjust the smoothing of the graph you change the window size over which the points on the graph are calculated, using the sliders shown below. If you are not familiar with any of these please ask.



Notice how several of the plots show 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.

Your Artemis window should now look similar to the one shown.

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 below. However, be careful zooming out quickly with all the features being displayed, as this may temporarily lock up the computer.

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.

No stop codons shown on frame lines

Menu item for de-selecting stop codons

1

Artemis Entry Edit: S_typhi.dna
 File Entries Select View Goto Edit Create Run Graph Display
 Entry: S_typhi.dna S_typhi.tab PF.tab
 Selected feature: bases 1053 amino acids 350 fbaB (/EC number="4.1.2.13" /class="3.5.5" /colour=1 /gene="fbaB"/gene="dhnA"/g
 GC Content (%) Window size: 500
 GC Deviation (G-C)/(G+C) Window size: 949
 Karlin Signature Difference Window size: 476
 STY2343 STY2345 STY2348
 STY2349
 STY2344 TYZ STY2347
 RBS isc feature feature RBS
 80200 |2182400 |2184600 |2186800
 misc
 STY2365 STY2372
 misc feature
 2195600 |2197800 |2200000 |2202200
 RBS RBS RBS
 361
 STY23 STY2366
 STY2362:1

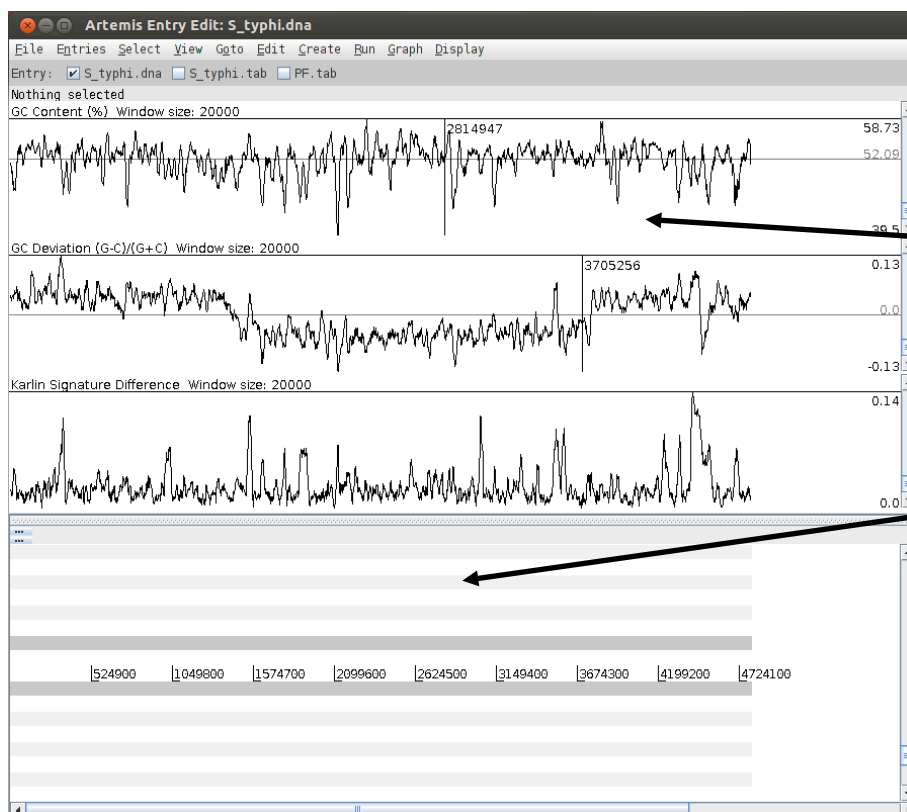
- Smallest Features In Front
- Set Score Cutoffs ...
- Raise Selected Features
- Lower Selected Features
- Zoom to Selection
- Select Visible Range
- Select Visible Features
- Frame Line Features ...
- Entries
- Select
- View
- Goto
- Edit
- Create
- Write
- Run
- Start Codons
- Stop Codons
- Feature 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

3 Graph scaling menu

4 Slider for zooming out

The screenshot shows the Artemis software interface with three stacked line graphs: GC Content (%), GC Deviation (G-C)/(G+C), and Karlin Signature Difference. A right-click context menu is open over the top graph, with 'Set the Window Size...' selected. A dialog box titled 'Set Window Size' is open, showing a text input field with the value '20000' and 'Set Window Size' and 'Cancel' buttons. A slider on the right side of the window is being moved to zoom out.

3. 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).
4. You are now ready to zoom out by dragging or clicking the slider indicated above. 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.