WHO/TDR Bioinformatics Workshop at ICGEB, New Delhi (2005)
Index

Module 1
Artemis: Prokaryotic

Introduction

Artemis (Rutherford *et al*., (2000) is a DNA viewer program, written by Kim Rutherford, and used for both Prokaryotic and Eukaryotic annotations. It allows the user to get away from the relatively faceless EMBL and Genbank style database files and view the sequence in a graphical and highly interactive format. Artemis is designed to present multiple lines of information within a single context. This manifests itself as being able to zoom in to look for fine DNA motifs as well as being able to zoom out and bring into view operons, several kilobases of a genome or in fact to view an entire genome in one screen. It is also possible to perform quite sophisticated analyses and store the output within the 'Artemis environment' to be accessed later.

Aims

The aim of this Module is for you to become familiar with the basic functioning of Artemis by 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 and gain a basic understanding of Artemis. Like all the Modules in this workshop, the key is ‘if you don’t understand please ask’.

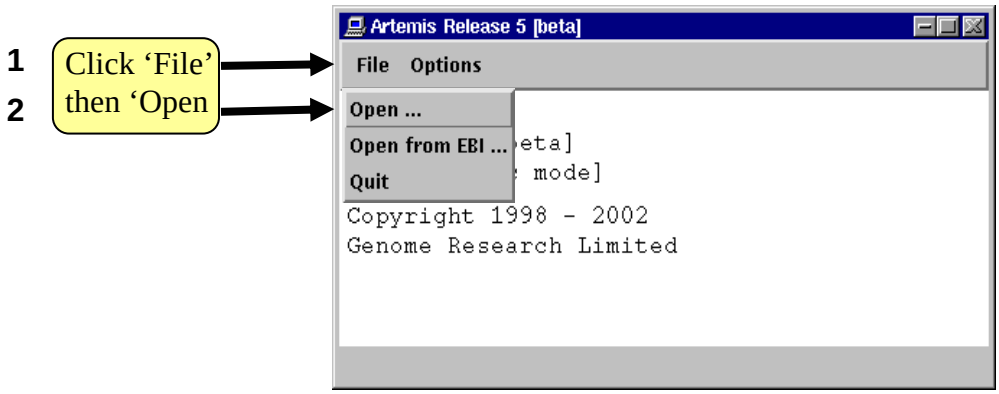
Artemis Exercise 1 Part I

1. Starting up the Artemis software

Navigate your way into the correct directory for this module
Then type:

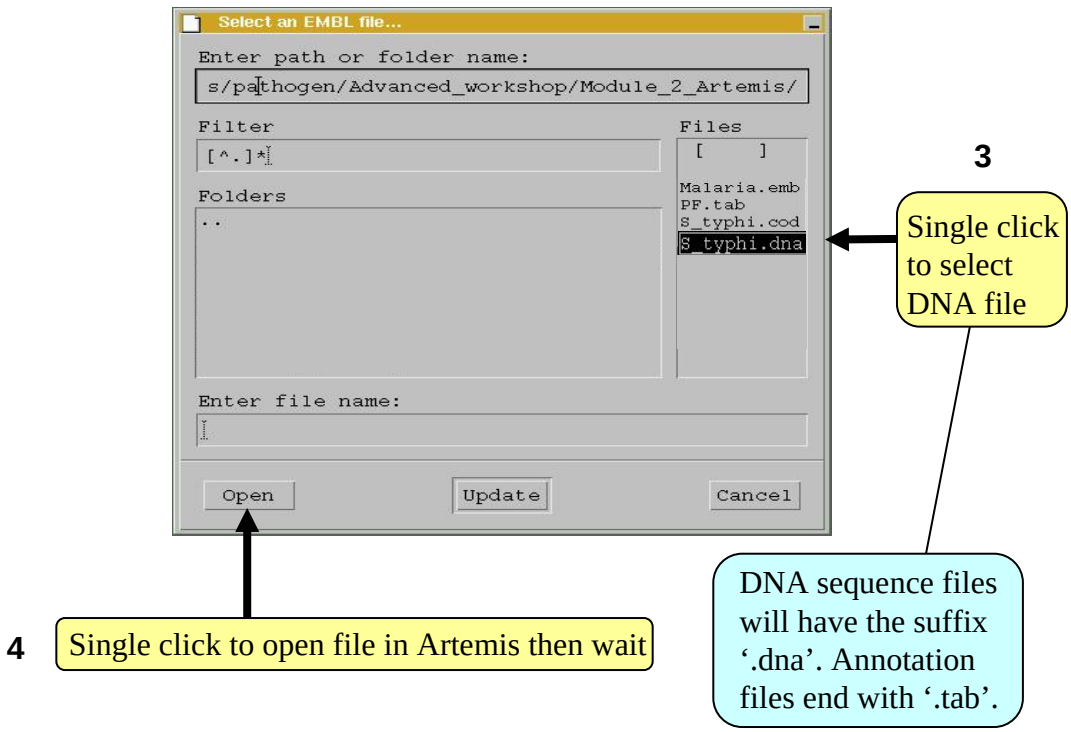
art & [return]

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.



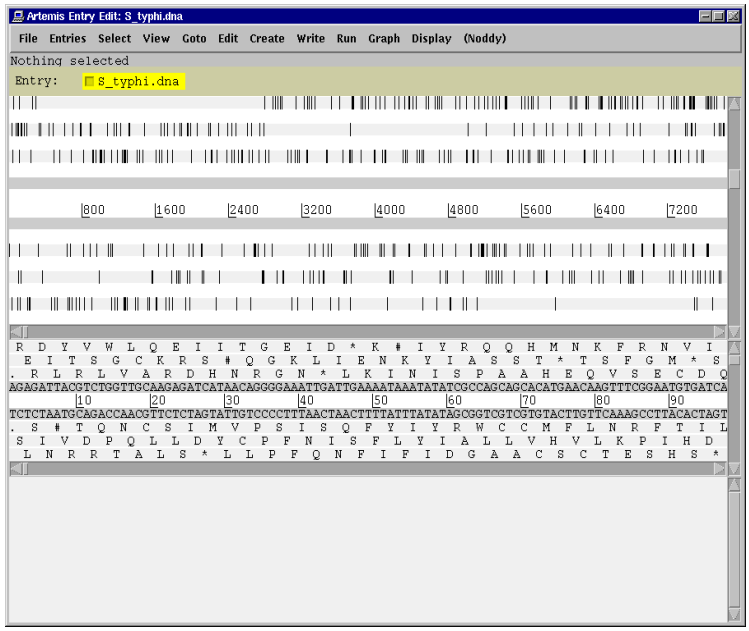
For simplicity it is a good idea to open a new start up window for each Artemis session and close down any sessions once you have finished an exercise.

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

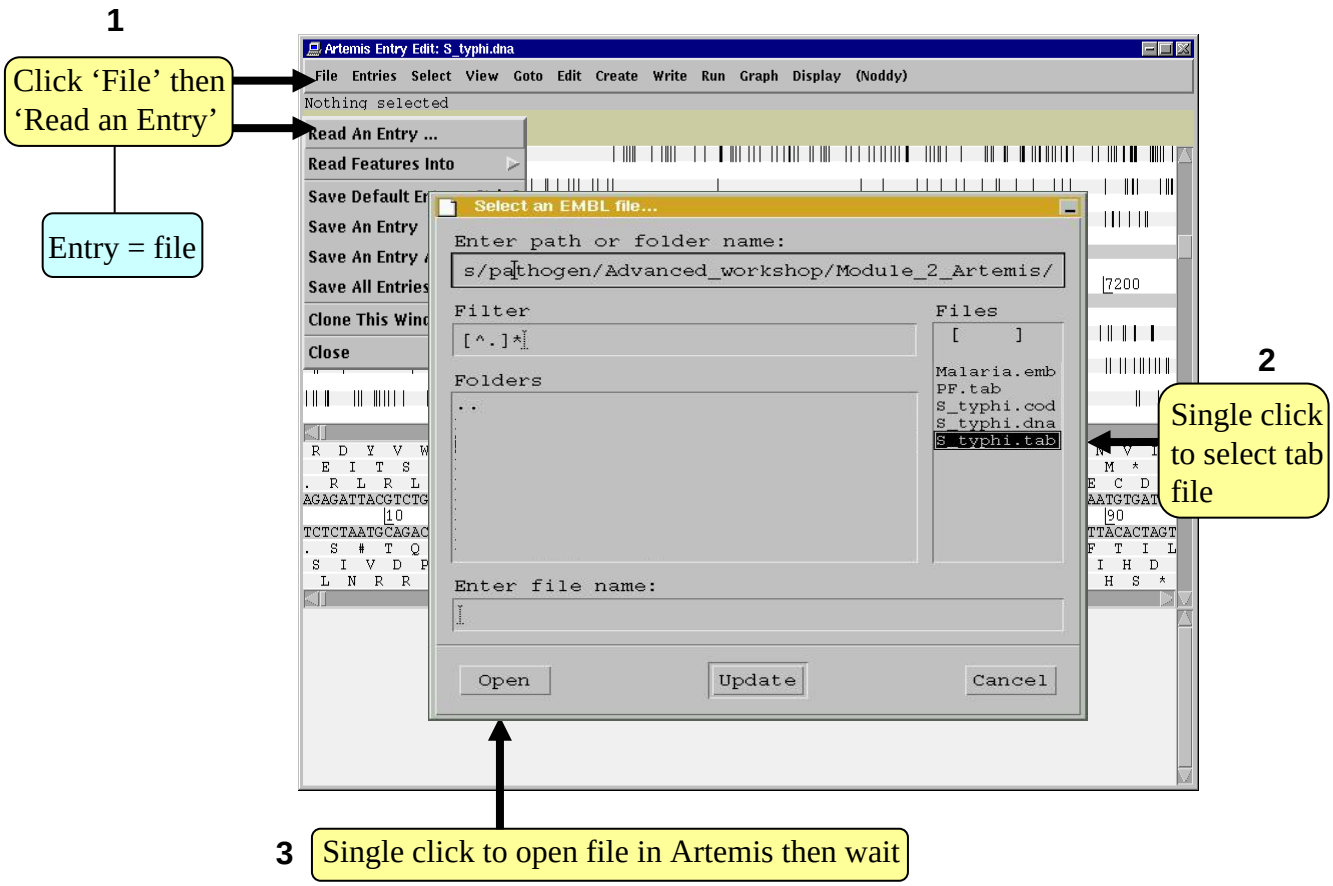


2. Loading annotation files (entries) into Artemis

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



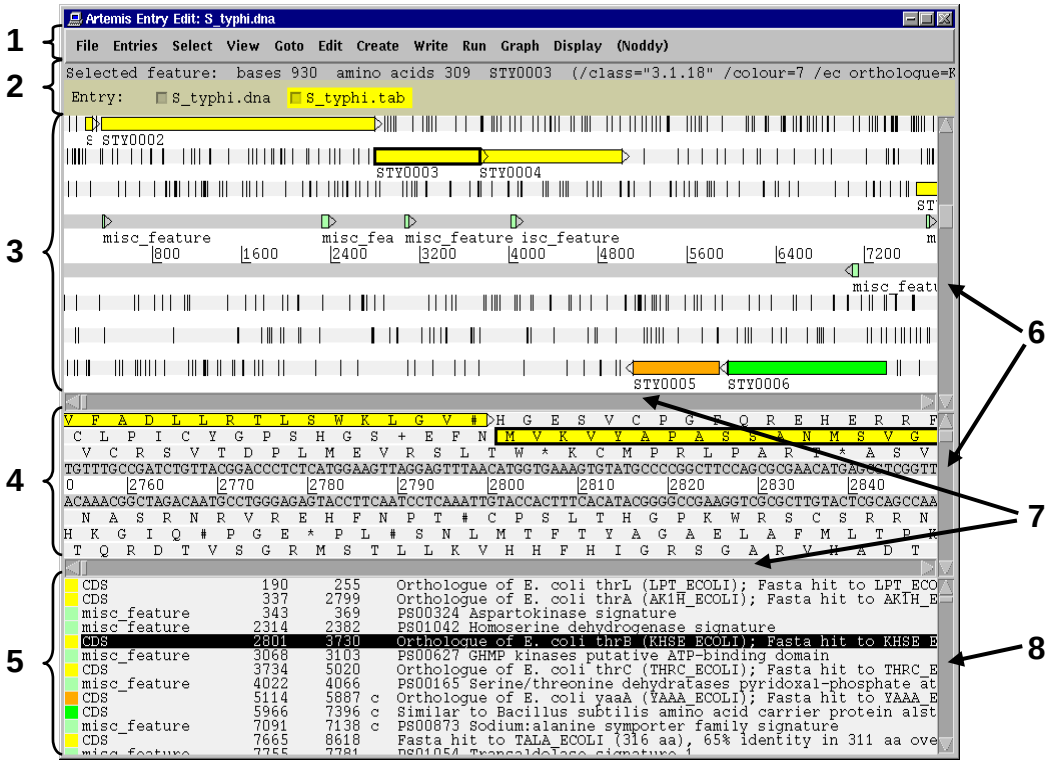
Now follow the numbers to load up the annotation file for the *Salmonella typhi* chromosome.



What's an "Entry"? It's a file of DNA and/or amino acid features which can be overlaid onto the sequence information displayed 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.



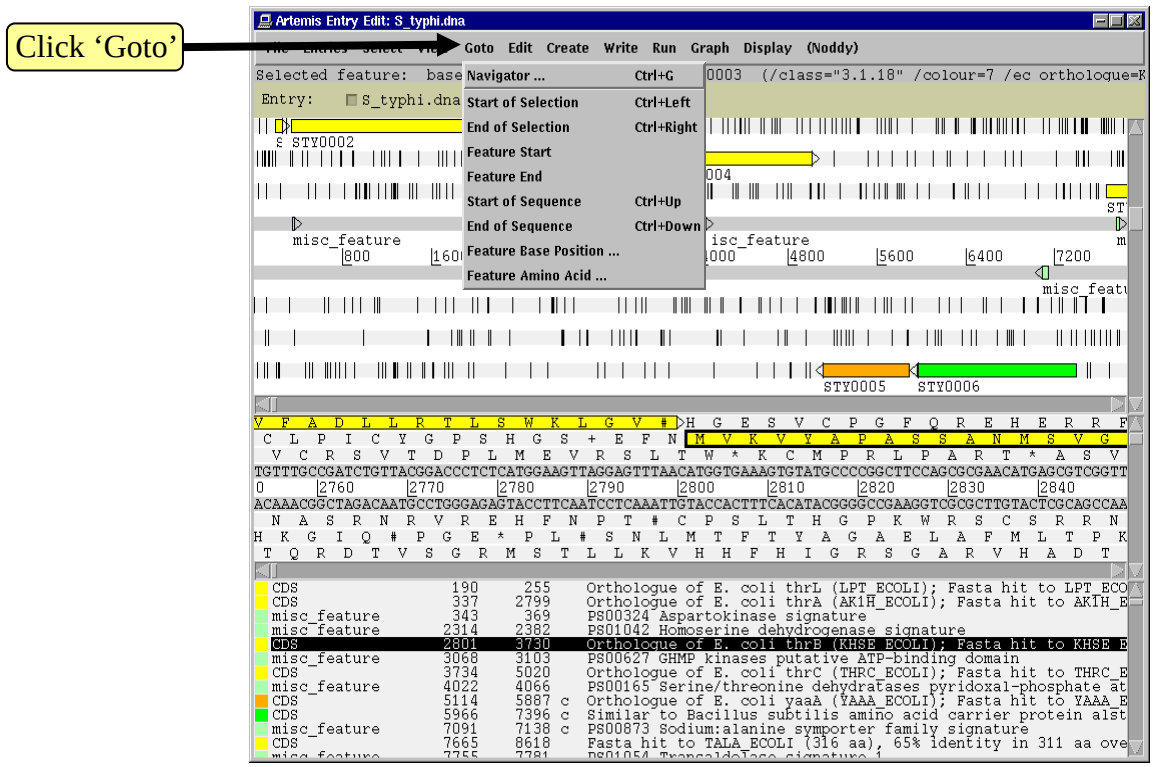
1. Drop-down menus. There's lots in there so don't worry about them right now.
2. Shows what entries are currently loaded (bottom line) and gives details regarding the feature selected in the window below; in this case gene STY0003 (top line).
3. This is the main sequence view panel. The central 2 grey 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. We will refer to genes as coding sequences or CDSs from now on.
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.

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!



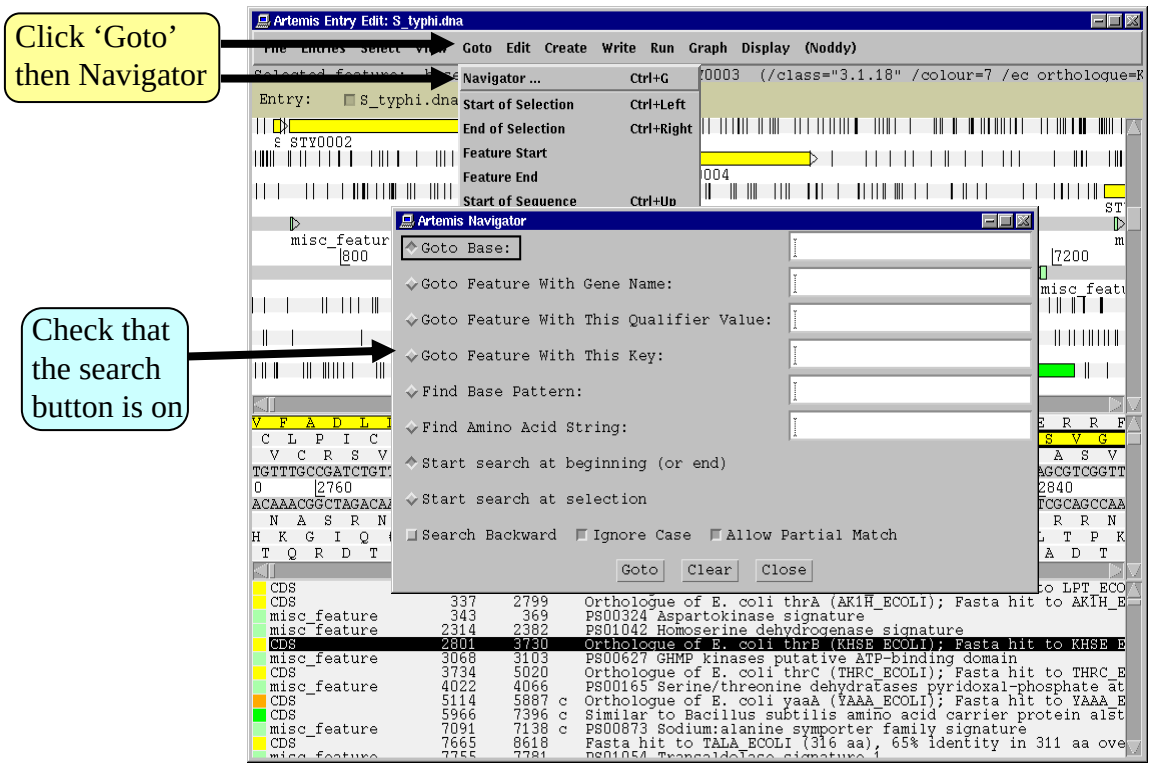
It may seem that 'Goto' 'Start of Selection' and 'Goto' 'Feature Start' do the same thing. Well they do if you have a feature selected but 'Goto' 'Start of Selection' will also work for a region which you have highlighted by click-dragging in the main window. So yes, give it a try!

Suggested tasks:

1. Zoom out, highlight a large region of sequence 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.



Suggestions of where to go:

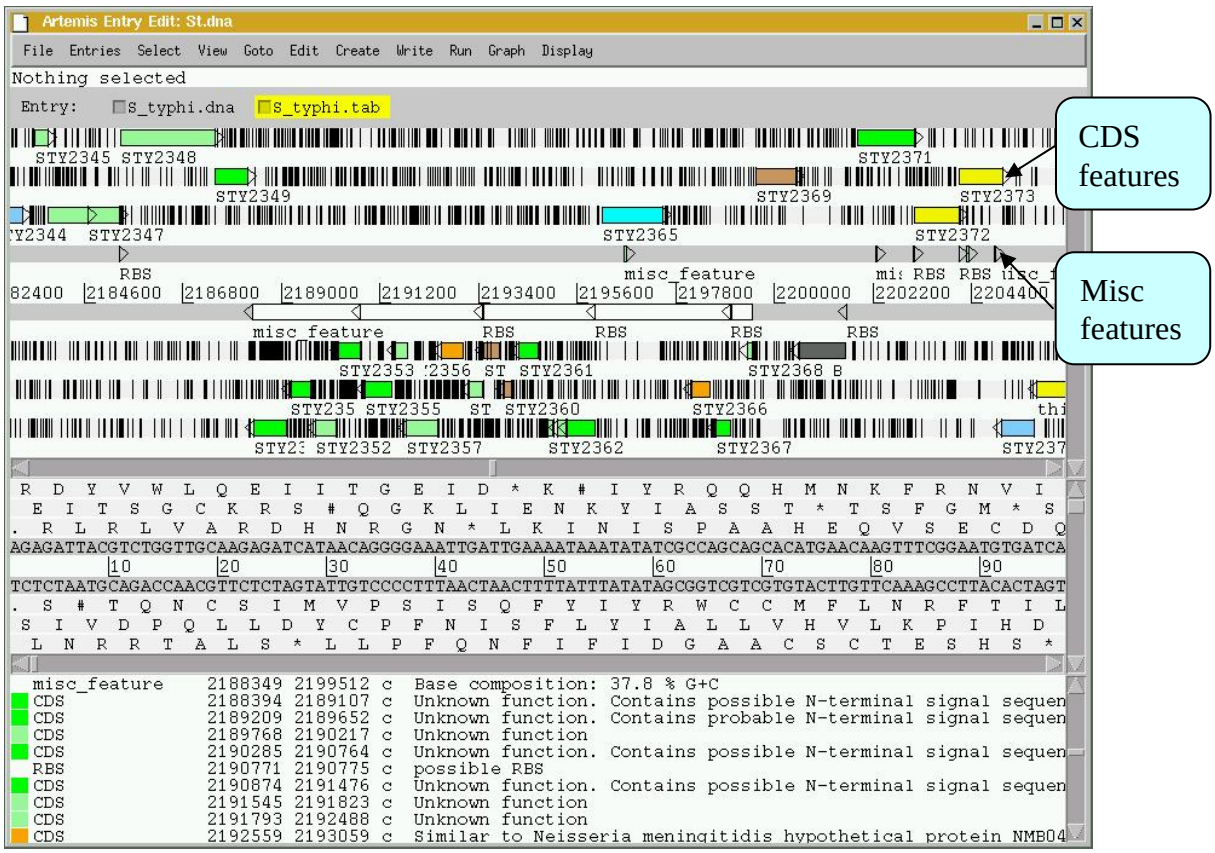
1. Think of a number between 1 and 4809037 and go to that base (notice how the cursors on the horizontal sliders move with you).
2. Your favourite gene name (it may not be there so you could try 'fts').
3. Use 'Goto Feature With This Qualifier value' to search the contents of all qualifiers for a particular 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 VIII).
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 III

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 1 Part II

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 located between 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 either the Navigator, Feature Selector or Goto functions discussed previously to get there. The region you arrive at should look similar to that shown below.



Once you have found this region have a look at some of the information that is available to you:-

Information to view:

Annotation

If you click on a particular feature you can view the annotation attached to it: select a CDS feature (or any other feature) and click on the Edit menu and select Edit Selected Feature. 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 as seen in Module 1.

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 '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 run on the translation of each CDS should already be on display as pale-green boxes on the grey DNA lines. The results from the Pfam protein motif searches are not 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 'View Selection' or click 'Edit' then Edit Selected Features'. Please ask if you are unsure about Prosite and Pfam.

Viewing the results of database searches

Click the 'View' menu, then select 'Search Results' and then 'Fasta results'. The results of the database search will appear in a scrollable window. If you click on the button at the bottom of this window labelled 'view in browser', then the results will be posted into an internet browser window. Within this window there are many active links (coloured blue), to external sources of information such as the original database entries for all those aligning to your sequence, as well as information stored in PubMed, PFAM and many others. This is your opportunity to explore some of the other features of Artemis whilst we are here to help.

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

Module 2
Artemis: Eukaryotic

Introduction

Following a similar format to Module 1, this Module will introduce Eukaryotic sequence analysis using Artemis. This exercise will look at a section of the Malaria genome. Your task is to assess the gene models that we have given you and to assess whether they are acceptable or in need of modification. To do this you will use G+C content to identify possible missing exons and then run database searches in order to see if there are similar CDS in the public databases. Note that there is not always a perfect answer when creating gene-models and a certain amount of subjectivity can be involved.

Aims

The aim of this Module is for you to become familiar with creating CDS features and merging them to create multi-exon gene models for this region of sequence. You will also find out how to run database searches against a locally installed public sequence database.

Exercise 1

This exercise will look at a section of the Malaria genome. You will need to close down the last Artemis exercise if you haven't already done so. Then start a new Artemis Session, as before, using the file 'Malaria.embl' in the current directory (Module_2_Artemis). Unlike the Salmonella exercise, in this instance the annotation and sequence are contained within the same file 'Malaria.embl'
The sequence you are going to look at is a small region of contrived sequence (~21 kb) taken from *Plasmodium falciparum* chromosome 13. You will see 7 CDSs, some with multiple exons. As a gentle introduction to splicing we would like you to look at the genes named , PF13_0119, MAL13P1.294 and PF13_0061. They have only been partially characterised and may in fact be missing exons. Have a look at these CDSs and confirm, edit or dismiss the proposed gene models by using G+C content, database searches and looking for splice sites (**Appendix IX**).

G+C content is a very good indicator of coding capacity in Malaria. On average, the coding regions are ~23% G+C and the non-coding regions are ~19%. Have a look at the G+C content for this region by selecting the appropriate graph. Left click within the graph window and then select by clicking on the exons to see how this relates to the G+C peaks on the graph.

Note, we will cover the principals and methods of gene prediction in much more detail in a module 3.

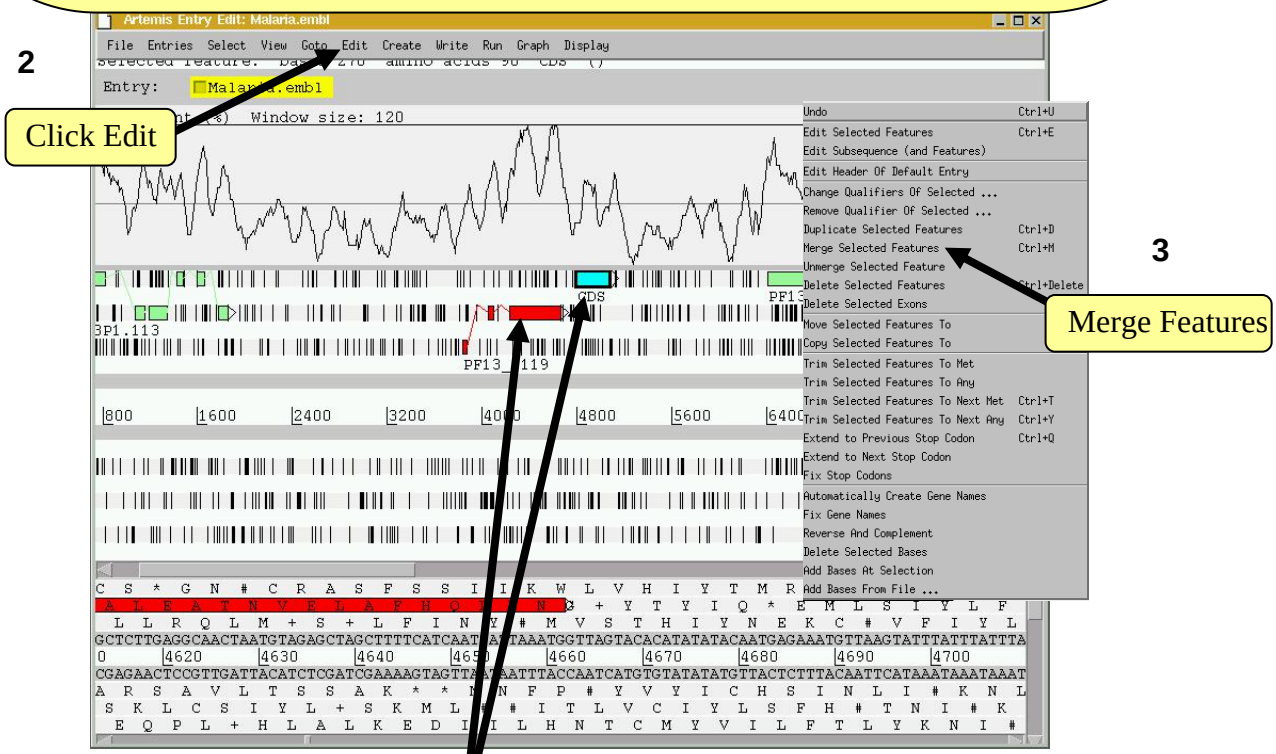
To compare the three CDS with others currently in the public databases run a fasta search. Left click the CDS, click on the 'Run' menu and then 'Run fasta on selected features'. When the search is finished, a banner will appear saying 'fasta process completed' (see above). The search may take a couple of minutes to run.

To view the search results click 'View' then 'Search Results' then 'fasta results'. The results will appear in a scrollable window. You could also view these results in your Netscape Browser window as in the previous exercise.

How does your predicted gene model for this CDS compare with proteins pulled out of the public databases? Is it possible that there are additional exons not featured in the current model.

If you think that there are additional exons that should have been included in the gene model you should add them to it. Using GC content and results from your database search as guides roughly draw in where you think the additional exon(s) lie:

To create additional exons:
Select the region you think represents the exon by holding down the left mouse button and dragging the curser over the region of interest. Then click the 'Create' menu and select 'Create feature from base range'. A new blue CDS feature will appear on the appropriate frame line (See below).

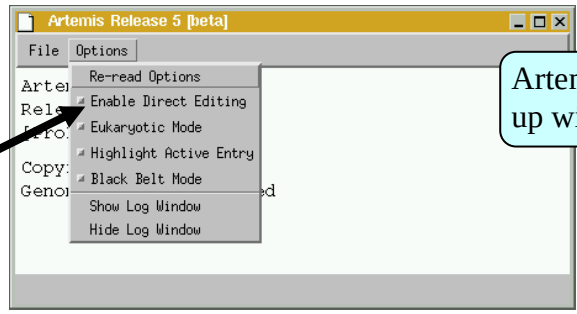


Select both the original gene-model and the new CDS feature, which is to be merged with it to form a new exon.

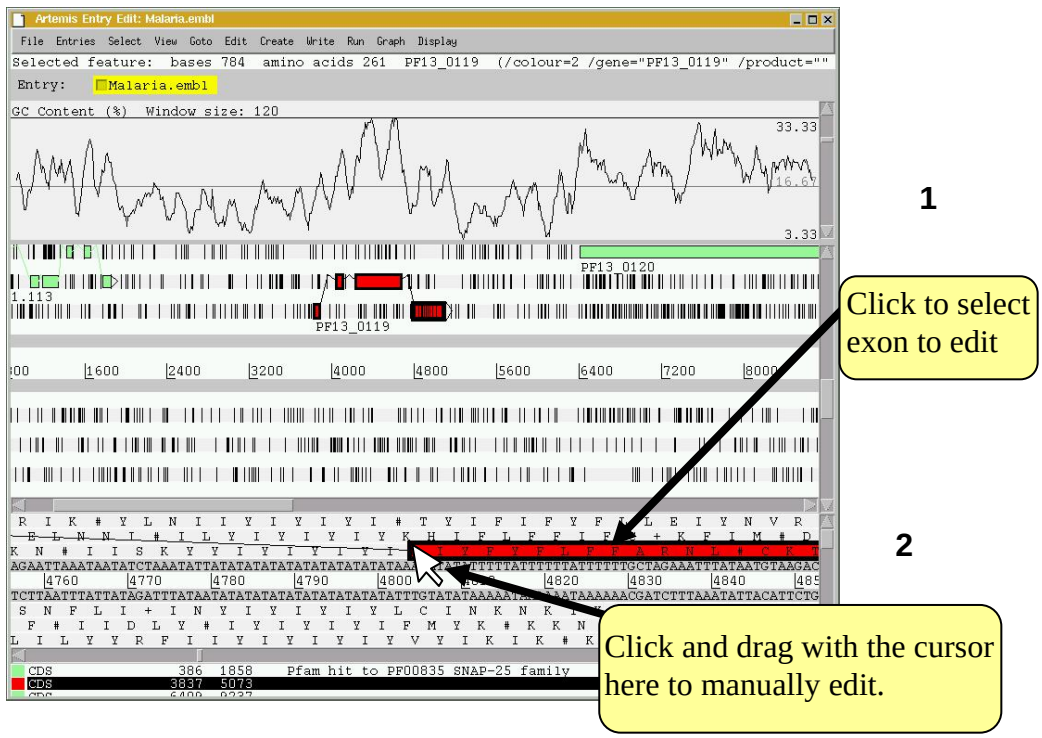
Tip, to select more than one feature (of any type) you must hold the shift key down.

The new CDS feature can then be merged with the original gene model as shown above.

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 that you started with plus the merged feature) will be retained.



You may noticed after you performed the merge function that one of the exons has subsequently jumped into another reading frame. Artemis automatically splices the CDS and so if the exon boundaries have an additional partial codon then any following exon will be pushed into another reading frame to account for this. To correct this you can edit the exon boundaries directly by turning on manual editing in the options menu of the Artemis start-up window, (as shown above). This will now allow you to edit the start and end positions of the feature boxes by using the left mouse button. Click and hold down the curser over the first or last base of any feature and then drag the mouse. The feature box should move as you drag it (see below. This can be a little tricky so please ask)



When manually editing your exons you should look out for appropriate splice donor and acceptor sites. See below for a small list and **Appendix IX** for details of known acceptor and donor motifs for Malaria splice sites.

Once you are happy with your newly created exon re-run the fasta search and see how this compares with the other hits in the public databases. If there are more exons to mark up try and complete the gene model.

The three example CDS to analyse were selected because they have very good database hits. This obviously makes the task of making the gene model far easier. However, several of the other CDS in this region have no significant database hits. If you have time you may want to have a look at these too.

Module 3
Artemis Advanced

Introduction

This Module builds on the Prokaryotic exercise we completed in Module 1. Like Module 1 you will be looking at the *Salmonella typhi* genome sequence. *Salmonella typhi* is the causative agent of Typhoid fever. It has been known for some time that *S. typhi* has evolved into a potent pathogen by acquiring large regions of DNA from other bacteria by a process called lateral gene-transfer. Many of these laterally acquired DNA regions encode genes that are important for virulence and consequently some of these regions have been called *Salmonella* pathogenicity islands.

Aims

The aim of this Module is extend your knowledge of Artemis. You will identify regions within the *Salmonella* genome that may have been acquired by lateral gene-transfer and then edit one of these regions as a subsequence and to save this information to a newly created file.

Artemis Exercise 1

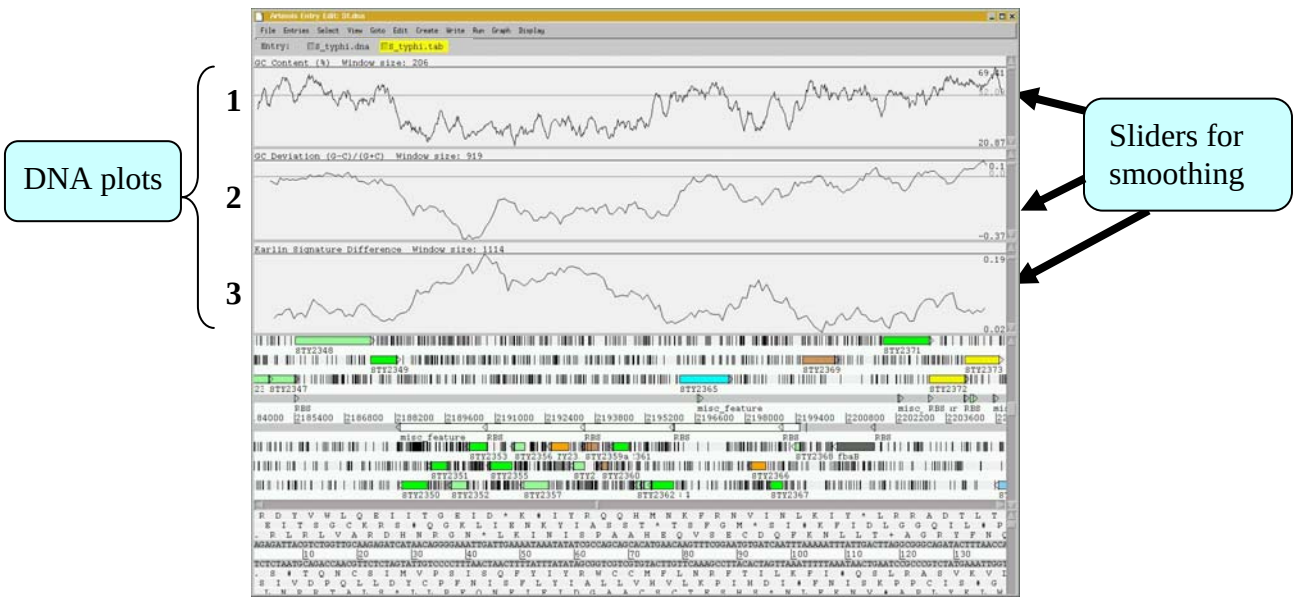
Follow the same procedures for starting Artemis as described in Module 1. All the files (S_typhi.dna and S_typhi.tab) you will need are contained in the directory: **Module_3_Whole_genome_analysis.**

By a method of your choice (i.e. use Navigator, Feature Selector or Goto) go to the region located between 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 either the Navigator, Feature Selector or Goto functions discussed previously to get there. The region you arrive at should look similar to that shown below.

In addition to looking at annotation for this region it is also possible to look at the characteristics of the DNA displayed. This can be done by adding in to the display various plots showing different characteristics of the DNA. This 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 many take a little time so be patient.

To view the graphs:

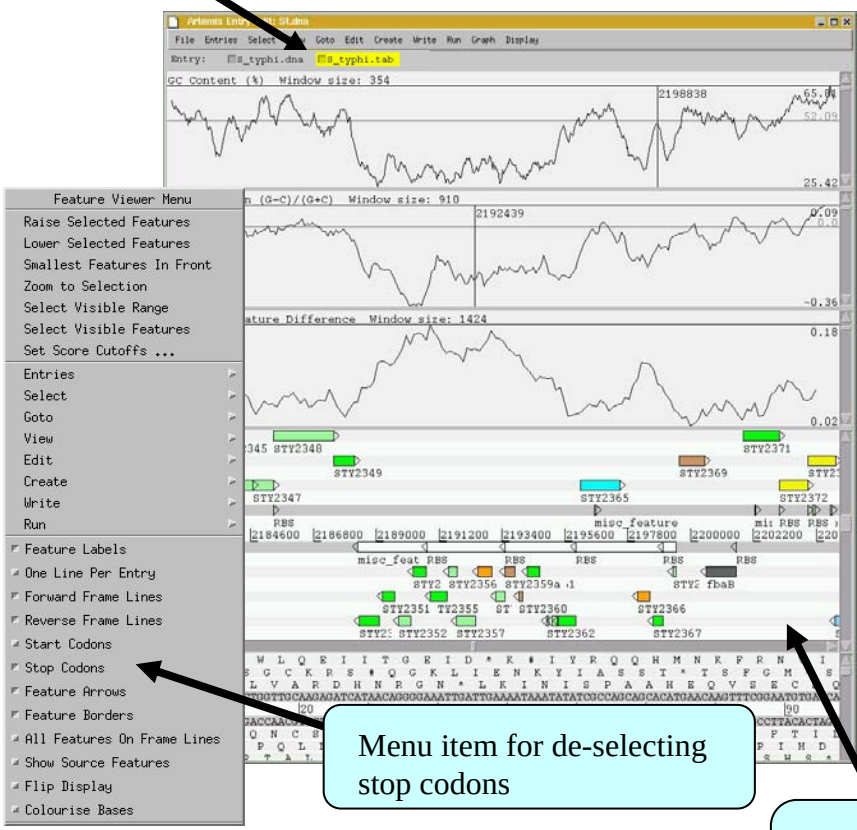
Click on the 'Graph' menu to see all those available. Perhaps some of the most useful plots are the 'GC Content (%)' (**1**) 'GC Deviation' (**2**) and 'Karlin signature plots' (**3**) as shown below. To adjust the smoothing of the graph you change the window size over which the points on the graph are 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.

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 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. 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). If you have any problems ask a demonstrator.

To de-select the annotation click here.

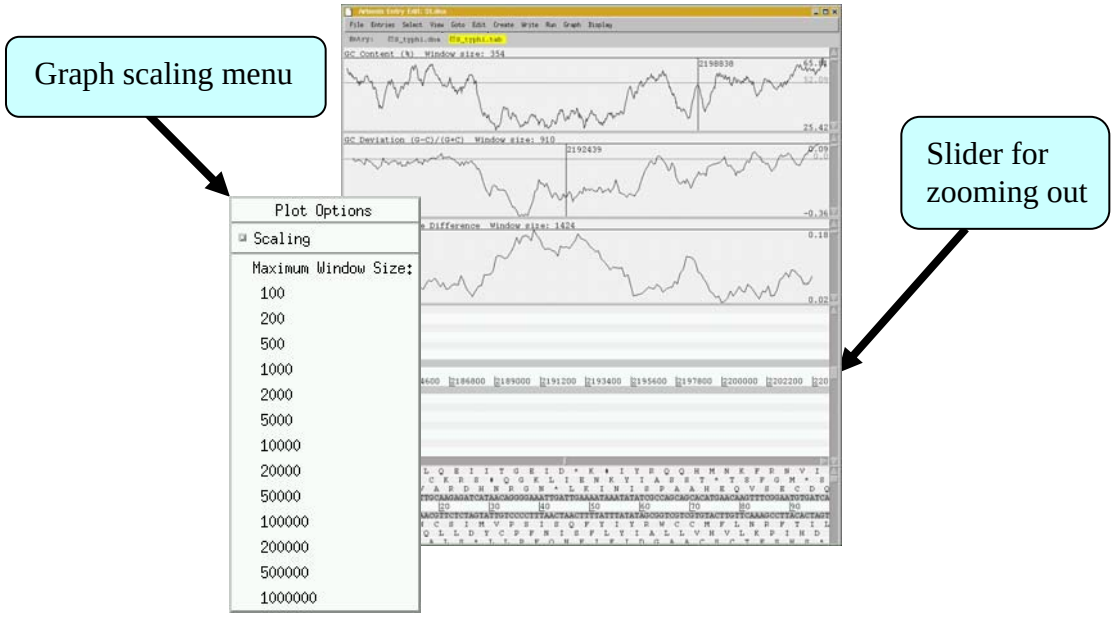


Menu item for de-selecting stop codons

No stop codons shown on frame lines

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.

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



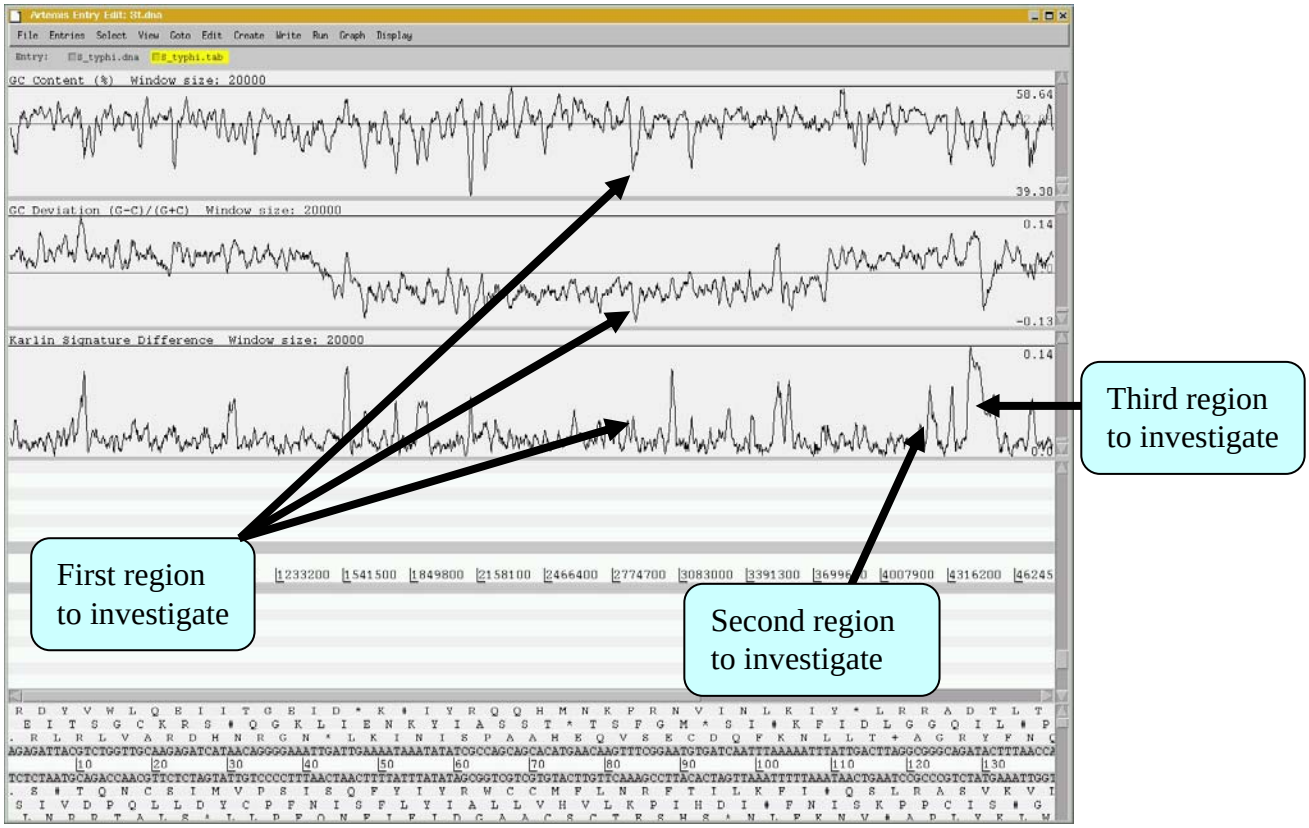
One final tip is to adjust the scaling for each graph displayed before zooming out. This increases the maximum window size over which a single point for each plot is calculated. To adjust the scaling click with the right mouse button over a particular graph window. A menu will appear with a series of values for the maximum window size (see above), select 20000. You should do this for each graph displayed.

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 should now be marked in the graph windows that you previously clicked in.

Artemis Exercise 1 Part III



There are many examples where these anomalous regions of DNA within a genome have been shown to carry laterally acquired DNA. In this part of the exercise we are going to look at several of these regions in more detail. Starting with the whole genome view, note down the approximate positions and characteristics of the three regions shown above. Remember the locations of the peaks are given in the graph window if you click with the left mouse button within it.

	Genome location	Characteristics of DNA plots
Region 1 :	*2,860,000 bps*	*peak - karlin, troughs for G+C and CG deviation*
Region 2 :		
Region 3 :		

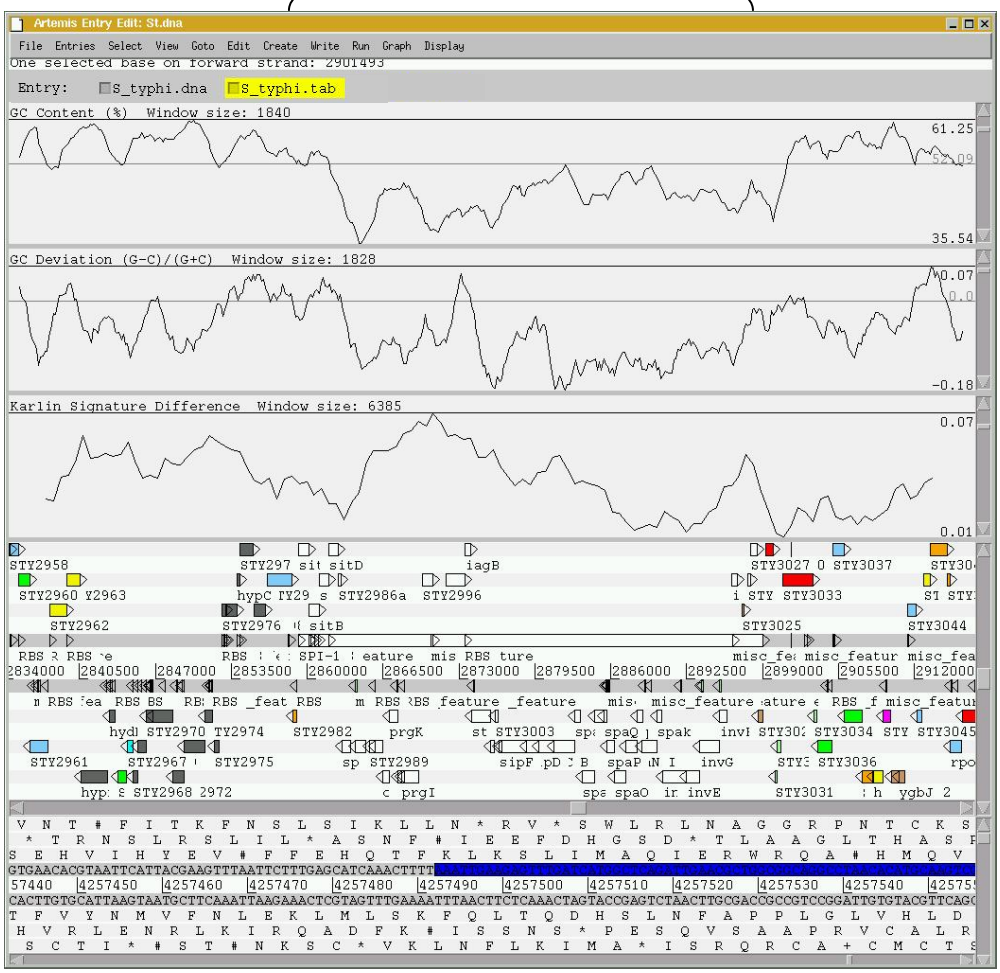
We will now zoom back into the genome to look in more detail at the first of these three peaks. Zoom into this position by first clicking on the DNA line at approximately the correct location. If you then use the vertical side slider to zoom back in, Artemis will go to the location you selected. Remember that in order to see the CDS features lying within this region you will need to turn the annotation (*S*_typhi.tab) entry back on.

The region you should be looking at is shown below and is a classical example of what is referred to as a *Salmonella* pathogenicity island (SPI). The definitions of what actually 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.

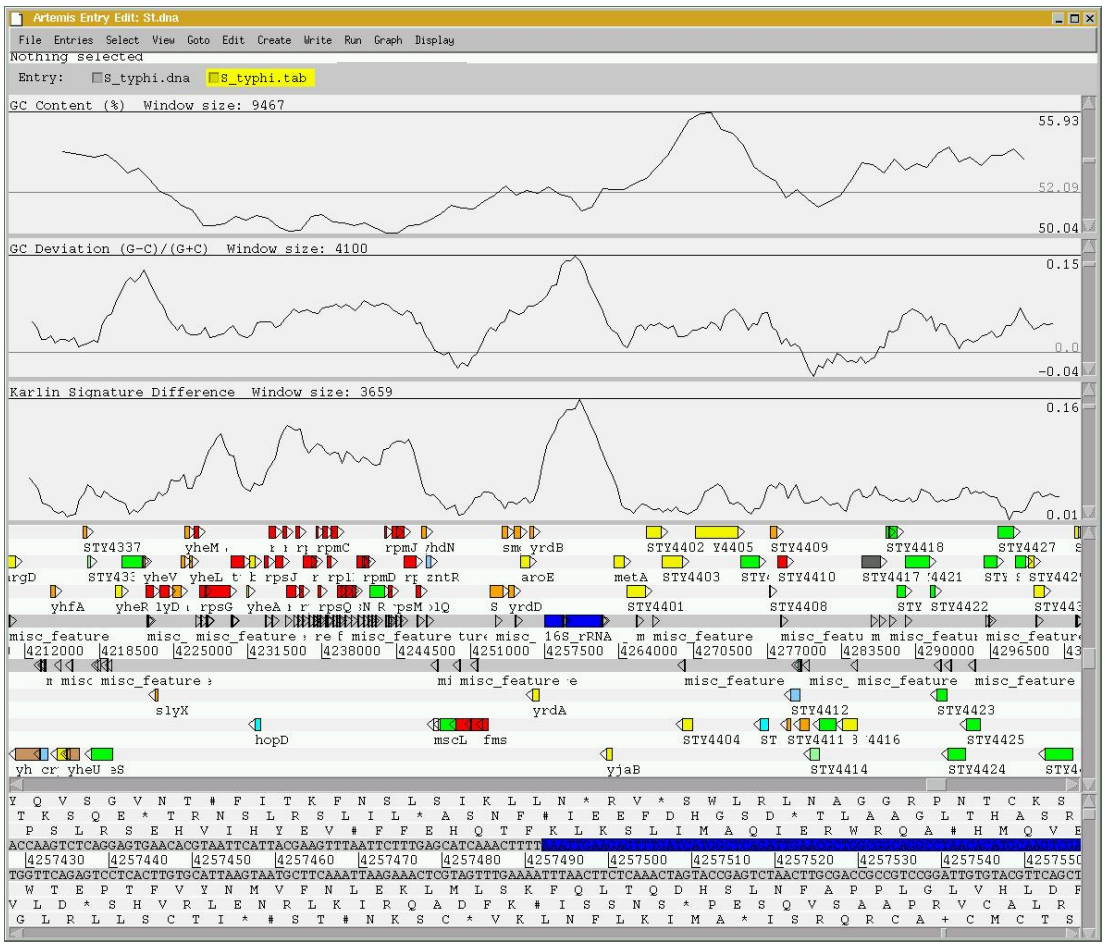
1. Often inserted alongside stable RNA's
2. Atypical G+C contents.
3. Carry virulence-related functions
4. Often carry genes encoding transposase or integrase-like proteins
5. Unstable and self-mobilisable
6. Of limited phylogenetic distribution

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

Region 1 SPI-1



Region 2



Use one of the methods you have already used to take you to the second region of interest that you noted down.

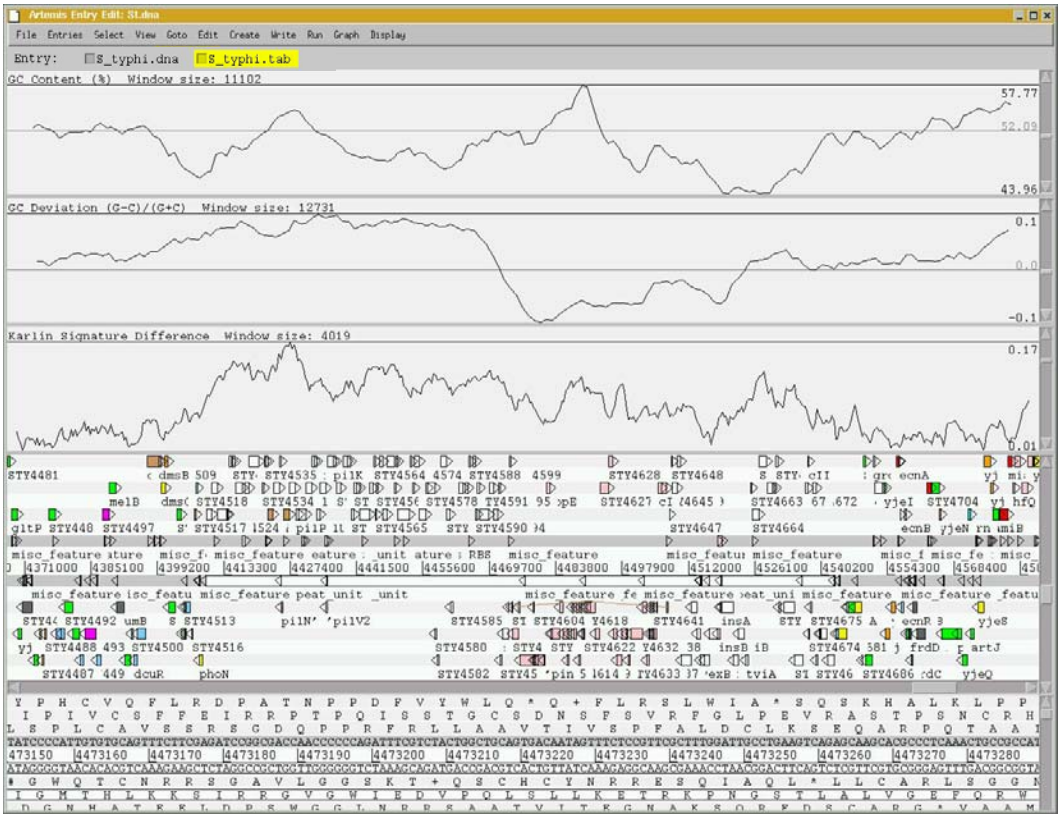
Region two acts as a cautionary note when looking at anomalous regions within a genome. Have a look at the CDSs within this region.

Does this region:

- have any of the characteristics of pathogenicity island
- are the genes within this region essential or dispensable.

Is it possible that the atypical base composition of this region is not a consequence of having originated from a foreign host. The base composition may actually be reflective of the tight sequence constraints under which this region has been maintained, in contrast to the background level sequence variation in the rest of the genome.

Region 3



Go to region 3 as before.

Like region 1, this region is also referred to 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. Have a look at the CDSs within this region. As before notice any stable RNAs that may have acted as the phage integration site.

Artemis Exercise 1 Part IV

Continuing on from the analysis of Region 3 or SPI-7 (the major Vi-antigen pathogenicity island) 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.

2 Click 'Edit'

3 Click 'Edit subsequence (and features). '

1 Select region by clicking with the left mouse button & dragging

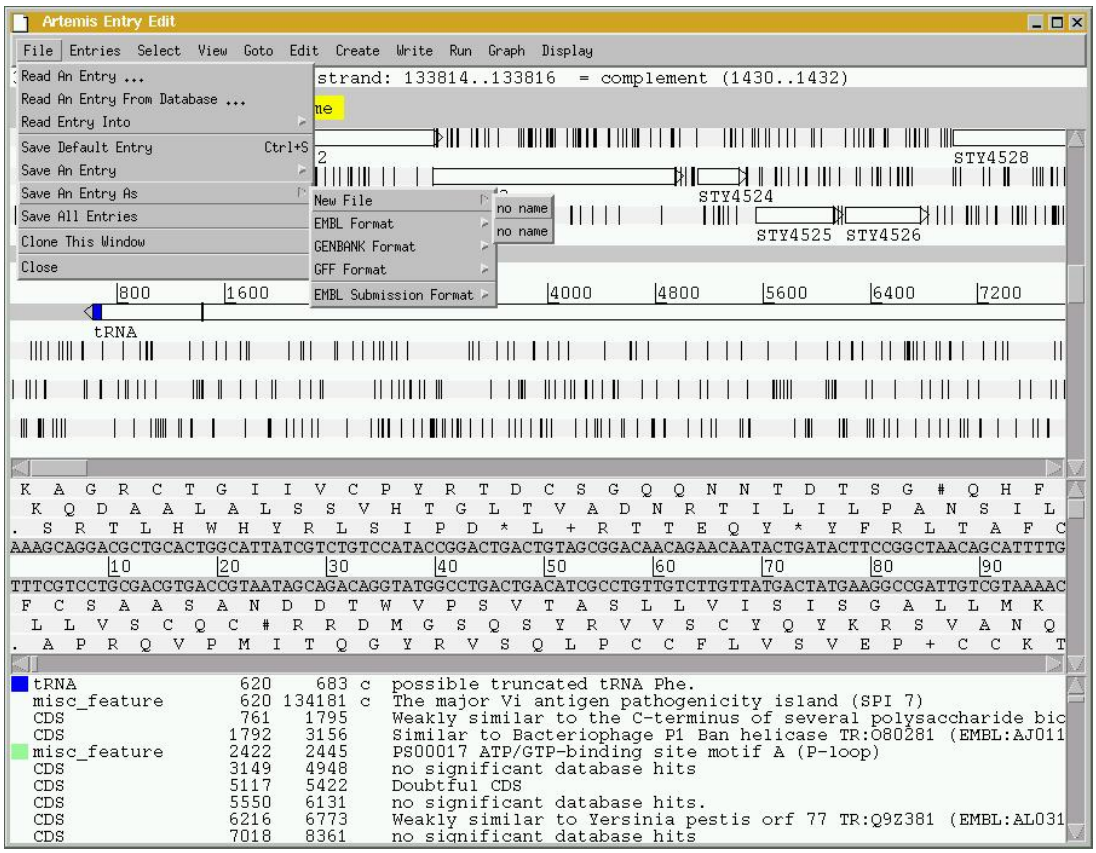
Note the entry names have changed

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

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

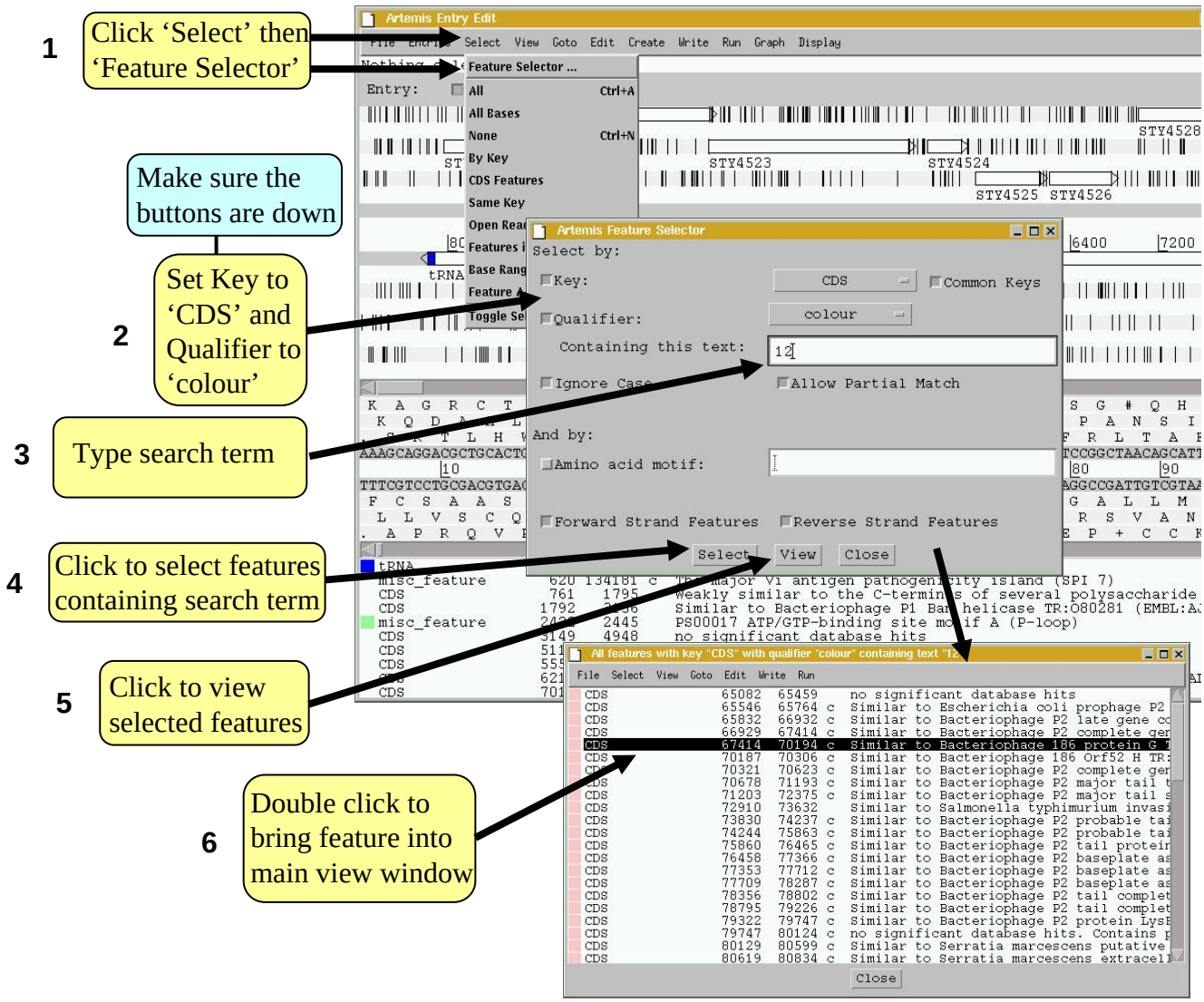
Note that the two entries on the grey Entry line are now denoted ‘no name’, they represent the same information in the same order as the original Artemis window but simply have no assigned name. Because the sub-sequence is now viewed in a new Artemis session, this prevents the original files from being over written (i.e. S_typhi.dna and S_typhi.tab). We will now save them as new files 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 other 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). We are going to use this information to select all the relevant phage genes using the Feature selector as shown below and then to 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.

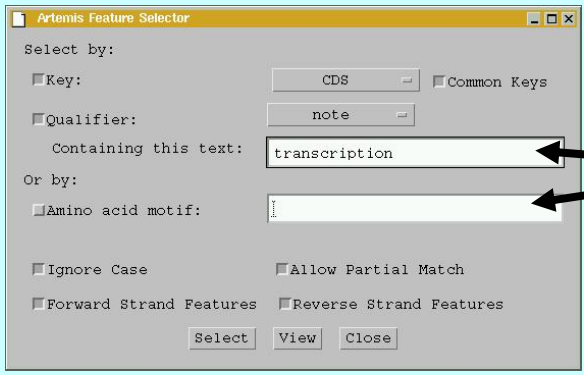


The genes listed in **6** are only those fitting your selection criterion. They can be copied or moved in to a new entry so we can view them in isolation from the rest of the information within spi7.tab.

Firstly in window **6** select all of the CDS 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 'move 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 you did before. 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 of selecting/extracting features using the Feature Selector

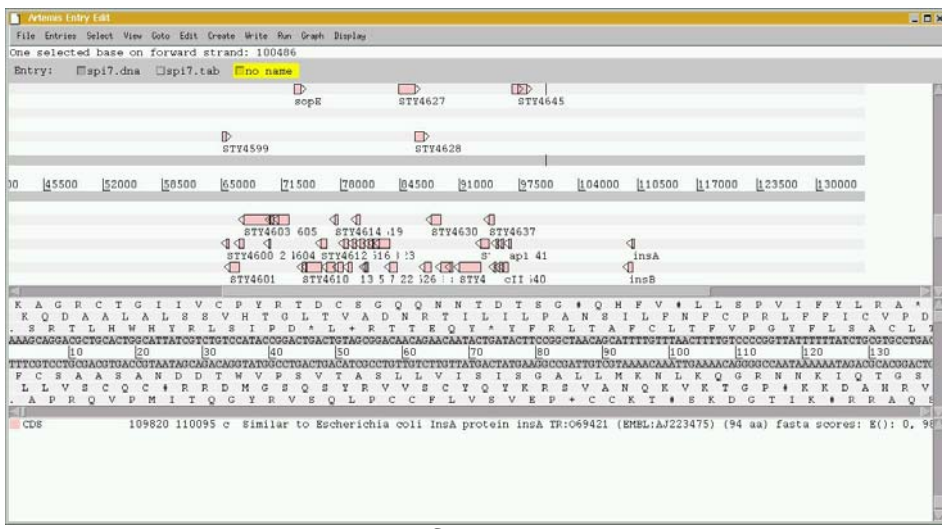
It is worth noting that the feature selector can be used in many other ways to select and extract subsets of features from the genome. If you have a closer look at the Feature selector you will also see that you can use search terms to select a class or all those features with a particular amino acid motif.



Space for a search term or amino acid motif

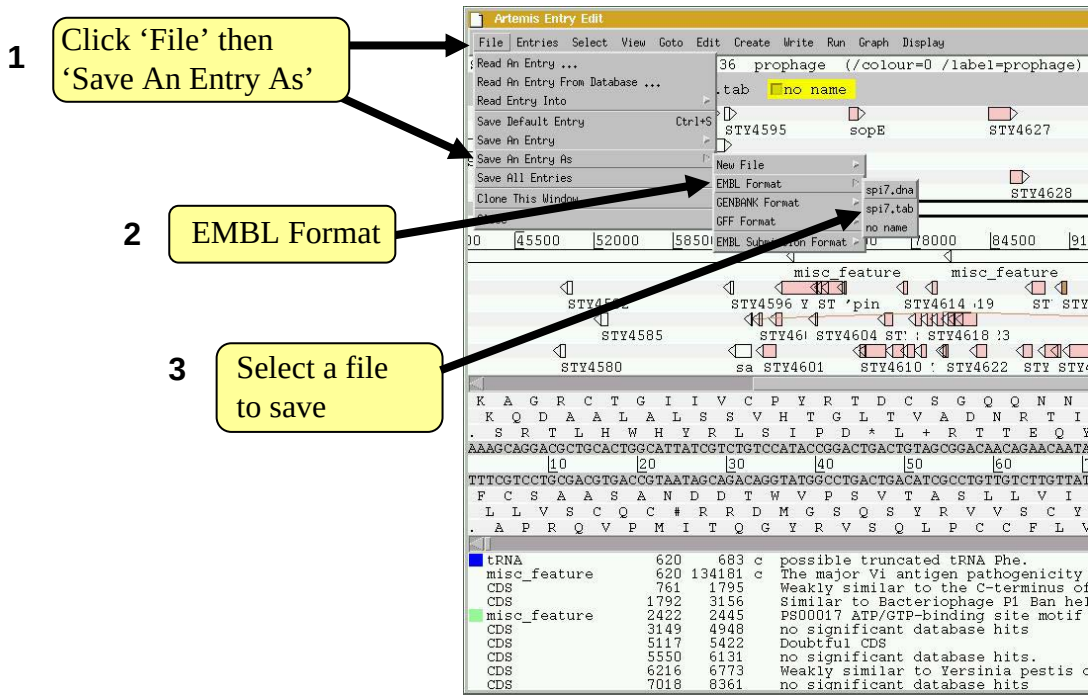
Defining the extent of the prophage.

Even from this very cursory analysis it is clear from the selection that the prophage occupies a fairly discrete region within SPI-7 (see below). It is often useful to create a 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. 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 it to 'misc_feature' (an option in the key menu top left hand corner at the edit window) Artemis will place this feature on the DNA line. This is perhaps more appropriate and is easier to visualise. If you also add in a qualifier, such as ' /label ' and add text following the /label= ????, 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 turn back on the spi7.tab and have a look at the genes located at either side of your selection. Go to and look at the CDS *samA*. In reality this gene was disrupted by the insertion of this bacteriophage. If you look at the FASTA results for this CDS you may be able to track the bases between which this phage inserted.

Your final task is to write out these files in EMBL format and create a merged annotation and sequence file in EMBL format:



This will create two files one with the sequence and the other with the annotation in the directory within which you started Artemis. To create a complete EMBL file use the UNIX you covered earlier and 'cat' the files together

Module 4
Gene Prediction

Introduction

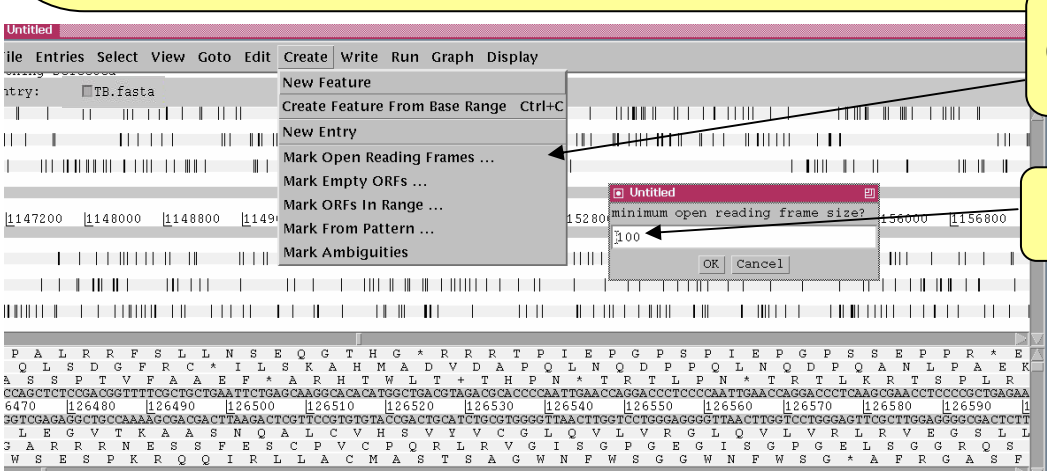
There are many automated gene prediction programs commonly used for both Manual and Automated annotation protocols. Most of these programs use different algorithms, data sets and criteria for gene-calling. Consequently, if you ran all of these different gene prediction programs on the same piece of DNA they would all come up with different solutions (sometimes markedly different) describing the coding capacity of that section of DNA. The importance of this should not be underestimated when you consider that many of these automatically assigned genes may find their way into the public databases and subsequently influence experimental design.

Aims

The aim of this module is to compare the results generated by several gene prediction programs. We will also use several other metrics with which to validate the output of these programs and finally generate a gene model for a given region of DNA. We will cover both Prokaryotic and Eukaryotic worked examples.

Gene Identification

Exercise 1
Finding the open reading frames
This exercise is designed to introduce the different methods used to identify genes in genomic sequence
To start open the file TB.fasta (*Mycobacterium tuberculosis* genome) in Artemis.
A quick method to identify all possible genes is to identify all possible open reading frames. To do this select 'Create' and then 'Mark open reading frames' (see below)
You can choose a minimum size of open reading frame that you want to create. Try typing 100. Notice that a new entry will appear on the entry line called ORFS 100+. This will contain all the ORFS you have just created. Turn it on and off to check.
Then go to 'Select' and 'All'. Then 'Edit' and 'Trim Selected Features to Any' this will give you all the possible open reading frame with a bacterial start codon.



Go to Mark Open Reading Frames **1**

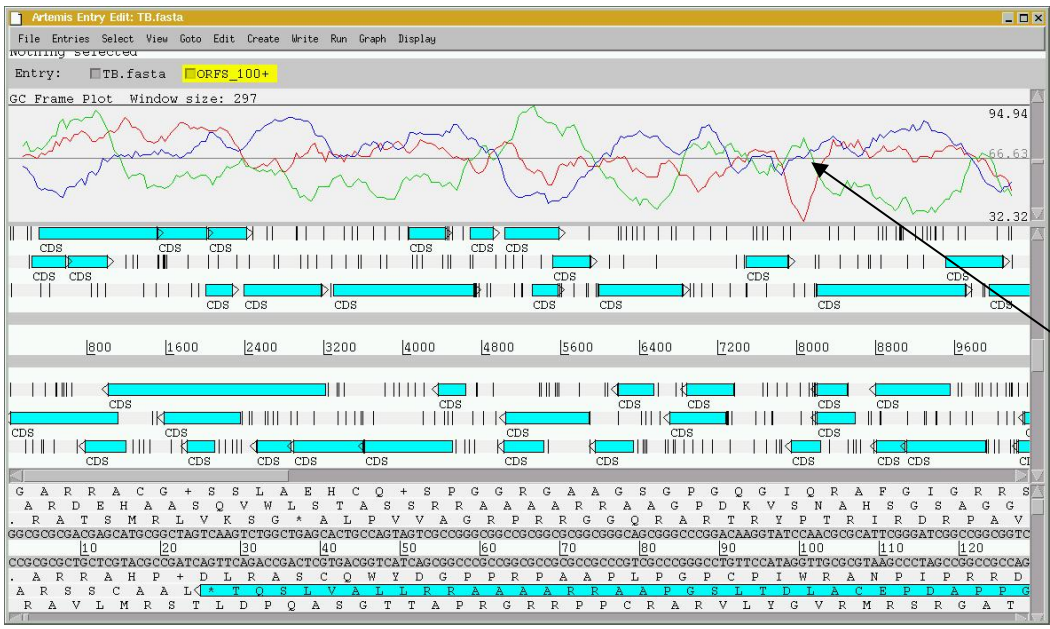
Type 100 **2**

Delete features that can't be trimmed to a start codon. These will be selected.

Trimmed ORFs

Exercise 2
Using DNA content plots

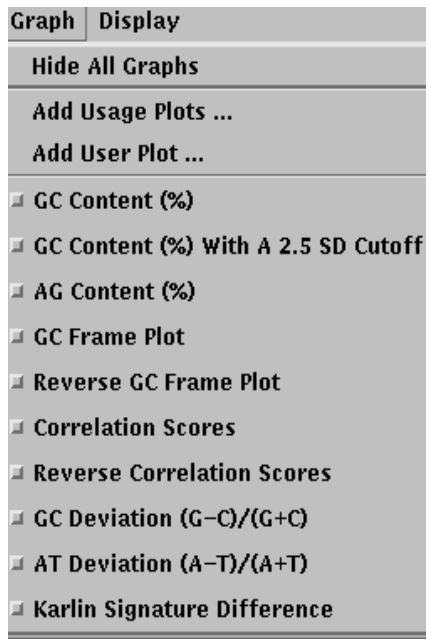
The G+C content in each reading frame of a piece of DNA is often different in coding regions compared to non-coding regions due to the limitations imposed by codon usage. You can visualise the codon usage for a DNA sequence in Artemis.
Go to ‘Graph’ and ‘GC frame plot’.
Compare this to the overall GC content by going to ‘Graph’ and ‘Choosing G+C content’.
Try to optimise the graphs by moving the slide bar on the right.



Move slide bar to change window size

GC frame plot

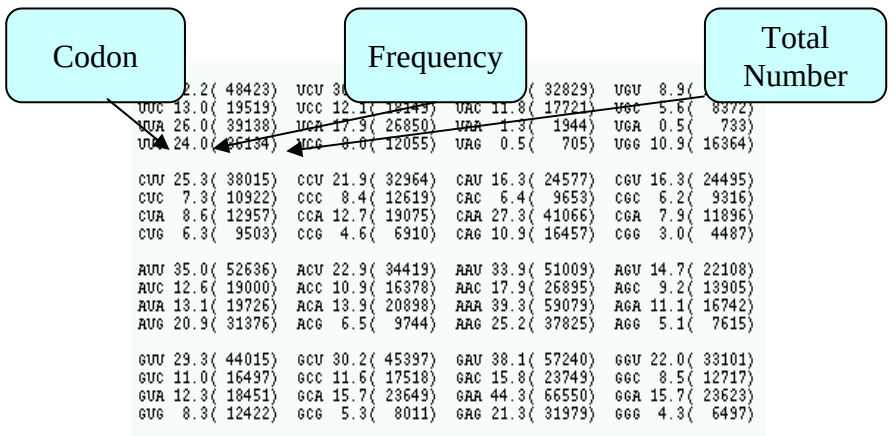
If you go to the ‘Graph’ menu you will see that Artemis can display many different graphs of different DNA properties. Try as many as you can and decide which ones may be useful for predicting genes. A description of what each of these can be found at www.sanger.ac.uk/software/artemis/v4/manual.



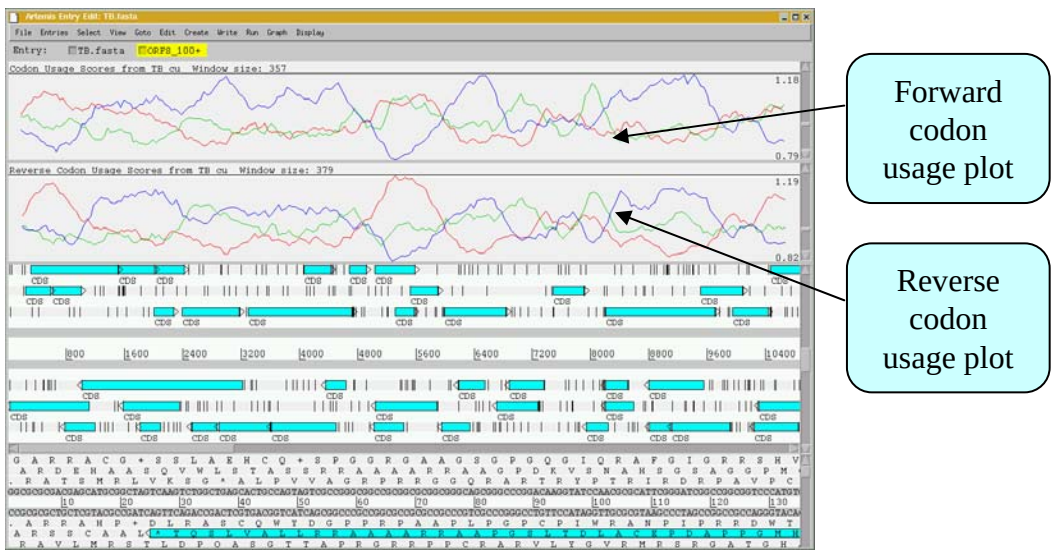
Codon usage

Most organisms have a preferred set of codons for protein coding DNA. These are represented by codon usage tables (see below) and can be used to predict what regions of a genome are coding using a codon usage plot. To do this:

- go to 'Graph' and 'Add usage plots'
- Choose the file TB_cu in the current directory
- Two graphs will appear (see below) use the vertical slider to scroll the graphs
- Decide which CDSs agree with the codon usage and delete those that definately do not.



UUU	2.2(	48423)	UCU 3		(	32829)	UGU 8.9(
UUC	13.0(	19519)	UCC 12.1(	18145)	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)



Introduction

Automated Gene Prediction

Gene finding software (Glimmer and Orpheus) that has been trained for *M. tuberculosis* has been pre run for you. To see the gene finding predictions:

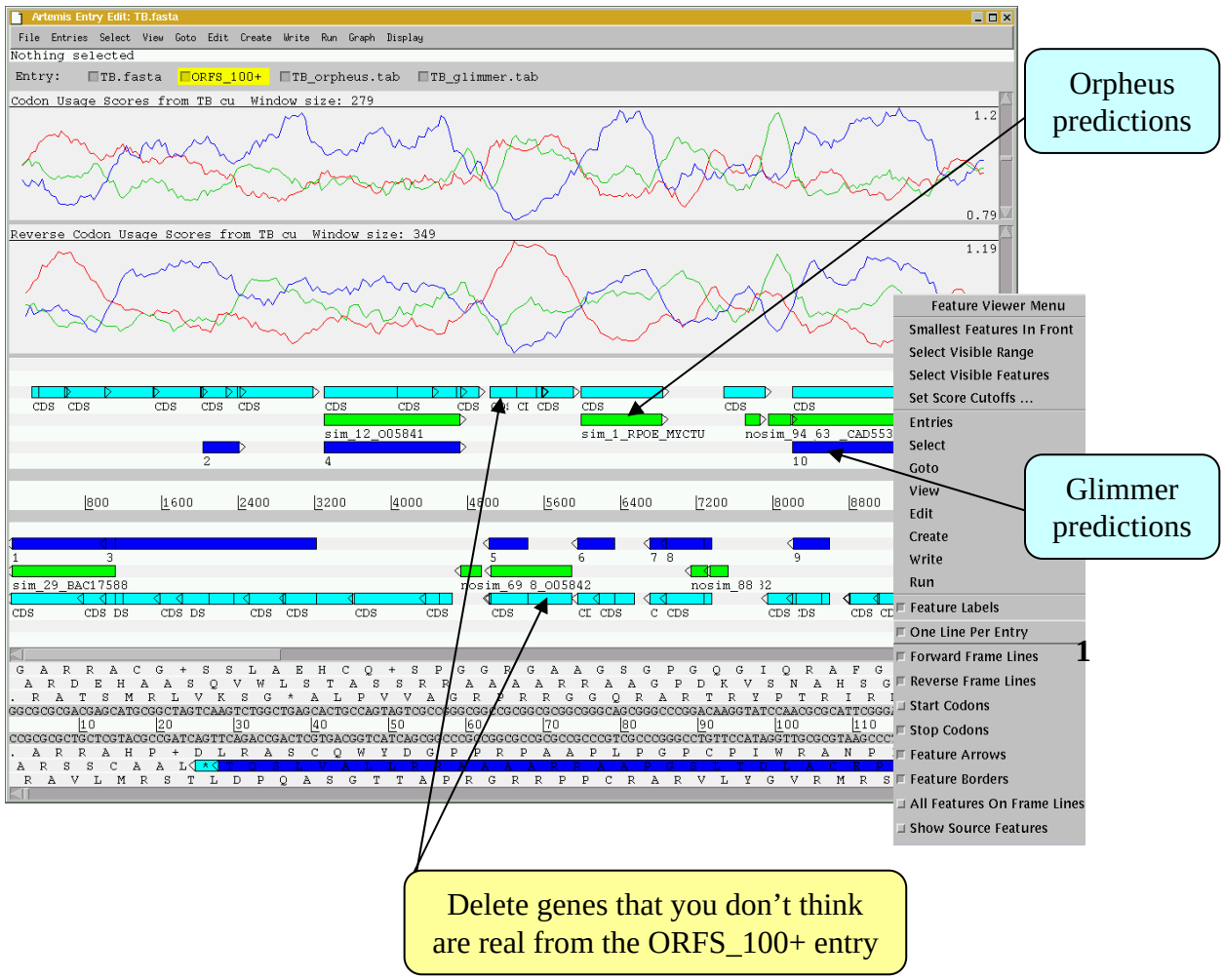
•Go to the 'File' menu and choose 'Read an entry'

•Read in both files called TB_orpheus.tab and TB_glimmer.tab.

•You can show all of the evidence (ORFS +100, orpheus and glimmer) on separate lines by right clicking on the frame lines and selecting 'One line per entry' from the menu that appears (see below; **1**)

•Compare the different predictions and using the plot information. Remove any genes that you think are not ' real ' from the ORFS_100+ entry you created earlier. Left click on them and press delete.

•You can be conservative at this stage.



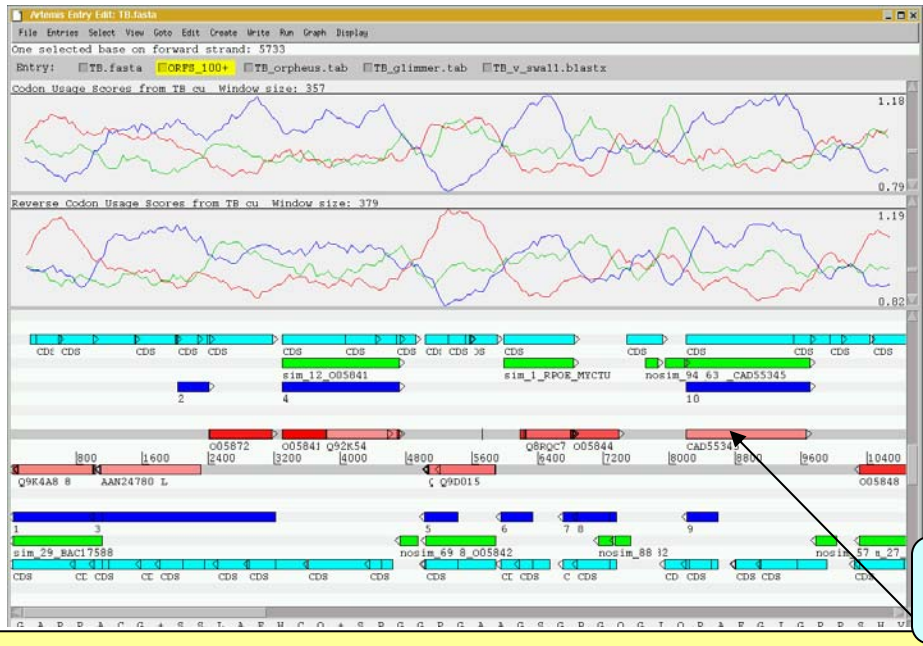
Exercise 3
Using homology data

The tools you have used so far have only looked at sequence properties to predict if a region is coding. However you may also use homology evidence to identify if it is coding. Evidence from Blast searches can be read into Artemis for this purpose. To do this:

•Go to 'Read an entry' in the 'File' menu and select TB_v_swall.blastx. This file contains BLASTX hits from the *M.tuberculosis* DNA sequence searched against the SWALL non-redundant protein database.

•From this evidence you will be able to remove more genes that are incorrect from your ORFS_100+ file.

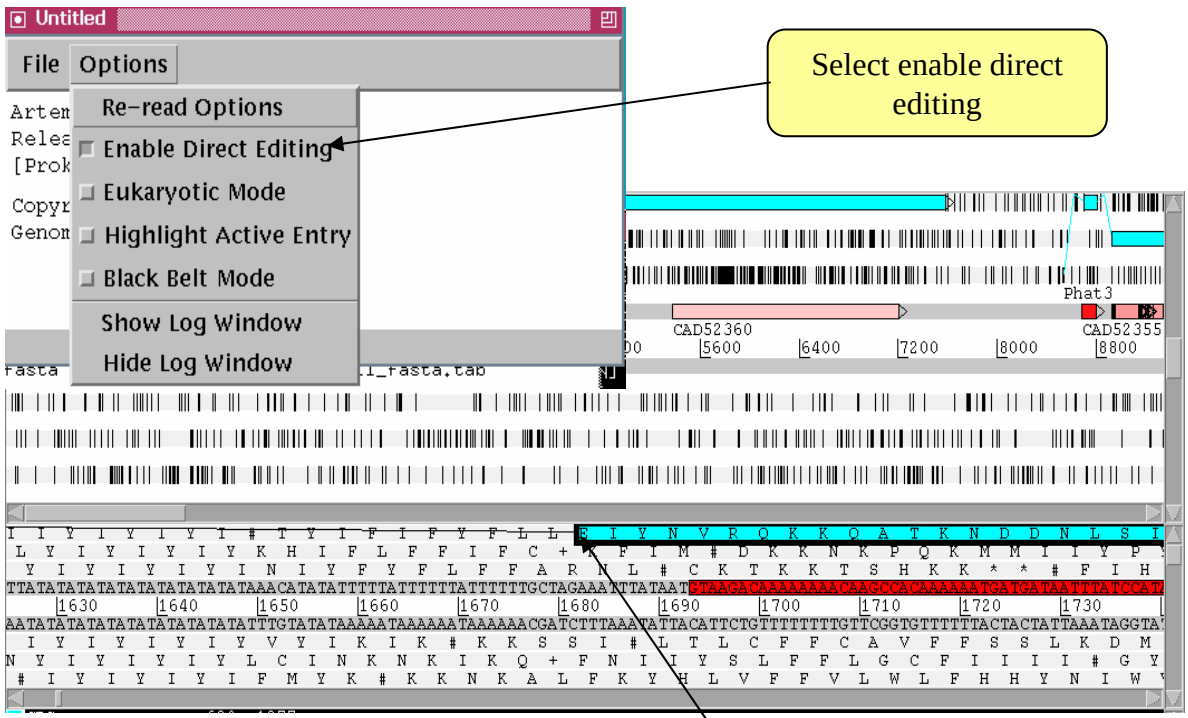
•At this point you can run FASTA searches of the remaining ORF sequences using the 'Run' menu. Use this evidence to help you predict which genes are real and remove any others. Also remember that bacterial CDSs rarely overlap by more than 3-5 codons



Check your predictions against the Sanger annotations by reading the entry TB.tab.

Gene-prediction for *M. tuberculosis* was a relatively simple, although time consuming, task. Once you have predicted several CDS for this bacterium, repeat the same steps for *M. leprae*. All the files that you will need are in the current directory and named using the same conventions as the *M. tuberculosis* files e.g. LEPRAE.fasta and LEPRAE_glimmer.tab etc. The exception is the BlastX file (LEPRAE_v_TB.blastx) which is the results of a search of the *M. leprae* proteins against those of *M. tuberculosis*. The reason for this is that the *M. leprae* genome has undergone reductive evolution leaving many pseudogenes and gene fragments that remain intact in the closely related and larger *M. tuberculosis* genome sequence. Many of these can be seen using the BlastX comparison data. The region you will look at is equivalent to that you have just been looking at in *M. tuberculosis*. Note gene-prediction in *M. leprae* is very difficult.

Exercise 4
Gene finding for spliced genes
In many Eukaryotic organisms the principles covered in the earlier exercises still hold, however, some genes may contain introns hence gene identification becomes more complicated. For the next exercises you will need to close the previous Artemis session
•Start Artemis and load the sequence file Pfal_subseq.embl
•Load the Phat gene predictions pfal_subseq_phat.tab
•Find which sequence plot would be most useful for this organism (*Plasmodium falciparum*).
•Load Blastx file swall_blastx.crunch.
•Using the Fasta searches and information you have loaded edit the gene models to fit the evidence you have



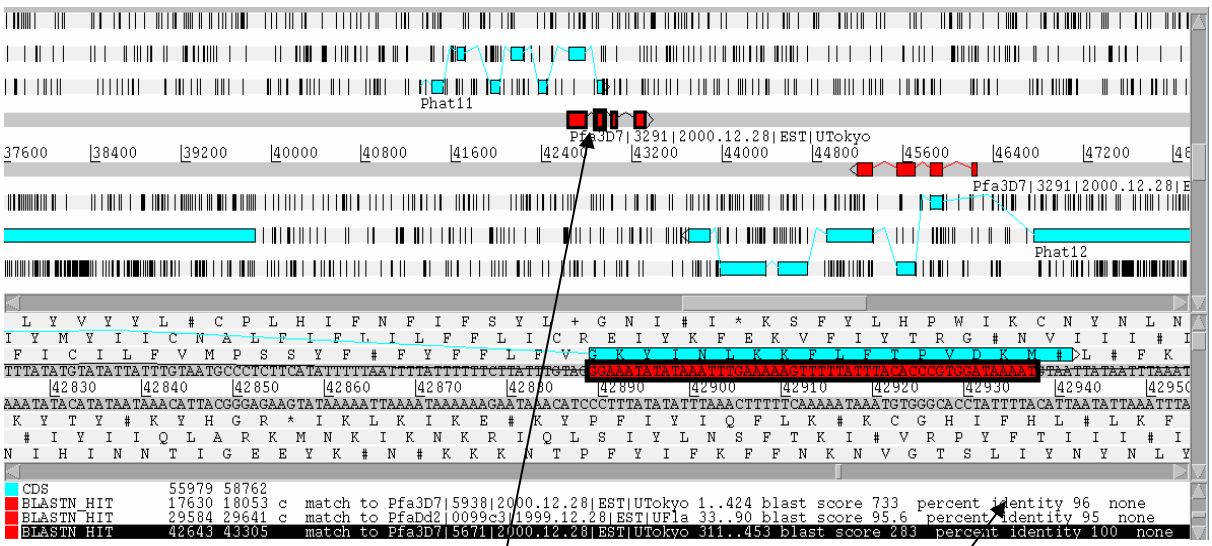
Select enable direct editing

Click and drag the last/first base of exon

To Change a gene model
•In the startup window make sure you have 'Enable direct editing' selected.
•In the DNA window click the base at the beginning or end of an exon and drag it to where you want it to go.
•If you want to merge two exons, select them and, go to the 'Edit' menu and select 'Merge selected features'
•You may want to refer back to **Artemis Module 2, Exercise 2, Part I** for help.

ESTs

ESTs (Expressed Sequence Tags) are sequences of cDNA derived from mature transcripts, hence they give useful information about splice site boundaries. Remember that they will contain UTRs and hence will not help find start and stop sites.
You can see the *Plasmodium* ESTs by reading in the entry EST.blastn.tab. This compares BlastN results of the *Plasmodium* sequence against a DNA database of all *Plasmodium* EST sequences. View *BlastN* hits from Plasmodium ESTs by reading the entry EST_blastn.tab. Try and use this information to help refine your gene models. Remember, ESTs are clear evidence that this region is transcribed and is useful for finding missing exons.



EST Blast Hit

Check quality of hit here

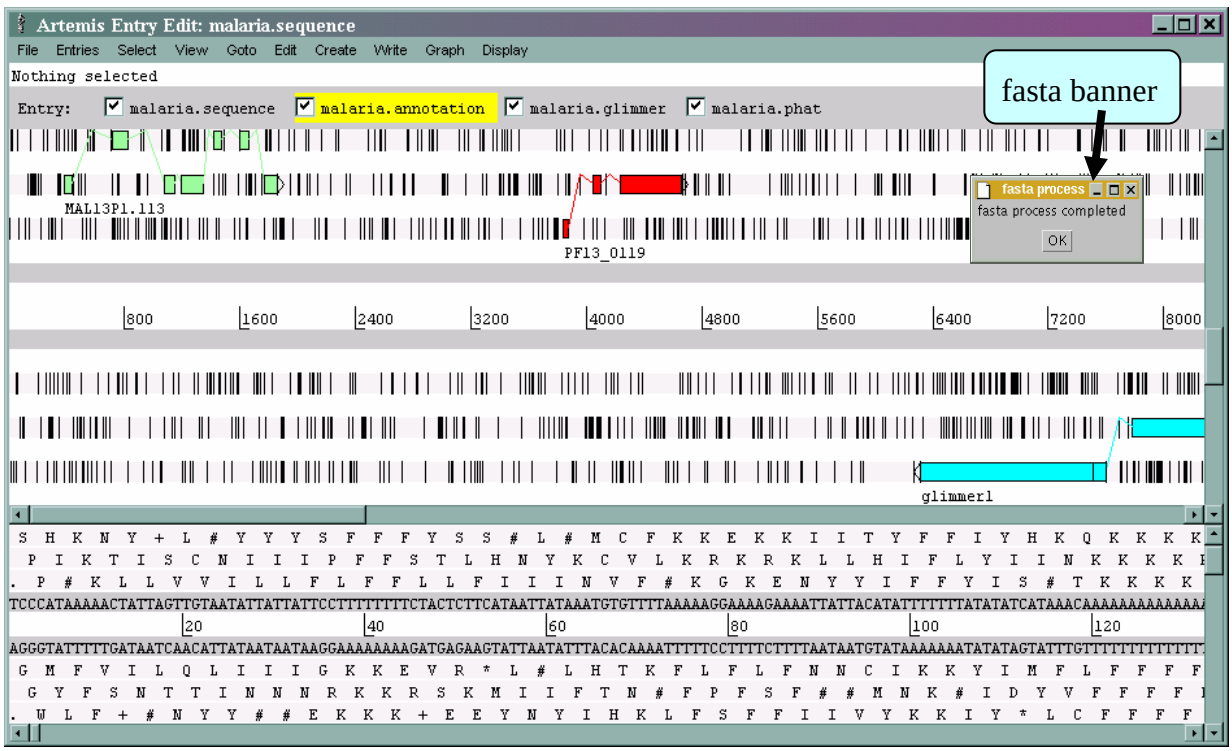
Once you have finished you may check how your gene prediction of this region compares to the final Sanger annotation by reading in the file Pfal_subseq.tab

Exercise 5
Gene finding for spliced genes (malaria)

You will need to start a new Artemis Session, as before, using the files malaria.sequence, malaria.annotation, malaria.glimmer and malaria.phat in the current directory.

The sequence you are going to look at is a small region of contrived sequence (~24 kb) taken from *Plasmodium falciparum*. The file malaria.annotation contains two annotated CDSs with multiple exons. See if you agree with them – one has only been partially characterised and may in fact be missing exons. You will also see predicted CDSs from the algorithms Glimmer and Phat. Make you own gene models based on the predictions in the tab file called malaria.annotation. Use the strong G+C bias of malaria to guide your decisions

G+C content is a very good indicator of coding capacity in Malaria. On average, the coding regions are ~23% G+C and the non-coding regions are ~19%. Have a look at the G+C content for this region by selecting the appropriate graph. Left click within the graph window and then select by clicking on the exons to see how this relates to the G+C peaks on the graph.



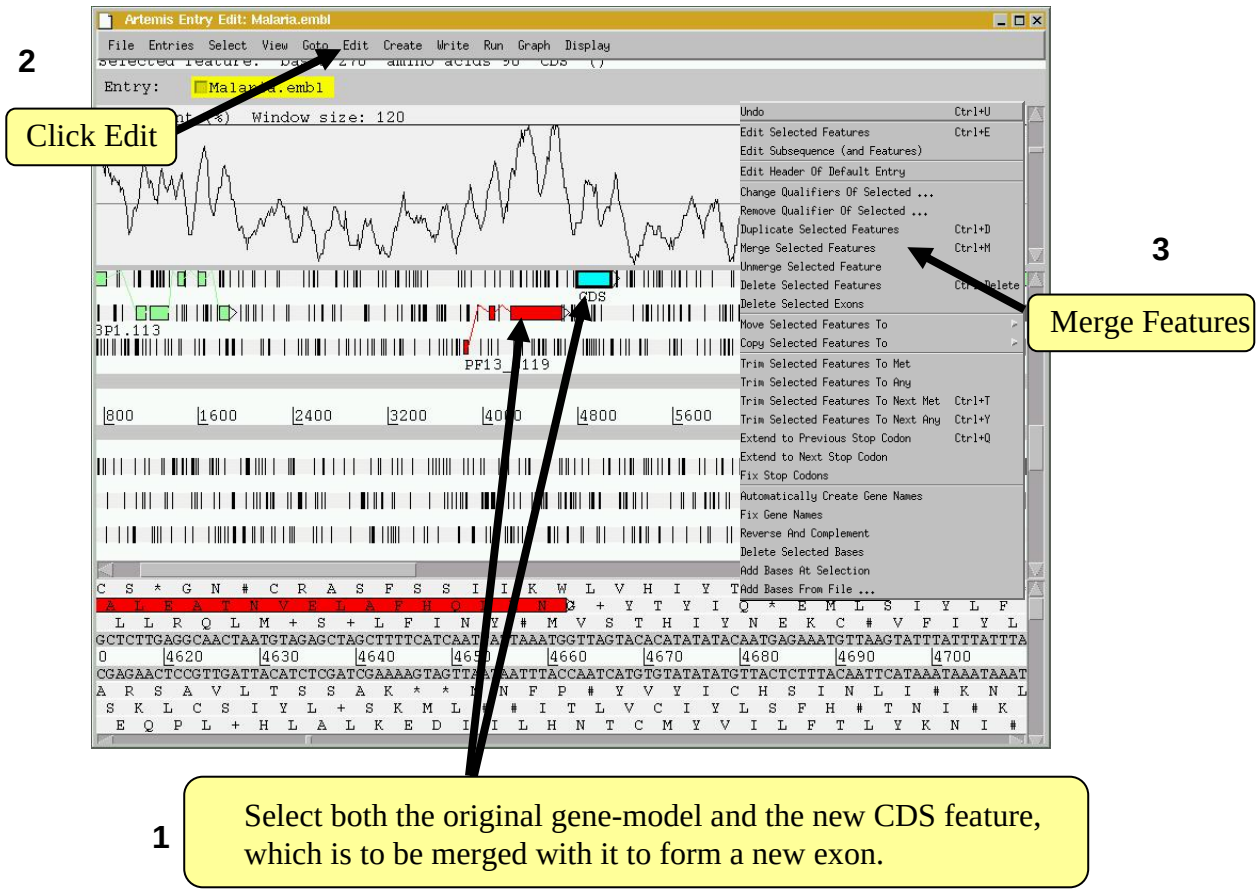
To compare CDSs with others currently in the public databases run a fasta search. Left click the CDS, click on the 'Run' menu and then 'Run fasta on selected features'. When the search is finished, a banner will appear saying 'fasta process completed' (see above). The search may take a couple of minutes to run.

To view the search results click 'View' then 'Search Results' then 'fasta results'. The results will appear in a scrollable window. You could also view these results in your Netscape Browser window as in the previous exercise.

How does your predicted gene model for this CDS compare with proteins pulled out of the public databases? Is it possible that there are additional exons not featured in the current model.

If you think that there are additional exons that should have been included in the gene model you should add them to it. Using G+C content and results from your database search as guides roughly draw in where you think the additional exon(s) lie:

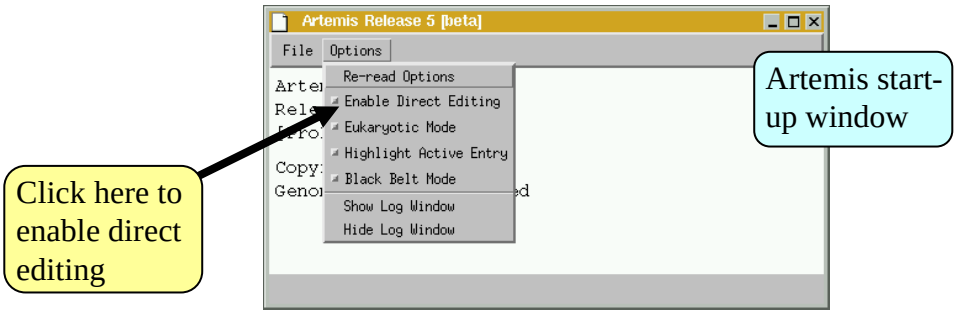
To create additional exons:
Select the region you think represents the exon by holding down the left mouse button and dragging the curser over the region of interest. Then click the 'Create' menu and select 'Create feature from base range'. A new blue CDS feature will appear on the appropriate frame line (See below).



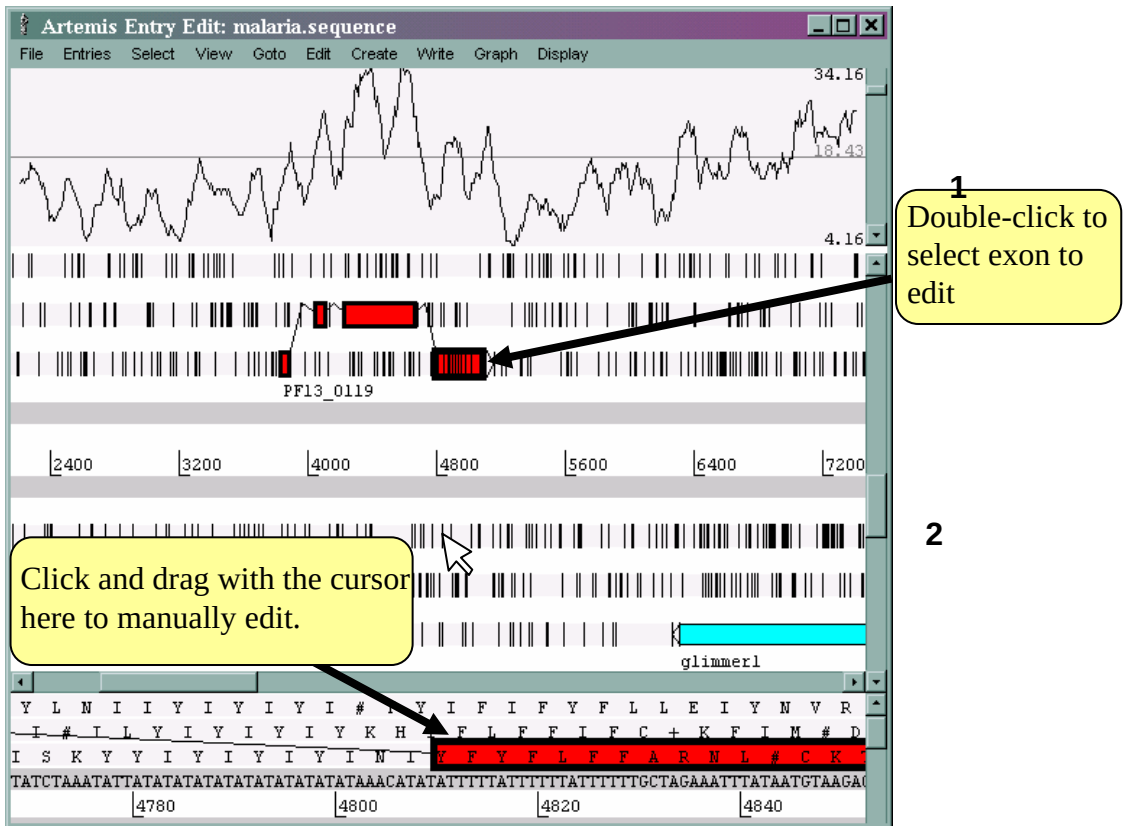
Tip, to select more than one feature (of any type) you must hold the shift key down.

The new CDS feature can then be merged with the original gene model as shown above.

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 that you started with plus the merged feature) will be retained.



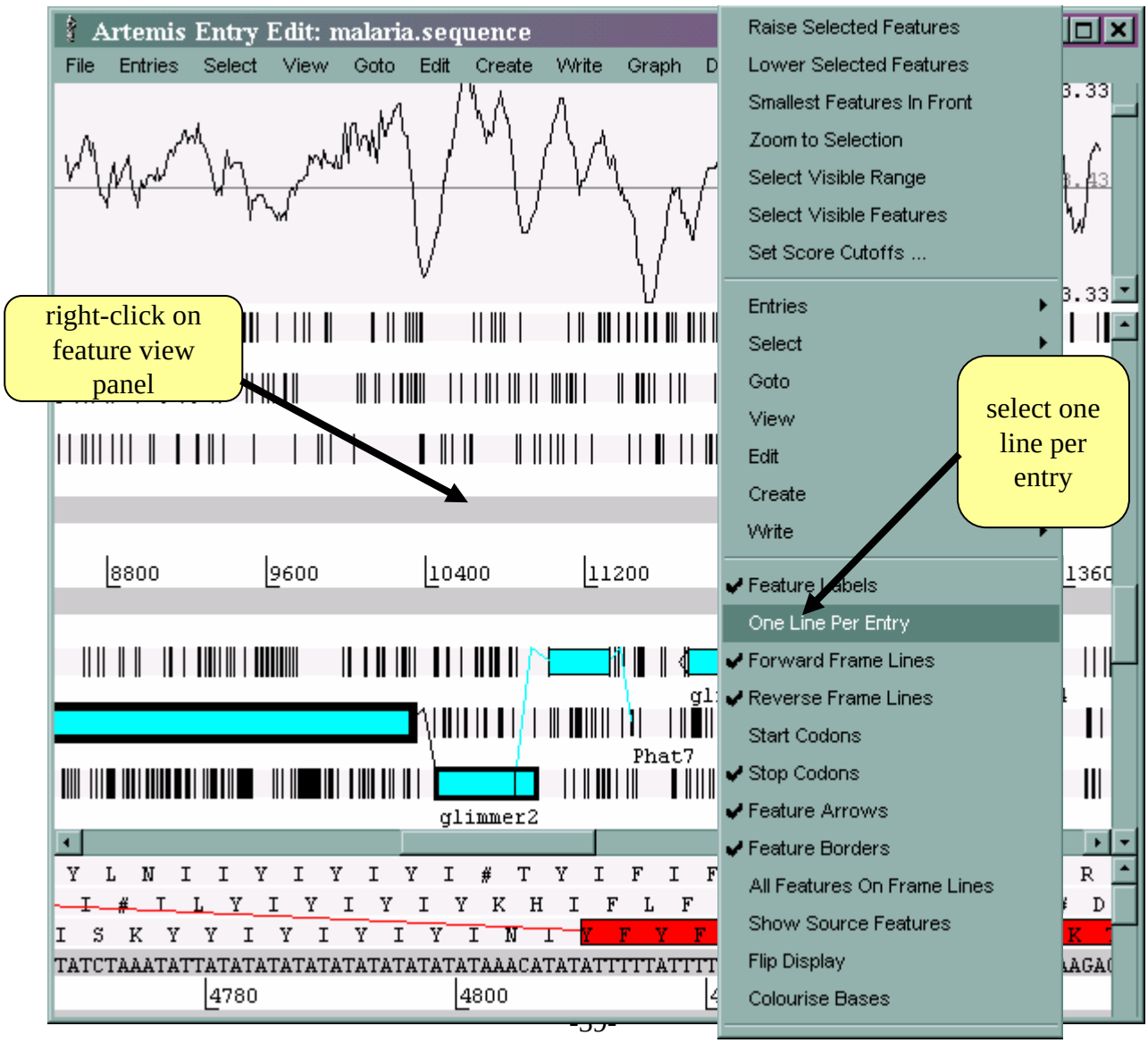
You may noticed after you performed the merge function that one of the exons has subsequently jumped into another reading frame. Artemis automatically splices the CDS and so if the exon boundaries have an additional partial codon then any following exon will be pushed into another reading frame to account for this. To correct this you can edit the exon boundaries directly by turning on manual editing in the options menu of the Artemis start-up window, (as shown above). This will now allow you to edit the start and end positions of the feature boxes by using the left mouse button. Click and hold down the curser over the first or last base of any feature and then drag the mouse. The feature box should move as you drag it (see below. This can be a little tricky so please ask)



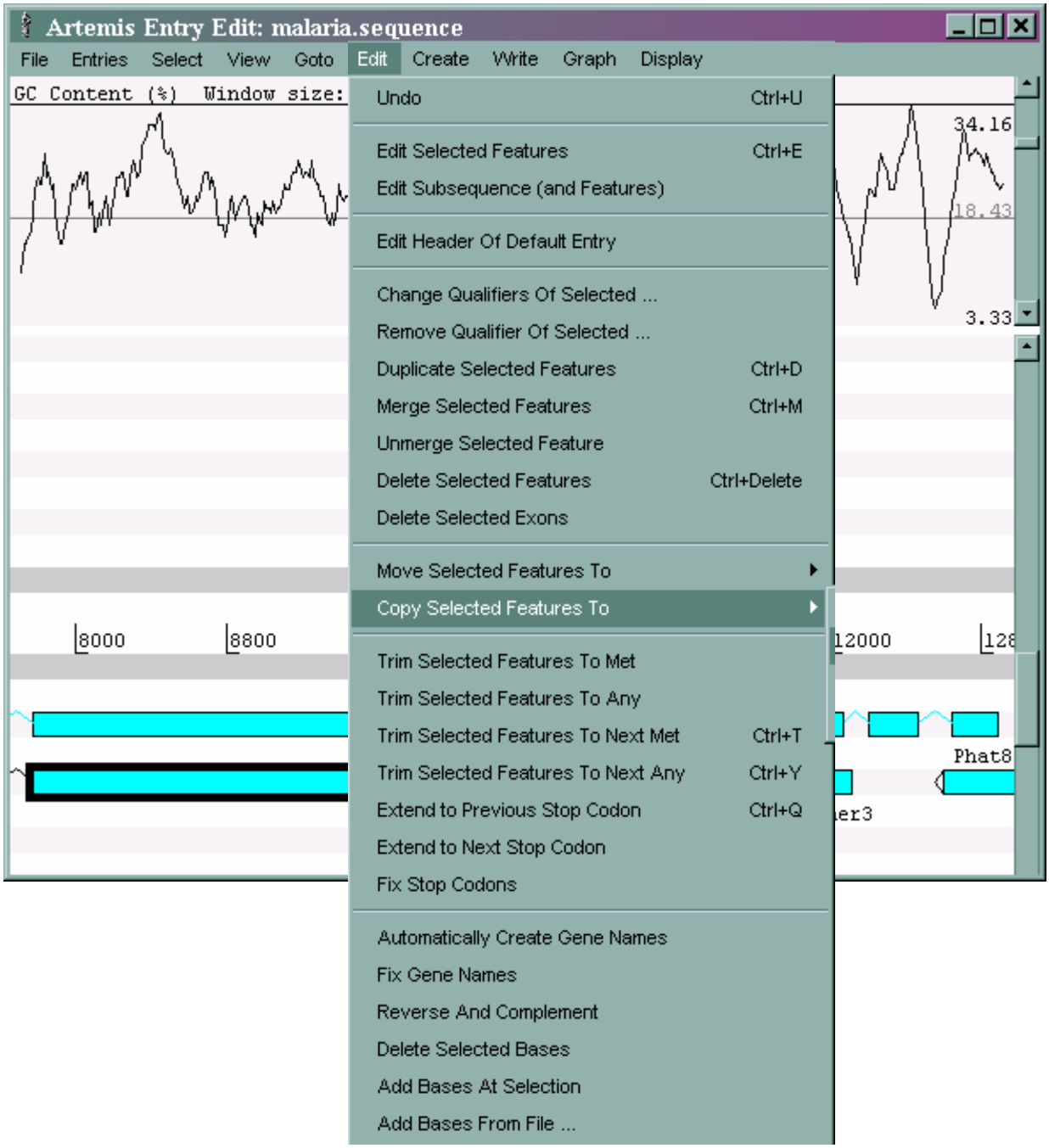
When manually editing your exons you can should look out for appropriate splice donor and acceptor sites.

Once you are happy with your newly created exon re-run the fasta search and see how this compares with the other hits in the public databases. If there are more exons to mark up try and complete the gene model.

To compare the output of different algorithms alongside each other, it is necessary to use a different view in Artemis – “One line per entry”.



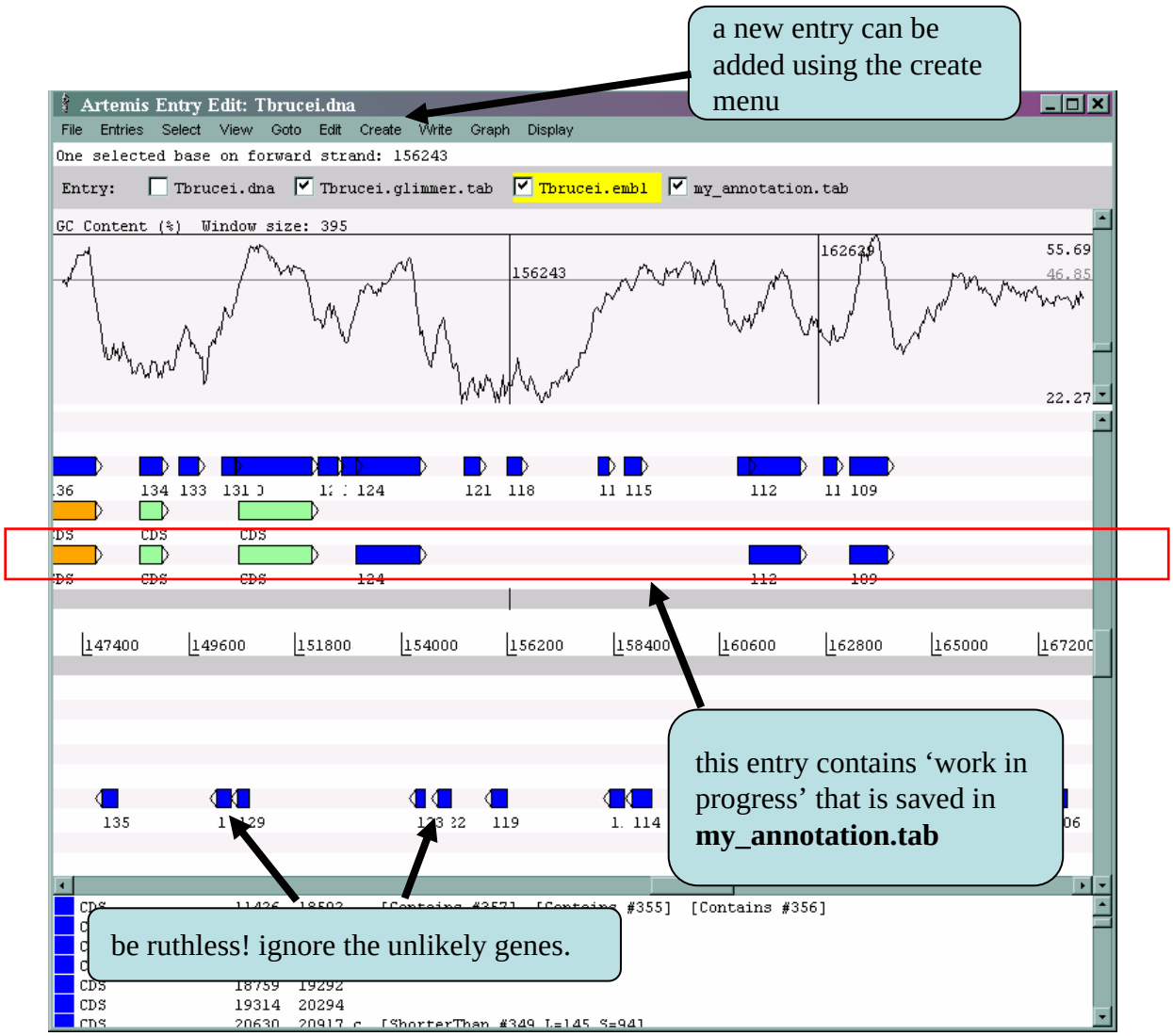
Now feature coordinates can be directly compared against each other. After running fasta, you can copy a feature that you are happy with to the malaria.annotation file



Exercise 6
Gene finding in kinetoplastid parasite *Trypanosoma brucei*

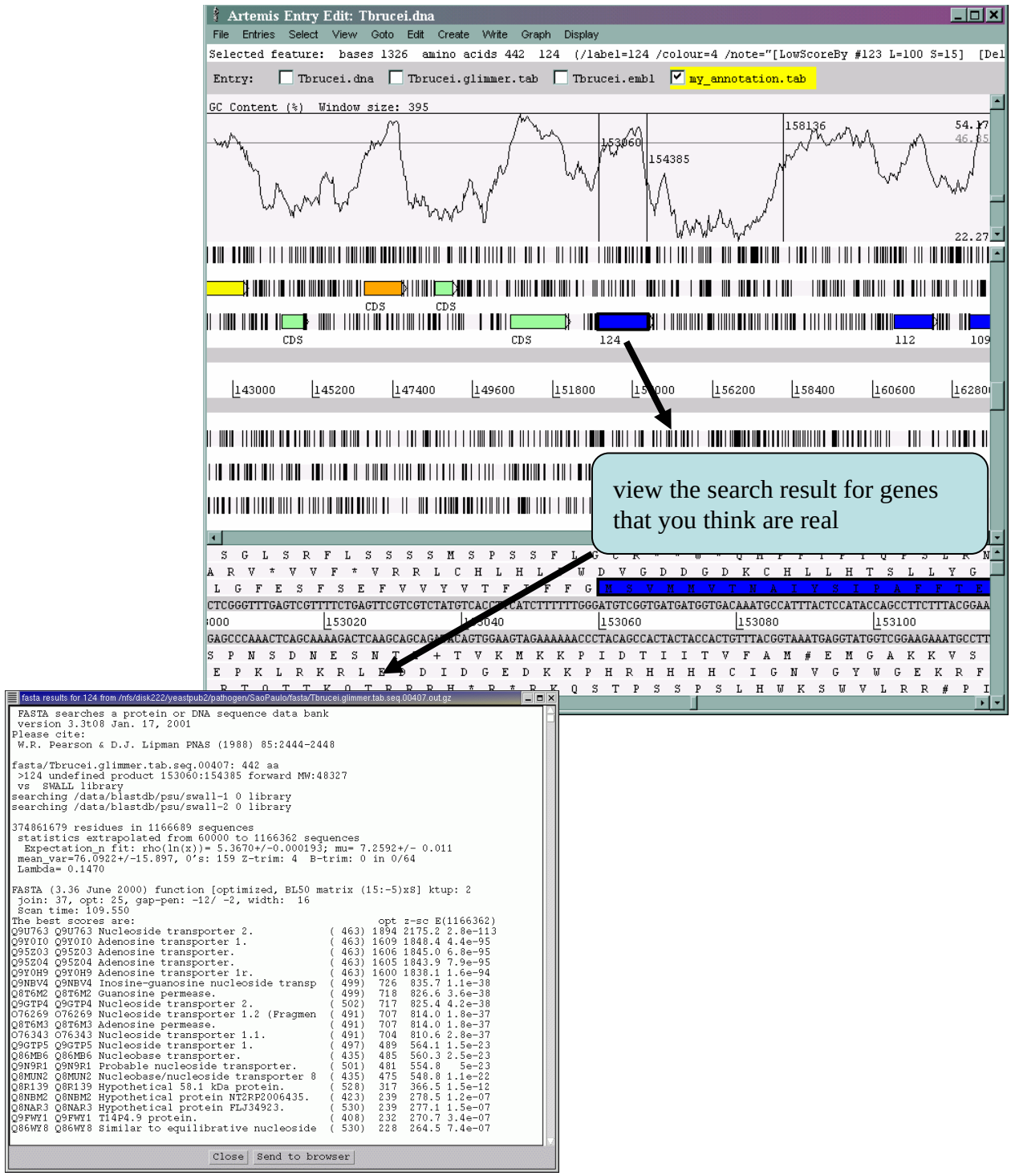
You will need to start a new Artemis session and open the file called **Tbrucei.dna**. The sequence you are going to look at is a large region from chromosome 9 of *T. brucei* (~242kb). Add to the sequence a graph of G+C content, as before, and open up the file **Tbrucei.glimmer.tab**, which contains Glimmer prediction for this region. What can you already see about the sequence that will help you decide which genes are real?

Some annotation has been provided to get you started. Open the file called **Tbrucei.embl**. Now create a new entry to store your own annotation. You can copy whichever genes you believe are real from **Tbrucei.embl** and **Tbrucei.glimmer.tab**. You will need to use bothe the 6-reading frames and the One Line Per Entry views.

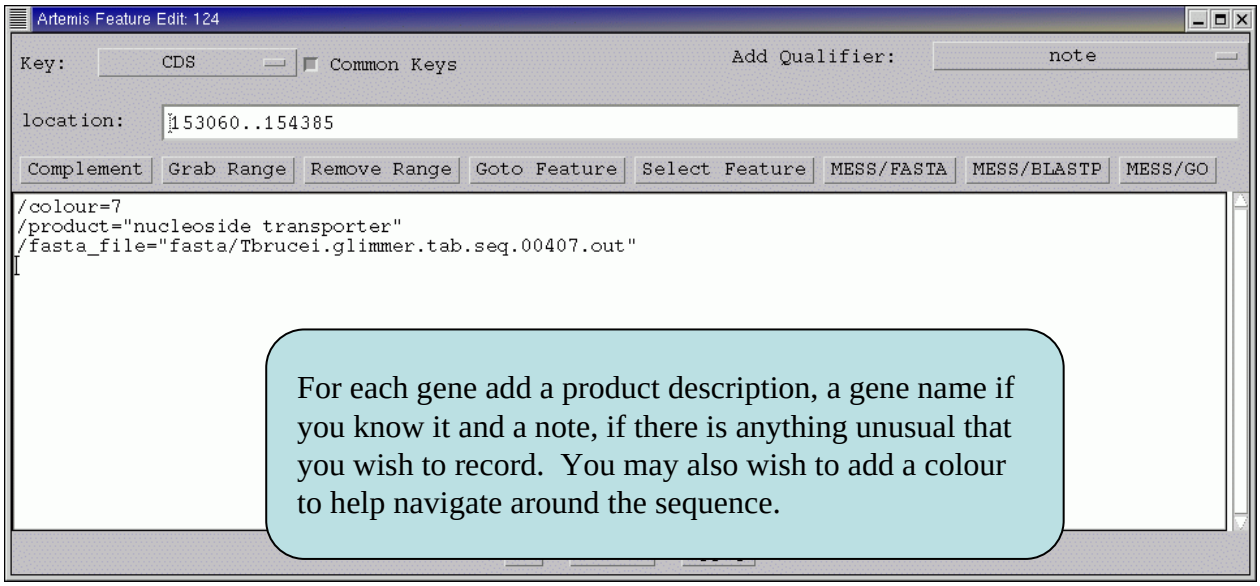


Glimmer is designed for prokaryotic gene prediction, so you will need to check that each gene starts with a Methionine codon. If it does not, trim it to the nearest methionine. This can be done easily from the **Edit** menu

When you have decided that a gene is real, you need to annotate it. If you haven't already run searches do so and view the results. More information on this will be given in the following Module.



Based on the alignment, make sure that the gene model is satisfactory before adding your annotation



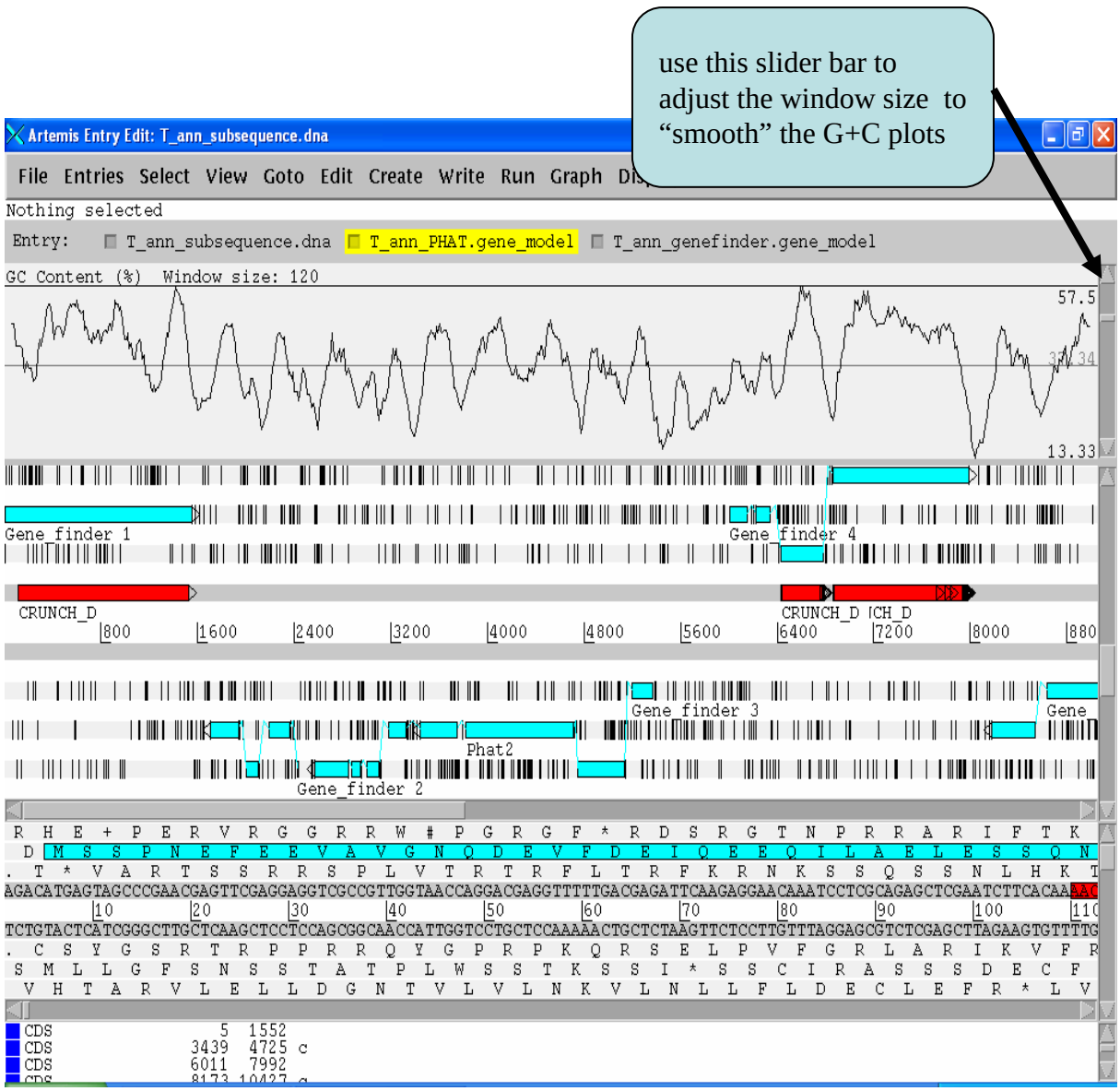
For each gene add a product description, a gene name if you know it and a note, if there is anything unusual that you wish to record. You may also wish to add a colour to help navigate around the sequence.

Using "view selection" from the view menu, you will see your annotation for a given feature in EMBL format. This is the information that Artemis actually records.
For example:

```
FT   CDS             153060..154385
FT                   /product="nucleoside trransporter"
FT                   /colour=4
```

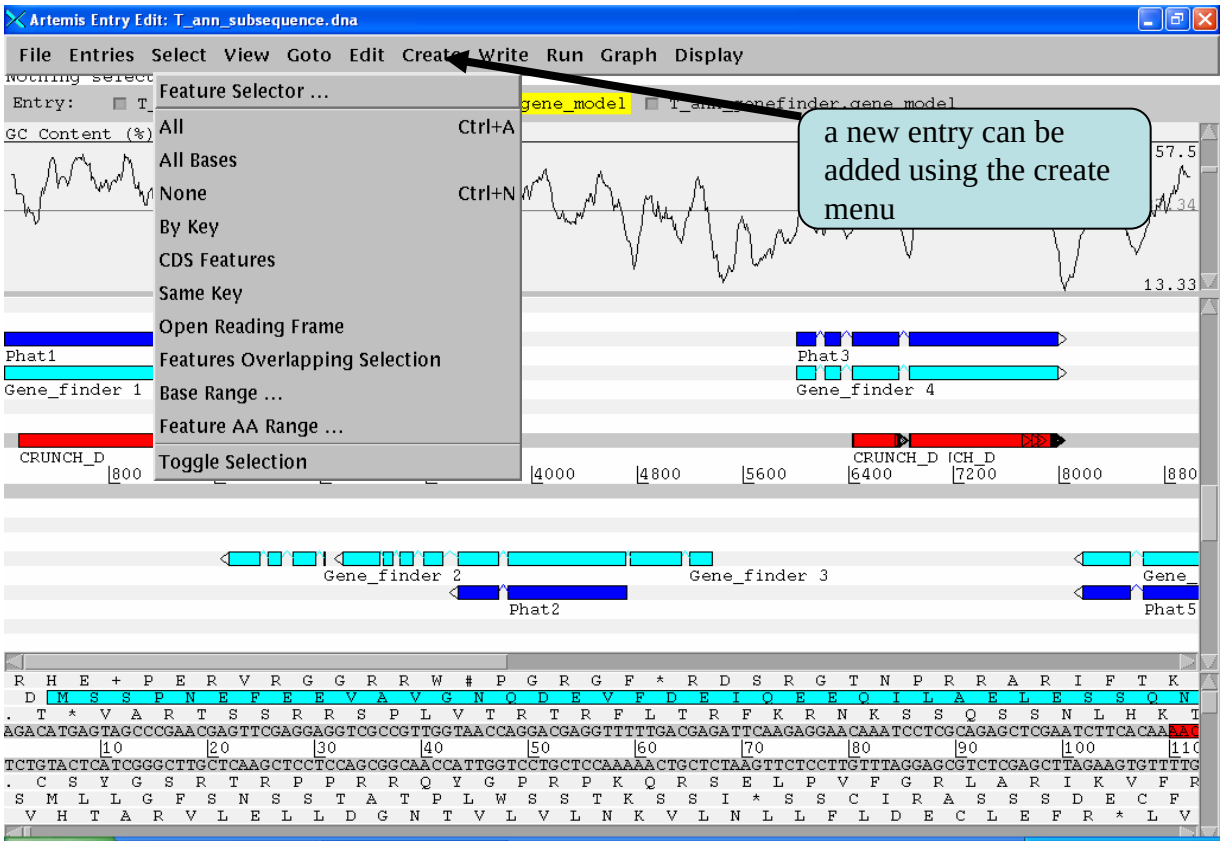
Exercise 7
Gene finding for spliced genes (*Theileria annulata*)

You will need to start a new Artemis session and open the file called **T_ann_subsequence.dna**. The sequence you are going to look at is a ~17 kb region from c*T. annulata genomic DNA*. Add to the sequence a graph of G+C content, as before, and open up the following files one-by-one: **T_ann_PHAT.gene_model** (which contains PHAT gene predictions for this region), **T_ann_genefinder.gene_model** (which contains genefinder gene predictions for this region), and **T_ann_blastsearch_SWALL**. (which contains the blastx results against the SWALL database) for this region. Add the G+C plot to the window as you did before.



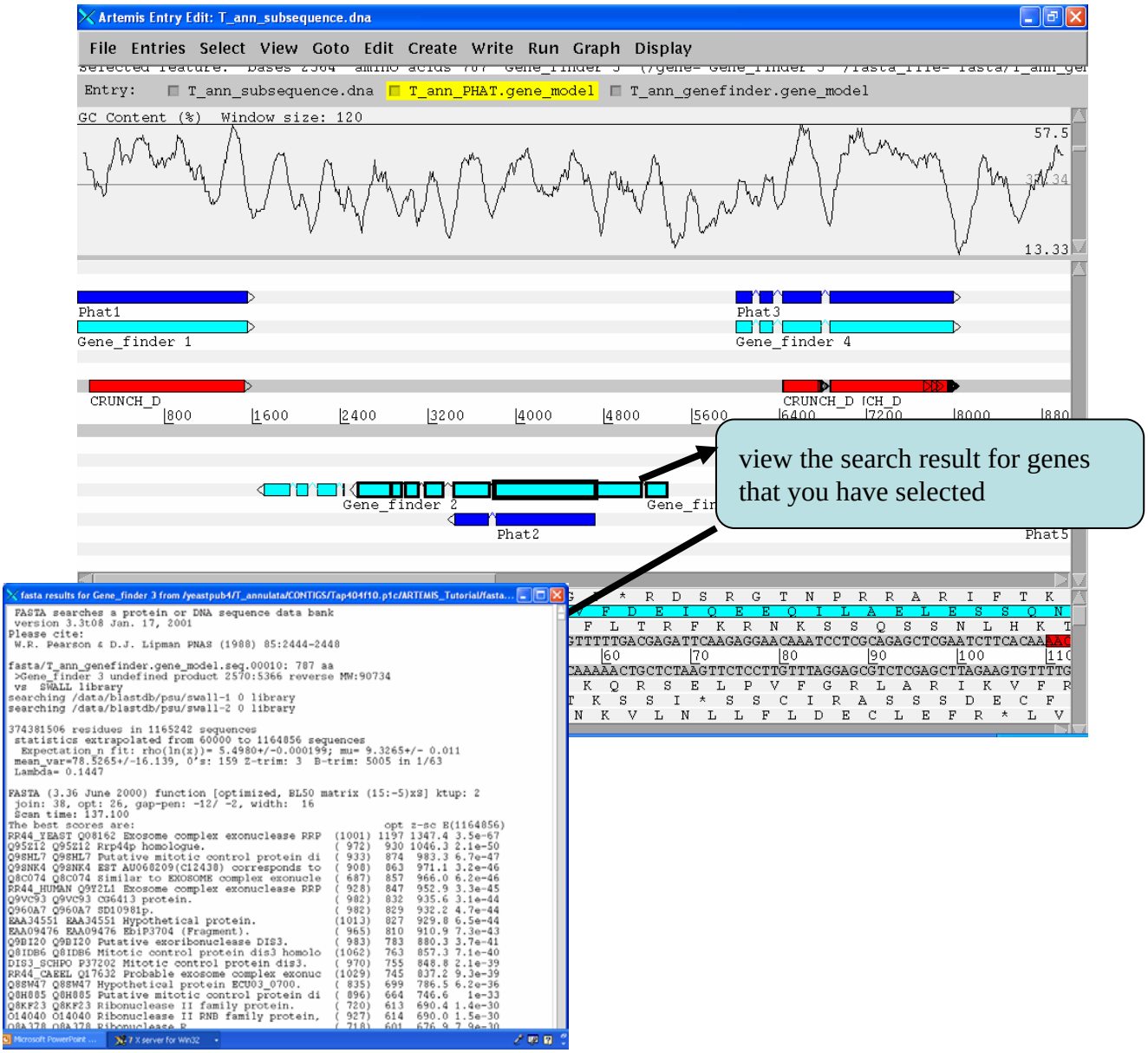
Select 'One Line Per Entry view as you did in previous exercises. Then select all 'CDS Features' from 'Select' drop-down menu and after selecting all CDS features select 'Run Fasta (%L) on selected features' under the 'Run' drop-down menu. Similarly run blastp against SWALL on all of the selected CDS features. It will take some time to run the Fasta and blastp searches and it will report to you when the fasta / blastp searches are completed.

Create a new blank entry to store your own annotation and save your entry as my_annotation2.tab. Now for check the Fasta and blastp results for each CDS feature and decide about the correct gene model, based on the results of your searches. You may need to combine the results from both the automated gene prediction algorithms (such as PHAT and genefinder in this particular case) to reach to a consensus gene model which you think is the most likely gene model for a particular gene and copy the model to your own entry **my_annotation2.tab** and add your annotation (such as the gene product and a specific colour, based on the colour scheme mentioned later in the exercise). You will need to use both the 6-reading frames and the One Line Per Entry views as and when required and also check the blastx hits (if any) for a given gene prediction from the file **T_ann_blast_search_SWALL**. You can copy whichever genes you believe are real from the **T_ann_PHAT.gene_model** and **T_ann_genefinder.gene_model** to your own annotation file..

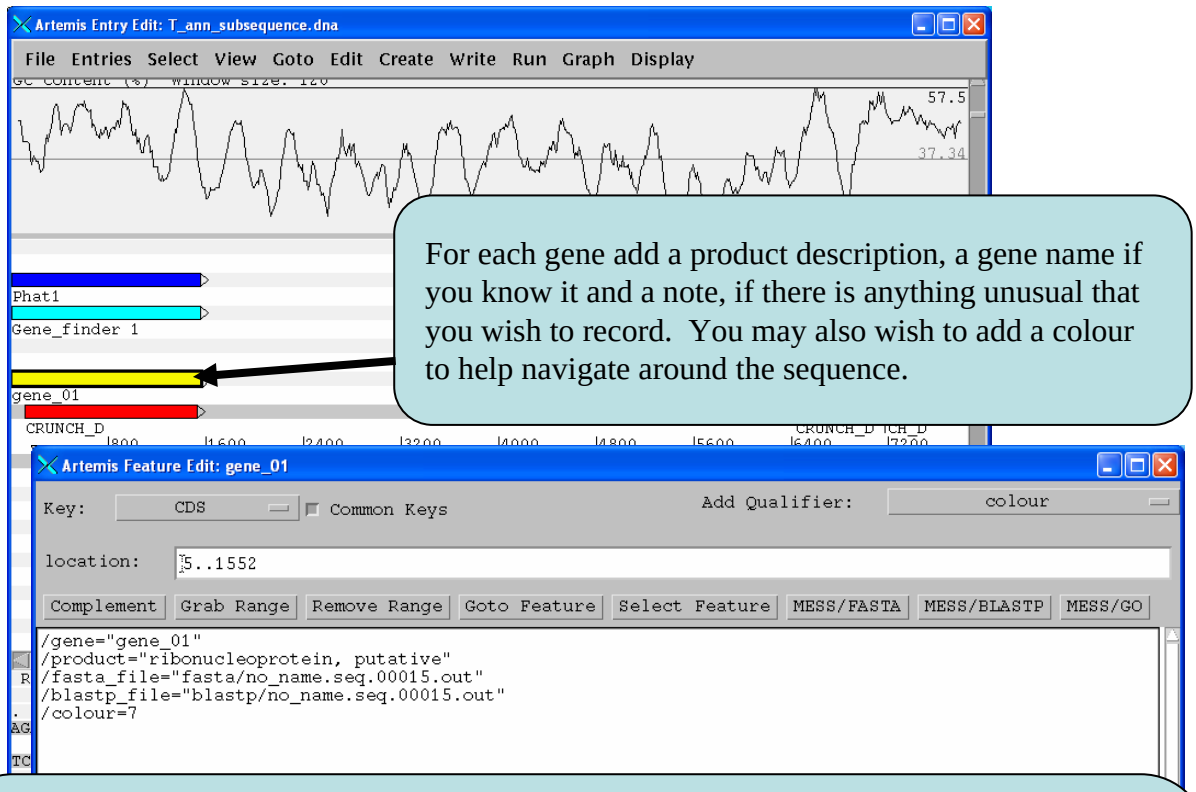


For every gene prediction, check the splice-boundaries are correctly predicted (following the GT-AG rule) and also check that the every gene you predict in you're my_annotation2.tab file starts at a start codon and ends at a stop codon.

You view the fasta or blastp search results by first selecting a CDS feature and then selecting appropriate search results from the 'Search Results' menu under the 'View' drop-down menu. More information about these will be given in the following Module.



Based on the alignment, make sure that the gene model is satisfactory before adding your annotation



For each gene add a product description, a gene name if you know it and a note, if there is anything unusual that you wish to record. You may also wish to add a colour to help navigate around the sequence.

Using "View Selection" from the "View" menu, you will see your annotation for a given feature in EMBL format. This is the information that Artemis actually records.
For example:

```
FT    CDS              5..1552
FT                     /product="RNA-binding protein, putative"
FT                     /gene="gene_01"
FT                     /colour=7
```

After you have finalised about the prediction of the gene models in your own annotation file and have added preliminary annotation, compare your own annotation with the **T_ann_curated_gene_model** annotation file (which contains preliminary annotation of corrected (by an annotator) gene models) by uploading the file in Artemis.

For colouring the genes on this DNA contig, use the following scheme:
Hypothetical protein: colour 8; Protein with known homologues other organisms: colour 7; conserved hypothetical protein: colour 10; Protein, already known in Theileria species: colour 2

Module 5
Small Scale Annotation

Introduction

In this short Module you will attempt to annotate a small region of genomic DNA. Using the web analysis tools covered in the previous Module (such as Prosite and Pfam) cut and paste the nucleotide or amino acid sequence into the submission box of the relevant web page(s).

Aims

This Module is also your opportunity to have a go at annotating one , or hopefully more genes that have been predicted in the genomic segments detailed over the page.

Note:

It is not practical to rely on cut and paste searches for the analysis of whole genomes and so for large scale genome analysis these programs must be installed and run locally on your own computer. This has the added advantage of allowing you to feed the input to these programs in batch i.e. sending off hundreds of CDS/proteins in one operation. This also makes it possible to convert the output of these searches into a form that can be read directly into Artemis, examples to which will be included.

Unfortunately, local installation of these software falls outside of the scope of this workshop. To do this you need to have systems administrators clearances at your home institute and a detailed knowledge of your computer operating environment. However, so as not to dodge this important issue we can give you details of how to approach doing this and so I urge you to speak to the demonstrators about this during the course

Exercises:

The choices are:

Own Sequence Please ask a demonstrator.

Exercise 1

Aspergillus - Eukaryotic filamentous fungi

Or

Exercise 2

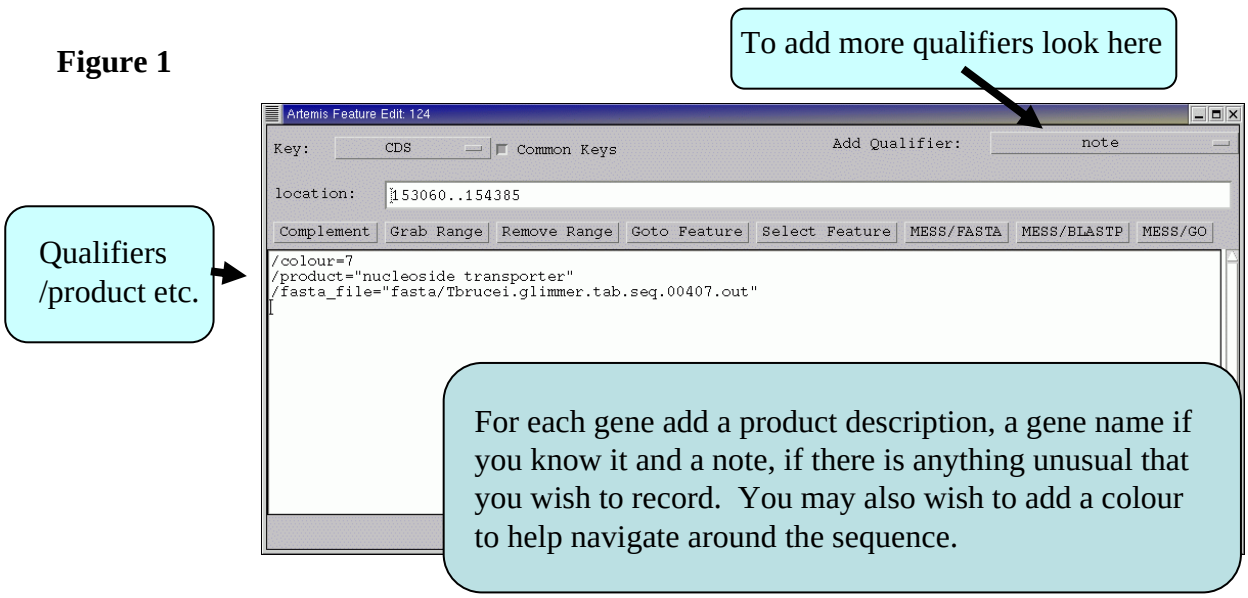
Yersinia - Gram-negative Prokaryote

Once in the correct directory start up an Artemis session using the appropriate files (detailed in the exercise details themselves).

Although you have only been given minimal 'Manual-based' direction for these exercises there will be plenty of demonstrators around to help you out if you are stuck, so please ask. The details of the file for each exercise and what they contain are given below.

An example of a partially annotated CDS (figure 1). Aim to fill in some of the qualifier fields: gene, product etc. For the Prokaryotic exercise there is a colour code and classification system you may want to use in Appendix VI and VII. Eukaryotic protein classification will be mentioned in the final Module of this course.

Figure 1



Exercise 1

Aspergillus fumigatus is the most common mould pathogen of human and usually causes both invasive aspergillosis and allergies in immunocompromised patients and allergic diseases in patients with atopic immune systems. We have provided you with a part of *Aspergillus* genomic DNA sequence that originated from a pilot project to sequence part of *A. fumigatus* genome by construction of a bacterial artificial chromosome (BAC) library and subsequent BAC-end sequencing and analysis (done at the Sanger Institute in collaboration with the University of Manchester).

All you have to do is to open the main DNA file (containing the DNA sequence and the curated gene models) and annotate at least one gene. To extract the sequence click on the CDS feature you wish to annotate and click on 'view' and 'view bases of selection' or 'view amino acids of selection'. Note that you can view the sequence in different formats. By cutting & pasting the sequence into the 'Web tools you were introduced to in one of the previous modules.

The file you will need is within the exercise directory:

1. Af_2004.genemodels.embl (The *A. fumigatus* DNA with gene models).

As mention in the introduction to this Module for larger scale analysis we cannot use the 'Cut and Paste' approach and need to install and run these search programs locally. The output can then be converted directly into a format that Artemis can read. There are additional files in the current directory which contain this type of search so have a look after you have had a bash at cut and paste.

Pre-run search files:

1. Af_2004_blastx_swall.crunch (Blastx comparison file against all proteins in the public database)
2. Af_2004_blastx_nidulans.crunch (Blastx comparison file against all *A. nidulans* proteins
3. Af_2004_signalp.tab (The SignalP output file).
4. Af_2004_tmhmm.tab (The TMHMM output file).
5. Af_2004_pfam.tab (The Pfam output file).
6. Af_2004_tigr.tab (The TIGRfam output file).

Exercise 2

This exercise is centred on a segment of bacterial DNA taken from the genome sequence of Yersinia sp. X. The file Yersinia_2004.embl contains the sequence and predicted CDS for this region. All you have to do is to open the main DNA file (containing the DNA sequence and the curated gene models) and annotate at least one gene. To extract the sequence click on the CDS feature you wish to annotate and click on 'view' and 'view bases of selection' or 'view amino acids of selection'. Note that you can view the sequence in different formats. By cutting & pasting the sequence into the 'Web tools you were introduced to you in one of the previous modules.

The file you will need is within the exercise directory:

1. Yersinia_2004.embl (The *Yersinia* DNA with gene models).

As mentioned in the introduction to this Module for larger scale analysis we cannot use the 'Cut and Paste' approach and need to install and run these search programs locally. The output can then be converted directly into a format that Artemis can read. There are additional files in the current directory which contain this type of search so have a look after you have had a bash at cut and paste.

Pre-run search files:

1.	Yersinia_2004.embl	The *Yersinia* DNA with gene models.
2.	SignalP_2004.tab	The output of a SignalP search (signal sequences)
3.	Prosite_2004.tab	The output of a Prosite search (protein motifs)
4.	TMHMM_2004.tab	The output of a TMHMM search (membrane domains)
5.	PF_2004.tab	The output of a Pfam search (protein motifs)
6.	Yersinia.cod	The *Yersinia* codon usage table
7.	Blastx_2004.tab	The output of a blastX search of *Yersinia* against SWALL

Module 6
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 can be obtained from the ACT user manual:
http://www.sanger.ac.uk/Software/ACT /manual/).

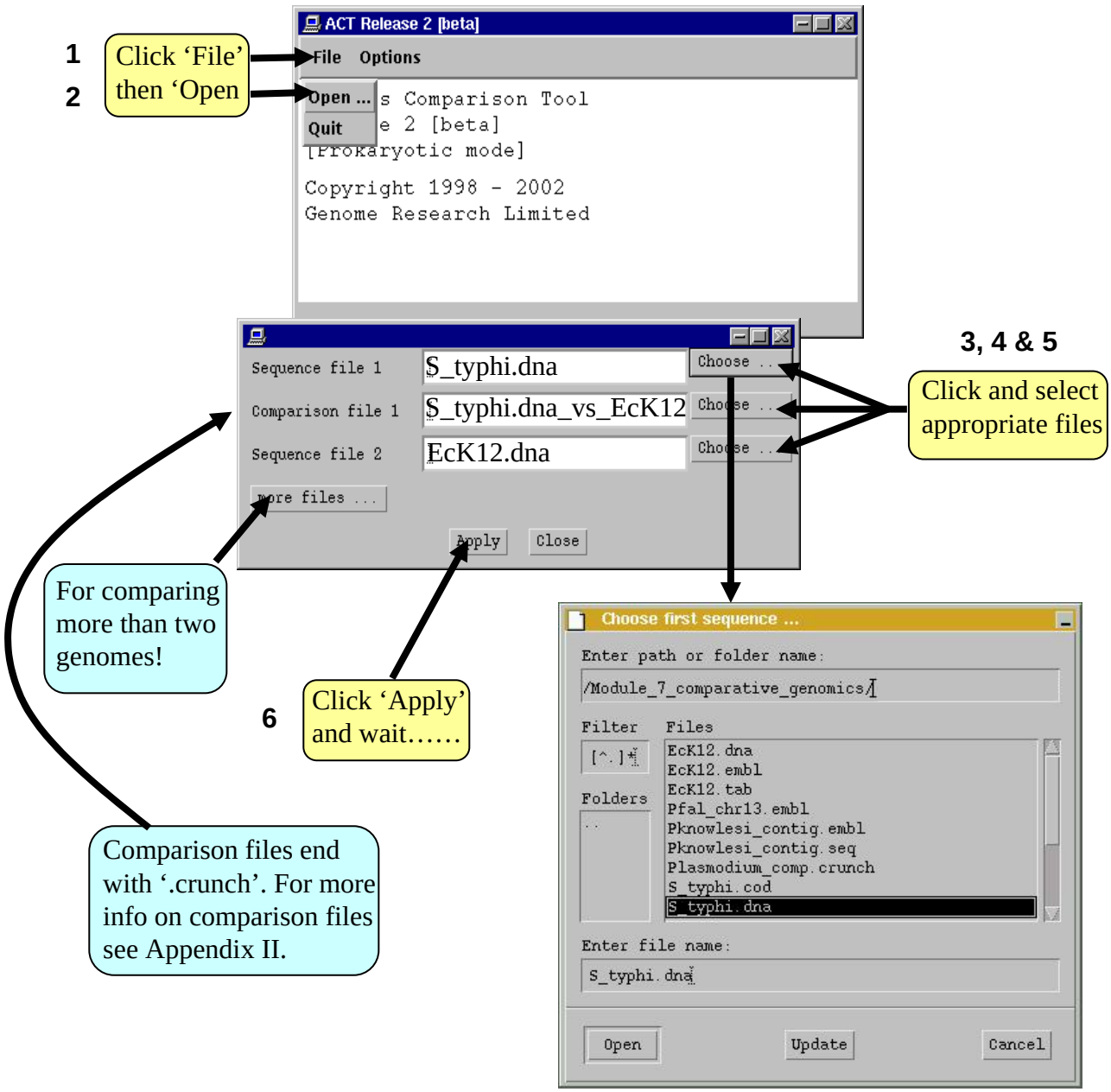
Aims

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

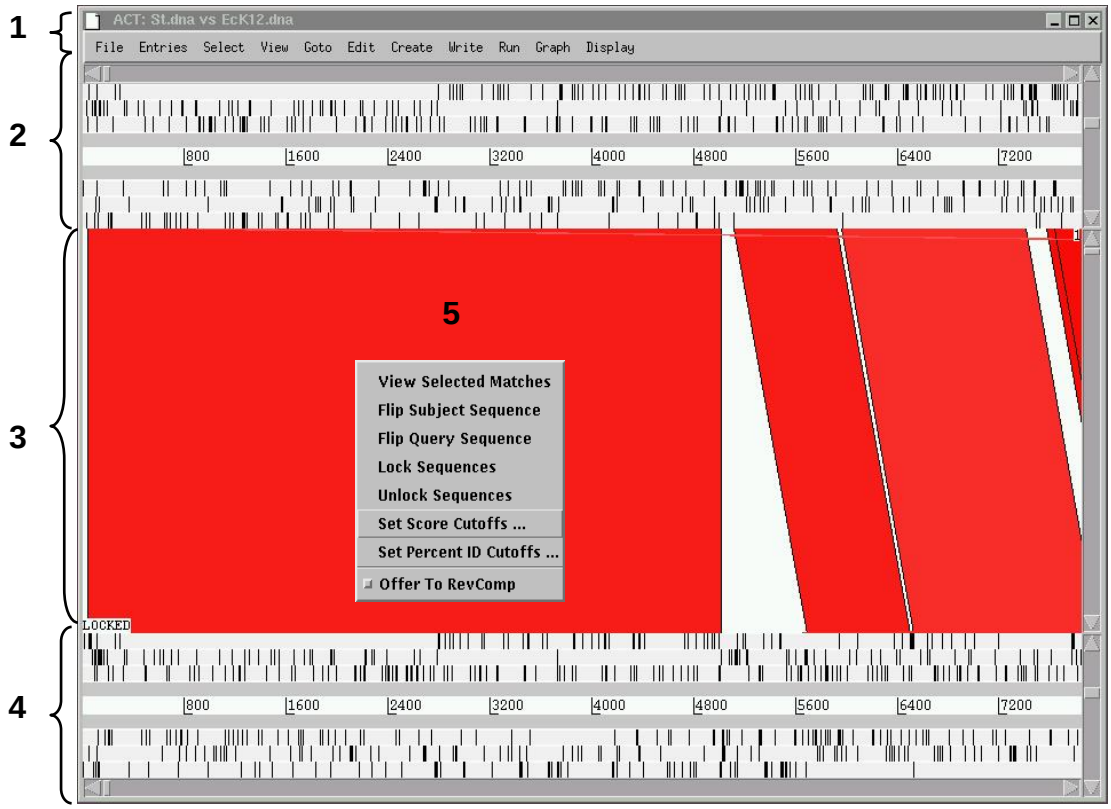
Make sure you're in the correct directory.
Then type
act & [return]
A small start up window will appear.
Now let's load up a *S. typhi* versus *Escherichia coli* comparison.

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



2. The basics of ACT

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



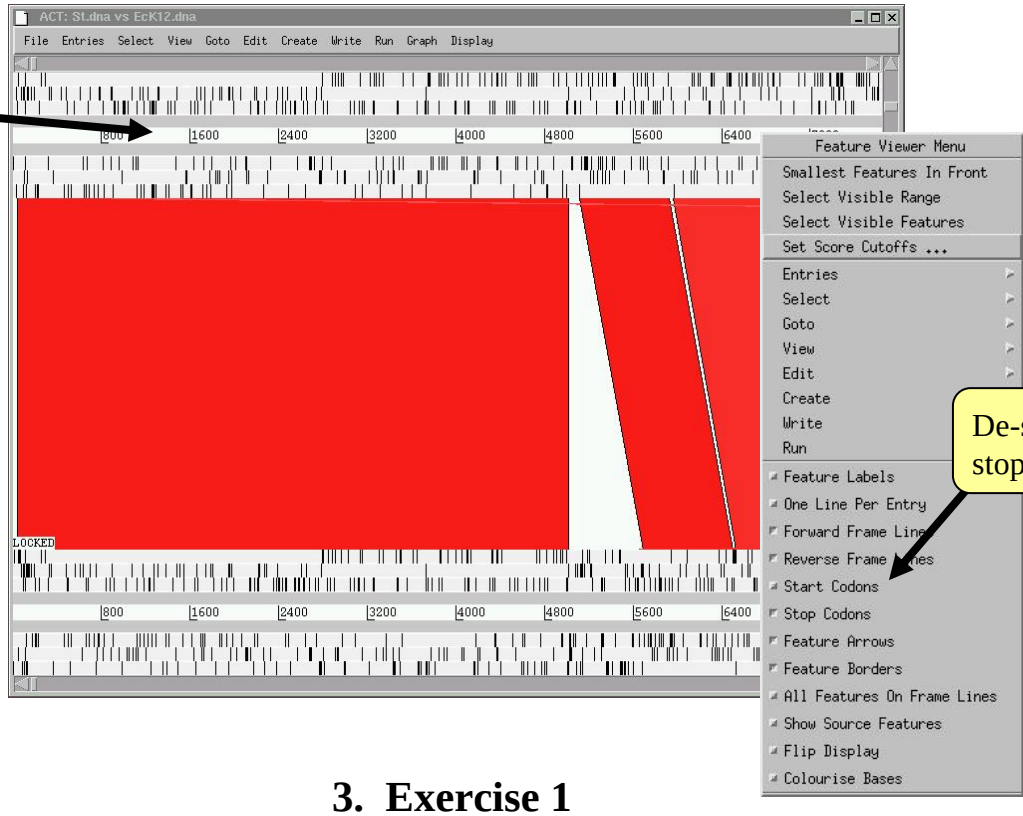
1. Drop-down menus. These are mostly the same as in Artemis. The major difference you'll find is that after clicking on a menu header you will then need to select a DNA sequence before going to the full drop-down menu.
2. This is the Sequence view panel for 'Sequence file 1' (Subject Sequence) you selected 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.
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

2

De-select stop codons



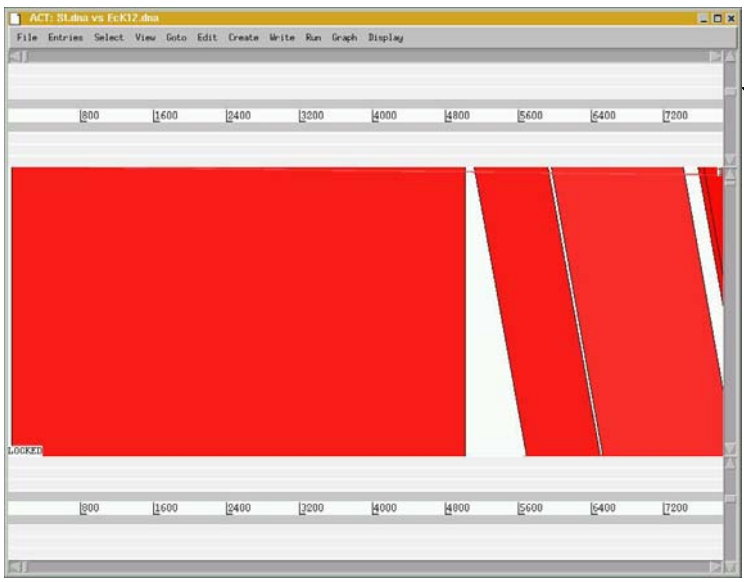
3. Exercise 1

Introduction & Aims

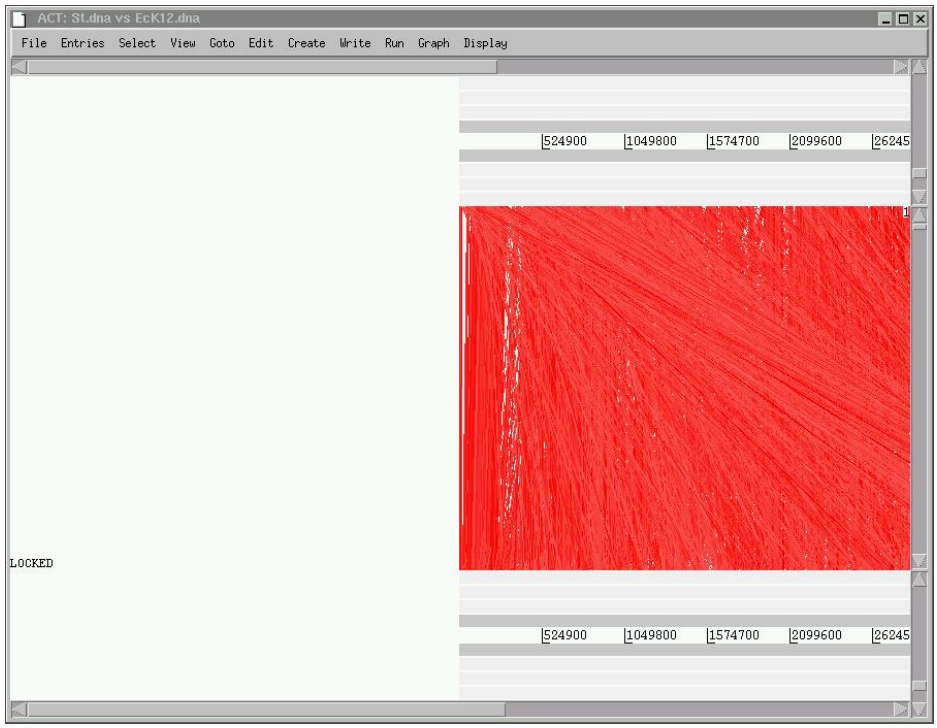
In this first exercise we are going to explore the basic features of ACT. Using the ACT session you have just opened we firstly are going to zoom outwards until we can see the entire *S. typhi* genome compared against the entire *E. coli* K12 genome. As for the Artemis exercises we should turn off the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

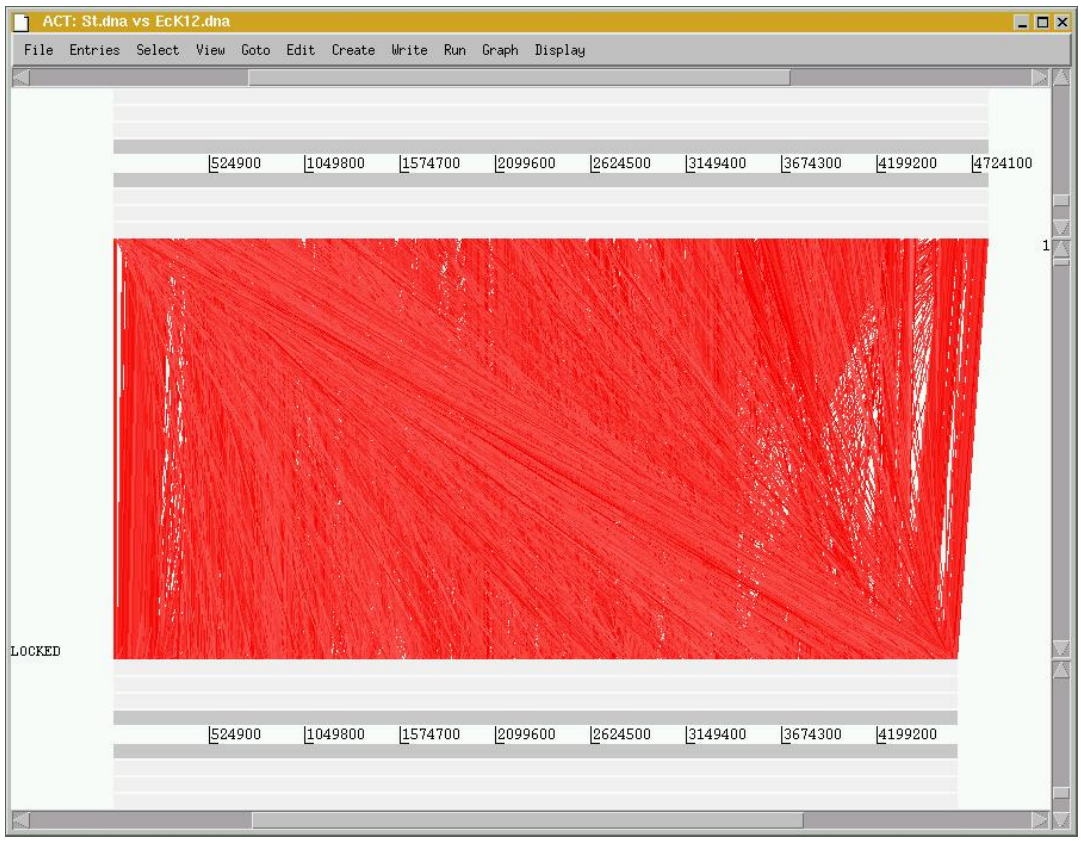
Now turn the stop codons off in the other sequence too. Your ACT window should look something like the one below:



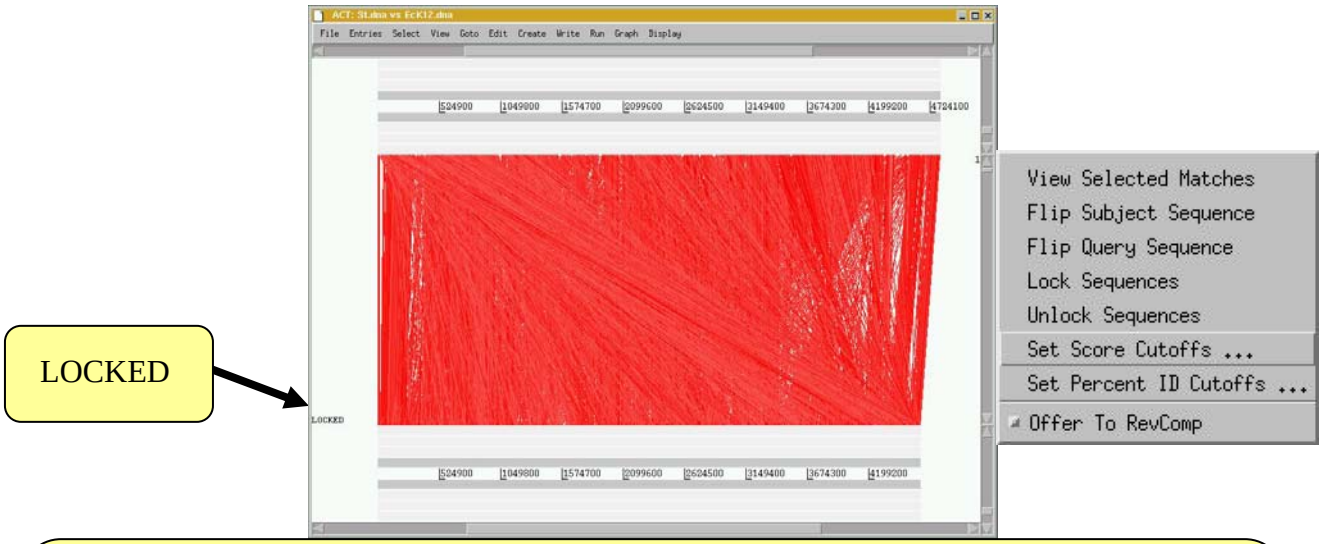
Use the vertical sliders to zoom out. Drag or click the slider downwards from one of the genomes. The other genome will stay in synch.



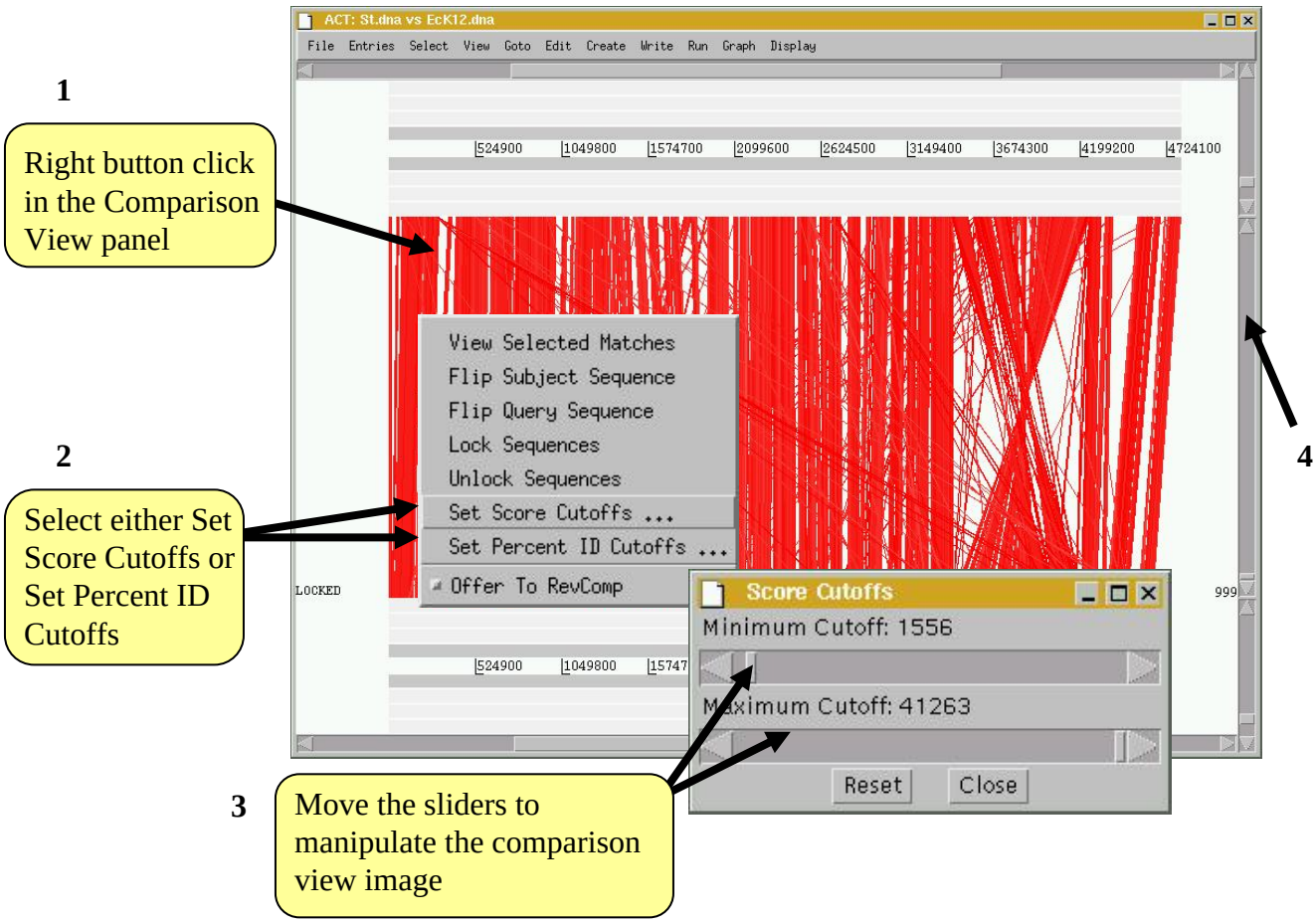
Once zoomed out your ACT window should look similar to the one shown above. If the genomes in view fall out of view to the right of the screen, use the horizontal sliders to scroll the image and bring the whole sequence into view, as shown below. You may have to play around with the level of zoom to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

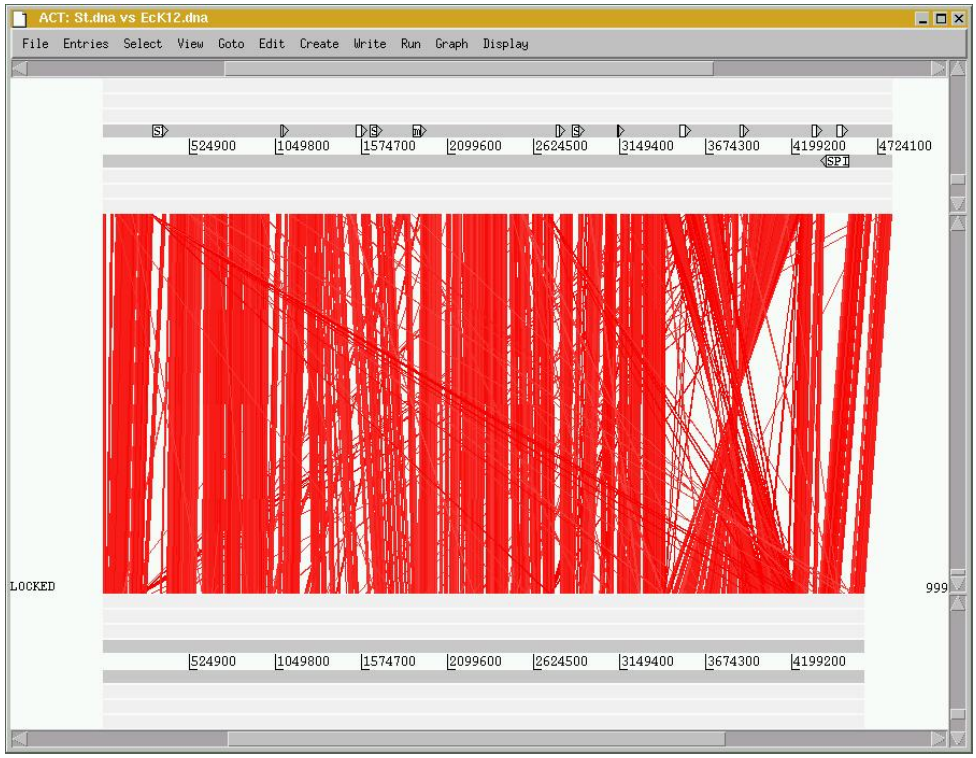


Notice that when you scroll along with either slide both genomes move together. This is because they are 'locked' together. Right click over the middle comparison view panel. A small menu will appear, select Unlock sequences and then scroll one of the horizontal sliders. Notice that 'LOCKED' has disappeared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



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".





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

6. Exercise 2 Part I:

Plasmoidum falciparum and Plasmodium knowlesi: Genome Comparison

Introduction

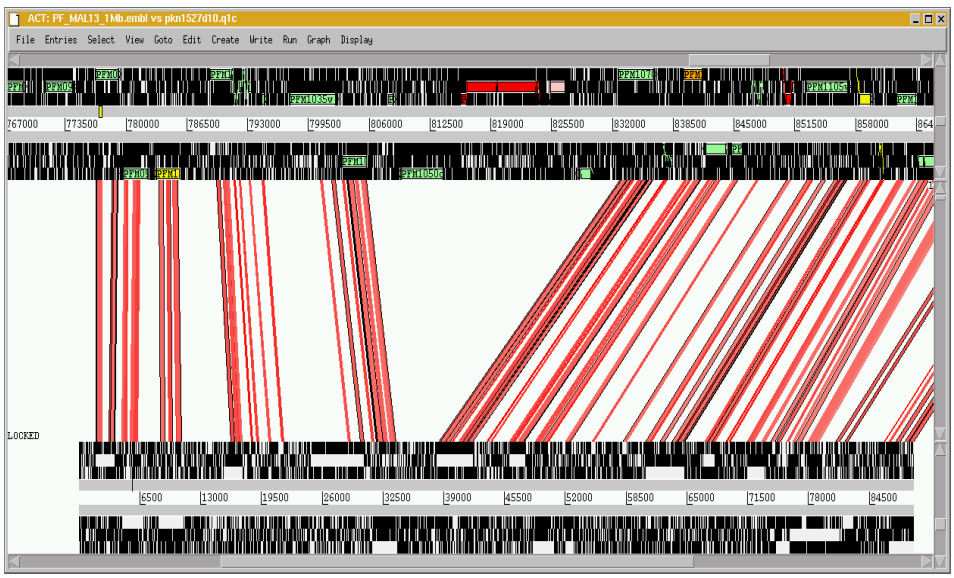
The parasite *P. falciparum* is responsible for hundreds of millions of cases of malaria and causes over 1 million deaths every year. Treatment and control have become difficult with the spread of drug-resistant malaria strains across the endemic countries in the world and there has been a major emphasis on research as part of our search for new drugs / vaccine candidates to fight against malaria. The analysis of the whole genome of *P. falciparum* has been completed and is made publicly available by the Malaria Genome Sequencing Consortium . Several animal models of malaria have also been used by researchers to study several aspects of malaria biology / host-parasite interactions. Sequences representing partial genomes of some of these model malaria parasites are also available now. This allows us to perform comparative analysis of the genomes of malaria parasites and understand the basic biology of their parasitism, based on the similarities / dissimilarities between the parasites at DNA / predicted protein level.

Aim

You will be looking at the comparison between a genomic DNA fragment of the primate malaria *P. knowlesi* and the previously annotated chromosome 13 of *P. falciparum* . By comparing the two genomic fragments you will be able to study the degree of conservation of gene order and identify new genes in *P. knowlesi* genome. As part of the exercise you will also identify any gross dissimilarity visible between the the two genomic fragments and finally, predict/ modify the gene model for one multi-exon gene in *P. knowlesi* genomic fragment.

The files that you are going to need are:

Pfal_chr13.embl	- annotation file with sequence
Pknowlesi_contig.seq	- sequence file (without annotation)
Pknowlesi_contig.embl	- annotation file with sequence
Plasmodium_comp.crunch	- tblastx comparison file



P. falciparum chr 13 (fragment)

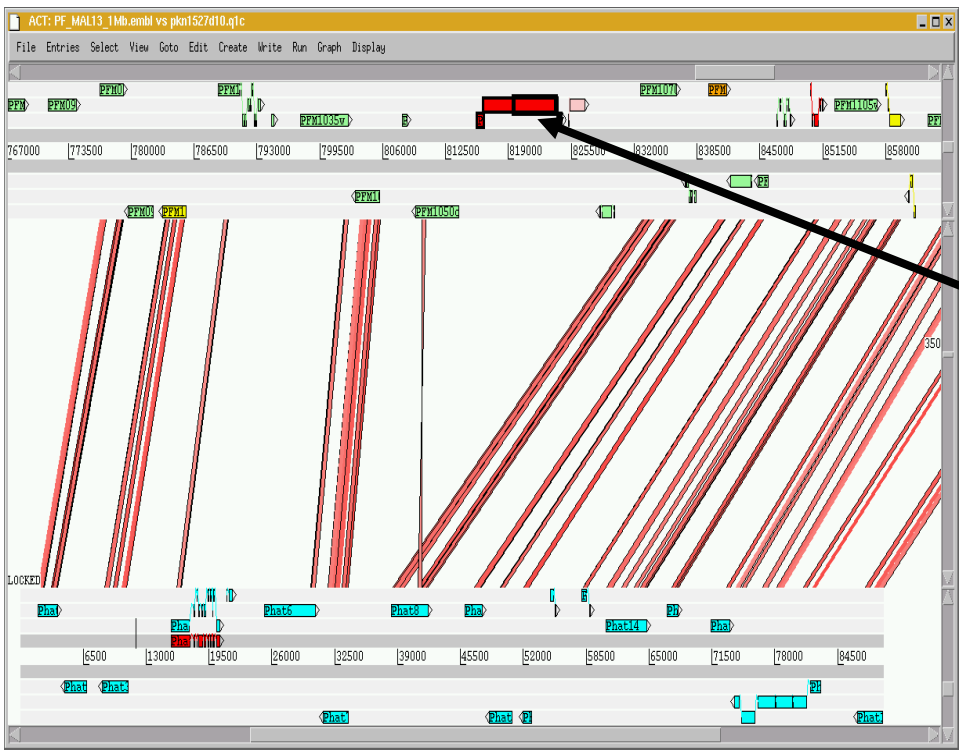
P. knowlesi contig

Comparison of *P. knowlesi* contig and the annotated chromosome 13 fragment of *P. falciparum*

Exercise 2 Part II

Conservation of gene order (synteny)

- In the ACT start up window load up the files Pfal_chr13.embl, Pknowlesi_contig.seq and the comparison file Plasmodium_comp.crunch
- Use the slider on either sequence view panel to obtain a global view of the genome comparison. Also used the slider on the comparison view panel to remove the 'shorter' similarity hits. What effects does this have?
- Can you see conserved gene order between the 2 species?
- Can you see any region where similarity is broken up? Zoom in and look at some of the genes encoded within this unique region in file: Pfal_chr13.embl (top sequence)
- Example location: **Pfal_chr13.embl**, 815823..829969
- What are the predicted products of the genes assigned to this unique location? View the details by clicking on the feature, and then select *`Edit selected feature'* from the *'Edit'* menu after selecting the appropriate CDS feature.
- Can you identify a few putative genes in *P. knowlesi* contig, based on their conserved and syntenic nature with *P. falciparum* chromosome 13? Activate / inactivate stop / start codons in an entry, using the right click button on the mouse. This will allow you to see any potential ORFS.
- Any thoughts about the possible biological relevance of the comparison?



P. falciparum
Pfal_chr13.embl

What is the gene product?

P. knowlesi
Pknowlesi_contig.embl

Exercise 2 Part III

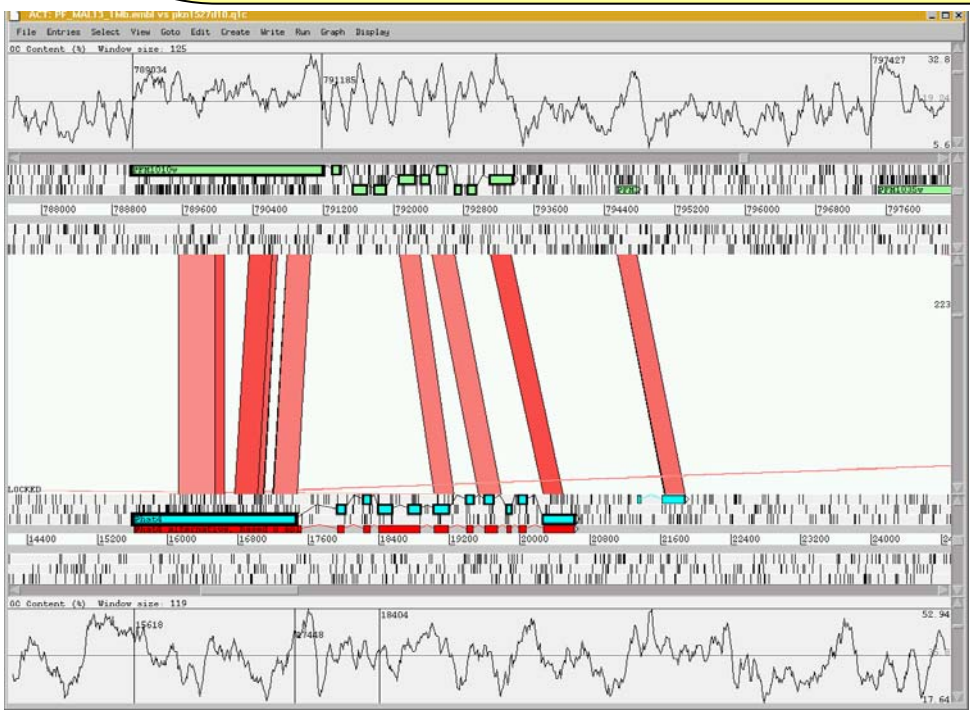
Prediction of gene models:

There are several computer algorithms covered earlier in Module 3 that predict gene models, based on training the algorithm with previously known gene sets with previously known experimentally verified exon-intron structures (in eukaryotes). However, no single programme can predict the gene structure with 100% accuracy and one needs to curate / refine the gene models, generated by automated predictions. We have generated automated gene models for the *P. knowlesi* contig, using PHAT (Pretty Handy Annotation Tool, a gene finding algorithm, see in Mol. Biochem. Parasitol. 2001 Dec;118(2):167-74) and the automated annotation is saved in Pknowlesi_contig.embl.

- Zoom into the *P. falciparum* gene labelled PFM1010w shown below. Can you compare the 2 gene models and identify the conserved exon(s) between the 2 species?
- Use the slider on the comparison view panel to include some 'shorter' similarity hits. Can you now identify all the conserved exons of the PFM1010w orthologue in the *P. knowlesi* contig? (For the time being, disregard the misc_feature for 'Phat4', coloured in red in the 'Pknowlesi_contig.embl' file)
- Open the '*GC Content (%)*' window from '*graph*' menu for both the entries. Can you relate the exon-intron boundaries to GC-content for the *P. falciparum* gene labelled PFM1010w? Is it also applicable to the gene model 'Phat4' in the *P. knowlesi* contig?
- Example regions:

  Pfal_chr13.embl, 789034..793351

  Pknowlesi_contig.embl, 15618..20618



P. falciparum
Pfal_chr13.embl

P. knowlesi
Pknowlesi_contig.embl

Comparison between orthologous genes in *P. falciparum* and *P. knowlesi*

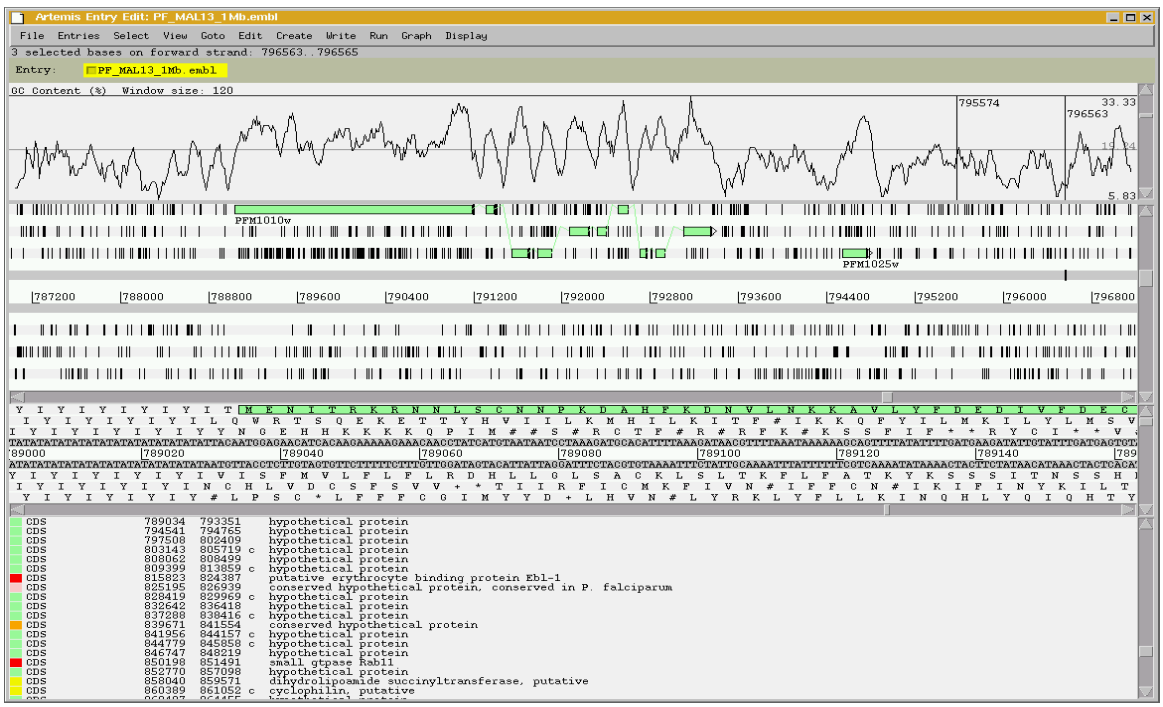
Exercise 2 Part IV

Gene models for multi-exon genes in *P. falciparum*:

- Use '*File*' menu to select entry 'Pfal_chr13.embl' and select '*Edit In Artemis'* to bring up an Artemis window.
- In Artemis window, use '*Graph*' menu and switch 'on' the '*GC Content (%)'* window.
- Use '*Goto*' menu to select '*Navigator*' window and within the Navigator window, select '*Goto Feature With This qualifier value'* and type 'PFM1010w', click then close the dialogue box.
- Go through the annotated gene model for 'PFM1010w' and have a look at the the exon-intron boundaries and compare with the splice site sequences from *P. falciparum* given in Appendix IX.
- Also have a glance through a few other gene models for multi-exon genes and have a look at the intron sequences as well. Can you find any common pattern in the putative intron sequences? Hint – look at the complexity of the sequence
- You can delete exon(s) of any gene by selecting the exon(s) and then choosing '*Delete Selected Exons'* from '*Edit*' menu. Similarly, you can add an exon to a particular gene by co-selecting the exon and the gene (CDS features) followed by selecting '*Merge Selected Features'* from the '*Edit*' menu.

- Example regions:

  Pfal_chr13.embl, 789034..793351, 657638..660023, 672361..673753



Example location: 789034..793351, in Pfal_chr13.embl

Exercise 2 Part V

Curation of gene models in *P. knowlesi*:

We are now going to edit the gene model for *P. knowlesi*.

•Use '*File*' menu from the ACT displaying *P. falciparum* and *P. knowlesi to* select entry 'Pknowlesi_contig.embl' and select '*Edit In Artemis'* to bring up an Artemis window.

•Within the Artemis window, use '*Graph*' menu and switch 'on' the '*GC Content (%)*' window.

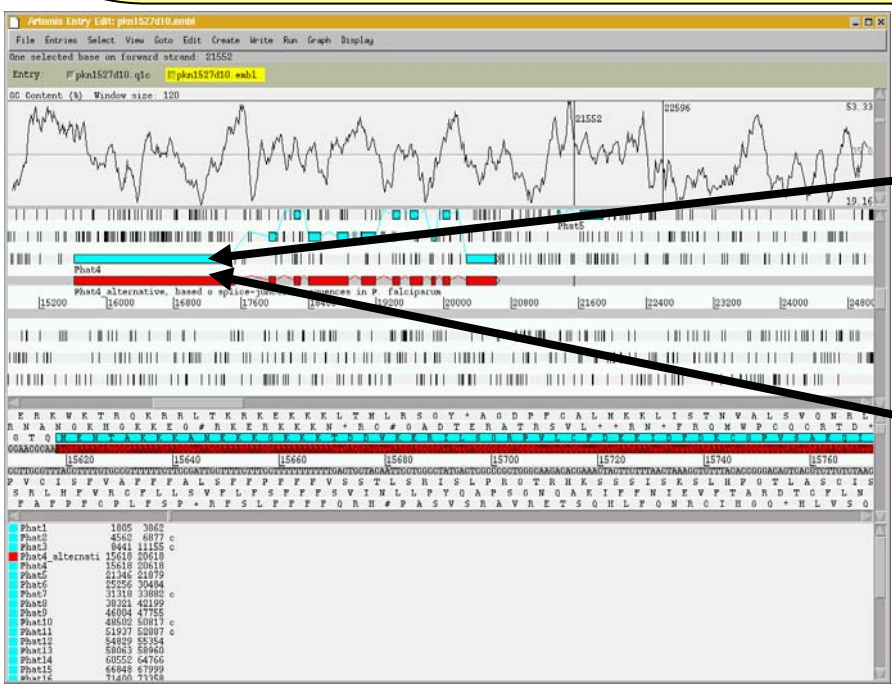
•Use '*Goto*' menu to select '*Navigator*' window and within the Navigator window, select '*Goto Feature With This Text'* and type 'Phat4'.

•Go to the first ACT window, and use the '*Options*' menu to select '*Enable Direct Editing'*

•Go through the gene model of 'Phat4' and have a glance through the exon-intron boundaries. Can you suggest any alternative gene model, after consulting the Table provided in Appendix IX, containing several examples of experimentally verified splice site sequences for *P. falciparum*?

•Example modifications:

Have a look at the '*misc_feature'*, coloured in red (location: 15618..20618). Can you spot any difference in the red gene model of 'Phat4' at the exon-intron boundaries? Select the red feature, click on '*Edit*' menu and select '*Edit Selected Features'* and in the new window that pops out, change the 'Key' from misc-feature to 'CDS' and click on 'OK' button to close the window. Now you can compare the automatically created blue gene model and the curated red gene models at protein level and predict any alternative splicing pattern.



Automated gene prediction for hypothetical gene 'phat4'

Can you curate the 'Phat4' gene model and suggest any alternative splicing pattern such as the red model?

Example location: 15618..20618, in Pknowlesi_contig.embl

Exercise 3

Introduction

Having familiarised yourselves with the basics of ACT, we are now going to use it to look at a region of synteny between *T. brucei* and *Leishmania*.

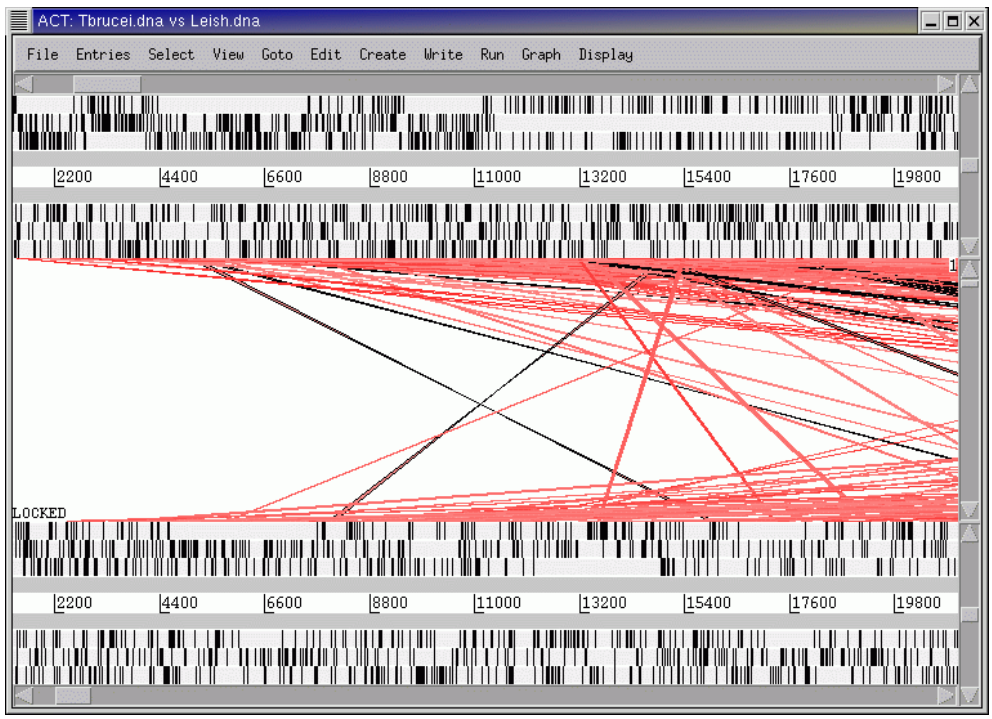
Aim

By looking at a comparison of the annotated sequences of *T. brucei* and *L.major* you will be able to analyse, in detail, those genes that are found in both organisms as well as spot the differences. You will also see how act can be used to study the different chromosome architecture of these two parasite species.

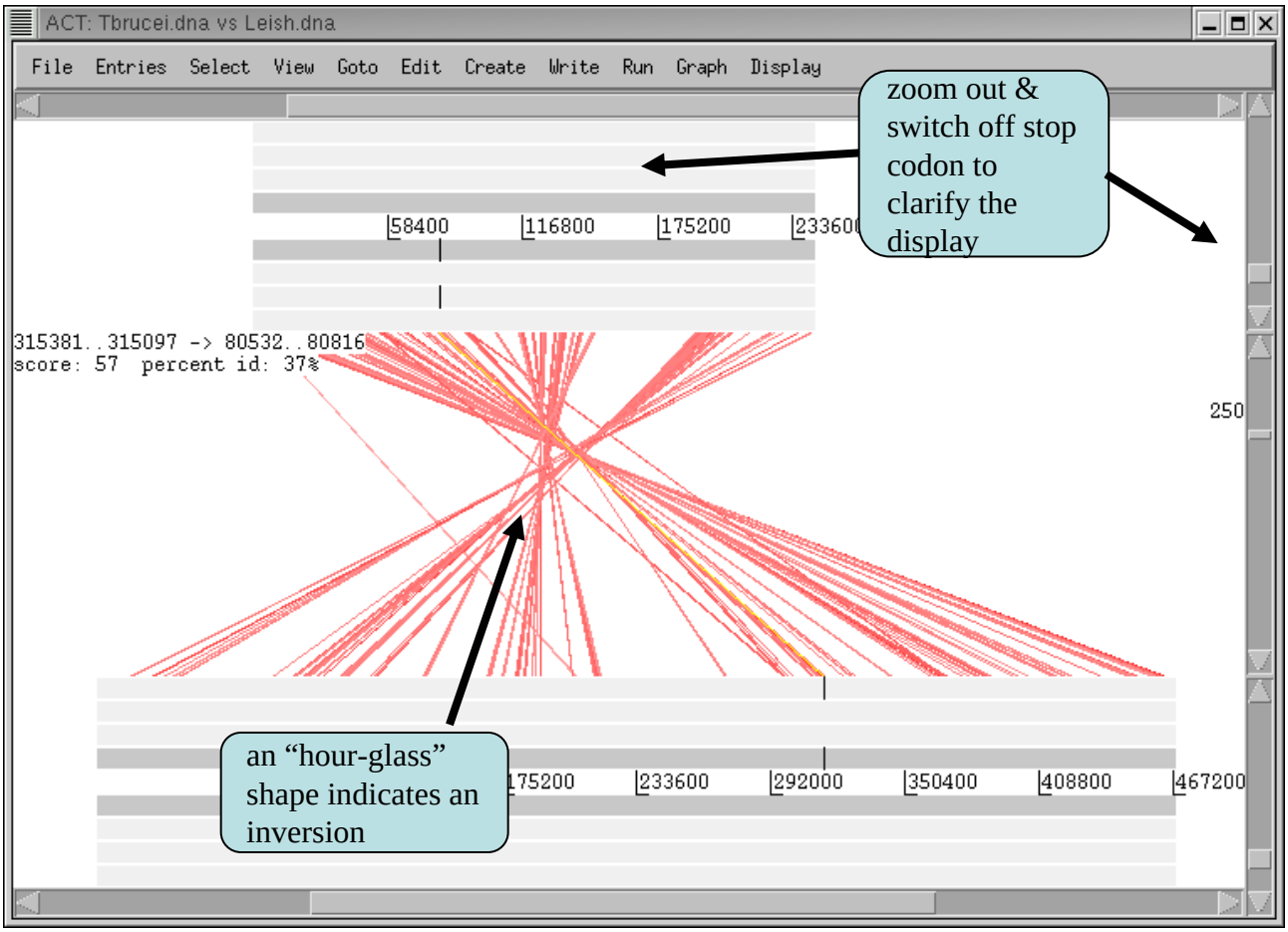
The files that you are going to need are:

Tbrucei.dna	- *T. brucei* sequence
Tbrucei.embl	- *T. brucei* annotation
Leish_vs_Tbrucei.tblastx	- comparison file
Leish.dna	- *L. major* sequence
Leish.embl	- *L. major* annotation

First, load up the sequence files for *T. brucei* and *L. major* and the comparison file in ACT.



Next, you need to find the regions of synteny between the sequences.



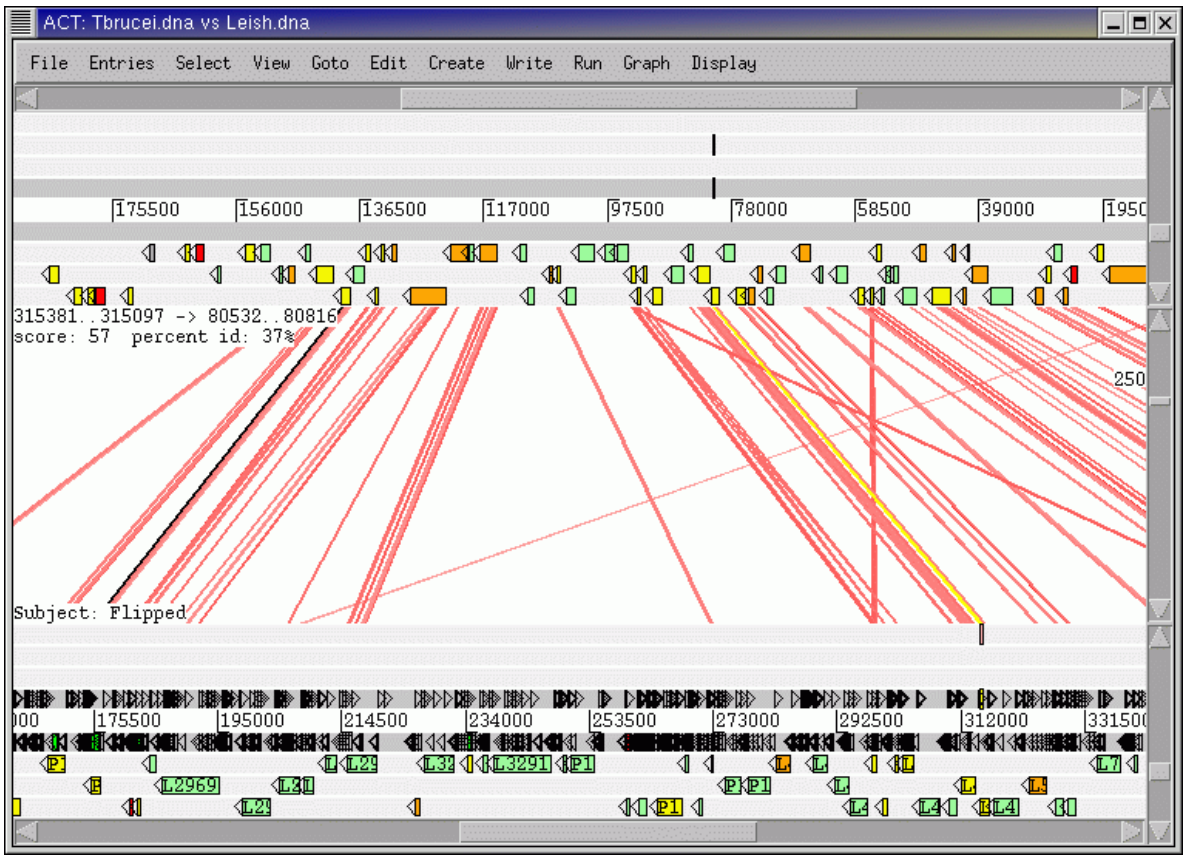
When you have determined where there is synteny, zoom in to the region for a detailed look. At this point you can add the annotation from the files called **Leish.embl** and **Tbrucei.embl**.

Can you see conserved gene order between the 2 species?

Can you see any region where similarity is broken up? Zoom in and look at some of the genes encoded within theses regions.

What are the predicted products of the genes assigned to these locations? View the details by clicking on the feature, and then select *`Edit selected feature'* from the *'Edit'* menu after selecting the appropriate CDS feature.

Can you identify any genes in one organism that don't appear to be predicted in the other? If so, add these to your annotation.



Exercise 4

Introduction

The quinic acid gene cluster (the *qut* cluster) is present among many filamentous fungi including including *Aspergillus fumigatus*, *Neurospora crassa*, *Aspergillus nudulans* and *Podospora anserina*. Although these fungi belong to the same fungal taxonomic family (Ascomycetes), they vary greatly in their biological characteristics. In this exercise you will be studying and comparing the organisation of *qut* gene cluster among these 4 fungi, using ACT.

Aim

By looking at a comparison of the annotated sequences of *N. crassa*, *A. fumigatus* and *A. nidulans* you will be able to first, add annotations to qut cluster genes in P. anserina sequence and second compare those genes that are found in all 4 organisms as well as spot the differences and study the synteny.

The files that you are going to need are:

1) N_crassa_qut.embl - sequence & annotated file for *N. crassa*
2) A_fum_qut.embl - sequence & annotation file for *A. fumigatus*
3) A_nid_qut.embl - sequence & annotation file for *A. nidulans* (artificially joined contig)
4) P_anserina_qut.embl - sequence & gene model file for *P. anserina* (without annotation)

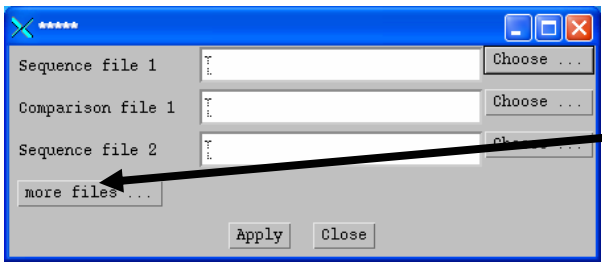
5) A_fum_N._crassa.comp - tblastx comparison file of *A. fumigatus* & *N. crassa*
6) A_fum_A_nid.comp - tblastx comparison file of *A. fumigatus* & *A. nidulans*
7) A_nid_P_anserina.comp - tblastx comparison file of *A. nidulans* & *P. anserina*
8) P_anserina_N_crassa.comp- tblastx comparison file of *P. anserina* & *N. crassa*.

First, open an ACT window and then open the annotation and the appropriate comparison files in the order of 1 – 5 – 2 – 6 – 3 – 7 – 4 – 8 – 1 (the numbers are designated above).

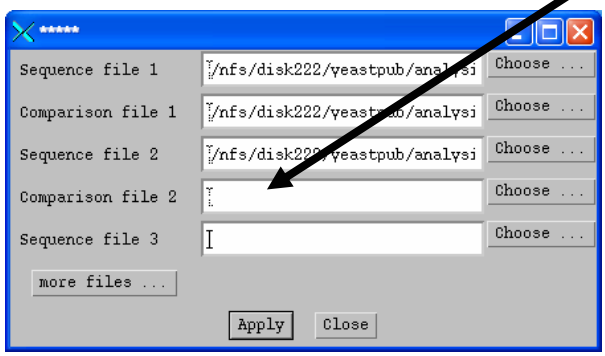
You will need to click on ‘more files’ to upload more than 2 sequences and the comparison flies.

Click on ‘apply’ after you have uploaded all the files.

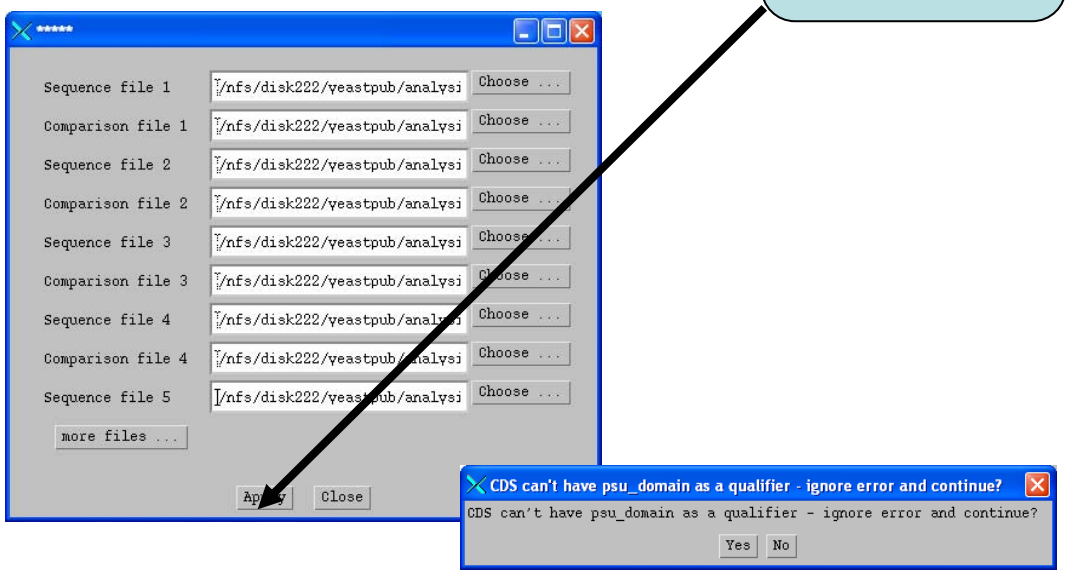
Upload the files in sequential order as described in the previous page



Click on here to load more files and select the appropriate file



Click on here to read all the files that you have selected.



Click on 'yes' if any small dialogue window appears while reading / opening the files.

Can you see any conserved gene order between the *A. fumigatus* & *A. nidulans* in the qut gene cluster?

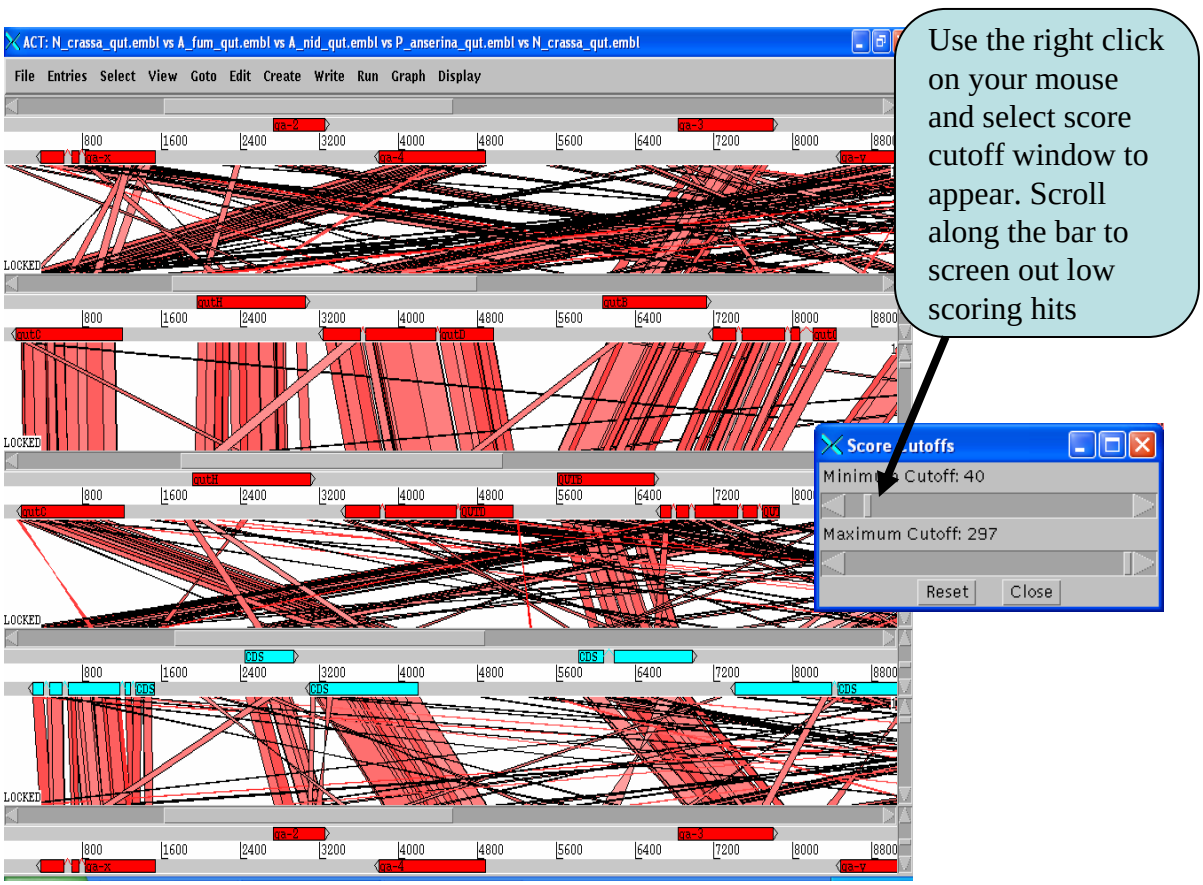
Can you obtain a clearer picture of the ACT 4-way comparison figure by filtering out the low scoring segments, using the blast score cut off feature which you have used previously.

Zoom in and look at some of the genes encoded within theses regions. View the details by clicking on the feature, and then select *`Edit selected feature'* from the *'Edit'* menu after selecting the appropriate CDS feature.

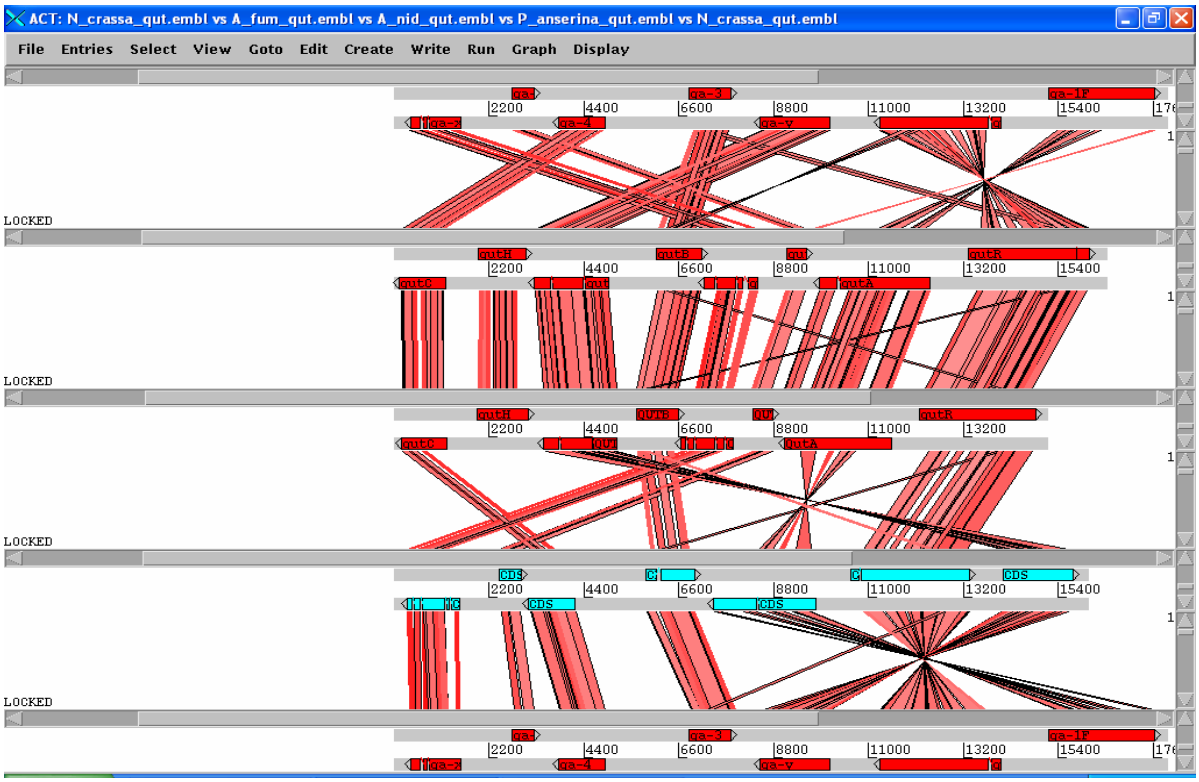
By comparing the blast similarity matches, assign your own annotation (gene product) to the predicted gene models (the blue genes) on the *P. anserina* gene model file.

Can you identify any gene NOT present in the qut cluster of ALL four fungi?

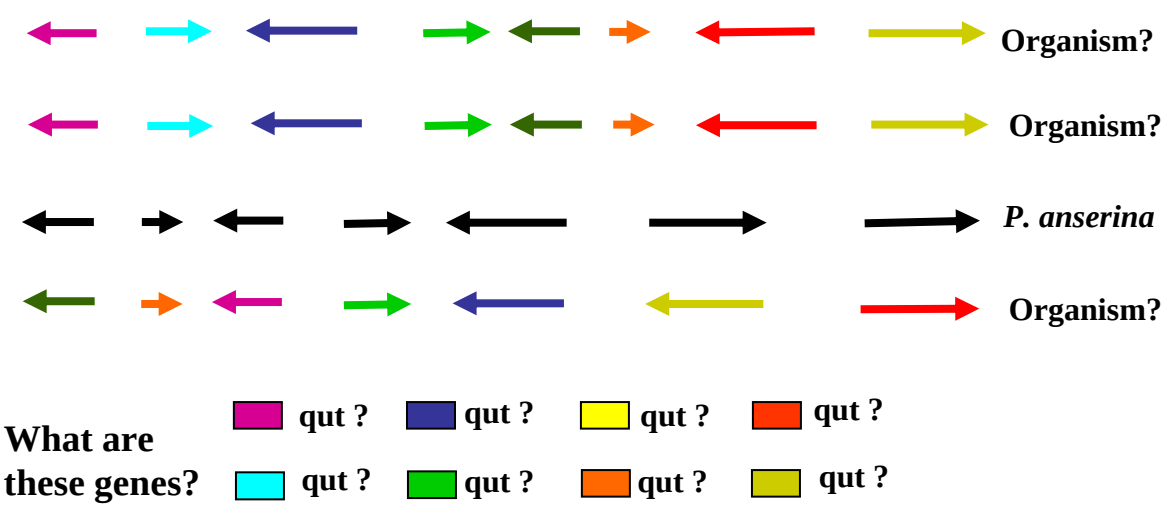
Note down the gene order (and direction of transcription) in each after you have completed annotation of the *P. anserina* gnes in the *qut* cluster.



After filtering out the low-scoring blast matches, you should be able to see a figure like the image below.



After comparing the arrangement of genes in the *qut* cluster in these fungi, do you agree with the schematic diagram (not in scale) below where each colour represents a specific type of gene in the quinic acid utilisation gene cluster and each set of clustered genes represents the *qut* cluster one of the organsims. Before you do this you need to annotate the *P. anserina* genes shown as black arrows.



Module 7
Generating ACT comparison files using BLAST

Introduction

In the previous module you used ACT to visualize pairwise BlastN or TBlastX comparisons between DNA sequences. In order to use ACT to investigate your own sequences of interest you will have to generate your own pairwise comparison files. ACT is written so that it will read the output of several different comparison file formats; these are outlined in appendix II. Two of the formats can be generated using Blast software freely downloadable from the NCBI (appendix X). Both Windows and Linux versions of the software are available which can be loaded onto a PC or Mac.

For the purposes of this module the NCBI Blast distribution software has already been installed locally and therefore ready to use. To give you an idea of how easy it is to download and install the software on a PC we have included a step-by-step guide in the appendixes (Appendix X). The example shown in appendix X is for downloading onto a PC with Windows XP. The exercises in this module are based on the Linux version of the Blast software. Although the operating systems are different, the command lines used to run the programs are the same. One of the main differences between the two operating systems is that in Windows the Blast program command line is run in the DOS Command Prompt window, whereas in Linux it is run from a Xterminal window.

Aims

The aim of this module is to demonstrate how you can generate you own comparison files for ACT from a stand-alone version of the Blast software. In this module you will use Blast to generate comparison files for sequences that you have downloaded from the EBI genomes web resource. A copy of the Blast software has been installed locally. You will run Blast from the command-line using two different programs from the NCBI Blast distribution to generate ACT-readable comparison file for two small sequences (plasmids), and for two large sequences (whole genomes).

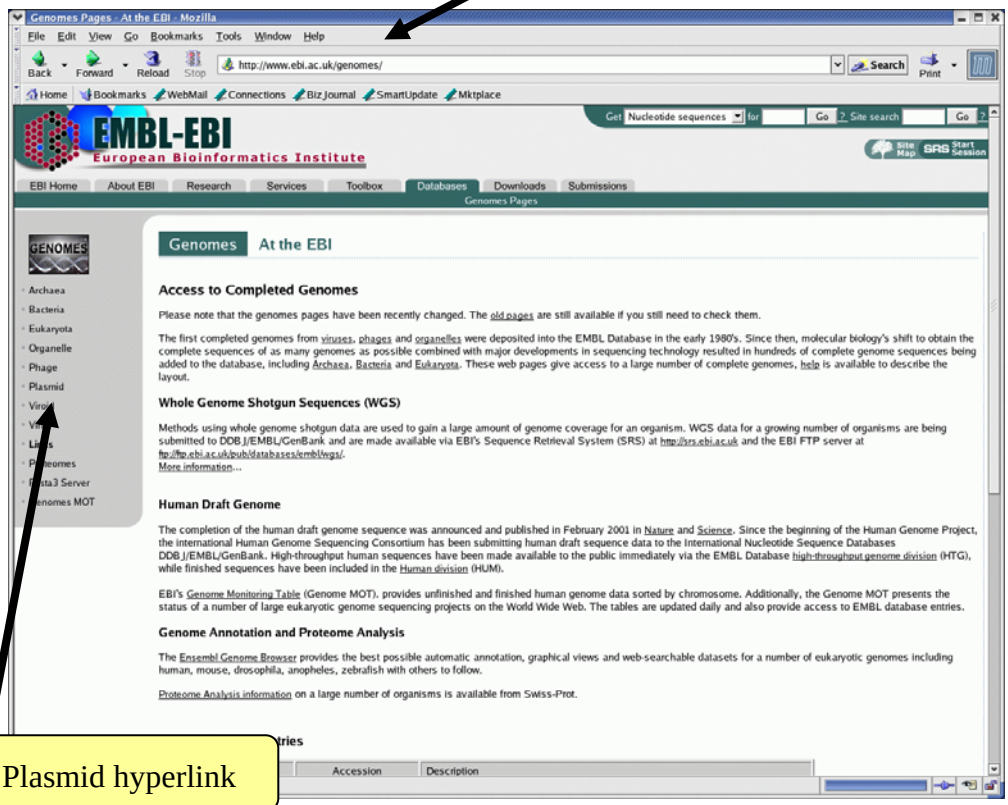
Exercise 1

In this exercise you are going to download two plasmid sequences in EMBL format from the EBI genomes web page. You are then going to use Artemis to write out the DNA sequences of both plasmids in FASTA format. These two FASTA format sequences will then be compared using BlastN to identify regions of DNA-DNA similarity and write out a ACT readable comparison file.

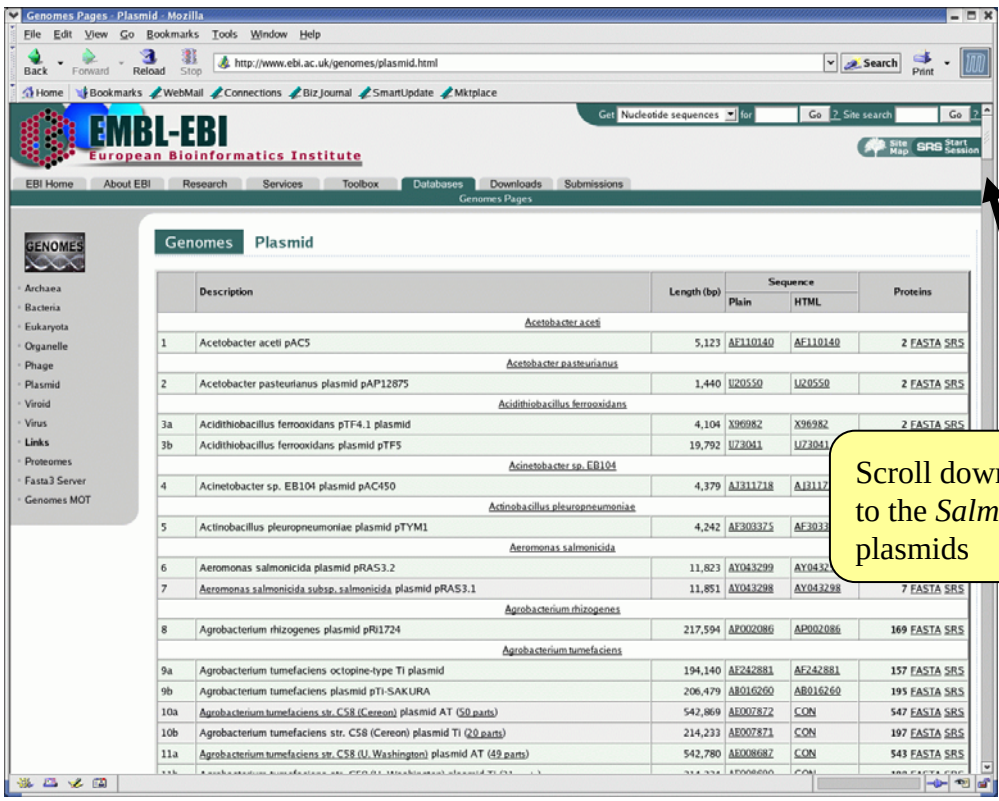
The plasmids chosen for this comparison are the multiple drug resistance incH1 plasmid pHCM1 from the sequenced strain of *Salmonella typhi* CT18 originally isolated in 1993, and R27, another incH1 plasmid first isolated from *S. typhi* in the 1960s.

Downloading the *S. typhi* plasmid sequences

Go to the EBI genomes web page (http://www.ebi.ac.uk/genomes)

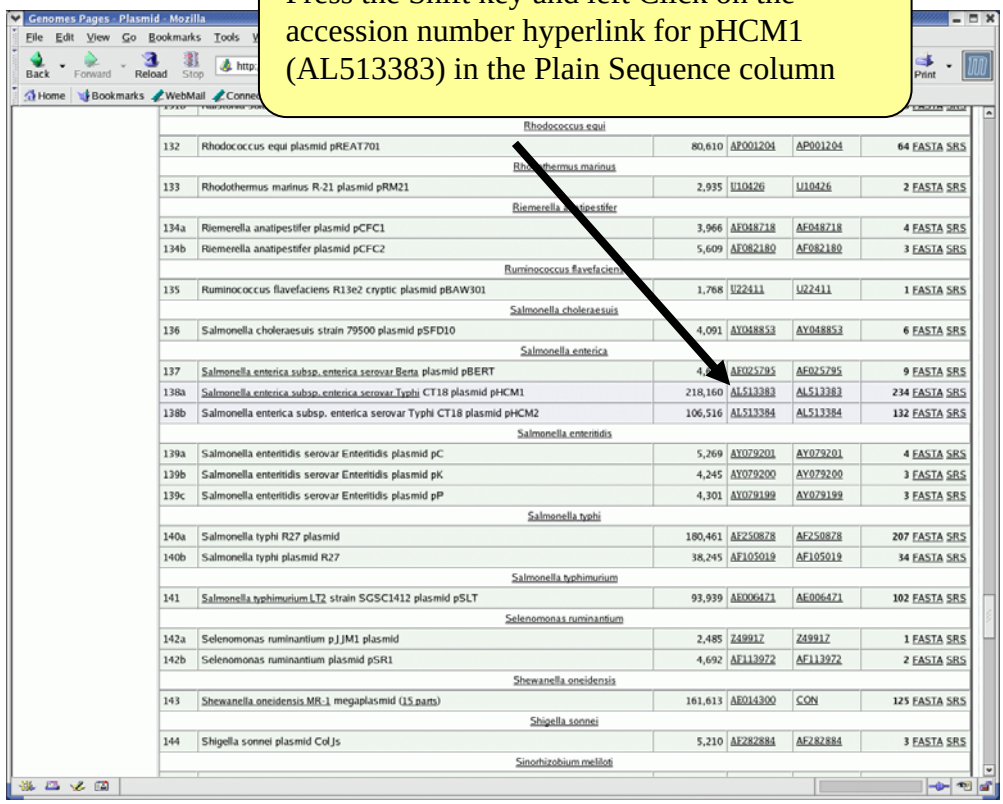


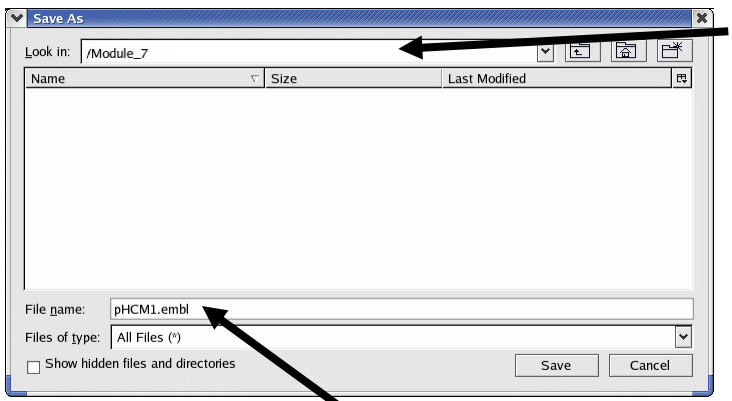
Click on the Plasmid hyperlink



Scroll down the page to the *Salmonella* plasmids

Press the Shift key and left Click on the accession number hyperlink for pHCM1 (AL513383) in the Plain Sequence column





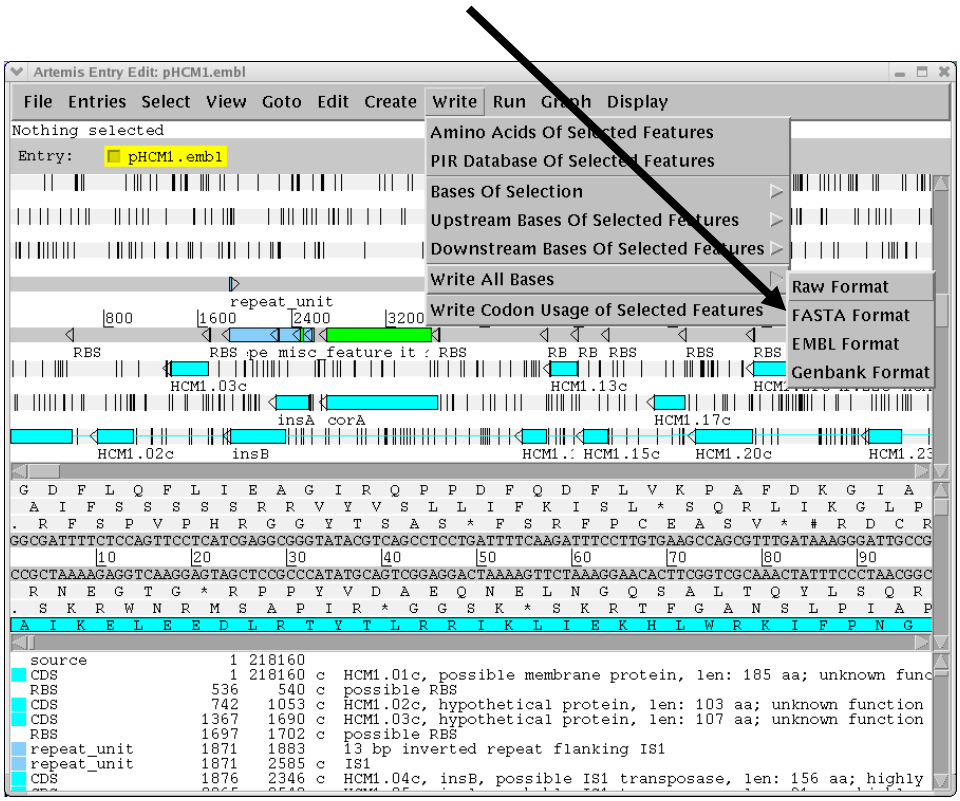
Save the EMBL sequence in the **Module_7** directory

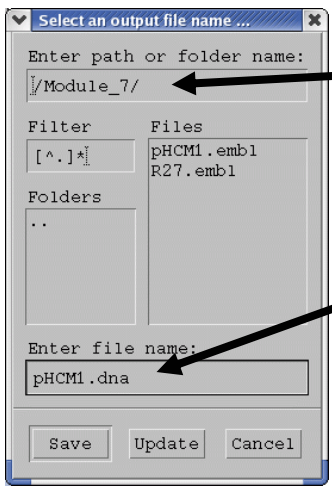
Save the file as pHCM1.embl

Repeat for the *Salmonella typhi* R27 plasmid (AF250878). Be careful when choosing the plasmid to download as there is also a *Salmonella typhi* plasmid R27 entry (AF105019), the one that you want is the larger of the two, 180,461 kb as opposed to 38,245 kb. Save as R27.embl.

In order to run BlastN you require two DNA sequences in FASTA format. The pHCM1 and R27 sequences previously downloaded from the EBI are EMBL format files, i.e. they contain protein coding information and the DNA sequence. In order to generate the DNA files in FASTA format, Artemis can be used as follows.

Load up the plasmid EMBL files in **Artemis** (each plasmid requires a separate Artemis window), select **Write**, **Write All Bases**, **FASTA format**.





Save the DNA sequence in the **Module_7** directory

Save as pHCM1.dna

Also do this for R27.embl

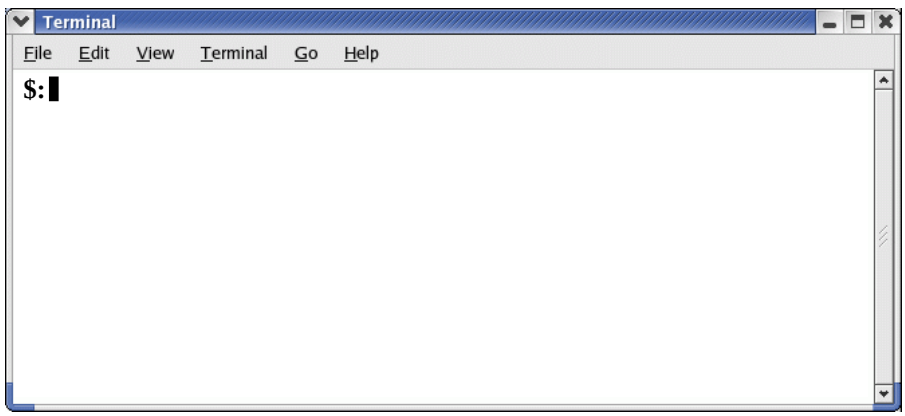
Running Blast

There are several programs in the Blast package that can be used for generating sequence comparison files. For a detailed description of the uses and options see the appropriate README file in the Blast software directory (see appendix X).

In order to generate comparison files that can be read into ACT you can use the **Blastall** program running either BlastN (DNA-DNA comparison) or TBlastX (translated DNA-translated DNA comparison) protocols.

As an example you will run a BlastN comparison on two relatively small sequences; the pHCM1 and R27 plasmids from *S. typhi*. In principle any DNA sequences in FASTA format can be used, although size becomes and issue when dealing with sequences such whole genomes of several Mb (see exercise 2 in this module). When obtaining nucleotide sequences from databases such as EMBL using a server such as SRS (http://srs.ebi.ac.uk), it is possible to specify that the sequences are in FASTA format.

To run the blast software you will need an Xterminal window like the one below. If you do not already have one opened, you can open a new window by clicking on the Xterminal icon on the menu bar at the bottom of your screen.



Make sure you are in the Module_7 directory. You should now see both the new FASTA files for the pHCM1 and R27 sequences in the Module_7 directory as well as their respective EMBL format files.
(Hint: You can use the **pwd** command to check the present working directory, the **cd** command to change directories, and the **ls** command will list the contents of the present working directory).

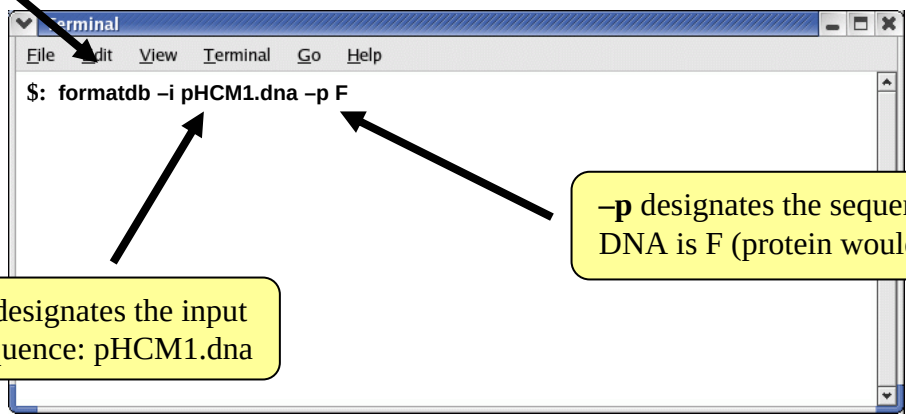
When comparing sequences in Blast, one sequence is designated as a **database** sequence, and the other the **query** sequence. Before you run Blast you have to format one of the sequences so that Blast recognises it as a database sequence. **formatdb** is a program that does this and comes as part of the NCBI Blast distribution.

You will treat pHCM1.dna as the **database** sequence and R27.dna as the **query** sequence

At the Command Prompt type:
formatdb –i pHCM1.dna –p F

Press **Return**

formatdb is the database format program



–p designates the sequence type: DNA is F (protein would be T)

–i designates the input sequence: pHCM1.dna

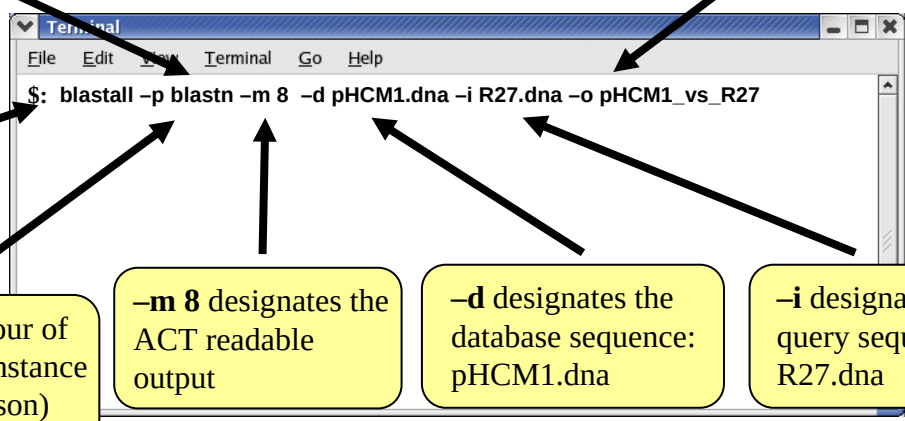
Now you can run the Blast on the two plasmid sequences. The program that you are going to use is **blastall**. In addition to the standard command line inputs we have to add an additional flag (**-m 8**) to the command line so that the Blast output can be read by ACT. This specifies that the output of Blast is in one line per entry format (see appendix II).

At the Command Prompt type:
blastall –p blastn –m 8 –d pHCM1.dna –i R27.dna –o pHCM1_vs_R27

Press **Return**

tblastx could be substituted here if a translated DNA-translated DNA comparison was required

–o designates the output file: pHCM1_vs_R27



blastall is the Blast program

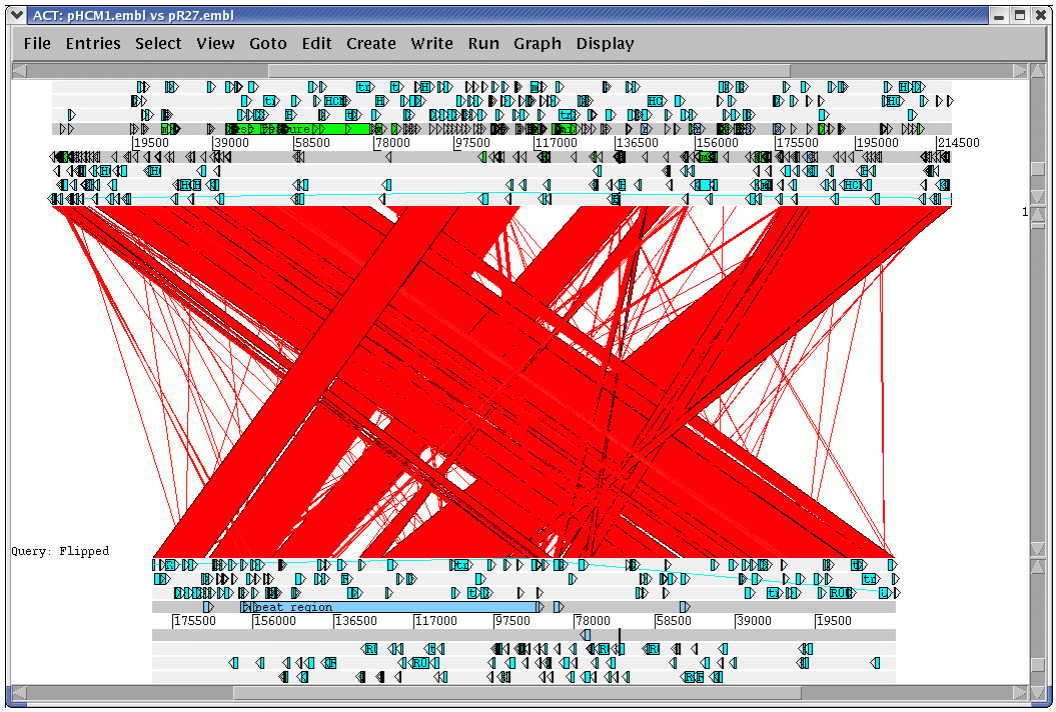
–p designates the flavour of Blast: **blastn** (in this instance a DNA-DNA comparison)

–m 8 designates the ACT readable output

–d designates the database sequence: pHCM1.dna

–i designates the query sequence: R27.dna

The pHCM1_vs_R27 comparison file can now be read into ACT along with the pHCM1.embl and R27.embl (or pHCM1.dna and R27.dna) sequence files.



The result of the BlastN comparison shows that there are regions of DNA shared between the plasmids; pHCM1 shares 169 kb of DNA at greater than 99% sequence identity with R27. Much of the additional DNA in the pHCM1 plasmid appears to have been inserted relative to R27 and encodes functions associated with drug resistance. What antibiotic resistance genes can you find in the pHCM1 plasmid that are not found in R27?

The two plasmids were isolated more than 20 years apart. The comparison suggest that there have been several independent acquisition events that are responsible for the multiple drug resistance seen in the more modern *S. typhi* plasmid.

Exercise 2

In the previous exercise you used BlastN to generate a comparison file for two relatively small sequences (>500,000 kb). In the next exercise we are going to use another program from NCBI Blast distribution, **megablast**, that can be used for nucleotide sequence alignment searches, i.e. DNA-DNA comparisons. If you are comparing large sequences such as whole genomes of several Mb, the **blastall** program is not suitable. The Blast algorithms will struggle with large DNA sequences and therefore the processing time to generate a comparison file will increase dramatically.

Megablast uses a different algorithm to Blast which is not as stringent which therefore makes the program faster. This means that it is possible to generate comparison files for genome sequences in a matter of seconds rather than minutes and hours.

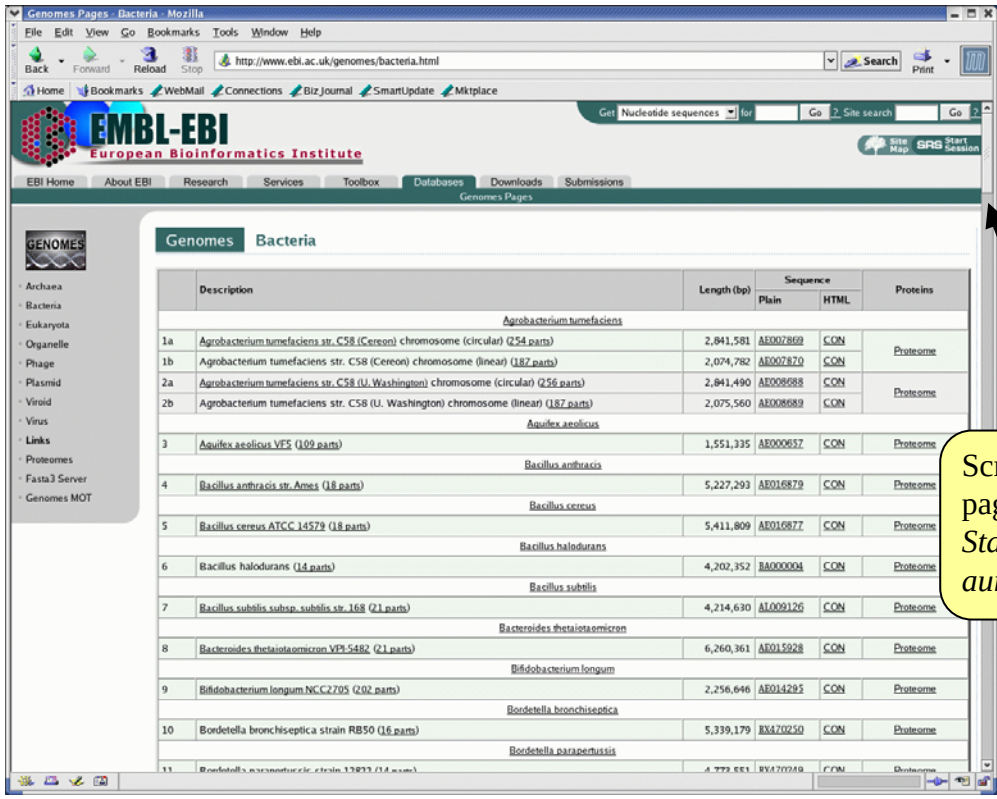
There are some drawbacks to using this program. Firstly, only DNA-DNA alignments (BlastN) can be performed using **megablast**, rather than translated DNA-DNA alignments (TBlastX) as can be using **blastall**. Secondly as the algorithm used is not as stringent, megablast is suited to comparing sequences with high levels of similarity such as genomes from the same or very closely related species.

In this exercise you are going to download two *Staphylococcus aureus* genome sequences from the EBI genomes web page and use Artemis to write out the FASTA format DNA sequences for both as before in exercise 1. These two FASTA format sequences will then be compared using **megablast** to identify regions of DNA-DNA similarity and write out an ACT readable comparison file.

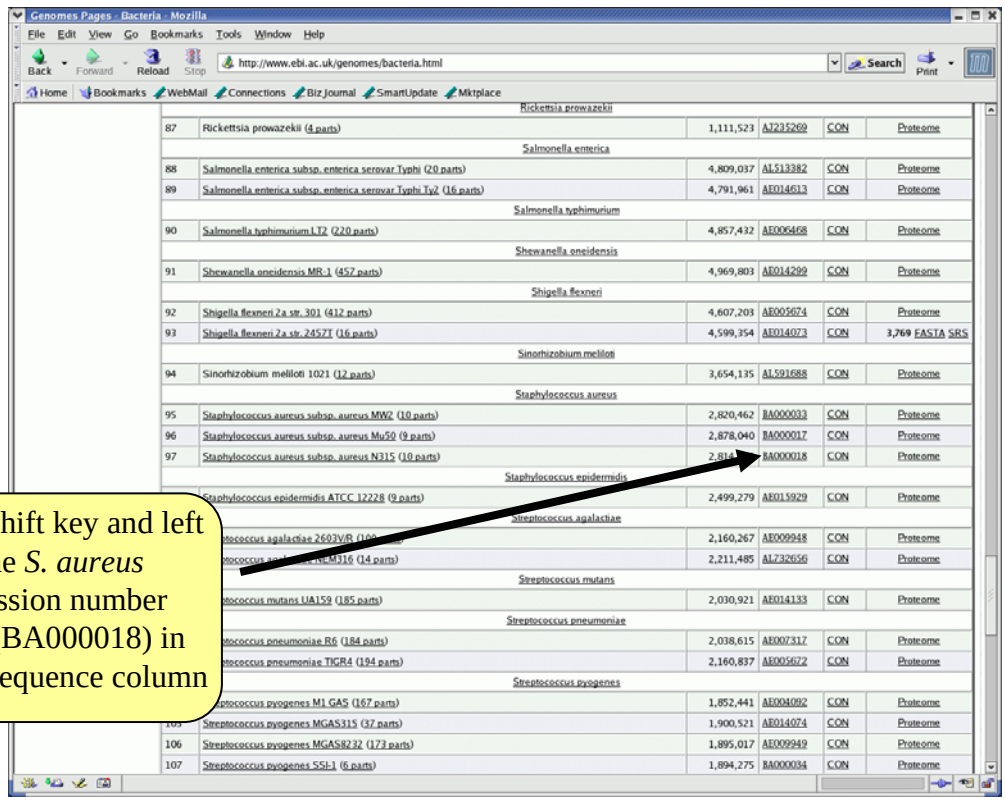
The genomes that have been chosen for this comparison are from a hospital-acquired methicillin resistant *S. aureus* (MRSA) strain N315 (BA000018), and a community-acquired MRSA strain MW2 (BA000033).

Downloading the *S. aureus* genomic sequences

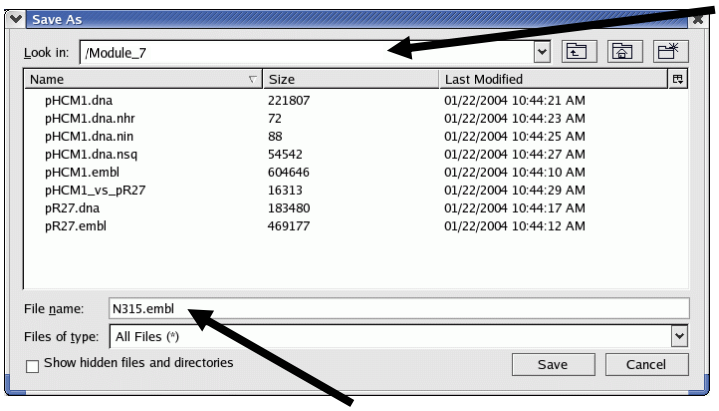
Go to the EBI genomes web page (http://www.ebi.ac.uk/genomes) as before in exercise 1, and click on the **Bacteria** hyperlink



Scroll down the page to the *Staphylococcus aureus* genomes



Press the Shift key and left Click on the *S. aureus* N315 accession number hyperlink (BA000018) in the Plain Sequence column



Save the EMBL sequence in the **Module_7** directory

Save the file as N315.embl

Repeat for the *S. aureus* MW2 genome (BA000033). Be careful when choosing the genome to download as there is another *S. aureus* genome entry for strain Mu50 (BA000017). Save as MW2.embl.

Generate DNA files in FASTA format using Artemis for both the genome sequences as previously done in exercise 1.

(Hint: In **Artemis** (each genome requires a separate Artemis window), select **Write**, **Write All Bases**, **FASTA format**).

Save the DNA sequences as N315.dna and MW2.dna for the respective genomes.

Running Blast

In the previous exercise you used the **blastall** program to run BlastN on two plasmid sequences. As the genome sequences are larger (~2.8 Mb) you are going to run **megablast**, another program from the NCBI Blast distribution that can generate comparison files in a format that ACT can read (see appendix II). For a detailed description of the uses and options in **megablast** see the megablast README file in the Blast software directory (appendix X).

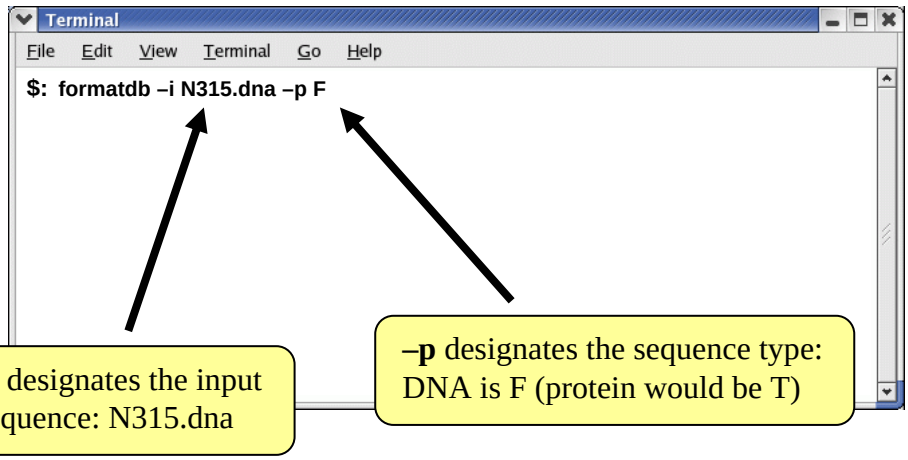
As before you will run the program from the command line in an Xterminal window.

Like Blast, **megablast** requires that one sequence is designated as a **database** sequence and the other the **query** sequence. Therefore one of the sequences has to be formatted so that Blast recognises it as a database sequence. This can be done as before using **formatdb.**

We will treat N315.dna as the **database** sequence and MW2.dna as the **query** sequence

At the Command Prompt type:
formatdb –i N315.dna –p F

Press **Return**



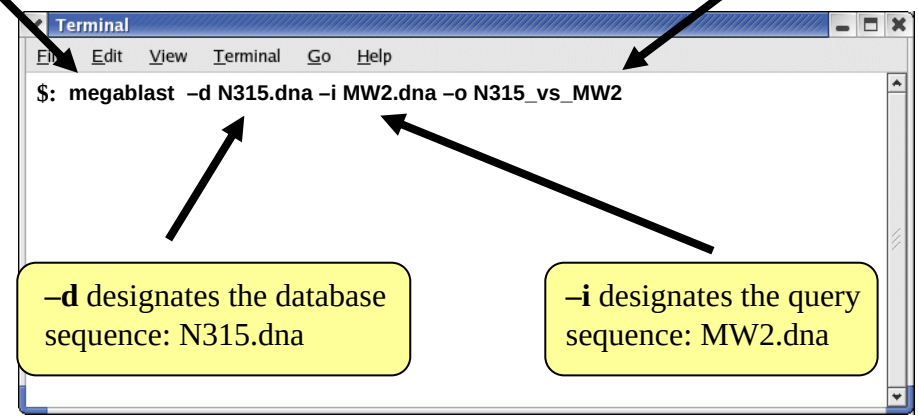
Now we can run the **megablast** on the two MRSA genome sequences. The default output format is one line per entry that ACT can read, therefore there is no need to add an additional flag to the command line (see appendix II).

At the Command Prompt type:
megablast –d N315.dna –i MW2.dna –o N315_vs_MW2

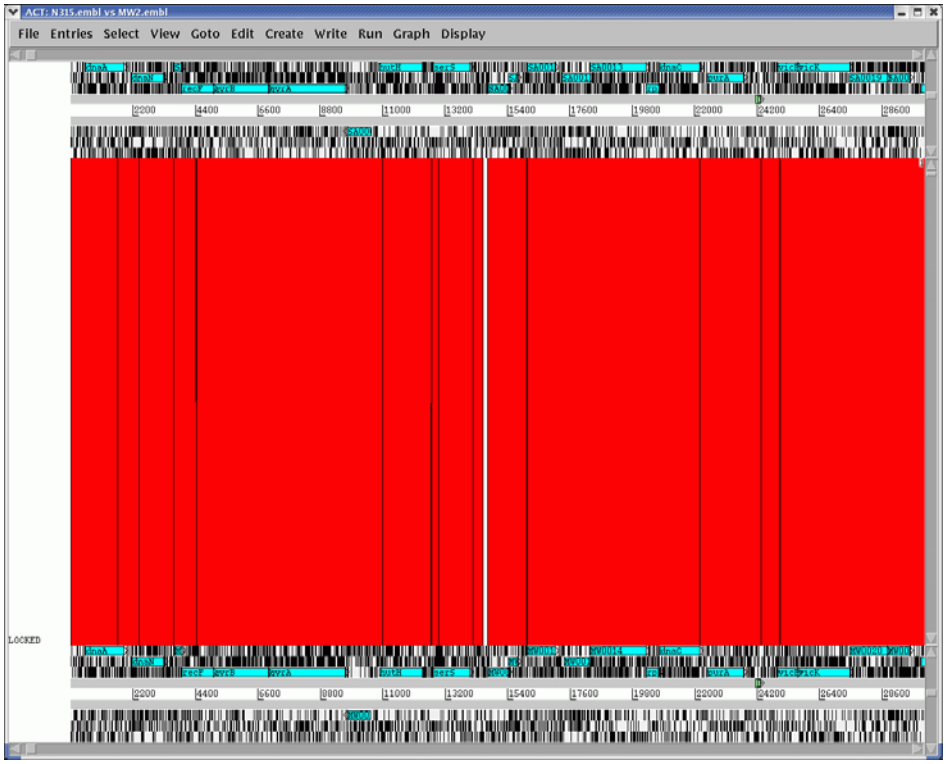
Press **Return**

megablast is the program

–o designates the output file: N315_vs_MW2



The N315_vs_MW2 comparison file can now be read into ACT along with the N315.embl and MW2.embl (or N315.dna and MW2.dna) sequence files.



A comparison of the N315 and MW2 genomes in ACT using the **megablast** comparison reveals a high level of synteny (conserved gene order). This is perhaps not unsurprising as both genomes belong to strains of the same species. Using results of comparisons like these it is possible to identify genomic differences that may contribute to the biology of the bacteria and also investigate mechanisms of evolution.

Both N315 and MW2 are MRSA, however N315 is associated with disease in hospitals, and MW2 causes disease in the community and is more invasive. Scroll rightward in both genomes to find the first large region of difference. Examine the annotation for the genes in these regions. What are the encoded functions associated with these regions? What significance does this have for the evolution of methicillin resistance in these two *S. aureus* strains from clinically distinct origins?

References

Abbot, J. C. et al. (2005) Bioinformatics 21(18) 3665-3666
Web AT – an online companion for the Artemis Comparison Tool

Carver et al. (2005)

Hacker, J., Blum-Oehler, G., Muhldorfer, I., and Tschape. (1997) Pathogenicity islands of virulent bacteria:structure, function and impact on microbial evolution. *Mol Microbiol* **23** 1089-97.

Berriman, M., and K. Rutherford (2003) Brief Bioinform **4** (2) 124-132
Viewing and annotating sequence data with Artemis.

Hacker, J., Blum-Oehler, G., Muhldorfer, I., and Tschape. (1997) Pathogenicity islands of virulent bacteria:structure, function and impact on microbial evolution. *Mol Microbiol* **23** 1089-97.

Majoros et al. (2003) *Nucleic Acids Research* **31** (13) 3601-3604
GlimmerM, Exonomy and Unveil: three *ab initio* eukaryotic genefinders

Parkhill, J. (2002) *Methods in Microbiology* **33** 1-26
Annotation of Microbial Genomes

Rutherford et al.(2000) *Bioinformatics* **16** (10) *944-945*
Artemis: sequence visualization and annotation

References

Hacker, J., Blum-Oehler, G., Muhldorfer, I., and Tschape. (1997) *Mol. Microbiol.* **23**: 1089-97.
Pathogenicity islands of virulent bacteria:structure, function and impact on microbial evolution.

Berriman, M., and K. Rutherford (2003) Brief Bioinform. **4** (2): 124-132.
Viewing and annotating sequence data with Artemis.

Majoros *et al.* (2003) *Nucleic Acids Research* **31** (13): 3601-3604.
GlimmerM, Exonomy and Unveil: three *ab initio* eukaryotic genefinders

Parkhill, J. (2002) *Methods in Microbiology* **33:** 1-26.
Annotation of Microbial Genomes

Rutherford *et al*. (2000) *Bioinformatics* **16** (10): 944-945.
Artemis: sequence visualization and annotation

Abbot, J. C. *et al.* (2005) Bioinformatics **21**(18): 3665-3666.
Web ACT – an online companion for the Artemis Comparison Tool

Carver, T. J. *et al.* (2005) Bioinformatics **21**(16): 3422-3423.
ACT: the Artemis Comparison Tool

Appendices

Appendix I: Artemis minimum hardware and software requirements.

Artemis and ACT will, in general, work well on any standard modern machine and with most common operating systems. It is currently used on many different varieties of UNIX and Linux systems as well as Apple Macintosh and Microsoft Windows systems.
Note that the ability to run external programs (such as BLAST and FASTA) from within Artemis and ACT is available only on UNIX and Linux systems.
Minimum memory requirements for people working on whole genomes are approximately 128 megabytes for Artemis and 128 megabytes per genome for ACT. Analysis of cosmid sized sequences can comfortably be achieved with less memory.

Appendix II: ACT comparison files

ACT supports three different comparison file formats:

1) BLAST version 2.2.2 output: The blastall command must be run with the -m 8 flag which generates one line of information per HSP.
2) MEGABLAST output: ACT can also read the output of MEGABLAST, which is part of the NCBI blast distribution.
3) MSPcrunch output: MSPcrunch is program for UNIX and GNU/Linux systems which can post-process BLAST version 1 output into an easier to read format. ACT can only read MSPcrunch output with the -d flag.

Here is an example of an ACT readable comparison file generated by MSPcrunch -d.

```
1399 97.00 940 2539 sequence1.dna 1 1596 AF140550.seq
1033 93.00 9041 10501 sequence1.dna 9420 10880 AF140550.seq
828 95.00 6823 7890 sequence1.dna 7211 8276 AF140550.seq
773 94.00 2837 3841 sequence1.dna 2338 3342 AF140550.seq
```

The columns have the following meanings (in order): score, percent identity, match start in the query sequence, match end in the query sequence, query sequence name, subject sequence start, subject sequence end, subject sequence name.

The columns should be separated by single spaces.

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

1 – Feature 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 coding sequence.
➔ RBS: Ribosomal binding site
➔ misc_feature: Miscellaneous feature in the DNA
➔ rRNA: Ribosomal RNA
➔ repeat_region
➔ repeat_unit
➔ stem_loop
➔ tRNA: Transfer RNA

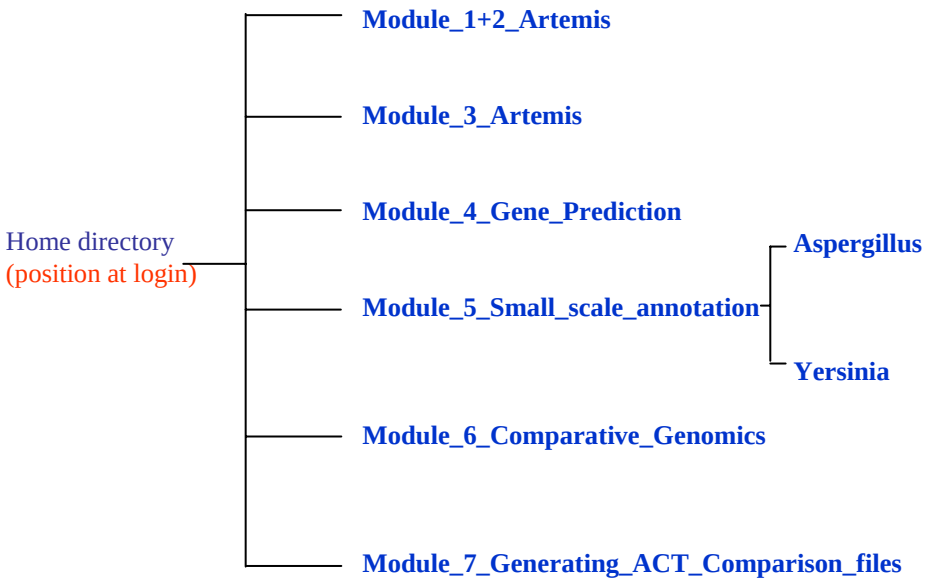
2 – Qualifiers: They describe features with protein coordinates. Once marked they appear in the lower part of the Artemis window. They describe the gene whose coordinates appear in the 'location' part of the editing window. The ones we commonly use for annotation at the Sanger Institute are:

➔ Class: Classification scheme we use "in-house" developed from Monica Riley's MultiFun assignments (see Appendix VI).
➔ 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 to have been mutated.

The list of keys and qualifiers accepted by EMBL in sequence/annotation submission files are list at the following web page:
http://www3.ebi.ac.uk/Services/WebFeat/

Appendix IV: Schematic of workshop files and directories

Key:
Directories and subdirectories



Appendix V: 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.sanger.ac.uk/Software/Pfam/index.shtml
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/
Glimmer	http://www.tigr.org/software/glimmer/
Orpheus	http://pedant.gsf.de/orpheus

Appendix VI: Prokaryotic Protein Classification Scheme used within the PSU

This scheme was adapted for in-house use from the Monica Riley's protein classification <http://genprotec.mbl.edu/riley-lab.html>).
More classes can be added depending on the microorganism that is being annotated (e.g secondary metabolites, sigma factors (ECF or non-ECF), etc).

0.0.0 Unknown function, no known homologs
0.0.1 Conserved in Escherichia coli
0.0.2 Conserved in organism other than Escherichia coli

1.0.0 Cell processes
- 1.1.1 Chemotaxis and mobility
- 1.2.1 Chromosome replication
- 1.3.1 Chaperones

1.4.0 Protection responses
- 1.4.1 Cell killing
- 1.4.2 Detoxification
- 1.4.3 Drug/analog sensitivity
- 1.4.4 Radiation sensitivity

1.5.0 Transport/binding proteins
- 1.5.1 Amino acids and amines
- 1.5.2 Cations
- 1.5.3 Carbohydrates, organic acids and alcohols
- 1.5.4 Anions
- 1.5.5 Other

1.6.0 Adaptation
- 1.6.1 Adaptations, atypical conditions
- 1.6.2 Osmotic adaptation
- 1.6.3 Fe storage

1.7.1 Cell division

2.0.0 Macromolecule metabolism

2.1.0 Macromolecule degradation
- 2.1.1 Degradation of DNA
- 2.1.2 Degradation of RNA
- 2.1.3 Degradation of polysaccharides
- 2.1.4 Degradation of proteins, peptides, glycoproteins

2.2.0 Macromolecule synthesis, modification
- 2.2.01 Amino acyl tRNA synthesis; tRNA modification
- 2.2.02 Basic proteins - synthesis, modification
- 2.2.03 DNA - replication, repair, restriction./modification
- 2.2.04 Glycoprotein
- 2.2.05 Lipopolysaccharide
- 2.2.06 Lipoprotein
- 2.2.07 Phospholipids
- 2.2.08 Polysaccharides - (cytoplasmic)
- 2.2.09 Protein modification
- 2.2.10 Proteins - translation and modification
- 2.2.11 RNA synthesis, modif., DNA transcrip.
- 2.2.12 tRNA

3.0.0 Metabolism of small molecules

3.1.0 Amino acid biosynthesis
- 3.1.01 Alanine
- 3.1.02 Arginine
- 3.1.03 Asparagine
- 3.1.04 Aspartate
- 3.1.05 Chorismate
- 3.1.06 Cysteine
- 3.1.07 Glutamate
- 3.1.08 Glutamine
- 3.1.09 Glycine
- 3.1.10 Histidine
- 3.1.11 Isoleucine
- 3.1.12 Leucine
- 3.1.13 Lysine
- 3.1.14 Methionine
- 3.1.15 Phenylalanine
- 3.1.16 Proline
- 3.1.17 Serine
- 3.1.18 Threonine
- 3.1.19 Tryptophan
- 3.1.20 Tyrosine
- 3.1.21 Valine

3.2.0 Biosynthesis of cofactors, carriers
- 3.2.01 Acyl carrier protein (ACP)
- 3.2.02 Biotin
- 3.2.03 Cobalamin
- 3.2.04 Enterochelin
- 3.2.05 Folic acid
- 3.2.06 Heme, porphyrin
- 3.2.07 Lipoate
- 3.2.08 Menaquinone, ubiquinone
- 3.2.09 Molybdopterin
- 3.2.10 Pantothenate
- 3.2.11 Pyridine nucleotide
- 3.2.12 Pyridoxine
- 3.2.13 Riboflavin
- 3.2.14 Thiamin
- 3.2.15 Thioredoxin, glutaredoxin, glutathione
- 3.2.16 biotin carboxyl carrier protein (BCCP)

Appendix VI (cont):

3.3.0 Central intermediary metabolism
- 3.3.01 2'-Deoxyribonucleotide metabolism
- 3.3.02 Amino sugars
- 3.3.03 Entner-Douderoff
- 3.3.04 Gluconeogenesis
- 3.3.05 Glyoxylate bypass
- 3.3.06 Incorporation metal ions
- 3.3.07 Misc. glucose metabolism
- 3.3.08 Misc. glycerol metabolism
- 3.3.09 Non-oxidative branch, pentose pathway
- 3.3.10 Nucleotide hydrolysis
- 3.3.00 other
- 3.3.11 Nucleotide interconversions
- 3.3.12 Oligosaccharides
- 3.3.13 Phosphorus compounds
- 3.3.14 Polyamine biosynthesis
- 3.3.15 Pool, multipurpose conversions of intermed. metabol'm
- 3.3.16 S-adenosyl methionine
- 3.3.17 Salvage of nucleosides and nucleotides
- 3.3.18 Sugar-nucleotide biosynthesis, conversions
- 3.3.19 Sulfur metabolism
- 3.3.20 Amino acids

3.4.0 Degradation of small molecules
- 3.4.1 Amines
- 3.4.2 Amino acids
- 3.4.3 Carbon compounds
- 3.4.4 Fatty acids
- 3.4.5 Other
- 3.4.0 ATP-proton motive force

3.5.0 Energy metabolism, carbon
- 3.5.1 Aerobic respiration
- 3.5.2 Anaerobic respiration
- 3.5.3 Electron transport
- 3.5.4 Fermentation
- 3.5.5 Glycolysis
- 3.5.6 Oxidative branch, pentose pathway
- 3.5.7 Pyruvate dehydrogenase
- 3.5.8 TCA cycle

3.6.0 Fatty acid biosynthesis
- 3.6.1 Fatty acid and phosphatidic acid biosynthesis

3.7.0 Nucleotide biosynthesis
- 3.7.1 Purine ribonucleotide biosynthesis
- 3.7.2 Pyrimidine ribonucleotide biosynthesis

4.0.0 Cell envelop
- 4.1.0 Periplasmic/exported/lipoproteins
- 4.1.1 Inner membrane
- 4.1.2 Murein sacculus, peptidoglycan
- 4.1.3 Outer membrane constituents
- 4.1.4 Surface polysaccharides & antigens
- 4.1.5 Surface structures

4.2.0 Ribosome constituents
- 4.2.1 Ribosomal and stable RNAs
- 4.2.2 Ribosomal proteins - synthesis, modification
- 4.2.3 Ribosomes - maturation and modification

5.0.0 Extrachromosomal

5.1.0 Laterally acquired elements
- 5.1.1 Colicin-related functions
- 5.1.2 Phage-related functions and prophages
- 5.1.3 Plasmid-related functions
- 5.1.4 Transposon-related functions

6.0.0 Global functions
- 6.1.1 Global regulatory functions

7.0.0 Not classified (included putative assignments)

7.1.1 DNA sites, no gene product

7.2.1 Cryptic genes

Appendix VII: List of colour codes

0 (white) - Pathogenicity/Adaptation/Chaperones
1 (dark grey) - energy metabolism (glycolysis, electron transport etc.)
2 (red) - Information transfer (transcription/translation + DNA/RNA modification)
3 (dark green) - Surface (IM, OM, secreted, surface structures
4 (dark blue) - Stable RNA
5 (Sky blue) - Degradation of large molecules
6 (dark pink) - Degradation of small molecules
7 (yellow) - Central/intermediary/miscellaneous metabolism
8 (light green) - Unknown
9 (light blue) - Regulators
10 (orange) - Conserved hypo
11 (brown) - Pseudogenes and partial genes (remnants)
12 (light pink) - Phage/IS elements
13 (light grey) - Some misc. information e.g. Prosite, but no function

Appendix VIII: 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 IX Splice site information

Gere	No.	Exon | Intron | Exon	Size (bp)
41-3	1	GAA|**GTA**CACA..CCTTCTTTTTCCATATT**TAG**|CAA	152
	2	AAT|**GTT**AAAA...TTTTTTTTTTTAAACT**TAG**|CCG	208
	3	GAG|**GTA**AGAA...ATTCATTATATATTTA**TAG**|GGA	86
	4	TCG|**GTA**TGGA...TTTTGAAATACTTCCT**CAG**|TTA	152
	5	ACT|**GTA**ATAT..TTTTTTTTTTTATTTCC**TAG**|ATG	112
	6	CAG|**GTA**AATA..ATAATGACATTTTGATA**CAG**|ATT	120
	7	AAT|**GTA**CATT..TTATTTTTATTTATTTA**TAG**|AAA	81
	8	TAG|**GTA**TTTG..ATATTTTTTACTTATGA**TAG**|TTA	96
RhopH3	1	AGG|**GTA**ATAT..TTTATTTTATTTTTTTT**TTA**|TTT	150
	2	GGA|**GTA**AGAG..TTTTTATTATTTTATTG**TAG**|TCC	442
	3	GGA|**GTA**AGAG..TTTTTATTATTTTATTG**TAG**|TCC	199
	4	CAG|**GTA**YGCT..TTTAATTTTTTTTTCCT**TCA**|TCA	160
	5	AAA|**GTA**AGAA..TATTTTTTTACAATTTT**TAG**|TTC	206
	6	AAG|**GTA**AAAG..TTTTTTTTTTTTTGTTT**CAG**|TTT	142
RNA pol III	1	CAG|**GTA**CATA..TTTTTTTTTTTTTTTTTT**TAG**|GTG	158
	2	CAA|**GTA**ATTA..TATATTTTATTTTTTCT**TAG**|GTT	113
	3	TAC|**GTT**AGTT..TTTTTTTTTTTTTTTTT**TAG**|TGG	169
	4	ATT|**GTA**AGTT..TATTTTTTTTTTTTTTTT**TAG**|TGA	112
SERA	1	TGT|**GTA**AGAA..TTGTCATTATTTTTTTT**TAG**|GTG	158
	2	AAA|**GTA**TAAA..TTTATTTATTTTTTTTT**TAG**|ATA	175
	3	CAG|**GTA**AATA..TTTTAATTTTTTTGTTT**TAG**|AAA	129
SERP H	1	CTG|**GTT**TGTC..CATATATTTCTTTATTT**TAG**|ATA	345
	2	AGA|**GTA**AAAA..TTTCTTATATTTTCTTT**TAG**|GTG	92
	3	CTG|**GTT**TGTC..CATATATTTCTTTATTT**TAG**|ATA	116
Ag15	1	ATG|**GTA**AGAG..TATTTTTGATACCTTTA**TAG**|AGT	214
	2	AAA|**GTA**ATTA..CAATCATATTAACACAA**AAG**|ATG	280
PfGPx	1	GAG|**GTA**TACA..TTATTATTCCCTTGCTT**TAG**|ATC	208
	2	TCG|**GTT**AGTA..TATTTATCATTTTTTTC**CAG**|ATG	168
Calmodulin	1	GAA|**GTA**AATC..TTTTTTATTTTTCTCAT**TAG**|CTA	480
PfPK1	1	TAG|**GTG**TGTT..TCATTACATTTTTACCT**TAG**|GAT	101
MESA	1	TTA|**GTA**AGTT..CGTAATATATTTTTTTT**TAG**|GAT	122
Aldolase	1	ATG|**GTA**AGAA..TATTTTTATATTTTTTT**TAG**|GCT	452
KAHRP	1	AAC|**GTA**AGTT..TTATTTTTTTTTTCATA**TAG**|TGC	430
GBPH2	1	TTG|**GTA**TGCC..TTTGTATTATTTAATTT**TAG**|AAT	157
GBP	1	TTG|**GTA**TG....TGTGTATTGTTTATTTT**TAG**|AAT	179
FIRA	1	TGT|**GTA**AGGA..TTTTTATATTTTTTCTT**TAG**|CGA	175
GARP	1	AAG|**GTA**ACAA..TATATGTATTTTTTTTT**TAG**|TGC	214

↑ **Donor motif** ↑ **Acceptor motif**

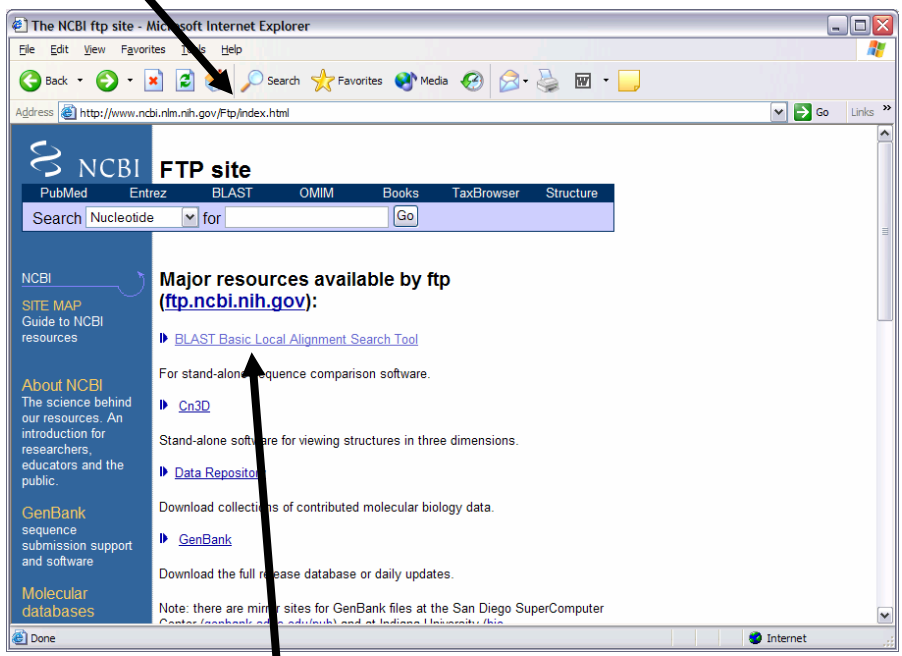
The splice acceptor and donor sequences for several *P. falciparum* genes: adapted from Coppel and Black(1998). In "Malaria:Parasite Biology, Pathogenesis and Protection", I.W. Sherman (ed.); ASM Press; Washington DC; pp185-202

Appendix X: Downloading and installing BLAST on a Windows PC

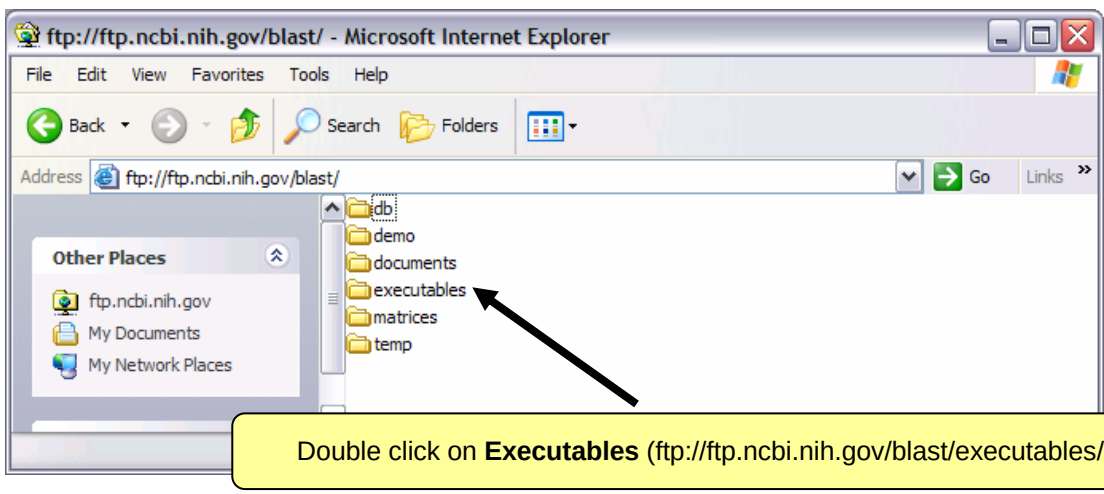
The following pages describe downloading BLAST onto a computer running Windows XP. Downloading onto computers with other versions of Windows should be essentially the same but the windows will look different to the screen shots used here.

Go to NCBI home page (http://www.ncbi.nlm.nih.gov/)

Scroll to bottom, Click on **FTP Site** (left hand side of the screen; http://www.ncbi.nlm.nih.gov/Ftp/index.html)

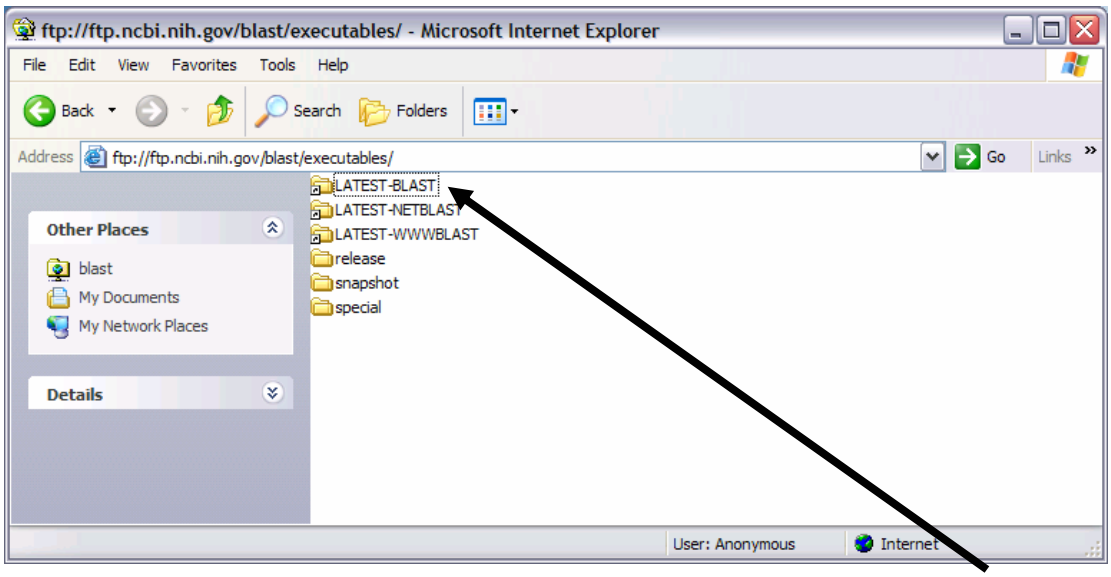


Click on **BLAST Basic Local Alignment Search Tool** (ftp://ftp.ncbi.nih.gov/blast/)

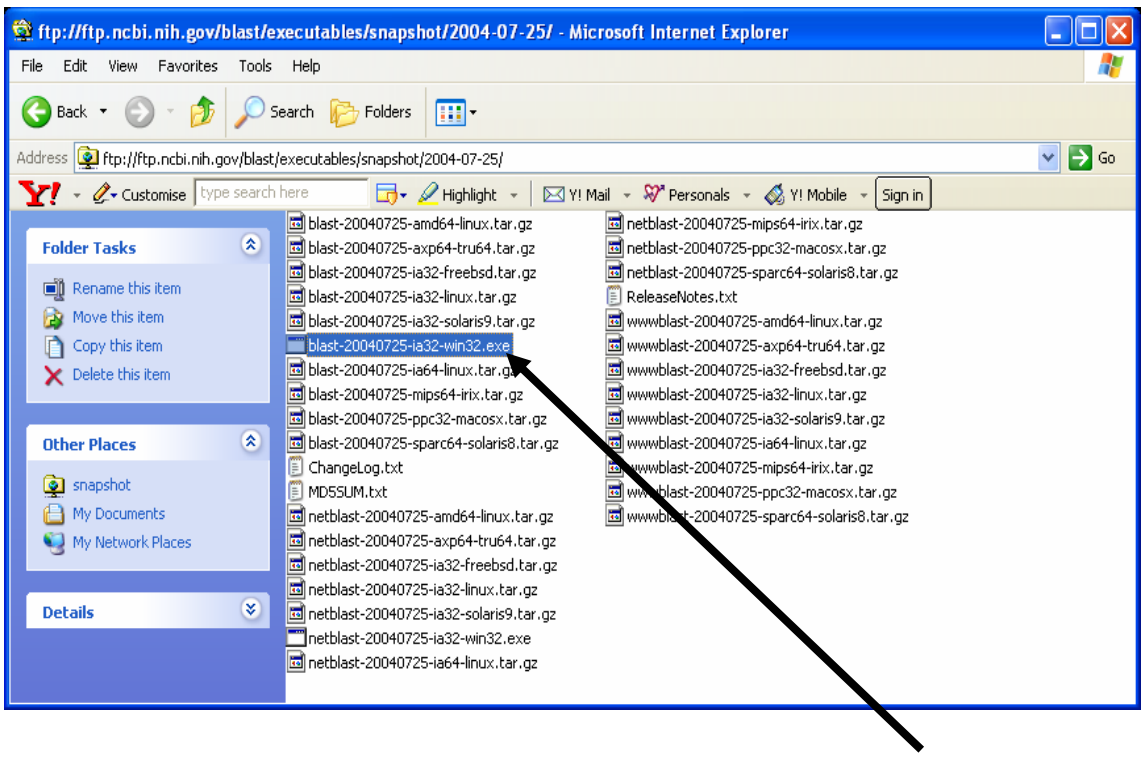


Double click on **Executables** (ftp://ftp.ncbi.nih.gov/blast/executables/)

This page may appear slightly different if you are using Netscape



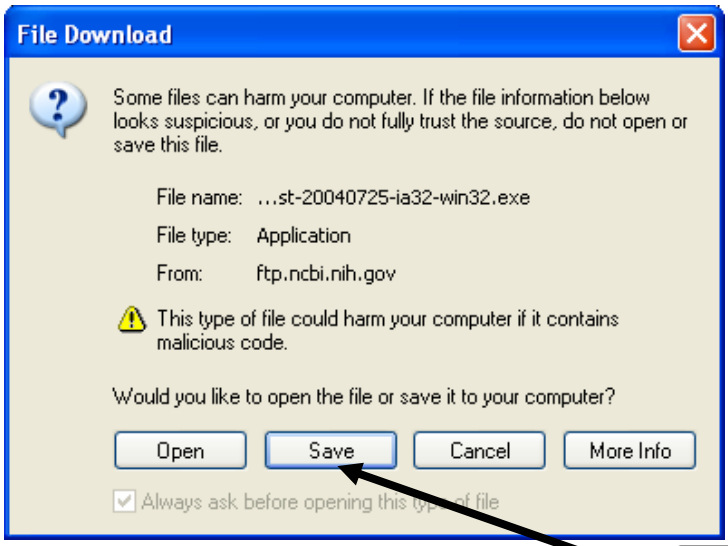
Double click on the **LATEST-BLAST** shortcut



Double click on **blast-20040725-ia32-win32.exe**

Blast-20040725-ia32win32.exe is the blast exe file for windows

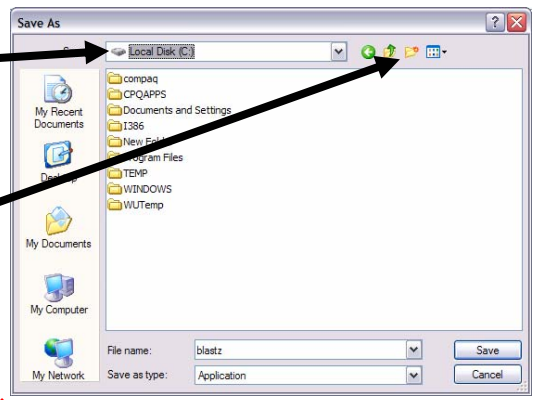
You now need to save the **blast-20040725-ia32-win32.exe** file in a new directory, blast, on to the hard drive of your PC

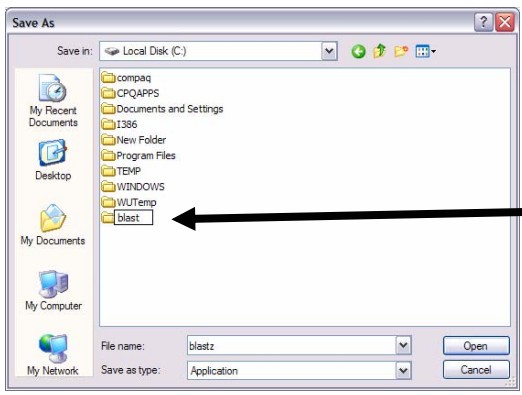


Click on **Save**

Click on **local disk C:**

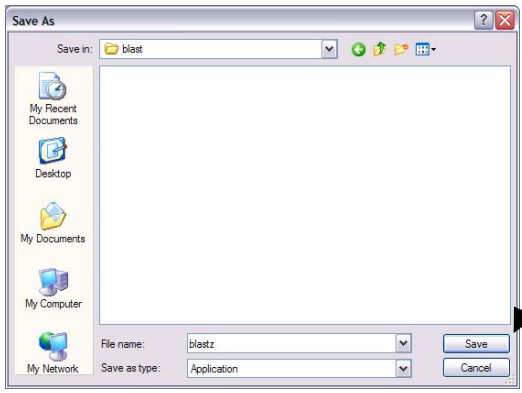
Click on **new directory icon**





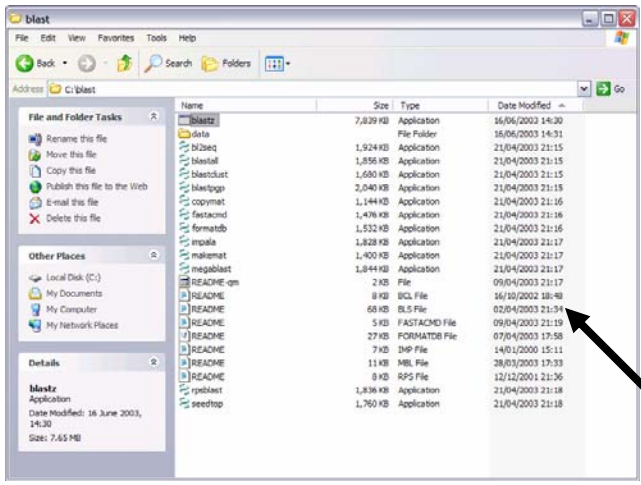
Type **blast** in the name box, press **Enter** key.

Double click on the new **blast** directory



Click on **Save**

Once downloaded view the contents of the blast directory by clicking on the open folder button



blast-2.2.6-ia32-win32.exe is a compressed file that contains a host of other files.

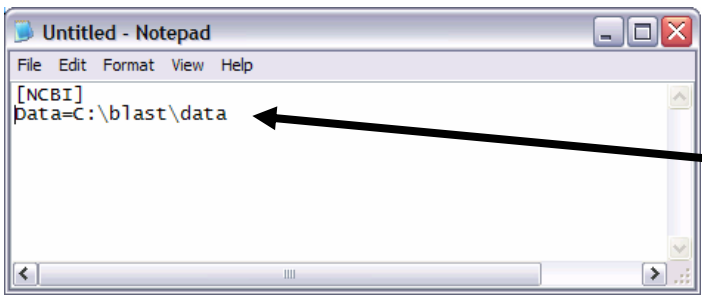
Now double click on the **blast-2.2.6-ia32-win32.exe** file to extract and unpack the rest of the BLAST download files

Included in the directory that has now been unpacked are several README files that describe the various programs in the BLAST software package. These files also provide descriptions of the command line options that you can set when you run the programs. To read these files double click on the icon or view them in notepad.

The **README.BLS** file contains details of the main BLAST program and how to format DNA sequences prior to running BLAST

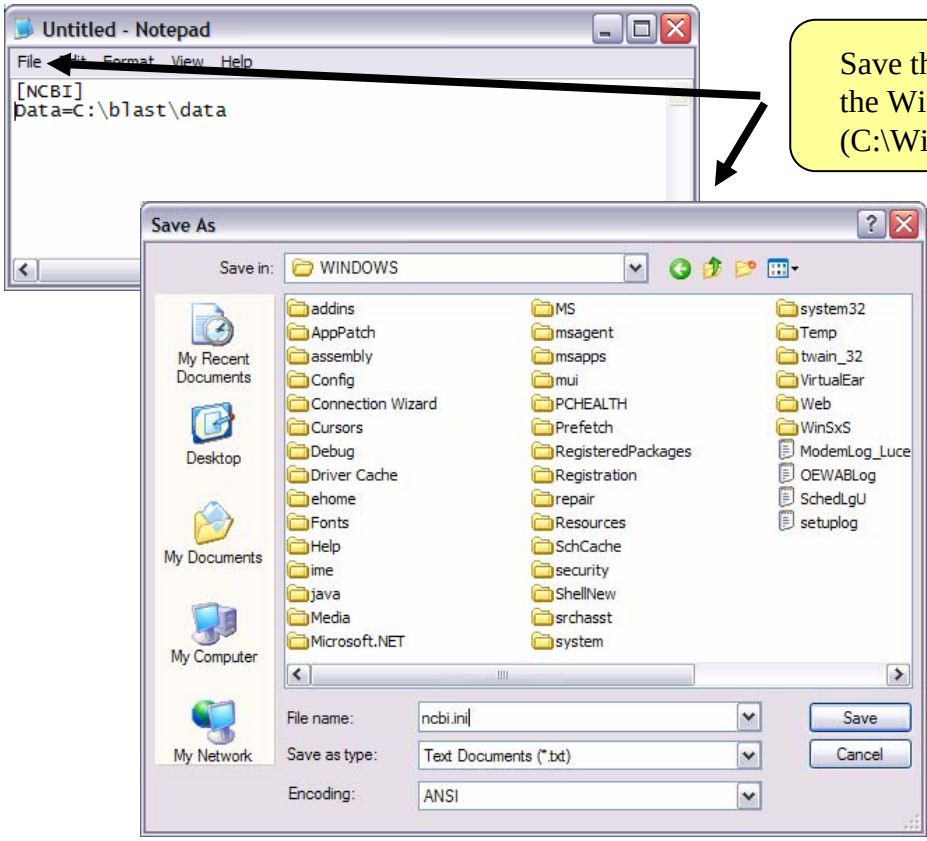
Before you can run BLAST you will need to create an **ncbi.ini** file containing the following lines:

[NCBI]
Data=C:\blast\data



Open **Notepad** (All programs, Accessories menu). Type in the text:

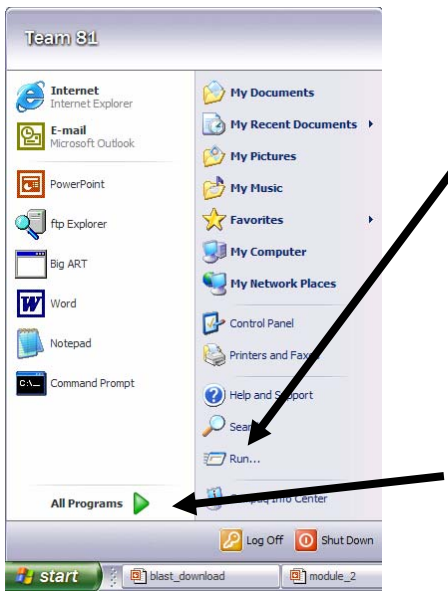
[NCBI]
Data=C:\blast\data



Save the file as **ncbi.ini** in the Windows directory (C:\Windows\)

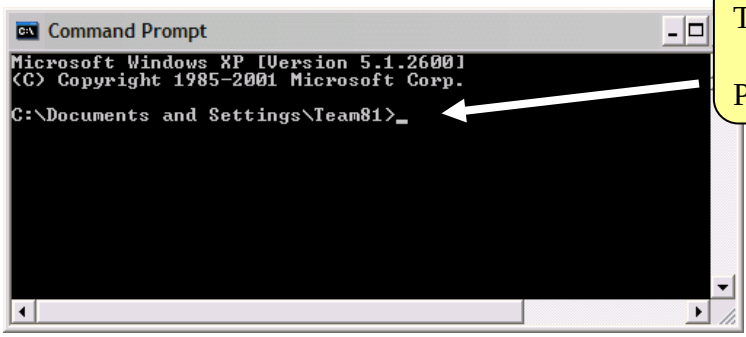
Running BLAST

The BLAST software does not run in Windows, but DOS, an operating system that Windows runs in. When you want to run blast you will need a DOS window a.k.a. Command Prompt



To get to a DOS window select from the Start menu, **Run** and type **Command**

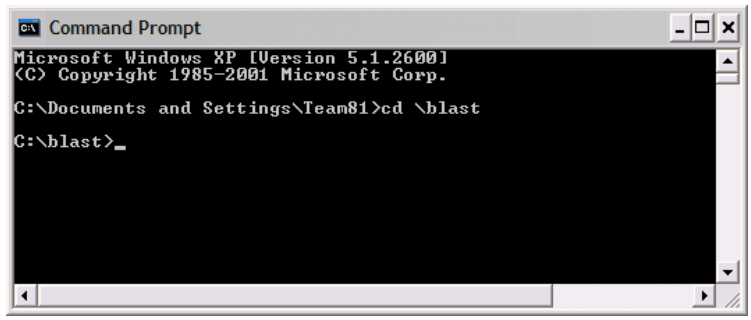
Alternatively the Command prompt icon can be found in Accessories, under All programs



Type **cd \blast**

Press **Return**

This changes the directory to the blast folder which you have just down loaded and unpacked blast-2.2.6-ia32-win32.exe



Now that that you are in the blast directory you can start to run BLAST from the command line

There are several programs in the BLAST package that you have now downloaded that can be used for sequence comparison. For a detailed description of the uses and options see the appropriate README file.