Workshop

‘Genomics of Clostridia’

January 24-26, 2007

Göttingen, Germany

Local organizer:
Prof. Dr. Wolfgang Liebl
Georg-August University Göttingen
Institute of Microbiology and Genetics
Table of contents

Workshop schedule and programme

Contact data of workshop tutors

Contact data of workshop participants

Manual “Annotating Genomes – on the reliability of data from genome projects”

Manual “Use of ARTEMIS and ACT for clostridial genome analysis”

Review “DNA microarray technology for the microbiologist: an overview”
Workshop ‘Genomics of Clostridia’

Göttingen, January 24-26, 2007

Location:
Lectures Seminar room, 1st floor (room No. 140), Institute of Microbiology and Genetics, Grisebachstraße 8
Practical Courses Computer facility, room MN05, Department of Bioinformatics, Institute of Geosciences, Goldschmidtstraße 1

Schedule and Programme:

Day 1 (Wednesday, Jan 24th):
Arrival in Göttingen
19:30 Dinner, Restaurant “Thalassa”, directly behind the hotel

Day 2 (Thursday, Jan 25th):
8:30 Welcome address
(Wolfgang Liebl, Genomic and Applied Microbiology Dept., Göttingen)
8:45 Genome sequencing strategies (Axel Strittmatter, G2L Göttingen)
9:30 Gene prediction in prokaryotes
(Burkhard Morgenstern, Bioinformatics Dept., Göttingen)
10:15 Coffee/Tea break
10:30 Comparative genomics of clostridia I
Exercises in comparative genomics and annotation using the ERGO software package (Heiko Liesegang, G2L Göttingen)
13:30 Lunch
14:30 Tour through the G2L genome sequencing facility
(Axel Strittmatter, G2L Göttingen)
15:30 Comparative genomics of clostridia II
Use of ARTEMIS and ACT for clostridial genome analysis
(Mohammed Sebaihia, Sanger Institute, Cambridge, UK)
19:30 Dinner, Restaurant “Rathskeller”, Markt 9, downtown Göttingen

Day 3 (Friday, Jan 26th):
8:30 Lessons from the Clostridium tetani and other clostridial genomes
(Gerhard Gottschalk, G2L Göttingen / Holger Brüggemann, Institut Pasteur)
9:30 To be announced (Peter Dürre, Microbiology Dept., Ulm)
10:15 Coffee/Tea break
10:30 From genome sequences to microarrays;
Introduction to procedures and software for microarray analysis
(Armin Ehrenreich, Microarray Facility, Göttingen)
12:30 Concluding remarks (Wolfgang Liebl, Göttingen)
13:00 Lunch
14:00 Departure
Contact data of workshop tutors

**Prof. Dr. Wolfgang Liebl**  
Universität Göttingen  
Institut für Mikrobiologie und Genetik  
Grisebachstr. 8  
D-37077 Göttingen  
Tel. +49-551-39-3795  
email: wliebl@gwdg.de

**Prof. Dr. Gerhard Gottschalk**  
Universität Göttingen  
Institut für Mikrobiologie und Genetik  
Grisebachstr. 8  
D-37077 Göttingen  
Tel. +49-551-39-4041  
email: ggottsc@gwdg.de

**Prof. Dr. Peter Dürre**  
Universität Ulm  
Institut für Mikrobiologie und Biotechnologie  
D-89069 Ulm  
Tel. +49-731-50-22711  
email: peter.duerre@uni-ulm.de

**Prof. Dr. Burkhard Morgenstern**  
Universität Göttingen  
Institut für Mikrobiologie und Genetik  
Abteilung Bioinformatik  
Goldschmidtstr. 1  
D-37077 Göttingen  
Tel. +49-551-39-14628  
email: bmorgen@gwdg.de

**Dr. Mohammed Sebaihia**  
Senior Computer Biologist  
The Wellcome Trust Sanger Institute  
Genome Campus, Hinxton  
Cambridge, CB10 1SA.  
Tel. 01223 494783  
email: ms5@sanger.ac.uk

**Dr. Armin Ehrenreich**  
Universität Göttingen  
Institut für Mikrobiologie und Genetik  
Grisebachstr. 8  
D-37077 Göttingen  
Tel. +49-551-39-3833  
email: aehrenr@gwdg.de

**Dr. Axel Strittmatter**  
Universität Göttingen  
Institut für Mikrobiologie und Genetik  
Grisebachstr. 8  
D-37077 Göttingen  
Tel. +49-551-39-3841  
email: astritt@gwdg.de

**Dr. Maike Tech**  
Universität Göttingen  
Institut für Mikrobiologie und Genetik  
Abteilung Bioinformatik  
Goldschmidtstr. 1  
D-37077 Göttingen  
Tel. +49-551-39-14297  
email: maike@gobics.de

**Dr. Heiko Liesegang**  
Universität Göttingen  
Institut für Mikrobiologie und Genetik  
Grisebachstr. 8  
D-37077 Göttingen  
Tel. +49-551-39-3843  
email: hlieseg@gwdg.de

**Rasmus Steinkamp**  
Universität Göttingen  
Institut für Mikrobiologie und Genetik  
Abteilung Bioinformatik  
Grisebachstr. 8  
D-37077 Göttingen  
Tel. +49-551-39-14297  
email: rsteink@gwdg
Contact data of workshop participants

**Jana Avbersek**, University of Ljubljana  
jana.avbersek@vf.uni-lj.si

**Tim. T. Binnewies**,  
Technical University of Denmark, Lyngby  
tim@cbs.dtu.dk

**Eve Bird**, University of Edinburgh  
e.bird@napier.ac.uk

**Benjamin Blount**, University of Nottingham  
nixbb@nottingham.ac.uk

**Jean-Charles Blouzard**, CNRS Marseille  
blouzard@ibsm.cnrs-mrs.fr

**Lukasz Dabrowski**, University of Norwich  
lukasz.dabrowski@bbsrc.ac.uk

**Cecile Deneve**, University of Paris Sud  
cecile.deneve@cep.u-psud.fr

**Anouk Lanckriet**, University of Ghent  
anouk.lanckriet@ugent.be

**Katherine McCurrie**, University of London  
k.mccurrie@UCL.ac.uk

**Oliver Riebe**, University of Rostock  
oliver.riebe@uni-rostock.de

**Janez Simenc**, University of Ljubljana  
j.simenc@volja.net

**Marcel Verhaart**, University of Wageningen  
marcel.verhaart@wur.nl

**Clare Cooksley**, University of Nottingham  
paxcmc1@nottingham.ac.uk

**Katja Hinderink**, University of Helsinki VetMed  
katja.hinderink@helsinki.fi

**Miia Lindström**, University of Helsinki  
miia.lindstrom@helsinki.fi

**Antonella Maugliani**,  
National Center for Food Quality and Risk Assessment, Rom  
mauganto@iss.it

**Christos Michaelidis**, University of Athen  
michaelidis@aau.gr

**Malgorzata Romanik**, University of Katowice  
mikrobiologia@slam.katowice.pl

**Elisabeth Steiner**,  
University of Wales, Aberystwyth  
eis@aber.ac.uk
Annotating Genomes – on the reliability of data from genome projects

Heiko Liesegang, Göttingen Genomics Laboratory
The annotation gap: thousands of new genes in a short time

- The availability of DNA sequence data increased dramatically in the last decade.
- In 1991 microbiologists expected *E. coli* to consist about 1,000 genes.
- The corresponding cellular functions were experimentally well characterized.
- The publication of *E. coli* K12 revealed about 4,317 genes.
- Thus 75% of the genes were new and unexpected!
The genomic era: annotation by comparison vs. experimental characterization

- The capacity of biochemical labs to characterize proteins is limited to a few proteins per researcher and year.
- In the characterized proteins no pair of sequences with a certain value of similarity have different functions.
  => a certain similarity points at a common function!
- Most genes are annotated by a database comparison to known proteins.
- The reliability of functional data from the databases is decisive to the quality of annotation.
Typical annotation process

- ORF finding
- Manual ORF curation
- Automated annotation
- Manual reannotation
- Share data with project partners
- Reevaluate data
- Construct metabolic network

↓

- “Reliable” annotated genome
YACOP - Combined ORF finding

- Glimmer
- Orpheus
- Critica
- Zcurve

ORF finding is based upon statistical properties of coding regions.

Experts have to curate ORFs according to frameshifts, start codons, artifact ORF finding.
ORF to gene

- *Thermus thermophilus* sequences bare 2 – 3 fold coverage of ORFs per sequence

- Automated ORF finding uses codon biases und HMM to identify potential “coding” ORFs

- Accuracy: ca. 97% relevant ORFs can be identified, ca. 50% to 65% of gene starting points can be correctly predicted

- Prediction of gene starting points is curated by RBS-finder and TICO

- All ORFs have to be curated by human experts.
**T. thermophilus**

- *T. thermophilus* contains two replicons:
  - Chromosome (1,894,877 bp, 69.2% GC content)
  - pTT27 (232,605 bp, 69.4% GC content)

Orffinding,  
manual curation

<table>
<thead>
<tr>
<th>ORFs</th>
<th>Chromosome</th>
<th>pTT27</th>
</tr>
</thead>
<tbody>
<tr>
<td>with assigned function</td>
<td>1397 (70%)</td>
<td>85 (37%)</td>
</tr>
<tr>
<td>conserved hypothetical</td>
<td>192 (10%)</td>
<td>56 (24%)</td>
</tr>
<tr>
<td>no database match</td>
<td>399 (20%)</td>
<td>89 (39%)</td>
</tr>
<tr>
<td>Total</td>
<td>1988</td>
<td>230</td>
</tr>
<tr>
<td>rRNA – Cluster</td>
<td>2</td>
<td>-</td>
</tr>
<tr>
<td>tRNA</td>
<td>47</td>
<td>-</td>
</tr>
<tr>
<td>IS-Elements</td>
<td>30</td>
<td>23</td>
</tr>
</tbody>
</table>
The first annotation steps - similarity based automated assignments

- **FASTA - automated assignments**
  - for all proteins against a high quality dataset SwissProt, PIR
- **BLAST, PSI-Blast**
  - Providing data for manual control
Transitive error propagation

- In automated annotation system as
  - GenDB
  - Genequiz
  - PEDANT
  - ERGO/WIT
- Annotations are assigned if proteins share a minimum similarity
- \( \Rightarrow \) original errors thus multiply by each new entry
Error in published genomes: Enzymes with an EC classification

- The EC classification for enzymes characterizes an enzyme by a four digit number.
- EC 6.1.1.19 (=) Arginyl-tRNA Synthetase (first characterized as Arginyl-tRNA Transferase, a complete different reaction type).
- (enzyme type).(cofactors).(reaction type).(substrate)
- Errors found:
  - First digit class of enzyme (~8 % errors)
  - Last digit substrate specificity (>= 33 % errors)
Source of annotation errors

- ORF-finding errors
- Sequencing errors
- Artifact similarity
- annotation errors in reference database
  - Misspelling
  - Under-annotation
  - Over-annotation
  - Miss-annotation
Manual curation of automated annotations

- A team of annotation experts has to check each automated annotation
- ERGO provides additional non similarity based means to evaluate annotations
- Tmpred, Pfam, COGs, Prosite are used to achieve a consistent functional assignment
- The kind of assignment evidence is a way of classifying annotation reliability
Ranking of reliability

1. Biochemical experimental characterization
2. Genetic experimental data from mutants
3. Consistent properties to a highly similar database hit
4. Good database hit
5. Consistent frame according to the used ORF finder
6. Non overlapping ORF of minimum size
Manual annotation using comparison data

- Experimental data give the highest reliable evidence for a protein-function. (not necessarily the best annotation)
- A single hit to a nucleotide sequence is not sufficient evidence to set CDS length.
- If possible take the annotation from best SWISS-PROT hit.
Manual annotation evaluating similarities

- An e-value is no absolute value, it depends on the protein length and the database size.
- The protein hit with the lowest e-value has the highest similarity.
- Similarity should cover the whole protein length.
- Cut off values:
  - \( e\text{-val} < 1.0\times 10^{-50} \) very similar protein
  - \( 1.0\times 10^{-50} < e\text{-val} < 1.0\times 10^{-10} \) moderate similarity
  - \( 1.0\times 10^{-10} < e\text{-val} \) non significant similarity
Functional coupling in genomes

- Most bacterial genes are clustered.
- Related organisms share the organization of gene clusters
- Cluster provide a context of genes for reannotation
- Cluster provide a source of data for gap closure
nomenclature

- Follow the newest highly similar SWISS-PROT hit.
- Gene names shall be used from older papers or the newest SWISS-PROT hit.
- Insecure annotations shall be marked with predicted (not 100% certain) or putative (possible but uncertain).
- Information based on uncertain data can be used in an addition like:
  - hypothetical protein, similar to hydrogenase
  -> low similarity hit
  - hypothetical protein, associated with ABC-transporter
  -> CDS co localised with ABC-transporter genes
  - hypothetical protein, related to Hydrogenase
  -> protein contains partial similarity to a hydrogenase
Gene product; annotation (alternative annotation)(..)(EC number)
Start with gene product when known AbcA
Avoid capital letters where possible
Include EC number
Example1:
AroA; 3-phosphoshikimate 1-carboxyvinyltransferase (5-enolpyruvylshikimate-3-phosphate synthase) (EPSP synthase) (EPSPS) (EC2.5.1.19)
Example2:
LolD; lipoprotein releasing system, ATP binding protein LolD
Metabolic reconstruction

- Based on annotated genes metabolic pathways can be predicted
- Pathway finding is based upon EC categorisation
- Missing EC functions might be found in the “hypothetical” genes
- Thus pathway reconstruction leads to a reannotated genome
Proteomics high throughput protein characterization

- A proteomic lab produces N-terminal sequences for all expressed proteins under certain conditions.
- Comparison with the genome data gives the corresponding gene
Genomic properties

- Genome sequences provide physiological predictions
- Predicted functions have to be proven experimentally
- High throughput evidence methods are transcriptomics and proteomics
Artemis

Introduction

Artemis is a free DNA viewer and annotation tool written by Kim Rutherford (Rutherford et al., 2000). It is routinely used by the Sanger Institute Pathogen Sequencing Unit for annotation and analysis of both prokaryotic and eukaryotic genomes. The program allows the user to view simple sequence files, EMBL/Genbank entries and the results of sequence analyses in a highly interactive and intuitive graphical format. Artemis is designed to present multiple sets/types of information within a single context. This manifests itself as the ability to zoom in to inspect DNA sequence motifs and zoom out to view local gene architecture (e.g. operons), several kilobases of a genome or even an entire genome in one screen. It is also possible to perform some analyses within Artemis with the output stored for later access.

Aims

The aim of this Module is for you to become familiar with the basic functions of Artemis using a series of worked examples. These examples are designed to take you through the most immediately useful functions. However, there will be time, and encouragement, for you to explore other menus; nooks and crannies of Artemis that are not featured in the exercises in this manual. The key is ‘if you don’t understand please ask’.
Artemis Exercise Part I

1. Starting up the Artemis software
Go to the ‘Artemis.dir’ 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 *Clostridium difficile* chromosome sequence.

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.

DNA sequence files will have the suffix ‘.dna’. Annotation files end with ‘.tab’. Use this feature to select the type of file displayed in this panel.
2. Loading annotation files (entries) into Artemis
Hopefully you will now have an Artemis window like this! If not, ask for assistance.

Now follow the numbers to load up the annotation file for the *C. difficile* chromosome.

1. Click ‘File’ then ‘Read an Entry’
   - Entry = file

2. Single click to select tab file

3. Single click to open file in Artemis then wait

What’s an “Entry”? It’s a file of DNA and/or features which can be overlaid onto the sequence information displayed in the main Artemis view panel.
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 *recF* (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 (CDSs) and other features (eg. Pfam and Prosite matches) are displayed as coloured boxes.
4. This panel has a similar layout to the main panel but is zoomed in to show nucleotides and amino acids. Double click on a gene in the main view to see the zoomed view of the start of that gene. Note that both this and the main panel can be scrolled left and right (7, below) zoomed in and out (6, below).
5. This panel lists the various features in the order that they occur on the DNA with the selected gene highlighted. The list can be scrolled (8, below).
6. Sliders for zooming view panels.
7. Sliders for scrolling along the DNA.
8. Slider for scrolling feature list.
4. Getting around in Artemis

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

4.1 The ‘Goto’ menu

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

Click ‘Goto’

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.

Click ‘Goto’ then Navigator

Check that the search button is on

Suggestions of where to go:
1. Think of a number between 1 and 4290252 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 chromosome.
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 VII).
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
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 1284327..1314871 on the DNA sequence. This region is adjacent to the zupT gene which encodes a zinc transporter. 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.

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

**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 yellow 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 Pfam.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.

Further information on specific Prosite or Pfam entries can be found on the web at http://ca.expasy.org/prosite and http://www.sanger.ac.uk/software/Pfam/tsearch.shtml
In addition to looking at the fine details of the annotated features it is also possible to look at the characteristics of the DNA covering the region displayed. This can be done by adding 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 may take a little time so be patient.

**To view the graphs:**

Click on the ‘Graph’ menu to see all those available. Perhaps some of the most useful plots are the ‘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 all 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. The *C. difficile* genome carries several genomic islands similar to the one shown above. We are going to explore one of these in details later.
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, 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 CD.tab entry button (1) on the grey entry line of the Artemis window shown above. You will also need to switch off stop codons by clicking with the right mouse button in the main view panel (2). A menu will appear with an option to de-select stop codons (3).
One final tip is to adjust the scaling for each graph displayed before zooming out. This increases the maximum window size over which a single point for each plot is calculated. To adjust the scaling click with the right mouse button over a particular graph window (1). A menu will appear with a series of values for the maximum window size (2), select 20000. You should do this for each graph displayed. You are now ready to zoom out by dragging or clicking the slider (3).

Right-click anywhere in the graph window.

Graph scaling menu

Slider for zooming out

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. Your Artemis window should now look similar to the one shown below. If you have any problems ask.

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 your previously clicked in.
Like the region that we have investigated before, there are several other regions in the *C. difficile* genome that display characteristic features of laterally acquired DNA. In this part of the exercise we are going to look at one of these regions in more detail. Starting with the whole genome view, note down the approximate positions and characteristics of the region indicated above (*). Remember the locations of the peaks are given in the graph window if you click with the left mouse button within it.

You can make a note of the genome location here:

We will now zoom back into the genome to look in more detail at this region. 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 (CD.tab) entry back on.
The region you should be looking at is shown below and corresponds to the \textit{C. difficile} conjugative transposon (CTn) \textit{Tn5397}, which mediates tetracycline resistance. The genome of \textit{C. difficile} carries several CTn. CTn are genetic elements that are capable of integrating into, and excising from, the host genome and transferring themselves, \textit{via} a conjugation like mechanism to other, often genetically unrelated bacteria. CTn are usually composed of modules involved in conjugation, regulation, integration and excision, and accessory genes that frequently encode antibiotic resistance. Occasionally they insert within, and disrupt, a gene.

Have a look in and around this region and look for some of these features.
Continuing on from the analysis of Tn5397 we are going to extract this region from the whole genome sequence and save it as a separate file which will include just the annotation and DNA for this region.

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

2. Click ‘Edit’

3. Click ‘Edit subsequence (and features)’

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

Note the entry names have changed.

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. CD.dna and CD.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 Tn937.dna

Do the same again for the other unnamed entry and save it as Tn937.tab

You have now successfully extracted a region of DNA and its associated annotation, and saved them as separate files.
Artemis Exercise Part V

In this exercise we are going to search for a group of genes using a keyword (e.g. sporulation), extract them from the genome and save them as a separate file. We are going to use the Feature selector to search for these genes as shown below.

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 these genes into here.

1. Click ‘Select’ then ‘Feature Selector’
2. Make sure the buttons are down
3. Set Key to ‘CDS’ and Qualifier to ‘product’
4. Type search term
5. Click to select features containing search term
6. Click to view selected features
7. Double click to bring feature into main view window
The genes listed in 6 are only those fitting your selection criterion (in this case sporulation). They can be copied or moved in to a new entry so we can view them in isolation from the rest of the genome.
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 ‘Copy Selected Features To’ then ‘no name’. Close the two smaller feature selector windows and return to the Artemis window. You could rename the ‘no name’ entry as you did before. Temporarily remove the features contained in ‘CD.tab’ file by left clicking on the entry button on the grey entry line. Only the genes that you selected 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 such as text or amino acid searches.
Artemis Comparison Tool (ACT)

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 sequences from closely related organisms (Carver et al. 2005). ACT is based on Artemis, and so you will already be familiar with many of its core functions, and 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 DNA sequences with their associated features. The middle window shows red and blue blocks, which span this middle layer and link conserved regions within the two sequences, in the forward and reverse orientation respectively. Consequently, if you were comparing two identical sequences in the same orientation you would see a solid red block extending over the length of the two sequences in this middle layer. If one of the sequences was reversed, and therefore present in the opposite orientation, there would be a blue ‘hour glass’ shape linking the two sequences. Unique regions in either of the sequences, such as insertions or deletions, would show up as breaks (white spaces) between the solid red or blue blocks.

In order to use ACT to investigate your own sequences of interest you will have to generate your own pairwise comparison files. Data used to draw the red or blue blocks that link conserved regions is generated by running pairwise BLASTN or TBLASTX comparisons of the sequences. 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, which can be loaded and run on a PC or Mac. A step-by-step guide of how to download and install the BLAST software on a PC is outlined in Appendix VIII. Instructions on how to generate comparison files from BLAST can be found in Appendix IX. Whilst having a local copy of BLAST to generate ACT comparison files can be very useful, it means that you are tied to a particular computer. Another way of generating comparison files for ACT is to use either of the WebACT or Double ACT web resources (see Appendix IV for URLs). Both of these sites allow you to cut and paste or upload your own sequences, and generate ACT readable BLASTN or TBLASTX comparison files.

Aims
The aim of this Module is for you to become familiar with the basic functions of ACT by using a series of worked examples. Hopefully, as well as introducing you to the basics of ACT this Module will also show you how ACT can be useful for a wide range of tasks during genome analysis. In this module you will also use a web resource, WebACT, to generate your own comparison files and view them in ACT.
ACT Exercise Part I

1. Starting up the ACT software
Make sure you’re in the ‘ACT.dir’ directory.
Then type
`act & [return]`
A small start up window will appear.
Now let’s compare the *C. difficile* and *C. botulinum* genomes.

The files you will need for this exercise are:

- CD.dna (*C. difficile* genome)
- CB_vs_CD.crunch (Comparison file)
- CB.dna (*C. botulinum* genome)

Click ‘File’ then ‘Open’

1. Click ‘File’ then ‘Open’
2. Sequence files
3, 4 & 5
4. Click and select appropriate files
5. Comparison files end with ‘.crunch’. For more info on comparison files see Appendix II.
6. For comparing more than two genomes!

Click ‘Apply’ and wait……
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). It’s a slightly compressed version of the Artemis main view panel. The panel retains the sliders for scrolling along the genome and for zooming in and out.

3. The Comparison View. This panel displays the regions of similarity between two sequences. Red blocks link similar regions of DNA with the intensity of red colour directly proportional to the level of similarity. Double clicking on a red block will centralise it. Blue blocks link regions that are inverted with respect to each other.

4. Artemis-style Sequence View panel for ‘Sequence file 2’ (Query Sequence).

5. Right button click in the Comparison View panel brings up this important ACT-specific menu which we will use later.
**ACT Exercise Part II**

**Introduction & Aims**
In this 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 *C. difficile* chromosome compared against the *C. botulinum* chromosome. 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 genome 1 as shown below, and repeat the process for genome 2.

1. Right button click here
2. De-select stop codons

Your ACT window should look 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 below. 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 slider both genomes move together. This is because they are ‘locked’ together. Right click over the middle comparison view panel (1). A small menu will appear, de-select lock sequences (2) 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”.

1. Right button click in the Comparison View panel
2. Select either Set Score Cutoffs or Set Percent ID Cutoffs
3. Move the sliders to manipulate the comparison view image
4. Slider for filtering regions of similarities according to size
Things to try out in ACT

Load into the top sequence (C. difficile) the annotation file ‘CD.tab’. You will need to use the ‘File’ menu and select the correct genome sequence (‘CD.dna’) before you can read in an entry. Repeat the process with the bottom genome (C. botulinum) by loading ‘CB.tab’.

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. Also try using some of the other Artemis features eg. graphs etc.
4. Find an inversion in one genome relative to the other then flip one of the sequences. To do that you have to click with the right-mouse button within the Comparison View panel and select ‘Flip Subject Sequence’ or ‘Flip Query Sequence’.
Aim

In this exercise you will be investigating a small region in the *C. difficile* and *C. botulinum* genomes. As part of the exercise you will be asked to compare and identify any visible similarities/differences between these two regions, and if possible, draw any conclusions.

Using one of the two genomes, and a method of your choice, go to:
- gene CD1844, at position 2136229, in the *C. difficile* genome
- or
gene CBO1056, at position 2137467, in the *C. botulinum* genome

After removing the smaller matches, the region you arrive at is shown below.
What immediate observations can you make?

Can you identify any genes in one genome that don’t appear in the other? If so, in which genome. What are the functions of the predicted products of these genes? 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 see any similarities between the two regions? If so, how many?

Double click on the regions of similarities

What are the predicted product(s) of the gene(s) displaying similarity?

Can you identify any apparent characteristics of the predicted product(s) of the gene(s) displaying similarity?

What overall conclusion(s) can you draw from this analysis?
Generating ACT comparison files

Introduction

There are two ways to generate ACT comparisons files.

1. You can generate your own comparison files for ACT from a locally-installed stand-alone version of the BLAST software. Details on how to download and install the NCBI BLAST distribution on a PC with Windows XP are described in Appendix VIII. A step by step guide on how to generate ACT comparison files using a locally-installed BLAST are in Appendix IX.

2. If you do not have access to BLAST software running on a local computer, there are a couple of web resources (WebACT or Double ACT; see Appendix IV for URLs) that can be used for generating ACT comparison files. Both of these sites allow you to cut and paste, or upload, your own sequences, and generate ACT readable BLASTN or TBLASTX comparison files. WebACT also has a large selection of pre-computed comparison files for bacterial genomes, which can be downloaded along with the EMBL sequence entries and viewed in ACT.

In this section, if you have time, you are going to use WebACT to generate ACT comparison files between C. perfringens and C. tetani.
Exercise
Generating ACT comparison files exercise using WebACT

Aim

For the purposes of this exercise you are going to use a web resource, WebACT, to generate a comparison file between the chromosomes of *C. perfringens* and *C. tetani*. From the WebACT site you will download a pre-computed ACT comparison file, along with the appropriate EMBL sequence and annotation files, which you will then open in ACT.
Open up a web browser and go to the URL: www.webact.org

Click on the Pre-computed tab.

The ‘Pre-computed’ page contains genomic sequences that have been compared using BLASTN to each other. By selecting the desired sequences from the sequence lists, the appropriate sequence and comparison files can be downloaded.

WebACT can display pairwise comparison between up to 5 sequences. Click here if you want to increase the number from the default of 2.

In addition to the chromosome sequences, plasmids can also be displayed by clicking in this box.
In the **Sequence 1** list select *C. perfringens* chromosome (accession number BA000016).

In the **Sequence 2** list select *C. tetani* chromosome (accession number AE015927).

Once you have selected the sequences click the **Next** button.

In this window you can specify the regions in the selected sequences to generate the comparison over. It is possible to query the sequences on gene name or coordinates. The default setting is for the whole sequence, and this is what we want for this exercise as you are going to compare the whole chromosomes.

Click the **Next** button.
In the Overview of Selection you can see a schematic representation of the relative size of the two sequence that have been chosen to be compared.

The Expect (E) value cut-off can be changed in this box. The default value is 0.01, but the range is from 10.0 to 0.0001.

Click the Download files button.

In addition to downloading the comparison file and sequence files it is also possible to view the comparison in a web start version of ACT. This will run locally on your machine and does not require ACT to be previously loaded, as a web start version of ACT will be included in the download. You are not going to use this option in this exercise.

The comparison file and sequences files will be contained in a folder. For the ease of downloading, the folder is zipped.

In the filename box you can type the file name of the zip file containing the sequence and comparison files. For this exercise call the file: CP_CT_comp.zip

Click the Download files button.
You may get a window appearing asking you what Mozilla should do with the CP_CT_comp.zip file? Save the file to disk.

Click the OK button

Select a ‘ACT.dir’ directory

Click the Save button

CP_CT_comp.zip should now be saved to ‘ACT.dir’ directory. To extract a folder containing the comparison file and the appropriate EMBL sequences you will have to use the UNIX command, unzip. Type:

```
unzip CP_CT_comp.zip
```

If in future you download comparisons from WebACT onto a Windows PC or Mac, the comparison folder file may automatically be unzipped and therefore you will be able to see the folder and the contained files. If not you can un-zip it using the appropriate software (double clicking on the zipped file icon may work).

The files contained in the directory should include:
comparison1_BA000016_vs_AE015927, sequence1_BA000016.embl and sequence2_AE015927.embl. These are the ACT comparison file and the C. perfringens and C. tetani chromosome annotation and sequence files, respectively.
Open up ACT, and load up the comparison (comparison1_BA000016_vs_AE015927) along with the two EMBL sequence and annotation files (sequence1_BA000016.embl and sequence2_AE015927.embl). If you get a warning window asking if you want to read warning, click No.
Now remove the stop codons for both entries, and then zoom out as you did before. Your ACT window should look like the one shown below.

Move the slider down to eliminate smaller matches.
In WebACT you can also generate, and view in ACT, comparisons for your own and/or other sequences. Sequences in various formats can be cut and pasted, or uploaded onto the WebACT site. In addition, if you know the accession number of the sequence that you want to compare, you can use that.

- Cutting and pasting a sequence
- Uploading a file
- Entering an EMBL or RefSeq Accession number

Once you have added the relevant sequence information, submit your query. The comparison file or files are downloaded as shown in the exercise above, and ready to be loaded into ACT.
References

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

ACT: the Artemis Comparison Tool.

Viewing and annotating sequence data with Artemis.

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
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 ones 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 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**: Descriptive gene a name, eg. ilvE, argA etc.
\**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.
\**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.
\**locus_tag**: Systematic gene number, eg SAS1670, Sty2412 etc.

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: 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
- Genomes at the EBI http://www.ebi.ac.uk/genomes/

Microbial Genome Databases Resources
- Sanger Microbial Genomes http://www.sanger.ac.uk/Projects/Microbes/
- GeneDB http://www.genedb.org/
- TIGR Microbial Database http://www.tigr.org/tdb/mdb/mdbcomplete.html
- Institute Pasteur GenoList databases http://genolist.pasteur.fr
  - Including: SubtiList, Colibri, TubercuList, Leprona, PyloriGene, MypuList, ListiList, CandidaDB
- Pseudomonas Genome Database http://www.pseudomonas.com/
- ScoDB (S. coelicolor database) http://streptomyces.org.uk/sco/index.html
- GenProtEC http://genprotec.mbl.edu

Protein Motif Databases
- Prosite http://www.expasy.ch/prosite/
- Pfam http://www.sanger.ac.uk/Software/Pfam/
- BLOCKS http://blocks.fhcrc.org
- InterPro http://www.ebi.ac.uk/interpro/
- PRINTS http://umber.sbs.man.ac.uk/dbbrowser/PRINTS/
- SMART http://smart.embl-heidelberg.de

Protein feature prediction tools
- TMHMM Transmembrane helices prediction http://www.cbs.dtu.dk/services/TMHMM-2.0/
- 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/
- MetaCyc http://metacyc.org/

Miscellaneous sites
- EBI FASTA website http://www.ebi.ac.uk/ fasta33/index.html
- The tmRNA website http://www.indiana.edu/~tmrna/
- tRNAScan-SE Search Server http://www.genetics.wustl.edu/eddy/ tRNAscan-SE/
- Rfam http://www.sanger.ac.uk/Software/Rfam/
- Codon usage database http://www.kazusa.or.jp/codon/
- 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/
- WebACT http://www.webact.org/WebACT/home
- Double ACT http://193.129.245.227/pise/double_act.html
- Glimmer http://cbcb.umd.edu/software/glimmer/
- Orpheus http://pedant.gsf.de/orpheus
- EasyGene http://www.cbs.dtu.dk/services/EasyGene/
- String http://string.embl.de
- EMBOSS http://emboss.sourceforge.net/
Appendix V: 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/files/Multifun.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).

<table>
<thead>
<tr>
<th>Function Categories</th>
<th>Subcategories</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0.0 Unknown function, no known homologs</td>
<td></td>
</tr>
<tr>
<td>0.0.1 Conserved in Escherichia coli</td>
<td></td>
</tr>
<tr>
<td>0.0.2 Conserved in organism other than Escherichia coli</td>
<td></td>
</tr>
<tr>
<td>1.0.0 Cell processes</td>
<td></td>
</tr>
<tr>
<td>1.0.1 Chemotaxis and mobility</td>
<td></td>
</tr>
<tr>
<td>1.0.2 Chromosome replication</td>
<td></td>
</tr>
<tr>
<td>1.0.3 Chaperones</td>
<td></td>
</tr>
<tr>
<td>1.4.0 Protection responses</td>
<td></td>
</tr>
<tr>
<td>1.4.1 Cell killing</td>
<td></td>
</tr>
<tr>
<td>1.4.2 Detoxification</td>
<td></td>
</tr>
<tr>
<td>1.4.3 Drug/analog sensitivity</td>
<td></td>
</tr>
<tr>
<td>1.4.4 Radiation sensitivity</td>
<td></td>
</tr>
<tr>
<td>1.5.0 Transport/binding proteins</td>
<td></td>
</tr>
<tr>
<td>1.5.1 Amino acids and amines</td>
<td></td>
</tr>
<tr>
<td>1.5.2 Cations</td>
<td></td>
</tr>
<tr>
<td>1.5.3 Carbohydrates, organic acids and alcohols</td>
<td></td>
</tr>
<tr>
<td>1.5.4 Anions</td>
<td></td>
</tr>
<tr>
<td>1.5.5 Other</td>
<td></td>
</tr>
<tr>
<td>1.6.0 Adaptation</td>
<td></td>
</tr>
<tr>
<td>1.6.1 Adaptations, atypical conditions</td>
<td></td>
</tr>
<tr>
<td>1.6.2 Osmotic adaptation</td>
<td></td>
</tr>
<tr>
<td>1.6.3 Fe storage</td>
<td></td>
</tr>
<tr>
<td>1.7.1 Cell division</td>
<td></td>
</tr>
<tr>
<td>2.0.0 Macromolecule metabolism</td>
<td></td>
</tr>
<tr>
<td>2.1.0 Macromolecule degradation</td>
<td></td>
</tr>
<tr>
<td>2.1.1 Degradation of DNA</td>
<td>2.1.3 Degradation of polysaccharides</td>
</tr>
<tr>
<td>2.1.2 Degradation of RNA</td>
<td>2.1.4 Degradation of proteins, peptides, glycoproteins</td>
</tr>
<tr>
<td>2.2.0 Macromolecule synthesis, modification</td>
<td></td>
</tr>
<tr>
<td>2.2.01 Amino acyl tRNA synthesis; tRNA modification</td>
<td>2.2.07 Phospholipids</td>
</tr>
<tr>
<td>2.2.02 Basic proteins - synthesis, modification</td>
<td>2.2.08 Polysaccharides - (cytoplasmic)</td>
</tr>
<tr>
<td>2.2.03 DNA - replication, repair, restriction/modification</td>
<td>2.2.09 Protein modification</td>
</tr>
<tr>
<td>2.2.04 Glycoprotein</td>
<td>2.2.10 Proteins - translation and modification</td>
</tr>
<tr>
<td>2.2.05 Lipopolysaccharide</td>
<td>2.2.11 RNA synthesis, modif., DNA transcrip.</td>
</tr>
<tr>
<td>2.2.06 Lipoprotein</td>
<td>2.2.12 tRNA</td>
</tr>
<tr>
<td>3.0.0 Metabolism of small molecules</td>
<td></td>
</tr>
<tr>
<td>3.1.0 Amino acid biosynthesis</td>
<td></td>
</tr>
<tr>
<td>3.1.01 Alanine</td>
<td>3.1.08 Glutamine</td>
</tr>
<tr>
<td>3.1.02 Arginine</td>
<td>3.1.09 Glycine</td>
</tr>
<tr>
<td>3.1.03 Asparagine</td>
<td>3.1.10 Histidine</td>
</tr>
<tr>
<td>3.1.04 Aspartate</td>
<td>3.1.11 Isoleucine3.1.18 Threonine</td>
</tr>
<tr>
<td>3.1.05 Chorismate</td>
<td>3.1.12 Leucine</td>
</tr>
<tr>
<td>3.1.06 Cysteine</td>
<td>3.1.13 Lysine</td>
</tr>
<tr>
<td>3.1.07 Glutamate</td>
<td>3.1.14 Methionine</td>
</tr>
</tbody>
</table>
Appendix V (cont):

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 Lipoyte
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)

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 Glucoseogenesis
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
3.3.11 Nucleotide interconversions
3.3.12 Oligosaccharides
3.3.13 Phosphorus compounds
3.3.14 Polyamine biosynthesis
3.3.15 Pool, multipurpose conversions of intermed. metab.
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
3.3.21 other

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
3.4.6 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
5.1.5 Pathogenicity island-related function

6.0.0 Global functions
6.1.1 Global regulatory functions

7.0.0 Not classified (included putative assignments)
Appendix VI: 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 VII: 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 VIII: 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. N.B. The version of BLAST available from the NCBI will have changed since these screen shots were taken, therefore the file names will not be same as those show below.

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 shortcut

Double click on blast-20041205-ia32-win32.exe

blast-20041205-ia32win32.exe is the blast exe file for windows.
N.B. The file name of the latest will have changed since these screen shots were taken
You now need to save the `blast-20041205-ia32-win32.exe` file in a new directory, blast, on to the hard drive of your PC.

Click on local disk C:

Click on new directory icon (folder with a sun peeking through)

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-20041205-ia32-win32.exe is a compressed file that contains a host of other files.

Now double click on the blast-20041205-ia32-win32.exe file to extract and unpack the rest of the BLAST download files.

README files that describe the various programs in the BLAST software package are to be found at ftp://ftp.ncbi.nih.gov/blast/documents. 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 blast.txt 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
Running BLAST on a Windows PC

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.

Save the file as `ncbi.ini` in the Windows directory (C:\Windows\)
Now that you are in the blast directory you can start to run BLAST from the command line. See next section (Appendix IX) for examples of running BLAST from the command line.

Type `cd \blast`

Press Return

This changes the directory to the blast folder which you have just downloaded and unpacked blast-20041205-ia32-win32.exe.
Appendix IX. Generating ACT comparison files using BLAST

There are several programs in the BLAST package that you have downloaded 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.

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.

The following pages demonstrate how you can generate your own comparison files for ACT from a stand-alone version of the BLAST software. You need two sequences, seq1 and seq2 in FASTA format. 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.

![Xterminal window](image)

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 seq1.dna as the database sequence and seq2.dna as the query sequence.

At the Command Prompt type:
```
formatdb -i seq1.dna -p F
```
Press Return.

formatdb is the database format program.

-\textit{i} designates the input sequence: seq1.dna

-\textit{p} designates the sequence type: DNA is F (protein would be T)

Now you can run the BLAST on the two 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 (\textbf{-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 -pblastn -m 8 -d seq1.dna -i seq2.dna -o seq1_vs_seq2
```
Press Return.

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

blastall is the BLAST program.

-\textit{p} designates the flavour of BLAST: blastn (in this instance a DNA-DNA comparison)

-\textbf{m 8} designates the ACT readable output

-\textit{d} designates the database sequence: seq1.dna

-\textit{i} designates the query sequence: seq2.dna

-\textit{o} designates the output file: seq1_vs_seq2