BugView

Users' Manual

David P. Leader

d.leader@bio.gla.ac.uk
http://www.gla.ac.uk/~dpl1n/BugView/

Version 1.3.3

March 2006

CONTENTS

Citation in Publications

If you publish work in which you have used *BugView*, please cite:

Bioinformatics **20**, 129-130, 2004

What's new in version 1.3.3?

Version 1.3.3 of *BugView* includes further enhancements to v.1.3.1, which had introduced improved cross-platform support (including a native Mac OS X version) compared to v. 1.3. (An applet version of *BugView* was also released about this time.) Version 1.3.2 was primarily a bug-fix version, but, in addition to fixing a couple of further bugs, v. 1.3.3 simplifies the creation of both Comparison files and Comparison pairs, and provides percentage identity information on Comparison pairs to help one choose which to align or delete. The Find and Search functions now default to the user's previous choice.

This version of the BugView manual also provides information on a new web facility for creating comparison pairs (Appendix III).

1. Introduction

BugView is an application to help experimental scientists visualize and compare bacterial genomes. For a particular genome the user may work with individual genes and:

- View their position in relation to other genes
- View information on their gene products
- Edit gene information or make annotations
- View the nucleic acid sequence of the gene and its conceptual translation
- Print this textual information or the graphic view, or save it to file.

BugView has been tailored to allow the user to compare the genes of two different genomes. A special comparison file may be loaded or created that lets the user:

- View information on related genes
- Perform or view pairwise comparisons by the Smith–Waterman method
- Create new relationships or delete old ones
- Adjust the overall view in different ways to facilitate comparison.

Although *BugView* was developed for bacterial genomes, it will also display eukaryotic chromosomes, and will represent individual exons. Currently the largest genome successfully tested is the 30 Mb chromosome I of *A.thaliana*.

System Requirements

BugView is a written in a quite basic version (1.1) of the cross-platform programming language, Java. Hence there should be a 'flavour' of *BugView* to suit the personal computers of most users.

Nevertheless, there are problems that might be encountered. Java programs require a so-called 'Java Virtual Machine' to run. This is provided by default on Macs (Apple manufactures its own JVM) but may not be installed on some versions of Windows and Unix/Linux. In that case it should be downloaded from Sun's website (e.g. at http://java.sun.com/j2se/1.4.2/download.html).

A second possible problem may be encountered with older machines. Although small sections of genomes can be handled on machines with relatively slow processors, the ability to handle large genomes requires a processor running at 500 MHz at least.

Mac

Separate double-clickable versions of *BugView* are provided for MacOS 9 (or 8) and Mac OS X. Users are advised to remove previous versions, and, on MacOS 9, to rebuild the desktop.

Windows

A simple double-clickable executable of *BugView* is available for Windows (95 to XP).

Unix/Linux

For Unix/Linux a 'jar' file, *BugView.jar*, is provided, together with a shell script, *bugview.sh*, providing the necessary command-line arguments, including extra memory assignment:

```
java -jar -Xms200m -Xmx400m BugView.jar
```

Acknowledgements

BugView was developed from the version of Andrei Gregoriev's Der Browser applet (*Bioinformatics* **14**, 252-258, 1997) modified by David Leader (http://www.gla.ac.uk/~dpl1n/derBrowser/). The sequence alignment code is courtesy of Peter Sestoft (http://www.dina.dk/~sestoft/), the gif preparation code is courtesy of Acme (http://www.acme.com/java/), and the PostScript preparation code is the copyright of E.J. Friedman-Hill (Sandia National Laboratories).

2. Loading and Unloading Genome Files

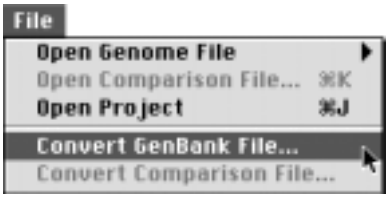
File Format

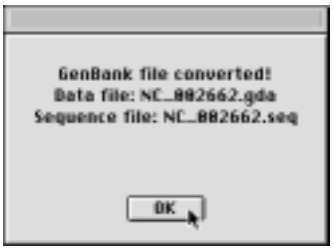
BugView uses files in a format based on that of *GenBank*, annotated in the manner currently standard for bacterial genomes. The essential features of this are described in Appendix I, which the user is strongly advised to consult before trying to use the program on files from other organisms.

For speed in saving annotations to file, and because repeated manipulation of the nucleic sequence can lead to corruption, the *GenBank* data for each genome is converted to two files — a Data file and a Sequence file. User annotations and edits are written to the Data file when they are saved to disc. The Sequence file, however, is left unchanged.

Converting *GenBank* files

1. Select 'Convert GenBank File' in the File menu, the first time you work on a genome, and open the *GenBank* file from the standard selection window.
2. When the file has loaded and been converted you will receive a message notifying you of the filenames of the Data and Sequence files generated. These are based on the accession number of the genome and are given the extensions .gda and .seq, respectively. It is not advisable to alter or remove these extensions. For genomes larger than 15 Mb, a Data file but no Sequence file is generated. In these circumstances one uses (and is able to load) the *GenBank* file instead of the Sequence file.
3. You will be able to continue working on the genome without quitting the application, but any changes will be saved to the Data file. The *GenBank* file will remain unchanged and is not required in subsequent sessions, unless it is being used in place of a Sequence file.



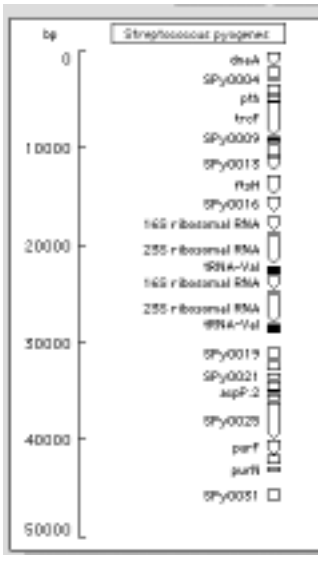


Loading Data and Sequence files.

The Data file for a genome must always be loaded before any Sequence file. It is not, in fact, necessary to load a Sequence file in order to view and annotate a genome, although in that case sequence-dependent information and procedures will obviously not be available. Any number of files may be loaded into a *BugView* window, but the same file may not be loaded twice.

1. To load a Data file select 'Open File...' > 'Open Data File' from the File menu and navigate to the file and open it. If successful you will see a display of the genome of the type shown.
2. To load a Sequence file select 'Open File...' > 'Open Sequence File' from the File menu and navigate to the file and open it. You will be informed when the file has loaded.

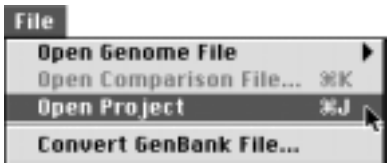
We reiterate that the Sequence file is only read by (not written to) *BugView*, and will remain unaltered at the end of your session. If, for any reason, you wish to edit a sequence you will need to open the Sequence file in a program that accepts and edits plain text.

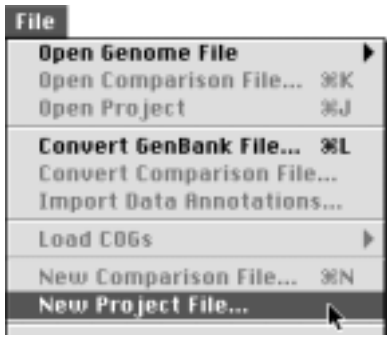


Loading a Project

To avoid the chore of loading five separate files for a comparison of two genomes, one can load them automatically from a single 'Project File' (extension '.prj') specifying the names of constituent files. This is done by selecting 'Open Project' from the File menu. The five files specified in the Project file (or aliases of them) must be in the same folder as the Project file.

The format for a Project file is very simple, and described in Appendix I. However one can generate a Project file automatically after loading the five constituent files by selecting 'New Project File' from the File menu.

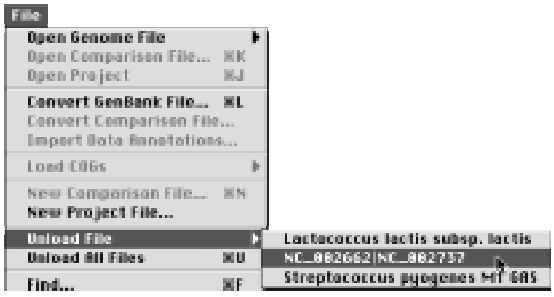




Unloading Files

Individual files (e.g. ones loaded by mistake or no longer required) can be removed completely from *BugView* without quitting by using 'Unload File' from the File menu. You will be asked if you wish to save any changes you have made (unless you have set your session to 'Autosave'). Occasionally you may be asked whether you wish to make changes when you do not think you have made any. This is because of limitations in the software which does not detect whether you have made any changes after you have opened a dialogue box.

The 'Unload All Files' menu item (immediately below 'Unload File') allows one to remove all the files from *BugView* in a similar manner



Renaming Genomes

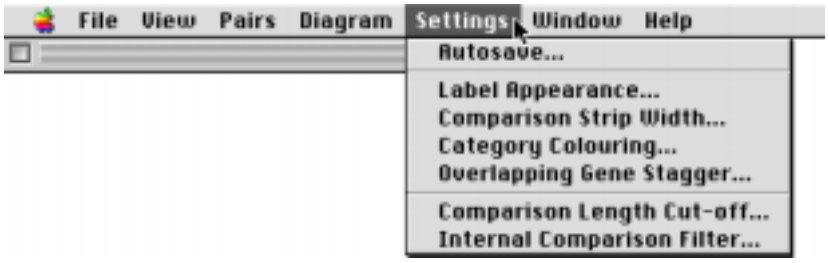
If you load two files which specify the genomes of different strains of the same organism, the names appearing initially in the graphical display may be the same, although the dynamic menus in which the names appear will have the accession number appended so that you can distinguish between them. You will find it preferable to edit the names so that you can see which genome is which on the graphical display. To do this:

1. Select 'Rename Genome' from the View menu.
2. Select the current name of the genome from the drop-down menu, edit its name in the edit field, click 'Update', and then 'Close', unless you wish to edit the names of any other loaded genomes.

3. Working in the *BugView* Browser

The main area of *BugView* is a resizeable scrollable *window* in which the graphic representation of the genome(s) resides. Operations are performed using pull-down menus from the menu bar, the controls on the control console, contextual pop-up menus, or with mouse/keyboard shortcuts.

The Menu Bar



General commands are found in the pull-down menus under the menu bar: the *File* menu for file manipulation, the *View* menu for altering the general appearance of the window and the objects in it, the *Pairs* menu for commands relating to gene comparison pairs, the *Settings* menu for setting user preferences, the *Window* menu for moving between different windows (section 11), and the *Help* menu for help in using the program. The menu items are described in detail in other sections.

The Control Console

Most of the controls in the control console are used with genes and other objects in the window that have been selected with the cursor. They are dimmed if a particular option is not available. Those concerned with editing or viewing object data are described in a later section. Others concerned with the size and position of objects are summarized here.

- The Zoom scroller increases the linear magnification of objects in the window up to 512×.
- The 'Focus On' button zooms a selected gene so that it is at its maximum size within the window, and centres it there. The 'Focus Off' button returns the zoom to 1×, whether or not the gene is still selected.
- The 'Centre' button centres a selected gene in the window, to the extent that the current magnification allows. It is not available at 1× magnification, and will not fully centre peripheral genes at low magnifications.

In addition, there is a text area in which the name of a selected object is displayed.

The Window

The only obvious controls in the window are the scrollbars, which become available when part of the genome map lies outside the visible area. However contextual pop-up menus can be invoked to allow the user who is familiar with the controls to access them with less physical movement. The pop-up menus are invoked by pressing the platform-specific 'modifier' after selecting an object or before holding down the cursor elsewhere in the window. The modifier keys are:

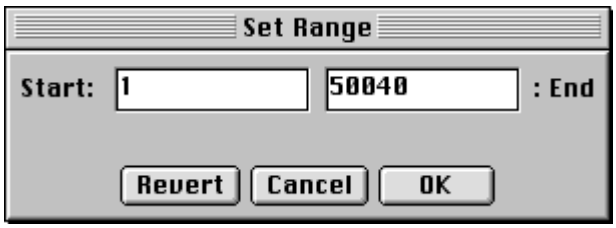
Mac:	Control Key
Windows:	Right Mouse Button
Unix/Linux:	Third Mouse Button

Alternative ways of Scrolling

As an alternative to using the scroll bars, the window can be scrolled interactively by simple mouse-dragging up or down. This allows the finest control at higher magnifications. One can also use the keyboard up-arrow and down-arrow keys to scroll half a page at a time, or the page-up and page-down keys to scroll a full page at a time. It is possible to centre on a particular point in the window by double-clicking at that point.

Viewing a Sub-Section of a Genome

As the genomes handled by *BugView* increase in size, the visibility of individual genes at maximum zoom size may become inadequate. It is possible to overcome this problem by restricting the display to a sub-section of the total genome. To do this:



1. From the View menu, select 'Set Genome Display Range'.
2. A dialogue box appears with the current start and end co-ordinates of the genome in the display.
3. Type in co-ordinates for the new range you wish to view and click 'OK'.

Temporarily Hiding a genome

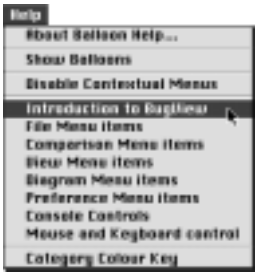
It is possible to temporarily hide a genome or the comparison pairs without having to remove them from *BugView* and reload them from disc later.

'Select Genomes to Display' in the View menu presents you with a dialogue from which you can choose what to hide or show again. Note that if a genome to which a comparison file refers is hidden, the comparison strip is also hidden. If a comparison strip is shown again, its associated genes are shown in concert.

Altering Label Size and Visibility

Selecting 'Label Appearance' in the 'Settings' menu presents a dialogue box in which the user can change aspects of the appearance of the name labels on genes. The labels can be hidden and shown again, or the size of the text can be changed from the default point size of 9 to values of 10, 11 or 12, to allow for differences in screen resolution and the user's eyesight. The background on labels of paired genes can be suppressed, e.g. for generating clearer gif images.

Help



The Help menu contains brief descriptions of the functions of the different menu items and the Control Console buttons. It includes a list of mouse and keyboard short-cuts under 'Mouse and Keyboard control', which can also be accessed from the keyboard with Help (Mac and Unix) or F1 (PC).

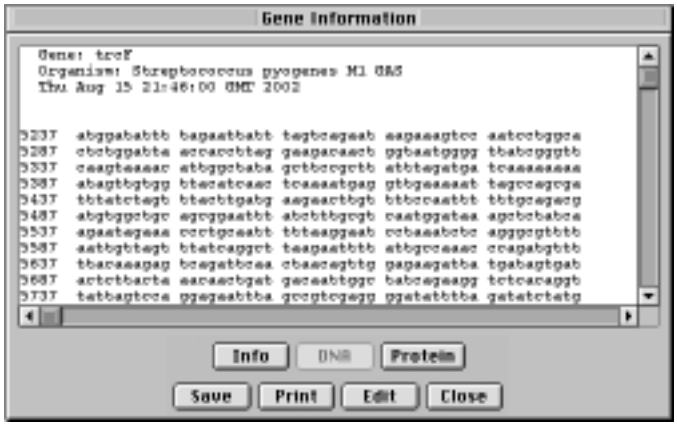
4. Viewing and Editing Gene Information

Viewing Gene Information

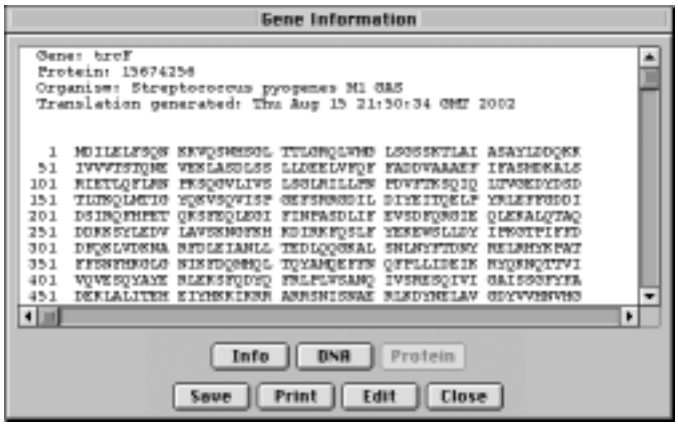
Genes are selected by single-clicking, when they become highlighted in red. With a gene selected, an information dialogue box can be invoked from the 'Gene Info' button in the 'View' group on the second row of the control panel. Double-clicking the gene has the same effect.

- The default view is of the information regarding the gene. This currently includes its position in the genome, its name, its ID, its product, the functional category to which it has been assigned, the best (external) BlastP hit, user comments, and user experimental data.

- The 'DNA' button displays the numbered DNA sequence of the coding strand.



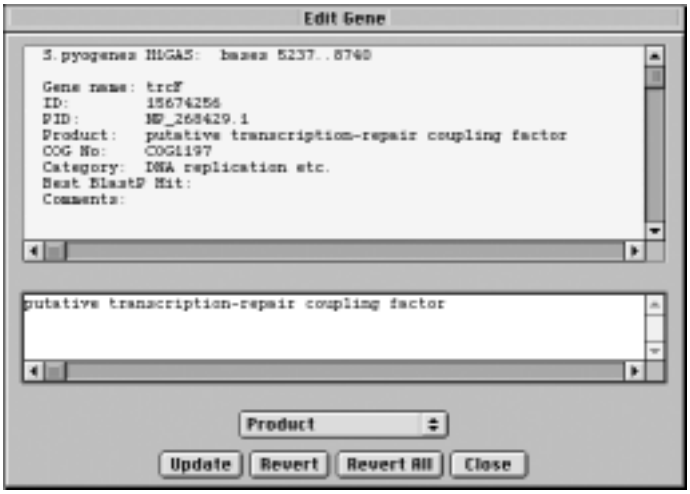
- The 'Protein' button displays the protein translation (spliced if appropriate).



- 'Save' and 'Print' bring up dialogues for saving or printing the contents of the text area.

Editing Gene Information

1. To add or edit gene information one presses the ‘Edit Info’ button in the ‘Gene’ group, This launches a dialogue with two text areas, the upper one containing the information that cannot be edited from within *BugView*, the lower showing the first field available for editing and its current content. (It is also possible to transfer directly from gene ‘view’ to gene ‘edit’ mode.)



2. A pull-down list allows one to select the field one wishes to edit. To insert a paragraph break in the edited text, press the ‘return’ key twice.
3. Changes are reflected in the upper ‘info’ window on clicking ‘Update’ or on changing to another edit field. It is possible to revert to the original text of a single field or all fields.
3. On finishing editing a gene, one clicks ‘Close’, which automatically implements the last edit. Note that all changes must be saved to disc manually, unless ‘Autosave’ is in use (below).

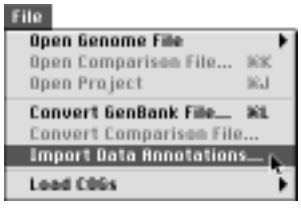
Saving Gene Information

So as to speed up the operation of the program, changes made are *not* automatically saved to disc unless ‘Autosave’ has been selected in the Settings menu

You may manually save Data file(s) to disc at any time (and are advised to do so periodically). You will always be prompted to save changes on unloading a genome or quitting. To save manually, select ‘Save File’ from the File menu (or use its keyboard equivalent). There is also the standard option, ‘Save File As’, to save the Data file for a particular genome (but not the Sequence file) under a different name.

Incorporating User Edits into Updated Genome Files

If an updated version of a genome appears, it will be necessary to generate new *BugView* Data and Sequence files. User edits can be imported from the old Data file using ‘Import Data Annotations’ from the File menu. The user is presented with the option of which fields of editable data to import. Note that revised genes have new *GenBank* GI numbers, and version numbers of ‘protein_id’ will also change.



5. Assigning a Category to a Gene

An important piece of information regarding a gene is the function of its product. In *BugView* this is recorded in the 'Category' field associated with a gene. There are 21 categories available and each is colour-coded to aid identification when viewing the genome.

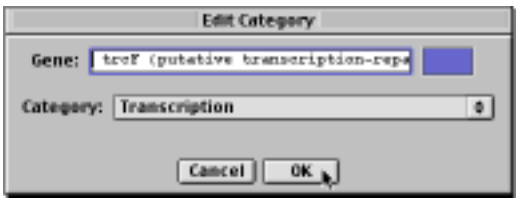
Category Descriptions and Colour-coding

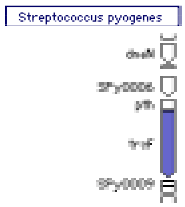
The categories are based on those used in COG — Cluster of Orthologous Groups of proteins (http://www.ncbi.nlm.nih.gov/COG/) — but have been extended to accommodate structural RNAs, inactive genes and some other descriptors. Two colour schemes are available. There is the one used on the web for COG, but, because the colour differentiation in this is quite subtle, *BugView* has its own colour scheme which is the default.

The list of categories and two colour schemes can be viewed by selecting 'Category Colour List' from the Help menu. The scheme used can be changed in a dialogue box accessed by selecting 'Category Colouring…' from the Settings menu.

Assigning a Colour Category to a Gene

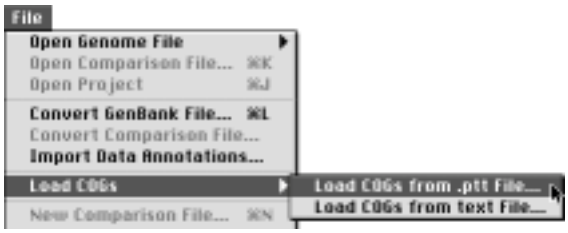
To assign a category to a single gene one presses the 'Edit Category' button in the 'Gene' group, makes a selection from the available list, and clicks 'OK'. The colour of the gene in the window changes to reflect the new category. GenBank rRNA and tRNA features and CDS '/pseudo' features are automatically assigned an appropriate category when the files are converted.





Assigning Colour Categories

In some cases COG categories are included in the .ptt file for a genome that is found with the .gbk file on the *GenBank* FTP site. In other cases one can assemble a file oneself from data on the COG website.



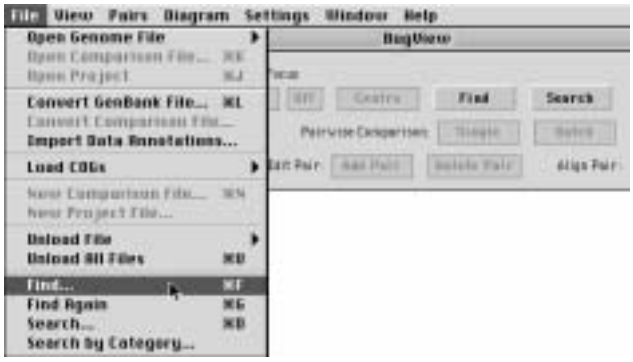
To load the COG categories from a .ptt file into your *BugView* genome data file, first load that file and then select 'Load COGs from .ptt File…' from the 'Load COGs' sub-menu in the File menu. If you collect the data yourself, prepare a text file in the following tab-separated format:

```
GeneName   COGNo  COGCategoryLetter
```

and load it by selecting 'Load COGs from text File…' from the sub-menu. When next you save the genome, the information will be written to the data file.

6. Finding and Searching for Genes

The ‘Find’ and ‘Search’ facilities in *BugView* allow one to look for genes corresponding to a particular text string or item category. In ‘Find’ the dialogue box is dispelled and one goes immediately to the found gene. In ‘Search’ one obtains a list of ‘hits’ which match the criteria, and one is then able to select from the list. The ‘Find’ and ‘Search’ options can all be accessed through pull-down or pop-up menus.



A related facility for obtaining a list of Comparison Pairs is described in section 8.

Find and Find Again

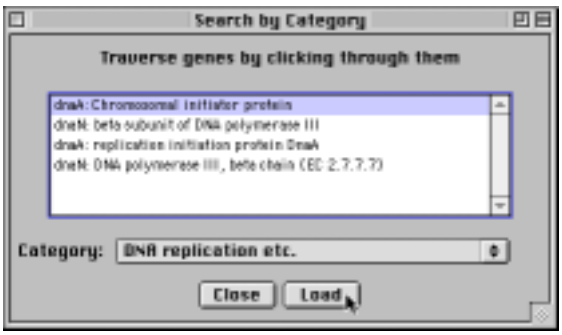
1. All the genes loaded into *BugView* can be searched by name using the Find button, the ‘Find’ item in the File menu or in a context-sensitive menu, or its keyboard equivalent (command-F or control-F for Mac or Windows, respectively) . This brings up a dialogue box into which a query term is entered and a gene characteristic (name, ID or product) selected. The first gene found with the description of the selected characteristic containing the query is highlighted and, if possible, centred in the window. ‘Find’ is not case-sensitive and does not require a complete word.
2. ‘Find Again’ from the File menu, or more conveniently its keyboard equivalent (command-G/ control-G), allows one to move successively through all the genes found by the query.

Search

1. The dialogue box for ‘Search’ is similar to that for ‘Find’, but has an additional text area to display the results of the search.
2. One may select a result of interest with the cursor or step through the list using the ‘down arrow’ key. The selected gene will, if possible, be moved to the centre of the window.
3. The ‘Search’ window remains open to allow one to return to the found list if one wishes. (Tip: If you click within the window the found gene will become de-selected. To prevent this, click within the control panel.)

Search by Category

This is similar to ‘Search’, but presents a drop-down list of gene categories for one to choose from. There is no control panel button for this option — it must be accessed from the menus.

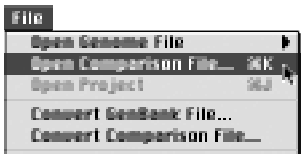


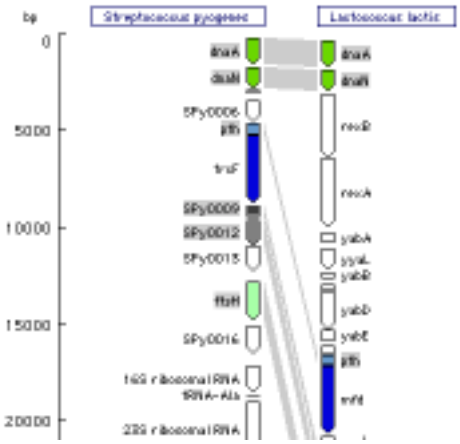
7. Loading and Creating Comparison Files

To compare genomes, a ‘Comparison File’ is needed specifying which genes in the two genomes are related, together with percentage identity scores. The file format is described in Appendix I.

Loading Comparison Files

1. Only one Comparison file may be loaded into *BugView* at any time, and Comparison files may only be loaded if the genomes to which they refer are already loaded.
2. Comparison files, which should have the file extension ‘.gcf’, are loaded by selecting ‘Open Comparison File’ in the File menu.





3. The type of display is illustrated. The order in which the genomes are displayed in the window is that specified in the comparison file, and they may rearrange after loading the latter. Where a comparison pair has been assigned to genes it is represented by a grey strip between the genes. (The labels of the second genome are automatically reoriented to the right to accommodate these strips.)

Converting Comparison Files from pre-1.2 Format

A change in the format of comparison files was introduced in *BugView* 1.2 . To convert pre-1.2 files, load the corresponding data files and then select ‘Convert Comparison File’ from the File menu.

Removing Comparison Files

Comparison files can be completely removed from *BugView* without quitting using ‘Unload File’ from the File menu, as for genome files. They are also automatically removed if either genome to which they refer is removed.

Creating a Comparison File from within *BugView*

The most practical way to create Comparison files for large genomes is externally, using standalone *Blast* (see below). However there may be circumstances where the user wishes to create a new Comparison file from within *BugView*.

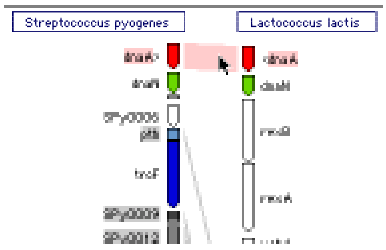
1. Select ‘New Comparison File...’ from the File menu.
2. If there are two Data files (and no Comparison file) present, a standard dialogue for saving the Comparison file will appear.
3. The Comparison file contains no information other than the genomes to which it refers. To populate it the user assigns gene pairs. (Before version 1.3.3 the comparison file had to be loaded after creation. This is no longer necessary.)

Creating a Comparison File with *Blast*

NCBI’s standalone *Blast* may be used to compare the proteins on one genome with those in another. This can be run so as to provide a Comparison file for loading into *BugView*. A web GridBLAST facility can also be used for this purpose. Details are given in Appendices II and III.

8. Viewing and Editing Comparison Pairs

Selecting Comparison Pairs



Comparison pairs, like genes, are selected by single-clicking on the strip joining them, when they become highlighted in pink. The names of the genes in the selected pair appear in the display box. Comparison pairs are also highlighted if either of the genes that they join is selected, which is useful when the comparison strip is at an oblique angle, making it difficult to select directly.

Because of the density of genes when the display is zoomed out, and the fact that overlapping labels are suppressed, it may be difficult to identify the component genes of a comparison pair. To help overcome this problem, the labels of genes in comparison pairs have a grey background by default. Their labels appear to move outwards, as well as changing colour to pink, when the comparison pair is selected.

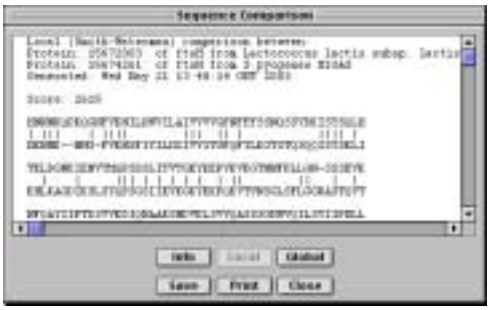
Changing the Width of the Comparison Strip

The user can specify a width for the comparison strip of between 40 and 160 pixels. (The default is 80 pixels.) A larger value can make oblique comparison strips easier to view and select. Changes are made in a dialogue box obtained by selecting 'Comparison Strip Width' from the Settings menu.

Viewing Information about Comparison Pairs

When a Comparison strip (or one of its constituent genes) is selected, the 'View Comparison Information' option become available in the buttons of the control panel or on the context-sensitive pop-up menus. (Double-clicking the pair has the same effect.)

- The default view is of the information about the comparison pair. Currently this is merely their names and parent genomes.
- The 'Local' button runs a Smith–Waterman local pairwise alignment on the protein products of the two genes. This is similar to the program 'BestFit' in the GCG package, and only displays those regions with a similarity above a certain cut-off value.
- The 'Global' button runs a Needleman–Wunsch global pairwise alignment on the protein products of the two genes. This is similar to the program 'Gap' in the GCG package, and produces an alignment of the whole of the two gene-products.
- 'Save' and 'Print' will bring up standard dialogues for saving or printing whatever is in the dialogue text area.



The total number of comparison pairs and the proportion of genes assigned to comparison pairs can be found by selecting 'Genome & Pair Summaries' from the View menu.

Adding a New Comparison Pair

Comparison pairs can be created between any two genes in the genomes specified in the comparison file, and any gene is allowed as many 'mates' as one wishes.

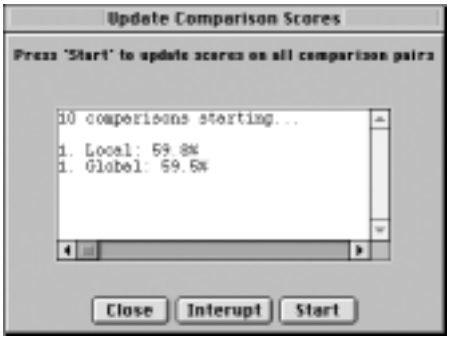
1. With an individual unpaired gene selected, click the 'Add Pair' button in the 'Edit Pair' group on the control panel. (Alternatively, a contextual pop-up menu provides this option.) A dialogue box will appear with fields for the IDs of the two genes. The left-hand field contains the ID of the selected gene, and cannot be edited. In the right-hand field you enter the ID of the gene 'mate' to be assigned from the comparison genome (this can be copied from 'Gene Info').
2. Alternatively perform a single pairwise alignment between the genes of interest (p. 17), and use the option to create a pair from the results. The IDs are then entered in the dialogue box for you.
3. On clicking 'OK', the comparison strip appears immediately to reflect the new comparison pair. (Changes must be saved manually unless 'Autosave' has been selected in 'Settings'.)

Deleting a Comparison Pair

1. With either the comparison pair selected or one of its constituent genes, click the 'Delete Pair' button in the 'Edit Pair' group on the control panel. (Alternatively, a contextual pop-up menu will provide this option.)
2. If a strip is selected or a gene with only a single 'mate', a dialogue box will appear asking you to confirm the deletion. If a gene with more than one 'mate' is selected, the dialogue will ask you to select the pair that you wish to delete. (The indication of percentage identity may be helpful here.)
3. Make a selection, if necessary, and click 'OK'. The comparison strip will be removed. (Changes made are *not* automatically saved to disc unless 'Autosave' has been selected in 'Settings'.)

Updating the Alignment Values for a Comparison Pair

When a new Comparison Pair is created, local and global alignments are run automatically (section 10), and, on saving, the values are included in the Comparison file. This is necessary to allow selective display of pairs on the basis of their similarity (section 9). When an externally edited file (e.g. from *Blast*) is first loaded, the comparison pairs will normally be set at identity values of 0%. The percentage identities for all comparison pairs in a file can be calculated (or recalculated) by selecting 'Update Pair Scores' from the Pairs menu. This may take an hour or so for a .gcf file of two bacterial genomes, so make sure you have a free machine. You should also turn off any time-dependent screen-saver, as this will slow the run down. Proteins larger than the 'Maximum Protein Length' (default 2000 amino acids) will be skipped. You can try to align these individually in *BugView* afterwards. Depending on the speed of your machine, you may wish to change this default value using the 'Comparison Length Cut-off' item in the Settings menu (section 10).



9. Aids to Working with Comparison Pairs

There are two main problems in visualizing Comparison pairs — their density and the relative alignment of partners. The following options are designed to alleviate these problems.

Setting a Comparison Pair Display Range

It is possible to set a restricted range for display of Comparison pairs. This allows one, for example, to focus on the most closely related pairs.

1. From the Pairs menu select ‘Set Pair Display Range’.
2. A dialogue box will appear. Choose either ‘Local’ or ‘Global’, depending on the basis of cut-off you require.
3. Enter a minimum and maximum value for display, and click ‘OK’.

Traversing Displayed Pairs

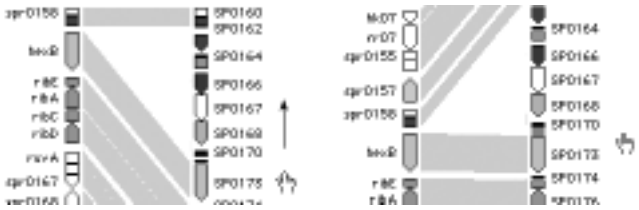
Even with the Pair Display Range set quite fine, it may be difficult to locate pairs in genome of 2000 or more genes. The Pairs menu allows access to a facility similar to ‘Search’ that lets one view a list of displayed pairs and ‘step’ through them.

1. From the Pairs menu select ‘Traverse Pairs’.
2. A dialogue box will appear. Click ‘Load’ to load all the displayed pairs in the text area. Wait. After loading, the constituent genes of the pairs are identified by their IDs.
3. Select with the mouse or ‘click-through’ the list using the ‘down-arrow’ key. The behaviour is analogous to that of the ‘Search’ function.

Related to this, and under the same menu, is an option to traverse a list of *unpaired* genes.

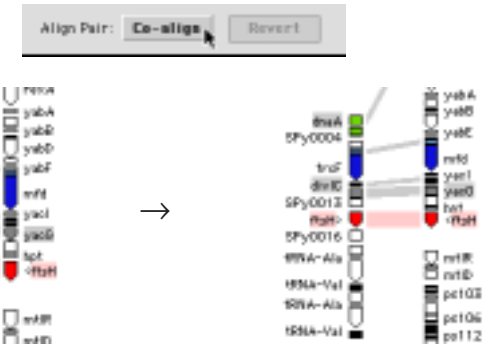
Adjusting the Relative Alignment of Genomes

It is easiest to examine the genomic regions containing related genes if they are adjacent. There are several ways of so positioning them:



- By dragging with the alt (option) key held down, fine adjustment of the right-hand genome can be performed interactively. (N.B. This feature does not allow overall realignment of genomes, as it causes one of the extremities to go off screen.)

- To co-align two genes exactly one can use the ‘Co-align’ button or the ‘Co-align Pair’ item in a pop-up menu after selecting the comparison strip or one of its constituent genes. A confirmation dialogue will appear, after which the alignment will occur.

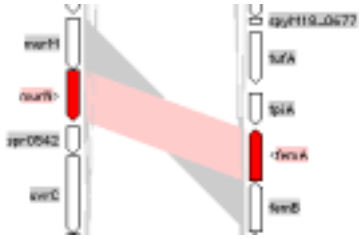


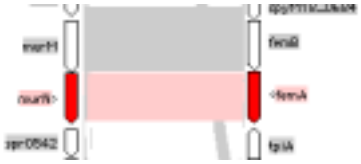
- Alternatively one can select ‘Co-align Genes’ or ‘Co-align Coordinates’ from the View menu, and in the dialogue box that appears enter the appropriate information.

The genomes can be reset to their original alignment by the ‘Revert’ button (and also from the View menu).

Reversing the Relative Orientation of Genes

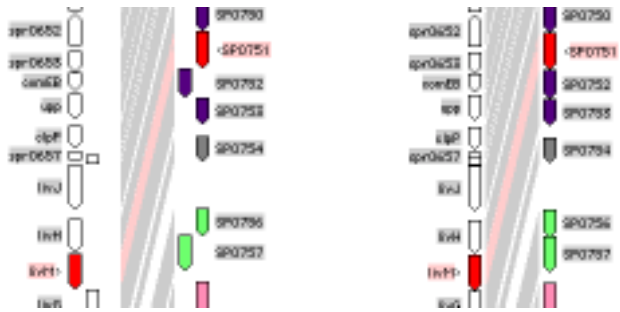
Sometimes blocks of genes in related bacteria have undergone inversion during evolution. This makes it more difficult to examine them side by side. In such circumstances the relative orientations of the genomes can be reversed by selecting ‘Reverse Directions’ from the View menu.





Turning off Gene Stagger

By default, overlapping genes are displayed ‘staggered’ in ranks so that they can be distinguished from one another. However, in certain circumstances this can make it less easy to see to which genes a particular comparison strip relates. If the genes of interest have little or no overlap, the user may wish to turn off staggering from the option in the Settings menu.



10. Running Pairwise Alignments

A simple pairwise alignment facility for gene-products was built into *BugView* for several reasons. It allows batch determination of the percentage identity of comparison pairs determined externally; it allows one to search the whole of one genome for the best match to a gene from the other genome, albeit relatively slowly; and it allows direct comparison of the products of any two genes of interest in the loaded genomes.

Although not directly related to comparison pairs, *BugView* also has facilities to search for proteins similar to a particular protein within the same genome (Internal comparison), or to compare an external protein with the product(s) of a gene or genome being displayed (Custom comparison).

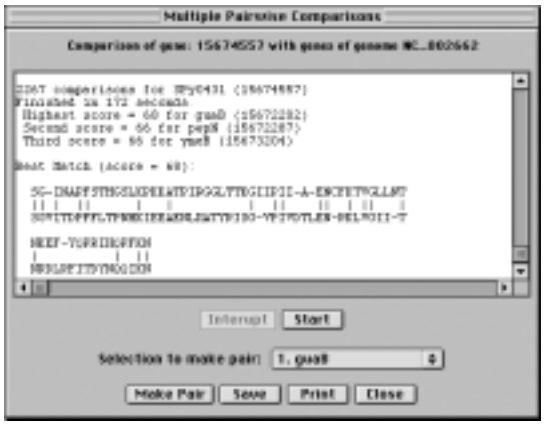
The second bank of buttons contains the pairwise comparison functions.

'Single' Pairwise Alignments

1. Select one of the genes to be compared and press 'Single' in the 'Pairwise Comparison' group. In the dialogue box enter the ID of the gene to be compared. This can be any gene displayed, including genes in the same genome as the first gene.
2. On pressing 'Start', a Smith–Waterman local alignment is performed and the 'score' and a comparison displayed.
3. The output may be saved to disc or printed. There is also the option to create a pair from the genes.

'Batch' Pairwise Alignments

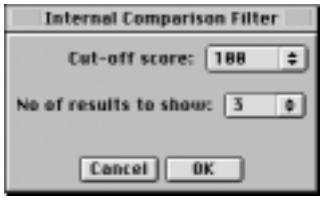
1. Select the query gene and press 'Batch' in the 'Pairwise Comparison' group.
2. On pressing 'Start', a Smith–Waterman local alignment is performed against the protein product of each gene in the other genome specified in the comparison file, with details of the best three matches presented.
3. One has the option of choosing one of the matches to create a pair with the query sequence.



'Internal' Pairwise Alignments

1. Select the query gene and press 'Internal' in the 'Pairwise Comparison' group.
2. On pressing 'Start' a Smith–Waterman local alignment is performed against the protein product of each gene in the same genome. (This type of alignment can be performed with a single genome loaded.)
3. The number of sequences displayed in the results is decided on a different basis from the other Batch

alignments — a chosen number above a particular cut-off score is used, with the default being the best three alignments with a score above 100. This allows, for example, more results to be displayed when there is a large number of related genes. The display can be altered from the default by selecting 'Internal Comparison Filter' from the Settings menu.



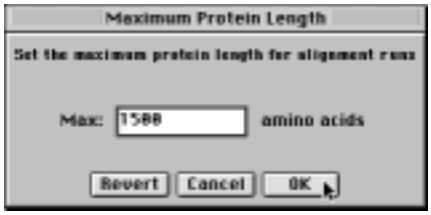
'Custom' Pairwise Alignments

1. Press 'Custom' in the 'Pairwise Comparison' group. If you have a gene selected, you will be offered the option of aligning this to a custom sequence, otherwise you select one of the genomes from the pull-down menu.
2. Paste* your protein sequence into the window. (There is no need to edit out spaces and numbers — they will be removed automatically.)
3. On pressing 'Start', a Smith–Waterman local alignment is performed between the external sequence and the local sequence or sequences, with the results displayed as for single and batch alignments, above.

Alignments and Protein Length

Even on modern machines it is possible to run into memory problems aligning very long proteins, especially in comparisons to proteins with which they have poor homology. This can cause the unsatisfactory situation that a batch comparison or identity update will fail. For this reason *BugView* has been programmed to skip alignments where one partner is longer than a particular maximum, the default being 2000 amino acids. The user is informed which proteins or pairs have been skipped and can either try them individually at a different cut-off value or on a faster machine.

The dialogue box for changing the default cut-off length is accessed from the 'Comparison Length Cut-off' item in the Settings Menu.

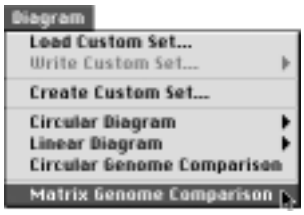


*Java 1.4 on Mac OS X has a bug that prevents one pasting sequences into this window. Users of Mac OS X 10.4 can circumvent this by installing Java 1.5 (a download on Apple's website). A future version of *BugView* will provide a file-load alternative for users of earlier versions of Mac OS X.

11. Alternative Genome Views

In *BugView* one works in the main window, which shows genomes in vertical orientation. There are, however, alternative views available, which, although not allowing access to the general *BugView* feature set, may aid orientation. These are all accessed from the 'Diagram' menu. It is generally possible to move directly from a particular region in one of these alternative views (diagrams) to the corresponding region in the main window.

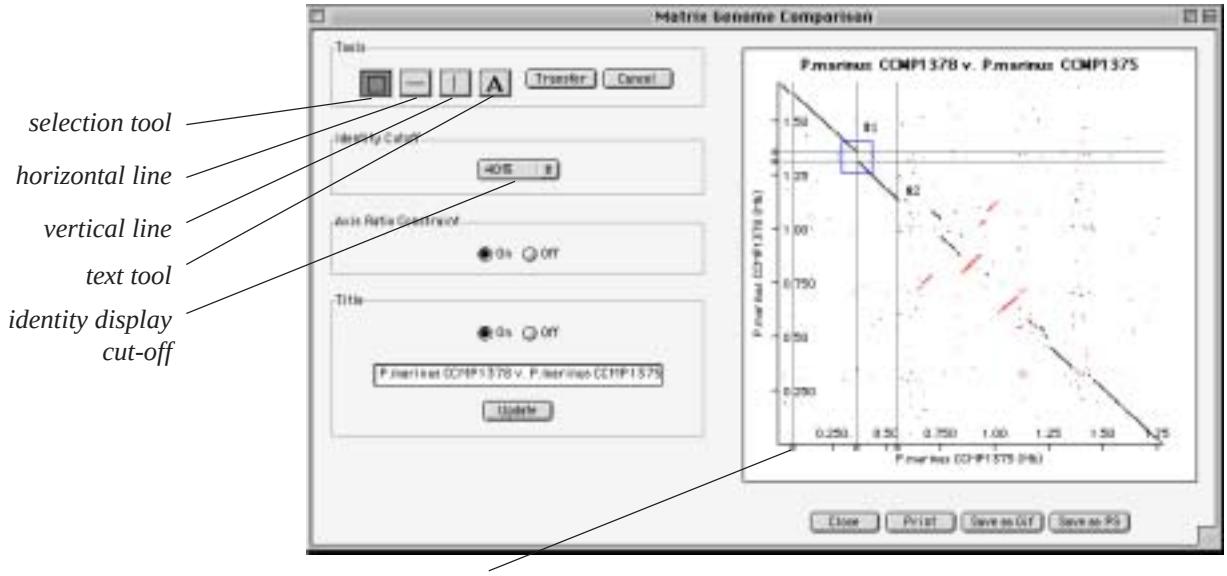
Matrix Genome Comparison



The Matrix Comparison Diagram is a simple dot-plot of the comparison pairs of two genomes. It should be emphasized that this matrix is based on the user's pre-assigned comparison pairs — it is *not* based on any programmatic whole-genome comparison within *BugView*.

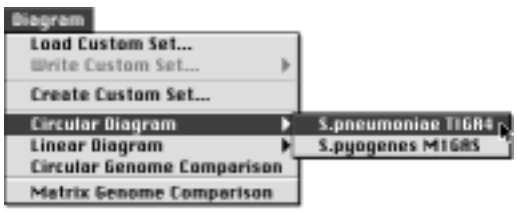
This view allow the user to locate regions of similarity between two genomes, and is especially valuable where the order of genes in the genomes have diverged somewhat. One can select an area of interest and then click the 'Transfer' button to move to the corresponding region (zoomed) in the main window. (The 'Cancel' option allows a selection to be discarded.)

One can draw guidelines on the plot to mark one's position as one works through a genome — select the horizontal or vertical guideline tool and click where you wish the line to be. The position of a guideline can be changed by dragging the square handle, and a line can be removed by option (alt) clicking the handle (cursor should change from cross-hair to arrow first). One can also make simple editable textual annotations after selecting the text tool.



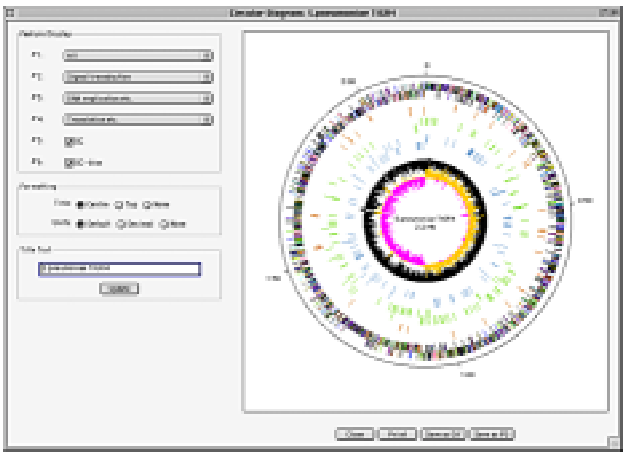
handle for moving or deleting a line (with tool selected)

Circular Genome Diagrams



The Circular Genome Diagrams present the familiar circular views of bacterial chromosomes, and were originally introduced for publication and slides. They now have some additional features that may help orientation. There is a choice between creating a circular diagram of a single genome, or a circular diagram comparing two or three genomes.

The Circular Diagram option allows the position of user-selected gene types to be visualized in relation to the genome, and also allows display of GC statistics if the sequence has been loaded.

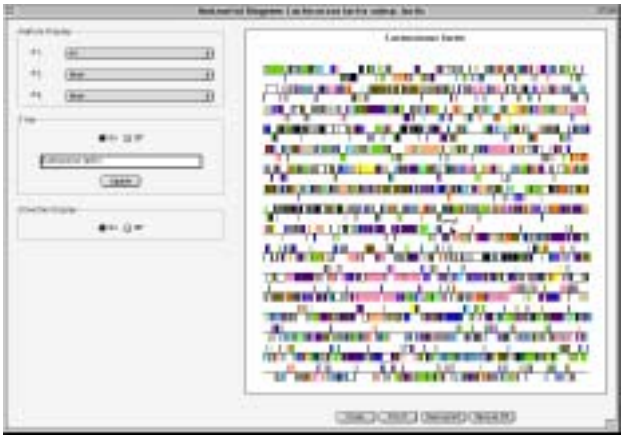


The Circular Genome Comparison allows two or three genomes to be viewed side-by-side in a circular presentation. The user can select whether to view all the genes or just genes of a particular type.

The default categories of gene that can be selected are those described in section 5. There is also a facility for users to create Custom Sets of genes to use during a session, and the option to save them to file for import and use in a subsequent session. The colouring of Custom Sets is determined dynamically. The Custom Sets only apply to the Diagram windows, not to the main program window.

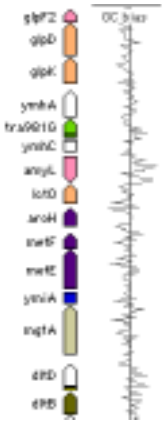
If one clicks on a point within the inner circle in the single-genome Circular Diagram one is transferred to the corresponding region in the main window. If one has made the transfer because one is interested in the GC-bias, one can display this feature in the main window (*see* below).

Linear Genome Diagram



The Linear Diagram has options quite similar to those for circular diagrams, but does not display GC statistics. Its main potential is to visualize the genome as a whole at higher resolution than in the main window or in the circular diagrams. The names of the genes appear above them on 'mouse-over', and if one clicks on a gene of interest the main window comes to the fore with that gene zoomed to the centre.

Viewing Additional Genomic Features in the Main Window



The default presentation of genomes in the main window is restricted to representation of the genes and their names, together with the relationship between genes in genomes, where this has been determined. As GC-bias values are available in the Circular Diagram, it was felt that there should be an option to display these in the main window too, and this was implemented in version 1.3.

The GC-bias display can be switched on via a dialogue box invoked from 'Display Other Features' in the View menu. At the moment GC-bias is the sole additional feature that can be displayed, although others may be added in future versions of *BugView,* should there be a demand.

Printing and Saving Graphic Views of Genomes

Any view of the genomes or their comparison can be printed or saved to disc in Gif or PostScript format. In the main window these options are in the File menu, in the other graphic displays the options are in the buttons below the display.

'Print' presents the user with a standard print dialogue box. The graphical area, only, is printed from the alternative genomic views, but the print-out from the main window is of the whole application, including the controls.

'Save as Gif' presents the user with a standard 'Save' dialogue box. The gif file produced contains only the map area without the names of the genomes. Graphics in the Gif format are bitmap images, especially suited for the web (although of unsatisfactory resolution for print work unless manipulated by a specialist).

'Save as PostScript' ('Save as PS') presents the user with a standard 'Save' dialogue box. In the main window the file produced contains only the map area without the names of the genomes. PostScript is a vector graphic format, which can be resized and edited for print publication using either specialist Unix PostScript tools or a Mac or PC vector graphics application such as Adobe *Illustrator*. Unfortunately the PostScript implementation in *BugView* is not very sophisticated, so that the output from some views may be less than ideal.

Appendix I: File Formats

Genome files and GenBank format

As bacterial genomic data is placed in the public domain in *GenBank* format, *BugView* had to be able to read *GenBank* files. This posed a problem as the *GenBank* format is flexible enough that information about genes may be presented differently by different authors. *BugView* parses *GenBank* files in what appears to be a standard format for bacterial genomes. Although *BugView* has been shown to parse *GenBank* files for over 50 bacterial genomes and several eukaryotic chromosomes, an understanding of the parsing is useful.

The minimum requirement for holding information on genomes and genes is that they each have a unique identifier (ID) and the extents of their associated sequences are specified. They also need names of some sort, but these need not be unique for the genes. The only other type of useful information about the genes that can easily be gleaned from the *GenBank* files is their products.

1. The genome information and its source in the *GenBank* file are as follows (see also illustration):

ID:	the *GenBank* Accession Number (RefSeq) read from the ACCESSION line.
Extent:	read from the 'source' line following the FEATURES line. (The No. of bp on the LOCUS line is used as a backup in case of a 'join' set on the 'source' line.)
Name:	read from the '/organism' qualifier on the following line.
The sequence:	read from the line following the ORIGIN line to the terminating '//'.

```
LOCUS       NC_003098            2038615 bp    DNA     circular BCT 03-OCT-2001
DEFINITION  Streptococcus pneumoniae R6 complete genome.
ACCESSION   NC_003098
VERSION     NC_003098.1  GI:15902044
KEYWORDS    .
SOURCE      Streptococcus pneumoniae R6.
  ORGANISM  Streptococcus pneumoniae R6
            Bacteria; Firmicutes; Bacillus/Clostridium group; Lactobacillales;
            Streptococcaceae; Streptococcus.
REFERENCE   1  (bases 1 to 2038615)
  AUTHORS   Hoskins,J.A., Alborn,W. Jr., Arnold,J., Blaszczak,L., Burgett,S.,
            DeHoff,B.S., Estrem,S., Fritz,L., Fu,D.-J., Fuller,W., Geringer,C.,
            Gilmour,R., Glass,J.S., Khoja,H., Kraft,A., LaGace,R.,
            LeBlanc,D.J., Lee,L.N., Lefkowitz,E.J., Lu,J., Matsushima,P.,
            McAhren,S., McHenney,M., McLeaster,K., Mundy,C., Nicas,T.I.,
            Norris,F.H., O'Gara,M., Peery,R., Robertson,G.T., Rockey,P.,
            Sun,P.-M., Winkler,M.E., Yang,Y., Young-Bellido,M., Zhao,G.,
            Zook,C., Baltz,R.H., Jaskunas,S.Richard., Rosteck,P.R. Jr.,
            Skatrud,P.L. and Glass,J.I.
  TITLE     Genome of the bacterium Streptococcus pneumoniae strain R6
  JOURNAL   J. Bacteriol. 183 (19), 5709-5717 (2001)
  MEDLINE   21429245
   PUBMED   11544234
REFERENCE   2  (bases 1 to 2038615)
  AUTHORS   NCBI Microbial Genomes Annotation Project.
  TITLE     Direct Submission
  JOURNAL   Submitted (02-OCT-2001) National Center for Biotechnology
            Information, NIH, Bethesda, MD 20894, USA
COMMENT     PROVISIONAL REFSEQ: This record has not yet been subject to final
            NCBI review. The reference sequence was derived from AE007317.
            COMPLETENESS: full length.
FEATURES             Location/Qualifiers
     source          1..2038615
                     /organism="Streptococcus pneumoniae R6"
                     /strain="R6"
                     /db_xref="taxon:171101"
```

2. The gene information and its source in the *GenBank* file are as follows (see also illustration). Reading is done in the order indicated and starts on recognition of a gene by the line flags ‘CDS’, ‘tRNA’ or ‘rRNA’, which also act as terminators.

Co-ordinates:	read from any line starting with ‘CDS’, ‘tRNA’ or ‘rRNA’. ‘Joins’ are handled to produce the correct translation, and introns are represented graphically. Greater- or less-than characters are ignored.
Name:	read from any ‘/gene’ qualifier on the next line or computer-generated.
ID:	read from any ‘/db_xref="GI:…’ qualifier or computer-generated.
Product:	read from any ‘/product’ qualifier or set as the RNA name or set at “”.
PID:	read from any ‘/protein_id’ qualifier.
Code:	read from any ‘/transl_table=’ qualifier.

```
gene            2722..2916
                /gene="spr0003"
CDS             2722..2916
                /gene="spr0003"
                /codon_start=1
                /transl_table=11
                /product="Hypothetical protein"
                /protein_id="NP_357597.1"
                /db_xref="GI:15902047"
                /translation="MYQVGNFVEMKKPHACTIKSTGKKANRWEITRVGADIKIKCSNC
                EHVVMMGRYDFERKMNKIID"
gene            3000..4115
                /gene="spr0004"
```

These are the only fields used from the *GenBank* qualifiers (except for any ‘/pseudo’ associated with a CDS feature). Note that from version 1.2 the unique ID is the GI (GeneInfo Identifier).

3. *BugView* reads the following additional fields that the user may define:

Category:	read from any ‘/category’ qualifier.
Best BlastP hit:	read from any ‘/bestBlastP’ qualifier.
Comments:	read from any ‘/comments’ qualifier.
Experimental:	read from any ‘/experimental’ qualifier.

Thus when the .gbk file has been converted to a .gda file it will be simplified to the type of format shown below:

```
ACCESSION   NC_003098
FEATURES             Location/Qualifiers
     source          1..2038615
                     /organism="S.pneumoniae R6"
     CDS             4382..4951
                     /gene="SP0005"
                     /protein_id="NP_344558.1"
                     /product="peptidyl-tRNA hydrolase"
                     /db_xref="GI:15899954"
                     /category="Not yet assigned"
     CDS             4952..8461
...
```

The .seq file is in the format:

```
SEQUENCE   NC_003098
ORIGIN
      1  ttgaaagaaa aacaattttg gaatcgtata ttagaatttg cacaagaaag actgactcga
     61  tccatgtatg atttctatgc tattcaagct gaacttatca aggtagagga aaatgttgcc
....

2038501  tggataaagt ttggtaacat tgtggattat ttttcacagc ttgtggaaaa ttcttgctat
2038561  ctatggtaaa atatctctag tattaaactt ttaaatagta aaggaggaga aagga
//
```

Comparison file format

This is a simple format containing:

A first line in the form: HOMOLOG *genomeID1|genomeID2*

Any number of subsequent lines in the form *geneID1|geneID2|local score|global score*

where the local and global scores are according to Smith &Waterman, and Needleman & Wunsch, respectively.

```
HOMOLOG NC_003098|NC_003028
15902045|15899950|0.9977925|0.9977925
15902046|15899951|0.994709|0.994709
15902047|15899952|0.984375|0.76829267
15902048|15899953|0.99730456|0.99730456
15902049|15899954|0.98941797|0.98941797
15902050|15899955|0.9948674|0.9948674
...
```

Note that from version 1.2 of *BugView* the geneIDs (1590245 etc.) are the *GenBank* GI numbers. Older comparison files need to be converted to the new format (section 7).

Project file format

This is a simple format containing five separate lines:

Filename of data file from Genome 1
Filename of sequence file from Genome 1
Filename of data file from Genome 2
Filename of sequence file from Genome 2
Filename of comparison file for two genomes

e.g.

```
NC_003098.gda
NC_003098.seq
NC_003028.gda
NC_003028.seq
NC_003098-NC_003028.gcf
```

How *BugView* reads .ptt files

The header and first data line of a typical .ptt file is shown below.

```
Lactococcus lactis subsp. lactis, complete genome - 0..2365589
2267 proteins
  Location Strand Length     PID      Gene Synonym  Code COG     Product

  358..1725   +     455    15671983   dnaA  L0274    L  COG0593 replica…
```

BugView parses the .ptt files in the following way. It uses its PID (actually the *GenBank* GI number and *not* the field of the same name used in *BugView* for the GenBank ‘protein_id’) to identify the particular protein, picks up the category code from the ‘Code’ field, and converts this to a text entry for the ‘/category’ line. The ‘COG’ field is also imported into *BugView*, and is incorporated into a new ‘/db_xref="COG:’ line:

```
CDS             4382..4951
                /gene="SP0005"
                /protein_id="NP_344558.1"
                /product="peptidyl-tRNA hydrolase"
                /db_xref="GI:15899954"
                /db_xref="COG:COG0193"
                /category="Translation etc."
```

Appendix II: Creating Comparison Files from Standalone *Blast*

Obtaining *Blast*

Standalone *Blast* (including documentation) can be downloaded from:

`ftp://ftp.ncbi.nih.gov/blast/executables/`

and installed on a local Unix machine. This includes several flavours of *Blast* (*Blastall* is what is actually used here) and the *formatdb* program for creating databases. If you are installing it yourself, follow the instructions in the 'Read Me' for creating an .ncbirc file to point to the *Blast* directory.

Obtaining Genome Protein Sequence Files

Genome sequence files are at `ftp://ftp.ncbi.nih.gov/genomes/`. There are about a dozen files for each genome, but the .faa file is all that you need for a protein database. In this file you will find translations of all the genes (from the .gbk file) in FastA format. The only identifier that you can rely on being on the description line is the 'gi' number. Originally *BugView* used the 'ref' identifier, but its omission from more recent *GenBank* .faa files provoked the format change in version 1.2.

```
>gi|15902045|ref|NP_357595.1| DNA biosynthesis, initiation, binding pr...
MKEKQFWNRILEFAQERLTRSMYDFYAIQAELIKVEENVATIFLPRSEMEMVWEKQLKDIIVVAGFEIYD
AEITPHYIFTKPQDTTSSQVEEATNLTLYDYSPKLVSIPYSDTGLKEKYTFDNFIQGDGNVWAVSAALAV
SEDLALTYNPLFIYGGPGLGKTHLLNAIGNEILKNIPNARVKYIPAESFINDFLDHLRLGEMEKFKKTYR
SLDLLLIDDIQSLSGKKVATQEEFFNTFNALHDKQKQIVLTSDRSPKHLEGLEERLVTRFSWGLTQTITP
PDFETRIAILQSKTEHLGYNFQSDTLEYLAGQFDSNVRDLEGAINDITLIARVKKIKDITIDIAAEAIRA
RKQDVSQMLVIPIDKIQTEVGNFYGVSIKEMKGSRRLQNIVLARQVAMYLSRELTDNSLPKIGKEFGGKD
HTTVIHAHAKIKSLIDQDDNLRLEIESIKKKIK
>gi|15902046|ref|NP_357596.1| DNA biosynthesis; sliding clamp subunit,...
MIHFSINKNLFLQALNITKRAISSKNAIPILSTVKIDVTNEGVTLIGSNGQISIENFISQKNEDAGLLIT
SLGSILLEASFFINVVSSLPDVTLDFKEIEQNQIVLTSGKSEITLKGKDSEQYPRIQEISASTPLILETK
```

You need to transfer this file to a directory in your Unix filespace where you do all the preparatory work.

Creating Databases

Making sure that you are working in the directory with the .faa files, you run a program to create a database that *Blast* can use. The *BugView* website provides a simple script, *blastdbprep*, that will prompt you for the parameters needed for this if you wish. A typical session with it is shown:

```
bugs> blastdbprep

--- CREATE A DATABASE FOR BLASTING ---

Enter the name of the input file (.faa): PneuTigNC_003028.faa
Is this a Protein file? Y/N: Y
```

When you see the prompt again, list the files in your directory (ls -l), and you should see that a set of files of the following type has been created:

```
-rw-r-----   1 dpl1n    strep     191537 May 30 12:46 PneuTigNC_003028.faa.phr
-rw-r-----   1 dpl1n    strep      16812 May 30 12:46 PneuTigNC_003028.faa.pin
-rw-r-----   1 dpl1n    strep      16752 May 30 12:46 PneuTigNC_003028.faa.pnd
-rw-r-----   1 dpl1n    strep        116 May 30 12:46 PneuTigNC_003028.faa.pni
-rw-r-----   1 dpl1n    strep     185004 May 30 12:46 PneuTigNC_003028.faa.psd
-rw-r-----   1 dpl1n    strep       4249 May 30 12:46 PneuTigNC_003028.faa.psi
-rw-r-----   1 dpl1n    strep     595759 May 30 12:46 PneuTigNC_003028.faa.psq
```

Running *Blast* with *gcfprep*

Having prepared databases for the two genes for which you wish to construct a .gcf file, you can now blast them against each other. The *BugView* website provides a downloadable Perl script, *gcfprep*, to automate this for you.*

```
bugs:> gcfprep

--- CREATE A BUGVIEW COMPARISON FILE (v.1.3.1) ---

Enter the Query Database FastA filename: PneuR6NC_003098.faa
Enter the Accession Number of the query genome: NC_003098

Enter the Target Database FastA filename: llactisNC_002662.faa
Enter the Accession Number of the target genome: NC_002662

Please wait while the data are processed...

50 sequences processed
```

This 'blasts' each protein sequence in one genome against each in the other and records the pair with the best score. It then 'blasts' the genes in the second strand against the first, eliminates any duplicates and adds more pairs to the list. (This is because the best match for A could be X, although X could be more closely related to B, which in turn is actually more closely related to Y. Without the second run the X–B match would be missed.) The resulting .gcf file (in the example it would have the name NC_003098-NC_002662.gcf) should be transferred to your desktop machine for using in *BugView*.

It should be emphasized that the .gcf file generated by *gcfprep* does *not* contain values for percentage identities of the pairs. These are calculated when the file is first used in *BugView* (p. 14).

The *gcfprep* script seems to miss a few very short proteins. If there any of these a log file (NC_003098-NC_002662.log in this case) will be created, listing them. You can then run an individual batch comparison in *BugView* for each of these proteins that have been missed.

*This script was revised at the same time as the release of *BugView 1.2* to identify proteins on the basis of 'gi' numbers, rather than 'ref' numbers, which *GenBank* stopped using in .faa files. It has been changed again for the release of *BugView 1.3.1* to use an e-value cut-off of 0.05 for *Blast*, rather than the default of 10. This eliminates a large number of spurious matches. For those who prefer to make their own decisions on the e-value cut-off, an alternative version of *gcfprep — gcfprepE —* is available.

Appendix III: Creating Comparison Files from *GRIDBlast*

For those users who are not in a position to set up standalone Blast, web access to a BLAST grid service has been provided by BRIDGES, a UK e-Science project. This allows users to cross-blast two bacterial genomes. The output must then be parsed to convert it to *BugView* format, and the author has provided a desktop utility to do this.

Using *GRIDBlast*

Before starting, ensure that you have available the GenBank .faa files for at least one of the genomes that you wish to compare.

1. Connect to http://cassini.nesc.gla.ac.uk:9081/wps/portal.
2. You are required to register before being able to use this Grid service. There is a small link — 'Sign up' — at the top right of the page for doing this.
3. After registration you should click on 'Log in' at the extreme top right corner, when you will be taken to the login page. There you should enter your User ID and Password in the appropriate fields, and click on the 'Log in' button.
4. On the page that appears, click on the blue 'Computational Resources' tab on the horizontal bar.
5. Next click on 'GRIDBLAST Job Submission', which will take you to the page where you can run your Blast genome comparison.
6. In the first two fields, respectively, enter a job name and, if you prefer not to wait while your job runs, your e-mail address for notification of completion.
7. Clear the contents of the third field and leave it empty. Instead of pasting the large genome .faa file here, upload it from your filespace at the 'Select input file' option using the 'Browse' button. You should make a note of the RefSeq number of this file, and the fact that it will subsequently be referred to as the 'Query' genome.
8. Choose the second genome from the list on the pull-down menu. The names of the genomes, rather than their RefSeq numbers, are listed on the menu, so you will need to check to reconcile these yourself, referring to the Genbank website if necessary. Make a note of this RefSeq number as that of the 'Database' genome.
9. None of the default values of the pull-down menus is appropriate. Carefully select the following:

Blast Program	blastp
e-value	0.1 or 0.01
word size	3
generate alignments	no
include gi numbers in output	yes
output format	txt

10. Click the button entitled 'Submit Job'. It typically takes about ten minutes for a comparison of genomes with 2000 genes to run, generating an output file of about 5 Mb in size.

Creating a Comparison file from the *GRIDBlast* output

The relevant information from this output file is converted to a *BugView* Comparison file using a small utility, *BlastToGCF* (available from the *BugView* website). Launch this, choose 'Load Blast File' from the File menu, and locate and load the *GridBLAST* output file. After a short delay you should receive a message that the file has been read, with an invitation to view the list of protein pairs that has been generated. If all appears satisfactory at this stage, choose 'Write gcf File' from the file menu. You need to enter the RefSeq numbers of the 'Query' and 'Database' Genomes (as above) and then save with a suitable name and .gcf extension. The resulting file will now typically be only 50K in size.