MuGeN User Manual

Mark Hoebeke
INRA - Unité MIG (http://www-mig.jouy.inra.fr)

mhoebeke@jouy.inra.fr

MuGeN User Manual

by Mark Hoebeke

Revision History

Revision 1.0 2002-06-19 Revised by: mh
Initial release of user manual

Table of Contents

List of Tables

List of Figures

What is MuGeN ?

The *Multi-Genome Navigator*, or MuGeN, is a bioinformatics software package providing tools for exploring multiple annotated genomes along with in silico analysis results. It offers two distinct programs, one for interactive vizualization and navigation and another for the generation of images in various formats. Both programs can load annotated sequence data from a local file or retrieve it from databases across the network. Most of the parameters governing the way annotations and analysis results are displayed are customizable, either through the the graphical user interface or with command-line parameters. The following sections show how to install and to use MuGeN before describing how to format home-made analysis results in order to integrate them in MuGeN.

Installing MuGeN

System Requirements

MuGeN has been used successfully on Intel/Linux and Sparc/Solaris platforms. It is not intended for use on Windows machines. Disk space requirements are minimal (less than 1 megabyte) but memory footprint can grow as more feature-rich genomes and/or complex analysis results are loaded. For example, the display of two reasonably-sized microbial genomes (about 4 Mb each) and a box plot of the conserved portions between the two of them consumes about 70 Mb of RAM on an Intel/Linux workstation. Line plots having one or more data points per base can be especially memory-consuming.

Software dependencies

MuGeN's programs and modules are all written in Perl. They rely on a set of more or less specialized third-party modules whose list is given in Table 1 (required modules) and Table 2 (optional modules). Notice that these tables only list modules not commonly found in Perl distributions. All of these components are freely available, mostly from CPAN.

Table 1. Modules Needed by MuGeN

Component	Release	Available at
bioperl	1.0	bio.perl.org (http://bio.perl.org)
Error	0.15	CPAN (http://www.cpan.org/modules/by-module/ERROR/)
Gtk	0.7008	CPAN (http://www.cpan.org/modules/by-module/GD/)
libxml	0.07	CPAN (http://www.cpan.org/modules/by-module/XML/)
PodParser	1.18	CPAN (http://www.cpan.org/modules/by-module/Pod/)
Usage	0.10	CPAN (http://www.cpan.org/modules/by-module/Usage/)
XML-Writer	0.4	CPAN (http://www.cpan.org/modules/by-module/XML/)

Table 2. Optional Modules for MuGeN

Component	Release	Available at
EMBL data retrieval		
CORBA-ORBit	0.4.3	CPAN (http://www.cpan.org/modules/by-module/CORBA/)
Micado data retrieval		
DBD-Pg	1.01	CPAN (http://www.cpan.org/modules/by-module/DBD/)
DBI	1.20	CPAN (http://www.cpan.org/modules/by-module/DBI/)
IO-String	1.01	CPAN (http://www.cpan.org/modules/by-module/DBI/)
XEMBL data retrieval		
SOAP-Lite	0.55	CPAN (http://www.cpan.org/modules/by-module/DBI/)

Installation procedure

MuGeN is available as a gzipped archive. Download the archive in an installation directory and expand its contents by issuing:

```
gzip -dc mugen-XXXXXXXX.tgz | tar xf -
```

where `XXXXXXXX` stands for the release number. This will create a subdirectory called `mugen-XXXXXXXX` containing all of MuGeN's programs and modules. `cd` in this directory and type:

```
perl install.pl
```

This will check for the required and optional Perl modules and configure MuGeN's scripts for execution. The absence of one or more optional modules will not prevent the program's installation, only a warning will be issued.

The executables **mugenb** and **mugenv** are located in the MuGeN installation directory. By adding this directory to the $PATH variable the executables can be run from any directory.

MuGeN relies on a preferences file to fix some display and database connection parameters. By default, it looks for a file named `.mugenrc` in the user's $HOME directory. A template preferences file, called `mugenrc_template.xml` and located in the `Data` subdirectory can be used for a start and copied to the $HOME directory.

A set of example files, used throughout this document, is available in the `mugen-data-XXXXXXXX.tgz` archive. This archive can be extracted anywhere and creates a `mugen-data-XXXXXXXX` directory containing several annotated genomes in GenBank format, as well as some computer anlysis results.

Using MuGeN

The following sections offer a guided tour of MuGeN's main features and of how to make them work[1]. The examples use the data files found in the MuGeN data archive. To run the commands given in these sections, make the `mugen-data-XXXXXXXX` directory your current directory, and make sure the **mugenv** and **mugenb** commands are located in a direcory accessible through your $PATH.

Using mugenv for Interactive Genome Exploration

To start a visual exploration, just specify the name of a file containing annotated sequence data after the **-d** option. This option can be used more than once to load several files. Running **mugenb** to navigate through the complete genomes of *Bacillus subtilis* and *B. halodurans* (contained in the `Bsub.gbk` and `Bhal.gbk` GenBank formatted files) is performed by typing:

```
mugenv -d Bsub.gbk -d Bhal.gbk
```

After a (short) while, the three windows of the graphical user interface pop up.

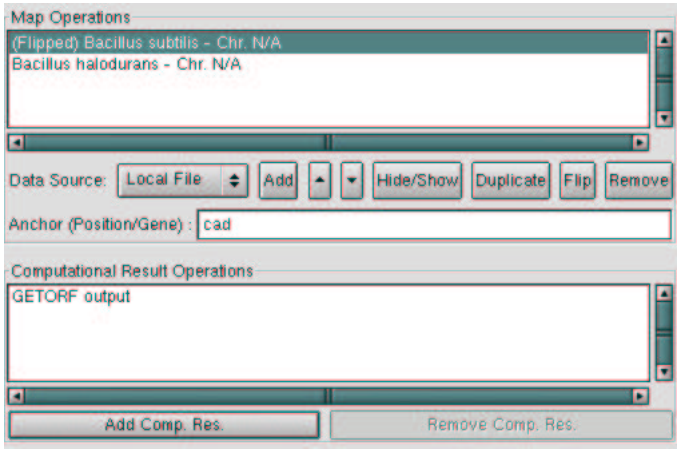
The Map List Window

This window (Figure 1) displays all loaded maps and computer analysis results. It also allows the manipulation of these maps and associated analysis results with the button row located below the map list. To load a new map, select a data source from the available sources in the popup menu, the click on the Add button. Depending on the datasource, some additionnal information will be requested (typically a filename or an access number). It may be useful to work with several copies of the same map (for instance to compare different portions of the same genome). The order in which the maps are displayed can be modified with the two arrow buttons. They shift the currently selected map up or down. A map can be hidden and redisplayed with the Hide/Show button. To generate a new copy of a given map, select it in the map list and click on the Duplicate button. Any map can be "flipped" with the Flip button, meaning that the base positions decrease from left to right, instead of increasing, and that the strands of the features are switched: features ont the forward strand move to the reverse strand and vice versa. This feature is useful to compare genome portions which are conserved but whose directions are opposite. Finally a map can be removed using the Remove button. Notice that if there is only one map in the list, it cannot be removed.

Below the map operations panel, an Anchor textfield can be found. Each map can have it’s own anchor which "fixes" its relative position. An anchor is either an integer value, representing a base position, or a gene name. In the latter case, the start position of this gene (if it exists in the selected map) will be used as anchor. Moreover, the map will be flipped if the gene is on the reverse strand. Anchors are useful to simultaneously display distant portions of genome maps. For example, after loading two genome maps of closely related organisms, the context of a gene bearing the same name in the two organisms can be examined by selecting each map in turn and entering the common gene name in the anchor textfield. In the case of *B. subtilis* and *B. halodurans*, a possible anchor for both genomes is the *cad* gene.

The remaining part of the map list window contains a list of computer analysis results loaded for the currently selected genome map. Such results can be added (respectively removed) through the Add Comp. Res. (resp. Remove Comp. Res.) button. A sample analysis result file `Bsub_orfs.xml` contains all ORFS over 300 bp detected by the **getorf** program included in the EMBOSS package.

Figure 1. The Map List Window



The map list window with two genome maps (*Bacillus subtilis* and *Bacillus halodurans*). The selected map (*B. subtilis*) is anchored on the *cad* gene which has caused the map to be flipped. A computer analysis result, *GETORF output* has been added to the *B. subtilis* map.

The Map Drawing Window

This window gives a graphical display of the annotated genome maps along with the computer analysis results. The main area is divided in "strips" or lines. Each strip represents either a portion of an annotated genome or a portion of a computer analysis result. When several annotated maps are loaded, their strips are displayed one above the other (i.e. the first strip of the first map followed bu the first strip of the second map followed by the second strip of the first map etc.). In that case, each map will have a different background color, ranging from white to light grey. When computer analysis results exist for a given map, they are displayed either on the map itself, or they are allocated separate strips immediately below the map they belong to.

By default, six lines per strip are used to draw CDSs, on for each reading frame of each strand. Other features are drawn either on the axis, if they are positional features (promoters, terminators, RBSs), or on a separate line below the CDS lines if they extend more than a dozen bp. (different RNAs, miscellaneous features and others). Also by default, CDSses are colored according to the strand they are located on, and filled if they have a known function, and empty otherwise.

Two other view modes are also available:

- a *bird's eye* view: this view is adapted to display large portions of genome maps. It is automatically activated when more than 50 Kb per line are shown. In bird's eye view mode, all features are drawn as simple boxes, or little sticks and are no more reactive.
- a *sequence* view: this is the view mode for lines smaller than 100 b. It shows the nucleotide sequence as well as its translation in the six reading frames.

The majority of display settings can be modified with the user controls at the bottom of the Map Drawing Window or with the menu entries it offers. The topmost row of user controls contains arrow buttons to

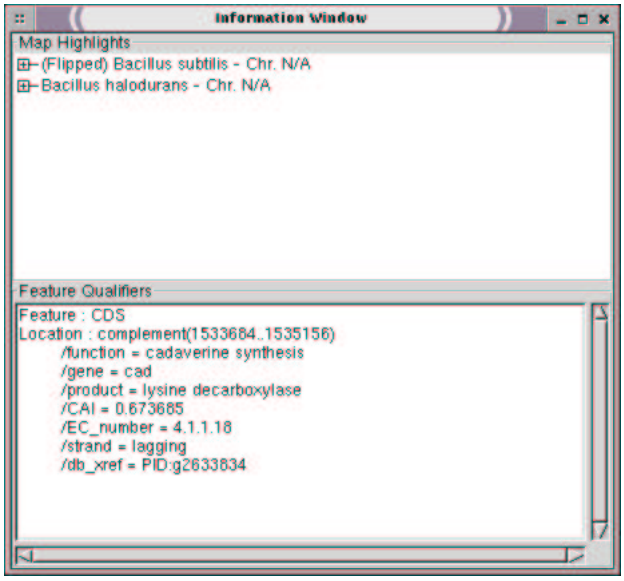
move forward or backward alon the maps. The row below allows them be to zoomed in or out. Precise starting points, number of lines and bases per line can be set with the text fields below the zoom buttons. Finally, the thresholds for switching between the different view modes can be fixed with the sliders at the bottom of the window. The rightmost slider defines the minimum relative size for features whose names will be displayed. The Preferences menu offers several items influencing the map display:

- Expand Strands: When checked, features belonging to different strands will be displayed on separate lines. Otherwise they will be displayed on the same line.
- Show Frames: When checked, CDSs are displayed on different lines acccording to their reading frame.
- Visible Features: This submenu offers one entry per feature type. Only the checked featured are displayed on the map.
- Map Area Width: The width in pixels of the area on which the maps are drawn can be selected in this submenu.
- Save Preferences: The current settings of the Preferences menu are saved in the default preferences file (`$HOME/.mugenrc`).

Figure 2. The Map Drawing Window

The Information Window

Figure 3. The Information Window



Generating Annotated Genome Images

The MuGeN preferences file

Computer Analysis Result formats

MuGeN Option List

Table 3. Options common to mugenb and mugenv

Option	Multi [a]	Functionality

Option	Multi [a]	Functionality
-d *source:id*	Yes	Specifies a resource from which to load annotated genome maps. Each resource consists of two parts, a *source* and an *id*. The source can be one of **file, genbank, embl, xembl** or **micado**. When no source is specified, **file** is taken as default. The id points to the specific map in the source. When the latter is a file, the id is simply the filename (in GenBank, EMBL, BSML or fasta format). When the source is a database (**genbank, embl, xembl, micado**) the id is the access number of the database entry. Maps will be displayed from top to bottom in the order they are entered on the command line. If the *id* start with a "!" the map will be flipped.
-f *firstbase*	No	Specifies the starting point of the image to build. In the absence of any reference points, this is the first base of the map that will be located in the upper left corner of the image. If a reference point is given, the upper left corner will be the reference point offset by the amount specified by this option.
-l *lastbase*	No	Specifies the ending point of the image to build. In the absence of any reference points, this is the last base of the map that will be located in the upper lower right corner of the image. If a reference point is given, the lower right corner will be the reference point offset by the amount specified by this option.
-s *step*	No	Specifies the number of bases per display line.
-r *refpos*	Yes	Specifies a *reference position* or *anchor* for a genome map. If the reference position is an integer, the start of the displayed image will be computed by adding the value of the -f option to the integer. If the reference position is a string, MuGeN will look for a CDS feature having a gene qualifier whose value equals the given string. If such a CDS is found, it's start base will be used to compute the start of de displayed image as explained above. Moreover, if the gene is on the reverse strand, the map will be flipped. The genome map for which the reference position is defined is determined by the index of the -*r* option wrt. the -*d* option (i.e. the first -*r* option will be applied to the map defined by the first -*d*option, the second -*r* applies to the second -*d* and so on).
-c *filename[,index]*	Yes	Specifies a computational analysis results file to display with a genome map. If a comma and an *index* are appended to the filename, the result will be applied to the genome map of the corresponding index. Index 1 is the genome map loaded by the first -*d* option, index 2 the map corresponding to the second -*d* and so on.
-e *filename*	No	Specifes a file containing a color scheme to apply to displayed features.
-w *n*	No	Specifes the width in pixels of the drawing area

Option	Multi [a]	Functionality
-p *filename*	No	Specifes the preferences file to load. If no *-p* option is given, the preferenes file will be set to `${HOME}/.mugenrc`.
Notes: a. Multi options are options that can be used several times on the command line.		

Table 4. Options specific to mugenb

Option	Multi	Functionality
-o *format*	No	Specifies the output format of the image file to be generated. Valid formats are : PNG, IMAP, PS, EPS, XFIG.
-m *mediatype*	No	Specifies the media type, for PS or EPS output files. Valid types are : a7, a6, a5, a4, a3, a2, a1, a0, b7, b6, b7, b4, b3, b2, b1, b0, lettern legal, executive, ledger.
-u *urlprefix*	No	Specifies the root URL for client-side image maps in IMAP format. Parameters relative to dislayed features will be appended to this root URL. For instance, given a root URL of `http://www.somewhere.org/cgi-bin/myscript.pl?myid=xyz&`, and an image containing a CDS feature, whose name is abcX positioned from base 1234 to base 5678, the URL generated for it's clickable area will be `http://www.somewhere.org/cgi-bin/myscript.pl?myid=xyz&tag=CDS&n`

Notes

1. Table 3 details all command line options supported by MuGeN.