MCRL - Metagenomic Clustering by Reference Library

SOFTWARE INSTALLATION INSTRUCTIONS

Contents

1. Introduction

MCRL is a data mining tool that can be used to probe a metagenome for homologs of a pre-defined reference library. The input to MCRL is an assembled metagenome in nucleotide or amino acid format and a library of reference sequences in amino acid format. MCRL will then perform iterative clustering of the reference library with respect to the given metagenome and provide as output a list of nonredundant reference genes that have homologous counterparts in the metagenome.

MCRL can be downloaded from https://github.com/a-tadmor/MCRL.

2. Files provided

`MCRL_readme_vXXX.pdf`	This readme file
`MCRL_vXXX_MATLAB.zip`	MATLAB sources (zip)
`MCRL_vXXX_MATLAB.tar.gz`	MATLAB sources (tar)
`MCRL_vXXX_installer_WIN64.exe`	Windows distribution (mrc environment & MCRL)
`MCRL_vXXX_Linux64.tar.gz`	Linux distribution (MCRL)
`Plot_networks_vXXX_WIN64.exe`	Windows executable to plot reference gene networks
`Plot_networks_vXXX_Linux64.tar.gz`	Linux executable to plot reference gene networks
`license.txt`	MCRL license

3. Installation & system requirements

MCRL can be run on Windows or Linux operating systems as executables or through MATLAB. After MCRL has been installed and the user interface loads, the user can download and install the latest viral RefSeq library from NCBI by pressing the button "Download and assemble the latest viral RefSeq reference library".

3.1 Windows distribution

Requirements:

1. 64-bit Windows
2. An internet connection and administrator privileges are required for installation

Instructions:

1. Download and install Visual C++ Redistributable (required for DIAMOND) from: https://support.microsoft.com/en-us/help/2977003/the-latest-supported-visual-c-downloads
2. Download `MCRL_vXXX_installer_WIN64.exe` to a local directory such as `d:\tmp`.
3. Double click the executable. This executable will automatically install the MATLAB Compiler Runtime (mcr) library and copy the MCRL installation files to a folder designated by the user. Install MCRL in the same local directory (`d:\tmp`)
4. Create a local directory `MCRL\bin` (e.g., `d:\MCRL\bin`) and copy `MCRL.exe` from the `application` folder to the `bin` folder.
5. To complete the installation either double click `MCRL.exe` to start the graphical user interface, or use MCRL in command line mode (see below). Follow the instructions and prompts to install BLAST and MCRL (accept default entries when prompted). Depending on your system it may take a couple of minutes for the installation to begin. Note: **prompts requiring user response can be minimized in the taskbar.** If you are having trouble getting BLAST to install see section 4 for troubleshooting tips, or instructions for installing BLAST manually. Once the GUI loads you are ready to use MCRL.

Command line mode in windows

To run MCRL in command line mode navigate to the `bin` folder and type in the shell

```
$MCRL "<arguments>"
```

For example, to get the help menu type

```
$MCRL "--h"
```

Input parameters for MCRL include the following:

```
--m  <metagnome FASTA file name>                          FASTA file is assumed to be in the ..\metagenomes\ folder (do not provide path)
--f  <reference library FASTA file name>                  .faa file is assumed to be in the ..\RefSeq_databas\ folder (do not provide path)
--g  <reference library GenPept file name>        (optional)   .gpff file is assumed to be in the ..\RefSeq_databas\ folder (do not provide path)
--o  <output tag>                                 (optional)   output string to append to output files
--E  <E value threshold, Eth>                     (optional)   default 1e-7
--e  <signature E value threshold, E0>            (optional)   default 1e-3
--O  <overlap condition>                          (optional)   stringent or inclusive (default inclusive)
--T  <overlap threshold>                          (optional)   0 to 100 (default 50)
--s  <sequence type in metagenome>                (optional)   nucl or prot (default prot)
--A  <aligner>                                    (optional)   diamond_sensitive, diamond_more_sensitive, diamond_default, blast (default blast)
--p  <precompute tracks>                          (optional)   1 for yes, 0 for no (default 0)
--a  <add annotation to reference_gene_clusters_nr> (optional)   1 for yes, 0 for no (default 0)
--t  <number of threads>                          (optional)   default 1
--h                                               (optional)   List of arguments to MCRL is provided
--v                                               (optional)   Version
--d  <alinger (optional)>                         (optional)   Demo run, aligner=blast, diamond (default blast)
```

To run a short demo with BLAST type

```
$MCRL "--d"
```

The demo will create the metagenome and reference library files in the appropriate MCRL folders. If BLAST is not already installed, the BLAST installation process will automatically begin.

To run a short demo with DIAMOND type

```
$MCRL "--d diamond"
```

To execute MCRL using 4 threads on your data (without providing a `gpff` file), type for example

```
$MCRL "--m my_metagenome.faa --f my_reference_library.faa --t 4"
```

3.2 Linux distribution

Requirements:

1. 64-bit Linux*
2. An internet connection and root privileges are required to install the MATLAB Compiler Runtime (mcr) library

*MCRL was tested on CentOS Linux 7

Instructions:

1. Download and install the 2019b MATLAB Compiler Runtime (mcr) library from: https://www.mathworks.com/products/compiler/mcr/index.html
2. Create a local folder `/home/user/MCRL/bin` and download `MCRL_vXXX_Linux64.tar.gz` to the `bin` folder (the parent folder name can be arbitrary).
3. Extract the tarball in the `bin` folder by typing in the shell from within the `bin` folder

```
$ tar -xvf MCRL_vXXX_Linux64.tar.gz
```

4. To complete the installation and start using MCRL, type in the shell from within the `bin` folder:

```
$ ./run_MCRL.sh <mcr_directory>
```

where `<mcr_directory>` is the directory where the MATLAB Runtime library was installed (or the directory where MATLAB is installed on the machine). For example: `$ ./run_MCRL.sh /code/MATLAB/2019b/`

Follow the instructions and prompts to install blast and MCRL. Depending on your system it may take a couple of minutes for the installation to begin. **Note: prompts requiring user response can be minimized in the taskbar.** Alternatively, use MCRL directly from the commend line (see below).

Once the main interface of MCRL loads you may start using MCRL.

Command line mode in Linux

To run MCRL in command line mode navigate to the `bin` folder and type in the shell

```
$ ./run_MCRL.sh <mcr_directory> "arguments"
```

For example, to get the help menu type (see list of input arguments in section 3.1)

```
$ ./run_MCRL.sh <mcr_directory> "--h"
```

To run a short demo with BLAST type

```
$ ./run_MCRL.sh <mcr_directory> "--d"
```

The demo will create the metagenome and reference library files in the appropriate MCRL folders. If BLAST is not already installed, the BLAST installation process will automatically begin.

To run a short demo with DIAMOND type

```
$ ./run_MCRL.sh <mcr_directory> "--d diamond"
```

To execute MCRL using 4 threads on your data (without providing a `gpff` file), type for example

```
$ ./run_MCRL.sh <mcr_directory> "--m my_metagenome.faa --f my_reference_library.faa --t 4"
```

3.3 Running MCRL from MATLAB

Requirements:

1. MATLAB 2016a or later version
2. Supported operating systems include: Windows (32 bit, 64 bit), Linux* (64 bit)
3. Optional: Parallel Computing toolbox v4.2 or higher
4. Internet connection
5. For installation administrator privileges are required

*MCRL was tested on CentOS Linux 7. MacOS is not officially supported.

For Linux RedHat users: prior to installation of MCRL, manually download and install blast version 2.2.22+ for your OS from (see instructions for manually installing blast below).

For MAC users: prior to installation of MCRL, install conda (https://docs.conda.io/) in order to enable MCRL to install DIAMOND.

Instructions:

1. Download and install Visual C++ Redistributable (required for DIAMOND) from:

   https://support.microsoft.com/en-us/help/2977003/the-latest-supported-visual-c-downloads

2. Download `MCRL_vXXX_MATLAB.zip` (for Linux `MCRL_vXXX_MATLAB.tar.gz`)
3. Extract files to a **local** directory such as `d:\MCRL`. For Linux type

```
-xvf MCRL_vXXX_MATLAB.tar.gz
```

4. Start MATLAB, navigate to the `bin` folder in the MCRL installation and run from the MATLAB command prompt `MCRL_EXE`.
5. Follow the instructions and prompts to install blast, clicking "next" and accepting all of the default entries. **Note: prompts requiring user response can be minimized in the taskbar.**

Once the main interface of MCRL loads you may start using MCRL.

Command line mode in MATLAB

To run MCRL from the MATLAB command line on a single metagenome use `command_line_MCRL_EXE.m`. To run MCRL in batch mode on many metagenomes in parallel use `par_command_line_MCRL_EXE.m` (requires the parallel processing toolbox). Note that MCRL needs to be installed first in order to use the MATLAB command line mode.

3.4 Preinstalled demo

For demonstration purposes, MCRL comes preinstalled with a demo mini-metagenome (`demo_contigs.faa`) and demo mini-reference library (`demo_reference_library.faa`, `demo_reference_library.gpff`). The user can test run the default demo to see that MCRL is properly installed. Run time of demo is <1min.

4. Troubleshooting

Problems with MCRL installation (Windows)

1. Installation files should be downloaded to a local folder (e.g., `d:\`) and not a network drive.
2. MCRL should also be installed in a local folder (e.g., `d:\MCRL`) and not on a network drive.
3. Administrator privileges are required for installation.
4. Make sure there is no program blocking running executables (e.g., Windows Defender or some other antivirus software).

Problems with MCRL installation (Linux)

The installation tarball should be extracted in a local `bin` folder within the MCRL installation folder (e.g., `/home/user_name/MCRL/bin/`).

blast cannot be downloaded via ftp

1. Make sure you have a working internet connection
2. MCRL should be installed in a local drive and not a network drive
3. Make sure the ftp port is not blocked by the firewall, Windows Defender or some other antivirus software
4. Disconnect any active VPNs
5. Contact your IT manager to check the settings on your computer
6. If an ftp connection still fails, manually download and install blast (see instructions below)

blast cannot be executed/installed

1. Make sure there is no program blocking running executables (e.g., Windows Defender or some other antivirus software).
2. On Windows you must have administrator privileges to install blast
3. On Windows, make sure you install blast in the default folder (`C:\Program Files (x86)\NCBI`)
4. On Windows there might be a preexisting installation of blast interfering with MCRL. If Windows is installed on the `c:\` drive, check the folder `c:\Windows\` for a file called `ncbi.ini` and temporarily rename it.
5. Install blast manually (see below)

Parallel processing fails

1. If you are using the deployed version of MCRL download the most recent release of the MATLAB Compiler Runtime (mcr) library
2. If you are using the MCRL from MATLAB use the current release of MATLAB (2021 and onward).

DIAMOND cannot be executed/installed

1. Make sure there is no program blocking running executables (e.g., Windows Defender or some other antivirus software). Contact your IT manager to check the settings on your computer.

2. DIAMOND executables are provided for Windows and Linux in the local `diamond` folder in the MCRL installation and require no installation. If MCRL is not able to install DIAMOND for your OS, install DIAMOND manually for your OS following the instructions here: http://www.diamondsearch.org/index.php. For Linux, the binary files should copied to the local `diamond` folder in the MCRL installation. For MacOS DIAMOND needs to be on the global path.

Manually installing blast

If there is a problem downloading or installing the blast software, blast v2.2.22+ can be downloaded manually for your OS from the NCBI website: ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/2.2.22/. Once blast is installed, start MCRL, when prompted, click "Locate sources on computer" and locate the `bin` folder of the blast installation. For example, if blast was manually installed in `c:\NCBI\blast-2.2.22+\`, upon rerunning MCRL, locate with the browser the `bin` folder within this installation: `c:\NCBI\blast-2.2.22+\bin\`. In Linux, if blast was locally installed in `/home/user_name/ncbi-blast-2.2.22+/`, locate with the browser the folder `/home/user_name/ncbi-blast-2.2.22+/bin/`.

5. Plotting reference gene networks

5.1 Configuration file

To plot reference gene networks, first configure the setup file `config_file_network.txt` found in the `networks` folder of the MCRL installation. This configuration file has three parts:

1. Select between four options for labeling nodes:
    a. `none`: no labels are plotted
    b. `all`: all nodes are labeled with reference gene IDs
    c. `auto`: optimize label plotting to image complexity
    d. `epicenter`: plot only labels of reported reference genes

2. Enter the name of the `MCRL_table_nr.txt` output file using a full path

3. Enter one or more reported reference genes for which you wish to plot the networks. These reference genes must be reported reference genes included in the list provided in the `MCRL_table_nr.txt` output file.

Example of `config_file_network.txt` file format:

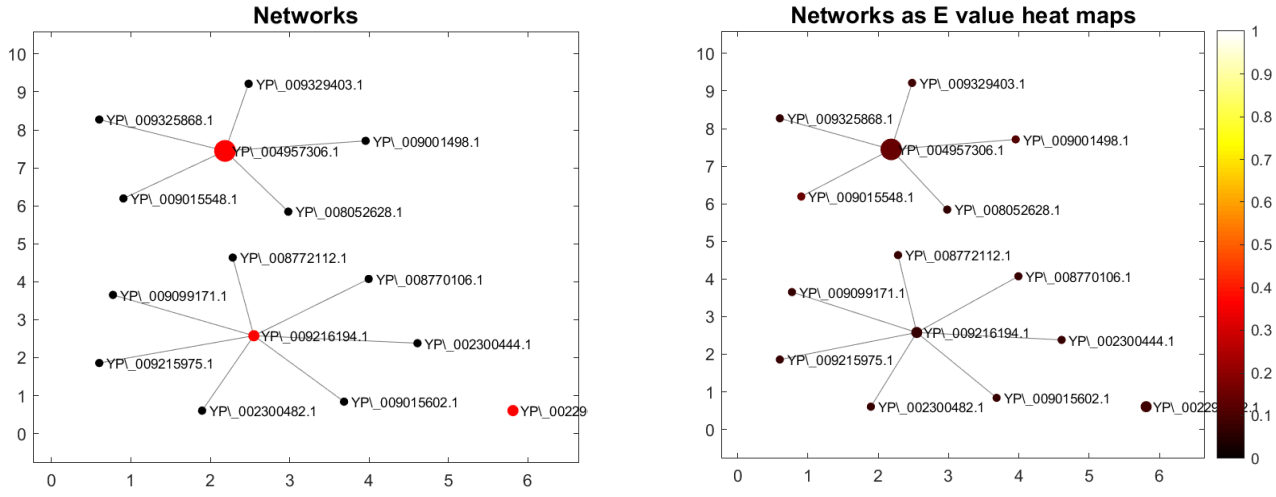
```
# Node labels (all/none/epicenter/auto)
auto

# MCRL_table_nr.txt file
C:\MCRL\FFFFFFFF_BLAST_table_EXAMPLE_MCRL_Table_nr.txt

# Reported reference gene(s)
YP_004957306.1
YP_008770526.1
YP_009238667.1
```

Initial calculation may take a long time but subsequent runs are immediate. It is possible to eliminate the initial computation time if the checkbox “Precompute tracks” is toggled in the user interface before running MCRL. In this case computation of all network tracks will be performed on the fly. Fig. 1 shows an example of networks plots.

Figure 1. Example of reference gene networks plots computed by MCRL. The scale of the heat map for the E value is logarithmic (0 corresponds to an E value of 1, and 1 corresponds to an E value of 1e-100 or lower).



5.2 Creating plots from MATLAB

To plot reference gene networks from MATLAB:

1. Edit the `config_file_network.txt` file found in the `networks` folder
2. Execute the source `Plot_networks_EXE.m` from the `bin` folder in MATLAB.

Figures will be saved to the `output` folder in the MCRL installation path with the appropriate <RUN_ID>. This script produces both MATLAB figure files (`.fig`) and tif files. If opening the `.fig` file in MATLAB does not show a figure, type `set(gcf,'visible','on')` in the MATLAB command line.

5.3 Creating plots using the Windows distribution

To plot reference gene networks using the Windows distribution:

1. Copy the binary file `plot_networks_EXE_WIN64.exe` to the local `bin` folder.
2. Edit the `config_file_network.txt` file found in the `networks` folder
3. Double click `plot_networks_EXE_WIN64.exe`,
   or, alternatively, from the command line (from the `bin` folder) type:

```
Plot_networks_WIN64.exe <MCRL path>\networks\config_file_network.txt
```

where <`MCRL_path`> is the installation path of MCRL. Figures will be saved to the `output` folder in the MCRL installation path with the appropriate `<RUN_ID>`.

5.4 Creating plots using the Linux distribution

To plot reference gene networks using the Linux distribution:

1. Extract the tar file `plot_networks_LINUX64.tar` in the local `bin` folder.
2. Edit the `config_file_network.txt` file found in the `networks` folder
3. From the command line (from the `bin` folder) type:

```
$ run_Plot_networks_LINUX64.sh <mcr_directory>
<MCRL_path>\networks\config_file_network.txt
```

5.5 Demo plot

For demonstration purposes, the default file `config_file_network.txt` can be run with the default demo files provided with the MCRL installation, only the path of the `*MCRL_table_nr.txt` file needs to be updated.

6. Support

To report bugs or request support please contact Arbel Tadmor (arbel.tadmor@tron-mainz.de).