CAMBerVis User's Manual

Release 1.0

Michal Wozniak

19 July 2011

Contents

1 About the software

1.1 Background

We observe a lot of inconsistencies in the genome structure annotations among bacterial strains (John Dunbar et al. BMC Genomics, 12(1):125, 2011). This inconsistency is a frustrating impedance to effective comparative genomic analysis of bacterial strains in promising applications such as gaining insights into bacterial drug resistance.

CAMBer (Wozniak M, Wong L, Tiuryn J. BMC Genomics 2011) is an approach to support comparative analysis of multiple bacterial strains. CAMBer unifies annotations of closely related species by homology transfer. It produces what we called multigene families. Each multigene family reveals genes that are in one-to-one correspondence in the bacterial strains, thereby permitting their annotations to be integrated. We present results of our method applied to three human pathogens: *Escherichia coli*, *Mycobacterium tuberculosis* and *Staphylococcus aureus*.

1.2 CAMBerVis

The CAMBerVis software is a tool to visualize inconsistencies in genome structure annotations among bacterial strains. It is based on the idea introduced by CAMBer (Wozniak et al. BMC Genomics 2011).

1.3 Availability

This software is an open source application (GPL 3 license). Sources are available at the code.google website:

`http://code.google.com/p/camber2/`

Please don't hesitate to contact us with any comments and suggestion or if you are interested in co-developing this software.

1.4 About the authors

This software was implemented by Michal Wozniak. Project idea and guidance came from Limsoon Wong and Jerzy Tiuryn.

Affiliation: Institute of Informatics, University of Warsaw

E-mail: `m.wozniak@mimuw.edu.pl`

2 Manual

2.1 Software requirements

- Java 1.6

The software was tested on Windows, Linux and Mac platforms.

2.2 Installation

Download the software zip package or the installer from our project's website:

`http://bioputer.mimuw.edu.pl/camber`

3 Input files

CAMBerVis requires two files to determine the list of strains and the structure of unified annotations. It also requires a set of FASTA files with genome sequences. See below for more format details.

3.1 Strains and plasmids

The file format describes a list of strains and plasmids which we want to visualize. Each line represents one strain or plasmid. For strains the line contains only the strain name, but for plasmids it contains both the plasmid name and the name of the strain, which contains the plasmids.

Example

```
ED98          -
JH1           -
JH9           -
ED98-pAVY     ED98
ED98-pAVX     ED98
ED98-pT181    ED98
JH-pSJH101    JH1
JH9-pSJH901   JH9
```

3.2 Unified annotations

Since CAMBer (Wozniak et al. BMC Genomics 2011) introduces new concepts to represent unified annotations among bacterial strains, we require the specific file format described below. The files used in the examples on *M. tuberculosis* and *S. aureus* have been generated by CAMBer, however the file format is generic and not dependent on CAMBer.

Each line of the file represents one multigene. The first column its the connected component (gene family) identifier. Second column its the unique multigene identifier created by concatenation on stop codon position, strand and strain or plasmid id. Third column its the annotated multigene id, or 'x' in the case when the multigene was predicted. The following columns contain all annotated and predicted gene lengths for the multigene, where the originally annotated one is marked by a star.

Example

```
6   473984.-.JH9   SaurJH9_0442    153   *162
6   399845.-.ED98  SAAV_0362      *153   162
6   474054.-.JH1   SaurJH1_0454    153   *162
24  208618.+.JH9   x               102
24  208688.+.JH1   x            102
24  177436.+.ED98  SAAV_0157   *102
```

3.3 Genome sequences

Genome sequences for both main genomes and plasmid genomes should be in the FASTA format.

3.4 Examples

Two examples of data on *M. tuberculosis* and *S. aureus* are integrated with the software. There are located in the `examples/` folder.

4 CAMBerVis features

Below we list the most important features of CAMBerVis. The features list with examples by figures is integrated with the CAMBerVis help system.

4.1 Multiple genome visualization

CAMBerVis provides multiple visualization all bacterial genomes supporting all standard genome navigation options like zoom in/out, move left/right. Another useful feature is flip of the genome strand. Moreover, CAMBerVis supports smooth changing of visualization between main chromosomes and plasmids, which are specific to bacteria.

The tool also provides special features to visualize differences in the structure of annotations. On the pictures below multigenes are draw as horizontal bars, since TISs are vertical ticks (black for predicted and red for originally annotated ones).

4.2 Multigene color schemas

There are two schema colors available. First (default) uses the colors accordingly to types of the connected components:

- RED — ANCHOR connected components (gene families) with elements in all strains
- YELLOW — ANCHOR connected components (gene families) with elements not in all strains
- GRAY — ANCHOR connected components (gene families) with exactly one element (in only one strain)
- GREEN — NON-ANCHOR connected components (gene families) with elements in all strains
- BLUE — NON-ANCHOR connected components (gene families) with elements not in all strains

In the second color schema multigene colors are assigned randomly to gene families.

4.3 Statistics

In order to support analysis CAMBerVis provides variety of statistics that are managed by sortable tables. First of the tables manages the list of connected components (gene families) by many different features like: number of multigenes, multigene length or number of strains with annotated elements.

The second table manages statistics of the visualized genomes.

4.4 On-the-fly computations

CAMBerVis supports also on-the-fly computations of multiple alignments and queries to the NCBI databases like Swissprot or Nr.

Results of the computations are kept in the *TasksTable* window. Double click on a row opens the corresponding task report. The original file storing results is also accessible.

Both ClustalW and queries to NCBI databases can be configures, a user can choose DNA or amino acid queries, TIS or the length of promotors.

4.5 Configuration

CAMBerVis is expected to be a fully user configurable tool. So far it provides configuration of paths to external tools like CLUSTALW, necessary for computing of multiple alignments.