BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 23 2011, pages 3313–3314
doi:10.1093/bioinformatics/btr561

Genome analysis

Advance Access publication October 8, 2011

CAMBerVis: visualization software to support comparative analysis of multiple bacterial strains

Michał Woźniak[1,2,*], Limsoon Wong[2] and Jerzy Tiuryn[1]

[1]Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Warsaw, Poland and [2]School of Computing, National University of Singapore, Singapore, Singapore

Associate Editor: Martin Bishop

ABSTRACT

Motivation: A number of inconsistencies in genome annotations are documented among bacterial strains. Visualization of the differences may help biologists to make correct decisions in spurious cases.

Results: We have developed a visualization tool, CAMBerVis, to support comparative analysis of multiple bacterial strains. The software manages simultaneous visualization of multiple bacterial genomes, enabling visual analysis focused on genome structure annotations.

Availability: The CAMBerVis software is freely available at the project website: http://bioputer.mimuw.edu.pl/camber. Input datasets for *Mycobacterium tuberculosis* and *Staphylocacus aureus* are integrated with the software as examples.

Contact: m.wozniak@mimuw.edu.pl

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on August 8, 2011; revised on September 19, 2011; accepted on October 5, 2011

1 INTRODUCTION

The large amount of genomic data that is being generated enables exciting new ways for comparative analysis of genomes. In particular, revealing drug resistance mechanisms in pathogenic bacteria is currently one of the important applications of comparative genomics. Systematic comparison of drug resistant and drug susceptible strains can bring us new clues on how drug resistance emerges. However, since the presence of a single gene or mutation can lead to drug resistance, we may require very precise genome structure annotations to carry out reasonable drug resistance studies.

A recent comparative study (Wozniak *et al*., 2011) considered three pathogenic bacteria: *Mycobacterium tuberculosis*, *Staphylocacus aureus* and *Escherichia coli*. The results showed many inconsistencies in genome structure annotations. Moreover, lots of the inconsistencies are not due to real sequence differences, but are apparently caused by the use of different gene annotation methodologies by different laboratories.

The problem of inconsistencies in genome annotations of translation initiation start sites (TISs) was also reported by another recent study (Dunbar *et al*., 2011). The authors compared gene starts annotations of orthologous gene families among five bacterial strains of *Burkholderia*. The analysis showed that 53% of the ortholog gene families have inconsistent gene starts annotations in GenBank. Furthermore, inconsistencies for only 17% of all the ortholog gene families could be explained by sequence divergences.

Therefore, manual curation by biologists is necessary in cases where standard genome annotation tools produce inconsistencies. In order to better support this type of analysis, we have implemented CAMBerVis—a software that allows for visual comparison of the genome structure annotations of multiple bacterial strains.

*To whom correspondence should be addressed.

2 METHODS

2.1 Basic concepts

CAMBerVis is a software designed to visualize genome structure annotations (both original and predicted) according to the concepts introduced in our previous work (Wozniak *et al*., 2011) (called the CAMBer approach). Here we briefly introduce the basic required notations.

CAMBer is an approach to support comparative analysis of multiple bacterial strains. As input it uses genome sequences and *original annotations* of the bacterial strains considered. Then, it iteratively transfers gene annotations, until the transitive closure (of the proposed homology relation) is computed. Thus, during the procedure, new open reading frames (ORFs) become annotated. We call the final resulting structure annotations the *predicted annotations*. Furthermore, in order to manage the problem of inconsistencies in TIS annotations, CAMBer introduces the concept of a *multigene*, which represents a set of gene annotations with the same stop codon. Then, multigenes (nodes) are linked by homology relationships (edges) between their elements. We call the structure a *consolidation graph*. Multigenes in the same *connected component* of the consolidation graph are proposed to be gene families.

2.2 CAMBerVis

The input to CAMBerVis consists of genome FASTA files and a file with predicted genome structure annotation. The file format is generic and not dependent on CAMBer. A user may find more details of the format in the software documentation and learn from the integrated examples on *M.tuberculosis* and *S.aureus*.

Here we describe the main features of CAMBerVis based on a typical usage. In the first step, we identify a gene family of interest with some annotation inconsistencies. CAMBerVis manages statistics for every gene family in a table in the *ComponentsStats* window. Using this table, we can easily find gene families with missing gene annotations or inconsistencies among annotated TISs.

Second, the visualization is automatically focused on the selected gene family showing simultaneously its multigenes (with both annotated and predicted TISs) in all strains. We may also see their neighborhood in different scales using intuitive genome navigation.

Third, we use on-the-fly comparative analysis supported by CAMBerVis. For example in the case of inconsistently annotated TISs, we may compare promotor regions by multiple alignments using the integrated CLUSTALW.

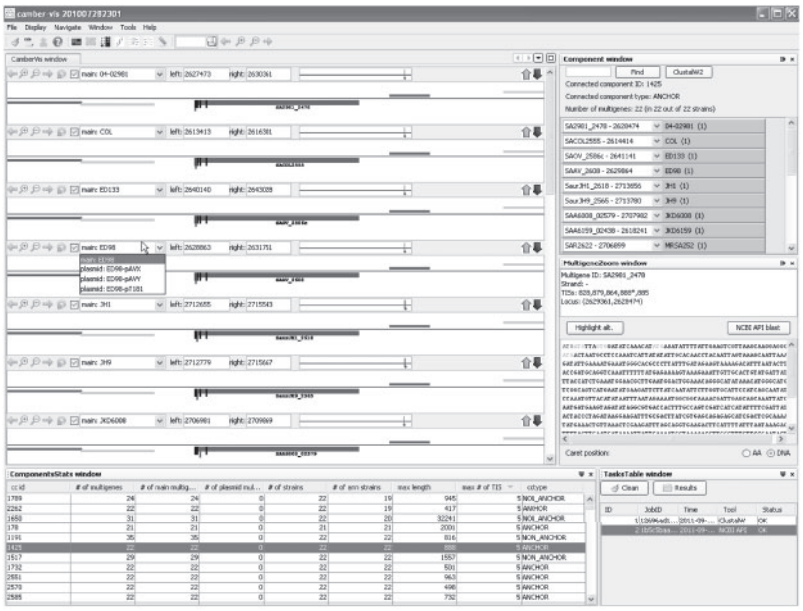

Fig. 1. The main view of the CAMBerVis interface with loaded example data for 22 strains of *S. aureus*. The view is focused on a highly conserved connected component (gene family) with five different TISs in each multigene, selected from the list in the *ComponentsStats* window. Multigenes are visualized as horizontal rectangles, with TISs presented as vertical ticks (originally annotated TISs are red and long). The window *TasksTable* keeps track of results obtained on-the-fly by ClustalW or NCBI BLAST API.

CAMBerVis also enables external queries via NCBI BLAST API, which can be applied to check which TIS is the most often annotated in external databases like, for example, NCBI non-redundant (NR) database.

Figure 1 presents a screen shot of the running application. The visualization is focused on a gene family identified by *ComponentsStats* table sorted by the number of TISs. There are five different TISs annotated in GenBank among the 22 fully sequenced strains of *S.aureus*, annotated with the following frequencies: 2,1,7,4,8 (ordered from the TIS giving rise to the shortest gene to the TIS giving rise to the longest gene). An analysis of the multiple alignment computed by CLUSTALW showed that the gene family is highly conserved among strains and only four of the strains have SNP in the 100 bp long promotor region. Queries to the NCBI NR database showed that the TIS that yields the longest gene is the most often annotated.

CAMBerVis is a stand-alone application written in Java, which makes it a cross-platform application, tested on Windows, Mac and Linux. Notably, it is implemented based on the Netbeans IDE platform, which makes the application flexible and easy to extend. Another benefit is that the windows manager allows user to customize the window localization.

3 CONCLUSION

The amount of data that is being generated stimulates active development of visualization techniques and softwares, which are invaluable to scientist for manual curation of results (Nielsen *et al*., 2010). The most notable are VISTA (Frazer *et al*., 2004), Microbial Genomes (Dehal *et al*., 2010), Integrated Microbial Genomes (Markowitz *et al*., 2008), SEED Viewer (Overbeek *et al*., 2005), K-BROWSER (Chakrabarti *et al*., 2004) and Artemis (Rutherford *et al*., 2000). However, only VISTA and K-BROWSER allow multiple genome visualization.

Here we present CAMBerVis, a new genome browser which allows simultaneous visualization and comparative analysis of multiple bacterial strains. Moreover, it is the first visualization software distinguishing original and predicted genome structure annotations. Another advantage of CAMBerVis over existing softwares is its intuitive management of plasmids, which are common in bacteria.

CAMBerVis is an open-source application freely available at the project website and integrated with two example datasets for *M.tuberculosis* and *S.aureus*.

Funding: Polish Ministry of Science and Higher Education grant no. (N N301 065236); Singapore Ministry of Education Tier-2 grant (MOE2009-T2-2-004), in part.

Conflict of Interest: none declared.

REFERENCES

Chakrabarti,K. and Pachter,L. (2004) Visualization of multiple genome annotations and alignments with the K-BROWSER. *Genome Res.*, **114**, 716–720.

Dehal,P.S. *et al.* (2010) MicrobesOnline: an integrated portal for comparative and functional genomics. *Nucleic Acids Res.*, **38**, D396–D400.

Dunbar,J. *et al.* (2011) Consistency of gene starts among Burkholderia genomes. *BMC Genomics*, **12**, 125.

Frazer,K.A. *et al.* (2004) VISTA: computational tools for comparative genomics. *Nucleic Acids Res.*, **32** (Suppl. 2), W273–W279.

Markowitz,V.M. *et al.* (2008) IMG/M: a data management and analysis system for metagenomes. *Nucleic Acids Res.*, **36** (Suppl. 1), D534–D538.

Nielsen,C.B. *et al.* (2010) Visualizing genomes: techniques and challenges. *Nat. Methods*, **7** (Suppl. 3), S5–S15.

Overbeek,R. *et al.* (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res.*, **33**, 5691–5702.

Rutherford,K. *et al.* (2000) Artemis: sequence visualization and annotation. *Bioinformatics*, **16**, 944–945.

Wozniak,M. *et al.* (2011) CAMBer: an approach to support comparative analysis of multiple bacterial strains. *BMC Genomics*, **12** (Suppl. 2), S6.