BIOINFORMATICS APPLICATIONS NOTE Vol. 22 no. 2 2006, pages 242–244
doi:10.1093/bioinformatics/bti759

Genome analysis

COMPAM :visualization of combining pairwise alignments for multiple genomes

DoHoon Lee[1,2,*], Jeong-Hyeon Choi[2], Mehmet M. Dalkilic[2,3] and Sun Kim[2,3]

[1]School of Computer Engineering, Miryang National University, Miryang, Kyungnam 627-702, Korea, [2]School of Informatics and [3]Center for Genomics and Bioinformatics, Indiana University—Bloomington, IN 47404, USA

Received on June 3, 2005; revised on October 13, 2005; accepted on November 2, 2005
Advance Access publication November 3, 2005
Associate Editor: Christos Ouzounis

ABSTRACT

Summary: COMPAM is a tool for visualizing relationships among multiple whole genomes by combining all pairwise genome alignments. It displays shared conserved regions (blocks) and where these blocks occur (edges) as block relation graphs which can be explored interactively. An unannotated genome, e.g. can then be explored using information from well-annotated genomes, COG-based genome annotation and genes. COMPAM can run either as a stand-alone application or through an applet that is provided as service to PLATCOM, a toolset for whole genome comparative analysis, where a wide variety of genomes can be easily selected. Features provided by COMPAM include the ability to export genome relationship information into file formats that can be used by other existing tools.

Availability: http://bio.informatics.indiana.edu/projects/compam/

Contact: dohhlee@indiana.edu; sunkim2@indiana.edu

INTRODUCTION

As more whole genomes become available, there is an opportunity to better and more deeply understand relationships among organisms. Although a number of multiple genome alignment algorithms have been developed, they can miss a large number of common protein coding genes (Choi *et al*., 2005a). One way these kinds of problems can be addressed is to use pairwise genome alignments directly, but visualizing all pairwise alignments of n genomes, which results in $(n/2) = [n \times (n-1)]/2$ pairs, in a single display window is very challenging. We have developed a visualization tool to combine pairwise alignments of multiple genomes (COMPAM) to address these challenges.

There are existing tools that either enable visualizing multiple genome sequence alignments (Chakrabarti and Pachter, 2004; Shah *et al*., 2004; Kozik *et al*., 2002; Darling *et al*., 2004; Choudhuri *et al*., 2004; Carver *et al*., 2005) or enable comparative analysis of whole genomes (Rasko *et al*., 2005; Choi *et al*., 2005b; Nix and Eisen, 2005).

FEATURES

For input COMPAM requires the genome sequence in FastaA format (`.fna` files), gene/COG family information (.ptt files) and all pairwise comparisons of the genomes by Blastz. COMPAM can run as either a standalone application or as a service to PLATCOM (Choi *et al*., 2005b). Users can select from a wide assortment of genomes. In this section we briefly present three important features of COMPAM: (1) it summarizes significant elements of the alignments; (2) summarizes $n-1$ genome alignments with respect to a particular genome and (3) displays simultaneously the collection of n whole aligned genomes. We will sketch how COMPAM works.

In step 1, a genome is compared with all other genomes using Blastz. The $n-1$ results are then combined to compute overlapped regions. In step 2, summaries are generated for all n genomes from step 1 and the resulting overlapped regions found. In step 3, regions are selected to make block nodes. A block node is defined as a connected component in the interval graph. Given a genome sequence in G_i, each interval in G_i in a pair of intervals from two genomes, say G_i and G_j $(i \neq j)$, becomes a node in the interval graph and an edge is created when two overlapping intervals in G_i share subsequences of a user-defined length or longer. Block nodes include information about overlaps, the genomes, COGs and genes. In step 4 a block relation graph (block graph) is constructed where nodes are block nodes in the interval graph and edges are matched blocks from pairwise alignments. Users can interact and explore the block graph by selecting blocks. Users can adjust block depth, which is defined as the depth of children visited in the block graph while performing breadth-first search.

COMPAM provides a number of features, some of which we highlight here:

- support for comparative annotation of new genomes in comparison with well-annotated genomes;
- interactive exploration of block graphs including paths up to length 3 from any particular block (depth control);
- divided dynamic zoom (from low to high resolution) of genomic regions;
- ability to export DNA sequences of a block from any block graph path and perform a multiple sequence alignment of them;
- display of various annotations from COGs and genes including frequency of overlapping regions and average scores;
- ability to select/hide matches with three different filtering methods;
- display of various statistics of the selected genomes.

*To whom correspondence should be addressed.

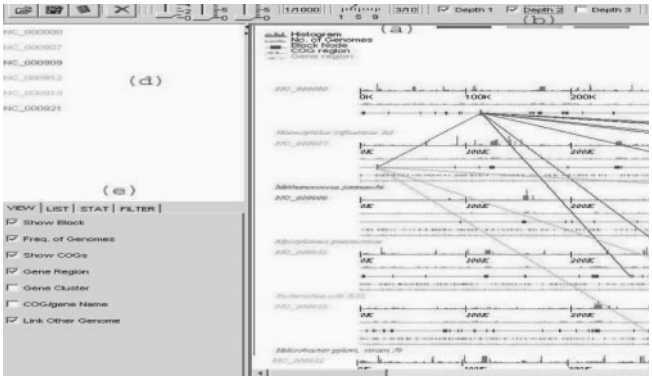

Fig. 1. A screen shot from COMPAM analyzing an unannotated genome enclosed by the red rectangle. COMPAM extracts all conserved regions of the new genome using other annotated genomes and displays paths from the reference block. We have indicated interesting elements with letters. (**a**) Shows the divided zoom bar. (**b**) Allows the user to select the depth of a block graph, blue for depth one, dark yellow for depth two and gray for depth three. (**c**) Is a button to export genome relationship information with respect to the current reference genome into a text file on a local machine. (**d**) Shows the list of genomes being compared. (**e**) Gives three different viewing options: the VIEW tab lists different elements to view: block, frequency of overlapping, the number of genomes, COG, gene and the relation graph. The LIST tab allows users to display information about the block graph, and the STAT tab generates simple statistics among the compared genomes. The FILTER tab allows users to control the connectivity of relation graph.

PROCESSING AND IMPLEMENTATION

Our running example will use five whole bacterial genomes. COMPAM provides a quick overview of multiple pairwise genome alignments, which can be useful to annotate a new genome in relation to other genomes. Assume we are given a new genome. COMPAM executes steps one and two resulting in a block graph. Then, as depicted in Figure 1, COMPAM shows all overlapped regions with the previously annotated genomes.

Figure 2 provides a more detailed example of exploring an unknown genomic region. Consider when we observe a particular region in this new genome that is not annotated with any COG or gene information, but has a high overlap frequency with other annotated genomes. This first observation is shown in (a). The region is denoted as a red rectangle. We next explore the block graph of depth one shown in (b). To check further relationships we examine the block graph to depths of two and three as shown in (c). Two of the paths have been selected and shown in (d). Observe that a region in *Haemophilus influenza* is related to other genomes *Staphylococcus aureus* and *Methanococcus jannaschii* through the block graph. The multiple alignment of selected regions in the path using Clustal-W as shown in (e).

DATA PREPARATION

COMPAM can be coupled with any pairwise local genome alignment programs as long as the comparison result is in COMPAM format. We currently provide a script to convert the alignment result with Blastz to a COMPAM input data. Users can download precomputed genome alignment data from our web server. Otherwise, users can compute COMPAM input data using the genomic DNA sequences with the scripts provided on the web server.

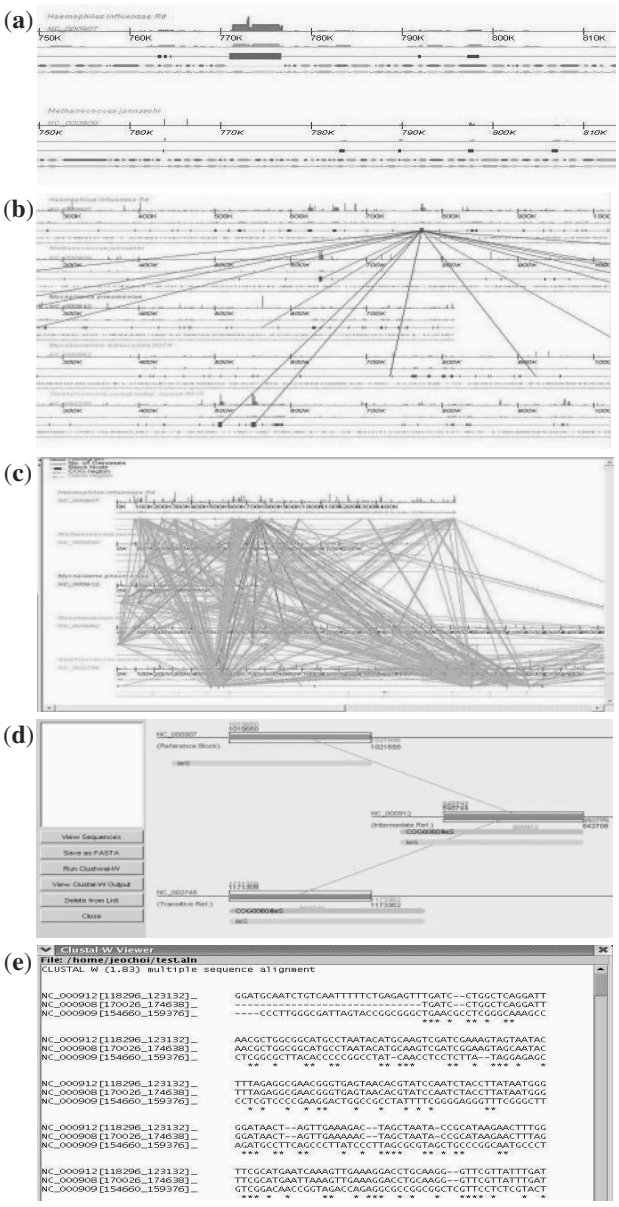

Fig. 2. These series of screen shots from COMPAM show a typical exploration of an unannotated region from a new genome. The sequence of steps is shown with (a) through (e). (**a**) Highlights the region of interest shown enclosed by a red rectangle. (**b**) Displays the block graph of depth one. Exploring depths two and three is shown in (**c**). Zoom (1/4000 ratio) was used additionally to view all the block relations. (**d**) Displays three blocks in detail and their respective COG and gene information. A region in *Haemophilus influenza* has overlap with block nodes matched to two other genomes *Staphylococcus aureus* and *Methanococcus jannaschii*. (**e**) Shows the multiple alignment of selected regions using Clustal-W.

PLATFORM SPECIFICATIONS

COMPAM consists of two modules: (1) data processing and extraction (2) visualization. The data processing and extraction module is responsible for calculating frequency of overlapped regions, creating the block graphs, simple statistics and so forth. This module is written in Perl 5.8.0. The visualization module is implemented using

JBuilder9. The minimal hardware requirements are P4 CPU with 512 MB of RAM.

SUMMARY AND FUTURE WORK

COMPAM is an easy-to-use tool that allows quick and efficient exploration of genomic sequence. We are currently working on improving the efficiency of block discovery. We will provide a set of scripts to convert outputs from existing genome alignment programs to the COMPAM input data. Further, we are making a library of COMPAM visualizations and block graphs.

ACKNOWLEDGEMENTS

We thank Kwangmin Choi for his suggestions and comments. We also appreciate anonymous reviewers' comments which improved the previous version of the paper significantly. This work is partially supported by NSF CAREER Award DBI-0237901 to S.K., NSF IIS-0082401 to M.M.D. and overseas research fund of Miryang National University to D.H.L.

Conflict of Interest: none declared.

REFERENCES

Carver,T.J. *et al*. (2005) ACT: the Artemis comparison tool. *Bioinformatics*, **21**, 3422–3423.

Chakrabarti,K. and Pachter,L. (2004) Visualization of multiple genome annotations and alignments With the K-BROWSER. *Genome Res*., **14**, 716–720.

Choi,J.-H., Choi,K., Cho,H.-G. and Kim,S. Multiple genome alignment by clustering pairwise matches. In Lagergren,J. (ed.), *Proceedings of the Comparative Genomics Satellite Workshop,* number 3388 in Lecture Note in Bioinformatics (LNB1), Bertinoro, Italy. Springer, Berlin, pp. 30–41.

Choudhuri,J.V. *et al*. (2004) GenAlyzer: interactive visualization of sequence similarities between entire genomes. *Bioinformatics*, **20**, 1964–1965.

Darling,A.C. *et al*. (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res*., **14**, 1394–1403.

Kozik,A. *et al*. (2002) GenomePixelizer—a visualization program for comparative genomics within and between species. *Bioinformatics*, **18**, 335–336.

Nix,D.A. and Eisen,M.B. (2005) GATA: a graphic alignment tool for comparative sequence analysis. *BMC Bioinformatics*, **6**, (http://www.biomedcentral.com/1471-2105/6/9).

Rasko,D.A. (2005) Visualization of comparative genomic analyses by BLAST score ratio. *BMC Bioinformatics*, **6**, 1471–2105.

Shah,N. *et al*. (2004) Phyto-VISTA: interactive visualization of multiple DNA sequence alignments. *Bioinformatics*, **20**, 636–643.