**GeneclusterViz: a tool for conserved gene cluster visualization, exploration and analysis**

Vikas R. Pejaver¹, Jaehyun An², SungMin Rhee², Ankita Bhan³, Jeong-Hyeon Choi⁴, Boshu Liu⁵, Heewook Lee³, Pamela J. Brown³, David Kysela³, Yves V. Brun³,* and Sun Kim²,∗

¹School of Informatics and Computing, Indiana University Bloomington, IN 47404, USA, ²School of Computer Science and Engineering, Bioinformatics Institute, Seoul National University, Seoul, Korea, ³Abbott Laboratories, Chicago, IL, ⁴Cancer Center and Biostatistics, Georgia Health Sciences University, Augusta, GA 30912, ⁵Department of Biology, Indiana University Bloomington, IN 47405-3700, USA

Abstract:

Motivation: Gene clusters are arrangements of functionally related genes on a chromosome. In bacteria, it is expected that evolutionary pressures would conserve these arrangements due to the functional advantages they provide. Visualization of conserved gene clusters across multiple genomes provides key insights into their evolutionary histories. Therefore, a software tool that enables visualization and functional analyses of gene clusters would be a great asset to the biological research community.

Results: We have developed GeneclusterViz, a Java-based tool that allows for the visualization, exploration and downstream analyses of conserved gene clusters across multiple genomes. GeneclusterViz combines an easy-to-use exploration interface for gene clusters with a host of other analysis features such as multiple sequence alignments, phylogenetic analyses and integration with the KEGG pathway database.


Contact: sunkim.bioinfo@snu.ac.kr; ybrun@indiana.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

Received on November 2, 2011; revised on February 27, 2012; accepted on April 3, 2012

1 INTRODUCTION

Recent advances in sequencing technology and ortholog detection methods have given rise to several methods for the detection of gene clusters conserved across two or more genomes. Calabrese et al. (2003), Tsibulsky et al. (2004), Hahs et al. (2004), Kim et al. (2007), Fang and Sun (2008), Zheng et al. (2008). However, gene cluster information obtained from these programs has been very difficult to analyze for bench scientists due to the lack of a more intuitive analysis tool. The visualization of gene clusters in multiple genomes, with respect to their spatial and functional features is a fundamental problem in this area.

There are currently several visualization tools that have been developed for a variety of applications in comparative genomics (reviewed in Nielsen et al. (2010)). A subset of these have been designed for the visualization of gene clusters. However, these tools are limited in different ways. In some cases, only paired-genome visualization is supported. Even in multi-genome visualization, scalability is a challenge and, thus, limits the utility of web-based tools. Moreover, web-based tools rely on server-hosted data and cannot utilize data provided by users. This is true even in the case of Absynte, a tool that was recently developed specifically for the task of visualizing bacterial and archaeal clusters (Despalins et al. 2011). Standalone applications overcome these drawbacks but involve cumbersome installation procedures or database setups that demand programming skills beyond that of most end-users. Finally, existing gene cluster visualization tools are limited in their support for further downstream analyses. We have implemented GeneclusterViz, a robust and dynamic standalone tool that provides a global and local view of gene clusters. Moreover, with a host of flexible sequence and function analysis features, we believe GeneclusterViz can be a very useful tool in comparative genomics.

2 IMPLEMENTATION

GeneclusterViz has been implemented in Java (JDK 1.6). For the construction and depiction of phylogenetic trees, the Phylogenetic Analysis library was used (Drummond and Strimmer 2001). To establish a server connection and to communicate with the back-end CGI program, the Jakarta Commons HTTPClient Java library from Apache Commons has been wrapped. The server-side CGI programs and wrappers that run CLUSTAL W (Thompson et al. 1994), HMMER v2.3.2 (Eddy, 1998) and KEGG pathway searches have been written in Perl.

3 FEATURES

The features of GeneclusterViz can be summarized into four broad categories—input, visualization, exploration and analyses features.

Input: GeneclusterViz accepts output files from the EGGs (Kim et al. 2004) and PhylotreeGGS (part of ISGA; Hemmerich et al. 2010) algorithms. These files are mainly tab-delimited plain-text formats that contain NC numbers and NCBI GI numbers as identifiers for genomes and individual gene products, respectively.
Fig. 1. Screenshots of the main and detailed views of GeneclusterViz for a dataset of eight alphaproteobacteria genomes. In the main view, clusters 13 (NADH dehydrogenase complex cluster) and 21 (translation-related cluster) are close together in the Caulobacterales genomes but are separated in the Rhizobiales genomes. They are even farther apart in the outgroup (*Escherichia coli*). This shows how separations between clusters can correlate with phylogeny—(A) cluster list with position information, (B) viewing area, (C) viewing and exploration options and (D) connected genes within the clusters. The detailed view shows the NADH dehydrogenase cluster and some analysis features—(E) cluster table, (F) viewing area, (G) viewing and export options, (H) phylogenetic tree for selected gene, (I) browser connection to KEGG pathway with the selected gene highlighted, (J) strand ‘flip’ button and (K) gene selection (red).
GeneclusterViz also accepts gene family files in an in-house file format (GFAM) that was developed to represent genes from common COG families.

Visualization: Users can easily zoom-in/out of cluster visualizations with respect to both the X and Y-axes. The ‘Connect Clusters’ feature displays all the connections between clusters across multiple genomes. Individual clusters can be accessed by clicking on a cluster in the table-pane on the left of the GeneclusterViz window. This highlights the particular cluster on the genome. The ‘Connect Genes’ feature displays how individual genes are connected in the highlighted cluster. The order in which the genomes are displayed can also be changed through a draggable list. Given a Newick tree file, the ‘Phylogeny View’ allows for the visualization of genomes in terms of their clades. These orderings are particularly useful if a user is interested in focusing on a subset of the input genomes.

Analyses: At the main GeneclusterViz window, upon double-clicking on a cluster entry in the table-pane, a new window displays a zoomed-in detailed view of that particular cluster. The genes are color-coded as per COG broad functional categories. One can view annotation information for a gene by simply mousing over that particular gene. Clicking on a particular gene in the cluster highlights the connection across the multiple genomes. If this highlighted connection is double-clicked, a new window with the sequences of the gene products is opened. These sequences can then be run through CLUSTAL W to generate a multiple sequence alignment and a phylogenetic tree can be built for their gene products. The corresponding pathway from KEGG can also be obtained.

Exploration: Users can search for their genes of interest by providing their names or locus tags. Geneclusterviz can identify which clusters the specified genes are in and generates a table with the input genes against the corresponding cluster number. The clusters can then be accessed directly from the table. Another exploration feature is the capacity to add a new genome, given an existing multi-genome cluster prediction. GeneclusterViz allows for the input of a new genome and incorporates a feature which checks whether a given cluster is present in the new genome or not. It achieves this by establishing a connection to a back-end server that performs a profile-model-based HMM search (using HMMER) to identify conserved cluster members in the new genome. In this case, relaxed criteria for cluster identification have been adopted (HMMER E-value <1e-10 which is the default and no proximity constraint on significant hits). Once the cluster has been identified in the newly input genome, it is displayed in the detailed view, along with the cluster in the other genomes for manual investigation.

4 DISCUSSION

GeneclusterViz serves not only as a user-friendly tool for the visualization of gene clusters across multiple genomes but also as a workspace for extensive research on these clusters. It tackles the non-trivial task of providing access and information about multiple conserved clusters across multiple genomes. There are several applications that GeneclusterViz can be used for including phylogenetic studies, pathway studies and gene function assignment.

Thus, it is expected to help bench scientists in understanding the functional relationship between genes in conserved clusters across the evolutionary spectrum of genomes.

Funding: NSF MCB-0731950 (to Y.Y.B. and S.K.), Seoul National University BioMAX Institute (to S.K.), and US National Institutes of Health GM51986 (to Y.V.B.)

Conflict of Interest: none declared.

REFERENCES