Genome analysis

Advance Access publication April 11, 2012

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

Vikas R. Pejaver[1], Jaehyun An[2], SungMin Rhee[2] Ankita Bhan[3], Jeong-Hyeon Choi[4], Boshu Liu[5], Heewook Lee[1], Pamela J. Brown[6], David Kysela[6], Yves V. Brun[6,*] and Sun Kim[2,*]

[1]School of Informatics and Computing, Indiana University Bloomington, IN 47404, USA, [2]School of Computer Science and Engineering, Bioinformatics Institute, Seoul National University, Seoul, Korea, [3]Abbott Laboratories, Chicago, IL, [4]Cancer Center and Biostatistics, Georgia Health Sciences University, Augusta, GA 30912, [5]Center for Genomics and Bioinformatics, Indiana University Bloomington, IN 47404, [6]Department of Biology, Indiana University Bloomington, IN 47405-3700, USA

Associate Editor: John Quackenbush

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.

Availability: http://biohealth.snu.ac.kr/GeneclusterViz/; http://microbial.informatics.indiana.edu/GeneclusterViz/

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; Fujibuchi *et al.*, 2000; Haas *et al.*, 2004; Kim *et al.*, 2007; Yang and Sze, 2008; Zheng *et al.*, 2005). 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

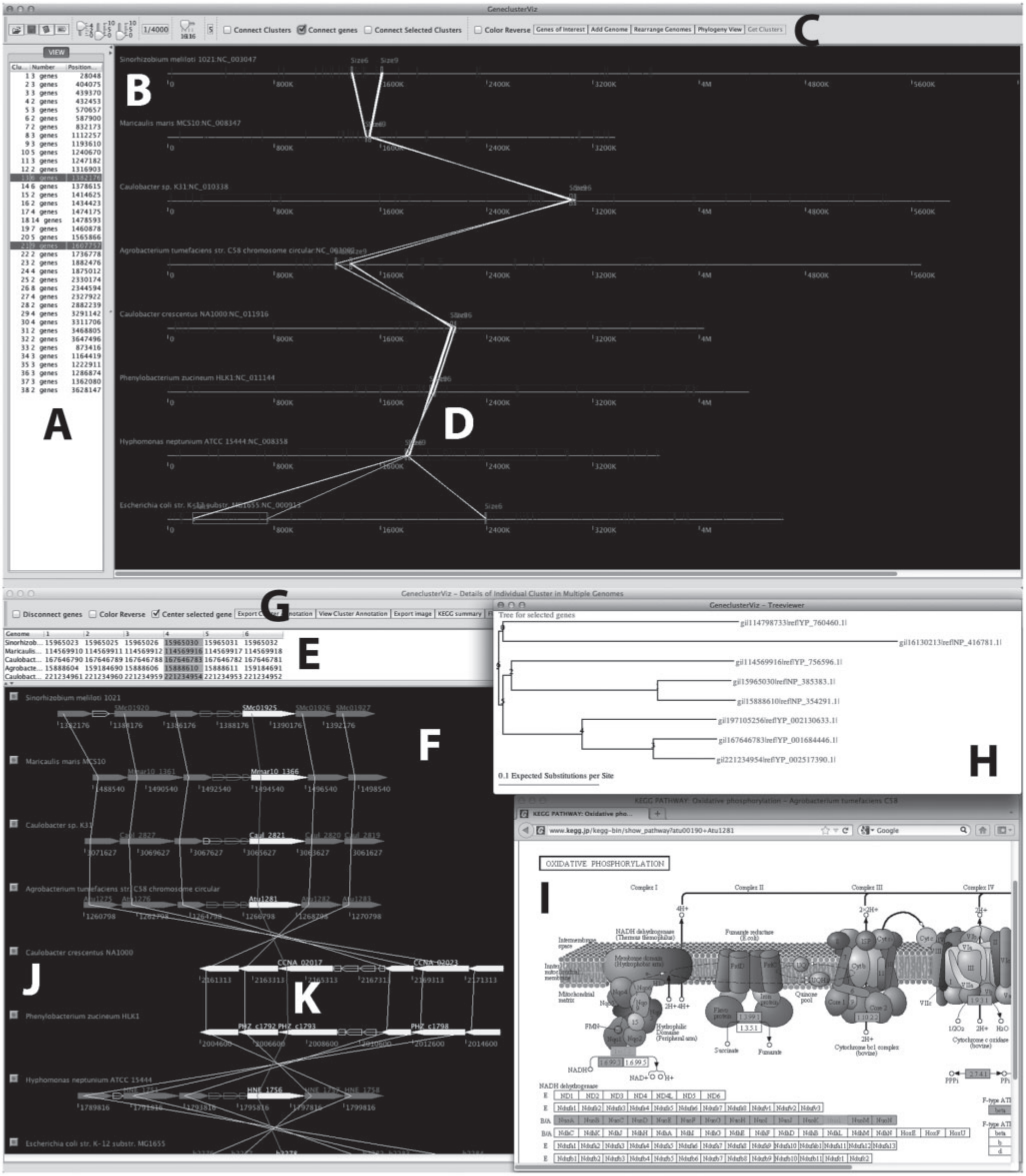
GeneclusterViz has been implemented in Java (JDK 1.6). For the construction and depiction of phylogenetic trees, the Phylogenetic Analysis library was used (Drummond and Strimmer, 2001). To establish a server connection and to communicate with the back-end CGI program, the Jakarta Commons HTTPClient Java library from Apache Commons has been used. 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.* (2007)] and PhyloEGGS [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.

*To whom correspondence should be addressed.


GeneclusterViz
1/4000
□ Connect Clusters ☑ Connect genes □ Connect Selected Clusters □ Color Reverse
Genes of Interest | Add Genome | Rearrange Genomes | Phylogeny View | Get Clusters
C
VIEW
A
B
D
Sinorhizobium meliloti 1021:NC_003047
Maricaulis maris MCS10:NC_008347
Caulobacter sp. K31:NC_010338
Agrobacterium tumefaciens str. C58 chromosome circular:NC_003062
Caulobacter crescentus NA1000:NC_011916
Phenylobacterium zucineum HLK1:NC_011144
Hyphomonas neptunium ATCC 15444:NC_008358
Escherichia coli str. K-12 substr. MG1655:NC_000913
0
800K
1600K
2400K
3200K
4M
4800K
5600K
Size6
Size9
GeneclusterViz – Details of Individual Cluster in Multiple Genomes
□ Disconnect genes □ Color Reverse ☑ Center selected gene
Export Cluster Annotation | View Cluster Annotation | Export image | KEGG summary
G
E
F
J
K
H
I
GeneclusterViz – Treeviewer
Tree for selected genes
gi|114798733|ref|YP_760460.1|
gi|16130213|ref|NP_416781.1|
gi|114569916|ref|YP_756596.1|
gi|15965030|ref|NP_385383.1|
gi|15888610|ref|NP_354291.1|
gi|197105256|ref|YP_002130633.1|
gi|167646783|ref|YP_001684446.1|
gi|221234954|ref|YP_002517390.1|
0.1 Expected Substitutions per Site
KEGG PATHWAY: Oxidative phosphorylation – Agrobacterium tumefaciens C58
www.kegg.jp/kegg-bin/show_pathway?atu00190+Atu1281
OXIDATIVE PHOSPHORYLATION
Complex I
Complex II
Complex III
Complex IV
NADH dehydrogenase
Sinorhizobium meliloti 1021
Maricaulis maris MCS10
Caulobacter sp. K31
Agrobacterium tumefaciens str. C58 chromosome circular
Caulobacter crescentus NA1000
Phenylobacterium zucineum HLK1
Hyphomonas neptunium ATCC 15444
Escherichia coli str. K-12 substr. MG1655
SMc01920
SMc01925
SMc01926
SMc01927
Mmar10_1361
Mmar10_1366
Caul_2827
Caul_2821
Caul_2820
Caul_2819
Atu1275
Atu1276
Atu1281
Atu1282
Atu1283
CCNA_02017
CCNA_02023
PHZ_c1792
PHZ_c1793
PHZ_c1798
HNE_1751
HNE_1756
HNE_1757
HNE_1758

Clu...	Number	Position...
1	3 genes	28048
2	3 genes	404075
3	3 genes	439370
4	2 genes	432453
5	3 genes	570657
6	2 genes	587900
7	2 genes	832173
8	3 genes	1112257
9	3 genes	1193610
10	5 genes	1240670
11	3 genes	1247182
12	2 genes	1316903
13	6 genes	1382176
14	6 genes	1378615
15	2 genes	1414625
16	2 genes	1434423
17	4 genes	1474175
18	14 genes	1478593
19	7 genes	1460878
20	5 genes	1565866
21	9 genes	1607757
22	2 genes	1736778
23	2 genes	1882476
24	4 genes	1875012
25	2 genes	2330174
26	8 genes	2344594
27	4 genes	2327922
28	2 genes	2882239
29	4 genes	3291142
30	4 genes	3311706
31	2 genes	3468805
32	2 genes	3647496
33	2 genes	873416
34	3 genes	1164419
35	3 genes	1222911
36	3 genes	1286874
37	3 genes	1362080
38	2 genes	3628147

Genome	1	2	3	4	5	6
Sinorhizob...	15965023	15965025	15965026	15965030	15965031	15965032
Maricaulis...	114569910	114569911	114569912	114569916	114569917	114569918
Caulobact...	167646790	167646789	167646788	167646783	167646782	167646781
Agrobacte...	15888604	159184690	15888606	15888610	15888611	159184691
Caulobact...	221234961	221234960	221234959	221234954	221234953	221234952

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 <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.V.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

Calabrese,P.P. *et al.* (2003) Fast identification and statistical evaluation of segmental homologies in comparative maps. *Bioinformatics*, **19**(Suppl 1), i74–i80.

Despalins,A. *et al.* (2011) Absynte: a web tool to analyze the evolution of orthologous archaeal and bacterial gene clusters. *Bioinformatics*, **27**, 2905–2906.

Drummond,A. and Strimmer,K. (2001) PAL: an object-oriented programming library for molecular evolution and phylogenetics. *Bioinformatics*, **17**, 662–663.

Eddy,S.R (1998) Profile hidden markov models. *Bioinformatics*, **14**, 755–763.

Fujibuchi,W. *et al.* (2000) Automatic detection of conserved gene clusters in multiple genomes by graph comparison and p-quasi grouping. *Nucleic Acids Res.*, **28**, 4029–4036.

Haas,B.J. *et al.* (2004) DAGchainer: A tool for mining segmental genome duplications and synteny. *Bioinformatics*, **20**, 3643–3646.

Hemmerich,C. *et al.* (2010) An Ergatis-based prokaryotic genome annotation web server. *Bioinformatics*, **26**, 1122–1124.

Kim,S. *et al.* (2007) EGGS: Extraction of gene clusters by iteratively using genome context based sequence matching techniques. In *Proceedings of the 2007 IEEE International Conference on Bioinformatics and Biomedicine*. IEEE Computer Society, San Jose, CA, USA, pp. 23–28.

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

Thompson,J.D. *et al.* (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.

Yang,Q. and Sze,S. (2008) Large-scale analysis of gene clustering in bacteria. *Genome Res.*, **18**, 949–956.

Zheng,Y. *et al.* (2005) Phylogenetic detection of conserved gene clusters in microbial genomes. *BMC Bioinformatics*, **6**, 243.