BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Software

Volume 9(17)

GIV: A Tool for Genomic Islands Visualization

Dongsheng Che* & Han Wang

Department of Computer Science, East Stroudsburg University of Pennsylvania, East Stroudsburg, PA 18301, USA; Dongsheng Che - Email: dche@po-box.esu.edu; Phone: 1-570-422-2731; *Corresponding author

Received September 30, 2013; Accepted October 01, 2013; Published October 16, 2013

Abstract:
A Genomic Islands (GI) is a chunk of DNA sequence in a genome whose origin can be traced back to other organisms or viruses. The detection of GIs plays an indispensable role in biomedical research, due to the fact that GIs are highly related to special functionalities such as disease-causing GIs - pathogenicity islands. It is also very important to visualize genomic islands, as well as the supporting features corresponding to the genomic islands in the genome. We have developed a program, Genomic Island Visualization (GIV), which displays the locations of genomic islands in a genome, as well as the corresponding supportive feature information for GIs. GIV was implemented in C++, and was compiled and executed on Linux/Unix operating systems.

Availability: GIV is freely available for non-commercial use at http://www5.esu.edu/cpsc/bioinfo/software/GIV.

Keywords: Prokaryotic genomes; Genomic islands; Sequence analysis; Visualization.

Background:
With the advances of high throughput sequencing technologies, lots of genomes have been sequenced and need to be analyzed. The huge amount of genomic data has in turn led to the development of visualization tools. Visualization tools can make the large data sets more meaningful, and mitigate the difficulties in detecting, filtering and classifying patterns within large gene sequences. Circos is one of popular visualization tools that are used to display genome information in circular ideogram with various configurations **[1].** This tool provides a great help for visually identifying and analyzing of similarities and differences across multiple genomes. It is also capable of generating data as line, histogram plots, heat maps, tiles, connectors, scatter, and text. We have recently developed a tool named GIV, a customized Circos, for displaying the locations of genomic islands and corresponding feature values in genomes. Genomic islands are chromosomal regions in a genome that have the origin of horizontal transfer. The stabilized GIs in the host genome can help itself adapt its new environment or condition, or even make it more competitive. For instance, the incorporation of drug-resistance genes in GIs can protect itself being killed, while the recruitment of secondary metabolite genes in GIs can help itself use surrounding resources more efficiently. The display of the locations of such GIs will make it extremely helpful for microbiologists and evolutionary biologists to study GIs, such as the study of the mechanism of forming GIs, or establishing the evolutionary relationships across genomes based on these GIs.

Methodology:
Software Input
The main purpose of generating GI visualization is to illustrate the positions of predicted GIs in the genome. In addition, we want to show the evidences of predicted GIs region by displaying all GI-associated feature values in the corresponding positions. To this end, we can display all feature values by different circular ideograms. By aligning different feature circles along with the original predicted GI circle, we can also identify which feature values are important when used in predicting GIs. Therefore, our tool GIV requires two input files: 1) genomic island locations; and 2) eight genomic island associated feature values (including IVOM, HEG, tRNA, Density, Phage, Integrase, Intergenic Distance and Transposases).

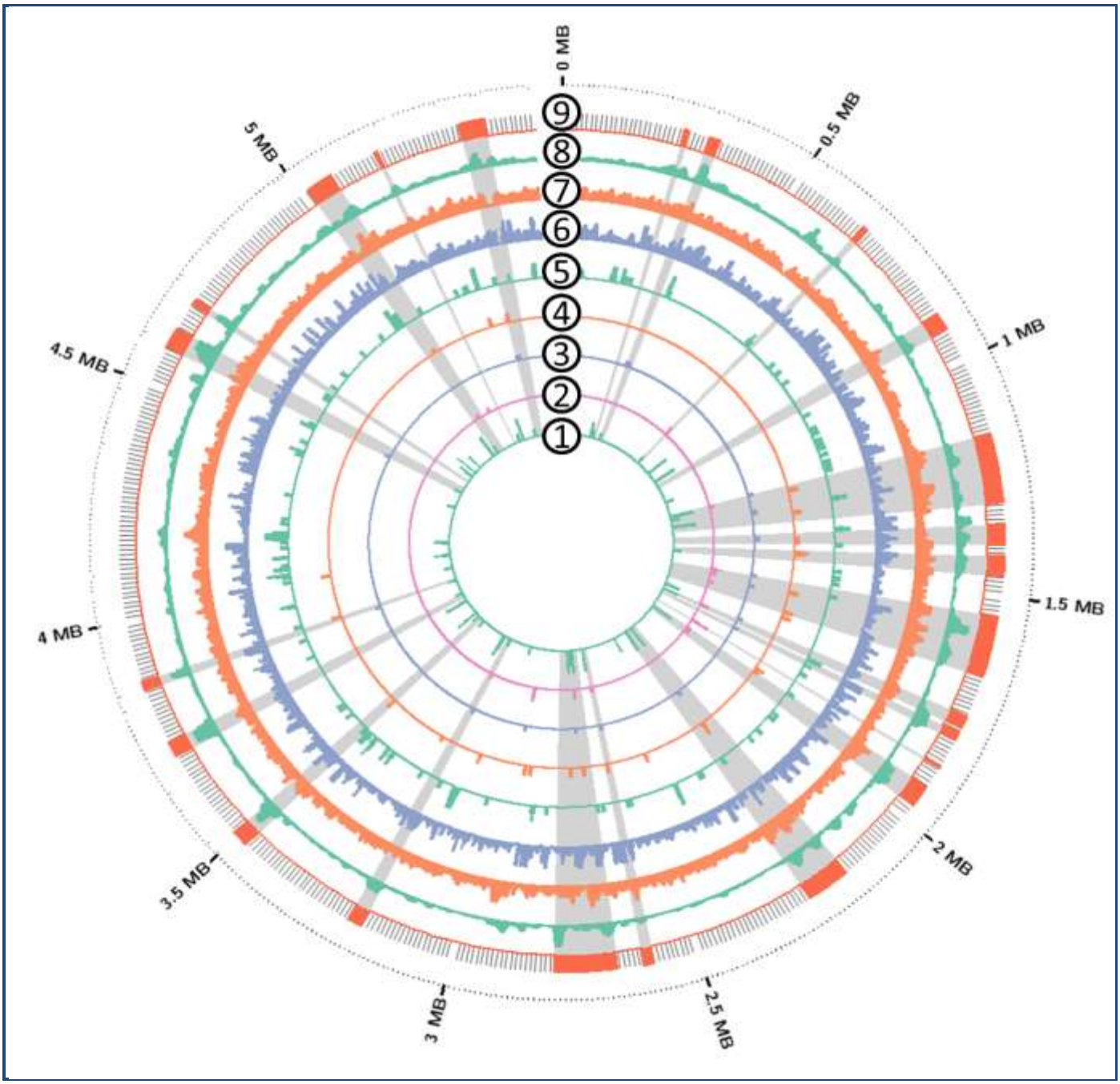

Figure 1: Genomic island visualization of Escherichia coli O157:H7 str. Sakai. There are eight circles (label 1-8), representing each of GI-associated features. The order of features from inside to outside is: **1)** tRNA, **2)** Phage, **3)** Integrase, **4)** Transposons, **5)** HEG, **6)** Intergenic-Distance, **7)** Density, and 8) IVOM. The most outside circle colored with orange color indicates the predicted GIs locations. The shaded areas cover the corresponding feature values for each of GI regions.

Raw Data Collection

For any fully sequenced genome, all eight GI-associated feature values can be generated by AlienHunter **[2],** and our in-house program GIHunter. Since AlienHunter is embedded in our program GIHunter, GIHunter can be used to calculate all eight feature values, including the IVOM score, which is actually generated by AlienHunter. GIHunter can be obtained at http:// www5.esu.edu/cpsc/bioinfo/software/GIHunter. **Table 1 (see supplementry material)** shows an example of part of raw data generated by GIHunter, which consists of start positions and end position of all genomic regions, and eight feature values corresponding to each of the regions.

Data Conversion

In order to use the Circos program to create GI images, we must convert all GI raw data into Circos' formatted data. Specifically, we read the original data file **(Table 1)** and separate them into different feature data files. In our study, eight features were used, and thus there are eight corresponding data files, with each of them with the Circos

formatted, which include the fields of chromosomal ID, start position, end position and value.

Configuration Setting

In Circos, each image should have a main configuration file, which includes general parameter setting, input files and output file. The parameter setting may include global color, font settings and tick mark setting, and they are defined in a separate configure file. In addition, the user is also required to define all input data files for the genome. In this study, we need to include eight GI feature-related data files, as well as predicted GI location file. Each of input data can have a choice of several Circos data type formats. In this study, we used the histogram representation, which includes chromosomal ID, start position, end position and value.

Sample Output

The output of our program will be an image, which will be saved in the same folder of input files, with the output name given by its genome name. Such information was automatically generated from our program input, where we provide genome information. A sample GI image is shown in **Figure 1,** where nine concentric circles were represented, with the most outside one representing the locations of predicted GIs, and the remaining circular ideograms representing eight feature values. Therefore, we can clearly see the relationships between the feature values and predicted GI region. We have run our GIV tool on more than 2000 genomes, converted the predicted GIs and GI-associated feature values by GIHunter to Circos' file format, and generated GI images for each of the genomes. The generated GI images have been uploaded to our genomic island database website: http://www5.esu.edu/cpsc/bioinfo/dgi.

Conclusion and Future Work

In this paper, we report the development and the usage of our genomic island visualization tool, GIV. We believe that this visualization tool will be helpful for medical and microbial scientific communities to study horizontal gene transfer, evolutionary microbial genomes. We also hope such a kind of visualization tool can serve as a model for visualizing other kinds of genomic structures with supporting feature values in the same genome for future genome analyses and studies.

Acknowledgement:

This research was partially supported by President Research Fund, FDR major grant, and FDR mini grant at East Stroudsburg University of Pennsylvania, USA.

References:

[1] Krzywinski M *et al. Genome Res*. 2009 **19**: 1639 [PMID: 19541911]

[2] Vernikos GS & Parkhill J, *Bioinformatics* 2006 **22**: 2196 [PMID: 16837528]

Edited by P Kangueane

Citation: Che & Wang, Bioinformation 9(17): 879-882 (2013)

BIOINFORMATION

open access

Supplementary material:

Table 1: An Example of Raw GI Data File**.** The first column contains genomic regions, while the remaining eight columns contain corresponding GI-associate feature values.

Location	IVOM	Density	Integrase	Phage	tRNA	HEG	Integenic Distance	Transposons
756572..764572	7.00	1.00	0	0	0	0	231.50	1
757873..765873	6.91	0.88	0	0	1	0	198.00	1
758670..766670	6.91	0.88	0	0	0	0	144.71	1
759701..767701	7.44	0.88	1	0	0	0	144.71	1
760633..768633	7.50	0.88	0	0	0	0	100.71	0
761130..769130	7.50	1.00	0	0	0	0	102.50	0