BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Software

Volume 8(4)

GIST: Genomic island suite of tools for predicting genomic islands in genomic sequences

Mohammad Shabbir Hasan[1], Qi Liu[2], Han Wang[1], John Fazekas[1], Bernard Chen[3], & Dongsheng Che[1]*

[1]Department of Computer Science, East Stroudsburg University, East Stroudsburg, PA 18301, USA; [2]College of Life Science and Biotechnology, Tongji University, Shanghai, 200092, China; [3]Department of Computer Science, University of Central Arkansas, Conway, AR, 72035, USA; Email: dche@po-box.esu.edu; Phone: 1-570-422-2731; *Corresponding author

Received February 06, 2012; Accepted February 07, 2012; Published February 28, 2012

Abstract:
Genomic Islands (GIs) are genomic regions that are originally from other organisms, through a process known as Horizontal Gene Transfer (HGT). Detection of GIs plays a significant role in biomedical research since such align genomic regions usually contain important features, such as pathogenic genes. We have developed a use friendly graphic user interface, Genomic Island Suite of Tools (GIST), which is a platform for scientific users to predict GIs. This software package includes five commonly used tools, AlienHunter, IslandPath, Colombo SIGI-HMM, INDeGenIUS and Pai-Ida. It also includes an optimization program EGID that ensembles the result of existing tools for more accurate prediction. The tools in GIST can be used either separately or sequentially. GIST also includes a downloadable feature that facilitates collecting the input genomes automatically from the FTP server of the National Center for Biotechnology Information (NCBI). GIST was implemented in Java, and was compiled and executed on Linux/Unix operating systems.

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

Keywords: Prokaryotic genomes, Genomic islands, Sequence analysis

Background:
Some prokaryotic genomes contain genomic sequences with different patterns than the remaining parts of the host genomes. Such differences may include GC content bias **[1],** codon usage bias **[2, 3],** k-mer nucleotide frequency bias **[4],** and the existence of mobile genes such as integrase genes and transposes genes **[5].** In some other cases, such regions are also bordered by transfer RNAs (t-RNA) **[6].** The abnormal regions that contain such types of characteristics are known as Genomic Islands (GIs). Research on identifying genomic islands has become more important as the scientific community can be significantly benefitted from such findings. Biomedical researchers and microbiologists can use the results to explain the pathogenicity of organisms, or discover industrial important metabolic components from GIs. Based on such findings, pharmacists can use them to design corresponding vaccines and antibiotics, and eventually promote pharmaceutical companies to produce medicines at a large scale. As it is generally believed that each genome contains unique genomic sequence signature, some computation tools based on sequence signature have been developed. Such sequence composition based tools include AlienHunter **[7]**, COLOMBO SIGI-HMM **[8],** GIDetector **[9],** IslandPath **[10],** INDeGenIUS **[11],** and PAI-IDA **[12].** Recent studies have shown that none of these tools can predict GIs accurately in all genomes **[13].** Hence it necessary to develop a computational framework that produces a better prediction results by combining the results of existing programs **[14].** We have

recently developed a tool, EGID, which has shown to optimize the results of individual tools, and produce a better prediction result for all genomes **[15].**

We realize that the majority users of these tools are biologists. Unfortunately, these programs are command line based, and different programs usually require different inputs to predict GIs, thus making it difficult for such group of people to use these tools for genomic island analyses. To this end, we have developed a user friendly graphic user interface, GIST, which contains a suite of tools for GI prediction. GIST provides a feature that allows user to download the necessary files required to run the tools automatically from the FTP server of the National Center for Biotechnology Information (NCBI) ftp://ftp.ncbi.nih.gov/genomes/Bacteria). Depending on the user's interest, GIST allows the user to select any combination of the tools, invokes and runs selected programs in the back end, generates and organizes prediction results. We believe that the development of GIST should benefit the scientific communities for easy use in studying genome evolutions and gene transfer mechanisms.

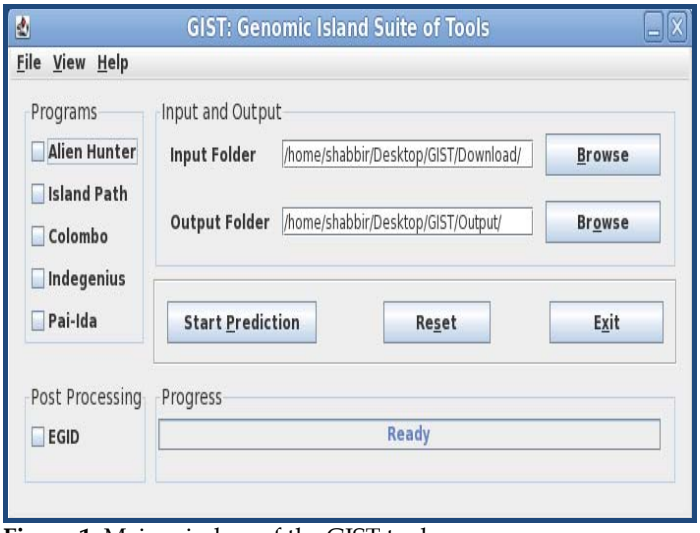

Figure 1: Main window of the GIST tool.

Software Input and Output:
GIST includes five individual GI prediction programs, as well as the optimization tool EGID, which uses the prediction results of any combination of individual programs as the inputs, and produces consensus predicted GIs. The GUI layout of GIST is shown in **Figure 1.** GIST requires five different types of files for any single genome for GI prediction. These file types include FNA, FAA, FFN, GBK and PTT, where the required information such as k-mers, G+C content, codon usage, and dinucleotide frequency can be extracted. For the same genome, all of these files need to be saved in the same directory that is used as the input for that genome. If users are only interested in a particular program, they can select the program from the 'Programs' panel and hit the 'Start Prediction' button. It is important to note that if EGID is selected, it executes other tools along with itself thereby produces the optimized prediction results.

Users can specify the output folder location; otherwise the output files are saved into the default output directory. The output file for each tool is a text file containing the start and stop positions of the genomic island regions for the input genome. For the detailed usage of GIST for GI prediction, please refer to the user guide of our website (http://www5.esu.edu/cpsc/bioinfo/software/GIST).

Automatic Genome Download Feature:
One of the most important features of GIST is the functionality of automatic connecting and downloading of the required genomic files through the FTP server of NCBI, as shown in **(Figure 2).** The panel 'FTP Directory' contains the tree representation of the organisms available in the FTP server of NCBI. User can select any genome that belongs to any of the organisms by exploring the tree node of that organism.

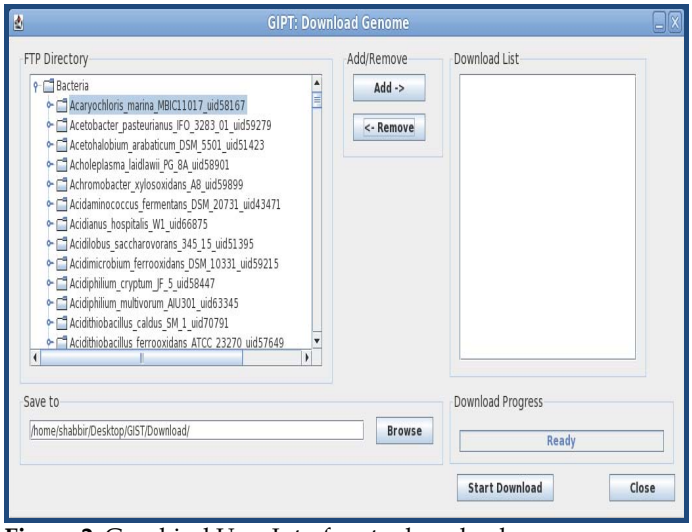

Figure 2: Graphical User Interface to download genomes

To add a genome into the download list, the user can double click on that genome name or use the 'Add' button in the 'Add/Remove' after selecting that genome. On the other hand, to remove any genome from the download list, the user can use the 'Remove' button. When the 'Start Download' button is pressed, necessary files of all genomes in the download list are downloaded automatically and the progress bar shows the download progress status. Downloaded files are saved into the corresponding directory of each genome. User can specify the directory location to save the downloaded files. By default, this program saves the downloaded files in the 'Download' directory (GIST_1.0/Download) if the location is not specified by the user.

Caveat and Future Development:
The current version of GIST produces prediction results in text file. In the next version, we will integrate the visualization feature such as circular representation, so that users can easily compare the results.

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

References:
[1] Hacker J *et al*. *Mol Microbiol*. 1997 **23**: 1089 [PMID: 9106201]
[2] Lawrence JG & Ochman H, *J Mol Evol*. 1997 **44**: 383 [PMID: 9089078]
[3] Karlin S *et al*. *Mol Microbiol*. 1998 **29**: 1341 [PMID: 9781873]

[4] Tsirigos A & Rigoutsos I, *Nucleic Acids Res*. 2005 **33**: 922 [PMID: 15716310]
[5] Schmidt H & Hensel M, *Clin Microbiol Rev*. 2004 **17**: 14 [PMID: 14726454]
[6] Hacker J & Kaper JB, *Annu Rev Microbiol* 2000 **54**: 641 [PMID: 11018140]
[7] Vernikos GS & Parkhill J, *Bioinformatics*. 2006 **22**: 2196 [PMID: 16837528]
[8] Waack S *et al*. *BMC Bioinformatics*. 2006 **7**: 142 [PMID: 16542435]
[9] Che D *et al*. *BMC Genomics*. 2010 **11**: S1 [PMID: 21047376]
[10] Hsiao W *et al*. *Bioinformatics*. 2003 **19**: 418 [PMID: 12584130]
[11] Shrivastava S *et al*. *J Biosci*. 2010 **35**: 351 [PMID: 20826944].
[12] Tu Q & Ding D, *FEMS Microbiol Lett*. 2003 **221**: 269 [PMID: 12725938]
[13] Langille MG. *BMC Bioinformatics*. 2008 **9**: 329 [PMID: 18680607]
[14] Langille MG *et al*. *Nat Rev Microbiol*. 2010 **8**: 373 [PMID: 20395967]
[15] Che *et al*. *Bioinformation*. 2011 **7**: 311 [PMID: 3280502]

Edited by P Kangueane
Citation: Hasan *et al*. Bioinformation 8(4): 203-205 (2012)