Proceedings of the 2012 IEEE 12th International Conference on Bioinformatics & Bioengineering (BIBE), Larnaca, Cyprus, 11-13 November 2012

GIDT - A Tool For The Identification And Visualization Of Genomic Islands In Prokaryotic Organisms

Haswanee D. Goodur, Vyasanand Ramtohul, Shakuntala Baichoo
Department of Computer Science and Engineering
University of Mauritius
shakunb@uom.ac.mu

Abstract— **For decades, it has been a challenge for biologists to identify genomic islands (GIs) within a bacterial genome as they usually rapidly evolve. The purpose of this research is to develop an application which will analyse DNA sequences, enabling researchers to be up-to-date with bacterial evolution. A Java-based GI detection tool, "Genomic Islands Detection Tool (GIDT)" is introduced to detect GI regions by using a number of nucleotide-based statistical methods and genic methods (including GC-content variation, codon usage bias, dinucleotide frequency bias, tetranucleotide frequency bias, and k-mer signature analysis) and identification of mobility genes. It takes as input genome files in embl/genbank -file formats and returns probable GI regions in a tree-view display along with a circular-view display. GIDT is a simple tool which uses six GI identification algorithms and visually displays probable GI regions in a given genome. It runs on Microsoft windows, MacOS and Linux.**

Keywords- ***bioinformatics; genomic island; prokaryotic organism; GI detection***

I. INTRODUCTION

Over the past few years, researchers have discovered that apart from the fundamental genes encoding essential metabolic functions, genomes also harbour a variable amount of accessory genes acquired by horizontal gene transfer (HGT) that encode adaptive traits, which might be beneficial for the species under certain growth or environmental conditions [1]. This has aroused new challenges in the medical as well as the agricultural sector and for this reason, the analysis of bacterial genome has become a major application in the bioinformatics field. A significant part of HGT is or has been assisted by GIs (syntenic blocks formed by many accessory genes). GIs are generally recognized as discrete DNA segments between closely related strains.

One of the emerging ideas is that GIs cover an overarching family of elements, including mobile genetic elements (MGEs) such as integrative and conjugative elements (ICEs), conjugative transposons and some prophages.

GIs have many general features; they are often inserted at tRNA genes and are flanked by 16-20 bp perfect direct repeats (DR). They are normally large (10-200kb) with small genomic islets (<10kb). Moreover, GIs may be predicted by nucleotide statistics that generally differ from the rest of the genome.

Using these general features, GI regions can be predicted. The most common GI identification methods are the diversities in sequences between the GI and the host DNA, including codon usage, Guanine-Cytosine (GC) content, k-mer signature analysis and the frequency of specific dinucleotides and tetranucleotides.

II. IMPLEMENTATION

The Genomic Island Detection Tool (GIDT) was implemented to identify GIs, using GC content, codon usage bias, dinucleotide frequency bias, tetranucleotide frequency bias, k-mer signature analysis (2-mer, 3-mer, 4-mer, 5-mer, and 6-mer) and presence of mobility genes. GIDT is a stand-alone application which has a simple graphical interface; where disclosed GIs are displayed in a tree-view and a circular graph. In the tree-view, the GI starting and ending position along with the genes it contains are plotted, while, in the circular graph, the results of each method along with the integrated one are drawn. The different methods implemented in GIDT are described below:

A. Guanine-Cytosine Content Variation

The guanine-cytosine content (GC-content) is the percentage of guanine or cytosine nitrogenous bases in DNA or RNA molecule. GC pairs are bound by three hydrogen bonds while AT pairs are bound by two only. For that reason, DNA with high GC-content is more stable. In addition, GC-content is highly affected by the environment, for example, "the bacteria from a sample of surface sea water had a median GC-content of 34%, while a soil sample had a median of 61." [2], thus dissimilar genomes will have different GC-contents.

GC-content is usually referred to as a percentage value but can sometime be represented as a ratio (G+C ratio). The formula for calculating GC-content is [3]:

$$\%GC = \frac{G+C}{A+T+G+C} * 100$$ **Percentage GC-content**

$$GC - ratio = \frac{A+T}{C+G}$$ **Ratio of GC-content**

B. Codon Usage Bias

A codon is a triplet of nucleotides that encodes for an amino acid. There are four nucleotides namely Adenine (A),

978-1-4673-4358-9/12/$31.00 ©2012 IEEE

Cytosine (C), Guanine (G), and Uracil (U), which give a total of 64 possible codon combinations. Each codon codes for one specific amino acid, but there can be several codons coding for the identic amino acid, for instance, GCU, GCC, GCA and GCG would all code for Alanine.

In a particular genome, one of these codons which encode for the same amino acid will have a higher preference than the others and will appear more often due to the abundance of its specific transport ribonucleic acid (tRNA). Genes that come from other genomes (i.e. GI) will have codons according to the codon preference of their source genome. Hence, using this codon preference idea, genes, which are foreign (GI), can be identified.

The relative synonymous codon usage (RSCU) value for each amino acid is used to observe the affinity for a definite codon since distinct organism has unusual affinity to different tRNA. The 'relative adaptedness' value, W_i, of a specific codon is calculated as:

$$W_i = \frac{RSCU_i}{RSCU_{MAX}}$$

Relative Adaptedness

where $RSCU_i$ = frequency of codon i in the subset of highly expressed genes and $RSCU_{MAX}$ = frequency of codon most often used to code for the considered amino acid in the subset of highly expressed genes.

The Codon Affinity Index (CAI) for a gene is then defined as the geometric mean of W_i values for codons in that gene. Genes with low ***CAI*** value can probably be a GI gene where ***CAI*** is computed as:

$$CAI = \left(\prod_{i=1}^{L} W_i\right)^{1/L} \; or \; exp\left(\frac{1}{L}\sum_{i=1}^{L} ln(W_i)\right)$$

Codon Affinity

where L is the number of codons in the gene excluding the start codon (methionine), tryptophan and the stop codons [4].

C. Dinucleotide Frequency Bias

In 1995, Karlin and Burge [5] described dinucleotide bias, a genome signature which is remarkably stable within a genome. A dinucleotide consists of two nucleotide bases, which can be a combination of either A, Thymine (T), C, or G. Hence, there can be 16 possible dinucleotides. The dinucleotide composition of a particular genome is said to be constant throughout the full genome. That is, if the percentage of a dinucleotide ***XY*** in an entire genome is Z, then a subset of this genome should also have around the same percentage composition of this dinucleotide. Different genomes have different dinucleotide compositions; thence, a gene which comes from another genome (GI) will have dinucleotide composition similar to its source genome rather than the one it is currently in. That's why; using this information it is possible to detect genome segments, which are foreign (GI). This is calculated by ascertaining relative abundance values:

$$P_{xy} = \frac{f_{XY}}{f_X f_Y}$$

Relative Abundance

where f_{XY} is the frequency of a dinucleotide in a region and f_X and f_Y is the frequency of the mononucleotides in the dimer. The frequencies of both strands of the DNA sequence region are calculated in order to compensate for any asymmetry. In 2001, Karlin [6] also reported that a helpful way of calculating the differences between the relative abundance value for a given region and the value of the whole genome is through:

$$\sigma(f,g) = \frac{1}{16}\sum_{XY} |P_{XY}(a) - P_{XY}(b)|$$

Average difference between genomes

where ***a*** would be the query region and ***b*** would be another region in the genome.

D. Tetranucleotide Frequency Bias

A tetranucleotide consists of four nucleotide bases, which can be a combination of either A, T, C, or G, and, thus, there can be 256 possible tetranucleotides. The latter is very alike to dinucleotide but Pride et al. [7] stated that tetranucleotide-based clustering was more relevant than dinucleotide-based clustering. So, they suggested that using tetranucleotide usage departure from expectations (TUD) could help to disclose GIs. TUD is the ratio ***F(W)*** where it is calculated as such:

$$F(W_i) = \frac{O(W_i)}{E(W_i)}$$

Departure from expectations

where $O(W_i)$ is the observed occurrence value, and $E(W_i)$ is the expected occurrence value of a tetranucleotide W_i. whereby the $E(W_i)$ value is calculated by:

$$E(W = w_1w_2w_3w_4) = f(w_1)f(w_2)f(w_3)|S|$$

Expected occurrence of tetranucleotide (method1)

where ***Wi*** is the i^{th} nucleotide of W; ***f(A), f(T), f(G),*** and ***f(C)*** are nucleotide frequencies for the sequence S and $|S|$ is the length of the sequence.

$$E(W = w_1w_2w_3w_4) = \frac{O(w_1w_2w_3)O(w_2w_3w_4)}{O(w_2w_3)}$$

Expected occurrence of tetranucleotide (method2)

In order to identify GIs, the divergence between observed and expected tetranucleotide frequency is calculated using the z-score approximation.

$$Z(W = w_1w_2w_3w_4) = \frac{O(w_1w_2w_3w_4) - E(w_1w_2w_3w_4)}{\sqrt{varO(w_1w_2w_3w_4)}}$$

Zscore of tetranucleotides

where the ***varO(W)*** can be approximated as follows:

$$varO(W) = E(W)\frac{|O(w_2w_3) - O(w_1w_2w_3)||O(w_2w_3) - O(w_2w_3w_4)|}{O(w_2w_3)^2}$$

Variance of tetranucleotides

The Pearson correlation coefficient (r) for the z-scores is used to determine whether the two genomic sequences exhibit a similar pattern for over- or under-represented tetranucleotides. It is defined as follows:

$$r = \frac{\sum Z_x Z_y}{N}$$

Pearson correlation coefficient

where ***r*** is the Pearson correlation coefficient.

Genomic fragments with similar patterns are determined by a high correlation coefficient while distinct patterns are the one with low correlation coefficients [8]. Therefore, it is obvious that the dissimilar patterns are foreign, hence credible GIs.

E. Presence of Mobility Genes

During HGT, MGEs such as integrase and transposase genes are acquired [9] along with some virulence factor genes [10]. These cluster genes are of probable horizontal origin and may be identified using Annotation in .embl and .gbk files, thus, help in disclosing possible GIs.

F. K-mer Signature Analysis

K-mer mostly refers to a specific n-tuple or n-gram for nucleic acid or amino acid sequences, which are used to identify certain regions within biomolecules such as DNA or proteins respectively. K-mer analysis is commonly used to predict biological meaningful clusters of DNA words (k-mers) and genomic entities. “Genome entities as diverse as genes, CpG dinucleotides, transcription factor binding sites (TFBSs) or ultra-conserved non-coding regions usually form clusters along the chromosome sequence" [11].

K-mer analysis algorithm detects the distance between a cluster of words in DNA sequence and neighbouring DNA sequences. Benjamin [12] stated that k-mer frequency analysis have been used to identify lateral gene transfer and since k-mer frequency, signatures are generally distinct across distinctive species, the frequency signatures of segments of a sequence can be compared with the signature of whole genome of the organism. If these are significantly different, they may be probable GIs.

To calculate the distance between the query segment and the whole genome sequence, the Euclidean distance algorithm is used. The formula is given as:

$$d(p,q) = d(q,p)$$

Euclidean distance (Equation 1)

such that

$$d(q,p) = \sqrt{((q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2)}$$

Euclidean distance (Equation 2)

such that

$$\sqrt{((q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2)} = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$$

Euclidean distance (Equation 3)

where ***p*** is the array of k-mer frequency signatures and ***q*** is the k-mer frequency signature of the whole genome.

III. THE INTERFACE OF GIDT

GIDT provides an easy-to-use interface where users are first prompted to load a bacterial file and thereafter requested to choose one/more methods to identify GIs (Fig. 1). A help button is also provided for the convenience of users, such that clicking on the same a user manual is displayed.

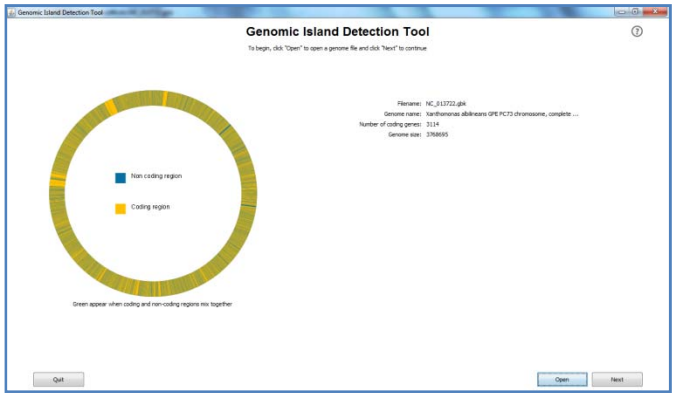

Figure 1. GIDT LoadFile Interface

After choosing the method/s, GIDT will display the list of probable GIs as per each method and the list of probable GIs identified by more than one method. For each probable GI region, the list of coding sequences found in each GI is also displayed in a tree view (Fig. 2). User can view the contents of each region and if required can choose and export a specific region to fasta. The interface also provides the facility to export the circular graph to bitmap or vector graphics.

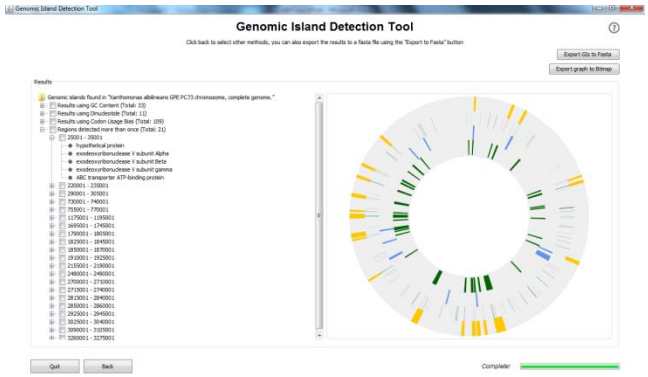

Figure 2. GIDT Output for Xanthomonas albilineans

If the user chooses mobility genes as one of the methods to identify GIs, a predefined list of mobility genes is provided. It must be noted that for the mobility genes a number of synonyms may exist and thus users of the system may configure the provided list of mobility genes (e.g. add more).

GIDT's interface is better represented than the existing software as it has an animated circular graph which gives the methods used as well as the GI's position. Besides, the tree-view graph allots detailed information about the GI's position by stating its genetic content. Figure 2 gives the result of GIDT using Xanthomonas albilineans in GenBank format.

IV. Results and Discussion

The reliability of GIDT is checked by comparing its outputs with that of IslandViewer [13] and SeqWord Sniffer GI Browser [14]. A sample output of running the software for Xanthomonas Albilineans (NC_013722) is shown in Fig. 3.

IslandViewer is a web-site developed for researchers to view and download GIs. The facility of uploading any unpublished and yet unknown genome is provided. The latter comprises of many refined practices such as IslandPick [9], IslandPath-DIMOB [15] and SIGI-HMM [16] which are very resource extensive. On the other hand, SeqWord is an online tool for the identification and visualization of GI regions of bacterial genomes through oligonucleotide usage. It can also be downloaded and installed.

Despite using very sophisticated algorithms, these two applications have some drawbacks. Unlike GIDT, IslandViewer depends upon Internet connection. SeqWord, on the contrary, can be used locally but the interface of the software is not very user-friendly.

For comparison purposes, a case study on "Xanthomonas albilineans" is carried out. The outputs of IslandViewer, SeqWord Sniffer GI Browser and GIDT are given in figure 3.

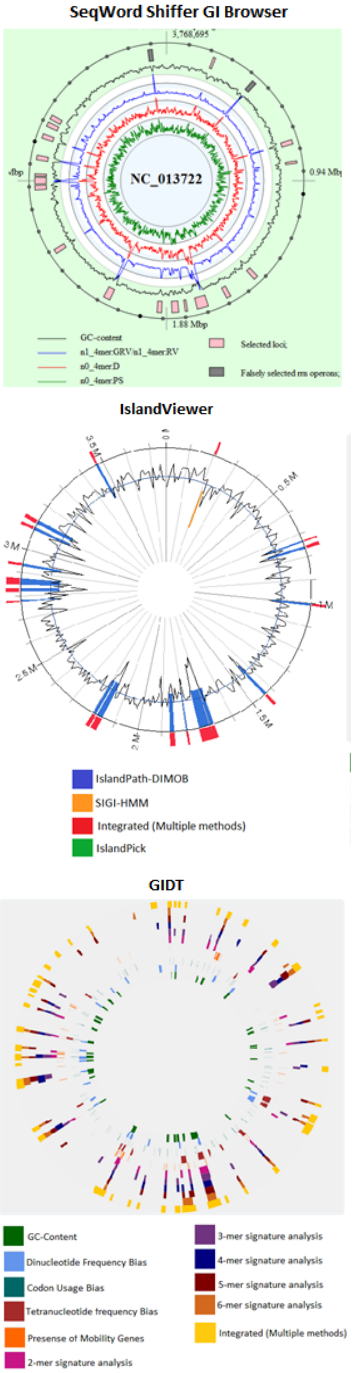

Figure 3. GIs in NC_013722.gbk

It is clearly shown that the results of the two existing software are present in GIDT. But, GIDT identifies more GI regions as its resulting GI regions are predicted by combining the outcomes of two or more algorithms. GIDT is stand-alone software; it does not need any database unlike the other two.

Its output can be exported to a .fasta file which can later be BLAST to find the origin of the GI segment.

These detailed data may help a researcher to know the purpose for which the foreign segment was inserted into the DNA sequence of the host prokaryote. One last point is that, IslandViewer takes GenBank and Embl files as input, while SeqWord takes GenBank and Fasta only. Conversely, GIDT accepts all these three file formats as input.

V. CONCLUSION

In this paper, a GI identification tool, GIDT, was developed which uses nucleotide-based statistics, for instance, CG content, codon usage bias, genome signature (dinucleotide frequency bias), tetranucleotide frequency bias, and k-mer signature analysis along with the presence of mobility genes. The outcome of GIDT is very similar to that of SeqWord Sniffer GI Browser and IslandViewer. Moreover, GIDT gives an integrated result which is accurate and it runs locally as compared to IslandViewer.

Despite, using multiple methods, categorising GIs are not that easy as the foreign DNA sequences get adapted to the new host and evolve to incorporate the genome, making it difficult to identify. Gene amelioration [17] [19] (the process whereby the sequence of the island becomes similar to that of the host in GC content and codon usage due to mutational biases of the host) may occur and obscure the GI, and for this reason it is less likely to be identified as an island. For detecting ameliorated GIs a phylogenomics approach is definitely necessary, and this issue is not addressed by GIDT.

Moreover, methionine (ATG) is the start codon for genes, but, in prokaryotes, there are two alternate start codons namely GTG and TTG, which are basically Valine and Leucine respectively [18]. This complicates the identification of GI by considering the genes. These problems can be solved by using the genome comparison method.

ACKNOWLEDGMENT

The authors wish to thank Prof. Y.Jaufeerally-Fakim from the Faculty of Agriculture, University of Mauritius for her constructive advice on this work. A special thanks goes to the BioJava team for creating such a powerful and yet easy to use library.

REFERENCES

[1] Juhas Mario, van de Meer Jan Roelof, Gaillard Muriel, Harding Rosalind M, Hood Derek W, and Crook Derrick W, March 2009, "Genomic islands: tools of bacterial horizontal gene transfer and evolution", FEMS Microbiol Rev, no. 33(2), pp. 376-393

[2] Hildebrand F, Meyer A, Eyre-Walker A , 2010, "Evidence of Selection upon Genomic GC-Content in Bacteria", PLoS Genet 6(9): e1001107. doi:10.1371/journal.pgen.1001107

[3] Hurst Laurence D. and Merchant Alexa R., March 2, 2001, 'High guanine-cytosine content is not an adaptation to high temperature: a comparative analysis amongst prokaryotes', Proc. R. Soc. Lond. B 2001 268, 493-497 doi: 10.1098/rspb.2000.1397

[4] Moriyama Etsuko N, 2003, 'Codon Usage', University of Nebraska, Lincoln, Nebraska, USA

[5] Karlin S and Burge C., July 1995, 'Dinucleotide relative abundance extremes: a genomic signature', TRENDS in Microbiology,, vol.11, issue.7, pages.283-290

[6] Karlin Samual, July 2001, 'Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes', TRENDS in Microbiology, vol. 9. No. 7, pages.335-343

[7] Pride David T, Meinersmann Richard J, Wassenaar Trudy M, and Blaser Martin J, January 14, 2003, 'Evolutionary Implications of Microbial Genome Tetranucleotide Frequency Biases', Spring Harbor Laboratory Press, vol.13, pages.145-158

[8] Bolshoy Alexander, Volkovich Zeev (Vladimir), Kirzhner Valery and Barzily Zeev, 2010, 'Genome Clustering: from Linguistic Models to Classification of Genetic Texts', Scientific Publishing Services Pvt. Ltd., Chennai, India

[9] Langille Morgan Gavel Ira, 2009, 'Computational Prediction and characterisation of Genomic Islands: Insights into Bacterial Pathogenicity', Simon Fraser University

[10] Sui Shannan J.Ho, Fedynak Amber, Hsiao William W.L., Langille Morgan G.I. and Brinkman Fiona S.L., 6 October 2007, 'The Association of Virulent Factors with Genomic Islands', PLoS ONE 4(12): e8094. doi:10.1371/journal.pone.0008094

[11] Hackenberg Michael, Carpena Pedro, Bernaola-Galvan Pedro, Barturen Guillermo, Alganza Angel M and Oliver Jose L, January 24, 2011, "WordCluster: detecting clusters of DNA words and genomic elements", Algorithms Molecular Biology, Vol. 6, issue 2. doi:10.1186/1748-7188-6-2

[12] Benjamin Ashlee, May 2009, "Genetic Elements of Microbes: A Comprehensive and Integrated Genomic Database Application", Rochester Institute of Technology

[13] Morgan G.I. Langille and Fiona S.L. Brinkman, January 61, 2006, "IslandViewer: an integrated interface for computational identification and visualization of genomic islands", Bioinformatics, vol. 25, issue. 5, pages.664-665, doi: 10.1093/bioinformatics/btp030

[14] Reva, O.N., and Tümmler, B, 2005, "Differentiation of regions with atypical oligonucleotide composition in bacterial genomes", BMC Bioinformatics, vol. 6, No.251, doi:10.1186/1471-2105-6-251

[15]Hsiao William, Wan Ivan, Jones Steven J., and Brinkman Fiona S.L., Feb 12, 2003, 'IslandPath: aiding detection of

genomic islands in prokaryotes', Bioinformatics, vol.19 no. 3 2003, pages 418-420

[16] Waack Stephan, Keller Oliver, Asper Roman, Brodag Thomas, Damm Carsten, Fricke Wolfgang Florian, Surovcik Katharina, Meinicke Peter, and Merkl Rainer, 2006, 'Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models', BMC Bioinformatics, v.7., issue.1

[17] Lawrence JG and Ochman H, April 1997, "Amelioration of bacterial genomes: rates of change and exchange", Journal of Molecular Evolution, vol. 44, issue. 4, pages.383-397

[18] Elzanowski Andrzej (Anjay) and Ostell Jim, 2010, "The Genetic Codes", Available: http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi [Accessed: July 3, 2011]

[19] Marri P., R., Golding G.,B., "Gene amelioration demonstrated: The journey of nascent genes in bacteria", Genome, 2008, VOL 51; NUMB 2, pages 164-168

SAMPLE GENOME FILE USED FOR TESTING FROM NCBI

Xanthomonas albilineans str. GPE PC73, chromosome, complete genome, NC_013722 (GenBank format)

AVAILIBILITY OF GIDT

GIDT can be obtained by sending an email to author shakunb@uom.ac.mu.