BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Web server

Volume 7(6)

IGIPT - Integrated genomic island prediction tool

Ruchi Jain, Sandeep Ramineni, Nita Parekh*

Centre for Computational Natural Sciences and Bioinformatics International Institute of Information Technology, Gachibowli, Hyderabad, India; Nita Parekh - Email:nita@iiit.ac.in.; *Corresponding author

Received November 03, 2011; Accepted November 08, 2011; Published November 20, 2011

Abstract:
IGIPT is a web-based integrated platform for the identification of genomic islands (GIs). It incorporates thirteen parametric measures based on anomalous nucleotide composition on a single platform, thus improving the predictive power of a horizontally acquired region, since it is known that no single measure can absolutely predict a horizontally transferred region. The tool filters putative GIs based on standard deviation from genomic average and also provide raw output in MS excel format for further analysis. To facilitate the identification of various structural features, viz., tRNA integration sites, repeats, etc. in the vicinity of GIs, the tool provides option to extract the predicted regions and its flanking regions.
Availability: http://bioinf.iiit.ac.in/IGIPT/

Keywords: genomic islands, horizontal gene transfer

Background:
A horizontally transferred event is defined as movement of genetic material between phylogenetically unrelated organisms by mechanisms other than vertical descent. These regions from diverse organisms, called Genomic Islands (GIs), are typically 10-200Kb in size (containing clusters of genes). Any biological advantage provided to the recipient organism by transferred DNA creates selective pressure for its retention in the host genome and several pathways of horizontal transfer have been established influencing traits such as antibiotic resistance, symbiosis and fitness, virulence and adaptation **[1]**. For example, horizontal gene transfer has been demonstrated in many pathogenic strains of bacteria and shown to be responsible for its virulence. The identification of genomic islands also forms the first step in the annotation of newly sequenced genomes. Various bioinformatics approaches have been proposed in their identification **[2]**. In the genomic era, with availability of large number of bacterial genomes, the preferred methods are based on nucleotide base compositions and comparative genomics. In IGIPT, we have implemented thirteen measures that capture anomaly in nucleotide composition, providing both genome-based and gene-based search on a single platform.

Methodology:
In any genome, vertically transmitted genes experience a particular set of directional mutation pressures mediated by the specific features of the replication machinery of the cell, such as balance of dNTP pools, mutational biases of the DNA polymerases, efficiency of mismatch repair systems and so on **[3]**. As a result each genome exhibits its own unique signatures, viz., distinct variations in the GC content, dinucleotide relative abundance, variations in usage of k-mer words, codons and amino acids. These measures, called parametric methods, are the most widely used approaches as the putative transferred genes can be identified without relying on comparisons with other organisms, thus providing an independent means of assessing the impact of gene transfer across lineages. The parametric measures implemented in IGIPT are broadly classified as genome-based or gene-based, depending on the analysis (shown as left- and right panel in **Figure.1**). These measures are computed in a sliding window and regions deviant from the genomic average by user defined standard deviation (default 1.5σ) are identified as probable GIs.

Measures at Genome Level
The major advantage of these measures is that they do not require pre-existing annotation or comparison of homologous

sequences, and can, therefore, be applied directly to newly sequenced genomes. The input to these measures is the complete genome/contig in Fasta format.

GC content
It computes the frequency of G and C nucleotides, called the GC content **[4]**.

Genomic signature
The set of dinucleotide relative abundance values constitutes a "genomic signature" of an organism. **Please see supplementary material.**

k-mer Distributions
It has been proposed by Karlin that most horizontally acquired genomic regions have distinct word (*k-mer*) compositions **[5]. Please see supplementary material**

Measures at the Gene Level
This module identifies horizontally acquired genes in a fully annotated gene set of the organism (in multi-fasta format). In the absence of this information, IGIPT provides comparison of two gene sets, one a representative gene set of the organism and the other whose horizontal acquisition needs to be confirmed (e.g., genes in predicted GIs from genome-based measures). This feature also allows comparison of predicted gene(s) with highly expressed genes of the organism, e.g., ribosomal genes, chaperon genes, etc. to reduce false predictions.

Codon usage Bias
The unequal usage of synonymous codons has been extensively studied and virtually every codon has been shown to be preferentially used in some organisms and rarely used in others. **Please see supplementary material.**

Amino Acid Bias
This bias refers to the deviation in the frequency of usage of individual amino acids over the average usage of all 20 amino acids. **Please see supplementary material.**

GC content at Codon Positions
This involves comparing the frequency of G or C at the three codon positions, GC1, GC2 and GC3, for a given gene set with the core gene set (or genomic average or highly expressed genes) of the organism **[8]**.

IGIPT provides an option to download the predicted horizontally transferred regions/genes and its flanking regions (lower panel in **Figure 1**) to facilitate analysis of conserved structural features in the vicinity of probable GIs, e.g., genes coding for integrases or transposases required for chromosomal integration and excision are flanked by direct repeats and are inserted in the vicinity of tRNA and tmRNA genes **[9]**. This feature is also useful for further analysis such as comparative genomics or phylogenetic analysis of putative GIs. The output of IGIPT is windows/genes filtered based on standard deviation and also provides option to download unfiltered output in MS excel format.

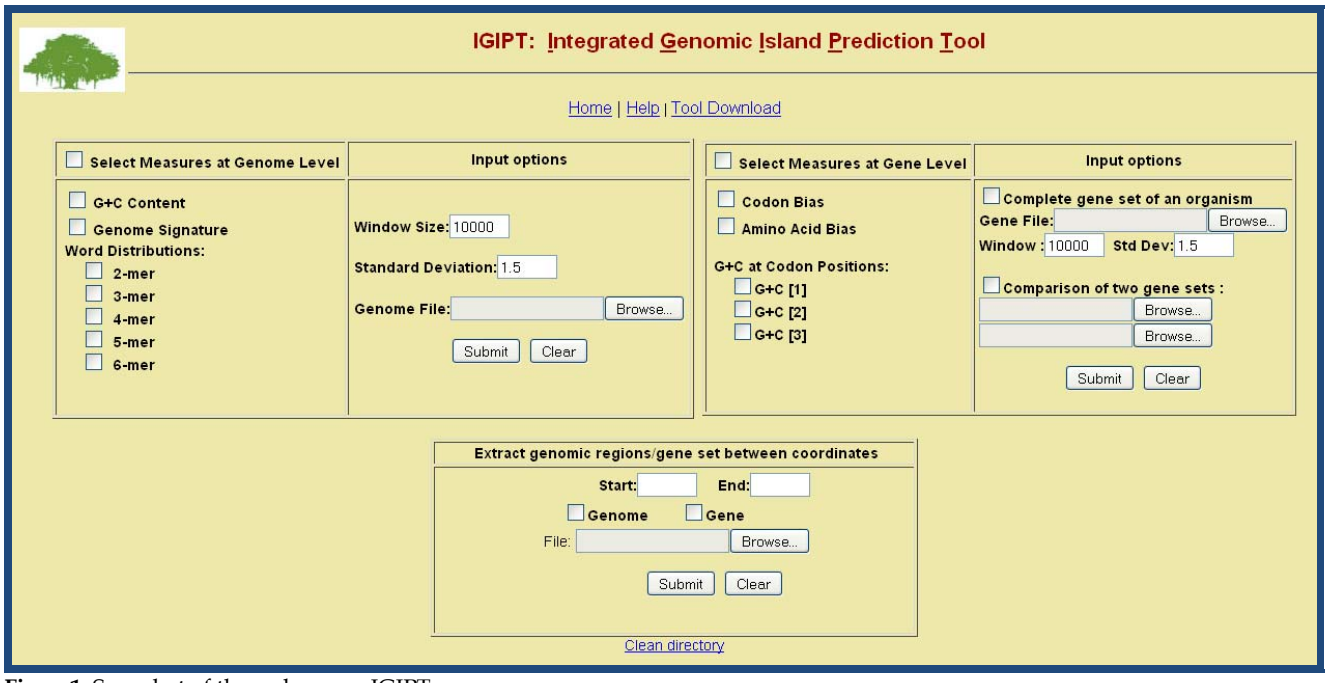

Figure1: Snapshot of the web-server IGIPT.

Conclusion:
Evolution of species by horizontal gene transfer is very common not only in prokaryotes but also in eukaryotes. It gives unique functionality to the organism to adapt to different environmental conditions and their identification is particularly useful in pathogens for identifying virulent genes. Since no single measure truly identifies a horizontally acquired region, by integrating numerous parametric measures on a single platform, IGIPT allows the users to analyze the predicted horizontally transferred regions/genes by thirteen different

measures simultaneously, thus greatly increasing the confidence of prediction. A drawback of these parametric methods is that regions acquired from donors with similar compositional bias as the host genome will not be identified.

References:

[1] Koonin EV *et al*. *Ann Rev Microbiol*. 2001 **55**:709 [PMID: 11544372].

[2] Langille MG *et al*. *Nat Rev Microbiol*. 2010 **8**:373 [PMID: 20395967].

[3] Lawrence J. *Curr Opin Genet Dev*. 1999 **9**:642 [PMID: 10607610].

[4] Gao F & Zhang CT, *Nucleic Acids Res*. 2006 **34**: W686 [PMID: PMC1538862].

[5] Karlin S & Mrazek J, *Proc Natl Acad Sci USA*. 1997 **94**:m10227 [PMID: 9294192].

[6] Pavlovic-Lazetic GM *et al*. *Comput Methods Programs Biomed* 2009 **93**: 241 [PMID: 19101056].

[7] Karlin S. *Trends Microbiol*. 2001 **9**:335 [PMID: 11435108].

[8] Yoon SH *et al*, *BMC Bioinformatics*. 2005 **6**: 184 [PMID: 16033657].

[9] Dobrindt U *et al*, *Nat Rev Microbiol*. 2004 **2**: 414 [PMID: 15100694].

Edited by P Kangueane

Citation: Jain *et al.* Bioinformation 7(6): 307-310 (2011)

Supplementary material:

Genomic signature
The dinucleotide biases assess difference between observed dinucleotide frequencies and that expected from random associations of component mononucleotide frequencies, defined as $\delta^*(W,G) = 1/16\sum|\rho^*_{xy}(W) - \rho^*_{xy}(G)|$, summed over all dinucleotide biases between sliding window (W) and whole genome value (G); $\rho^*_{xy} = f^*_{xy} / f^*_x f^*_y$; f^*_{xy} and f^*_x are frequencies of dinucleotide and mononucleotide, computed for the sequence concatenated with its inverted complement **[5]**.

k-mer Distributions
This module computes five measures based on the biases in average k-mer frequency of all possible words of size k (= 2 – 6) as $\delta_k(W,G) = 1/4^k \sum|f^k_i(W) - f^k(G)|$, $f^k_i(W)$ is the *k*-mer frequency in sliding window and $f^k(G)$ for the whole genome **[6].**

Codon Usage Bias:
The codon usage difference of gene family F relative to the genome or another gene set G is computed as $B(F|G) = \sum_a p_a(F) \sum_{(x,y,z)=a} |f(x,y,z) - g(x,y,z)|$; p_a(F) and f(x,y,z) are normalized amino acid and codon frequencies of gene family F and the sum extends over all synonymous codons **[7]**.

Amino Acid Bias
The amino acid bias between gene family F and genome (or second gene set) G is computed as $A(F|G) = (1/20)\sum_{i=1}^{20}|a_i(F) - a_i(G)|$, $a_i(F)$ being the average amino acid frequency of a_i in F **[7]**.

