Integrated Genomic Island Prediction Tool (IGIPT)

by

Ruchi Jain, Sandeep Ramineni, Nita Parekh

in


Report No: IIIT/TR/2008/175

Centre for Computational Natural Sciences and Bioinformatics
International Institute of Information Technology
Hyderabad - 500 032, INDIA
December 2008
Integrated Genomic Island Prediction Tool (IGIPT)

Ruchi Jain, Sandeep Ramineni and Nita Parekh
Centre for Computational Natural Sciences and Bioinformatics
ruchi_jain@research.iiit.ac.in, nita@iiit.ac.in

Abstract

We have developed a web based integrated platform for the identification of genomic islands in which various measures that capture bias in nucleotide compositions have been implemented, viz., GC content (both at the whole genome and at three codon positions in genes), genomic signature, k-mer distribution (k=2–6), codon usage bias and amino acid usage bias. The analysis carried out in sliding windows (default size 10Kb) is compared with the genomic average for each measure. The output is displayed in a tabular format for each window which may be filtered if the values of the measures differ by 1.5σ (standard deviations) from the genomic average.

Availability: http://ccnsb.iiit.ac.in/IGIPT/

1. Introduction

Genomic islands (GIs) are horizontally transferred regions, typically ~ 10 - 200 Kb in size. Any biological advantage provided to the recipient organism by the transferred DNA creates selective pressure for its retention in the host genome. Increasing evidence indicates that genomic islands are important in the evolution of bacteria, influencing traits such as antibiotic resistance, symbiosis and fitness, and adaptation in general.

Most widely used approaches for identifying recent horizontal transfers are based on the “genome hypothesis,” according to which codon usage and GC content are distinct signatures of each genome [1]. Here, we have implemented six such measures that capture anomaly in nucleotide composition on a single platform [2-4].

2. Materials and Methods

Various measures implemented in the IGIPT tool are briefly discussed below.

2.1 GC content anomalies: The frequency of G+C in a sliding window, W, (~10Kb) is compared to the average genomic G+C frequency [2]. Windows deviating significantly from the genomic average (~1.5σ) are identified as probable GIs. Similarly, GC content at the gene level is computed by computing the G+C frequency at the three codon positions, GC1, GC2, and GC3 [4]. In this case analysis on genes lying within window W is compared with the complete gene set of the organism.

2.2 Genomic signature: The dinucleotide bias of window W with respect to genome G is given by [2]

\[
\delta^* (W, G) = \frac{1}{16} \sum |\rho_{xy}^*(W) - \rho_{xy}^*(G)|
\]

where the sum extends over all nucleotides and the genome signature profile

\[
\rho_{xy}^* = \frac{f_{xy}^*}{f_x^* f_y^*},
\]

\[f_{xy}^*, f_x^*, f_y^*\] being the frequencies of dinucleotide XY and mononucleotide X, respectively, computed for the sequence concatenated with its inverted complement.

2.3 Word Distributions: Horizontally acquired regions have distinct word (k-mer) compositions and thus can be used for identifying probable GIs [3]. The frequency of all possible words of size k (= 2 – 6) are computed in sliding windows, W, and also for the whole genome. The difference in the average k-mer frequency is computed as

\[
\delta_k^* (W, G) = \frac{1}{4^k} \sum |\rho_{xy}^*(W) - \rho_{xy}^*(G)|
\]

where the sum extends over all codons translated to amino acid a. The codon usage difference of the gene

\[
\sum_{x,y,z=a} f(x,y,z) = 1
\]

and the sum extends over all codons translated to amino acid a. The codon usage difference of the gene

Table 1: Analysis of A. aurescens genome using IGIPT
2.5 Amino Acid Bias: This refers to the deviation in the frequency of usage of individual amino acids over the average usage of all 20 amino acids [2]. The amino acid bias between gene family \( F \) and the genome \( G \) is given by

\[
B(F \mid G) = \sum_{a} p_{a}(F) \left[ \sum_{(x,y,z)} f(x,y,z) - g(x,y,z) \right]
\]

where \( p_{a}(F) \) are the normalized amino acid frequencies of the gene family \( F \) [2].

3. Results and Discussion

A set of prokaryotic genomes and their gene sequences were downloaded from NCBI (ftp://ftp.ncbi.nih.gov/genomes/Bacteria) and analyzed using IGIPT. Here we present our results for *Arthrobacter aurescens* TC1 genome, in which 11 genes have been reported with atypical composition. The islands include transposons and related genes, transcriptional regulators, resistance genes, and genes involved in metabolism and transport of a wide range of substrates. Table 1 summarizes the output from IGIPT.

<table>
<thead>
<tr>
<th>Experimentally Verified GIs (Kb)</th>
<th>Size (Kb)</th>
<th>Measures at Genome Level</th>
<th>Measures at Gene Level</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>GC</td>
<td>GS</td>
</tr>
<tr>
<td>14.5</td>
<td>16.9</td>
<td>24</td>
<td>P</td>
</tr>
<tr>
<td>95.9</td>
<td>97.4</td>
<td>15</td>
<td>N</td>
</tr>
<tr>
<td>123.2</td>
<td>127.1</td>
<td>39</td>
<td>P</td>
</tr>
<tr>
<td>146.5</td>
<td>150.0</td>
<td>35</td>
<td>P</td>
</tr>
<tr>
<td>179.1</td>
<td>180.7</td>
<td>16</td>
<td>N</td>
</tr>
<tr>
<td>338.6</td>
<td>339.8</td>
<td>12</td>
<td>N</td>
</tr>
<tr>
<td>341.2</td>
<td>342.3</td>
<td>11</td>
<td>P</td>
</tr>
<tr>
<td>345.9</td>
<td>347.9</td>
<td>20</td>
<td>P</td>
</tr>
<tr>
<td>385.0</td>
<td>386.2</td>
<td>12</td>
<td>P</td>
</tr>
<tr>
<td>410.3</td>
<td>412.7</td>
<td>24</td>
<td>P</td>
</tr>
<tr>
<td>452.0</td>
<td>453.2</td>
<td>12</td>
<td>N</td>
</tr>
</tbody>
</table>

on *A. aurescens* genome. The results are shown only for windows differing by 1.5\(\sigma\) from the genomic average; here \(P\) – represents identified by a particular measure, and \(N\) – not identified. Regions acquired from donors with similar compositional bias as the host genome will not be identified by these measures. Since no single measure can truly identify a horizontally acquired region, we suggest users to confirm with two or more measures. The implementation of six different measures on a single platform greatly increases the predictability power of IGIPT.

4. References