An Anomaly Detection Algorithm for Identifying Alien Gene Clusters in Microbial Genomes

Dongsheng Che¹, Sai Vahini Manikonda¹, Zuqing Li¹, and Bernard Chen²
¹Department of Computer Science, East Stroudsburg University, East Stroudsburg, PA 18301, USA
²Computer Science Department, University of Central Arkansas, Conway, AR 72034, USA

Abstract—Genetic materials are often transferred between microbial organisms. The host genome regions that contain alien genetic materials may carry integration-related genes and products in some cases, but not all the time. Therefore, computational methods that use such information for detecting alien gene clusters in genomes have some limitation. The genome composition based approach have been applied using oligonucleotides and codon usage. In this paper, we report the development of our parametric genome composition-based anomaly detection algorithm for alien gene cluster finding, using the parameters of genome block size and k-mer word. We tested our algorithm on five genomic sequences. Our prediction results have shown that our algorithm can accurately detect alien gene clusters in these genomes.

Keywords: Alien gene cluster; Anomaly detection; Archaea; Bacteria; Gaussian distribution; Genome sequence

1. Introduction

An outlier is the observation that is significantly different than the majority of the observations, and there are many applications for outlier detection, such as fraud detection, and intrusion detection [1]. In bacterial and archaeal genomes, outlying genome segments are often seen. In this case, DNA patterns in the outlier regions are distinct from the remainder of the host genomes. Typically, two lineage distant organisms should have different genome patterns since each organism genome pattern is unique. Therefore, a genome with outlying regions in one organism could be caused by the gene inflow from other organisms, which is also known as horizontal gene transfer (HGT).

The gene inflow from the alien organism to the host organism can be carried out through: transformation [2], conjugation [3], transduction [4], and gene transfer agent [5]. During the integration of the alien genome sequences into the host genomes, integrase, transposon genes might participate in the integrating process in some cases [6]. The alien genome regions are sometimes also flanked by transfer RNAs [7].

The identification of alien gene clusters in the host genome is extremely important for biological communities. It helps evolutionary biologists understand how bacterial genomes are evolving in organisms’ evolutionary history. Such findings could also be used for studying donor-recipient relationship through gene cluster transfers, and for constructing donor-recipient networks across genomes [8]. In other cases, where alien gene clusters contain pathogenic genes, these can be used for explaining why some non-pathogenic bacterial genomes were found to cause diseases [9].

Alien gene clusters were initially found in Escherichia coli [10], Later on, more biological experiments and evolutionary studies showed the existence in other bacatical and arachaea genomes [11]–[14]. With the explosive growth of completed genome sequences, computational tools have been developed for predicting alien gene clusters. We can roughly categorize computational tools into three groups: sequence composition based, comparative genome analysis based, and ensemble based.

The sequence composition based approaches find the alien gene clusters based on the DNA signatures (such as G + C content, dinucleotide pattern), codon usage patterns, or integration associated genes (such as transposon genes and integrase genes, tRNA genes) in the query genome. The tools using sequence and gene information are AlienHunter [15], Centroid [16], GI Detector [17], GI Hunter [18], PAI-ID A [19], and SIGI-HMM [20].

For the comparative genome analysis based approach, the identification of alien gene clusters rely on the reference genomes, which are lineage closely related to the query genome. The gene clusters, that only exist in the query genomes but not in its reference genomes, are considered to be alien gene clusters [6]. Representative tools for this group of approaches include IslandPick [21] and MobileGeneFINDER [22]. Ensemble-based approaches integrate the prediction results of standalone tools to obtain census GIs. For instance EGIS [23] integrates the prediction of AlienHunter [15], SIGI-HMM [20], INDeGenIUS [24], IslandPath [25], and PAI-ID A [19].

Despite the development of computational tools for detecting alien gene clusters, none of them can accurately predicts for all genomes. Some prediction tools have high sensitivity but with low specificity such as AlienHunter [15], while other have high specificity but with low sensitivity such as IslandPath [25]. Some other tools that incorporates mobile gene, tRNA gene or virus gene information in the models also suffer the prediction accuracy, due to the fact that all genomes having alien gene clusters have such characteristics.
[26]. Therefore, there is still room for improvement in the
detection of alien gene clusters.

In this study, we will not consider genomic specific feature
information such as mobile genes, tRNAs genes, or virus
related genes for alien gene cluster prediction. Instead, we
will focus on the DNA sequence only to make our prediction
tool applicable to all genomes. Specifically, we will first
build DNA based features that rely on a parametric block
size (B) of the genome sequence. The DNA features are
built through counting the frequencies of k-mer nucleotides.
The anomaly detection algorithm will be implemented by
evaluating all combined feature values based on Gaussian
distribution model. The performance of the algorithm
will be evaluated by applying our algorithm on completely
sequenced genome datasets.

The remainder of this paper is organized as follows.
Section 2 describes the anomaly detection algorithm. Section
3 shows the prediction results of our detection algorithm on
five genomes. In Section 4, we will conclude the paper, with
the discussion of future work.

2. Methods

We detect outlying genomic regions through analyzing
each block of genomic segments. The block size (denoted
as B) can range from 5 kb, 10 kb, 20 kb and 50 kb. The
starting positions of each block are the multiple of B/2.
For instance, for the block size of 20 kb, we look into the
regions of (0-20), (10-30), (20-40) kb, etc. For each block
of genomic region, we extract the features as described below:

2.1 Feature Construction

A vector of the features for any block of genomic region
is constructed based on the k-mer words, where k can be 4,
5, or 6. Since there are four possible nucleotides (A, C, G and
T) for any position of the k-mer word, the total number of
possible words should be 4^k. For instance, for k = 5, then
there are 4^5 = 1024 different possible words, and thus the
feature vector size is 1024. Since our block size of genomic
region to be investigated are between 5kb to 50k, we choose
the our k be 4, 5, and 6. If k is too large, then the frequencies
of many of k-mer will be zeros, and thus making feature
values indistinguishable in the dataset.

Let i be the i^{th} block of genomic region, and m be the
total number of blocks in the entire genome sequence. Let
x_j^{(i)} be the frequency of the j^{th} word in the i^{th} block
of genomic region, then a vector of feature values x^{(i)} in the
i^{th} block can be represented as follows:

\[ x^{(i)} = (x_1^{(i)}, x_2^{(i)}, x_3^{(i)}, \ldots, x_{4^k-1}^{(i)}, x_{4^k}^{(i)}) \]

(1)

All m feature vectors (x^{(1)}, x^{(2)},... and x^{(m)}) are computed,
and they are used to be fed into the anomaly detection
algorithm.

2.2 Anomaly Detection Algorithm

The outliers can be detected based on Gaussian distribution
model. If a data point is far from the mean by more
than a couple of times the standard deviation, then it can be
treated as an outlier. For the problem of detecting outlying
genomic region, we compute a vector of 4^k features of
means and standard deviation as follows:

\[ \mu_j = \frac{1}{m} \sum_{i=1}^{m} x_j^{(i)} \]  

(2)

\[ \sigma_j^2 = \frac{1}{m} \sum_{i=1}^{m} (x_j^{(i)} - \mu_j)^2 \]  

(3)

The probability of a vector of 4^k features is the simply the
product of all individual probabilities, which are calculated
based on Gaussian distribution. The product of 4^k
probabilities will be very small, and could lead to the problem of
underflow error. To deal with this kind of problem, we could
do log transformation. So the log of the product becomes the
sum of logs. The probability for estimating the i^{th} block of
genomic region to be outlying or not can be given as follows:

\[ \log P(x^{(i)}) = \sum_{i=1}^{4^k} \log \left( \frac{1}{\sqrt{2\pi\sigma_j}} \exp \left( -\frac{(x_j^{(i)} - \mu_j)^2}{2\sigma_j^2} \right) \right) \]  

(4)

The i^{th} block of region is considered to be an outlying
regions if LogP(x^{(i)}) is less than a predefined threshold
value. The threshold value will be different for a different
k-mer word. The pseudocode of the anomaly detection
algorithm is given below:

**Algorithm 1 Anomaly Detection Algorithm**

1: m = the total number of blocks of genomic regions
2: for i ← 1, m do
3:  Compute the i^{th} block’s feature vector x^{(i)}
4:  \( x^{(i)} = (x_1^{(i)}, x_2^{(i)}, x_3^{(i)}, \ldots, x_{4^k-1}^{(i)}, x_{4^k}^{(i)}) \)
5:  \end for
6: for j ← 1, 4^k do
7:  \( \mu_j = \frac{1}{m} \sum_{i=1}^{m} x_j^{(i)} \)
8:  \( \sigma_j^2 = \frac{1}{m} \sum_{i=1}^{m} (x_j^{(i)} - \mu_j)^2 \)
9:  \end for
10: for i ← 1, m do
11:  \( \log P(x^{(i)}) = \sum_{i=1}^{4^k} \log \left( \frac{1}{\sqrt{2\pi\sigma_j}} \exp \left( -\frac{(x_j^{(i)} - \mu_j)^2}{2\sigma_j^2} \right) \right) \)
12: if LogP(x^{(i)}) < threshold then
13:  Report i^{th} block to be an outlying genomic region
14: \end if
15: \end for
3. Results

We have implemented our parametric anomaly detection algorithm for alien gene cluster finding, and tested the algorithm on five genomes. The genome sequences were collected from the National Center for Biotechnology Information (NCBI) FTP server (ftp://ftp.ncbi.nih.gov/genomes/Bacteria). These five genomes were Corynebacterium glutamicum (C. glutamicum) ATCC 13032, Corynebacterium diphtheriae (C. diphtheriae) NCTC 13129 chromosome, Helicobacter pylori (H. pylori) strain J99, Rhodopseudomonas palustris (R. palustris) CGA009, Vibrio vulnificus (V. vulnificus) CMCP6 chromosome I. The genome information is shown in Table 1.

We run our algorithm using the following nine parameter combinations, (5K, 4-mer), (10K, 4-mer), (20K, 4-mer), (50K, 4-mer), (10K, 5-mer), (20K, 5-mer), (50K, 5-mer), (20K, 6-mer) and (50K, 6-mer). Below we present the predicted alien gene clusters, and compare the predicted results with the results using other methods.

3.1 C. glutamicum ATCC 13032

C. glutamicum ATCC 13032 is bacterium that has a genome length of 3.28 Mb, and it contains 3002 protein-coding genes. The species is important for industrial production of amino acids [32], and its genome contains several genome regions that were acquired from C. diphtheriae through horizontal gene transfer.

The prediction results of our algorithm with nine combinations of two parameters are shown in Figure 1. In general, the smaller block size and the smaller k-mer word, the more predicted outlying regions. On the other hand, the larger block size and the larger k-mer word, the less predicted outlying regions. For instance, the combination of (5K, 4-mer) in Figure 1(a) shows many valleys, while the combination of (50K, 6-mer) in Figure 1(i) shows only one significant valley. This significant outlying region of 1.75-2.00 Mb covers 236 genes, which was consistent with previous study of 1.78-1.99 Mb. The predicted region is mainly composed of hypothetical proteins, but it also contains one integrase gene, one phage related genes, seven tRNAs genes, and two transposases [26].

For the combination of (20K, 6-mer), as shown in Figure 1(h), our algorithm predicted four significant regions with the probability value less than -4600, i.e., 0.37-0.39, 1.78-1.94, 2.70-2.72, 3.16-3.18. The genomic region of 0.37-0.39 contains a number of proteins related to lipopolysaccharide synthesis sugar transferase, glycosyltransferase, which explains its impact on the production of L-aspartate-derived amino acids and vitamins.

3.2 C. diphtheriae NCTC 13129 chromosome

C. diphtheria NCTC13129 is a gram-positive bacterium that produces diphtheria toxin (DT), which causes the symptoms of diphtheria [28]. The genome sequence has the size of 2.49 Mb, and it contains 2320 genes.

With the combination of (50K, 6-mer), we predicted four genomic regions that have the probability values less than -5200, i.e., 0.15-0.20, 0.25-0.30, 0.35-0.40 and 1.10-1.40 (Figure 2(a)). The region of 1.12-1.52 contains 351 genes, including nine tRNA genes, and four putative transposase genes. this predicted region was consistent with other prediction tools such GIHunter [26] and IslandViewer [33].

3.3 H. pylori strain J99

H. pylori is one of the most common human pathogens that colonizes the gastric mucosa [29]. The genome is 1.65 Mb long, and it was reported to contain the type IV secretion system for virulence of pathogens, and encoded on the cag pathogenicity island, which is acquired by horizontal transfer [34].

For the combination of (50K, 5-mer), we predicted two outlying regions that have the probability values less than -1800, i.e., 0.50-0.55, and 1.00-1.10 (Figure 2(b)). The region of 0.50-0.55 Mb contains 49 genes, with the majority of encoding cag island proteins. The region of 1.00-1.10 Mb contains 83 genes, with the majority of them encoding putative proteins, and with several integrases.

3.4 R. palustris CGA009

R. palustris CGA009 is a metabolically versatile photosynthetic bacterium that produces energy by using light and inorganic and organic compounds [30]. The genome is 5.46 Mb long, and it has 4,836 predicted genes. The

<table>
<thead>
<tr>
<th>Genome</th>
<th>Length (Mb)</th>
<th>Accession Number</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corynebacterium glutamicum ATCC 13032</td>
<td>3.28</td>
<td>NC_003450.3</td>
<td>[27]</td>
</tr>
<tr>
<td>Corynebacterium diphtheriae NCTC 13129 chromosome</td>
<td>2.50</td>
<td>NC_002935</td>
<td>[28]</td>
</tr>
<tr>
<td>H. pylori strain J99</td>
<td>1.65</td>
<td>NC_000921</td>
<td>[29]</td>
</tr>
<tr>
<td>Rhodopseudomonas palustris CGA009</td>
<td>5.46</td>
<td>NC_005296.1</td>
<td>[30]</td>
</tr>
<tr>
<td>Vibrio vulnificus CMCP6 chromosome I</td>
<td>3.25</td>
<td>NC_004459.3</td>
<td>[31]</td>
</tr>
</tbody>
</table>

Table 1: A list of genome datasets tested in this study
G + C content did not show apparent horizontal transferr islands, but Z-curved based method was able to identify three horizontal islands [35].

With the combination of (50\(K\), 6-mer), we predicted three genomic regions that have the probability values less than -5000, i.e., 2.40-2.60, 3.70-3.80 and 4.55-4.70 (Figure 2(c)). The region of 2.40-2.60 contains 160 genes, including type IV secretion system subunit and conjugal transfer proteins. The region of 4.55-4.70 Mb contains multidrug-efflux transport proteins, conjugal transfer proteins, and other integration supporting integrases and tRNAs. All these three predicted regions were consistent with the reported regions with Z-curve approach [35].

3.5 V. vulnificus CMCP6 chromosome I

V. vulnificus CMCP6 is a pathogenic bacterium with a genome sequence length of 3.28 Mb [36]. Previous studies have shown that this genome has three anomaly gene clusters, VVGI-1 (2,438,377-2,605,507bp), VVGI-2 (355,728-395,914bp) and VVGI-3 (3,248,897-3,281,945bp) [37].

With the combination of (50\(K\), 6-mer), we predicted four genomic regions that have the probability values less than -5500, i.e., 0.30-0.40, 0.75-0.80, 2.45-2.60 and 3.23-3.28 (Figure 2(d)). The region of 0.30-0.40 contains 93 genes, most of them are hypothetical proteins, but it also contains type IV secretory proteins, and phage integrase. The region of 2.45-2.60 contains 131 genes, where it contains nine transposase, integrase and prophage antirepressor.
4. Conclusions and Discussion

In this paper, we have implemented our genome anomaly detection algorithm for detecting alien gene clusters in bacterial and archaeal genomes. The algorithm is parameterized by the block size $B$ and $k$-mer word. Our experiments on five genome sequences have shown that the block size of 50 $K$, and 6-mer word (5-mer for small genome size) have better prediction power than smaller block sizes and smaller $k$-mer word. Most of our detected anomalous gene clusters are consistent with previous studies, which showed horizontal gene transfer evidence through case by case studies.

While the genome sequence based detection algorithm can be applied to any microbial genome, it remains to be a challenging task for the following scenarios: (1) The genome sequences of the donor and recipient species could be similar, thus, making it difficult to detect alien gene clusters in the core genome. (2) The alien genome sequence in the host genome can be ameliorated, a process that makes the sequence composition (or codon usage) of the alien genomic region be similar to that of the core genome. A recent large scale genomic study of bacterial and archaeal genomes have supported the existence of amelioration. Again, the anomaly detection algorithm is incapable of detecting such ameliorated regions. (3) Not all biased genome sequences are horizontally transferred. Instead, they could be essential to host genomes and happen to be biased. Highly expressed genes, such as ribosomal related genes, chaperonin genes, transcription and termination factor genes, energy
Table 2: Alien gene clusters detected by our algorithm with supporting information

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Cut-off</th>
<th>Predicted Region (Mb)</th>
<th>Genes Covered</th>
<th>Reported Region (Mb)</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>C. glutamicum</em> ATCC 13032</td>
<td>(50k, 6 mer)</td>
<td>-6000</td>
<td>1.75-2.00</td>
<td>236</td>
<td>1.78-1.99 [37]</td>
</tr>
<tr>
<td><em>C. diphtheriae</em> NCTC 13129 chromosome</td>
<td>(50k, 6 mer)</td>
<td>-5300</td>
<td>0.15-0.20</td>
<td>57</td>
<td>0.15-0.19 [28]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.25-0.30</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.35-0.40</td>
<td>53</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.12-1.52</td>
<td>351</td>
</tr>
<tr>
<td><em>H. pylori</em> strain J99</td>
<td>(50k, 5 mer)</td>
<td>-1800</td>
<td>0.50-0.55</td>
<td>49</td>
<td>0.50-0.54 [34]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.00 - 1.10</td>
<td>83</td>
</tr>
<tr>
<td><em>R. palustris</em> CGA009</td>
<td>(50k, 6 mer)</td>
<td>-5000</td>
<td>2.40-2.60</td>
<td>160</td>
<td>2.48-2.57 [35]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3.70-3.80</td>
<td>78</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4.55-4.70</td>
<td>177</td>
</tr>
<tr>
<td><em>V. vulnificus</em> CMCP6 chromosome I</td>
<td>(50k, 6 mer)</td>
<td>-5500</td>
<td>0.30-0.40</td>
<td>93</td>
<td>0.35-0.40 [37]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.75-0.80</td>
<td>55</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.45-2.60</td>
<td>131</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3.23-3.28</td>
<td>46</td>
</tr>
</tbody>
</table>

metabolism genes, recombination and repair genes, and electron transport genes, have the characteristics of sequence bias. In this scenario, the detected anomalous regions are not horizontally transferred. Given that, additional information might be incorporated into the algorithm to handle these scenarios.

Our experiments have shown that large block size (e.g., 50k) and 6-mer (or 5-mer) word have fairly good detection power. Such settings could be adjusted slightly for detecting any other genomes, depending on the genome sequence size. The threshold values for selecting alien gene clusters in the genomes is hard to determine since different genomes have various alien genome percentage. It has been reported that some genomes could have up to half of genome sequences transferred from other organisms. One of possible solution to that is to use supervised learning approach to obtain threshold value for new genomes.

References


