Identification of compositionally distinct regions in genomes using the centroid method

Issaac Rajan¹, Sarang Aravamuthan² and Sharmila S. Mande¹,*
¹Life Sciences Research and ²e-Security R&D, Advanced Technology Centre, Tata Consultancy Services, Hyderabad 500 081, Andhra Pradesh, India

Received on January 23, 2007; revised on July 16, 2007; accepted on August 6, 2007

Advance Access publication August 27, 2007

Associate Editor: Burkhard Rost

ABSTRACT

Motivation: It is known that most genomic regions of special interest, e.g. horizontally acquired sequences, genomic islands, etc. have distinct word (m-mer) compositions. Most of the earlier work along this direction, addressed di- and tri-nucleotide compositions. We present an approach that can be applied to analyze compositions of any given word size. The method, called the centroid approach, can reveal compositionally distinct regions in genomic sequences for any given word size.

Results: We applied our method to 50 bacterial genomes and demonstrated its ability to identify embedded sequences of varying lengths from distantly related organisms. We also investigated the genetic makeup of the regions identified as compositionally distinct by our method, for four organisms from our dataset. Pathogenicity island (PAI) components and genes encoding strain-specific proteins are all frequently seen to be constituents of these regions.

Availability: Program is available on request from the authors.

Contact: sharmila@tcs.com

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Following their discovery in uropathogenic Escherichia coli, pathogenicity islands (PAIs) (Hacker et al., 1990) have been identified and intensely studied in other bacterial genomes. Subsequently, other large segments, similar to PAIs, within prokaryotic genomes were observed encoding various specialized functions. Examples of these functions include secondary metabolism (metabolic islands), antibiotic resistance (resistance islands) and secretion (secretion islands). These genomic substructures are referred to as ‘genomic islands’. Islands often possess transposons, phage sequences and clusters of genes which perform related functions or participate in related pathways. In general, islands are flanked by direct/inverted repeats and have tRNA or tmRNA in their proximity. There is substantial evidence to suggest their acquisition through horizontal origin (Blum et al., 1994; Sullivan and Ronson, 1998).

Several groups have used annotation-based features to identify genomic islands. For instance, tRNA and tmRNA were used as initial leads to identify genomic islands (Mantri and Williams, 2004). Similar feature-based approaches include efforts by Ou et al. (2006) and Nag et al. (2006). Although these approaches are sound, an obvious constraint is set by the availability of well-annotated genomic sequences. An incorrect or less rigorous annotation can severely hamper the outcome of these methods. Another limitation is that since these methods seek certain biological features, islands devoid of these features are likely to be overlooked. Methods that use a more intrinsic attribute (such as genome composition) are devoid of these limitations. These methods are based on the hypothesis that genomic islands possess distinct composition, as compared to the rest of the genome. Karlin (2001) proposed several strategies based on the compositional aspects of the genome for the identification of anomalous gene clusters and PAIs in diverse bacterial genomes. Zhang et al. (2001) proposed a windowless method for the GC content computation, termed as the cumulative GC profile and applied it for the identification of genomic islands in Corynebacterium glutamicum and Vibrio vulnificus CMCP6 chromosome I (Zhang and Zhang, 2004). Tu and Ding (2003) used iterative discriminant analysis to define genomic regions that deviate the most from the rest of the genome based on three compositional criteria, namely, G+C content, dinucleotide frequency and codon usage. Besides, these successes in analyzing genomes using words of size 2 or 3, it is generally acknowledged that larger word sizes (5–9) characterize the genomic regions that deviate the most from the rest of the genome.

In this article, we present an approach (called the centroid method) that enables identification of compositionally distinct regions in genomes for any word size. We also show, through examples, that this method is able to identify embedded foreign sequences in genomes. Finally, we analyze the DNA content of the genome composition outlier bins for four of the organisms from our dataset and comment on the biological nature of the centroid-defined ‘alien DNA’.

2 METHODS

2.1 The centroid method

In the centroid method, we first partition the genome sequence into non-overlapping bins of equal length and associate an n-tuple with each bin. Here, n is the number of distinct m-words (words of length m). For
a given word, there are four possible symbols \{A, G, C and T\} for each letter of the word. Hence for a given word of size \(m\) letters, the base distribution frequency of a genomic fragment can be represented in an \(n\)-dimensional space, where \(n = 4^m\). For example, for a word of size 5 letters, \(n = 4^5 = 1024\). These vectors are viewed as points in an \(n\)-dimensional space. We then determine the centroid of these points. The distance from the centroid is used as the criterion for determining the outliers among these points. The outliers correspond to the compositionally distinct bins.

The steps in the centroid method are given below:

1. The genome of interest is partitioned into non-overlapping bins of equal size.
2. The frequencies of all possible words for a given word size are enumerated corresponding to each bin, considering words in both the DNA strands. This list is the word frequency vector for the bin.
3. The average frequency of each word across all bins is computed. The vector of these averages is the centroid.
4. For each bin, the distance between its word frequency vector and the centroid is computed (see below).
5. Based on the distribution of distances of the bins from the centroid, a suitable outlier selection criterion is defined in order to identify outliers among the bins.
6. Steps 1–5 are repeated for varying offsets from the start position of the genome. While doing so, the regions identified as compositionally distinct for all the different offsets should be combined.

The bins identified as outliers may be subject to further investigation for the biological context they define. Steps 1–4 are schematically depicted in Figure 1.

We applied this algorithm to 50 diverse bacterial genomes. The genome sequences were obtained from the NCBI GenBank (http://www.ncbi.nlm.nih.gov). We used the Manhattan distance metric \(L_1\) norm for distance calculations. We chose to use the Manhattan distance since it has been established that the Manhattan distance is preferable over the Euclidean distance \(L_2\) norm) for high-dimensional data (Aggarwal et al., 2001). In all the genomes studied, we considered the bin size to be 10 kb, the word size to be 5 (so that \(n = 1024\)) and an offset of 2 kb (which resulted in five scans of a given genome in our case).

It may be noted that the distance of a bin from the centroid reflects the extent of its word compositional distinction with reference to the word-frequencies profile defined by the centroid. That is, the greater the distance of the word frequency vector of a bin to the centroid, the more different the bin’s word composition is from the ‘average’ genomic bin (as defined by the centroid).

### 2.2 Outlier detection

Bins which were away from the mean by more than three times the SD were considered as outlier bins.

The efficacy of the centroid method was tested in terms of its ability to identify embedded sequences of foreign origin.

### 2.3 Identification of embedded sequences from distantly related organisms

In order to test the success rate of the centroid method in identifying embedded sequences of foreign origin, we inserted sequences from other organisms at a single as well as multiple locations in the genomes. In all the genome chimeras constructed, the implanted portion replaces the existing portion of the host genome sequence. For each of the genome chimeras, we constructed in-frame and out-of-frame chimeras.

In-frame constructs of chimeras are those where the beginning and the end of the inserted portion is in sync with the bin partitions, while out-of-frame constructs are those in which beginning and the end of the inserted portions are not in sync with the bin partitions.

Genome chimeras with single insert were constructed by replacing 20 kb fragment of the genome by a 20-kb fragment from *Brucella melitensis* 16M chromosome I. Genome chimeras with multiple inserts were constructed by embedding segments of varying lengths from three different bacteria, namely, *Acanthoaxor borkumensis* (10 kb insert), *B.melitensis* 16M chromosome I (20 kb insert) and *Azoarcus* sp. EbN1 (40 kb insert). A schematic representation of the various chimeras constructs is provided in Figure 2. The choice of *B.melitensis* 16M, *Azoarcus* sp. EbN1 and *A.borkumensis* as the sources for the implanted sequence is based on the fact that on the bacterial phylogenetic tree, these organisms belong to distinct branches of the alpha-, beta- and gamma-Proteobacteria, respectively.

Sensitivity in identifying the inserted foreign sequence is scored as the percentage of the number of bases of the insert identified with respect to the length of the inserted sequence. Using the sensitivity values thus obtained for each dataset, we computed the mean sensitivity.

### 2.4 Identification of an embedded sequence from a closely related organism

In order to test whether our method is able to identify an embedded sequence coming from a closely related organism, we constructed genomic chimeras, wherein we implanted a 20-kb insert from a closely related genome (donor organism) between positions 50 001 and 70 000 in the recipient genome. The chimeras (Recipient–Donor) constructed were: \((E.coli\; CFT073–E.coli\; K12), (Heliobacter\; pylori\; 26695–H.pylori\; 399), (Mycobacterium\; tuberculosis\; H37Rv–Mycobacterium\; leprae)\) and \((Yersinia\; pestis\; CO92–Yersinia\; pseudotuberculosis\; IP32953)\).

### 2.5 Benchmarking

Benchmarking of the centroid method was done against methods based on %G+C and genomic signature (Karlin, 2001). These methods were applied considering the same initial partitioning of the genome into non-overlapping bins. While evaluating these other approaches, we used the same statistical criterion for outlier detection, as was done in the case of the centroid method. Identical means of measuring sensitivity was followed for the identification of implanted portions in the genome chimeras. In order to test whether our method is able to identify an embedded sequence coming from a closely related organism, we constructed genomic chimeras, wherein we implanted a 20-kb insert from a closely related genome (donor organism) between positions 50 001 and 70 000 in the recipient genome. The chimeras (Recipient–Donor) constructed were: \((E.coli\; CFT073–E.coli\; K12), (Heliobacter\; pylori\; 26695–H.pylori\; 399), (Mycobacterium\; tuberculosis\; H37Rv–Mycobacterium\; leprae)\) and \((Yersinia\; pestis\; CO92–Yersinia\; pseudotuberculosis\; IP32953)\).

### 2.6 Effect of insert length

In order to test the impact of the insert length on the sensitivity of the centroid method, we constructed genome chimeras having a single inserted sequence fragment from *B.melitensis* 16M chromosome I of varying lengths \((5, 10, 20, 40, 80, 160\) and \(320\) kb). In order to represent a more realistic situation, these chimeras were constructed by implanting the inserted fragment in the out-of-frame configuration. Using each dataset of 50 organisms, comprising of insert of a particular length, centroid method’s sensitivity was computed.

### 3 RESULTS

#### 3.1 Detection of embedded sequences from distantly related organisms

We tested the ability of the centroid method to identify the bins containing the embedded foreign sequence in a given genome (Supplementary Table A, B). For the single insert data, both in the cases of in-frame as well as out-of-frame inserts, centroid method consistently performed better than the other
two methods, namely, %G+C and genomic signature approaches (Table 1). For example, in the case of single insert out-of-frame data, mean sensitivity calculated using centroid method was higher (89.8), compared to that calculated using %G+C (57.5) and the genomic signature (78.4) approaches.

In the case of multiple insert data, it was observed that the centroid method performed distinctly well for higher insert sizes (Table 2). This observation was true both for data obtained from in-frame as well as out-of-frame multiple inserts. For instance, we observed in the case of out-of-frame multiple insert data, centroid method detected 75 and 84% for inserts of sizes 20 and 40 kb, respectively. Calculation based on %G+C variation resulted in detection of 42 and 54%, for the same inserts of lengths 20 and 40 kb, respectively. Similarly, Karlin's genomic signature-based approach detection was 53 and 62%, in inserts of varying lengths using methods based on centroid, %G+C and genomic signature for all the 50 organisms is given in Supplementary Table B.

### 3.2 Detection of embedded sequence from a closely related organism

The four chimeras constructed were tested against the centroid method as well as the benchmarks. It was seen that only in the case of *M.tuberculosis* H37Rv chimera, the inserted portion was detected fully both by the centroid method as well as %G+C.

### 3.3 Effect of insert length

Table 3 provides data on the effect of insert length on the mean sensitivity of the centroid method. From the data it is clear that the mean sensitivity value is maximum for inserted fragment lengths in the range of 10–40 kb. The mean sensitivity is poor for shorter inserts. The mean sensitivity falls rapidly when the insert length is 80 kb and above. This is, as one would expect, due to the fact that when the inserted portion is significantly large in context of the size of the recipient genome, this portion contributes significantly to the average trend and is hence not easily detected as an outlier. Effect of insert size of the embedded sequences on the sensitivity of the centroid method for all the 50 organisms is given in Supplementary Table C.

### 3.4 Gene contents in the outlier bins

3.4.1 Overall aspects  
We analyzed the gene contents in the outlier bins identified by centroid method for four organisms using the annotations in the.ptt file of NCBI GenBank to gain some biological insights about the nature of compositionally distinct regions in genomes. The centroid analysis was performed using conditions mentioned in Sections 2.1 and 2.2. It identified 4.6, 2.1, 4.3 and 3.7% of the genome as compositionally distinct, in *H.pylori* 26695, *Mycoplasma pulmonis* UAB CTIP, *M.tuberculosis* H37Rv and *Y.pseudotuberculosis* IP32953, respectively. Here, we provide an outline about the general nature of the compositionally distinct bins for these organisms. Supplementary Table D provides a complete account of the genes present in each of the bins identified by our method.

3.4.2 *Helicobacter pylori* 26695  
The *cag* PAI in *H.pylori* 26695 is associated with the reduced expression of interleukin-4 (IL-4) mRNA in human gastric mucosa (Orsini et al., 2003). We observed a cluster of 20 *cag* genes (cag3-22) coding for the *cag* PAI proteins and the virB11 homolog present within the *cag* PAI region in *H.pylori*. Earlier shown to have a distinct dinucleotide composition (using genomic signature difference profile) and codon bias (Karlin, 2001). Several predicted coding regions with unknown
functions could be seen in the following regions of the genome: 456001–466000, 470001–484000, 1048001–1072000. These regions were identified by our method as compositionally distinct. Such regions will be interesting to investigate in terms of the functions they code for.

3.4.3 Mycoplasma pulmonis UAB CTIP

*M. pulmonis* is the etiologic agent of murine respiratory mycoplasmosis (Maureen *et al.*, 1988). It has been suggested that the *vsa* (variable surface antigen) genes encoding the V-1 family of surface proteins may contribute to the host specificity of the mycoplasma as well as chronicity and the severity of disease (Shen *et al.*, 2000). These *vsa* genes are known to differ between different strains of *M. pulmonis*, although the rest of the genome is mostly conserved (Shen *et al.*, 2000 and references therein). The genes encoding several *vsa*-lipoprotein fragments A, F, C, E, G, I and H as well as the lipoproteins D, B and A occurring in this organism were observed as part of the region 636001–656000 identified as compositionally distinct by our method.

3.4.4 Mycobacterium tuberculosis H37Rv

The PE and PPE gene families in *M. tuberculosis* encode large multi-protein families of unknown function (Camus *et al.*, 2002; Cole *et al.*, 1998). These proteins comprise ~10% of the coding region in the *M. tuberculosis* genome (Cole *et al.*, 1998). Due to the highly polymorphic nature of the C-terminal domain in both these families, they are thought to be involved in antigenic variation (Cole, 1999; Cole *et al.*, 1998). The PE-PGRS is the largest PE subfamily (Banu *et al.*, 2002). Thus PE-PGRS is a major source of polymorphism in the *M. tuberculosis* complex, which otherwise displays a high degree of genetic homogeneity and very few single nucleotide polymorphisms (Sreevatsan *et al.*, 1997). The following regions identified by our method contain gene(s) encoding PE-PGRS proteins: 330001–344000, 832001–842000, 1210001–1222000, 1628001–1640000, 2794001–2806000, 3728001–3768000 and 3922001–3956000. That is, they were observed in 7 out of 11 regions predicted by our approach in this organism.

Proteins with a general designation of PPE family proteins could be observed in the following regions: 330001–344000, 364001–378000, 420001–438000, 2162001–2172000 and 3728001–3768000. The region 1210001–1222000 also encodes two other PE family proteins, besides PE-PGRS proteins. Most of the regions identified also encode several conserved hypothetical proteins.

3.4.5 Yersinia pseudotuberculosis IP32953

A siderophore called yersiniabactin, chelates iron molecules bound to
The identified region 1920001–1948000 has five yersinia bacteriactin siderophore biosynthesis proteins. Several players involved in the sugar metabolism are observed in the identified region 1198001–1218000. For instance, in this region, we observed three glucose dehydrogenases, a paratose synthase and two putative mannosyltransferases. The identified region spanning the genome coordinates between 3600001 and 3624000 contains the following general secretion pathway proteins: L, K, J, and C. The identified region spanning 4002001–4016000 contains a putative type IV secretion ATPase, a possible Yersinia enterocolitica-like Orf1, six membrane proteins, a putative exported protein and an ExbB/TolR-family transport protein. Genome coordinates between 4110001 and 4122000 have a single annotation for putative hemolysin activator/exporter. The region spanning 4302001–4524000, identified as compositionally distinct by our method, has a putative membrane protein, a possible bacterial Ig-like domain (group 1) and an aspartate semialdehyde dehydrogenase. Conserved hypothetical and hypothetical proteins were seen in several of the regions identified as compositionally distinct. Description provided here pertains to 6 out of 11 regions identified for this organism. Refer to the Supplementary Table D for complete details about the gene contents.

4 DISCUSSION

The centroid method identifies compositionally distinct regions in a genome based on word frequency vectors corresponding to individual bins. This method has the ability to identify embedded sequences of foreign origins. Thus our method passes the necessary tests for the identification of compositionally distinct regions. In addition, we also show that the efficacy of this method over the currently widely used methods (%G+C and Karlin’s genomic signature approach) in identifying compositionally distinct genomic regions. Centroid method being designed primarily for the identification of compositionally distinct regions in a genome, one would expect at best only a weak detection of an embedded non-native sequence coming from a closely related organism. Our results are in agreement with this expectation. In just one out of the four chimeras constructed, our method could identify the embedded non-native sequence.

We also analyzed the gene contents of all the compositionally distinct bins identified by this method for 4 out of the 50 bacterial genomes chosen in this study. We observed examples of well known compositionally distinct and/or alien genes in the outlier bins of all four genomes examined. For instance, we observe the region comprising the genes encoding the cag PAI proteins identified by our method in the case of H. pylori 26695. In addition, other interesting regions characteristic of each organism are clearly identified. This includes, \textit{vsu} genes characteristic of \textit{M. pulmonis}, genes encoding PE-PGRS proteins of \textit{M. tuberculosis} H37Rv and the genes encoding yersiniabactin siderophore biosynthesis proteins characteristic of \textit{Yersinia}.

We have also compared the outcomes of our method with those of other methods based on G+C genome variation, genomic signature divergences, extremes of codon bias and anomalies of amino acid usage (Karlin, 2001). For example, in the case of \textit{H. pylori} J99 our method along with two of the methods (genomic signature difference and codon bias) identified cag island proteins. Similarly, for \textit{Pseudomonas aeruginosa}, our method identified the wbp operon region containing hisF2 and hisF3 genes. This wbp operon is a cluster of 16 genes involved in the synthesis of \textit{P. aeruginosa} PAO1 (serotype O5) O antigen (Burrows et al., 1996; Lightfoot and Lam, 1993). It was clearly seen earlier that the genomic signature difference method does not identify these regions effectively. In the case of \textit{M. tuberculosis}, previous methods including %G+C, genomic signature difference, codon bias and amino acid bias identified regions containing only five PE-PGRS genes. Using our method, we have been able to identify as many as 14 annotated PE-PGRS genes. In the case of \textit{E. coli} K12 MG1655, our method detects several CP4-6 prophage and Qin prophage proteins, which is comparable to the prophage components detected by genome signature difference and codon bias approaches in the case of \textit{E. coli} O157.

The centroid method that we describe is therefore able to identify additional compositionally distinct regions that have been missed by currently available analytical tools. It can be applied to any bacterial genome, independent of the availability of annotation data and will facilitate the identification of genomic islands and portions of genome that are horizontally acquired.

ACKNOWLEDGEMENTS

We thank Tata Consultancy Services (TCS) Limited, Hyderabad, India for providing the infrastructure and all the necessary support for carrying out this work. We also thank Rajgopal Srinivasan and Probah Chaudhuri for fruitful discussions and Vidya G. Krishnan for generating some data.

Conflict of Interest: none declared.

REFERENCES


Orsini, B. et al. (2003) *Helicobacter pylori* cag pathogenicity island is associated with the reduced expression of interleukin-4 (IL-4) mRNA and modulation of the IL-4 mRNA isoform in human gastric mucosa. *Infect. Immum.*, **71**, 6664–6667.


