journal homepage: www.intl.elsevierhealth.com/journals/cmpb

n-Gram characterization of genomic islands in bacterial genomes

Gordana M. Pavlović-Lažetić[a,*], *Nenad S. Mitić*[a], *Miloš V. Beljanski*[b]

[a] *University of Belgrade, Studentski trg 16, 11001 Belgrade, Serbia*
[b] *IGPC, Studentski trg 16, 11001 Belgrade, Serbia*

ARTICLE INFO

Article history:
Received 20 April 2008
Received in revised form
10 September 2008
Accepted 21 October 2008

Keywords:
n-Grams
Statistical analysis
Zipf-like analysis
Genomic islands
Horizontal gene transfer
Backbone sequence
Escherichia coli O157:H7 EDL933

ABSTRACT

The paper presents a novel, *n*-gram-based method for analysis of bacterial genome segments known as *genomic islands* (GIs). Identification of GIs in bacterial genomes is an important task since many of them represent inserts that may contribute to bacterial evolution and pathogenesis. In order to characterize and distinguish GIs from rest of the genome, binary classification of islands based on *n*-gram frequency distribution have been performed. It consists of testing the agreement of islands *n*-gram frequency distributions with the complete genome and backbone sequence. In addition, a statistic based on the maximal order Markov model is used to identify significantly overrepresented and underrepresented *n*-grams in islands. The results may be used as a basis for Zipf-like analysis suggesting that some of the *n*-grams are overrepresented in a subset of islands and underrepresented in the backbone, or vice versa, thus complementing the binary classification. The method is applied to strain-specific regions in the *Escherichia coli* O157:H7 EDL933 genome (O-islands), resulting in two groups of O-islands with different *n*-gram characteristics. It refines a characterization based on other compositional features such as G + C content and codon usage, and may help in identification of GIs, and also in research and development of adequate drugs targeting virulence genes in them.

1. Introduction

Many bacterial genomes have been shown to contain specific genomic regions, known as islands. Islands that were acquired by horizontal gene transfer (HGT) events among bacteria are designated as genomic islands (GIs) and may contribute to their adaptability. Genes encoded in GIs offer various functions, e.g., additional metabolic activities, the capability of symbiosis with other organisms, antibiotic resistance and secretion, etc. [1]. The group of GIs that contain a variety of virulence factors, providing for specific host recognition, penetration and colonization of the host organism, and the ability to overcome host defense systems, are known collectively as pathogenicity islands (PAIs). GIs are identified and characterized by different compositional features, such as biased G + C content, codon usage (CU), dinucleotide signature contrasts, amino acid contrasts [2], and different functional features such as the presence of virulence genes and mobility (e.g., integrases, transposases) genes, or structural features such as the presence of proximal tRNA and/or rRNA gene(s) and repeats at their boundaries, presence of insertion sequence elements, origin of plasmid replication, etc. [3,4].

Abbreviations: HGT, horizontal gene transfer; GI, genomic island; OI, O-island; PAI, pathogenesis island; CU, codon usage.

* *Corresponding author at*: Faculty of Mathematics, University of Belgrade, P.O.B. 550, Studentski trg 16, 11001 Belgrade, Serbia. Tel.: +381 11 2027801; fax: +381 11 2630151.
E-mail addresses: gordana@matf.bg.ac.yu (G.M. Pavlović-Lažetić), nenad@matf.bg.ac.yu (N.S. Mitić), mbel@matf.bg.ac.yu (M.V. Beljanski).

0169-2607/$ – see front matter © 2008 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.cmpb.2008.10.014

In this paper we apply a linguistic method – exhaustive *n*-gram analysis – to already annotated islands in an attempt to characterize GIs more precisely than proved possible using earlier techniques, and to understand their structure better. We illustrate the method on the *Escherichia coli* O157:H7 EDL933 genome, a member of genus *Escherichia* of *Enterobacteriaceae phylum* and a well known and important experimental, medical and biotechnological organism [3,5].

The paper is organized in the following way. Section 2 surveys different methods and algorithms for identification and prediction of GIs, including the *n*-gram technique and its applicability to characterization of different types of texts. It also outlines the authors' prior work in the field. Section 3 describes the three steps of the GI characterization procedure. We first perform *n*-gram statistical analysis of islands, for different *n*, in order to classify them according to (dis)agreement with the complete genome. Next, we apply other compositional features (G+C content, CU) to islands and calculate statistical measures—recall, precision, sensitivity and specificity for the results of *n*-gram classification so as to examine how the *n*-gram feature contributes to characterizing GIs. Then we identify significantly overrepresented and underrepresented *n*-grams based on the maximal order Markov model, which may be used as a basis for Zipf-like analysis and for classification of islands based on such *n*-grams [6]. Section 4 presents computational results obtained for the *E. coli* EDL 933 genome. In Section 5 we offer our conclusions and outline some future plans.

2. Background

2.1. *n-Grams*

An *n*-gram, as introduced by Shannon in 1948 [7], is a subsequence of length n of a sequence over the given alphabet. The sequence may be a message in a natural or artificial language, a discrete approximation of a continuous signal, e.g., speech, or any sequence of symbols generated by a stochastic process. Any such "text" can be approximated by the set of *n*-gram statistical data (e.g. frequency distribution and the respective mean and standard deviation), and two such texts may be compared based on the distance of such approximations. *n*-Grams of length 2, 3 and 4 are usually called *bigrams*, *trigrams* and *tetragrams*, respectively, and for higher values of *n*—simply *n*-grams.

Formally, as defined in Vinga and Almeida [8], a sequence X of length k is a linear succession of k symbols from a finite alphabet, A, of cardinality $|A| = r$. A segment of n consecutive symbols from the sequence X ($n \leq k$) is an n-gram (n-tuple, n-word, n-plet, n-mer) of the sequence X. There are $L = r^n$ different n-grams over the alphabet A, $\{w_1, w_2, \ldots w_L\}$. There are $k - n + 1$ overlapping n-grams in the sequence X. Some authors use n-grams in a broader sense not assuming contingency of symbols but a distance of a given length between them (spacer).

If c_i denotes the number of occurrences of the n-gram w_i ($i = 1, 2, \ldots, L$) in the sequence X, and f_i denotes relative frequency of the n-gram w_i in the sequence X ($f_i = c_i/(k - n + 1)$), then a vector of n-gram counts, $c_n^X = (c_1, c_2, \ldots, c_L)$, as well as a vector of n-gram frequencies, $f_n^X = (f_1, f_2, \ldots, f_L)$ may be associated with the sequence X. For example, for DNA sequences, $A = \{A, C, G, T\}$, $r = 4$; for $n = 2$, number of all possible bigrams is $L = r^n = 16$ and the set of all possible bigrams will be $\{AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG, TT\}$. For the sequence $X = ATATAC$, where $k = 6$, there are $k - n + 1 = 5$ bigrams, determined by sliding a two letter window: *AT*, *TA*, *AT*, *TA*, *AC*, so the vector of n-gram counts will be (0, 1, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0), and the vector of n-gram frequencies will be (0, 0.2, 0, 0.4, 0, 0, 0, 0, 0, 0, 0, 0, 0.4, 0, 0, 0).

The dissimilarity of two sequences, X and Y, may be defined by a distance function computed in the vector spaces of either n-gram counts or n-gram frequencies.

In general, n-gram analysis proves effective, regardless of the type and origin of the text analyzed. *n*-Grams were first used in the domain of natural language text analysis, for different tasks such as text compression [9,10], spelling error detection and correction [11,12], language identification [13], automatic text categorization [14], authorship attribution [15], etc. For example, Damashek [16] reports the automatic classification of a whole library of multilingual documents based on topical similarity which is determined by using *n*-gram s and Euclidian distance in the vector space of *n*-gram frequencies.

n-Gram-based methods have also been applied to protein, proteome and genome sequences, for different purposes—to measure sequence similarity and reconstruct phylogenetic trees without sequence alignment, for protein classification, genome characterization, etc. As early as in 1967, Krzywicki and Slonimski compared the expected with the observed frequencies of amino acid bigrams with distance and showed that for certain distances, statistically highly significant deviations are present in proteins [17]. Radomski and Slonimski applied bigram analysis to a set of ribosomal protein sequences some thirty years later [18] to develop the notion of the genomic "style" of proteins, and the concept of *n*-grams with a spacer was used by Rosato et al. [19] to analyze the thermal dependencies of different proteomes, with some observed anomalies in *n*-gram distribution at certain distances (spacer lengths).

Deviation of observed from expected *n*-gram frequencies further aided the investigation of overrepresented and underrepresented *n*-grams, both in nucleotide and in protein composition. Phillips et al. [20,21], Colosimo et al. [22], Schbath et al. [23], Gelfand and Koonin [24], Karlin et al. [25], Karlin and Burge [26], Rocha et al. [27], Pevzner et al. [28], Karlin et al. [29], Burge et al. [30] and Schbath [31], investigated identification of over- and underrepresented oligonucleotides and different methods for calculating the expected number of oligonucleotides and the comparison of the expected against the observed number of oligonucleotides. A comparison of different statistical measures of bias of oligonucleotide sequences in the DNA sequences of bacterial genomes is given by Elhai [32]. Noncontiguous sequences were considered since they exhibited significant bias, and the corresponding methods proved more efficient than Markov analysis at the highest order. Reinert et al. [33] reviewed statistical and probabilistic properties of words in sequences, with emphasis on the deductions of exact distributions and the evaluation of their asymptotic approximations.