Available online at www.sciencedirect.com
SCIENCE DIRECT
Gene 346 (2005) 173–185
GENE
SECTION
EVOLUTIONARY GENOMICS
www.elsevier.com/locate/gene

The spectrum of genomic signatures: from dinucleotides to chaos game representation

Yingwei Wang[a,*], Kathleen Hill[b], Shiva Singh[b], Lila Kari[a]

[a]Department of Computer Science, University of Western Ontario, London, Ontario, Canada N6A 5B7
[b]Department of Biology, University of Western Ontario, London, Ontario, Canada N6A 5B7

Received 8 March 2004; received in revised form 28 September 2004; accepted 21 October 2004
Available online 29 January 2005
Received by A.M. Campbell

Abstract

In the post genomic era, access to complete genome sequence data for numerous diverse species has opened multiple avenues for examining and comparing primary DNA sequence organization of entire genomes. Previously, the concept of a genomic signature was introduced with the observation of species-type specific Dinucleotide Relative Abundance Profiles (DRAPs); dinucleotides were identified as the subsequences with the greatest bias in representation in a majority of genomes. Herein, we demonstrate that DRAP is one particular genomic signature contained within a broader spectrum of signatures. Within this spectrum, an alternative genomic signature, Chaos Game Representation (CGR), provides a unique visualization of patterns in sequence organization. A genomic signature is associated with a particular integer order or subsequence length that represents a measure of the resolution or granularity in the analysis of primary DNA sequence organization. We quantitatively explore the organizational information provided by genomic signatures of different orders through different distance measures, including a novel Image Distance. The Image Distance and other existing distance measures are evaluated by comparing the phylogenetic trees they generate for 26 complete mitochondrial genomes from a diversity of species. The phylogenetic tree generated by the Image Distance is compatible with the known relatedness of species. Quantitative evaluation of the spectrum of genomic signatures may be used to ultimately gain insight into the determinants and biological relevance of the genome signatures.

Keywords: Dinucleotide Relative Abundance Profiles; Genomic signature distances; Phylogenetic trees; Organizational information of a DNA sequence

1. Introduction

Although efforts are continuously being made toward understanding the characteristics of genomes, any particular genome is too long and too complex for a person to directly comprehend its characteristics. In 1990, Jeffrey proposed using Chaos Game Representation (CGR) to visualize DNA primary sequence organization CGR (Jeffrey, 1990). A CGR is plotted in a square, the four vertices of which are labelled by the nucleotides A, C, G, T, respectively. The plotting procedure can be described by the following steps: the first nucleotide of the sequence is plotted halfway between the centre of the square and the vertex representing this nucleotide; successive nucleotides in the sequence are plotted halfway between the previous plotted point and the vertex representing the nucleotide being plotted. The major advantage of CGR is the use of a two-dimensional plot to provide a visual representation of primary DNA sequence organization for a sequence of any length, including entire genomes.

CGRs of DNA sequences show interesting patterns. Various geometric patterns, such as parallel lines, squares, rectangles, and triangles can be found in CGRs. Some of the CGRs even show a complex fractal geometrical pattern

Abbreviations: A, adenosine; C, cytidine; G, guanosine; T, thymidine.

* Corresponding author. Department of Computer Science and Information Technology, University of Prince Edward Island, Charlottetown, Prince Edward Island, C1A 4P3 Canada. Tel.: +1 902 566 0499; fax: +1 902 566 0466.
E-mail address: ywang@upei.ca (Y. Wang).

0378-1119/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2004.10.021

which is very similar to the Sierpinsky Triangle (Mandelbrot, 1982). These interesting features relevant to the DNA sequence organization attracted further research in CGR (Dutta and Das, 1992, Hill et al., 1992, Oliver et al., 1993).

In 1993 Goldman analyzed the patterns shown in CGRs and concluded that "it is unlikely that CGRs can be more useful than simple evaluation of nucleotide, dinucleotide and trinucleotide frequencies" (Goldman, 1993). According to this conclusion, CGR should be relegated to the status of a pictorial representation of nucleotide, dinucleotide and trinucleotide frequencies.

After this sobering conclusion, research on CGRs continued with less frequency. Hill and Singh (1997) compared CGRs of mitochondrial genomes and explored the evolution of species-type specificity in DNA sequences. Almeida et al. (2001) suggested that CGR is a generalization of Markov Chain probability tables that accommodates non-integer orders.

In parallel to CGR research, Karlin and Burge proposed the concept of *genomic signature* (Karlin and Burge, 1995). The key observation behind the genomic signature concept is that Dinucleotide Relative Abundance Profiles (DRAPs) of different DNA sequence samples from the same organism are generally much more similar to each other than to those of sequences from other organisms. In addition, closely related organisms generally have more similar DRAPs than distantly related organisms. It was concluded from these observations that the DRAP values constitute a genomic signature of an organism.

Since 1995, genomic signatures have been studied from a variety of perspectives, as witnessed by Karlin et al. (1997), Campbell et al. (1999), Deschavanne et al. (1999, 2000), Gentles and Karlin (2001), Sandberg et al. (2001), Edwards et al. (2002), and Hao et al. (2000). Campbell et al. (1999) compared genomic signatures of prokaryote, plasmid, and mitochondrial DNA. Deschavanne et al. (2000) showed that word usage in short fragments of genomic DNA (as short as 1 kb) is similar to that of the whole genome, thus providing a strong support to the concept of genomic signature. Gentles and Karlin (2001) looked at the genomic signature of various eukaryotes. Sandberg et al. (2001) proposed a method to classify sequence segments using genomic signatures. More recently, genomic signatures were used in phylogenetic analysis (Edwards et al., 2002).

In 1999, an interesting paper provided a link between CGRs and genomic signatures (Deschavanne et al., 1999). Experiments showed that variation between CGR images along a genome was smaller than variation among genomes. "These facts strongly support the concept of genomic signature and qualify the CGR representation as a powerful tool to unveil it" (Deschavanne et al., 1999).

In this paper, we discuss CGR and DRAP (currently proposed as genomic signature) from the following perspectives: In Section 2, we challenge the idea that CGR is merely a representation of nucleotide, dinucleotide, and trinucleotide frequencies. The aim of Section 2 is to provide evidence supporting the claim that CGRs have more complex features worth further investigation. In Section 3, we propose the idea of a spectrum of genomic signatures, and describe the common features as well as variations within this spectrum. Section 4 discusses various distance definitions between genomic signatures of two DNA sequences. *Order* is an integer number associated with a genomic signature to describe its granularity. In Section 5, we design an experiment to quantitatively analyze the information provided by the genomic signatures of different orders of a given DNA sequence. Section 6 presents our conclusions.

2. What determines the pattern in a CGR?

The interesting patterns in CGRs inspired exploration of the underlying determinants of these patterns in different ways. Hill et al. (1992) tried to use image analysis techniques to categorize and analyze CGRs. Goldman (1993) used Markov Chain model simulation to explore these determinants.

Goldman (1993) concluded that "the CGR gives no further insight into the structure of the DNA sequence than is given by the dinucleotide and trinucleotide frequencies" and "unless more complex patterns are found in CGRs, there is no justification for ascribing their patterns to anything other than the effects described in this paper." These conclusions had the effect that CGRs have subsequently been much less studied from this perspective. In this section we first present arguments supporting our claim that CGRs give more insight into DNA structures than those given by nucleotide, dinucleotide, and trinucleotide frequencies, and then present our answer to the question, "What determines the pattern in a CGR?"

2.1. Short nucleotide frequencies cannot solely determine the pattern in a CGR

The results reported in Goldman (1993) are obtained through DNA sequence simulation based on Markov Chain model. We first briefly introduce the first-order and second-order Markov Chain model. In the first-order Markov Chain model, successive bases in a simulated sequence depend only on the preceding base. A 4×4 matrix P defines the probabilities with which subsequent bases follow the current base in a DNA sequence. If the base labels A, C, G, and T are equated with the numbers 1, 2, 3, and 4, then P_{ij}, the jth element of the ith row of P, defines the probability that base j follows base i. The row-sums of P must equal 1. Using this matrix, a simulated DNA sequence is obtained by selecting a first base randomly, according to the frequencies of the bases in the DNA string under study; if this is base i, then the probabilities P_{i1}, P_{i2}, P_{i3}, and P_{i4} are used to select the next base, and so on until the simulated sequence is of the same length as the original DNA sequence.