Available online at www.sciencedirect.com
ScienceDirect
Gene 385 (2006) 75 – 82
GENE
SECTION
EVOLUTIONARY GENOMICS
www.elsevier.com/locate/gene

A new parameter to study compositional properties of non-coding regions in eukaryotic genomes

Emanuele Bultrini, Elisabetta Pizzi *

Dipartimento di Malattie Infettive, Parassitarie ed Immunomediate, Istituto Superiore di Sanità, Viale Regina Elena, 299, 00161 Roma, Italy

Received 18 January 2006; received in revised form 4 May 2006; accepted 19 May 2006
Available online 9 August 2006

Abstract

Genomes are characterized by global and local compositional properties that are interesting in an evolutionary perspective but also provide useful information for the identification of some functional elements. Following previous studies, in this work we investigated compositional properties of non-coding sequences in four eukaryotic genomes (*C. elegans*, *D. melanogaster*, *M. musculus*, *H. sapiens*). We developed a procedure based on Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) to identify pentamers that are over-represented in introns (intron vocabulary) and to define a new parameter (LD) that reflects oligonucleotide composition of a given sequence. We analyzed genomic sequences and we found that all non-coding parts of a genome are characterized by similar LD values. Furthermore, we used the new parameter to analyze potentially regulatory regions. We extracted non-redundant sets of promoter sequences for *D. melanogaster* and *H. sapiens* and we studied their compositional (G+C content and LD parameter) and conformational (bendability propensity) properties. We found that regions immediately surrounding transcription start sites are distinguishable because of their %G+C, LD and bendability values.

Keywords: *D. melanogaster*; *C. elegans*; *H. sapiens*; *M. musculus*; Principal component analysis; Linear discriminant analysis; Eukaryotic promoters

1. Introduction

Compositional properties of genomes have been under investigation long before the beginning of sequencing projects. It is known that genomes are characterized by specific composition at the level of nucleotides (Grantham et al., 1980) but also typical oligonucleotide usage was observed as species-dependent. Campbell et al. (1999) observed a consistent dinucleotide usage in genomes and defined it as a “genome signature”. They found that the signature is relatively constant throughout the genome and that comparisons of signatures provide a measure for similarity between genomic sequences independently of sequence alignments. Furthermore, it was shown that some genomes can be represented as a patchwork of different regions following specific nucleotide composition. Using a wide range of methods, Bernardi and co-workers observed that vertebrate genomes are mosaics of isochores that are long regions of DNA, homogeneous in base composition and correlated with gene distribution (for a review see Bernardi, 2000a). Analyses of compositional and statistical properties of DNA allowed them to shed light on possible mechanisms that shape genomes and to trace an evolutionary history of vertebrate genomes in general (Bernardi, 2000b), and of the human genome in particular (Bernardi, 1995).

Many of these compositional properties have been confirmed and thoroughly analyzed after the era of genome sequencing projects. Compositional and statistical properties have been used to undertake comparative analyses between genomes, to investigate evolutionary mechanisms (Karlin and Ladunga, 1994; Nekrutenko and Li, 2000; Gentles and Karlin, 2001) and to construct unconventional phylogenetic trees (Baronchelli et al., 2004; Chapus et al., 2005; Foerstner et al., 2005). Standard statistical techniques were successfully applied by Zinovyev’s group to investigate composition of bacterial genomes. They performed principal component analysis (PCA) on triplet content of genomic sequences and found that they can be described by a universal 7-cluster structure (Alexander et al., 2005). Furthermore they defined

Abbreviations: PCA, principal component analysis; LDA, linear discriminant analysis; LD, linear discriminator; TSS, transcription start site; EPD, Eukaryotic Promoter Database; EID, Exon–Intron Database; DBTSS, DataBase Transcription Start Sites.

* Corresponding author. Tel.: +39 0649902226; fax: +39 0649902226.
E-mail address: epizzi@iss.it (E. Pizzi).

0378-1119/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2006.05.030

codon bias signatures for a wide range of genomes (80 from Eubacteria; 16 from Archea) and correlated them with some features that characterize lifestyle of the organisms (optimal growth temperature, aerobic and anaerobic respiration) (Carbone et al., 2005).

On the other hand, studies of compositional properties within a genome provided information about regions that play functional and/or structural roles. For example, differences in hexamer composition became part of gene prediction methods (Cruveiller et al., 2003), and occurrences of CpG islands were detected to predict human PolII promoter (Pedersen et al., 1999; Hannenalli and Levy, 2001).

In recent works, evidence emerged that typical compositional features characterize specific regions such as promoters. Aerts et al. (2004) analyzed promoter sequences from several metazoan and found that a species-specific base composition characterizes regions around the transcription start sites (TSSs). Kanhere and Bansal (2005) found differences in compositional properties between prokaryotic and eukaryotic promoter sequences, even if promoters from different genomes show several common structural features. Kel-Margoulis et al. (2003) studied the G + C content variations along the human genome and found that promoters possess some characteristics that differentiate them from random sequences as well as from other functional parts of the genome, such as surrounding sequences, exons, introns and repetitive elements. Mariño-Ramirez et al. (2004) developed a statistical procedure based on *z*-scores to identify over- and under-represented oligonucleotides in human promoter sequences. They identified a set of oligonucleotides as potential regulatory elements, many of which were recognized to be real transcription factor binding sites. In a recent past, gene expression data at the genomic scale for *Saccharomyces cerevisiae* became available, and this led to the development of many statistical approaches for the identification of potential regulatory elements in upstream regions of co-regulated genes (Wolfsberg et al., 1999; Van Helden et al., 1998; Brazma et al., 1998; Cora et al., 2004).

In previous works (Frontali and Pizzi, 1999; Bultrini et al., 2003) we carried out compositional analyses on *C. elegans* and *D. melanogaster* genomes. We identified a set of pentamers that are over-represented in introns with respect to exons and random expectation and we defined it as intron vocabulary. We also showed that this is a global compositional property of all non-coding portions of a genome. However, further investigations revealed that our approach was not suitable to detect intron vocabularies in other more complicated cases such as human and mouse genomes.

In the present work we propose a new procedure that allowed us to identify oligonucleotides over-represented in non-coding tracts of any genome. By means of this new approach we were able to identify a parameter (LD) reflecting this signature. We applied this new procedure to four eukaryotic genomes for which complete sequences are available (*C. elegans*, *D. melanogaster*, *M. musculus*, *H. sapiens*). First of all, we used intron sequences, masked for repetitive elements (mini and microsatellites), and derived new intron vocabularies. Intergenic sequences were then analyzed by means of the new parameter LD and their accordance to the vocabulary was assessed.

Finally, since regions with functional and/or structural role are usually characterized by typical compositional properties (Aerts et al., 2004; Kanhere and Bansal, 2005), we decided to analyze these regions using the new parameter LD. We considered experimentally determined promoter sequences available for *D. melanogaster* and *H. sapiens* and showed that, on average, regions immediately surrounding the TSSs, beside being characterized by a typical nucleotide composition and bendability propensity, can also be clearly distinguished by their low LD values.

2. Methods

2.1. Data source

Experimentally determined intron and exon sequences (longer than 1 kb) of *C. elegans*, *D. melanogaster*, *M. musculus* and *H. sapiens* were extracted from Exon–Intron Database (EID; http://www.meduohio.edu/bioinfo/eid/) (Saxonov et al., 2000). In order to exclude splicing junction motifs and branch signals, corresponding sequences (8 bp at the 5′ end, 50 at the 3′ end) were eliminated. Furthermore all sequences were submitted to a RepeatMasker procedure at http://repeatmasker.org to discard all internal repetitive elements.

We downloaded sequences of experimentally determined promoters (−500 bp upstream and +500 bp downstream the TSS) for *D. melanogaster* and *H. sapiens* from Eukaryotic Promoter Database (EPD, http://www.epd.isb-sib.ch/) (Périer et al., 2000). In order to obtain non-redundant sets we discarded sequences more similar than 80% of identity within each set of data. We obtained 1921 sequences for *D. melanogaster*, and 1723 for *H. sapiens*. Furthermore we extracted available promoter sequences (−500 bp upstream and +100 downstream the TSS) from DataBase of Transcriptional Start Sites (DBTSS, http://dbtss.hgc.jp/) (Suzuki et al., 2004), for *M. musculus* and *H. sapiens*; non-redundant sets comprise 9934 and 10,564 sequences respectively.

Intergenic portions were extracted from *C. elegans* chromosome 1 (GenBank accession no. NC_003279), *D. melanogaster* chromosome 3R (GenBank accession no. NT_033777) and 3L (GenBank accession no. NT_037436), *M. musculus* chromosome 19 (GenBank accession no. NT_039687), *H. sapiens* chromosome 19 (GenBank accession no. NT_011109). In all cases, all non-coding sequences ranging from the stop codon to the first neighbouring ATG codon were extracted and then joined together in 5′–3′ direction to form a sole supersequence.

2.2. Principal component analysis and linear discriminant analysis

We constructed four datasets: *C. elegans* (70 introns, 70 exons), *D. melanogaster* (87 introns, 87 exons), *M. musculus* (40 introns, 40 exons), *H. sapiens* (163 introns, 163 exons). In each dataset shuffled versions of intron sequences were added.

For every sequence the 1024 frequencies of overlapped pentamers were calculated so that a sequence was considered as a vector in a 1024-dimensional space. In the first step we applied PCA; this analysis allows the identification of new