Journal of Theoretical Biology 363 (2014) 357–366

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Variation and constraints in species-specific promoter sequences

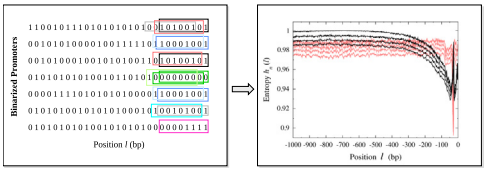
Elisa Calistri [a,b], Marcello Buiatti [b], Roberto Livi [a,c,*]

[a] Center for the Study of Complex Dynamics (CSDC), University of Florence, Via G. Sansone 1, 50019 Sesto Fiorentino (Fi), Italy
[b] Department of Biology, University of Florence, Via Madonna del Piano 6, 50019 Sesto Fiorentino (Fi), Italy
[c] Department of Physics, University of Florence and INFN Florence, Via G. Sansone 1, 50019 Sesto Fiorentino (Fi), Italy

HIGHLIGHTS

- Base composition and NPSE analyses of promoters highlight different structural classes.
- NPSE quantifies the conservation of motifs in specific positions across promoters.
- Promoter's regularities are correlated with spatial organization of W/S nucleotides.
- There are biases in the distributions of low-complexity sequences in human promoters.
- At promoter level, sequence structure is related to the function of the gene locus.

GRAPHICAL ABSTRACT



The **Normalised Positional Shannon Entropy** $h_N(l)$
as a measure of the information content in clusters of gene promoters.

ARTICLE INFO

Article history:
Received 26 March 2014
Received in revised form
30 July 2014
Accepted 4 August 2014
Available online 19 August 2014

Keywords:
Gene regulation
Base composition
Entropy
Low-complexity sequence
CpG

ABSTRACT

A vast literature is nowadays devoted to the search of correlations between transcription related functions and the composition of sequences upstream the Transcription Start Site. Little is known about the possible functional effects of nucleotide distributions on the conformational landscape of DNA in such regions. We have used suitable statistical indicators for identifying sequences that may play an important role in regulating transcription processes. In particular, we have analyzed base composition, periodicity and information content in sets of aligned promoters clustered according to functional information in order to obtain an insight on the main structural differences between promoters regulating genes with different functions. Our results show that when we select promoters according to some biological information, in a single species, at least in vertebrates, we observe structurally different classes of sequences. The highly variable and differentiated gene expression patterns may explain the great extent of structural differentiation observed in complex organisms. In fact, despite our analysis is focused on *Homo sapiens*, we provide also a comparison with other species, selected at different positions in the phylogenetic tree.

1. Introduction

A promoter is the region of DNA that contains the information required for transcriptional initiation and regulation. At its most fundamental level, the function of a promoter is to recognize information about the status of the cell and, accordingly, start transcription and regulate its rate (Core et al., 2012; The ENCODE Project Consortium, 2012). The activation and the rate of transcription initiation are determined both by a regulation *in cis*, mediated by DNA sequences, and *in trans*, mediated by the different signalling molecules, activators or repressors, that interact with a wide collection of different transcription-factor-binding-sites (TFBSs) (Carey et al., 2012). However, the transcriptional output of a promoter is not a simple

* Corresponding author. Tel.: +39 055 4572332; fax: +39 055 4572121.

http://dx.doi.org/10.1016/j.jtbi.2014.08.006
0022-5193/

function of its binding motifs that interact with regulatory proteins, but it depends also on the three-dimensional dynamics of the DNA string. The chemo-physical properties of the sequence of nucleotides in a promoter are crucial for determining the double helix dynamics and, accordingly, the transmission of the genetic message (Lavery et al., 2010; Alexandrov et al., 2010; Smith et al., 1996; Van Erp et al., 2005). This is the reason why the region near the Transcription Start Site (TSS) in eukaryotes exhibits a specific structural profile (Abeel et al., 2008; Fujimori et al., 2005; Louie et al., 2003). With these premises a systematic study of promoter structure looks timely and appropriate.

Our approach is in line with earlier theoretical studies aimed at understanding genomic architectures by measuring base composition, putative long- and short-range correlations, periodicities, the information content or "complexity" in DNA sequences, (Li, 1997; Louie et al., 2003; Aerts et al., 2004; Trifonov, 1998; Lió et al., 1994, 1996a, 1996b; Schieg and Herzel, 2004; Arneodo et al., 2011; Mantegna et al., 1994; Menconi et al., 2008). In a previous paper (Calistri et al., 2011), we used Base Composition Analysis (BCA) to identify differences among species, keeping trace of their phylogenetic relationship. In this paper we further investigate the putative existence of correlations between base composition along promoter sequences and the function of the genes they control. In particular, functional properties are expected to be related to the presence of specific motifs, i.e. over-/under-expressed nucleotide substrings which interact with regulatory proteins (Carey et al., 2012; Neph et al., 2012) or to peculiar structural and mechanical features of DNA (Eddy et al., 2011; Gemayel et al., 2010; Sela and Lukatsky, 2011). Then, we introduce a specific spatial entropy indicator, the Normalized Positional Shannon Entropy (NPSE), that is measured in the binary code based on the distinction between Weak (A and T) and Strong (C and G) bases. NPSE provides a measure of "conserved" as well as variable motifs of different lengths at each position in space along the whole strings of promoters, thus allowing a comparison between putatively function related entropy patterns in the different functional classes. We have further refined the statistical analysis of spatial organization in DNA by searching for specific low complexity sequences, i.e. homogeneous patterns of weak/strong bases and over-/under-expressed dinucleotides (e.g., CpG), known to be related to special structures in DNA as well as to biological functions (Sela and Lukatsky, 2011; Bird, 1986). Most of the results reported in this paper concern *Homo sapiens*, data of this species being the best documented ones. Due to a high tissue differentiation, this species offers a good model organism to measure the variability content among the promoters of different gene classes. A few other well-documented eukaryotic species have also been studied and the data obtained are reported in Supplementary material for a possible comparison with *H. sapiens*.

2. Materials and methods

2.1. Sequence databases

Analyses are worked out on a genome-wide set of human promoters made of 32,122 elements, downloaded from DBTSS (Version 6.0, based on UCSC hg18), a database of TSSs, obtained from a collection of experimentally-determined 5′-end sequences of full-length cDNAs (Yamashita et al., 2006). Each promoter is represented by the 1000 base pairs (bp) upstream the TSS of all annotated genes. The organisms analysed in Supplementary material have been chosen for their different positions along the evolutionary scale, they are the unicellular red algae *Cyanidioschyzon merolae*, the model plant *Arabidopsis thaliana* and three animals such as the zebrafish *Danio rerio* and the two model mammals *Rattus norvegicus* and *Mus musculus*. Apart *A. thaliana*, all data comes from the same database (DBTSS, Version 6.0) and, accordingly, they have been collected making use of a common procedure. For what concerns *A. thaliana*, promoter sequences have been downloaded from TAIR (The Arabidopsis Information Resource) web site (released in March 2008) and also in this case annotation of genes is largely based on sequenced cDNAs and ESTs alignments with the genome (Swarbreck, 2008).

2.2. Base composition analysis (BCA)

We have carried out the global analysis of the nucleotide composition according to spatial distribution in a set of N promoters, by computing the quantity (see Calistri et al., 2011)

$$\rho_k(l) = \frac{1}{N}\sum_{i=1}^{N} s_i(k;l) \tag{1}$$

where the index i labels the N promoter sequences, k = A, C, G, T denotes the different nucleotides and $l = -1$ bp, -2 bp, …, -1000 bp identifies the position of nucleotides in the promoters preceding the TSS (that is assumed to be located at the origin); the variable $s_i(k;l)$ is defined as follows:

$$s_i(k;l) = \begin{cases} 1 & \text{if nucleotide } k \text{ is at position } l \text{ in promoter } i, \\ 0 & \text{otherwise} \end{cases}$$

Accordingly, $\rho_k(l)$ measures the density of nucleotide k at position l along the promoter sequence, averaged over N promoters. We want to point out that the phenomenological quantity $\rho_k(l)$ exhibits fluctuations (e.g., see Fig. 1). It can be considered a good statistical indicator, provided the averaging over a set of N promoters exhibits stationary features. This can be checked by computing first the *average* density of nucleotides $\overline{\rho}_k(l)$: this is defined as the best-fit function of $\rho_k(l)$, averaged over the whole database of promoters. Then, one can compute the standard deviation from the average density $\overline{\rho}_k(l)$ of a set of N randomly selected promoters from the database. We have found that, as expected, the standard deviation scales as $N^{-1/2}$ at each position l (not shown). On the other hand, one should point out that there is no a priori argument for assuming that nature has selected nucleotides in promoter sequences according to any preassigned function $\overline{\rho}_k(l)$, nonetheless, this is a well-defined mathematical quantity.

2.3. Normalized Positional Shannon Entropy (NPSE)

We have introduced a more refined statistical indicator for characterizing the information content of promoter sequences (Shannon, 1948). This is the Normalized Positional Shannon Entropy (NPSE)

$$h_{n,b}(l) = -\frac{1}{n}\sum_{w_n} f_{w_n}(l)\log_b f_{w_n}(l) \tag{2}$$

where $f_{w_n}(l)$ is the frequency of occurrence in the set of promoters of a substring (word) w_n of length n bp at position l. The words w_n are expressed in an alphabet of b letters. The normalization factor $1/n$ allows to compare NPSE obtained for different values of n: one can easily realize that the relation $0 \le h_n(l) \le 1$ holds independently of n. Notice that for $n=1$ and $b=4$ (i.e., k=A, C, G, T) the NPSE essentially contains the same information of BCA, since $f_{(w_1=k)}(l) = \rho_k(l)$.

Since we are dealing with promoter sequences of finite length, namely 1000 bp, a major limitation of the NPSE is that it can be affected by poor statistics when large-n words are taken into account. On the other hand, direct inspection of BCA (e.g., see Fig. 1) shows that $\rho_A(l)$ and $\rho_T(l)$ exhibit common trends, as well as $\rho_C(l)$ and $\rho_G(l)$. This suggests that the quaternary ($b=4$) word-code