Linguistic analysis of promoter regions in eukaryotic genomes

Emanuele Bultrini and Elisabetta Pizzi
Istituto Superiore di Sanità, Viale Regina Elena 299 – 00144 Rome (Italy) – email:epizzi@iss.it

Keywords: Promoters, Principal Components Analysis, Linguistic Methods, Genomics

Introduction

Promoter recognition is one of the most difficult tasks in annotating eukaryotic genomes. Binding sites for transcription factors are very short sequences (5-15 bp) and not very well preserved in sequence. In addition, other signals can be associated with a regulatory region. For instance in vertebrates, some classes of promoters are associated with compositionally characterised regions (CpG islands) and there is also evidence that molecular conformation of human promoters is involved in the transcription activity [1, 2].
Following a previous investigation [3, 4], in the present work we propose a new procedure, based on well established statistical methods, to extract a set of oligonucleotides specifically characterising intron sequences.
Partitioning of genomic sequences, based on the accordance to the extracted "introns'vocabulary", reveals that intergenic DNA appears as a patchwork of different elements. The majority of them adopt the "introns' vocabulary", whereas some others (a small percentage) do not.
We hypothesise that the identified linguistic property is a sort of "background-noise" of a genome; in this perspective regions that play a functional and/or a structural role have probably to emerge from the background, adopting specific compositional properties.
The analysis of promoter sequences for the four examined genomes (*C. elegans, D. melanogaster, M. musculus, H. sapiens*) appears to confirm our hypothesis, as regions immediately surrounding the transcritpion start site deviate from the introns'vocabulary usage.
Furthermore, analyses on C+G composition, bendability propensity and torsional rigidity of promoter sequences are presented.

Results

In order to extract oligos that specifically characterise non-coding tracts in a genome, we developed a procedure based on well-established statistical methods. A reliable set of experimentally determined introns and exons from the Exon Intron DataBase (EID) [5] was selected and a set of randomised versions of intron sequences was added to take into account nucleotide compositional bias. Each sequence was then considered as a vector in a multidimensional space, the variables being the frequencies of the 1024 pentanucleotides.
A scheme of the procedure is reported below:

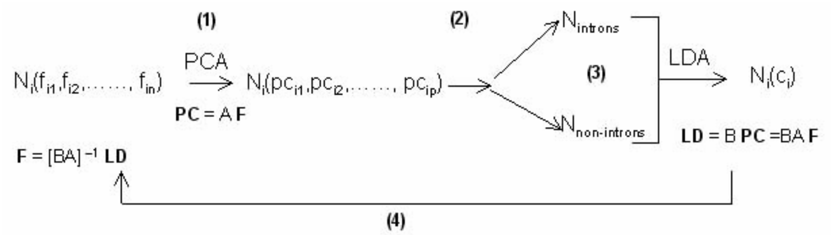

Fig.1: scheme of the procedure

In step (1) Principal Component Analysis (PCA), is applied to the data set composed by the three groups of sequences (introns, exons, and shuffled introns). PCA consists in a linear transformation in the variable space, providing an optimal reduction of space dimensionality.
In step (2) the first six Principal Components are selected.
In step (3) sequences are labelled as introns and non-introns and Linear Discriminant Analysis (LDA) is applied to the set of data in the new six-dimensional space. A unique variable is identified that discriminates between introns and other sequences (exons and randomised introns).
In step (4) the global transformation (PCA and LDA) is considered; its coefficients provide the contributions of old variables (pentamer frequencies) to the discrimination, and oligos contributing most are identified as the introns' vocabulary.

We applied this procedure to four completely sequenced genomes (*C. elegans, D. melanogaster, M. musculus, H. sapiens*). In order to test that vocabularies reflect a genome-wide linguistic property, correlation analyses were performed on whole chromosomes for each genome. Results confirm our hypothesis.

Further, it appears that intergenic and intron sequences share a common pentamer usage, except for a small population of intergenic tracts that deviate from the introns vocabulary.

Eukaryotic promoter sequences have been analysed in terms of their compositional and structural properties. We extracted from the Eukaryotic Promoter Database [6] a non-redundant set of promoters for the four examined species. For each sequence we constructed a moving window profile in which the degree of similarity with the "introns' vocabulary" is expressed by means of correlation coefficient (Corr) between vocabulary's words frequency distributions of each window and the average distribution in the original set of introns. Furthermore profiles based on C+G percentage, bendability propensity, and torsional rigidity were constructed.

Averaged profiles of Corr, %CG and bendability, concerning human and fly sequences are shown in figure 2.

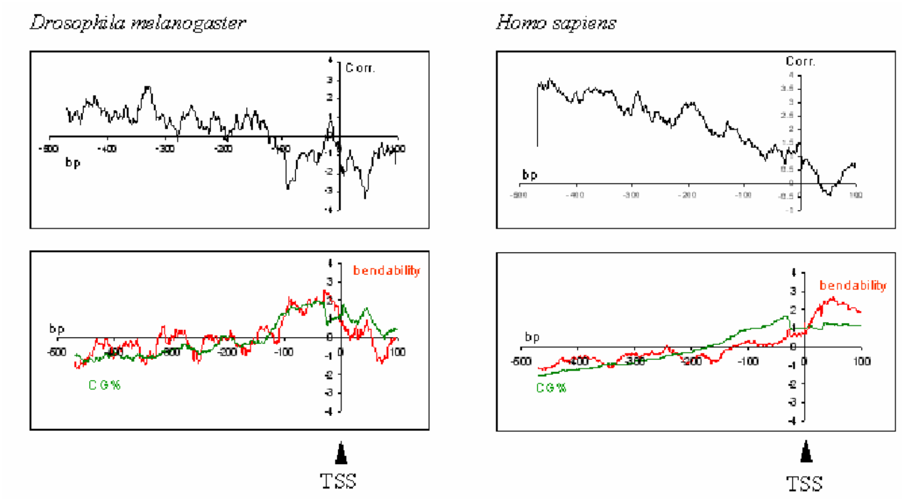

Fig.2: Profiles of correlation coefficients, bendability propensity and %CG of human and fly promoters

Regions immediately surrounding transcription start site (TSS) appear characterised by lower values of correlation coefficients. On the contrary CG% and bendability values increase. This suggests that regulatory regions (namely regions around TSS), besides being characterised by typical conformation, exhibit also typical compositional (CG%) and linguistic (Corr) properties.

These results support our initial hypothesis and suggest that genomic sequences adopting the identified genome specific vocabulary could be filtered out with the aim to isolate potential functional regions.

Acknowledgements

This work is supported by FIRB (Post Genomics Project).

References

[1] Pedersen, A.G., Baldi, P., Chauvin, Y. and Brunak, S. DNA structure in human RNA polymerase II promoters*. J. Mol. Biol*. 281: 663-673, 1998

[2] Pedersen, A.G., Baldi, P., Chauvin, Y. and Brunak, S. The biology of eukaryotic promoter prediction – a review. *Comput. Chem*. 15:191-207, 1999

[3] Frontali, C. and Pizzi, E. Similarity in oligonucleotide usage in introns and intergenic regions contributes to long-range correlation in the *Caenorhabditis elegans* genome. *Gene* 232: 87-95, 1999

[4] Bultrini, E., Pizzi, E., Del Giudice, P. and Frontali, C. Pentamer vocabulaires characterizing introns and intron-like intergenic tracts from *Caenorhabditis elegans* and *Drosophila melanogaster*. *Gene* 304: 183-192, 2003

[5] Saxonov, S., Daizadeh, I., Fedorov, A., and Gilbert, W. EID: The Exon-Intron Database-an exhaustive database of protein-coding intron-containing genes. *Nucleic Acids Res.* 28: 185-190, 2000

[6] Périer, R.C., Praz, V., Junier, T., Bonnard, C. And Bucher, P., The Eukaryotic Promoter Database (EPD). *Nucleic Acids Res.* 28: 302-303, 2000