

Compositional analysis of non-coding regions in eukaryotic genomes

Emanuele Bultrini, Paolo Del Giudice, Elisabetta Pizzi

Istituto Superiore di Sanità, Viale Regina Elena 299, 00161 Roma [Italy] - email:epizzi@iss.it

Keywords. Linguistic methods, Genomics, Molecular sequence analysis

Introduction

Whereas, in the last years, many efforts have been devoted to locate genes within genomes, relatively few tools have been developed to identify the regulatory regions required for the correct transcriptional activity of the genome. This task is particularly difficult in the case of eukaryotic organisms for which regulatory regions represent a small percentage, overwhelmed by –presumably- non-functional DNA. Recently, several computational procedures are emerging to solve this problem, including knowledge-based methods, comparative genome analysis as well as methods based on statistical-compositional properties of DNA sequences [for reviews see 1,2].

In this work we present a study of compositional properties of non-coding DNA in eukaryotic genomes. Specifically, we highlight statistical features characterizing introns and intergenic regions, and we formulate hypotheses concerning functional implications, with the aim to propose an approach suited to locate regulatory regions.

It is well known that genomes are characterised by species-specific compositional features, and that coding and non-coding DNA are distinguishable in terms of their pentamer and hexamer distributions [3, 4]. There is also evidence that some class of promoters are associated with compositionally characterised regions as in the case of CpG islands and bent DNA tracts [5-8].

By using recurrence quantitative analysis we were able to show that in some eukaryotic genomes, introns and intergenic tracts exhibit highly recurrent patterns with correlated properties distinguishing them from the low-recurrence regimen present in exons [9]. This observation was explained by assuming a preferential oligonucleotide usage in non-coding DNA. In a recent work [10] we confirmed this assumption by means of correlation analysis, using whole-genome data from the *C.elegans* genome. We studied relationships between pentamer frequency distributions in intron, exon and intergenic DNA and confirmed that non-coding DNA tracts share a similar oligonucleotide usage.

In order to characterise this usage, we applied principal component analysis (PCA) on pentamer distribution of experimentally confirmed introns and exons from *C.elegans*, *D. melanogaster*. We identified in either cases a subset of pentamers that significantly discriminates introns from their randomised counterparts and from exons. Results indicate that, at least in the two examined genomes, introns are characterised by a highly redundant usage of a limited number of pentamers (introns' vocabulary): 26/1024 for *C.elegans* and 40/1024 for *D.melanogaster*. Genome-wide analysis revealed a widespread but patchy usage of the introns' vocabulary along intergenic tracts, with a small percentage of interspersed elements emerging from the average genomic oligo composition.

Results

Our guess is that the genome-wide usage of the introns' vocabulary can be viewed as a sort of background noise. According to this hypothesis intergenic tracts characterised by a low vocabulary usage would be candidate regulatory regions. For a given genomic sequence we constructed a moving window profile reporting the total number of occurrences of vocabulary words in each window. Regions characterised by a low vocabulary usage appear as minima along these profiles.

We applied this procedure to the 5' upstream regions of members of the rifin multigene family from the *P.falciparum* genome. The rifin family includes about 149 members, organized in sub-telomerically located clusters [11,12]. We first extracted the introns' vocabulary for *P.falciparum* as in the case of *C. elegans* and *D. melanogaster*, and then constructed vocabulary-usage profiles for the 23 rifin upstream regions [1 kb long] from chromosome I and II. Each sequence was partitioned into windows 200 bp long, shifted by 50 bp. We found that four minima are common to all profiles and a subsequent analysis revealed that two of them correspond to regions containing motifs highly conserved among the 23 sequences.

We also found that a third common minimum corresponds to a local molecule conformation [bent DNA] even if does not imply any sequence similarity. Interestingly, it has been suggested that part of the sequence requirements for a functional transcription initiation site are of a structural nature and that the sequence heterogeneity of transcription factors binding sites might hide a conserved structural motif [8].

This preliminary results encourage us to extend our study to other genomes and to propose a new procedure for filtering out potential functional regions along eukaryotic genomes.

References

- [1] Fickett, J.W., and Hatzigeorgiou, A.G. Eukaryotic promoter recognition. *Genome Res.* 7:861-878. [1997].
- [2] Ohler, O. and Niemann, H. Identification and analysis of eukaryotic promoters: recent computational approaches. *TRENDS Genet.* 17:56-60. [2001].
- [3] Claverie, J.M., Sauvaget, I. and Bougueleret, L. K-tuple frequency analysis from intron/exon discrimination to T-cell epitope mapping. *Methods Enzimol.* 183: 237-252. [1990].
- [4] Fickett, J.W., Tung, C.S. Assessment of protein-coding measures. *Nucleic Acids Res.* 20:6441-6450. [1992].
- [5] Ihoshikhes, J.P. and Zhang, M.Q. Large-scale human promoter mapping using CpG islands. *Nat. Genet.* 26:61-63. [2000].
- [6] Benham, C.J. Computation of DNA structural variability – a new predictor of DNA regulatory sequences. *CABIOS*, 12: 375-381. [1996].
- [7] 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].
- [8] Pedersen, A.G., Baldi, P., Chauvin, Y. and Brunak, S. The biology of eukaryotic promoter prediction – a review. *Comput. Chem.* 15:191-207. [1999].
- [9] 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*: 87-95. [1999].
- [10] 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].
- [11] Gardner et al. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* 419: 498-511. [2002].
- [12] Pizzi, E. and Frontali, C. Fine structure of *Plasmodium falciparum* subtelomeric sequences. *Mol. Biochem. Parasitol.* 118: 253-258. [2001].