Computational analysis of non-coding regions in eukaryotic genomes - (session: Comparative Genomics and Molecular Evolution)

E. Pizzi, E. Bultrini, P. Del Giudice, C. Frontali

Istituto Superiore di Sanità, Roma

Genome sequencing projects determine a large amount of sequence data each year. One of the major challenges for computational biologists is to extract relevant biological information from billions of Megabases that have been stored in the databases so far. 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 genomics analysis as well as methods based on statistical-compositional properties of genomes.

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 regime present in exons. This observation was explained assuming a peculiar oligonucleotide usage in non-coding DNA and significant different in protein-coding regions. In order to characterise this oligonucleotide usage, we applied principal component analysis on pentamer distribution of experimentally introns and exons from C.elegans and D. melanogaster genomes. We found a subset of pentamers that significantly discriminate introns from their randomised counterparts and from exons. A genome-wide analysis of pentamer usage revealed that most introns and intergenic tracts utilize the identified subset of pentamers, whereas exons and a small percentage of non-coding fraction do not.

Our hypothesis is that genome pentamer-usage could be reviewed as a sort of genome background noise and hence functional sequences might emerge as regions having different compositional properties. In order to test our hypothesis, we analysed the 5, upstream regions of more than 100 members of a multigene family from P.falciparum genome. We identified four regions, within 1 kb, with an anomalous oligonucleotide-usage; we compared our results with those obtained through a multiple alignment performed on the same sequences.

The overall compositional property could be reviewed as a sort of genome background. Regulatory elements might take place within regions that adopt a different oligos usage.