Extraction of a pentanucleotide vocabulary shared by introns and intergenic elements

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

Istituto Superiore di Sanità - viale Regina Elena 299 - 00161 - Roma

Starting from the observation that similarity in oligonucleotide usage in introns and intergenic regions contributes to the correlation structure of the Caenorhabditis elegans, we applied principal component analysis to sets of experimentally confirmed C. elegans introns and exons in order to extract the subset of words (5 bp long) most relevant in differentiating the vocabulary of introns both from exons and from random permutations of intron sequences.

In C. elegans the extracted vocabulary is almost entirely composed of pairs of reverse complementary oligos. The extracted pentamer vocabulary is used to generate a probe suitable for scanning long genomic portions and scoring for consistent pentamer usage in regions not necessarily related by significant sequence similarity. This procedure identifies, in addition to introns, frequent intergenic elements sharing the same lexical features. These might mark the presence of as yet unpredicted genes (in current annotation), or might reflect the accumulation of intron-like elements in those non-coding regions that are under weak functional constraints.

Following the same procedure, a restricted vocabulary, partially overlapping the C. elegans intron vocabulary, is also found in non-coding regions of Drosophila melanogaster, thus revealing an interesting kind of inter-specific correlation, which might give some clues as to the evolutionary diversification of intron sequences.