A new strategy to identify novel genes and gene isoforms: whole genome comparison of human and mouse

Mignone F (1,2,*), Re M (1,*), Horner D (1), Pesole G (3)

(1) Dipartimento di Scienze Biomolecolari e Biotecnologie, Università degli studi di Milano
(2) Dipartimento di Chimica Strutturale e Stereochimica Inorganica, Università degli studi di Milano
(3) Dipartimento di Biochimica e Biologia Molecolare, Università di Bari
(*) These authors contributed equally to this work

Motivation

Despite consistent efforts to improve the annotation of the human genome, we are still far from either having a complete list of human genes or knowing the correct structure of already "annotated" genes. Moreover certain genes are characterized by a very low expression level which complicates the detection of their expression products. It is commonly accepted that one of the most reliable way to predict and identify novel protein-coding genes is the alignment of interspecific genomic sequences. This approach is constantly acquiring greater importance because of the increase in the rate of generation of complete (or near complete) genome sequences. Such an approach is also expected to generally improve the accuracy of gene annotation We present here an improvement of a methodology we previously developed for the identification of genome regions likely encoding protein-coding genes based on the detection of clusters of potentially coding conserved sequence tags (CSTs).

Methods

In a previous study, we presented a clustering method (benchmarked on human chromosomes 15, 21 and 22) able to highlight groups of coding CSTs characterized by a density comparable with that observed in annotated genes. CSTs are obtained using CSTminer, a tool that performs a crossgenome blast-like alignment and then calculates a Coding Potential Score (CPS) to discriminate between coding and non-coding sequences. We present here a new implementation of the clustering protocol with improved sensibility and selectivity. Moreover we have analyzed the entire collection of syntenic regions of H.sapiens and M.musculus genomes.

Results

A whole genome set of more than 130,000 coding Conserved Sequence Tags (CST) was obtained from our analysis generating more than 10,000 CST clusters. The accuracy of our observations has been assessed with respect to annotations contained in the latest human and mouse EnsEMBL release. Beside clusters of CSTs corresponding to genes annotated in both human and mouse genomes, we identified clusters corresponding to genes annotated in one genome only. We also identified CST clusters not corresponding to any annotated gene in either human and mouse genomes - potentially corresponding to unannotated genes. These clusters have been compared to expressed sequences databases to provide support to their genic nature.

Contact email: mailto: graziano.pesole@biologia.uniba.it