Prediction of new genes and transcriptome survey through an 'ab-initio'

ID - 203

Molineris Ivan[1], Sales Gabriele[1], Peyron Chiara[1], Di Cunto Ferdinando[2], Federico Bianchi[2], Michele Caselle[1]
[1]Department of Theoretical Physics, Università degli Studi, Torino
[2]Molecular Biotecknology Center, Università degli Studi, Torino

Motivation

Processed pseudogenes are DNA sequences which were generated through reverse transcription (typically LINE-mediated) of mature mRNAs. These sequences are usually considered as junk DNA since in most of the cases they lack a suitable promoter and are no longer transcribed. However, due to their origin, they represent a valuable source of information on the transcriptome (at least for cells belonging to the germ line). This type of information becomes of particular importance for organisms for which no EST database exist. However also for model organisms, like human or mouse, for which EST data exist processed pseudogenes can be used to refine the transcriptome picture obtained from EST clusters and allow to find new splicing variants and in some cases also new genes.

Methods

We constructed both for human and mouse a database of paralogous alignements (i.e. alignements of the human (mouse) genome with itself). We then filtered the alignements looking for clusters of nearby alignements which are the typical signature of retrotranscribed spliced sequences. Several further filters, described in the poster, allow us to reduce the number of false positive identifications.

Results

With the above analysis we were able to identify 2265 psrocessed psudogenes in human and 2063 in mouse. We compared our list of pseudogenes with those existing in the literature finding a very good overlap. It should be noticed that all the existing databases of processed pseudogenes were obtained starting from the set of known protein coding sequences and then looking for paralagous alignement. Our database is the first one obtained by directly starting from the DNA sequence and is thus completely independet from the functional annotation of the organism.

Using as reference the annotations reported in Ensemble and in the UCSC browser we were able to obtain a list of new alternative splicing versions of the known genes and a list of new candidate genes both in human and in mouse. We are presently performing the experimental validation of a few of these candidates.

Email: molineri@to.infn.it