Identification of candidate regulatory sequences in mammalian 3'-UTR regions by statistical analysis of oligonucleotide distributions

ID - 120

Cora' Davide[1], Di Cunto Ferdinando[2], Caselle Michele[1], Provero Paolo[2]
[1]Department of Theoretical Physics, University of Torino, Torino
[2]Department of Genetics, Biology and Biochemistry, University of Torino, Torino

Motivation

3'-UTR regions contain binding sites for several regulatory elements which play an important role in the post-transcriptional control of gene expression, regulating mRNA stability, localization and translation efficiency.

In particular, a mechanism of post-transcriptional regulation whose importance was realized in the last few years is mediated by a class of small RNAs called micro-RNAs.

With this abstract, we present a computational methodology aimed at the identification of putative regulatory elements in human 3'-UTR regions, with a focus for miRNA binding sites.

Methods

We propose and develop two complementary approaches to the statistical analysis of oligonucleotide frequencies in mammalian 3'-UTR regions aimed at the identification of candidate binding sites for regulatory elements. The first method is based on the identification of sets of genes characterized by evolutionarily conserved overrepresentation of an oligonucleotide, without requiring any alignment procedure. The second method is based on the identification of oligonucleotides characterized by statistically significant strand asymmetry in their distribution in 3'-UTR regions.

More precisely, we analyzed repeat-masked 3' UTR sequences of human and mouse genes using two different pipelines, both based only on the statistical properties of oligonucleotide frequencies:

- Conserved overrepresentation. We constructed, separately for human and mouse, sets of genes sharing overrepresented oligonucleotides. We then selected those oligos whose sets of genes in human and mouse contained a statistically significant fraction of orthologous genes. Oligos selected in this way are thus characterized by an evolutionary conserved overrepresentation in the 3'UTR sequences of selected sets of genes and can thus be considered as good candidate binding sites.
- Strand asymmetry. We identified those oligos whose frequency distribution shows a statistically significant strand asymmetry, that is a difference in frequency between the oligo and its reverse complement. Oligos which are binding sites of regulatory elements acting on single-stranded 3'-UTR sequences are expected to show such an asymmetry since, contrary to what generally happens for transcription factor binding sites, there is no functional equivalence between an oligonucleotide and its reverse complement.

Results

Concentrating on oligonucleotides of length 7, taken together, the methods identify a total of 610 7-mers as candidate binding sites.

Comparing the results with databases of known 3'-UTR regulators, we demonstrate that both methods are able to identify many previously known binding sites located in 3' UTR regions, in particular miRNA-binding sites.

Moreover, we obtain a subset of 59 7-mers showing strand asymmetry both in human and mouse and also identified by conserved overrepresentation.

We consider this last set of oligos as our best candidate binding sequences.

Detailed analysis on this subset of 59 oligos allowed us to identify 52 of these as known cis-acting elements. The remaining 7 are strong candidates to represent new cis-acting elements in mammalian 3' UTR and are promising candidates for experimental verification.

Availability: http://www.to.infn.it/ftbio/mirna_human/supplementary.html

Email: cora@to.infn.it