Discovery of conserved long non-coding RNAs in vertebrates

S. Basu, R. Sanges
Bioinformatics - Animal Physiology and Evolution, Stazione Zoologica Anton Dohrn, Napoli, Italy

Motivations
Long non-coding RNAs (lncRNA) have been reported as a major class of novel transcripts related to organism development and early neural expression pattern [1-4]. They are reported to be expressed in large numbers in the mammalian transcriptomes [5,6] and recently reported to be expressed in the teleost fishes [7,8]. Computational identification and characterization of lncRNA from public sequence resources have been performed by different groups [9-11]. The focus of attention has been on the mammalian genomes starting by the assumption that they are not well conserved in term of sequence. However, systematic studies measuring their levels of conservation among vertebrates are lacking. Hence we want to computationally evaluate the existence of vertebrate conserved IncRNAs through systematic conservation analyses of both sequence as well as genomic architecture.

Methods
Mouse IncRNAs reported in an earlier study [2] and predicted by the EnsEMBL pipeline were considered as a reference dataset. Homology search of the IncRNAs against the zebrafish conserved phastcons elements was performed with the BLAST program. The phastcons elements are regions of conservation in the zebrafish genome with human, mouse, western clawed frog and two teleost fishes, tetraodon and stickleback. The lack of selection pressure in IncRNAs as compared to the protein-coding genes required a calibration of BLAST parameters to define a cut-off score indicative of significant conservation. Using ROC analyses we calculated the best BLAST parameters able to select regions of IncRNA conserved in vertebrates. The predicted conserved candidates were also evaluated in terms of their RNA secondary structure using the RNAfold software. Gene ontology and expression pattern enrichment of flanking protein-coding genes was performed with DAVID software.

Results
Our results show that the usage of the alignment length as cut-off is sufficient to distinguish the conservation of mouse IncRNAs in zebrafish as compared to conservation of random genomic regions. The RNA secondary structure prediction was not able to define any threshold for conservation. From an initial dataset of ~2,800 IncRNAs we could predict that 235 are conserved using the defined cut-off on the alignment length. Gene ontology enrichment analyses, related to the protein-coding genes proximal to the region of conservation in mouse and zebrafish, highlighted corresponding GO classes such as regulation of transcription and central nervous system development. The proximal coding genes exhibited a similar enrichment for their tissue of expression where brain was highly enriched in both mouse as well as zebrafish. Two interesting candidate regions of conservation were chosen for future experimental validation based upon the presence of ESTs overlap and the function of the proximal proteins (in this case the interest being development and functioning of the nervous system). The analysis is poised as an initial pipeline to select interesting candidate IncRNAs conserved among vertebrates.
References


