


RESEARCH ARTICLE

Open Access



# Identification of long non-coding transcripts with feature selection: a comparative study

Giovanna M. Ventola<sup>1,2†</sup>, Teresa M. R. Noviello<sup>1,2†</sup>, Salvatore D'Aniello<sup>3</sup>, Antonietta Spagnuolo<sup>3</sup>, Michele Ceccarelli<sup>1</sup> and Luigi Cerulo<sup>1,2\*</sup> 

## Abstract

**Background:** The unveiling of long non-coding RNAs as important gene regulators in many biological contexts has increased the demand for efficient and robust computational methods to identify novel long non-coding RNAs from transcripts assembled with high throughput RNA-seq data. Several classes of sequence-based features have been proposed to distinguish between coding and non-coding transcripts. Among them, open reading frame, conservation scores, nucleotide arrangements, and RNA secondary structure have been used with success in literature to recognize intergenic long non-coding RNAs, a particular subclass of non-coding RNAs.

**Results:** In this paper we perform a systematic assessment of a wide collection of features extracted from sequence data. We use most of the features proposed in the literature, and we include, as a novel set of features, the occurrence of repeats contained in transposable elements. The aim is to detect signatures (groups of features) able to distinguish long non-coding transcripts from other classes, both protein-coding and non-coding. We evaluate different feature selection algorithms, test for signature stability, and evaluate the prediction ability of a signature with a machine learning algorithm. The study reveals different signatures in human, mouse, and zebrafish, highlighting that some features are shared among species, while others tend to be species-specific. Compared to coding potential tools and similar supervised approaches, including novel signatures, such as those identified here, in a machine learning algorithm improves the prediction performance, in terms of area under precision and recall curve, by 1 to 24%, depending on the species and on the signature.

**Conclusions:** Understanding which features are best suited for the prediction of long non-coding RNAs allows for the development of more effective automatic annotation pipelines especially relevant for poorly annotated genomes, such as zebrafish. We provide a web tool that recognizes novel long non-coding RNAs with the obtained signatures from fasta and gtf formats. The tool is available at the following url: <http://www.bioinformatics-sannio.org/software/>.

**Keywords:** lncRNA, Feature selection, Classification

## Background

The recent advances in whole transcriptome sequencing offers new opportunities for discovering novel functional transcript elements. In past decades only 2% of mammalian genome have been identified as coding for

proteins, while it is now known that a significant amount of the genome can be transcribed into different families of non-coding RNAs (ncRNAs) [1]. Such a high amount of transcripts demanded for the development of methods able to detect functional ncRNAs, and, among them, long non-coding RNAs (lncRNAs) which have emerged as important regulators of gene expression at several levels [2]. lncRNAs have been described in all taxa including plants, animals, prokaryotes, yeasts, and viruses [3] and their sequence conservation is usually lower than that of coding RNAs. Historically, they have been classified

\*Correspondence: lcerulo@unisannio.it

†Equal contributors

<sup>1</sup>Department of Science and Technology, University of Sannio, via Port'Arso, 11, 82100 Benevento, Italy

<sup>2</sup>BioGeM, Institute of Genetic Research "Gaetano Salvatore", c.da Camporeale, 83031, Ariano Irpino (AV) Italy

Full list of author information is available at the end of the article