Classification of mRNA and ncRNA sequences: a study based on complex networks and filter by exclusivity

Fabrício Martins Lopes, Murilo Montanini Breve

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ, Universidade Tecnológica Federal do Paraná (UTFPR)

Abstract

The large volume of biological data generated in recent decades and the need to analyze them to produce consistent discoveries has led to the development of a new area of knowledge, Bioinformatics. This is an interdisciplinary area that presents several important computational challenges related to Biology, one of them is the need to distinguish mRNAs and ncRNAs effectively. The correct identification of these RNA sequences is important due to the existence of thousands of non-coding RNAs, whose function and meaning are not yet known, as well as the challenge of understanding their genetic expression and possible regulatory action. This work adopts the complex network theory, which is being successfully used in many problems and, in several contexts. Thus, the proposed method consists of generating a complex network for each sequence of RNA, and then applying a filter in order to select the most exclusive edges of each class, increasing the distinction between the networks. For each filtered network, some topological measurements were extracted that are organized in a database, and thus used to classify each sequence in mRNA or ncRNA. For this purpose, the Random Forest classification method was adopted in the R project. Experiments were carried out to evaluate the proposed method considering a data set with six different species and comparing its acertivity with relevant methods in the literature such as CPC1, CPC2 and PLEK. The results indicated a high distinction between the RNA classes due to the filtering of the exclusive edges, which allowed the proposed method to reach average rates of accuracy higher than 98% in the mRNA and ncRNA classification considering all the adopted species. Finally, the proposed method presented less variations in its results when compared to the competing methods, indicating its adequacy for the classification of the RNA sequences. The application and a more in-depth study of this method can lead to a better understanding of non-coding RNA structure and as results its function.

Funding: Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)

Link to Video: