Distinguishing coding and non-coding RNA sequences and improving its functional annotation using machine learning approaches

Thaís de Almeida Ratis Ramos[1], Daniel Miranda de Brito[2], Raul Arias-Carrasco[3], Leonardo Vidal Batista[2], Thaís Gaudencio do Rêgo[2], Vinicius Maracaja-Coutinho[3,4,5]

[1]Universidade Federal do Rio Grande do Norte, [2]Universidade Federal da Paraíba, [3]Universidad Mayor, [4]Instituto Vandique, [5]Beagle Bioinformatics

Non-coding RNAs (ncRNAs) are important players in the cellular regulation in organisms from all domains of life. Its investigation is already routine in every transcriptome or genome project. Two key steps on the predicting process and functional assignation of ncRNAs, are (i) the ability to distinguish coding and non-coding sequences, followed by (ii) a functional assignation of RNA families based on sequences similarity searches or secondary structure predictions. Here, we applied different machine learning approaches in order to distinguish coding and ncRNA sequences, and to functionally predicted ncRNAs into known RNA families. The coding potential prediction was developed using different randomly selected sets of ncRNA sequences, extracted from Rfam database; and human RefSeq protein coding genes. Coding and ncRNAs had their tri-nucleotides counts analyzed using three different equally divided sets of 200, 400 and 1000 instances. For the functional classification of ncRNAs, we performed multiple alignments using sets of (i) 100 sequences and secondary structure models (SSMs) in Dot-Bracket Notation from 10 different families; (ii) 200 sequences and SSMs from 20 families; and (iii) three different sets of 500, 1000 and 2000 ncRNA sequences from 50 families. All RNA families were randomly selected from Rfam. Sequences and SSMs were filtered using a maximum similarity cutoff between them of 80% (Levenshtein distance); and a maximum length of 400nt. Then, we analyzed the counts of mono-, di- and tri-nucleotides on the primary sequences. Next, multiple alignments were performed using Clustal Omega and MARNA, respectively for primary sequences and SSMs. Finally, different classification tests were performed, using Naive Bayes, SMO, IBK, Multilayer Perceptron and Random Forest through WEKA tool. The coding potential evaluation using 200, 400 and 1000 sequences presented accuracies reaching 99%, 99% and 99.2%, respectively. The functional assignation of ncRNAs using 10 and 20 families, revealed results with an accuracy reaching 99% and 98.5%, respectively. Tests performed using 50 ncRNA families, resulted in an accuracy of up to 94.2%. These results outperforms predictions available in literature, which used a maximum of 25 RNA families. Tests were also performed in order to predict new ncRNAs families in different transcriptome data, opening new opportunities for the development of novel tools for nucleotides coding potential prediction and for the functional classification of ncRNA sequences. Future directions consists on the evaluation of our methodology performance using different sets of specie-specific nucleotide sequences and SSMs.