Extraction of features using topological measures of complex networks

Isaque Katahira[1], Eric Augusto Ito[1], Fábio Fernandes da Rocha Vicente[1], Fabricio Martins Lopes[1]

1 FEDERAL TECHNOLOGICAL UNIVERSITY OF PARANÁ

Abstract

The feature extraction methods are important for the study of the large amount of data produced by the high-performance sequencing techniques. The dimensionality reduction methods have been used to summarize the most significant characteristics of a data source. The goal is to represent a great volume of data from its characteristics, minimizing the information loss. Thus, the current research proposes a model of feature extraction based on the theory of complex networks for the representation of biological sequences. The proposed model consists on sequence mapping in graphs, in which the vertices are the segments of a sequence and the edges are defined by their structural organization (neighborhood). These edges are weighted by the pair occurrence frequency of adjacent segments in the input sequence. Then, topological measures of graphs are extracted: motifs, degree, minimum degree, maximum degree, standard deviation, cluster coefficient, average path length, proximity and intermediation. These measures compose a feature vector that represent a sequence, which is used to classify the input sequences. Coding and non-coding transcripts of nine species were used in order to verify the suitability of the proposed method, using the algorithms Random Forest, Naive Bayes, LibSVM and J48. A 10 fold cross-validation was performed to evaluate the predictors. The maximum accuracy for the coding transcripts identification was reached by the LibSVM with 100%; followed by J48 with 99.43%; Random Forest with 99.38%; and Naive Bayes with 93.64%. The best index related to the accuracy for the identification of non-coding transcripts was reached by Naive Bayes with 93.12%; followed by Random Forest with 81.43%; J48 with 79.41% and LibSVM with only 6%. The predictor that obtained the best accuracy average between the classification of coding and non-coding was the Naive Bayes with 93.38%. The results indicate the validity of the proposal, considering that the extraction of topological characteristics of complex networks got significant values ??of accuracy, which can be extended to the classification of other biological sequences like DNA and amino acids.

Funding: CAPES, Fundação Araucária