An Alternative Splicing Predictor in C.Elegans Based on Time Series Analysis

ID - 209

Ceccarelli Michele[1], Antonio Maratea[1]

[1]Research Center on Software Technologies, University of Sannio, Benvento, Italy

Motivation

Prediction of Alternative Splicing (AS) has been traditionally based on expressed sequences' study, helped by homology considerations and the analysis of discriminative features whithin the splice site. More recently, ab initio Machine Learning algorithms based only on the pre-mRNA sequence have shown good performances as predictors of some forms of alternative splicing, reducing the need to consider explicitly homology between species, but nonetheless giving special attention to the structure of the splice site. These approaches are based on ad hoc kernels aimed on one side at extracting implicitly a set of features from the variable-length genomic sequence and on the other side at the classification of sequences by evaluating carefully positions in the splice site. We show that it is possible to obtain results similar or better than the state of the art without any explicit modeling of positions in the splice site, nor any use of other local features. As a consequence, our method has a better generality and a broader and simpler applicability with respect to previous ones.

Methods

We approach the problem of ab initio AS prediction as a variable length sequence classification task and we extract a fixed-length set of features from the sequences to be used in the classifier. The proposed method can be resumed in three steps: i) coding of each exonic or intronic sequence; ii) feature extraction from coded sequences; iii) training a classifier on the features' vector previously obtained. Coding is done trivially substituting each of the four DNA bases A,C,G and T with one integer number from 1 to 4.

Feature extraction is done choosing the parameters of an Auto Regressive (AR) model as good descriptors of the dynamic of the phenomenon. In this way, each observed coded exonic or intronic sequence S is assumed to be the output of an order p AR model driven by a white noise process e(n); in other words the value in position n is assumed to be equal to a weighted sum of the p previous values plus a white noise term. To estimate model parameters we used the classical Yule-Walker method, actually minimizing the forward prediction error in a Least-Squares sense. As classifier we consider Support Vector Machine (SVM). It is a classical technique for Pattern Recognition and Data Mining classification tasks that aims to find a linear surface that splits the data in two groups according to the indicated labels, maximizing the margin, that is the distance from both sets of points. In this work we used a Gaussian (RBF) kernel.

Results

The labeled dataset used to test our method is a collection of 487 Exons for which EST show evidence of alternative splicing and 2531 Exons for which ther's no evidence of alternative splicing, for a total of 3018 labeled examples. All data regard C.Elegans and were obtained from the Wormbase, dbEST and UniGene. As this dataset was biased towards non splicing sequences, we generated a new dataset of alternative splicing Exons resampling 5 times the original one of 487 alternative splicing sequences and obtaining a total dataset of 4966 samples. To tune each parameter (AR model order, RBF width) we used 5 fold cross validation. We reapetedly splitted randomly the data in five blocks and used in turn four of them as the Trainig Set and the fifth one as Testing. Performances were evaluated in terms of the average AUC (Area Under Curve) index of Receiver Operating Characteristic (ROC) Curves. The method reaches an average AUC of over 91\% on Testing sets. The corresponding ROC curve does not rise immediately but has a shift from the y axis due to method's failure to classify a few high scoring points from the SVM. Ranking sequences on the base of the SVM predicted value, we noted that the top misclassified sequences at each run tend to be conserved. This fact suggests the opportunity that further biological verification is performed on these sequences, because their labeling, naturally prone to errors, may be wrong.

Email: ceccarelli@unisannio.it