Increasing the Perfomance of Splice Site Prediction: A Feature Selection Approach

Michiel Van Bel[1,2], Yvan Saeys[1,2] and Yves Van de Peer[1,2]

[1] Department of Plant Systems Biology, VIB, Technologiepark 927, 9052 Gent, Belgium

[2] Department of Molecular Genetics, Ghent University, Gent, Belgium

E-mail : Michiel.vanbel@psb.ugent.be

Splice Site Prediction

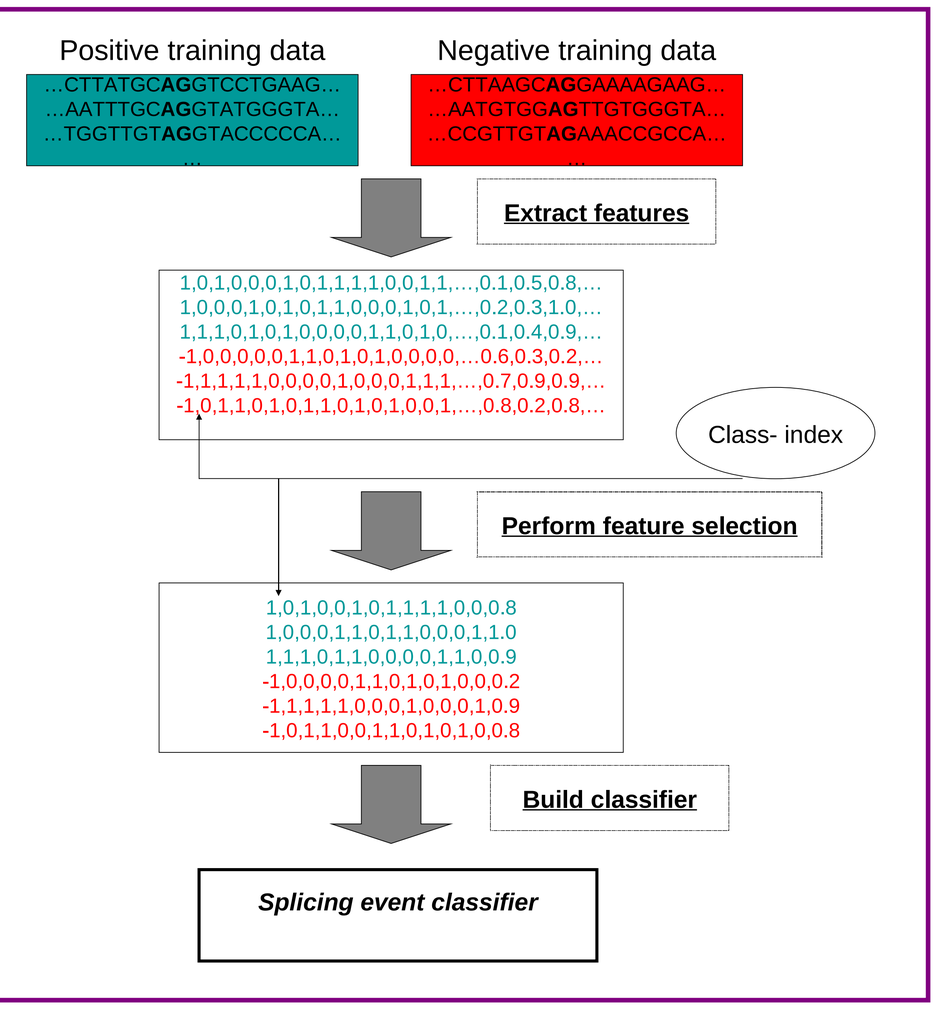
Our approach to splice site prediction involves the extraction of a high-dimensional feature vector from the local context around splice sites, and using these features to train a Support Vector Machine (SVM) classifier to create a splice site model for a particular class of splice sites (e.g. donor sites and acceptor sites).

The extracted features involve both positional based features and occurrence based features. By varying both the lengths of the features and the frame size from which these feature are extracted, we can perform a first optimization of the predictive Performance of the SVM models.

Feature Selection

In order to further increase the performance of the created SVM models, we extended the program with the capabilities to handle multiple types of feature selection algorithms. Applying these algorithms as filters for the extracted features, we hope to improve the predictive performance of the created SVM models.

We have chosen to use four univariate feature selection algorithms, and two multivariate feature selection algorithms. The results of applying these algorithms are compared to the baseline performance of the SVM's (with optimized lengths and frame sizes for all features). Because the optimization of the frame sizes may result in the loss of useful features, we also included the results of a single feature selection algorithm that was applied on semi-optimized data.



Feature Selection Results

	Arabidopsis				Human			
	Donor		Acceptor		Donor		Acceptor	
FS Algorithm	#Attributes	F1-Score	#Attributes	F1-Score	#Attributes	F1-Score	#Attributes	F1-Score
Baseline	22264	0,8664	23144	0,8313	14664	0,8664	14984	0,8502
Symmetrical Uncertainty	5721	0,905	5996	0,8547	3593	0,8713	5173	0,8697
Information Gain	5720	0,8983	5995	0,85	3592	0,8731	5172	0,8718
Gain Ratio	5721	0,9013	5996	0,8529	3593	0,872	5173	0,8722
Chi Square	5720	0,8971	5995	0,8545	3592	0,8746	5172	0,8752
FCBF	119	0,7954	125	0,7693	77	0,7862	111	0,7863
CFS	11	0,6483	106	0,7918	12	0,7487	164	0,7948
Semi-range Symmetrical Uncertainty	6486	0,9092	6526	0,8585	5561	0,8804	7182	0,8788

Methods

1. All results were acquired by applying 10-fold cross validation to the training sets.
2. All test sets consisted of 1000 positive examples and 10000 negative examples (in order to compensate for the overrepresentation of pseudo splice sites in DNA-sequences.
3. In order to compensate for a possible bias in the training examples, we choose to randomly extract the data from a larger set of training examples.
4. Every 10-fold cross validation was done 10 times with a new random extraction of data, in order to further minimize the risk of having a bias. The final result is the mean of these 10 randomizations.

Discussion Results

1. It is clear that the univariate techniques perform better than the baseline classifier, while the multivariate techniques perform much worse than the baseline classifier. The rather minimal amount of retained features with the multivariate algorithms points to the fact that these selection algorithms add too little features to the optimal subset.

2. The poor performance of the multivariate algorithms does not mean that there are no real dependencies between various parts of the DNA-sequence around the splice sites. Rather, it is a sign that the current types of feature extraction fail to capture these dependencies. Work on identifying these dependencies is a topic of current research.

3. The computational time for the different feature selection algorithms was not recorded, but it was clear that the multivariate algorithms took a longer time to complete.

4. The test to gain better results by skipping the first optimization (Semi-Range Symmetrical Uncertainty) yields some better results. However, these increases In predictive performance are minimal and they are the result of a substantial Increase in the number of retained features, thus leading to a slower computation.

References

1. Degroeve et al. **Predicting splice sites from high-dimensional local context representations.** Bioinformatics 21(8),1332-1340 (2005)

2. Saeys et al. **A review of feature selection techniques in Bioinformatics.** Bioinformatics (2007) (In Press).

3. Yu et al. **Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution.** ICML-03, 856-863