Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

SpliceIT: A hybrid method for splice signal identification based on probabilistic and biological inference

Andigoni Malousi [a,*], Ioanna Chouvarda [a], Vassilis Koutkias [a], Sofia Kouidou [b], Nicos Maglaveras [a]

[a] Lab. of Medical Informatics, Medical School, Aristotle University of Thessaloniki, Greece
[b] Lab. of Biological Chemistry, Medical School, Aristotle University of Thessaloniki, Greece

ARTICLE INFO

Article history:
Received 26 October 2008
Available online 30 September 2009

Keywords:
Splice sites
Predictive modeling
Feature selection
Biological inference
Hybrid classification

ABSTRACT

Splice sites define the boundaries of exonic regions and dictate protein synthesis and function. The splicing mechanism involves complex interactions among positional and compositional features of different lengths. Computational modeling of the underlying constructive information is especially challenging, in order to decipher splicing-inducing elements and alternative splicing factors. SpliceIT (Splice Identification Technique) introduces a hybrid method for splice site prediction that couples probabilistic modeling with discriminative computational or experimental features inferred from published studies in two subsequent classification steps. The first step is undertaken by a Gaussian support vector machine (SVM) trained on the probabilistic profile that is extracted using two alternative position-dependent feature selection methods. In the second step, the extracted predictions are combined with known species-specific regulatory elements, in order to induce a tree-based modeling. The performance evaluation on human and *Arabidopsis thaliana* splice site datasets shows that SpliceIT is highly accurate compared to current state-of-the-art predictors in terms of the maximum sensitivity, specificity tradeoff without compromising space complexity and in a time-effective way. The source code and supplementary material are available at: http://www.med.auth.gr/research/spliceit/.

1. Introduction

Pre-mRNA splicing is an essential step in gene expression, involving an RNA modification during which introns are excised in a two-step enzymatic procedure. In the first step, the adenosine corresponding to the branch site of the polypyrimidine track that precedes an acceptor splice site bonds covalently to the guanosine at the donor splice site. The second step involves the pairing of adjacent exons and the excision of the inner intron that is then degraded in the cell nucleus and the splicing product moves from the nucleus to the cytoplasm. In splice site forms, GT and AG dinucleotides signal the beginning and end of an intron, respectively. The canonical GT/AG splice site rule dominates on the overwhelming majority of splice sites in different species, e.g. more than 98% of confirmed human splice sites follow the canonical GT/AG splice site rule [1].

This strong conservation observed in splice junctions is not sufficient to accurately locate a splice site, due to the huge number of GT/AG-containing sequences and thus of false positive cases. To cope with this issue, a larger consensus sequence exhibiting weaker conservation is often modeled to discriminate an actual splice site from splice-like signals. As splice site identification is used to computationally localize protein-coding sequences within an uncharacterized DNA segment, being able to locate actual GT/AG splicing pairs is an important issue, in order to increase the predictive accuracy of whole gene sequences [2]. In addition, more accurate splice site predictions imply higher sensitivity to whatever positional variations are observed in their locality.

Splice site prediction has been elaborated by various computational techniques so far. Position-specific weight matrices (PWMs) and weight array models (WAMs) of various orders have been applied formerly for splice site prediction [3,4]. Over time, more sophisticated methods have been proposed that significantly increase the predictive power. For example, NNSplice employs a feedforward neural network with one hidden layer to identify splice sites [5], while Loi and Rajapakse introduced a hybrid method that combines Markov models and neural networks [6]. In addition, DGSplicer employs a dependency graph model to fully capture the intrinsic interactions among nucleotides in the locality of splice sites [7], while GeneSplicer combines Markov modeling with a maximal dependence decomposition method in order to capture the most significant dependencies among adjacent and non-adjacent residues [8]. Furthermore, the use of

[*] Corresponding author. Address: Lab. of Medical Informatics, Medical School, Aristotle University of Thessaloniki, Thessaloniki, Greece.
E-mail address: andigoni@med.auth.gr (A. Malousi).

1532-0464/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2009.09.004

decision trees has been proposed by Thanaraj and Robinson to build discriminative models between real and pseudo splice sites [9].

Support vector machines (SVMs) have also been employed for splice site identification. For example, the performance of a Bayesian feature mapping that fed a linear SVM proposed in [10] was fairly robust, when applied to large volumes of data. In addition, 1st order Markov descriptions of the input dataset with an SVM classifier using a polynomial kernel has been applied for splice site prediction in [11,12], while Zhang et al. used SVMs to extract the classification rules that best discriminate real from pseudo exons using a degree-2 polynomial kernel on different feature type combinations [13]. Likewise, SpliceMachine implements an efficient method for predicting splice sites that selects and merges positional and compositional features that are learned by a linear SVM classifier [14]. In practice, SVM learning has been proven to be highly accurate in various other biological classification, regression and novelty detection problems [15–18].

Recently, an alternative approach compared to such machine learning strategies has been proposed by Trapnell et al. for the ab initio identification of splice junctions relying on a novel similarity-based mapping algorithm that aligns short reads from RNA-Seq experiments against the whole reference genome [19].

Features encoding and selection in the abovementioned machine learning approaches play an important role in the classification performance. Most often, binary classifiers are combined with features representation and selection techniques by employing a pre-processing decision making on the type and number of most informative features, i.e. those preserving the underlying information of the learning problem. In this context, a number of alternative modeling techniques have been proposed, such as permuted variable length Markov model (PVLMM) for the identification of transcription factor binding sites and splice signals [20], the optimized mixture of Markov models (OMiMa) for modeling dependence structures within biological motifs [21], and generalizations of standard Markov models to characterize biological sequences [22]. As far as feature selection is concerned, Degroeve et al. proposed a wrapper-based feature selection method for splice site prediction that improved the classification performance compared to the use of all available features [23], while Saeys et al. applied a fast detection of relevant feature subsets using a heuristic method based on the estimation of distribution algorithms [24]. Both methods when combined with SVMs gave superior results in terms of accuracy and time-efficiency. Finally, Chen and Lin proposed alternative feature selection and ranking strategies, namely, *F*-score and random forest, that are well-suited for binary SVM-based classification problems [25].

The present work proposes a hybrid method for splice site prediction, following some logical hypotheses not previously considered in other computational methods. The first hypothesis involves the type of features that should be incorporated. Recently, several experimental or computational studies on the splicing-inducing factors have highlighted the importance of specific positional and compositional features that discriminate actual splice site from decoys [26–28]. These features have been associated with the presence of splicing regulatory elements and therefore could be important in predictive modeling. The idea in this case is that, instead of performing exhaustive searches for oligomers with high discriminative power on our dataset, we can exploit available evidence inferred from published computational or experimental studies. In this context, it is self-evident that feature extraction is radically more time-efficient than selecting and modeling features from scratch, and in fact more generic, since these features stem from various studies applied on different splice datasets. In addition, due to the limited number of known features, it is feasible to manually decide on the encoding scheme that is more suitable. In this work, two encoding types are used, namely local context (LC) and weighted distribution (WD). It has to be noted that the selected features are species-specific and different for donor and acceptor sites.

A second hypothesis examined in this work is that, although the aforementioned evidence-based features are highly informative, they are not sufficient to delineate the whole splice sequence profile and cannot fully capture the positional information of the sequence residues. To deal with this issue, we integrated the probabilistic profile of the splice residues and trained them independently in a preceding classification step. The extracted probability estimates [29] together with the evidence-based features discussed in the first hypothesis constitute the feature set in the subsequent training procedure.

A final motivation of this work involves the order of the probabilistic modeling that should be selected, i.e. the dependency length among residues that best discriminates positive instances from decoys. Generally, higher order Markov models perform better than low order Markov models at the expense of the state complexity [30], and reduced generalization performance [31]. On the other hand, low order Markov models do not look far into past events; nevertheless, they are often preferred since they require less training data, they are less state demanding, and often perform better on unseen data. Most techniques describing the positional content of a splice sequence use a single type of signal interactions in form of fixed-order Markov models [5,11]. The selection of the dependency length in these studies is poorly justified and the induced model partially captures the positional properties. A rather simplistic and straightforward solution to this problem would be to extract multiple orders of positional array models (WAM-*k*) in the "all-*k*th-order" feature representation [32]. This approach increases clearly the space complexity and most importantly has no or even negative influence on the prediction outcome, due to the abundance of redundant features [31].

In this work, we investigate the performance of two methods for selecting probabilistic features, that are alternatively used, namely the positional feature selection (PFS) [33] and the principal feature analysis (PFA) [34]. PFS selects the most informative positional description per residue according to specific optimality criteria, while PFA exploits the mutual information among residues and selects a subset of probabilistic parameters (principal features) of different orders following a PCA (principal component analysis) based selection method. PFA allows for a specific position to be multiply represented by different orders, while PFS associates a unique probabilistic parameter with each position residue having no additional cost on the space complexity, compared to individual positional models.

Following these hypotheses, we developed a hybrid splice site predictor, called SpliceIT (Splice Identification Technique). SpliceIT uses a Gaussian SVM to classify the PFS or PFA-based probabilistic descriptions used in the first classification step and a binary decision tree for the classification of the additional evidence-based features in the second classification step. In the following, we use the term *probabilistic sequence features* to refer to the Markov features employed in the first classification step and the term *evidence-based features* for the sequence motifs used in the second classification step.

2. Materials and methods

2.1. Evaluation datasets

SpliceIT was evaluated on 1115 human and 1323 *A. thaliana* non-redundant genes that were first used to build predictive models by GeneSplicer [8]. The training sequences make up a realistic