Journal of Biomedical Informatics 43 (2010) 208–217

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

splice dataset that allows for different window sizes adjacent to splice junctions to be incorporated in the classification scheme. Of the 315,615 splice-like donors and 417,939 acceptor sites contained in the *A. thaliana* genes a subset of 10,000 false splice sites were randomly selected to optimize the SVM training parameters. Similarly, 1000 donor and 1000 acceptor sites were randomly selected out of the 5440 and 5488 actual splice sites contained in the *A. thaliana* dataset, respectively. As regards the human splice sites, the same subset size was used to find the optimal parameters of the Gaussian SVM classifier out of the 5733 actual donor and acceptor sites and the 478,983 and 650,099 false donor and acceptor sites, respectively. All splice sites (actual and pseudo), follow the canonical GT/AG splice site rule.

A region of 50 nt upstream the splice junction and equal size of nucleotides downstream the consensus GT/AG sites were extracted from the actual and pseudo splice datasets for both human and *A. thaliana* genes. The results of the optimization step were used in the training procedure that was applied on a larger dataset containing 5000 true and 50,000 false donor and acceptors each, that were randomly selected from the initial human and *A. thaliana* datasets.

Aiming to assess the reproducibility and consistency of the results, we performed an additional evaluation on the NN269 dataset. The NN269 dataset is a compilation of human splice sites extracted from 269 genes (Genbank v.95) that was first used to train the NNSplice predictor [5,35] and in various other studies as benchmark dataset [2,12,11,15]. The NN269 dataset is a well-documented and widely-known benchmark dataset that contains 1324 sequences of actual donor and acceptor sites following the canonical GT/AG splice site rule. The pseudo splice site training and test datasets contain 4922 donor and 5553 acceptor sites. Donor sequences (5256 training and 990 test) contain 15 nucleotides with the splice junction located at positions 8 and 9. Similarly, the AG consensus is located at positions 69, 70 within acceptor sequences (5788 training and 1089 test) of 90 nt long.

2.2. Methodology overview

The predictive modeling employed in SpliceIT addresses several important aspects such as: (1) appropriate features encoding scheme, (2) feature selection/ranking method, and (3) parameters optimization. The basic subsequent processing steps are outlined in the following:

1. *Feature extraction*: Positional probabilistic descriptions of different orders are constructed and a pool of candidate features is generated.
2. *Feature selection*: The discriminative power of each feature is assessed and the most informative features are selected using either PFS or PFA.
3. *First classification step – SVM:* The SVM classifier is trained on the probabilistic parameters.
4. *Additional evidence-based feature encoding:* A set of evidence-based features is selected and appropriately represented.
5. *Second classification step – decision trees:* Ambiguous predictions are re-considered in a tree-based modeling based on the features obtained in the former step.

2.3. Feature extraction

Given a sequence of random variables $S = S_1, S_2, \ldots$, a Markov model is used to capture the inter-dependencies among successive states in order to extract a set of probabilistic features [36]. In a 1st order Markov chain model, for example, the transition probability of a state x to y at position i of a sequence S is defined as:

$$p_{x,y}(i) \triangleq P(S_{i+1} = y | S_i = x). \quad (1)$$

The resulting model assigns different transition probabilities for each position. PWMs that define the probability of a nucleotide to be observed at a specific position are equivalent to positional zero order Markov models. Likewise, WAMs correspond to non-zero order positional Markov models.

SpliceIT employs PWMs and WAMs to build the probabilistic feature set. In order to bypass zero probabilities, which are more frequently observed in higher order WAMs, we apply a simple smoothing method that adds a pseudocount to each actual probability. The so-called Laplace rule has its theoretical justification coming from the probability theory and is defined (assuming a prior uniform nucleotide distribution) as $p_k^L = \frac{f_{s_i}+1}{f_{s_i}+N}$, where N is the number of alphabet letters and f_{s_i} is the frequency of the nucleotide at position i following the k-mer sequence of frequency f_s.

2.4. Feature selection

Feature selection is a particularly important pre-processing step that is commonly used in pattern classification techniques, aiming to address the dimensionality reduction problem, reduce storage space and classification time, and improve the understanding of the problem and interpretability of the results [37,38].

Typically, feature selection methods are mostly applicable to problems where hundreds or thousands of features have to be considered. In this study, the feature space is not that complex; nevertheless, feature selection serves as a necessary mechanism that filters out redundant features and provides to some extent biological interpretation of the incorporated features. In this context, feature selection is employed in two alternative ways:

- decide on the best-fitting feature per position using PFS [33], or
- prune the original feature set to an optimal subset that retains or increases the classification performance using PFA [34].

A detailed analysis of the feature selection strategies employed is described in the following.

Positional feature selection (PFS). PFS defines the probability of observing a state at a certain position by selecting the best-fitting order of past events, i.e. the most informative sequence of preceding states for each residue [33]. The generative model uses the *F*-score value as a selection criterion for the best-fitting feature per site [25]. In a binary classification problem, *F*-score gives a measure of the discriminative power of each feature, using the averaged values of the positive and negative instances, and assigns a numeric value to each one of the features. Larger *F*-score values indicate more informative features.

Given a set of n_p and n_n positive and negative training vectors, respectively, the *F*-score of the ith feature corresponding to the kth order probabilistic parameter is defined by the formula:

$$F_k(i) \equiv \frac{\left(\bar{x}_k^p(i) - \bar{x}_k(i)\right)^2 + \left(\bar{x}_k^n(i) - \bar{x}_k(i)\right)^2}{A_p + A_n},$$
$$\text{with } A_m = \frac{1}{n_m - 1} \sum_{j=1}^{n_m} \left(x_k^m(j,i) - \bar{x}_k^m(i)\right)^2, \quad (2)$$

where $\bar{x}_k(i)$, $\bar{x}_k^p(i)$, $\bar{x}_k^n(i)$ are the average values of the total, positive and negative instances assigned to the ith feature, respectively, while $x_k^p(j,i)$ and $x_k^n(j,i)$ denote the probability estimates of the ith feature corresponding to the jth positive and negative instances, respectively.

For each Markov model, the corresponding F_k vector is defined as:

$$F_k = \{F_k(k+1), F_k(k+2), \ldots, F_k(N)\}, \quad (3)$$

with $k \geqslant 0,\ k+1 \leqslant i < N$, where N is the sequence length. For each order k of conditional probabilities, the corresponding F-score vector of $N-k$ length is generated. Given a kth order Markov model, the extracted feature vector of a training instance is defined as:

$$X_k = \{x_k(k+1), x_k(k+2), \ldots, x_k(N)\}, \quad 0 \leqslant k < N-1. \tag{4}$$

If $Y = \{y_1, y_2, \ldots, y_N\}$ is the vector of the PFS encoding for an instance, then y_i is given by the formula:

$$z = arg\max_k F_k(i), \quad y_i = x_z(i), \tag{5}$$

with $k \leqslant i$ and $1 \leqslant i \leqslant N$. PFS has linear computational cost with the number of features at the expense of discarding the mutual information among features. PFS also assumes that all features selected by the F-score criterion are equally important.

Pruning with principal feature analysis (PFA). PFA is an alternative approach that incorporates features' mutual information. Evidently, some features may have no or even negative effect on the classification performance. PFA is a variant PCA method that is used to cut out noisy and redundant variables [34]. PFA deals with the dimensionality reduction problem in a computationally cost-effective way by choosing a subset of the original feature vector that retains the underlying discriminative information using the same optimality criteria as in PCA. However, instead of finding a projection of all features included to the original feature space to a lower dimensional space, PFA exploits the properties of the principal components to select a subset of the original features. Contrary to PFS, PFA takes into account the mutual information among the selected features. In this study, the source features are the variable-order WAMs and the outcome is the principal feature subset that efficiently characterizes the initial pool of probabilistic parameters. To provide additional confirmatory evidence, the extracted components are independently studied for their statistical significance by performing the Wilcoxon rank sum test ($p < 0.05$). The Wilcoxon rank sum test for equal medians gives an insight of the differences between positive and negative instances of each feature and does not imply any assumption on the distribution of the tested features.

2.5. First classification step-support vector machines

SVMs are discriminative learning methods that are used for supervised learning in classification and regression problems. In binary classification, SVMs are used to learn a classification rule from labeled data by generating a hyperplane between the two classes that maximizes the margin to the closest points, called support vectors [39,40]. SVMs apply an implicit mapping Φ of the input data into a high-dimensional feature space using a function:

$$k(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle, \tag{6}$$

that gives the inner product and can be computed without having to explicitly project x_i and x_j into the feature space.

Non-linear SVMs use a kernel function instead of the inner products defined in Eq. (6). Considering their theoretical justification, SVMs are computationally very effective, since no computations are performed in the high-dimensional space that is used to map the input data. Moreover, different kernel methods can be designed and applied in SVM classification so that the decision function is interpretable and useful to extract biological knowledge [41,42].

SpliceIT employs the radial basis function (RBF) $K(x_i, x_j) = e^{-\gamma \|x_i - x_j\|^2}$, to map input data points into the feature space. RBF kernels give good generalization performance, need less hyperparameters than the polynomial kernel, and are considered as a generalization of the linear kernel function. The penalty parameter C that defines the tradeoff between training errors and stringent margins [40] and the γ value of the RBF kernel are user-defined and need to be fine-tuned in order to minimize the classification error. In this context, a grid-based hyperparameter selection is used that performs an exhaustive search in a predefined logscaled grid space, where each pair of parameter values is evaluated in terms of the overall classification accuracy. Grid search is computationally intensive compared to heuristic approaches; however, the latter do not guarantee the selection of the optimal C and γ pair within a certain range of values.

2.6. Additional evidence-based feature encoding

The features used in the subsequent classification step stem mainly from species-specific published studies and are described with the following encodings:

- weighted distribution (WD) of the conservation profile, and
- abundance of predefined oligomers (local context, LC).

WD is calculated by the weighting function `bin(code, index, dir)` that converts an oligomer sequence into a binary number and subsequently to the corresponding decimal number by applying a gradually vanished weighting as removed from splice sites where:

- `code` refers to the IUPAC symbols corresponding to the positively-weighted nucleotides,
- `index` gives the relative position with respect to the splice site, and
- `dir` is a binary variable that refers to the scanning direction either 5′ to 3′ (0) or 3′ to 5′ (1).

For example, `bin(Y, −3, 1)` returns the decimal number corresponding to the binary number that results by transforming the 3-mers preceding splice sites at the 3′ to 5′ direction, with $C, T \rightarrow 1$ and $G, A \rightarrow 0$.

LC measures the abundance of specific oligomers and is represented by their frequency in predefined regions adjacent to splice sites. Finally, the probability estimates provide a confidence metric of each prediction made by the SVM training procedure [29].

2.7. Second classification step – decision trees

Decision tree modeling in classification problems is often used to assess the importance of candidate features for discrimination and prediction [43]. One of its strengths lies in the interpretability of the constructed model that is especially useful when we need to ascribe particular meaning to the classification results. In this study, decision trees are employed in the second classification step, in order to train species-specific features that have been associated with the regulation of the splicing process in human and *A. thaliana*. Typically, a decision tree is gradually growing by splitting the class members according to some scoring criteria. SpliceIT uses maximum deviance reduction as a criterion for choosing a split [44]. In addition, to avoid limited generalization performance of an overgrown tree model, SpliceIT applies a post-pruning technique that shrinks the tree structure by turning a number of branch nodes into leaves[1] [44].

The number of the training instances in the tree-based modeling is defined by the probability estimates extracted by the SVM classifier [29]. All instances assigned with probabilities belonging to an ambiguous mid-region are re-considered in the tree-based modeling, while for the remaining instances no additional process-

[1] Implementation details available at: http://WWW.med.auth.gr/research/spliceit/.

ing is performed. The definition of this unclear region is based on the probability threshold that gives the optimal sensitivity and specificity tradeoff, i.e. the closest point of the ROC (receiver operating characteristic) curve from an absolutely well-trained SVM classifier. The amount of the training data is then defined by taking different probability ranges around this threshold. The overall classification performance is estimated by summing up the true positive and negative predictions of the SVM classifier for the instances that were not used in the tree-based training procedure and the corresponding true positive and negative predictions made by the induced decision tree. For each training dataset, the corresponding tree-based model is induced and the convex hull of all the sensitivity, specificity pairs is taken.

2.8. Performance measures – implementation issues

The classification performance of the proposed hybrid method is estimated based on the ROC curves and the associated area under ROC curves (AUC). ROC curves give a measure of the tradeoff between the false positive rate $FPR = 1 - Sp = \frac{FP}{FP+TN}$ and the true positive rate $TPR = Sn = \frac{TP}{TP+FN}$, where Sp, Sn correspond to specificity, sensitivity, respectively. $TP(TN)$ is the number of actual(pseudo) splice sites that are correctly predicted and $FP(FN)$ is the number of splice sites that are mistakenly labeled as actual(pseudo) splice sites. Accordingly, Sn is the probability of correctly predicting a positive instance and Sp is the probability that a positive prediction is correct. Correlation coefficient (CC) is also used as a comprehensive classification performance metric incorporating both sensitivity and specificity measures as defined by the formula [45]:

$$CC = \frac{TP * TN - FP * FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}. \quad (7)$$

CC ranges from -1 to 1, with 1 corresponding to a perfectly well-trained classifier. To estimate the optimal FPR and TPR pair, we used the Euclidean metric. Specifically, the best sensitivity, specificity tradeoff is defined by the coordinates of each point on the ROC curve with the minimum distance from a perfectly well-trained classifier, i.e. point $(0,1)$ on the false positive and true positive axes of the ROC plot, respectively.

To avoid deceptive comparisons on the maximum sensitivity and specificity, we tested equal number of points using cubic interpolation on the ROC curve defined by the sensitivity levels in Tables S1 and S2 (Additional file). Furthermore, to cope with overfitting the training data, a 10-fold cross-validation (CV) learning procedure is employed using LIBSVM [46]. Although not theoretically justified, 10-fold CV has been extensively used as a preferred way for measuring the error rate giving the best classification performance [47]. Tree pruning was also performed on 10-fold CV training data using MATLAB®. Finally, the best-fitting C and γ parameters of the SVM were identified within a range of $[-5,5]$ and $[-4,0]$, respectively, with step 1 in the logscale, using uniformly spaced sample values.

3. Results

3.1. Positional weight models

The performance of the Gaussian SVM classifier was tested on PFS and PFA using PWM, WAM-1 and WAM-2 descriptions of the 50 residues at either sides of the GT/AG consensus for the GeneSplicer datasets. PFS and PFA-based encodings offer two different position-dependent views of the same source of probabilistic parameters. PFS assumes that all residues are equally important and therefore need to be represented in the final encoding scheme. On the other hand, PFA goes deeper into investigating the local properties of the most informative residues (principal features) that can be multiply represented, while redundant residues are cut out.

Figs. 1 and 2 present the most informative features per position according to the weighting schemes followed in the PFS and PFA feature selection methods. In PFS, the most discriminative dependency length per position, i.e. with the highest F-score, is specified by different colors in the F-score chart. Evidently, the contribution of each residue diminishes as removed from splice sites; however, the distribution of the important features is differentiated between donor and acceptor sites. Acceptor prediction is mostly weighted by the intronic residues close to the AG consensus for both human and *A. thaliana*, while donor selection is mostly controlled by a smaller intronic and exonic region adjacent to the GT consensus. Interestingly, as shown in Figs. 1 and 2, positional WAM-1 in human and *A. thaliana* are barely selected by PFS to characterize exonic residues preceding donor sites and following acceptor sites. On the contrary, WAM-1 are more frequently selected in intronic regions compared to PWM and WAM-2.

It is interesting to note that, while in donors small differences can be identified in the conservation profile between human and *A. thaliana*, in acceptors the differences among the two species are considerable (Figs. 1 and 2). Specifically, in human WAM-2, which can detect longer preserved motifs, is suitable for identifying preserved patterns in exons. On the contrary, WAM-1 which identifies conserved dinucleotides appears to be more suitable for defining conservation in introns. Dinucleotide sequences have been mostly related to conservation and packing considerations [48]. In *A. thaliana*, conservation in exonic sequences also appears to be selected by more stringent sequence considerations, but it is only in certain intronic sequences where conservation is observed at the dinucleotide level. Contrariwise, conservation in *A. thaliana* is frequently described by nucleotide PWM models. The above data would indicate that specific intronic conformational characteristics are probably essential for the recognition of splice sites and that these are more variable among humans, particularly in introns compared to exons, possibly reflecting complex splice-site recognizing machinery (RNPs). On the contrary, in *A. thaliana*, site-specific residue recognition appears to be critical for splicing.

Figs. 1 and 2 also illustrate the number of principal features extracted for each position by applying the PFA feature selection. Grayscale bars are lighter in positions, where more than one probabilistic parameter is extracted. Generally, these positions match those with high F-score values, though a small number of high F-score positions are under-represented among the principal features and vice-versa. This is sporadically observed in human and *A. thaliana* splice features and is basically justified by the theoretical setting of the applied splice selection methods. The total number of features extracted is shown in Table 1. The number of principal features is moderately increased compared to individual fixed-order probabilistic parameters and PFS; however, the observed increase mostly characterizes human splice sites (25% increase in human vs. 13.7% in *A. thaliana*).

3.2. SVM classification

The training procedure of the first classification step begins with the parameter optimization step that comes up with the best-fitting hyperparameter pair for each Gaussian SVM classifier (Table S3). These values are used in the subsequent 10-fold CV training procedure. Table 1 lists the classification outcome in terms of the maximum CC value and the best sensitivity and specificity tradeoff. The latter corresponds to the minimum distance from the perfect classification. CC is the maximum value obtained and

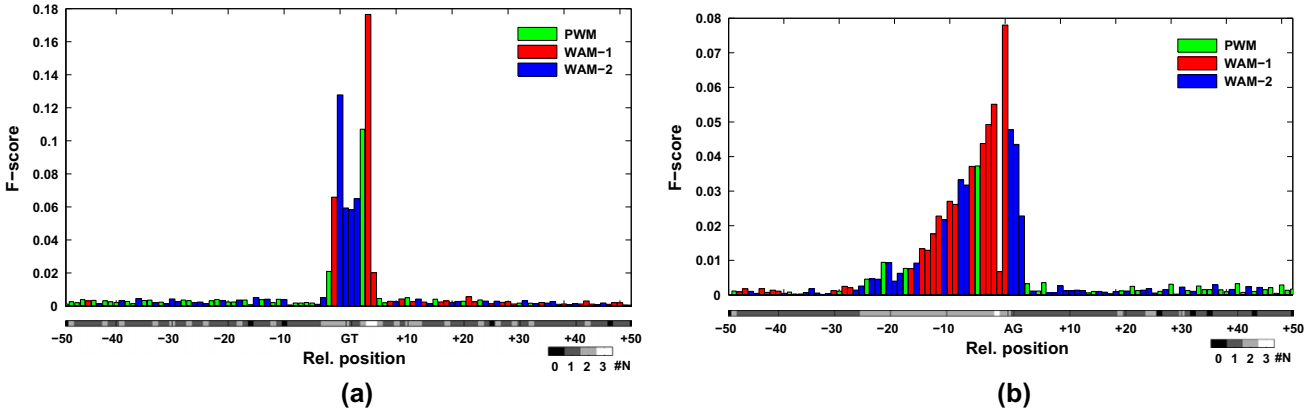

Fig. 1. Position-dependent evaluation of the discriminative power of each probabilistic model in human donor (a) and acceptor (b) sites. Colored bars correspond to the most discriminative dependency lengths (higher *F*-score value) that is selected by PFS to describe each position. The grayscale line underneath *F*-score chart illustrates the number of dependency lengths that are selected by PFA. White blocks indicate positions with all ($N = 3$) probabilistic models used as principal features, light gray are positions with two principal features, dark gray with one principal feature, and black blocks correspond to positions with no representative features ($N = 0$). (For interpretation of color mentioned in this figure the reader is referred to the web version of the article.)

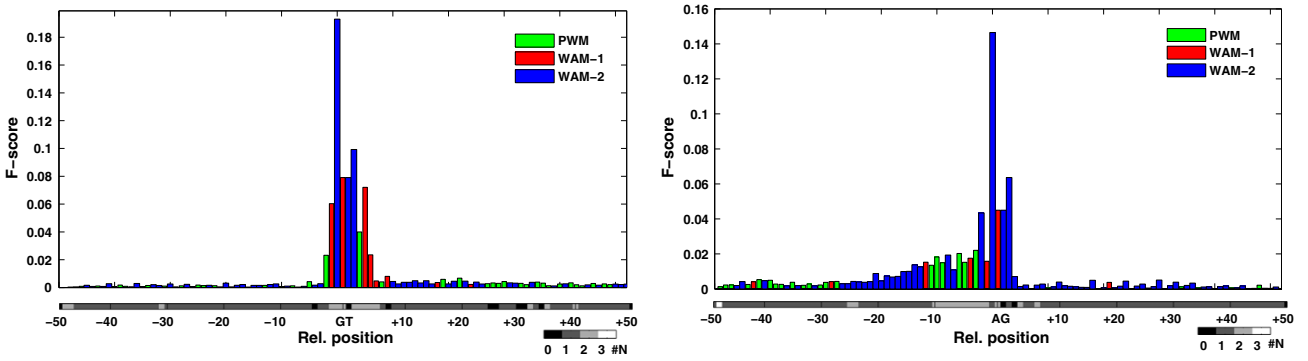

Fig. 2. Position-dependent evaluation of the discriminative power of each probabilistic model in *A. thaliana* donor (a) and acceptor (b) sites (see Fig. 1 legend for details).

Table 1
The optimal sensitivity (*Sn*) and specificity (*Sp*) pairs of the Gaussian SVM-PFS and SVM-PFA classifiers for human and *A. thaliana*. The optimal Sn, Sp pairs correspond to the minimum Euclidean distance $D(1-Sp, Sn)$ from a perfect classification. As for *D*, *CC* corresponds to the maximum value obtained from 100 points along the ROC curves.

	SVM–PFS						SVM–PFA							
	Donor			Acceptor			Donor				Acceptor			
	Sn (%)	*Sp* (%)	*CC*	*Sn* (%)	*Sp* (%)	*CC*	*Sn* (%)	*Sp* (%)	*CC*	Feats[a]	*Sn* (%)	*Sp* (%)	*CC*	Feats[a]
Human	94.96	92.97	0.782	93.56	92.20	0.751	95.88	92.20	0.802	126	93.32	92.76	0.748	129
A. thaliana	93.68	93.61	0.783	90.72	90.82	0.712	92.66	94.21	0.777	106	90.90	90.63	0.709	116

[a] Feats: number of selected features retaining 95% of the data variability.

hence it does not essentially correspond to the maximum sensitivity and specificity pair shown in the same row.

Evidently, PFS and PFA show satisfactory performance between actual splice sites and decoys. However, for both encoding schemes acceptor sites are clearly less identifiable compared to donors in human and especially in *A. thaliana*. This implies that donor identification is more consistent with the applied position-dependent probabilistic modeling than acceptor identification. As regards to the donor prediction, it is also evident that the classification performance on the PFA-based features in human is higher than in *A. thaliana*, while PFS slightly outperforms PFA in *A. thaliana*. The latter observation is possibly indicative of the complex interactions among successive nucleotides in human compared to *A. thaliana*; however, this came up as a result of the GeneSplicer dataset, so further studies are needed to assess this observation.

3.3. Profile of the classified instances

The profile of the classified instances was investigated by analyzing the sequence conservation in a small region flanking putative splice junctions. Sequence conservation at a specific position is defined by the difference between the maximum entropy and the entropy of the observed symbol distribution, i.e. $\log_2 N - \left(-\sum_{n=1}^{N} p_n * log_2 p_n\right)$, where p_n is the observed frequency of a specific residue *n* and *N* is the number of distinct sequence symbols.

Each sequence logo in the Supplementary Figs. S1–S4 (Additional file) represents the degree of conservation in the 20nt region of the classified splice sites [49]. These datasets come from SVM-PFS corresponding to the maximum sensitivity and specificity levels listed in Table 1. As expected, the correctly predicted positive

instances have increased information content compared to the corresponding negative instances. A small difference is observed between human and *A. thaliana*, basically on the G-rich +3 position following the GT consensus in donors and the abundance of the CT- and AT-content in the intronic region preceding AG consensus in acceptors, respectively. False positive splice sites are characterized by lower information content, though they have similar conservation profile with actual splice sites. The deficiency of the positional modeling in this case indicates that the splicing process is also controlled by other factors that obviously do not conform to the positional conservation modeled in the first processing step. The poor conservation of the sequences that are misclassified as decoys is also indicative of the presence of other weighting factors acting on the splicing potential of a sequence. These factors are filtered and appropriately modeled in the subsequent tree-based classification step.

3.4. Second step evidence-based features

Various experimental or computational studies on splicing regulatory elements have been reported on specific short motifs that enhance or inhibit splicing. Going through these studies, we extracted a set of motifs that has been associated with the splicing regulation and then we appropriately described them in the induced tree modeling.

Tables 2 and 3 list the oligomers used in this study along with the selected type of representation and search window. Taking into account the conservation profile of the classified instances (Additional file), human donor sites are represented by the weighted distribution of the C, G nucleotides in the 7-mer intronic and exonic region flanking GT consensus, as well as of the A, G nucleotides in the same region. In addition, G-triplets are known to act as a common intronic splicing enhancer in human [52], while G-repeats immediately downstream a donor site are also an important positive weighting factor for *A. thaliana* [27]. *A. thaliana* is further characterized by the abundance of intronic T-triplets and the concurrent scarcity of G-triplets which are however frequent in the exonic region upstream donor sites and downstream acceptors [27,28]. It was recently found that GAAG motifs are significantly abundant in exonic hexamers in *A. thaliana* that are predicted as potential exonic splicing enhancers (ESEs) [28]. In addition, a binary feature matching for the presence or absence of the consensus WNYAGR motif is used to describe acceptor sites in *A. thaliana* [53]. *A. thaliana* is finally characterized by the lack of AG intronic diplets and long AT stretches in introns flanking splice sites as opposed to the polypyrimidine track preceding human acceptors [27,53]. Recently, Dogan et al. reported on characteristic intronic tetramers close to acceptors in human, the abundance of which is highly differentiated between positive and negative instances [26].

3.5. Comparison of the overall predictive accuracy

The tree modeling elaborates on a subset of the original training dataset corresponding to ambiguous predictions, i.e. those associated with less confident SVM probability estimates. Considering different probability windows in the ambiguous mid-region, a 10-fold CV training procedure along with a post-pruning step is applied on the selected feature set defined in Tables 2 and 3. Figs. 3 and 4 illustrate the ROC curves for human and *A. thaliana*, respectively, that correspond to the overall classification performance of SpliceIT.

Donor prediction in human is more effectively addressed by SpliceIT-PFS and SpliceIT-PFA compared to GeneSplicer and SpliceMachine for sensitivity level over 90%. Acceptor sites are less identifiable by SpliceIT in human than donor sites. This is also reflected to the comparison results, since SpliceMachine gives superior classification performance in lower false positive rates (<3%). In all other cases, SpliceIT-PFS and SpliceIT-PFA perform similarly or higher than SpliceMachine and GeneSplicer.

As in human, SpliceIT is noticeably more accurate in predicting donor sites than acceptors in *A. thaliana* (Fig. 4), having approximately 1% increase of the acceptor false positive rates in sensitivity level that exceeds 95%. Evidently, SpliceIT-PFS and SpliceIT-PFA outperform GeneSplicer in donor and acceptor site prediction. Compared to SpliceMachine, the overall performance is differentiated between donor and acceptor sites. Donor prediction for *A. tha-*

Table 2
List of experimentally or computationally verified features that were used to induce the tree modeling in human. Relevant published studies are listed in the reference column.

Donor			Acceptor		
Feature	Type[a]	Reference	Feature	Type[a]	Reference
bin(S, −7, 1)	WD	[50]	CCTT	LC-i40	[26]
bin(S, +7, 0)	WD	[50]	TTTT	LC-i40	[26]
bin(R, −7, 1)	WD	[50]	TTTC	LC-i40	[26]
bin(R, +7, 0)	WD	[50]	CTTT	LC-i40	[26]
GGG	LC-i40	[26,51]	GAAG	LC-i40	[26]
GGGG	LC-i40, LC-e40	[26,52]	CT-content	LC-i20, LC-e20	[50]

[a] WD: weighted distribution, LC: local context.

Table 3
Experimentally or computationally verified features used to induce the tree modeling in *A. thaliana*. Relevant published studies are listed in the reference column.

Donor			Acceptor		
Feature	Type[a]	Reference	Feature	Type[a]	Reference
bin(W, +4, 0)	WD	[53]	AG	LC-i50	[27,53]
bin(R, −2, 1)	WD	[53]	WNYAGR	LC-splice consensus	[53]
AT-content	LC-i50	[27,53]	AT-content	LC-i50	[27,53]
AT-content	LC-e50	[27,53]	AT-content	LC-e50	[27,53]
GAAG	LC-e50	[28]	GAAG	LC-e50	[28]
TTT	LC-i50, LC-e50	[27]	TTT	LC-i50,LC-e50	[27]
GGG	LC-i50, LC-e50	[27]	GGG	LC-i50,LC-e50	[27]

[a] WD: weighted distribution, LC: local context.

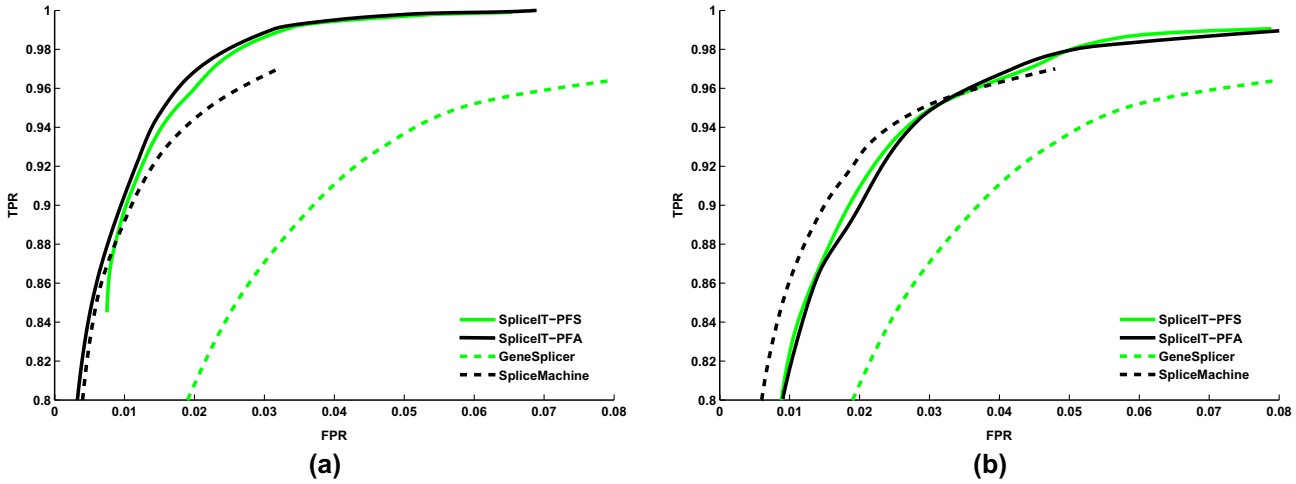

Fig. 3. ROC curves corresponding to the overall performance of SpliceIT-PFS and SpliceIT-PFA for human donor (a) and acceptor (b) sites. As for SpliceIT, the ROC curves for GeneSplicer and SpliceMachine are generated by the sensitivity, specificity pairs in Table S1, by applying the same cubic interpolation method.

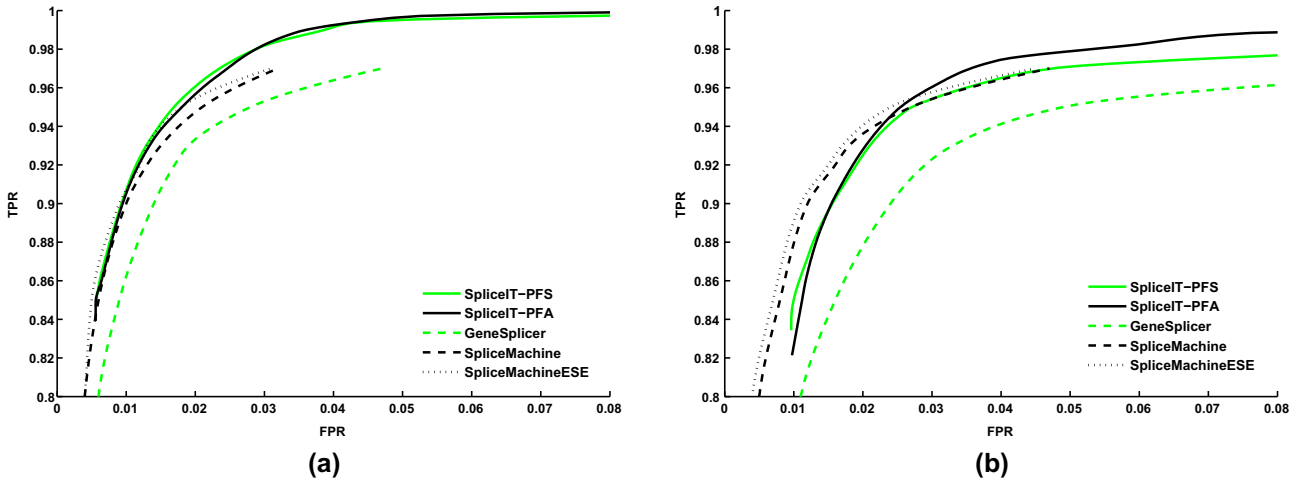

Fig. 4. ROC curves corresponding to the overall performance of SpliceIT-PFS and SpliceIT-PFA for *A. thaliana* donor (a) and acceptor (b) sites. As for SpliceIT, the ROC curves for GeneSplicer, SpliceMachine and SpliceMachineESE are generated by the sensitivity, specificity pairs in Table S2, by applying the same cubic interpolation method.

liana is clearly more precise in higher than 90% sensitivity levels for both SpliceIT-PFS and SpliceIT-PFA. On the contrary, SpliceIT is less accurate in predicting acceptor sites than SpliceMachine when the criterion is the minimum false positive rates, while more efficient in high sensitivity levels (>95%). Especially for *A. thaliana*, an additional comparison was made with a variant of SpliceMachine called SpliceMachineESE that incorporates a set of conserved exonic hexamers near splice sites corresponding to putative splicing enhancers [28]. SpliceMachineESE exhibits improved classification performance when compared to SpliceMachine in all cases, however as with SpliceMachine, SpliceIT has better tradeoff between false positive rates and sensitivity level when sensitivity exceeds 95%.

On the whole, SpliceIT exhibits superior performance in predicting donor sites, while acceptors are more identifiable in higher sensitivity levels compared to GeneSplicer and SpliceMachine. As regards the maximum sensitivity and specificity pairs, SpliceIT exhibits also superior classification performance. Table 4 lists the optimal sensitivity, specificity tradeoff of each method along with the distance from a perfectly well-trained classifier. Compared to the results of the SVM classification shown in Table 1, it is evident that in all cases the proposed hybrid approach clearly benefits from the two-step classification, validating this way the importance of the tested hypotheses. In addition, compared to GeneSplicer and SpliceMachine, either SpliceIT-PFS or SpliceIT-PFA, and most frequently both, give the optimal pair of false and true positive predictions for all datasets.

3.6. Evaluation on NN269

In NN269, the same experimental setup and parameterization was followed as with the GeneSplicer dataset, except for the examined length of the GGG and GGGG features in Table 2, as these exceed the total length of the donor sequences. The SVM optimization step was performed on the whole dataset and the optimal cost C and γ values are shown in Table S3.

Table S4 summarizes the predictive accuracy of SpliceIT in terms of the AUC and minimum Euclidean distance for the NN269 dataset. Evidently, SpliceIT yields similar minimum Euclidean distances on average when these are compared to corresponding values shown in Table 4 for human donor and acceptor sites. Accordingly, the maximum sensitivity/specificity levels are also similar to the GeneSplicer dataset for both PFS and PFA feature selection methods.

Table 4
Optimal sensitivity (Sn) and specificity (Sp) pairs in human and *A. thaliana* corresponding to the minimum Euclidean distance $D(1-Sp, Sn)$ of the ROC curve from the perfectly well-trained classifier.

	Human						*A. thaliana*					
	Donor			Acceptor			Donor			Acceptor		
	Sn (%)	*Sp* (%)	*D*	*Sn* (%)	*Sp* (%)	*D*	*Sn* (%)	*Sp* (%)	*D*	*Sn* (%)	*Sp* (%)	*D*
SpliceIT–PFS	97.00	97.40	0.040	96.50	96.40	0.050	97.00	97.90	0.037	96.00	96.47	0.053
SpliceIT–PFA	97.00	98.30	0.034	97.00	95.90	0.051	97.00	97.70	0.038	97.00	96.70	0.045
GeneSplicer	94.50	94.16	0.080	94.90	94.29	0.077	96.20	96.20	0.054	94.80	95.35	0.070
SpliceMachine	97.00	96.80	0.044	96.10	96.22	0.054	97.00	96.80	0.044	96.10	96.36	0.053
SpliceMachineESE	–	–	–	–	–	–	97.00	96.90	0.043	96.10	96.66	0.051

To further evaluate these results, we compared the AUC values obtained from NN269 (Table S4) with those recently published in [12]. The comparison of the best in terms of accuracy "Reduced MM1 SVM" model with SpliceIT shows that the latter increases the AUC performance by 1.3% for donor and 2% for acceptor sites on average.

3.7. Time estimations

The execution time is an important evaluation factor, as it can be seen as an index of the applicability of a method for larger datasets [8]. Table S5 lists the execution times of each one of the methodological steps for both the human GeneSplicer and NN269 datasets. The experiments run on the same P4 3.2 GHz/1GB system. The comparison between the two datasets shows that the execution time is considerably shortened for the NN269 dataset, due to the decreased amount of the training instances and features, especially for the NN269 donors as these are represented by significantly less positional features than NN269 acceptors. The most time-consuming task, i.e. 10-CV SVM training, takes 89% on average of the total elapsed time (including all methodological tasks), while the decision tree modeling requires only 6% on average. Compared to the training times (20.04/22.17 min for donors and acceptors, respectively) of the best in accuracy "Reduced MM1 SVM" method (AUC: 0.979 for donors, 0.974 for acceptors) provided in [12] for the NN269 dataset, SpliceIT is less time-demanding when PFS is used as feature selection method (6.87 min for donor, 19.67 min for acceptor sites). As expected, the training procedure on the PFA-based feature set needs less time for donors (6.64 min), compared to the acceptor sites (32.17 min), while the difference in the time estimations between PFS and PFA for acceptor sites is due to the increased number of the PFA features (116 vs. 90).

On the whole, the estimated elapsed time shows that SpliceIT can be affordably applied on larger volumes of training data and on wider feature sets, which is especially important considering the accelerating number of sequencing data that is constantly made available and the persistent need for processing massive amounts of experimentally-derived data.

4. Discussion and conclusions

SpliceIT investigates three major issues of the gene prediction discourse, i.e. the encoding and discriminative value of relevant computational and experimental findings, how these findings can be combined with the probabilistic profile of the sequence residues, and what is the role of the order and position in the probabilistic modeling.

The presented methodology constitutes an innovative approach, which drops light on these issues and also employs them towards effective splice site identification. Specifically, SpliceIT proposes a hybrid methodology that follows a two-step classification procedure, the first one employing probabilistic features, and the second combining the outcome of the first step with properly encoded evidence-based features.

The first classification step feeds a Gaussian SVM classifier that incorporates WAMs of different orders. In this phase, the two alternative feature selection methods tested, namely PFS and PFA, express two different working approaches; the first one assumes that each position is important, but the order of each position's Markovian is potentially different, while under the second approach particular positions, even with multiple structural relations, can be the most informative ones. Both approaches prove valuable, however, their classification outcome highlights possible differences between donor and acceptor sites, as well as differences among the species examined in this study, that provides the ground for further investigation.

The probabilistic modeling and the consequent classification shows that there are a number of misclassified negative instances that strongly resemble the positional profile of the actual splice sites. This observation is indicative of the presence of other compositional elements that play a regulatory role by either activating or inhibiting splicing. The so-called splicing enhancers and silencers are cis-acting, often antagonizing elements described by specific degenerate motifs that affect the splicing potential of a GT/AG-containing site [54]. The coverage and mode of dependency between these elements and the splicing process is conditioned by complex biological mechanisms that are yet poorly understood; nevertheless, it has been shown that these elements increase the predictive power of splice site classifiers [55].

The evidence-based features employed in the second step are aimed to capture complementary discriminative information stemming from the underlying splicing mechanism. In this scope, we associated important compositional features that were represented by their weighted distribution or local context. It has to be noted that SpliceIT allows also for additional species-specific features to be encoded, since this information is parameterized in its source code. The results of the second classification step suggest that the co-occurrence of these functional constituents along with other biology-driven compositional features makes actual splice sites significantly more identifiable. Compared to other state-of-the-art techniques, trained also on the same datasets, SpliceIT shows increased classification performance on human and *A. thaliana* splice sites in terms of maximum sensitivity and specificity pairs.

The time-complexity for PFS is linear with the number of the incorporated features, while for PFA the algorithmic complexity is of the order of PCA [34]. The number of the dependency lengths incorporated in the probabilistic modeling does not increase the time-complexity, since feature extraction is easily parallelized. In addition, SpliceIT copes with the shortcomings implied by the deficient justification, when deciding on which Markov order is most suitable to be modeled with no cost on the space complexity of the training procedure. Most importantly, the tree modeling

followed in the second classification step performed on predefined features originated from relevant published studies, hence no additional cost is introduced for deciding on the most representative oligomers and the most suitable search window.

Splice site prediction is becoming even more challenging considering the prevalence of alternative splicing events in complex organisms. Alternative splicing is affected by tissue-specific and developmentally-regulated factors that are poorly understood, yet is believed to be particularly significant in various pathological conditions [56]. Using suitable adjustments, involving the incorporation of additional/diverse biological signals, SpliceIT could facilitate the identification of decisive, splicing-inducing and regulating elements and promote our understanding of the alternative splicing process. In this regard, the source code of SpliceIT is made available to researchers who are interested in applying the whole training procedure to other splice site datasets or even to similar binary classification problems.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jbi.2009.09.004.

References

[1] Burset M, Seledtsov IA, Solovyev VV. Analysis of canonical and non-canonical splice sites in mammalian genomes. Nucleic Acids Res 2000;28(21):4364–75.
[2] Sonnenburg S, Schweikert G, Philips P, Behr J, Rätsch G. Accurate splice site prediction using support vector machines. BMC Bioinformatics 2007;8(Suppl. 10):S7.
[3] Staden R. Measurements of the effects that coding for a protein has on a DNA sequence and their use for finding genes. Nucleic Acids Res 1984;12(1):551–67.
[4] Zhang MQ, Marr TG. A weight array method for splicing signal analysis. Comput Appl Biosci 1993;9(5):499–509.
[5] Reese MG, Eeckman FH, Kulp D, Haussler D. Improved splice site detection in Genie. J Comput Biol 1997;4(3):311–23.
[6] Loi HS, Rajapakse JC. Splice site detection with neural networks/Markov model hybrids. In: Proceedings of the 9th international conference on neural information, vol. 5; 2002. p. 2249–53.
[7] Chen TM, Lu CC, Li WH. Prediction of splice sites with dependency graphs and their expanded Bayesian networks. Bioinformatics 2005;21(4):471–82.
[8] Pertea M, Lin X, Salzberg SL. GeneSplicer: a new computational method for splice site prediction. Nucleic Acids Res 2001;29(5):1185–90.
[9] Thanaraj TA, Robinson AJ. Prediction of exact boundaries of exons. Brief Bioinform 2000;1(4):343–56.
[10] Zhang Y, Chu CH, Chen Y, Zha H, Ji X. Splice site prediction using support vector machines with a Bayes kernel. Expert Syst Appl 2006;30(1):73–81.
[11] Baten AK, Chang BC, Halgamuge SK, Li J. Splice site identification using probabilistic parameters and SVM classification. BMC Bioinformatics 2006;7(Suppl. 5):S15.
[12] Baten A, Halgamuge SK, Chang BCH. Fast splice site detection using information content and feature reduction. BMC Bioinformatics 2008;9(Suppl. 12):S8.
[13] Zhang XH, Heller KA, Hefter I, Leslie CS, Chasin LA. Sequence information for the splicing of human pre-mRNA identified by support vector machine classification. Genome Res 2003;13(12):2637–50.
[14] Degroeve S, Saeys Y, De Baets B, Rouzé P, Van de Peer Y. SpliceMachine: predicting splice sites from high-dimensional local context representations. Bioinformatics 2005;21(8):1332–8.
[15] Rajapakse JC, Ho LS. Markov encoding for detecting signals in genomic sequences. IEEE/ACM Trans Comput Biol Bioinform 2005;2(2):131–42.
[16] Zien A, Rätsch G, Mika S, Schölkopf B, Lengauer T, Muller KR. Engineering support vector machine kernels that recognize translation initiation sites. Bioinformatics 2000;16(9):799–807.
[17] Bhasin M, Zhang H, Reinherz EL, Reche PA. Prediction of methylated CpGs in DNA sequences using a support vector machine. FEBS Lett 2005;579(20):4302–8.
[18] Krause L, McHardy AC, Nattkemper TW, Puhler A, Stoye J, Meyer F. GISMO-gene identification using a support vector machine for ORF classification. Nucleic Acids Res 2007;35(2):540–9.
[19] Trapnell C, Pachter L, Salzberg SL. Tophat: discovering splice junctions with RNA-Seq. Bioinformatics 2009;25(9):1105–11.
[20] Zhao X, Huang H, Speed TP. Finding short DNA motifs using permuted Markov models. In: Proceedings of the 8th annual conference on computational molecular biology (RECOMB); 2004.
[21] Huang W, Umbach DM, Ohler U, Li L. Optimized mixed Markov models for motif identification. BMC Bioinformatics 2006;7:279.
[22] Wang J, Hannenhalli S. Generalizations of Markov model to characterize biological sequences. BMC Bioinformatics 2005;6:219.
[23] Degroeve S, De Baets B, Van de Peer Y, Rouzé P. Feature subset selection for splice site prediction. Bioinformatics 2002;18(Suppl. 2):S75–83.
[24] Saeys Y, Degroeve S, Aeyels D, Van de Peer Y, Rouzé P. Fast feature selection using a simple estimation of distribution algorithm: a case study on splice site prediction. Bioinformatics 2003;19(Suppl. 2):II179–88.
[25] Chen Y-W, Lin C-J. Combining SVMs with various feature selection strategies. In: Guyon I et al., editors. Feature extraction: foundations and applications. Springer-Verlag; 2006. p. 315–23.
[26] Dogan RI, Getoor L, Wilbur WJ, Mount SM. Features generated for computational splice-site prediction correspond to functional elements. BMC Bioinformatics 2007;8:410.
[27] Saeys Y, Degroeve S, Aeyels D, Royzé P, Van de Peer Y. Feature selection for splice site prediction: a new method using EDA-based feature ranking. BMC Bioinformatics 2004;5:64.
[28] Pertea M, Mount SM, Salzberg SL. A computational survey of candidate exonic splicing enhancer motifs in the model plant *Arabidopsis thaliana*. BMC Bioinformatics 2007;8:159.
[29] Platt J. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In: Smola AJ, Bartlett PL, Scholkopf B, Schuurmans D, editors. Advances in large margin classifiers. MIT Press; 2000.
[30] Burge CB. Modeling dependencies in pre-mRNA splicing signals. In: Salzberg SL, Searls DB, Kasif S, editors. Computational methods in molecular biology. Elsevier; 1998. p. 129–64.
[31] Deshpande M, Karypis G. Selective Markov models for predicting web page accesses. ACM Trans Internet Techn 2004;4(2):163–84.
[32] Pitkow J, Pirolli P. Mining longest repeating subsequence to predict World Wide Web surfing. In: Second USENIX symposium on internet technologies and systems; 1999.
[33] Malousi A, Chouvarda I, Koutkias V, Kouidou S, Maglaveras N. Variable-length positional modeling for biological sequence classification. In: Proceedings of the American medical informatics association symposium (AMIA); 2008. p. 91–5.
[34] Cohen I, Tian Q, Zhou XS, Huang TS. Feature selection using principal feature analysis. In: Proceedings of the international conference on image processing; 2002.
[35] NNSplice Dataset. Available from: http://www.fruitfly.org/seq_tools/splice.html.
[36] Durbin R, Eddy SR, Krogh A, Mitchison G. Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge University Press; 1998.
[37] Neumann J, Schnorr C, Steidl G. Combined SVM-based feature selection and classification. Mach Learn 2005;61(1–3):129–50.
[38] Guyon IM, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res 2003;3:1157–82.
[39] Boser BE, Guyon IM, Bousquet O, Mukerjhee S. A training algorithm for optimal margin classifiers. In: Proceedings of the 5th annual ACM workshop on COLT; 1992. p. 144–52.
[40] Vapnik V. Statistical learning theory. Wiley; 1998.
[41] Rätsch G, Sonnenburg S, Schäfer C. Learning interpretable SVMs for biological sequence classification. BMC Bioinformatics 2006;7(Suppl. I):S9.
[42] Sonnenburg S, Rätsch G, Schäfer C, Schölkopf B. Large scale multiple kernel learning. J Mach Learn Res 2006;7:1531–65.
[43] Myles AJ, Feudale RN, Liu Y, Woody N, Brown SD. An introduction to decision tree modeling. J Chemom 2004;18:275–85.
[44] Breiman L, Friedman J, Olshen RA, Stone CJ. Classification and regression trees. Chapman and Hall; 1993.
[45] Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 1975;405(2):442–51.
[46] Chang CC, Liu CJ. LIBSVM: a library for support vector machines; 2001. Available from: http://www.csie.ntu.edu.tw/~cjlin/libsvm.
[47] Witten IH, Frank E. Data mining: practical machine learning tools and techniques. 2nd ed. Morgan Kaufmann; 2005.
[48] Nussinov R. Doublet frequencies in evolutionary distinct groups. Nucleic Acids Res 1984;12(3):1749–63.
[49] Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. Genome Res 2004;14:1188–90.
[50] Senapathy P, Shapiro MB, Harris NL. Splice junctions, branch point sites and exons: sequence statistics, identification, and applications to genome project. Methods Enzymol 1990;183:252–78.
[51] Louie E, Ott J, Majewski J. Nucleotide frequency variation across human genes. Genome Res 2003;13:2584–601.
[52] McCullough AJ, Berget SM. G triplets located throughout a class of small vertebrate introns enforce intron borders and regulate splice site selection. Mol Cell Biol 1997;17:4562–71.
[53] Hebsgaard S, Korning P, Tolstrup N, Engelbrecht N, Rouz P, Brunak S. Splice site prediction in *Arabidopsis thaliana* pre-mRNA by combining local and global sequence information. Nucleic Acids Res 1996;24:3439–52.
[54] Cartegni L, Chew SL, Krainer AR. Listening to silence and understanding nonsense: exonic mutations that affect splicing. Nat Rev Genet 2002;3(4):285–98.
[55] Churbanov A, Rogozin IB, Deogun JS, Ali H. Method of predicting splice sites based on signal interactions. Biol Direct 2006;1:10.
[56] Modrek B, Lee C. A genomic view of alternative splicing. Nat Genet 2002;30(1):13–9.