Exploring Alternative Splicing Features using Support Vector Machines

Jing Xia
Kansas State University
Manhattan, KS USA
xiajing@ksu.edu

Doina Caragea
Kansas State University
Manhattan, KS USA
dcaragea@ksu.edu

Susan Brown
Kansas State University
Manhattan, KS USA
sjbrown@ksu.edu

Abstract

Alternative splicing is a mechanism for generating different gene transcripts (called isoforms) from the same genomic sequence. Finding alternative splicing events experimentally is both expensive and time consuming. Computational methods, in general, and machine learning algorithms, in particular, can be used to complement experimental methods in the process of identifying alternative splicing events. In this paper, we explore the predictive power of a rich set of features that have been experimentally shown to affect alternative splicing. We use these features to build support vector machine (SVM) classifiers for distinguishing between alternatively spliced exons and constitutive exons. Our results show that simple linear SVM classifiers built from a rich set of features give results comparable to those of more sophisticated SVM classifiers that use more basic sequence features. Furthermore, we use feature selection methods to identify computationally the most informative features for the prediction problem considered.

1. Introduction

As genomes are sequenced, a major challenge is their annotation – the identification of genes and regulatory elements, their locations and functions. For years, it was believed that one gene corresponds to one protein, but the discovery of alternative splicing [10] provided a mechanism through which one gene can generate several distinct proteins. Years after its discovery, alternative splicing was still seen more as the exception than the rule [1]. Recently, however, it has become obvious that a large fraction of genes undergoes alternative splicing [11], suggesting the importance of this process. The task of accurately identifying alternative splicing isoforms is particularly intricate, as different transcriptional isoforms can be found in different tissues or cell types, at different development stages, or can be induced by external stimuli. Experimental methods for finding alternative splicing events are expensive and time consuming. Therefore, computational methods that can complement experimental methods are needed. Traditional computational methods rely on aligning expressed sequence tags (ESTs) and complementary DNA (cDNA) to genomic DNA to identify alternative splicing events [19, 16]. More recent machine learning approaches use various sequence features to predict alternative splicing events [23, 30, 27].

Although several types of alternative splicing events exist (e.g., alternative acceptor, alternative donor, intron retention), in this paper we focus on the prediction of cassette exons, one particular type of splicing event, where an exon is a cassette exon (or alternatively spliced) if it appears in some mRNA transcripts, but does not appear in all isoforms. If an exon appears in all isoforms, then it is called a constitutive exon. Several basic sequence features have been used to predict if an exon is alternatively spliced or constitutive, including: exon and flanking introns lengths and the frame of the stop codon. In particular, G. Rätsch et al. [23] have proposed a kernel method, which takes as input a set of local sequences represented using such basic features and builds a classifier that can differentiate between alternatively spliced and constitutive exons. In the process of building the classifier, this method identifies and outputs predictive *splicing motifs*, which are used to interpret the results. In this context, a motif is a sequence pattern that occurs repeatedly in a group of related sequences. The method in [23] is essentially searching for motifs within a certain range around each base. This range needs to be carefully chosen in order to obtain good prediction results[14].

Finding motifs that explain alternative splicing of premRNA is not surprising as it has been experimentally shown that alternative splicing is highly regulated by the interaction of intronic or exonic RNA sequences (more precisely, motifs that work as signals) with a series of splicing regulatory proteins [14]. Such splicing motifs can provide useful information for predicting alternative splicing events, in general, and cassette exons, in particular. Generally, computational identification of splicing motifs can be derived from patterns that are conserved in another organism [15, 26, 7]. However, since some exons and most introns are

not conserved, it is desirable to identify such motifs directly from local sequences in the organism of interest.

In addition to motifs, several other sequence features have been shown to be informative for alternative splicing prediction [14]. Among these, pre-mRNA secondary structure has been investigated to identify patterns that can affect splicing [13, 20]. It has been found that the pre-mRNA exhibits local structures that enhance or inhibit the hybridization of spliceosomal snRNAs to the pre-mRNA. In other words, the structure can affect the selection of the splice sites. As another feature, the strength of the general splice sites is very important with respect to the splicing process, as strong splice sites allow the spliceosomes to recognize pairs of splice sites between long introns [30, 8]. When the splice sites degenerate and weaken, other splicing regulatory elements (exon/intron splicing enhancers and silencers) [21] are needed. At last, one other feature that has been shown to be correlated with the spicing process is given by the base content in the vicinity of splice sites [14].

Although the method in [23] can *output* motifs that explain the classifier results, to the best of our knowledge there is no study that explores motifs (derived either using comparative genomics or local sequences) and other alternative splicing features (pre-mRNA secondary structure, splice site strength, splicing enhancers/silencers and base content) together as *inputs* to machine learning classifiers for predicting cassette exons. In this paper, we use the above mentioned features with state-of-the-art machine learning methods, specifically the SVM algorithm, to generate classifiers that can distinguish alternatively spliced exons from constitutive exons. We show that the classification results obtained using all these features with simple linear SVMs are comparable and sometimes better than those obtained using only basic features with more complex nonlinear SVMs. To identify the most discriminative features among all features in our study, we use machine learning methods (SVM feature importance and information gain) to perform feature selection.

The rest of the paper is organized as follows: We introduce the machine learning algorithms used to predict alternatively spliced exons and to perform feature selection in Section 2. In Section 3, we briefly describe the data set used in our experiments and explain how we construct the features considered in our study. We present experimental results in Section 4 and conclude with a summary and ideas for future work in Section 5.

2. Methods

2.1. Support Vector Machine Classifiers

The support vector machine (SVM) algorithm [29] is one of the most effective machine learning algorithms for many complex binary classification problems, including a wide range of bioinformatics problems [12, 17, 5, 21], and has been recently used to detect splice sites [22, 23, 25]. The SVM algorithm takes as input labeled data from two classes and outputs a model (a.k.a., classifier) for classifying new unlabeled data into one of those two classes. SVM can generate linear and non-linear models.

Let $E = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_l, y_l)\}$, where $\mathbf{x}_i \in R^p$ and $y_i \in \{-1, 1\}$, be a set of training examples. Suppose the training data is *linearly separable*. Then it is possible to find a hyperplane that partitions the pattern space into two half-spaces. The set of such hyperplanes is given by $\{\mathbf{x}|\mathbf{x} \cdot \mathbf{w} + b = 0\}$, where $\mathbf{x}$ is the p-dimensional data vector and $\mathbf{w}$ is the normal to the separating hyperplane. SVM selects among the hyperplanes that correctly classify the training set, the one that minimizes $\|\mathbf{w}\|^2$, subject to the constraints $y_i(\mathbf{x_i} \cdot \mathbf{w} + b) \leq 1$. This is the same as the hyperplane for which the *margin* of separation between the two classes, measured along a line perpendicular to the hyperplane, is maximized.

The algorithm assigns a weight α_i to each input point $\mathbf{x}_i$. Most of these weights are equal to zero. The points having non-zero weight are called *support vectors*. The separating hyperplane is defined as a weighted sum of support vectors. Thus, $\mathbf{w} = \sum_{i=1}^{l}(\alpha_i y_i)\mathbf{x_i} = \sum_{i=1}^{s}(\alpha_i y_i)\mathbf{x_i}$, where s is the number of support vectors, y_i is the known class for example $\mathbf{x}_i$, and α_i are the support vector coefficients that maximize the margin of separation between the two classes. The classification for a new unlabeled example can be obtained from $f_{\mathbf{w},b}(\mathbf{x}) = \text{sign}(\mathbf{w} \cdot \mathbf{x} + b) = \text{sign}(\sum_{i=1}^{l} \alpha_i y_i (\mathbf{x} \cdot \mathbf{x_i}) + b)$.

If the goal of the classification problem is to find a linear classifier for a non-separable training set (e.g., when data is noisy and the classes overlap), a set of *slack variables*, ξ_i, is introduced to allow for the possibility of examples violating the constraints $y_i(\mathbf{x_i} \cdot \mathbf{w} + b) \leq 1$. In this case the margin is maximized, paying a penalty proportional to the cost C of constraint violation, i.e., $C \sum_{i=1}^{l} \xi_i$. The decision function is similar to the one for the linearly separable problem.

If the training examples are not linearly separable, the SVM works by mapping the training set into a higher dimensional *feature* space, where the data becomes linearly separable, using an appropriate kernel function k.

We use the LIBSVM implementation of SVM, available at http://www.csie.ntu.edu.tw/ cjlin/libsvm/, in our study.

2.2. Feature Selection Methods

Feature selection methods are used to select the most informative features with respect to a prediction or classification problem. Eliminating redundant or uninformative features helps to enhance the generalization capability of machine learning algorithms and to improve the model in-

terpretability. In our study, we used two feature selection methods: (1) SVM feature importance [12] and (2) information gain [32], to identify the most relevant features for distinguishing alternatively spliced exons from constitutive exons. The weight vector $\mathbf{w} = \{|w_0|, |w_1|, ..., |w_n|\}$ (where n is the dimension of the feature vector) determined by the SVM algorithm is used as a heuristic to identify important features using the SVM feature importance method.

The information gain criterion also provides a simple way to determine feature importance. The information gain is the expected reduction in entropy caused by partitioning the training examples into classes, according to a certain feature (where the entropy measures the impurity of a sample E of training examples). One can rank all features in the order of decreasing information gain and select relevant features conservatively [32]. A more robust way of identifying important features is to use a decision tree algorithm, which iteratively selects the feature with the highest information gain at each node of the tree. The features that are nodes in the final decision tree are considered to be more informative than the others.

3 Data Set and Feature Construction

3.1. Data Set

The data set used in our experiments contains alternatively spliced and constitutive exons in *C.elegans*. It has been used in related work [23] and is available at http://www.fml.tuebingen.mpg.de/raetsch/projects/RASE. A detailed description of how this data set was generated can be found in [23]. Briefly, *C.elegans* EST and full length cDNA sequences were aligned against the *C.elegans* genomic DNA to find the coordinates of exons and their flanking introns. After finding these coordinates, pairs of sequences which shared 3' and 5' boundaries of upstream and downstream exons were identified, such that one sequence contained an internal exon, while the other did not contain that exon. This procedure resulted in 487 alternatively spliced exons and 2531 constitutive exons. The final data set was split into 5 independent subsets of training and testing files for cross validation purposes.

3.2. Feature Construction

Six classes of features that affect alternative splicing are considered in our study: (1) pre-mRNA splicing motifs, specifically (1a) motifs derived from local sequences using MAST (MAST) and (1b) intronic regulatory splicing (IRS) motifs derived using comparative genomics methods; (2) pre-RNA secondary structure related features, specifically (2a) the optimal folding energy (OFE) and (2b) a reduced motif set (RMS) obtained by taking the secondary structures into account; (3) exon splicing enhancers (ESE); (4) splice site strength (SSS); (5) GC-content (GCC) in introns; and (6) basic sequence features (BSF) used in [23], specifically exon and flanking introns lengths and stop codon frames.

We used the MEME [4] and MAST [3] tools available at http://meme.sdsc.edu/meme/intro.html to detect motifs based on local sequences. MEME is a statistical tool for discovering *unknown* motifs in a group of related DNA or protein sequences. Its underlying algorithm is an extension of the expectation maximization algorithm for fitting finite mixture models [2]. Optimal values for parameters such as the motif width and the number of motif occurrences are automatically found by MEME. Contrary to MEME, MAST is a tool for searching sequences with a group of *known* motifs. A match score is calculated between each input sequence and each given motif. To use the MEME/MAST system, we first constructed local sequences by considering (-100, +100) bases around the donor sites (splice sites of upstream introns) and acceptor sites (splice sites of downstream introns) of the sequences in the original data set. Then, we ran MEME to obtain a list of 40 motifs (20 motifs for donor sites and 20 motifs for acceptor sites). MAST was used to search each sequence with these 40 motifs to obtain their location in each sequence and the corresponding p-values. Finally, we represented each sequence as a 40-dimensional feature vector. Each dimension corresponds to one of the 40 MEME motifs and indicates how many times that specific motif appears in the sequence.

In addition to motifs identified by MEME/MAST based on local sequences, we also considered intronic regulatory (IRM) motifs found by comparative genomics in Nematodes [15]. The basic idea of the comparative genomics procedure here is to identify alternatively spliced exons whose flanking introns exhibit high nucleotide conservation between *C.elegans* and *C.briggsae*. Then, the most frequent pentamers and hexamers are extracted from the conserved introns. In our case, this procedure resulted in a list of 60 intronic regulatory motifs, 30 motifs for upstream introns and 30 motifs for downstream introns. For each sequence, we scanned the upstream intron with the upstream intronic motifs to find the number of occurrences of each motif. Each upstream intron is represented as a 30-dimensional vector, where each dimension indicates how many times the motif appears in the sequence. The same approach is applied to the downstream introns of each exons. Altogether, this set of features is represented as a 60-dimensional vector.

It is known that the splicing of exons can be enhanced or repressed by specific local pre-mRNA secondary structures around the splice sites [13, 20]. As shown in [13], motifs in single-stranded regions have more effect on the selection of splice sites than those in double-stranded regions. Following these ideas, we used the `mfold` software [18] available at http://mfold.bioinfo.rpi.edu/ to predict the

pre-mRNA folding (secondary structure formation) within a 100-base window around the acceptor and donor sites of each exon. `Mfold` parameters were chosen to prevent the formation of global double stranded base pairs. Thus, rather than folding the whole sequence, only local foldings were allowed. Two sub-classes of features related to the pre-mRNA secondary structure were considered in our study: (a) The *Optimal Folding Energy*, which roughly reflects the stability of the RNA folding; and (b) A *reduced motif set* derived, under the assumption that motifs on single stranded sequences are more effective than those on helices, from the set of MAST motifs by removing the motifs that are located on double stranded sequences with a probability greater than a threshold.

Although splicing regulators have been identified in both introns and exons, exon splicing regulators (ESR) are more common and better characterized than intron splicing regulators [6]. Exon splicing enhancers (ESE) affect the choice of splicing sites through recruiting arginine/serine dipeptide-rich (SR) proteins, which in turn bind other spliceosomal components through protein-protein interactions. We adopted the approach in [21] to search for specific ESEs in our data. Since recent studies show that ESEs tend to be less active outside the close vicinity of splice sites [21], we used a 50-base window around the splice sites to search for ESEs. We also considered the following two assumptions made in the RESCIE-ESE algorithm [9, 21] in our search: (1) ESEs appear much more frequently in exons than in introns and (2) ESEs appear much more frequently in exons with weak splice sites than in exons with strong splice sites. The following two difference distributions were computed in our study: (1) $\{|f_E^h - f_I^h||h \in \text{all possible hexamers}\}$, where f_E^h is the frequency of a given hexamer h in exon regions within the 50-base windows, and f_I^h is the frequency of a given hexamer h in intron regions; (2) $\{|f_W^h - f_S^h||h \in \text{all possible hexamers}\}$, where f_W^h is the frequency of a given hexamer in exons with weak splice sites, and f_S^h is the frequency of a given hexamer in exons with strong splice sites. Given these two difference distributions, we set a threshold and obtained 77 hexamers with high frequency in the two difference distributions. We scan the exon of each sequence for these motifs and represent the sequence as a 77-dimensional vector, where each dimension indicates how many times the corresponding hexamer appears in the sequence.

Another feature we used in our study is given by the strength of the splice sites, as the strength has been shown to be informative for identifying alternatively spliced exons [28, 30]. More precisely, the strength is expected to be lower for alternatively spliced sites compared to constitutive splice sites. We used a position specific scoring-based approach [8] to model the strength of splice sites, according to the following formula: $score = \sum_i log\frac{F(X_i)}{F(X)}$, where $F(X_i)$ is the frequency of the nucleotide X at position i, and $F(X)$ is the background frequency of the nucleotide X. As already known, in *C.elegans* the background frequency is 66% AT. We extracted a range of (-3, +7) around donor sites (3 exon bases and 7 intron bases) and a range of (-26, +2) around acceptor sites (26 intron bases and 2 exon bases), and used the formula above to obtain scores for the strength of the acceptor and donor sites. The two ranges above are chosen to cover the main AG dinucleotides, which are bound by splicing factors around acceptor sites and the adjacent polypyrimidine tracts (PPT) [30]. Because the acceptor and donor sites can be seen as a pair, their scores are summed together to obtain the overall splice site strength, which is represented as another feature.

The GC content of a sequence is another feature correlated with the selection of splice sites. Alternatively spliced exons occur more frequently in GC-poor flanking sequences [28]. We take into account this property by using a sliding window method to scan the GC content of each sequence within a range of (+100, -100) around donor and acceptor sites. The window size is set to 5, resulting in a 40-dimensional feature vector for each splice site. Each position indicates the ratio of GCs to the window size.

Last but not the least, sequence length has been shown to be a feature that can help distinguish alternatively spliced exons from constitutive exons [27, 7]. In [23], a feature vector consisting of upstream intron length, exon length, downstream intron length and the frame of the stop codon was constructed for each exon and its flanking introns. The length features were discretized into a logarithmically spaced vector consisting of 30 bins. The frame of the stop codons is represented using a 3D vector. In this study, we call this last set of features *basic features*.

4. Experimental Results

4.1. Motif Evaluation

The purpose of the motif evaluation in this section is to identify the splicing motifs that appear in several different sets, as those motifs are probably the most informative for alternative splicing. To do that, we first compared the set of 40 motifs identified by MEME/MAST with the set of putative motifs found in [23] and the ISR motifs found in [15]. The MAST motifs are represented as position-specific scoring matrices (PSSMs), shown as a two-level consensus sequences in Table 1. Upper-level bases have scores higher than or equal to the lower-level bases. A base is conserved if there is no lower-level base in its column. Eight motifs are found in all three sets compared, some of them (e.g., mast2

Table 1. The intersection between MAST motifs, motifs found in [23] and IRS motifs found in [15]. MAST motifs 1-20 are around 5' splice sites, while motifs 21-40 are around 3' splice sites. IRS motifs are italicized.

MAST motifs (Multilevel expression)	E-value	Contained hexamers	Number
TTTTTTTTTCA	4.8e-046	tttttt	mast2
GTGAGTTTTTT A	4.6e-033	tttttt	mast3
AAAAATTTTAAATTTTCAGG TT TTAAAATTT A	3.9e-030	tttttt, atatat tatata	mast4
ATTTTTCAAATTTTT T C T A C	1.6e-026	tttttt	mast6
GCCGGTGGAGCTGTCGTAGG A A CC CC GC GTAGC A	3.6e-026	gttgtc, *catcgc gtgttg*	mast9
AGCCGCCGAAGCCCTTGCCA CATT TA C AAAGCC GAG	1.0e-018	gttgtc , *ccctgg catcgc, cactgc*	mast14
CAGCACCAACAGCACCACCA TC TG G TT G A	1.4e-049	cagcag	mast22
TTTTTTTTTTCAAAATTTTA A TGG T CT	3.3e-038	tttaaa, aatttt atttta	mast23

and mast3) being highly conserved among the *C.elegans* sequences in our data set.

Second, we compared the 77 ESE hexamers, found as described in Section 3.2., with two sets of candidate human and mouse ESE hexamers proposed in [24]. Thirty two out of the 77 putative *C.elegans* ESE hexamers occur also in the human and mouse ESE sets, suggesting that the regulation of splicing, as well as the splicing process itself, are highly conserved in metazoans. Furthermore, a set of experimentally confirmed A. *thaliana* ESE ninemers [21] was used for comparison. The 32 conserved ESE hexamers are shown below; the A. *thaliana* ESE ninemers containing some of these hexamers are listed in brackets:

aatgga, aacaac, **aagaag** [GAAGAAGAA, GAGAAGAAG, TTGAAGAAG], **aaggaa** [GAAGGAAGA], **aaggag** [AAAGGAGAT], attgga, atgatg, atggaa, atggat, acaaga, **agaaga** [GAAGAAGAA, GAGAAGAAG], agaagc, tcatca, tgaaga, tgatga, tggaag, tggatc, **caagaa** [CAAGAAACA], **cagaag** [GAGCAGAAG], cgacga, gaaagc, **gaagaa** [GAAGAAGAA, GAGAAGAAG, GAAGAAAGA, TTGAAGAAG], **gaagat** [GAAGATGGA, GAAGATTGA], **gaagag** [GAAGAGAAA], **gaagga** [GAAGGAAGA], gatgat, **gatgga** [GAAGATGGA], gagaag, gaggag, ggaaga [GAAGGAAGA], **ggagaa** [ATGGAGAAA], ggagga.

It is worth mentioning that our study finds no intersection between the IRS motifs and the ESE motifs in *C.elegans*, suggesting that the two sets are functionally different.

Table 2. Results of alternatively spliced exons classification. All features, but IRS motifs, are included.

	C	Validation Score		Test score	
		fp 1%	AUC	fp 1%	AUC
Split1	0.05	35.36%	86.99%	44.44%	89.32%
Split2	0.05	36.50%	88.56%	46.92%	87.57%
Split3	0.1	35.27%	86.91%	47.31%	88.59%
Split4	0.01	37.56%	88.36%	26.88%	86.60%
Split5	0.1	39.80%	88.03%	29.47%	86.98%

4.2. Model Selection

The performance of a classifier depends on judicious choice of various parameters of the algorithm. For the SVM algorithm there are several inputs that can be varied: the cost of constraint violation C (e.g., $C = 1$), tolerance of the termination criterion (e.g., $\epsilon = 0.01$), type of kernel used (e.g., linear, polynomial, radial or Gaussian), parameters of the kernel (e.g., the degree or coefficients of the polynomial kernel), etc.

G. Rätsch et al. [23] have used basic features with several types of customized kernels, as well as an optimal subkernel weighting to learn SVM classifiers that differentiate between alternatively spliced and constitutive exons, and to identify motifs that can be used to interpret the results. In this section, we show that simple linear kernels can be used to obtain similar results if motifs are used as input features. In order to tune the cost C, we use 5-fold cross-validation for each training set, with $C \in \{0.01, 0.05, 0.1, 0.5, 1, 2\}$. We choose the value of C for which the area under curve (AUC) is maximized during the cross-validation. AUC is a global measurement which takes true positive ratio and false positive ratio into account. True positive ratio is the number of positively labeled examples classified by the algorithm as positive divided by the total number of positive examples. False positive ratio is the number of negatively labeled examples classified as positive divided by the number of negatively labeled examples.

Table 2 shows the results of classification of exons using all features described in Section 3, except conserved IRS motifs that need additional information from closely related organisms to be determined. Table 3 shows the results when the conserved IRS motifs described in Section 3.2 are also included. From Tables 2 and 3, we notice that on the average, the performance improves in terms of true positive rate at 1% false positive rate when IRS motifs are included, which means that IRS motifs conserved among several species contribute to better classification performance. Furthermore, the results are comparable and sometimes better than the results obtained by G. Rätsch et al. [23]. For example, when testing on the first data set we obtain a true

Table 3. Results of alternatively spliced exons classification. All features, including IRS motifs are used.

	C	Validation Score		Test score	
		fp 1%	AUC	fp 1%	AUC
Split1	0.05	32.45%	86.55%	56.48%	90.05%
Split2	0.05	39.33%	88.32%	52.04%	89.04%
Split3	0.1	37.56%	87.76%	38.71%	87.97%
Split4	0.01	40.86%	89.02%	37.63%	84.42%
Split5	0.1	36.48%	87.50%	35.79%	85.69%

positive rate of 56.48% at a fp rate of 1% and the AUC is 90.05%, thus improving the previous results of tp 51.85% at fp 1% and AUC 89.90%.

To evaluate how much the mixed features improve the performance of classification of alternatively spliced exons, we compared the AUC scores of classifiers trained on data sets with and without mixed features, respectively. Figure 1 shows the result of comparison between a data set with basic features only and a data set that includes the other features (except conserved IRS motifs).

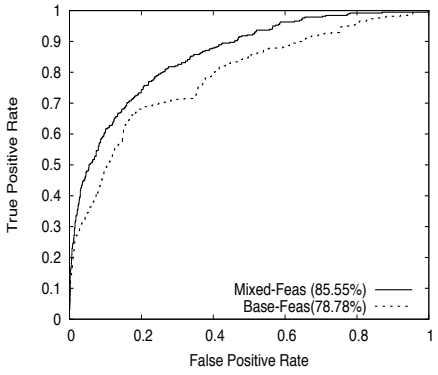

Figure 1. Comparison of ROC curves obtained using basic features only and basic features plus other mixed features (except conserved IRS motifs). Models trained by 5-fold CV with $C = 1$.

Figure 2 shows a comparison of the AUC scores for each data set. It can be seen that the SVM classifiers using MAST motif features return higher AUC scores than those considering only basic sequence features.

In order to evaluate the effect of pre-mRNA secondary structure features on classification of alternatively spliced exons, we performed two experiments, one using data sets considering pre-mRNA secondary structure features obtained as described in Section 3.2 and the other using data sets without secondary structure features. Figure 3 shows the results of the two experiments in which the classifiers

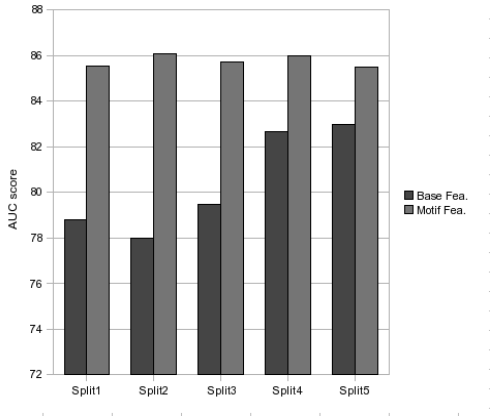

Figure 2. AUC score comparison between data sets with BSF and data sets including MAST motif features. AUC values were obtained based on 5-fold CV with $C = 1$.

were trained using 5-fold cross-validation with optimal cost parameters listed in Table 2. We can see the improvement obtained when considering secondary structure features.

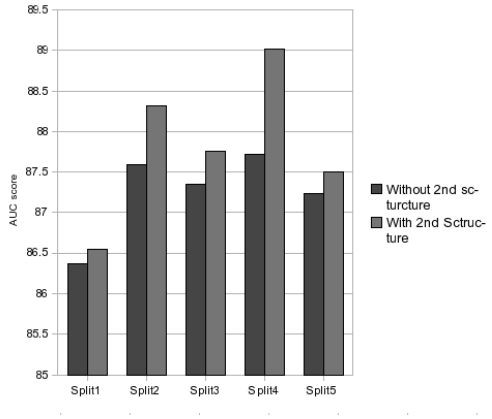

Figure 3. AUC scores comparison between data sets with features of secondary structure and data sets without features of secondary structure

4.3. Feature Selection

We used SVM feature importance and information gain criteria to order features according to their importance with respect to the problem of predicting alternatively spliced exons. First, a linear kernel SVM classifier with optimal cost value was learned for each dataset. The importance of each class of features was estimated by taking the average, across all features in a class, of the corresponding feature weight in

Table 4. Weight importance of the following features: 105 BSF, 1 SSS, 80 GCC, 60 IRS, 40 MAST, 77 ESE, 1 OFE.

Feature	Mean	Max	Min	Std. Dev.
BSF	16.61	27.48	0.13	9.87
SSS	51.05	51.05	51.05	0.00
GCC	10.60	14.90	6.65	1.77
IRS	10.14	25.93	3.43	4.41
MAST	2.06	3.80	0.27	1.02
ESE	1.08	2.13	0.45	0.32
OFE	0.18	0.12	0.24	0.06

the weight vector **w**. Table 4 shows the statistics obtained for the classes of features considered. It can be seen that SSS and BSF are the most informative classes of features. It is not surprising that these classes of features have high importance, as they were previously reported to be very informative for exon splicing prediction in [30] and [27], respectively. However, taken separately, the SSS features do not discriminate well between alternatively spliced and constitutive exons (results not shown), suggesting that they are highly correlated with the BSF features.

In Section 4.2., we have seen that IRS motifs, MAST motifs and ESEs provide useful information for classification, improving the results of classifiers that use only BSF and SSS features. To select the most informative motifs from these sets of features, we used the SVM-produced weight value to order the motifs and chose the best 20 motifs among these features. Most of the 20 best motifs were IRS motifs.

Furthermore, as described in Section 2.2, we also ran the J48 decision tree algorithm in the data mining package WEKA [31] to build a classifier for each data set. We analyzed the nodes in each constructed decision tree and extracted the motifs, namely nodes, occurring in all five trees. We consider these motifs as most informative motifs according to the information gain criterion. Table 5 shows the list of motifs found based on information gain. By comparing the set of the 20 best SVM motifs with the set of the best J48 motifs, we found that the IRS pentamers GCTTC and GTGTG in the upstream intron and GCATG in the downstream intron were included in both sets (bolded in Table 5). We also noted that ese65 (gatgat) was the most frequent hexamer among the selected ESEs.

Table 5. List of mast$k$, ese$k$ and irs$k$ motifs found by choosing nodes which occur in all decision tree classifiers, where k indicates the position in the corresponding list. Irs21,23,31 are IRS motifs identified by both J48 and SVM as important. The rank is based on SVM feature importance.

motifs	Location	Weight value	Rank
mast4	5' ss	1.59	272
mast17	5' ss	2.73	245
mast22	3' ss	3.35	238
mast23	3' ss	3.33	240
mast32	3' ss	1.34	283
ese20	5' ss	1.23	288
ese65	3 ' ss	1.85	262
irs7	5' intron	6.15	217
irs9	5' intron	10.18	134
irs14	5' intron	10.39	132
irs21	5' intron	16.05	62
irs23	5' intron	13.52	75
irs31	3' intron	11.76	109
irs49	3' intron	10.06	135

5. Conclusions and Future Work

The importance of identifying alternative splicing informative features and using them to predict alternative splicing events is reflected by the amount of recent research in this area [7, 23, 26, 27]. However, there is no comprehensive computational study that considers all the features that have been shown experimentally to contribute to the identification of alternatively spliced exons. In this paper, we have presented such a study.

More precisely, we have shown how to use computational methods to construct alternative splicing features and how to built simple SVM classifiers using the features constructed. Our ultimate goal was to gain insights into the most informative features for the prediction problem at hand. MEME/MAST tools were used to identify motifs from local sequences. We have demonstrated that the resulting motifs can aid the classification of alternatively spliced exons even when used with simple linear SVM classifiers, thus providing a good alternative to more sophisticated kernel methods [23]. We have also explored several other features, such as pre-mRNA secondary structure, exonic splicing enhancers, splice site strength and CG-content, which have been shown to be relevant to alternative splicing from a biological point of view. Our results indicate that these features can further improve the accuracy of classifiers that distinguish alternatively and constitutively spliced exons. Finally, we have shown how we can use features selection methods to identify informative features. The methods presented here will be useful for the analysis of predicted gene models in newly sequenced genomes with limited, but enough for training, ESTs and/or cDNA libraries.

Our future work will focus on identifying motifs more

accurately at first. We will also explore alternative ways to represent biological features, as well as relationships among biological features (e.g., pre-mRNA secondary structures and motifs) or between biological features and environment.

Acknowledgments: This work is supported by the National Science Foundation under Grant No. 0711396 to Doina Caragea. We thank Dr. William H. Hsu for providing financial support for Jing Xia.

References

[1] G. Ast. How did alternative splicing evolve? *Nat. Rev. Genet.*, 5(10):773782, 2004.

[2] T. L. Bailey and C. Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. of 2nd International Conf. on Intelligent Systems for Molecular Biology*, pages 28–36, 1994.

[3] T. L. Bailey and M. Gribskov. Combining evidence using p-values: application to sequence homology searches. *Bioinformatics*, 14(1):48–54, November 1998.

[4] T. L. Bailey, N. Williams, C. Misleh, and W. W. Li. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Research*, 34(Web Server issue):W369W373, July 2006.

[5] A. Ben-Hur and D. Brutlag. Remote homology detection: a motif based approach. *Bioinformatics*, 19(Suppl. 1):i26–i23, February 2003.

[6] L. Cartegni, S. L. Chew, and A. R. Krainer. Listening to silence and understanding nonsense: Exonic mutations that affect splicing. *Nature reviews Genetics*, 3(4):285298, November 2002.

[7] G. Dror, R. Sorek, and R. Shamir. Accurate identification of alternatively spliced exons using support vector machine. *Bioinformatics*, 21(7):897–901, November 2005.

[8] M. E. Fahey and D. G. Higgins. Gene Expression, Intron Density and Splice Site Strength in *Drosophila* and *Caenorhabditis*. *Journal of Molecular Evolution.*, 65(3):349–357, Sep 2007.

[9] W. G. Fairbrother, R.-F. Yeh, P. A. Sharp, and C. B. Burge. Predictive identification of exonic splicing enhancer motifs in human protein-coding genes. *Science*, 297(5583):10071013, August 2002.

[10] W. Gilbert. Why genes in pieces? *Nature*, 271(5645):501, 1978.

[11] B. Graveley. Alternative splicing: increasing diversity in the proteomic world. *Trends Genet.*, 17(2):100107, 2001.

[12] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46:389422, November 2002.

[13] M. Hiller, Z. Zhang, R. Backofen, and S. Stamm. Pre-mRNA Secondary Structures Influence Exon Recognition. *PLoS Comput Biol*, 3(11):e204, November 2007.

[14] D. Holste and U. Ohle. Strategies for Identifying RNA Splicing Regulatory Motifs and Predicting Alternative Splicing Events. *PLoS Comput Biol*, 4(1):e21, January 2008.

[15] J. L. Kabat, S. Barberan-Soler, P. McKenna, H. Clawson, T. Farrer, and A. M. Zahler. Intronic Alternative Splicing Regulators Identified by Comparative Genomics in *Nematodes*. *PLoS Comput Biol.*, 2(7):e86, July 2006.

[16] Z. Kan, E. C. Rouchka, W. R. Gish, and D. J. States. Gene Structure Prediction and Alternative Splicing Analysis Using Genomically Aligned ESTs. *Genome Res.*, 11(5):889–900, May 2001.

[17] C. S. Leslie, E. Eskin, A. Cohen, J. Weston, and W. S. Noble. Mismatch string kernels for discriminative protein classification. *Bioinformatics*, 20(4):467–476, August 2003.

[18] D. H. Mathews, J. Sabina, M. Zuker, and D. H. Turner. Expanded Sequence Dependence of Thermodynamic Parameters Improves Prediction of RNA Secondary Structure. *Journal of Molecular Biology*, 288(5):911940, May 1999.

[19] S. H. Nagaraj, R. B. Gasser, and S. Ranganathan. A hitchhiker's guide to expressed sequence tag (EST) analysis. *Brief Bioinform.*, 8(1):621, May 2006.

[20] D. J. Patterson, K. Yasuhara, and W. L. Ruzzo. PRE-mRNA Secondary Structure Prediction Aids Splice Site Prediction. *Proceedings of the Pacific Symposium on Biocomputing*, pages 223–234, 2002.

[21] M. Pertea, S. M. Mount, and S. L. Salzberg. A computational survey of candidate exonic splicing enhancer motifs in the model plant *arabidopsis thaliana*. *BMC Bioinformatics*, 8:159, May 2007.

[22] G. Rätsch and S. Sonnenburg. Accurate Splice Site Dectection for *Caenorhabditis Elegans* in Kernel Methods in Computational Biology. *Kernel Methods in Computational Biology, MIT press*, pages 277–298, 2004.

[23] G. Rätsch, S. Sonnenburg, and B. Schölkof. RASE: recognition of alternatively spliced exons in *c. elegans*. *Bioinformatics*, 21(Suppl 1):369–377, June 2005.

[24] Rescue-ese web server. [http://genes.mit.edu/burgelab/rescue-ese/].

[25] S. Sonnenburg, G. Schweikert, P. Philips, J. Behr, and G. Rätsch. Accurate splice site prediction using support vector machines. *BMC Bioinformatics*, 8(Suppl 10):S7, December 2007.

[26] R. Sorek and G. Ast. Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. *Genome Research*, 13(7):16311637, July 2003.

[27] R. Sorek, R. Shemesh, Y. Cohen, O. Basechess, G. Ast, and R. Shamir. A Non-EST based method for exon-skipping prediction. *Genome Res.*, 14(8):16171623, January 2004.

[28] T. A. Thanaraj and S. Stamm. Prediction and statistical analysis of alternatively spliced exons. *Prog Mol Subcell Biol.*, 31:131, 2003.

[29] V. N. Vapnik. *The Nature of Statistical Learning Theory (Statistics for Engineering and Information Science).* Springer Verlag, December 1999.

[30] M. Wang and A. Marin. Characterization and prediction of alternative splice sites. *Gene*, 366(2):219–227, February 2006.

[31] I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, second edition, 2005.

[32] E. P. Xing, M. I. Jordan, and R. M. Karp. Feature selection for high-dimensional genomic microarray data. *Proc. 18th International Conf. on Machine Learning*, pages 601–608, 2001.