Prediction of alternatively spliced exons using Support Vector Machines

Jing Xia* and Doina Caragea*

Department of Computing and Information Sciences,
Kansas State University,
234 Nichols Hall, Manhattan, KS 66506, USA
E-mail: xiajing@ksu.edu
E-mail: dcaragea@ksu.edu
*Corresponding authors

Susan J. Brown

Divison of Biology,
Kansas State University,
239A Chalmers Hall, Manhattan, KS 66506, USA
E-mail: sjbrown@ksu.edu

Abstract: Alternative splicing is a mechanism for generating different gene transcripts (called isoforms) from the same genomic sequence. In recent years, it has become obvious that a large fraction of genes undergoes alternative splicing. Finding alternative splicing events experimentally is both expensive and time consuming. Furthermore, traditional transcript-to-genome alignment-based approaches are limited to complete and annotated genomes for which a large amount of transcript data is also available. As a complement to experimental and traditional computational approaches, newer machine learning approaches have been successfully applied to the problem of predicting alternative splicing events. In particular, previous work has shown that machine learning algorithms that utilise sequence features can be used to learn classifiers for distinguishing alternatively spliced exons from constitutively spliced exons. In this paper, we explore the predictive power of a large set of gene features that have been experimentally shown to have effect on alternative splicing. We use such features to build Support Vector Machine (SVM) classifiers for predicting alternatively spliced exons and constitutive exons. Experimental results show that classifiers built from the extended set of features give better results than those that consider only Basic Sequence Features (BSF). Furthermore, we use feature selection methods to identify the most informative features for the prediction problem at hand.

AQ1

Keywords: pre-mRNA alternative splicing; feature construction; splicing motifs; SVMS; support vector machines; bioinformatics.

Reference to this paper should be made as follows: Xia, J., Caragea, D. and Brown, S.J. (xxxx) 'Prediction of alternatively spliced exons using Support Vector Machines', *Int. J. Data Mining and Bioinformatics*, Vol. x, No. x, pp.xxx–xxx.

Biographical notes: Jing Xia received his MS in Computer Science from Kansas State University. Currently, he is a PhD student in the Department of Computing and Information Sciences at Kansas State University. His research interests are in the areas of computational biology and bioinformatics, scientific computing and artificial intelligence.

Doina Caragea is an Assistant Professor in the Department of Computing and Information Sciences at Kansas State University. Her expertise is in the areas of artificial intelligence, machine learning, data mining, information integration and information visualisation, with applications to bioinformatics. Her PhD work at Iowa State University contributed to the design and implementation of a system for knowledge acquisition from autonomous, distributed, semantically heterogeneous data sources. Her recent work has been focused on the development of algorithms and tools for genome annotation.

Susan J. Brown, a full professor in the Division of Biology at Kansas State University, is a developmental geneticist who has been developing genetic and genomic tools for studies on the red flour beetle, *Tribolium castaneum*, which is now second only to *Drosophila melanogaster* among arthropods as a genetic model organism. She is the Director of the KSU Bioinformatics Center and the Director of the Center for Genomic Studies on Arthropods Affecting Human, Animal and Plant Health, leading efforts to provide bioinformatics expertise in genomics to researchers across the KSU campus.

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 (Gilbert, 1978) provided a mechanism through which one gene can generate several distinct proteins. The process is highly regulated by signals and regulators, known or unknown to scientists. Years after its discovery, alternative splicing was still seen more as the exception than the rule (Ast, 2004). Recently, however, it has become obvious that a large fraction of genes undergoes alternative splicing (Graveley, 2001). Early analyses suggested that at least 50% of human genes undergo alternative splicing (Venter et al., 2001a, 2001b). More recent studies have shown that approximately 75% of the human genes appear to be alternatively spliced (Johnson et al., 2003). Furthermore, alternative splicing occurs in many organisms and some instances of alternative splicing have been linked to pathological states such as cancer (Koslowski et al., 2002). These findings underscore the importance of identifying and understanding alternative splicing, both under normal and aberrant conditions. However, the problem of identifying alternative splicing events is particularly intricate, as different transcriptional isoforms can be found in different tissues or cell types, at different development stages or 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 methods for predicting alternative splicing events involve using Expressed Sequence Tags (ESTs) and complementary DNA (cDNA) to recover a gene's structure, and further to detect alternative splicing events based on the recovered information (Nagaraj et al., 2006; Kan et al., 2001). More recent computational approaches rely on machine learning algorithms to identify alternative splicing events based on sequence features (Rätsch et al., 2005; Sorek et al., 2004; Wang and Marin, 2006).

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: the exon and flanking introns lengths and the frame of the stop codon. In particular, Rätsch et al. (2005) 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, their method identifies and outputs predictive *splicing motifs*, which can be used to interpret and understand the classifier results. In this context, a motif is a sequence pattern that occurs repeatedly in a group of related sequences. Thus, the method in Rätsch et al. (2005) 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 (Holste and Ohle, 2008).

The fact that motifs can be used to explain alternative splicing of pre-mRNA 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 (Holste and Ohle, 2008). 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 other organisms (Kabat et al., 2006; Sorek and Ast, 2003; Dror et al., 2005). However, because 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 (Holste and Ohle, 2008). Among these, pre-mRNA secondary structure has been investigated with respect to its ability to help identify patterns that can affect splicing (Hiller et al., 2007; Patterson et al., 2002). The results have shown that the pre-mRNA exhibits local structures that enhance or inhibit the hybridisation of spliceosomal snRNAs to the pre-mRNA. In other words, the structure can affect the selection of the splice sites. Other studies (Wang and Marin, 2006; Fahey and Higgins, 2007) have shown that the strength of the general splice sites constitutes another important sequence feature with respect to the splicing process, as strong splice sites allow the spliceosomes to recognise pairs of splice sites between long introns. When the splice

sites degenerate and weaken, other splicing regulatory elements, such as exon or intron splicing enhancers and silencers, are needed (Pertea et al., 2007). At last, one other feature that has been shown to be correlated with the spicing process consists of the base content in the vicinity of splice sites (Holste and Ohle, 2008).

Although the method in Rätsch et al. (2005) 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) together with other alternative splicing features (pre-mRNA secondary structure, splice site strength, splicing enhancers/silencers and base content) 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 SVM algorithms, to generate classifiers that can distinguish alternatively spliced exons from constitutively spliced 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 non-linear 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 papers is organised 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 Machines

The Support Vector Machine (SVM) algorithm (Vapnik, 1999) is one of the most effective machine learning algorithms for many complex binary classification problems, including a wide range of bioinformatics problems (Guyon et al., 2002; Leslie et al., 2003; Ben-Hur and Brutlag, 2003; Pertea et al., 2007), and has been recently used to detect splice sites (Rätsch and Sonnenburg, 2004; Rätsch et al., 2005; Sonnenburg et al., 2007). The SVM algorithm takes as input labelled data from two classes and outputs a model (a.k.a., classifier) for classifying new unlabelled data into one of those two classes. SVM can generate linear and non-linear models.

Let $E = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_l, y_l)\}$, 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 input space into two half-spaces. The set of such hyperplanes is given by $\{\mathbf{x} \mid \mathbf{x} \cdot \mathbf{w} + b = 0\}$, 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 minimises $\|\mathbf{w}\|^2$, 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 maximised.

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 weights 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 maximise the margin of separation between the two classes. The classification for a new unlabelled example can be obtained from

$$f_{\mathbf{w},b}(\mathbf{x}) = \text{sign}(\mathbf{w} \cdot \mathbf{x} + b) = \text{sign}\left(\sum_{i=1}^{l} \alpha_i y_i (\mathbf{x} \cdot \mathbf{x_i}) + b\right). \quad (1)$$

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 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 maximised, 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,[1] in our study.

2.2 Feature construction

Eight classes of features that affect alternative splicing are considered in our study:

1 Motifs derived from local sequences using the MEME/MAST tools[2] (this set of motifs will be denoted by MEME/MAST).

2 Intronic Regulatory Splicing motifs derived using comparative genomics methods (denoted by IRS).

3 Hexamer motifs derived based on the mutual information between the class of a sequence from which a hexamer is drawn and a random variable over all hexamer occurrences (denoted by MI).

4 Pre-RNA secondary structure related features, specifically

- 4a the Optimal Folding Energy (denoted by OFE)
- 4b a reduced motif set obtained by filtering the MEME/MAST motifs based on secondary structure information (denoted by RMS).

5 Exon Splicing Enhancers (denoted by ESE).

6 Splice Site Strength (denoted by SSS).

7 GC-Content in the vicinity of the splice sites (denoted by GCC).

8 Basic Sequence Features used in Sorek et al. (2004), specifically exon and flanking introns lengths and stop codon frames (together denoted by BSF).

In what follows, we will describe the procedures that were used to construct the above features.

2.2.1 Motifs derived using the MEME/MAST tools

We used MEME (Bailey et al., 2006) and MAST (Bailey and Gribskov, 1998) tools together to detect motifs based on local sequences. MEME is a tool for discovering *unknown* motifs in a group of related DNA or protein sequences. Its underlying algorithm is an extension of the expectation maximisation procedure for fitting finite mixture modols (Bailey and Elkan, 1994). Optimal values for parameters such as the motif width and the number of motif occurrences can be automatically found by MEME. Complementary 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. Thus, MAST can be seen as a post-processing step for MEME, being used to filter the motif list found by MEME. To use the MEME/MAST system, in our study, 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). Next, 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 motif appears in the sequence.

2.2.2 Intronic regulatory motifs derived by comparative genomics

Given that our experimental study is performed on data from *C. elegans*, we derived intronic regulatory motifs based on comparative genomics in Nematodes (Kabat et al., 2006). The hypothesis that supports the use of comparative genomics features is that intronic elements that regulate alternative splicing are under selective pressure for evolutionary conservation, therefore putative motifs derived from these conserved intronic regions might have biological functionality. The basic idea of the comparative genomics procedure used in our study is to identify alternatively spliced exons whose flanking introns exhibit high nucleotide conservation between *C. elegans* and *C. briggsae*. Next, 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. Thus, each upstream intron is represented as a 30-dimensional vector, where each dimension indicates how many times the motif appears in the intron sequence. The same approach is applied to the downstream introns of each exon. Therefore, each sequence is represented as a 60-dimensional vector using this set of features.

2.2.3 Motifs derived based on mutual information

Mutual information represents a measure of the mutual dependence of two random variables (McCallum and Nigam, 1998). To identify motifs based on mutual information, in our study, we first constructed local sequences by considering $(-100, +100)$ bases around the donor and acceptor sites of the sequences in

the original data set (like in the case of MEME/MAST procedure). We used a sliding window to find all hexamer occurrences in our sequences corresponding to donor and acceptor sites, respectively. Then, we calculated the mutual information between the random variable X representing the class of a sequence from which a hexamer is drawn and the random variable Y_h over all hexamer occurrences, using the following formula (multinomial model):

$$I(X;Y_h) = \sum_{x \in X} \sum_{f_h \in H} P(x, f_h) \log \frac{P(x, f_h)}{P(x)P(f_h)}. \quad (2)$$

In our case, the class variable X takes two value $\{-1, 1\}$, where $+1$ corresponds to alternatively spliced exons and -1 corresponds to constitutively spliced exons. Probabilities $P(x, f_h)$, $P(x)$ and $P(f_h)$ are estimated by summing over all hexamer occurrences. Specifically, $P(x)$ is the number of hexamer occurrences appearing in sequences with class label x divided by the total number of hexamer occurrences; $P(f_h)$ is the number of occurrences of hexamer y_h divided by the total number of hexamer occurrences; and $P(x, f_h)$ is the number of occurrences of hexamer y_h that appear in sequences with class label x divided by the total number of hexamer occurrences. We calculated $I(X;Y_h)$ for all possible hexamers y_h and selected the 30 hexamers that have the highest mutual information with the class variable for both sequences corresponding to donor and to acceptor sites. As before, each sequence is represented as a 60-dimensional vector, using this set of features.

2.2.4 Features derived by using secondary structure information

As mentioned above, it is known that the splicing of exons can be enhanced or repressed by specific local pre-mRNA secondary structures around the splice sites (Hiller et al., 2007; Patterson et al., 2002). As shown in Hiller et al. (2007), 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 (Mathews et al., 1999) available at[3] 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 were constructed based on secondary structure information:

a The *Optimal Folding Energy*, which roughly reflects the stability of the RNA folding.

b Under the assumption that motifs on single stranded sequences are more effective than those on helices, a *reduced motif set* was derived from the set of MEME/MAST motifs, by removing the motifs that are located on double stranded sequences with a probability greater than a specific threshold.

In our study, the threshold was set so that 20 motifs were obtained for donors and acceptors, respectively. Therefore, a sequence is represented by a 40-dimensional vector, using this set of motifs.

2.2.5 Derivation of Exon Splicing Enhancers

Although splicing regulators have been identified in both introns and exons, Exon Splicing Regulators (ESR) are more common and better characterised than intron splicing regulators (Cartegni et al., 2002). 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 Pertea et al. (2007) 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 (Pertea et al., 2007), we used a 50-base window around the splice sites to search for ESEs. We also considered the following two assumptions made in the RESCUE-ESE algorithm (Cartegni et al., 2002; Pertea et al., 2007) in our search:

- ESEs appear much more frequently in exons than in introns
- 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:

- $\{|f_E^h - f_I^h| \,|\, h \in$ 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
- $\{|f_W^h - f_S^h| \,|\, h \in$ 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.

2.2.6 Calculating the strength of splice sites

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 (Thanaraj and Stamm, 2003; Wang and Marin, 2006). 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 (Fahey and Higgins, 2007) to model the strength of splice sites, according to the following formula:

$$score = \sum_i \log \frac{F(X_i)}{F(X)} \tag{3}$$

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 (Wang and Marin, 2006). 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 one numerical feature.

2.2.7 Calculating GC-Content features

The **GC Content** (GCC) of a sequence is another feature known to be correlated with the selection of splice sites. Alternatively spliced exons occur more frequently in GC-poor flanking sequences (Thanaraj and Stamm, 2003). We make use of this property by sliding a window to scan the GCC of each motif 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 GC bases to the window size. AQ2

2.2.8 Deriving basis features

Last, but not the least, sequence length has been shown to be a feature that can help to distinguish alternatively spliced exons from constitutive exons (Sorek et al., 2004; Dror et al., 2005). In Rätsch et al. (2005), 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 discretised into a logarithmically spaced vector consisting of 30 bins. The frame of the stop codons is represented using a 3D vector. In this study, use the same approach to derive this set of features, called *basic features*.

2.3 Feature selection methods

Feature selection methods are machine learning techniques used to select the most informative features with respect to a prediction or classication problem. Eliminating redundant or uninformative features can result in enhanced generalisation capability of machine learning algorithms and improved model interpretability. Given the large number of features in our study, to select the most informative features, we used the following two methods:

1 SVM variable importance (Guyon et al., 2002)

2 information gain (Xing et al., 2001).

The weight vector $\mathbf{w} = \{|w_0|, |w_1|, \ldots, |w_n|\}$ (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.

As we have seen above, the mutual information (or equivalently, the expected information gain) provides a simple way to decide the predictive value of a set of features (Mitchell, 1997). Thus, one can rank all features in the order of increasing information gain and select features conservatively. A more robust way is to use a decision tree algorithm, which iteratively selects the feature with the highest gain to build a decision tree. We select those features that are nodes in the decision tree built by considering all features (as those are assumed to be the most informative).

3 Experimental results

3.1 Dataset

The data set used in our experiments contains alternatively spliced and constitutive exons in *C. elegans*. It has been used in related work (Rätsch et al., 2005) and is available at.[4] A detailed description of how this data set was generated can also be found in Rätsch et al. (2005). 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 Motifs evaluation

The purpose of the motif evaluation experiment in this section is to identify the splicing motifs that appear in several different motif sets, as those motifs are presumably the most informative for alternative splicing.

MEME/MAST vs. hexamer motifs vs. IRS set. We first compared the set of 40 motifs identified by MEME/MAST with the set of hexamer motifs found in Rätsch et al. (2005) and the ISR motifs found in Kabat et al. (2006). The MEME/MAST motifs are shown in Table 1 in the form of regular expressions that correspond to multilevel consensus sequences determined by MEME. Figure 1 shows a sample output for MEME, including the multilevel consensus sequence representation of a motif and the corresponding regular expression. 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. As it can be seen in Table 1, eight motifs are found in all three sets compared, some of them (e.g., mast2 and mast3) being highly conserved among the *C. elegans* sequences in our data set.

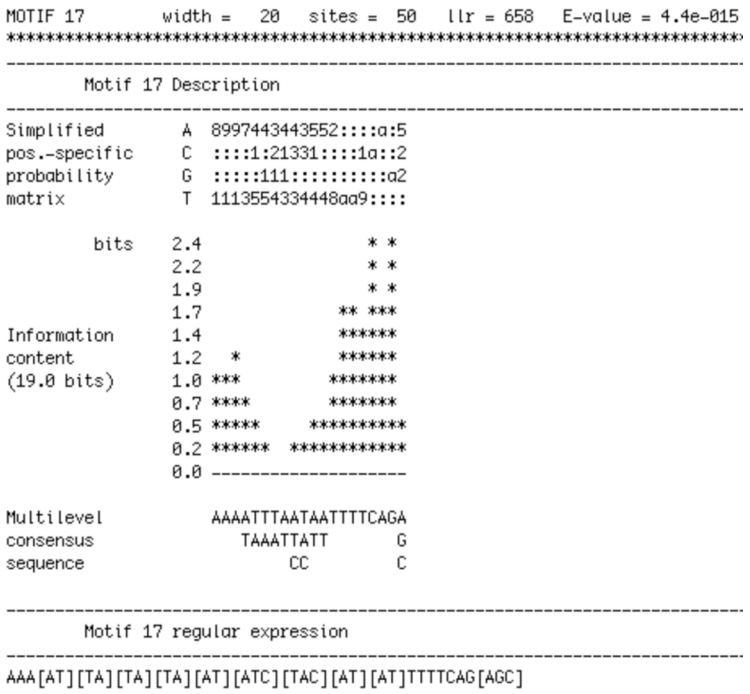
ESE hexamers vs. human and mouse hexamers. Second, we compared the 77 ESE hexamers, found as described in Section 2.2.5, with two sets of candidate human and mouse ESE hexamers proposed in Yeo et al. (2002). Thirty two out of the 77 putative *C. elegans* ESE hexamers occur also in the human and mouse 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 (Pertea et al., 2007) was also 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.

Table 1 Intersection of motifs identified by the MEME/MAST tools, hexamer motifs identified in Rätsch et al. (2005) and Intronic Regulator Splicing (IRS) motifs identified in Kabat et al. (2006). MEME/MAST motifs 1-20 are around 5′ splice sites, while motifs 21-40 are around 3′ splice sites. IRS motifs are italicised

MEME/MAST motifs (*Regular expression*)	*Contained hexamers*
mast2: TTTTTTTTTCA	tttttt
mast3: GTGAGTTTTT[TA]	tttttt
mast4: [AT][AT]A[AT][AT][TA][TA][TA][ATC][AT] [AT]TTTTCAG[GA]	tttttt, atatat tatata
mast6: [AT]TTTT[TC]C[AT]AA[TA]TTT[TC]	tttttt
mast9: [GA]C[CA]G[GC][TC]G[GC][AC]G[CG][TC]G[TG][CT][GA] [TGC][ACT]G[GAC]	gttgtc, *catcgc gtgttg*
mast14: [AC][GA][CT][CTA]G[CT][CA]G[AC]A[GA][CA][CA][CG] [TC][TC]G[CG][CA][AG]	gttgtc , *ccctgg catcgc, cactgc*
mast22: C[AT][GCT]C[AT][CGA]CA[AG]C[AT][GTACC]CA[CG] CA[CA]CA	cagcag
mast23: [TA]TTTTTTTTT[CT][AG][AG]A[AT]TTT[TC][AT]	tttaaa, aatttt atttta

Figure 1 MEME output. The width, number of occurrences in the training set (sites), log likelihood ratio (llr), and the E-value of the motif shown on the top line. The multilevel consensus sequence and the corresponding regular expression shown at the bottom



It is worth mentioning that our analysis did not find any common motifs between the IRS set and the ESE set in *C. elegans*, suggesting that the two sets are functionally different.

3.3 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: cost of the 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), kernel parameters (e.g., the degree or coefficients of the polynomial kernel), etc.

Rätsch et al. (2005) have used basic features with several types of customised kernels, as well as an optimal sub-kernel 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. We show that simple linear kernels that use splicing motifs as input features, can produce results similar to those in Rätsch et al. (2005). In order to tune the cost C, we used 5-fold cross-validation for each training set, with $C \in \{0.01, 0.05, 0.1, 0.5, 1, 2\}$. We chose the value of C for which the Area Under Curve (AUC) was maximised during the cross-validation. The classifier was trained on each training set with the selected cost value, respectively, and was evaluated on the corresponding test set. In preliminary work, we also tried Radial Basis Function kernel, with gamma $g = 1/k$, where k is the size of the data and $C \in \{1, 3, 5, 7, 11\}$, and obtained approximately equivalent results as for the linear kernel.

To evaluate the performance of the classifiers trained, we used ROC curves and AUC values, which capture the trade off between true positive rate and false positive rate. The true positive rate is given by the number of positively labelled examples classified by the algorithm as positive, divided by the total number of positive examples. The false positive rate is the number of negatively labelled examples classified as positive, divided by the total number of negatively labelled examples. We report the true positive rate, when the false positive rate reaches 1% (in following tables, the entry fp 1% shows such results).

3.4 Feature selection

To identify the most relevant features, 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 the weight vector $\mathbf{w}$. Table 2 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 Wang and Marin (2006) and Sorek et al. (2004), respectively. However, taken separately, the SSS features do not discriminate well between alternatively spliced and constitutive

exons (results not shown), suggesting that they encode information complementary to the information encoded in the BSF features.

Table 2 Weight importance of the following features: 105 BSF, 1 SSS, 80 GCC, 60 IRS, 40 MEME/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

In previous work (Xia et al., 2008), we have shown that IRS, MEME/MAST and ESE motifs 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 (results not shown).

Furthermore, we also ran the J48 decision tree algorithm in the data mining package WEKA (Witten and Frank, 2005) to build a classifier for each data set. We analysed 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 3 shows the list of best motifs according to the information gain criterion. 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 3). We also noticed that ese65 (gatgat) was the most frequent hexamer among the selected ESE motifs.

3.5 Performance results

As the mutual information motifs have not been used in our previous work (Xia et al., 2008), in this section we evaluate the performance of this set of motifs with respect to the classification problem at hand (when used by themselves or in combination with other features). In particular, we will use the GCC and Strength of Splice Sites (SSS) features together with the mutual information motifs, as these two features have been shown to have high SVM importance (Table 2). First, we did a comparison between classifiers trained with motifs obtained based on mutual information with BSF and classifiers trained with BSF, respectively. Table 4 shows the classification results when motifs features derived from mutual information are combined with BSF. Figure 2 shows a comparison between classifiers trained with motifs obtained based on mutual information and classifiers trained with BSF, respectively.

Table 3 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. Motifs irs21, irs23, irs31 are IRS motifs identified by both J48 and SVM as important. Weight values and ranking are 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

Table 4 Results of alternatively spliced exons classification when BSF and mutual information motifs are used as features

		Validation score		*Test score*	
	C	*fp 1%*	*AUC*	*fp 1%*	*AUC*
Split1	2	48.81%	89.83%	50.93%	89.25%
Split2	2	48.57%	88.80%	56.12%	92.20%
Split3	2	51.27%	88.78%	53.76%	90.27%
Split4	2	53.55%	89.88%	50.54%	90.38%
Split5	0.5	52.55%	89.97%	55.79%	89.34%

Second, Table 5 shows the results when GCC and SSS features are also included.

Tables 4 and 5 show that, on the average, the performance improves in terms of true positive rate at 1% false positive rate, when more features are included, which means that GCC and SSS features contribute to better classification performance (as expected). Furthermore, on the average, the results are comparable and sometimes better than the results obtained by Rätsch et al. (2005). For example, the average AUC is 90.98%, thus improving the previous AUC average which was 90.48%. Furthermore, the true positive rate 57.81% at false positive rate 1% has also increased by approximately 10% (previous value was 48.47%).

Figure 3 shows the result of comparison between a data set with basic sequence features only and a data set that includes the other features (motifs from mutual information, GCC and SSS). The results show significant improvement when these additional features are used.

Figure 2 Comparison between classifiers trained with mutual information motifs and classifiers trained with Basic Sequence Features (BSF) on the five split data sets

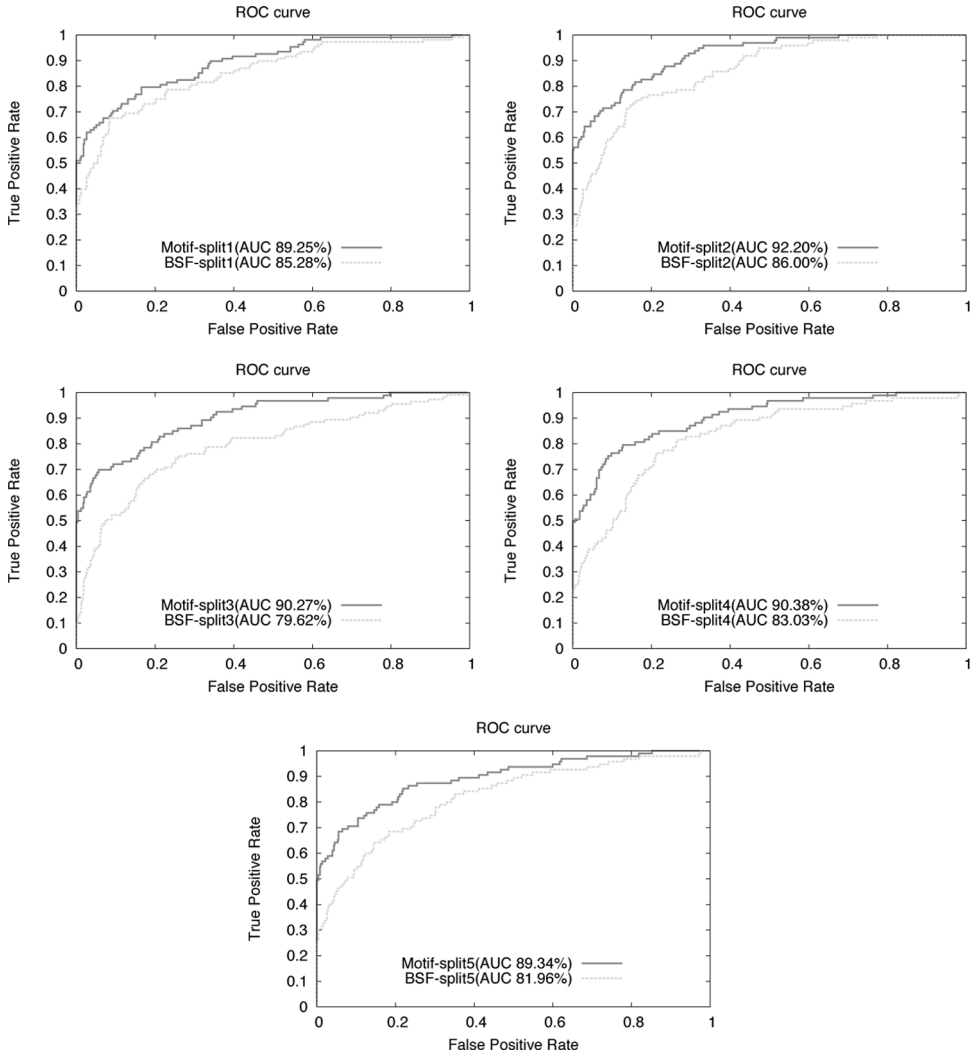
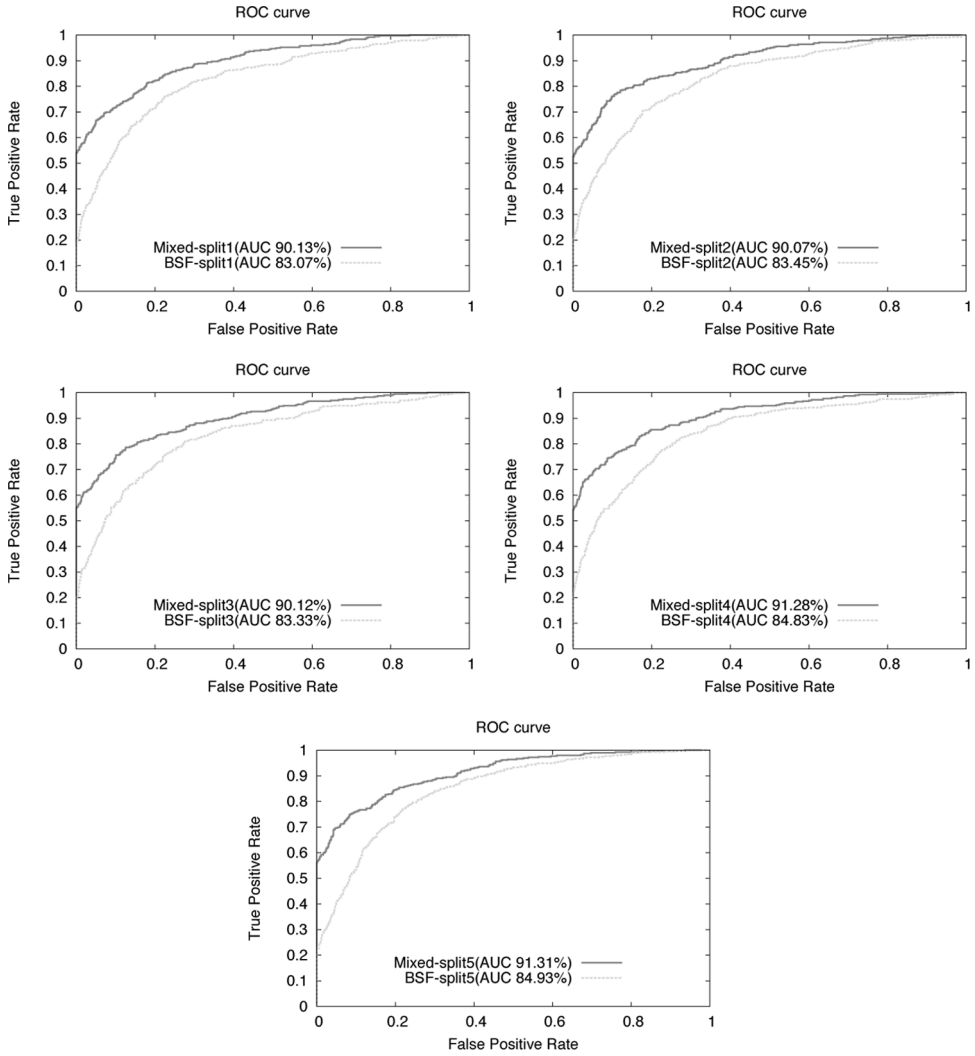


Table 5 Results of alternatively spliced exons classification when mutual information motifs, **GC Content** and Strength of Splice Sites are used as features

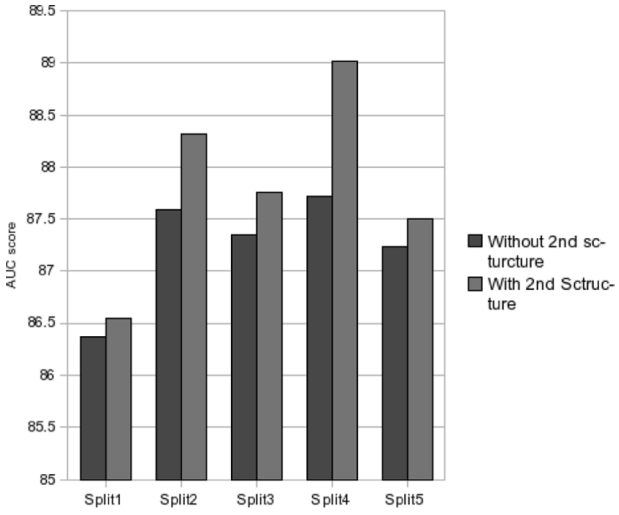
		Validation score		*Test score*	
	C	*fp 1%*	*AUC*	*fp 1%*	*AUC*
Split1	1	56.46%	90.13%	60.19%	92.10%
Split2	2	55.53%	90.07%	59.18%	92.36%
Split3	1	57.11%	90.12%	52.69%	90.98%
Split4	0.5	54.31%	91.28%	58.06%	90.30%
Split5	2	58.67%	91.31%	58.94%	89.25%

Figure 3 Comparison of ROC curves obtained using basic features only and basic features plus other features (motifs from mutual information, GCC and SSS), together denoted as mixed features. Models trained by 5-fold CV with parameters listed in Table 5



Finally, 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 and the other using data sets without secondary structure features. Figure 4 shows the results of the two experiments in which the classifiers were trained using 5-fold cross-validation with optimal cost parameters. We can see the improvement obtained when considering secondary structure features.

Figure 4 AUC scores comparison between data sets with features of secondary structure and data sets without features of secondary structure



4 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 work in this area (Dror et al., 2005; Rätsch et al., 2005; Sorek and Ast, 2003; Sorek et al., 2004). However, there has been no comprehensive computational study that considers all the features 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. Several methods 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 (Rätsch et al., 2005). 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. 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.

Acknowledgements

This work is supported by the National Science Foundation under Grant No. 0711356 to Doina Caragea. We thank Dr. William H. Hsu for providing financial support for Jing Xia.

References

Ast, G. (2004) 'How did alternative splicing evolve?', *Nat. Rev. Genet.*, Vol. 5, No. 10, pp.773–782.

Bailey, T. and Elkan, C. (1994) 'Fitting a mixture model by expectation maximization to discover motifs in biopolymers', *Proc. 2nd International Conf. on Intelligent Systems for Molecular Biology*, Vol. 271, No. 50, pp.28–36. **AUTHOR PLEASE SUPPLY LOCATION.** Stanford, California

Bailey, T. and Gribskov, M. (1998) 'Combining evidence using *p*-values: application to sequence homology searches', *Bioinformatics*, Vol. 14, No. 1, pp.48–54.

Ben-Hur, A. and Brutlag, D. (2003) 'Remote homology detection: a motif based approach', *Bioinformatics*, Vol. 19, Suppl. 1, pp.23–26.

Bailey, T., Williams, N., Misleh, C. and Li, W. (2006) 'MEME: discovering and analyzing DNA and protein sequence motifs', *Nucleic Acids Research*, Vol. 34, pp.369–373.

Cartegni, L., Chew, S. and Krainer, A. (2002) 'Listening to silence and understanding nonsense: exonic mutations that affect splicing', *Nature Reviews Genetics*, Vol. 3, No. 4, pp.285–298.

Dror, G., Sorek, R. and Shamir, R. (2005) 'Accurate identification of alternatively spliced exons using support vector machine', *Bioinformatics*, Vol. 21, No. 7, pp.897–901.

Fairbrother, W., Yeh, R.F., Sharp, P. and Burge, C. (2002) 'Predictive identification of exonic splicing enhancer motifs in human protein-coding genes', *Science*, Vol. 297, No. 5583, pp.1007–1013. **AUTHOR PLEASE SUPPLY LOCATION.** Washington, DC

Fahey, E. and Higgins, D.G. (2007) 'Gene expression, intron density and splice site strength in *drosophila* and *caenorhabditis*', *Journal of Molecular Evolution*, Vol. 65, No. 3, pp.349–357.

Gilbert, W. (1978) 'Why genes in pieces?', *Nature*, Vol. 271, No. 5645, p.50.

Graveley, B. (2001) 'Alternative splicing: increasing diversity in the proteomic world', *Trends Genet.*, Vol. 17, No. 2, pp.100–107.

Guyon, I., Weston, J., Barnhill, S. and Vapnik, V. (2002) 'Gene selection for cancer classification using support vector machines', *Machine Learning*, Vol. 46, pp.389–422.

Hiller, M., Zhang, Z.Y., Backofen, R. and Stamm, S. (2007) 'Pre-mRNA secondary structures influence exon recognition', *PLoS Comput. Biol.*, Vol. 3, No. 11, p.204.

Holste, D. and Ohle, U. (2008) 'Strategies for identifying RNA splicing regulatory motifs and predicting alternative splicing events', *PLoS Comput. Biol.*, Vol. 4, No. 1, p.21.

Johnson, J.M., Castle, J., Garrett-Engele, P., Kan, Z., Loerch, P.M., Armour, C.D., Santos, R., Schadt, E.E., Stoughton, R. and Shoemaker, D.D. (2003) 'Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays', *Science*, Vol. 302, No. 5653, pp.2141–2144.

Koslowski, M., Türeci, O., Bell, C., Krause, P., Lehr, H.A., Brunner, J., Seitz, G., Nestle, F.O., Huber, C. and Sahin, U. (2002) 'Multiple splice variants of lactate dehydrogenase C selectively expressed in human cancer', *Cancer Research*, Vol. 62, No. 22, pp.6750–6755.

Kan, Z.Y., Rouchka, E.C., Gish, W.R. and States, D.J. (2001) 'Gene structure prediction and alternative splicing analysis using genomically aligned EST', *Genome Res.*, Vol. 11, No. 5, pp.889–900.

Kabat, J.L., Barberan-Soler, S., McKenna, P., Clawson, H., Farrer, T. and Zahler, A.M. (2006) 'Intronic alternative splicing regulators identified by comparative genomics in nematodes', *PLoS Comput. Biol.*, Vol. 2, No. 7, p.86.

Leslie, C., Eskin, E., Cohen, A., Weston, J. and Noble, W. (2003) 'Mismatch string kernels for discriminative protein classification', *Bioinformatics*, Vol. 20, No. 4, pp.467–476.

McCallum, A. and Nigam, K. (1998) 'A comparison of event models for naive bayes text classification', *AAAI-98 Workshop on Learning for Text Categorization*. **AUTHOR PLEASE SUPPLY LOCATION AND PAGE RANGES.** Madison, Wisconsin. Tech. Rep. WS-98-05,

Mathews, D., Sabina, J., Zuker, M. and Turner, D. (1999) 'Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure', *Journal of Molecular Biology*, Vol. 288, No. 5, pp.911–940.

Maniatis, T. and Tasic, B. (2002) 'Alternative pre-mRNA splicing and proteome expansion in metazoans', *Nature*, Vol. 418, pp.236–243. **AUTHOR PLEASE CITE THIS REFERENCE IN TEXT.**

Mitchell, T. (1997) *Machine Learning*, McGraw-Hil. **AUTHOR PLEASE SUPPLY LOCATION.** Boston, MA

Nagaraj, S.H., Gasser, R.B. and Ranganathan, S. (2006) 'A hitchhiker's guide to expressed sequence tag (EST) analysis', *Brief Bioinform.*, Vol. 8, No. 1, pp.6–21.

Patterson, D.J., Yasuhara, K. and Ruzzo, W.L. (2002) 'PRE-mRNA secondary structure prediction aids splice site prediction', *Proceedings of the Pacific Symposium on Biocomputing*, pp.223–234. **AUTHOR PLEASE SUPPLY LOCATION.** Lihue, Hawaii

Pertea, M., Mount, S.M. and Salzberg, S.L. (2007) 'A computational survey of candidate exonic splicing enhancer motifs in the model plant arabidopsis thaliana', *BMC Bioinformatics*, Vol. 8, p.159.

Rätsch, G., Sonnenburg, S. and Schölkof, B. (2005) 'RASE: recognition of alternatively spliced exons in C. elegans', *Bioinformatics*, Vol. 21, Suppl. 1, pp.369–377.

Rätsch, G. and Sonnenburg, S. (2004) *Kernel Methods in Computational Biology*, MIT Press, pp.277–298. **AUTHOR PLEASE SUPPLY LOCATION.** Cambridge, MA

Sorek, R. and Ast, G. (2003) 'Intronic sequences flanking alternatively spliced exons are conserved between human and mouse', *Genome Research*, Vol. 13, No. 7, pp.1631–1637.

Sorek, R., Shemesh, R., Cohen, Y., Basechess, O., Ast, G. and Shamir, R. (2004) 'A non-EST based method for exon-skipping prediction', *Genome Res.*, Vol. 14, No. 8, pp.1617–1623.

Sonnenburg, S., Schweikert, G., Philips, P., Behr, J. and Rätsch, G. (2007) 'Accurate splice site prediction using support vector machines', *BMC Bioinformatics*, Vol. 8, Suppl. 10, p.7.

Thanaraj, T.A. and Stamm, S. (2003) 'Prediction and statistical analysis of alternatively spliced exons', *Prog. Mol. Subcell. Biol.*, Vol. 31, pp.1–31.

Vapnik, V.N. (1999) *The Nature of Statistical Learning Theory Statistics for Engineering and Information Science*, Springer-Verlag. **AUTHOR PLEASE SUPPLY LOCATION AND PAGE RANGE.** New York, NY, USA. pp. 1-314

Venter *et al.* (2001a) 'Initial sequencing and analysis of the human genome', *Nature*, The International Genome Sequencing Consortium, Vol. 409, No. 6822, pp.860–921. **AUTHOR PLEASE CHECK IF THE REFERENCE IS OK AND SUPPLY HIGHLIGHTED AUTHOR NAMES.** Eric S. Lander, Lauren M. Linton

Venter **et al.** (2001b) 'The sequence of the human genome', *Science*, Vol. 291, No. 5507, pp.1304–1351. **AUTHOR PLEASE SUPPLY HIGHLIGHTED AUTHOR NAMES.**

Witten, I.H. and Frank, E. (2005) *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann. **AUTHOR PLEASE SUPPLY LOCATION AND PAGE RANGE.** San Fransisco, CA pp. 1-525

Wang, M. and Marin, A. (2006) 'Characterization and prediction of alternative splice sites', *Gene*, Vol. 366, No. 2, pp.219–227.

Xing, E.P., Jordan, M. and Karp, R. (2001) 'Feature selection for high-dimensional genomic microarray data', *Proceedings of the Eighteenth International Conference on Machine Learning*, pp.601–608. **AUTHOR PLEASE SUPPLY LOCATION.** San Francisco, CA, USA

Xia, J., Caragea, D. and Brown, S. (2008) 'Exploring alternative splicing features using support vector machine', *Proceedings of IEEE Bionformatics and Biomedicine*, pp.230–238. **AUTHOR PLEASE SUPPLY LOCATION.** Philadelphia, PA

Yeo, G., Hoon, S., Venkatesh, B. and Burge, C.B. (2002) 'Variation in sequence and organization of splicing regulatory elements in vertebrate genes', *Proceedings of the National Academy of Sciences*, Vol. 101, No. 44, pp.15700–15700. **AUTHOR PLEASE SUPPLY LOCATION AND CHECK THE PAGE RANGE.**

Notes

[1]http://www.csie.ntu.edu.tw/cjlin/libsvm/
[2]http://meme.sdsc.edu/meme/intro.html
[3]http://mfold.bioinfo.rpi.edu/
[4]http://www.fml.tuebingen.mpg.de/raetsch/projects/RASE

Query

AQ1: AUTHOR PLEASE REDUCE ABSTRACT OF NO MORE THAN 100 WORDS.
AQ2: AUTHOR PLEASE CHECK IF THE EXPANSION FOR THE ACRONYM 'GCC' IS GC CONTENT OR GC-CONTENT.