BIOINFORMATICS ORIGINAL PAPER

Vol. 21 no. 7 2005, pages 897–901
doi:10.1093/bioinformatics/bti132

Genome analysis

Accurate identification of alternatively spliced exons using support vector machine

Gideon Dror[1,*], Rotem Sorek[2,3] and Ron Shamir[4]

[1]The Academic College of Tel-Aviv-Yaffo, Tel Aviv 4044, Israel, [2]Department of Human Genetics, Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv 69978, Israel, [3]Compugen, Tel Aviv 69512, Israel and [4]School of Computer Science, Tel Aviv University, Tel Aviv 69073, Israel

Received on September 7, 2004; revised and accepted on November 1, 2004
Advance Access publication November 5, 2004

ABSTRACT

Motivation: Alternative splicing is a major component of the regulatory action on mammalian transcriptomes. It is estimated that over half of all human genes have more than one splice variant. Previous studies have shown that alternatively spliced exons possess several features that distinguish them from constitutively spliced ones. Recently, we have demonstrated that such features can be used to distinguish alternative from constitutive exons. In the current study, we used advanced machine learning methods to generate robust classifier of alternative exons.

Results: We extracted several hundred local sequence features of constitutive as well as alternative exons. Using feature selection methods we find seven attributes that are dominant for the task of classification. Several less informative features help to slightly increase the performance of the classifier. The classifier achieves a true positive rate of 50% for a false positive rate of 0.5%. This result enables one to reliably identify alternatively spliced exons in exon databases that are believed to be dominated by constitutive exons.

Availability: Upon request from the authors.

Contact: gideon@mta.ac.il

1 INTRODUCTION

Alternative splicing is a process through which one gene can generate several distinct proteins. It occurs by the alternative usage of exons or parts of exons within pre-mRNA transcripts, and can be specific to a tissue, developmental stage or a condition such as stress (Maniatis and Tasic, 2002).

Computational prediction of alternative splicing usually involves the usage of expressed sequences, i.e. expressed sequence tags (ESTs) or cDNAs [reviewed in Graveley (2001) and Modrek and Lee (2002)]. Through such predictions, in addition to microarray analyses, several studies have estimated that alternative splicing occurs in 35–74% of all human genes (Brett *et al*., 2000; Kan *et al*., 2001, 2002; Lander *et al*., 2001; Mironov *et al*., 1999; Modrek *et al*., 2001; Johnson *et al*., 2003). However, ESTs and microarrays produce only a snapshot of the tissue they sample, in the specific time and condition it was sampled. Exons that are alternatively spliced in conditions other than the ones sampled will evade detection.

Recently, we have described several features in which alternative exons differ from constitutive ones. These features include the size of the exon, its divisibility by 3, the identity level when aligned to its mouse ortholog exon and the human/mouse conservation in the intronic sequences flanking the exon (Sorek and Ast, 2003; Sorek *et al*., 2004a). Using brute-force enumeration, we demonstrated that a combination of these features could be used to classify alternative exons with a true positive rate of $\sim$30% for a false positive rate $<$1%, regardless of their representation in ESTs (Sorek *et al*., 2004b).

In the current study we use state-of-the-art machine learning methods, along with additional sequence features, to generate a robust classifier of alternative exons. We achieved better sensitivity for a similar specificity performance—a true positive rate of 50% for a false positive rate of 0.5%. Moreover, not only is our performance measure more robust, but also we get much higher area under the ROC curve, which provides a proper measure for the quality of ranking of a classifier (Ling *et al*., 2003). We also report on the merit of many additional sequence features extracted from the vicinity of the exon.

2 METHODS

2.1 Dataset

The dataset comprises of 243 alternative and 1753 constitutive exons that are conserved between human and mouse. The data are described in detail in our previous studies (Sorek and Ast, 2003; Sorek *et al*., 2004b). Briefly, alternative exons in this set are exons that were found to be skipped both in the human and mouse transcriptomes; and constitutive exons are exons that are supported by at least four expressed sequences, with no evidence of ESTs skipping them, both in human and in mouse.

2.2 Data representation

For the current study, we used the seven features described in our previous study, as well as 221 additional new sequence features. The original features were (1) exon length, (2) exon divisibility by 3 (a Boolean feature), (3) percent identity when aligned to the mouse counterpart and (4) conservation in the upstream and downstream intronic sequences. Each of the two ‘intronic conservation’ features (upstream and downstream) was divided into two sub-features: (1) length of the best human/mouse local alignment in the 100 intronic nucleotides nearest to the exon (where only local alignments with at least 12 consecutive perfectly matching nucleotides were considered) and (2) identity level in this local alignment. Local alignments were performed using sim4 (Florea *et al*., 1998) as described by Sorek *et al*. (2004b).

Additional features tested here include 3-tuple counts, computed separately for the sequence of the exon, the 100 bases of the intron upstream of the exon

[*]To whom correspondence should be addressed.

(called 'pre') and the 100 bases of the intron downstream of the exon (called 'post'), adding up to $64 \times 3 = 192$ features.

We also used information from the 5′ splice site (5′ss, also called donor site) sequence. The nucleotide composition of the 5′ss reflects its base-pairing with small nuclear RNAs such as U1 (Zhuang and Weiner, 1986). It was previously shown that the composition of the 5′ss differs between alternative and constitutive exons (Clark and Thanaraj, 2002). It was also demonstrated that the alteration of 5′ss sequences can result in transition from alternative to constitutive splicing, or vice versa (Sorek *et al.*, 2004b). We therefore used position-dependent single base counts at the 5′ss sequence, ranging from the -3 to the $+6$ position relative to the splice site (not including invariant positions $+1$ and $+2$). This added up to $4 \times 7 = 28$ features.

The last feature used was the intensity of the poly-pyrimidine tract (PPT), which was defined as the number of pyrimidines (Cs and Ts) in a window of 15 bases in the last 19 nt of the upstream intron (not including the last 4 nt of the intron).

We also examined position-dependant base combinations of three bases at the splice site, that were shown to be the highly discriminating features for a similar task (Zhang *et al.*, 2003). However, preliminary analysis has indicated that the 3-base 5′ss combinations are not as informative for the present task and were therefore not included in the sequel.

We concatenated all features into one vector representation in $\mathbb{R}^N$ where $N = 7 + 192 + 28 + 1 = 228$. Since the features have very different distributions (binary, integer and real numbers), we standardized them such that each feature has a zero mean and variance of one. We denote the i-th standardized vector by $\mathbf{x}^i = (x^i_1, \ldots, x^i_N)$. Each example is labeled by $y^i = -1$ or $y^i = +1$, depending on whether it represents a constitutive or alternative exon, respectively.

2.3 Data partitioning

In the experiments reported here, we randomly split the dataset entries into a training and a testing set at a ratio of 2:1. Feature vectors as described above were used as examples for training various classifiers, while the testing examples were not exposed to the system during learning, feature selection and hyper-parameter selection phases.

2.4 Support vector machines

Support vector machine (SVM) learning is an area of statistical learning, subject to extensive research (Vapnik, 1998; Schölkopf *et al.*, 1999; Smola *et al.*, 2000). SVM has been used extensively for a wide range of applications in science, medicine and engineering and has shown excellent empirical performance. Recent bioinformatic investigations utilizing SVM include Brown *et al.* (1999), Zien *et al.* (1999), Jaakkola *et al.* (2000) and Leslie *et al.* (2004). More recently, SVM was used for the detection of splicing sites (Yamamura and Gotoh, 2003; Sun *et al.*, 2003; Zhang *et al.*, 2003). SVM has the following advantages for the present task:

(1) SVM is based on the principle of risk minimization and thus provides good generalization control. This allows one to work with datasets that contain many irrelevant and noisy features.

(2) Using non-linear kernels, SVM can model non-linear dependences among features and the target, which may prove advantageous for the problem at hand.

(3) SVM allows natural control on the relative cost of false positives and false negatives.

In the present research we used soft-margin SVM implemented in SVM$^{\text{light}}$ (Joachims, 1999). The latest version of this software is available at http://svmlight.joachims.org/.

2.5 Hyper-parameter selection

SVM training involves fixing several hyper-parameters. The values of these hyper-parameters determine the function that SVM optimizes and therefore have a crucial effect on the performance of the trained classifier. To identify an optimal hyper-parameter set we used 10-fold cross-validation on the training set, which is a robust method for hyper-parameter tuning (Duan *et al.*, 2003). The cross-validation was used also to tune the number of features used by the classifier, as discussed in the next subsection.

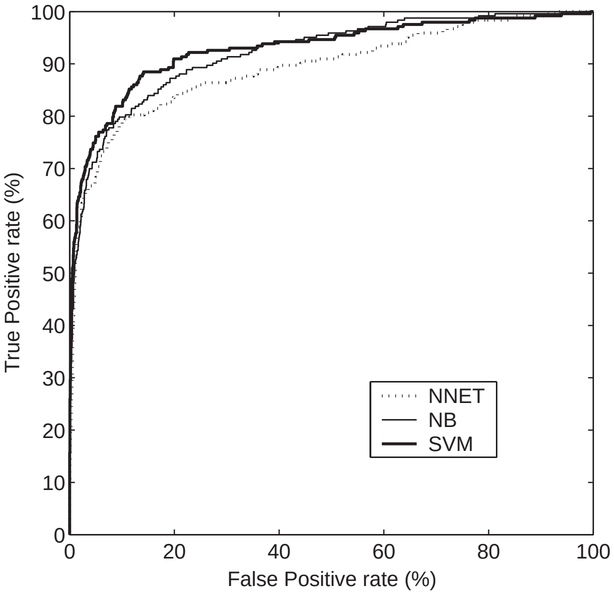


Fig. 1. The ROC curves of the three classifiers. The AUC for neural network (NNET), the Naive-Bayes (NB) and the SVM are 0.92, 0.89 and 0.93, respectively. The optimal performance of the first two classifiers was obtained with 11 features. SVM classifier uses a linear kernel with hyper-parameters $c = \sqrt{10}$ and $j = 0.5$. The SVM ROC with these hyper-parameters is quite insensitives to the number of features, as is shown below.

We used several kernels: linear, polynomials of degrees 2 and 3 and Gaussian kernel. For each kernel, we performed a grid search over the values of the slack parameter c, and the cost factor j, by which training errors on positive examples (false negatives) outweigh errors on negative examples (false positives). For the Gaussian kernel, we repeated the search for several values of r, the parameter that controls the width of the kernel.

For each hyper-parameter combination we measured the 10-fold cross-validation area under the ROC curve (AUC). AUC (Agarwal *et al.*, 2004) is a global performance measure since it is integrated over all threshold values. However, for the task of identifying alternative exons within a population in which the vast majority of exons are constitutive, one specifically needs high discrimination power at low false-positive rate. To this end, we also measured the true positive rate for a small value $0 < \alpha \ll 1$ of the false positive rate. We denote this performance measure as TP_α. For small values of α, TP_α is very sensitive to the minute details in the distribution of examples (e.g. the details of split between the training set and test set). Therefore we did not directly try to maximize it, so as to reduce the risk of severe overfitting. We selected the kernel and hyper-parameter set that gave the highest value of $\lambda\text{AUC} + (1 - \lambda)\text{TP}_\alpha$, where $0 < \lambda < 1$. We found that for the whole range $0.1 < \lambda < 0.9$ there was very good generalization, and that the final results varied only insignificantly.

The best cross-validation performance, for a value of $\lambda = 0.5$, was obtained by the Gaussian kernel, with intermediate slack parameter $c = \sqrt{10}$ and cost factor $j = 1/2$.

In addition to SVM, we also used Naive-Bayes and neural network classifiers. For training the neural network we used the Levenberg–Marquardt algorithm with Bayesian regularization. For both Naive-Bayes and neural network, we performed a search in hyper-parameter space and among several architectures to optimize performance. Figure 1 shows the ROC curves of the best SVM, Naive-Bayes and neural network classifiers. The values of AUC are quite close to each other.

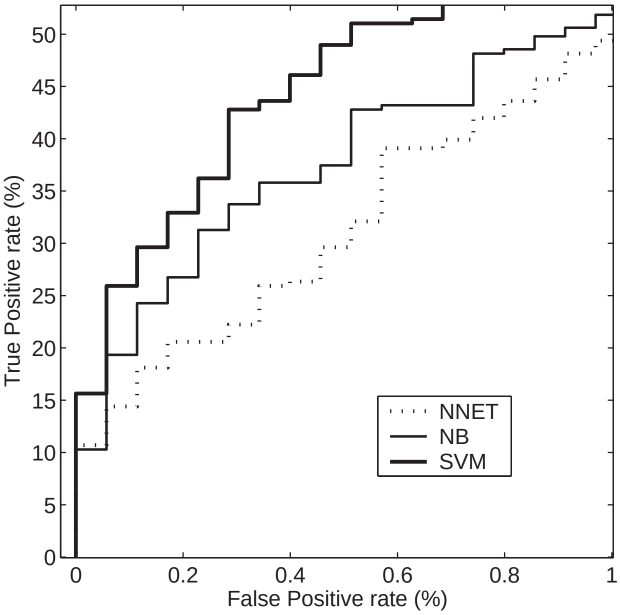


Fig. 2. The ROC curve of the three classifiers in the region of small false positive rate, FP $< 1\%$. It is evident that SVM considerably outperforms the neural network (NNET) and the Naive-Bayes (NB) classifiers.

Figure 2 depicts the ROCs of the three classifiers at the region of low false positive rate. It is clear that for all the ranges shown, $0 < \alpha < 1\%$, SVM achieves considerably higher TP_α and therefore better resolution in identifying alternative exons.

2.6 Feature selection

The potential benefits of feature selection are 3-fold: improving the performance of the classifier, producing a cost-effective classifier, and providing better understanding of the problem at hand. In our case, we used feature selection primarily for the purpose of enhancing the classifier's performance. Although state-of-the-art classifiers such as SVM and neural networks that incorporate regularization techniques can accommodate situations where many of the features are redundant or noisy, removing non-informative features can considerably enhance their performance (Guyon and Elisseeff, 2003).

Preliminary analysis of the data has shown that the seven features used in the original paper by Sorek *et al.* (2004c) are much more informative for the classification task than the vast majority of the remaining features. However, a χ^2-test showed that for several features, the distributions of the positive and negative examples are significantly different. Namely, they potentially convey useful information for the task of classification.

Our feature selection criterion is that used by Golub *et al.* (1999). For each feature x_j, $j = 1, \ldots, N$, we calculated the mean $\mu_j^+ (\mu_j^-)$ and standard deviation $\sigma_j^+ (\sigma_j^-)$ using only positive (negative) examples. The score

$$F(x_j) = \left| \frac{\mu_j^+ - \mu_j^-}{\sigma_j^+ + \sigma_j^-} \right| \tag{1}$$

serves as a simple heuristic for ranking the features according to how well they discriminate the positive and negative examples.

To avoid overfitting, we used the feature selection within the cross-validation loop. In other words, to estimate the performance of a classifier which uses n features, where $n \leq N$, we used Equation 1 on each split of the training set and simply took the n features with the highest $F(x_j)$ scores. Needless to say that this procedure produces a unique feature set for each split.

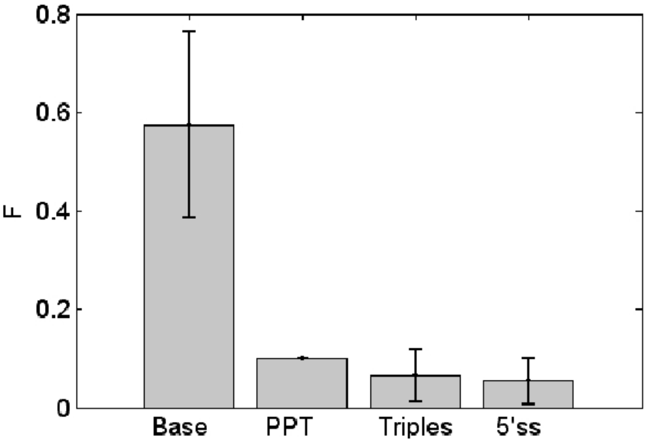


Fig. 3. The discriminative power of different feature types. For each set of features we plot the average values of F. Feature sets are: the original features of Sorek *et al.* (2004b) (Base—7 features), intensity of poly-pyrimidine tract (PPT—1 feature), triple counts (Triples—192 features) and position-dependent single base counts at the 5′ss (5′ss—24 features). The standard deviation of F within each set is expressed by the error bars.

Table 1. Most informative triples

Triple	Location	$\mu^+(\sigma^+)$	$\mu^-(\sigma^-)$	F	P-value
TTC	Pre	0.033 (0.021)	0.026 (0.016)	0.215	5.77E−7
AGG	Post	0.014 (0.017)	0.022 (0.020)	0.212	1.13E−9
GAG	Pre	0.008 (0.012)	0.014 (0.015)	0.210	5.94E−9
AGG	Pre	0.010 (0.014)	0.015 (0.016)	0.186	3.30E−7
GGA	Post	0.012 (0.015)	0.018 (0.017)	0.185	1.21E−7
GAG	Post	0.013 (0.016)	0.020 (0.019)	0.181	3.31E−7
TTT	Post	0.056 (0.055)	0.039 (0.042)	0.178	2.38E−6
TTT	Pre	0.070 (0.053)	0.052 (0.047)	0.178	1.99E−6
GTG	Exon	0.014 (0.016)	0.019 (0.015)	0.168	7.29E−7
AAG	Post	0.015 (0.014)	0.019 (0.014)	0.168	4.54E−6

The 10 most informative triples ranked by their F-value. For each triple, we specify its location relative to the exon (pre, exon, post) and its mean frequency among alternative and among constitutive exons, μ^+ and μ^-, respectively. The standard deviations of the latter quantities are listed in parentheses. For each feature we also list its F-value and the χ^2-P-value, which represents the probability that the distributions of the positive and negative class are sampled from a single distribution.

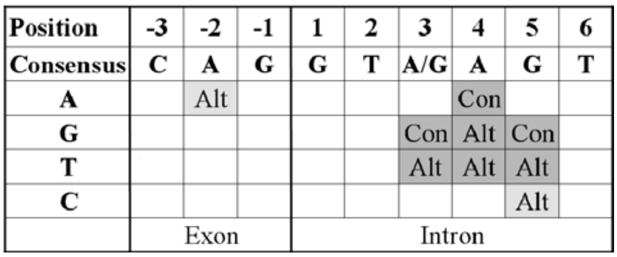
Figure 3 demonstrates the relative importance of the four parts comprising the feature vectors. It is evident that the set of seven features used by Sorek *et al.* (2004b) (Base) has a much higher discriminative power than the other sets. The F-values within this set fall between 0.287 and 0.834. It should be noted that although the average values of F of the remaining three sets of features, intensity of the PPT, triple counts (Triples) and position-dependent single base counts at the 5′ss are similar, the latter two sets contain many features that are much more informative than the single PPT feature.

In addition to the seven features reported by Sorek *et al.* (2004b), we discovered many features that convey useful information for the task of identifying alternative exons. Table 1 lists the 10 most informative features (all of them triples), together with their mean frequencies among alternative and constitutive exons, their F-score, and their significance level, as measured by χ^2-test. Interestingly, only one of these ten features is a tuple within the exon body, possibly indicating the significance of flanking intronic sequences in the regulation of alternative splicing. This tendency prevails also when inspecting

Table 2. Most informative single base features within the 5′ss region

Base	Position	$\mu^+(\sigma^+)$	$\mu^-(\sigma^-)$	F	P-value
A	4	0.56 (0.50)	0.71 (0.45)	0.163	8.41E−7
G	5	0.65 (0.48)	0.79 (0.41)	0.157	1.37E−6
T	4	0.21 (0.41)	0.12 (0.32)	0.129	3.73E−5
G	3	0.24 (0.42)	0.33 (0.47)	0.101	4.35E−3
T	5	0.13 (0.33)	0.07 (0.25)	0.098	1.45E−3
G	4	0.16 (0.37)	0.10 (0.30)	0.095	2.82E−3
T	3	0.05 (0.23)	0.02 (0.15)	0.078	8.38E−3
C	5	0.09 (0.28)	0.05 (0.21)	0.078	1.18E−2
A	−2	0.70 (0.46)	0.63 (0.48)	0.074	3.45E−2

Informative positions at the 5′ss, ranked by their F-value. For each feature we specify the base, its position relative to the actual splice site and its mean frequency among alternative and among constitutive exons, μ^+ and μ^-, respectively. The standard deviations of the latter quantities are listed in parentheses. For each feature we also list its F and χ^2 P-value.



Position	-3	-2	-1	1	2	3	4	5	6
Consensus	C	A	G	G	T	A/G	A	G	T
A		Alt					Con		
G						Con	Alt	Con	
T						Alt	Alt	Alt	
C								Alt	
	Exon			Intron					

Fig. 4. Different compositions of 5′ss in alternative and constitutive exons. Shown are positions −3 to +6 relative to the 5′ss. Positions −3 to −1 depict the end of the exon, and positions 1–6 are the beginning of the intron. Also shown is the consensus of the 5′ss. Each shaded frame indicates an informative nucleotide in the specific position, which is either over-represent in alternative exons (Alt) or constitutive exons (Con). Dark gray, alternative/constitutive difference is significance to $\alpha \leq 0.01$; light gray, $\alpha \leq 0.05$. For example, in position 4, A is over-represented in constitutive exons, while G and T are more pronounced in alternative ones.

a considerably larger number of top ranking triples, and is therefore a real characteristic of the data.

Importantly, the analysis has revealed biologically significant details. As seen in Table 1, 9 out of 10 informative triples were stretches of purines or pyrimidines in the upstream (‘pre’) or downstream (‘post’) introns. From these data, it is clear that there appears to be an under-representation of poly-purine stretches in the intronic sequence proximal to alternatively spliced exons, both upstream and downstream of the exon, and over-representation of poly-pyrimidine stretches in these same regions. Indeed, poly-purine stretches within exons are known to be composed of sequences that regulate splicing (both alternative and constitutive) (Cartegni *et al*., 2002). Therefore, it is possible that some of these discriminative features are parts of splicing regulatory motifs.

We were also able to identify informative features within the 5′ss sequence. Table 2 lists the most informative 5′ss features, ranked by their F-values. As shown in Figure 4, the most informative features lie in positions 3, 4 and 5 of the 5′ss. Such differences in the 5′ss composition of alternative versus constitutive exons were noted previously (Clark and Thanaraj, 2002).

To improve the results we also tried Recursive Feature Elimination (RFE), suggested by Guyon *et al*. (2002). In contrast to the ranking based on F, that considers each feature in isolation, RFE is capable of taking into account dependencies between features, and is therefore considered more sophisticated. However, no significant improvement has been observed in either AUC or the value of TP$_\alpha$. One possible explanation for this is the fact that each input vector x is actually a concatenation of several parts with significantly different distributions. This non-homogeneity introduces a bias which reduces the effectiveness of RFE.

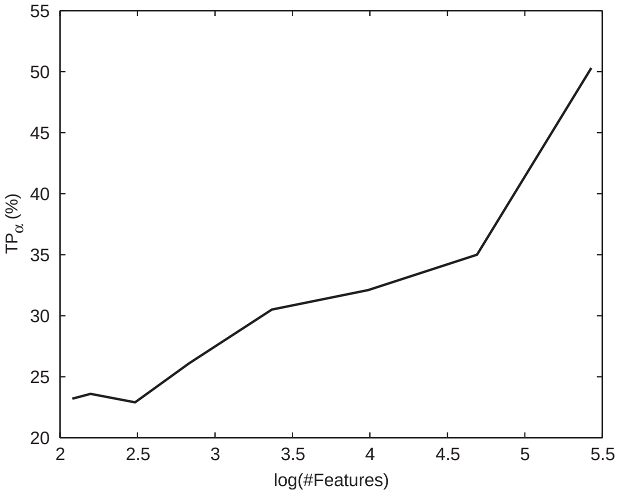


Fig. 5. The behavior of the true positive (TP) rate at a fixed false positive rate $\alpha = 0.5\%$ as a function of the logarithm of the number of features selected. The classifier uses a Gaussian kernel with $c = \sqrt{10}$ and $j = 0.5$.

2.7 Performance versus the number of features

To see the effect of the number of features, we used the optimal SVM hyper-parameters obtained by cross-validation and constructed eight classifiers. Each classifier was trained on a different feature subset, where the number of features was one of 8, 9, 12, 17, 29, 54, 109 and 228. The features were selected by their F-value. The performance (AUC, TP$_\alpha$) of each classifier was measured on the test set, to get an estimate of the performance of the SVM classifier as a function of the number of features selected. Figure 5 shows the dependency of TP$_\alpha$ on the number of features selected for $\alpha = 0.5\%$. Similar analysis of the AUC shows that it varies irregularly between 0.92 and 0.94, with no clear tendency, a behavior that probably originates from finite sample effects.

3 DISCUSSION AND CONCLUSION

Our aim in this paper was to build a classifier that robustly discriminates between constitutively and alternatively spliced conserved exons. To this end, we used a dataset comprising of constitutive and alternative exons in a 7:1 ratio, to train an SVM classifier.

Our feature selection procedure identified several new features whose alternative and constitutive distributions are significantly different. Those features might be involved in splicing regulation.

Using hyper-parameter selection and feature selection combined with cross-validation, a classifier with an AUC score of 0.93 was obtained. More importantly, this classifier is capable of rejecting constitutive exons very effectively at reasonable acceptance rates for true alternative exons. For example, with a false positive rate of 0.5% our classifier empirically achieved ∼50% true positive rate on an untouched test set.

It is important to note that our method is only capable of detecting exon-skipping in exons conserved between human and mouse genomes, because of its heavy reliance on conservation-based features. It is believed that a large proportion of functional alternative splicing is of the conserved type, but functional species-specific splice variants were also documented (Sorek *et al*., 2004a; Modrek and Lee, 2003). In our method, species-specific alternative splicing event will skip detection, as no conservation-based features can be calculated for them. Therefore, this set of exon-skipping events deserves specific solutions other than ours.

The results of this study are an improvement over our previous study, in which we used only seven features (five of them being conservation-based) to achieve a sensitivity of 30% at false positive rates similar to the ones in this study. The performance of the current study would enable effective scan of the exon database in search for novel alternatively spliced exons, in the human or other genomes.

ACKNOWLEDGEMENT

R.S. was supported by a fellowship from the Clore Israel Foundation.

REFERENCES

Agarwal,S., Graepel,T., Herbrich,R., Har-Peled,S. and Roth,D. (2004) Generalization bounds for the area under an ROC curve. *Technical Report UIUCDCS-R-2004-2433*, Department of Computer Science, UIUC, May 2004.

Brett,D., Hanke,J., Lehmann,G., Haase,S., Delbruck,S., Krueger,S., Reich,J. and Bork,P. (2000) EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. *FEBS Lett.*, **474**, 83–86.

Brown,M., Grundy,W., Lin,D., Christianini,N., Sugnet,C., Ares,M. and Haussler,D. (1999) Support vector machine classification of microarray gene expression data. *Technical Report UCSC-CRL 99-09*, University of California, Santa Cruz CA, June 1999.

Cartegni,L., Chew,S.L. and Krainer,A.R. (2002) Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat. Rev. Genet.*, **3**, 285–298.

Clark,F. and Thanaraj,T.A. (2002) Categorization and characterization of transcript-confirmed constitutively and alternatively spliced introns and exons from human. *Hum. Mol. Genet.*, **11**, 451–464.

Duan,K., Keerthi,S. and Poo,A. (2003) Evaluation of simple performance measures for tuning SVM hyperparameters. *Neurocomputing*, **51**, 41–59.

Florea,L., Hartzell,G., Zhang,Z., Rubin,G.M. and Miller,W. (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.*, **8**, 967–974.

Golub,T., Slomin,D., Tamayo,P., Huard,C., Gaasenbeek,M., Mesirov,J., Coller,H., Loh,M., Downing,J., Caligiuri,M., Bloomfield,C. and Lander,E. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Graveley,B.R. (2001) Alternative splicing: increasing diversity in the proteomic world. *Trends Genet.*, **17**, 100–107.

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

Guyon,I. and Elisseeff,A. (2003) An introduction to variable and feature selection. *J. Mach. Learning Res.*, **3**, 1157–1182.

Jaakkola,T., Diekhans,M. and Haussler,D. (2000) A discriminative framework for detecting remote protein homologies. *J. Comput. Biol.*, **7**, 95–114.

Joachims,T. (1999) Making large-scale SVM learning Practical. In *Advances Kernel Methods—Support Vector Learning*, MIT-Press, Chapter 11, pp. 169–184.

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*, **302**, 2141–2144.

Kan,Z., Rouchka,E.C., Gish,W.R. and States,D.J. (2001) Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. *Genome Res.*, **11**, 889–900.

Kan,Z., States,D. and Gish,W. (2002) Selecting for functional alternative splices in ESTs. *Genome Res.*, **12**, 1837–1845.

Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W., *et al*. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Leslie,C., Eskin,E., Cohen,A., Weston,J. and Noble,W. (2004) Mismatch string kernels for discriminative protein classification. *Bioinformatics*, **20**, 467–476.

Ling,C., Huang,J. and Zhang,H. (2003) AUC: a better measure than accuracy in comparing learning algorithms. In: *Proceedings of the 2003 Canadian Artificial Intelligence Conference*, pp. 329–341.

Maniatis,T. and Tasic,B. (2002) Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature*, **418**, 236–243.

Mironov,A.A., Fickett,J.W. and Gelfand,M.S. (1999) Frequent alternative splicing of human genes. *Genome Res.*, **9**, 1288–1293.

Modrek,B., Resch,A., Grasso,C. and Lee,C. (2001) Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res.*, **29**, 2850–2859.

Modrek,B. and Lee,C. (2002) A genomic view of alternative splicing. *Nat. Genet.*, **30**, 13–19.

Modrek,B. and Lee,C. (2003) Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nat. Genet.*, **34**, 177–180.

Smola,A., Bartlett,P., Scholkopf,B. and Schuurmans,D. (eds) (2000) *Advances in Large Margin Classifiers.* MIT Press, Cambridge, MA.

Schölkopf,B., Burges,C.J. and Smola,A. (eds) (1999) *Advances in Kernel Methods.* MIT Press, Cambridge, MA.

Sorek,R. and Ast,G. (2003) Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. *Genome Res.*, **13**, 1631–1637.

Sorek,R., Shamir,R. and Ast,G. (2004a) How prevalent is functional alternative splicing in the human genome? *Trends Genet.*, **20**, 68–71.

Sorek,R., Shemesh,R., Cohen,Y., Basechess,O., Ast,G. and Shamir,R. (2004b) Non-EST based method for exon-skipping prediction. *Genome Res.*, **14**, 1617–1623.

Sorek,R., Lev-Maor,G., Reznik,M., Dagan,T., Belinky,F., Graur,D. and Ast,G. (2004c) Minimal conditions for exonization of intronic sequences: 5′ splice site formation in alu exons. *Mol. Cell*, **14**, 221–231.

Sun,F., Fan,D. and Li,D. (2003) Identifying splicing sites in eukaryotic RNA: support vector machine approach. *Comput. Biol. Mach.*, **33**, 17–29.

Vapnik,V. (1998) *Statatistical Learning Theory*. Wiley, NY.

Zhang,X., Heller,K., Hefter,I., Leslie,C. and Chasin,L. (2003) Sequence information for the splicing of human pre-mRNA identified by support vector machine classification. *Genome Res.*, **13**, 2637–2650.

Yamamura,M. and Gotoh,O. (2003) Detection of the splicing sites with Kernel method approaches dealing with nucleotide doublets. *Genome Informatics*, **14**, 426–427.

Zhuang,Y. and Weiner,A.M. (1986) A compensatory base change in U1 snRNA suppresses a 5′ splice site mutation. *Cell*, **46**, 827–835.

Zien,A., Ratsch,G., Mika,S., Scholkopf,B., Lengauer,T. and Muller,K. (1999) Engineering support vector machine kernels that recognize translation initiation sites. *Bioinformatics*, **16**, 799–807.