Nucleic Acids Research, 2007, Vol. 35, Web Server issue ***W285–W291***
doi:10.1093/nar/gkm407

SplicePort—An interactive splice-site analysis tool

Rezarta Islamaj Dogan[1,2,*], Lise Getoor[1], W. John Wilbur[2] and Stephen M. Mount[3,4]

[1]Computer Science Department, University of Maryland, College Park, Maryland, 20742, [2]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland, 20894, [3]Department of Cell Biology and Molecular Genetics and [4]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, Maryland 20742, USA

Received January 31, 2007; Revised April 18, 2007; Accepted May 3, 2007

ABSTRACT

SplicePort is a web-based tool for splice-site analysis that allows the user to make splice-site predictions for submitted sequences. In addition, the user can also browse the rich catalog of features that underlies these predictions, and which we have found capable of providing high classification accuracy on human splice sites. Feature selection is optimized for human splice sites, but the selected features are likely to be predictive for other mammals as well. With our interactive feature browsing and visualization tool, the user can view and explore subsets of features used in splice-site prediction (either the features that account for the classification of a specific input sequence or the complete collection of features). Selected feature sets can be searched, ranked or displayed easily. The user can group features into clusters and frequency plot WebLogos can be generated for each cluster. The user can browse the identified clusters and their contributing elements, looking for new interesting signals, or can validate previously observed signals. The SplicePort web server can be accessed at http://www.cs.umd.edu/projects/SplicePort and http://www.spliceport.org.

INTRODUCTION

Accurate splice-site prediction is a critical component of eukaryotic gene prediction. Whole genome analysis of a single organism or comparison of genomes depends on accurate gene annotation. However, annotation is still limited by our ability to properly identify splice sites (1). We have developed a feature generation algorithm (FGA) for sequence classification (2). FGA automatically searches through a large space of sequence-based features to identify the predictive features. The identified features are used by a support vector machine classifier and produce accurate splice-site prediction on human pre-mRNA sequence data. In this work, we present a web-based interactive tool, SplicePort, which allows the user to explore the FGA features and allows the user to make splice-site predictions for submitted sequences based on these features.

Existing Internet resources, such as GeneSplicer (3), NetGene (4,5), MaxEntScan (6) and SplicePredictor (7), offer online splice-site prediction, providing the user with a list of predicted constituent splice sites for each input pre-mRNA (or genomic) sequence. However, a researcher may also be interested in identifying the signals used by the computational method to predict the splice site. Any element in the DNA sequence of a gene that helps to specify the accurate splicing of the pre-mRNA sequence is a splicing signal. Branch sites, pyrimidine tracts, exon splicing enhancers and silencers are all examples of known functional signals in the neighborhood of splice sites in eukaryotic genomes (see (8) for review). SplicePort, besides splice-site prediction, allows the user to explore all the FGA-generated features. We hope this will provide a useful resource for the identification of signals involved in specific splicing events, and possibly for the discovery of previously unappreciated splicing motifs.

THE FEATURE GENERATION ALGORITHM

In earlier work, we developed the FGA framework, which automatically identifies sequence-based features important for a sequence classification task (2). We applied this method to the task of splice-site prediction for the human genome (formally, the classificiation of AG dinucleotides into acceptors and non-acceptors and the classification of GT dinucleotides into donors and non-donors). FGA achieves very high accuracy compared to GeneSplicer (3), one of the leading programs in splice-site prediction. At the 95% sensitivity level, we were able to achieve improvements of 43.0% and 50.7% in the reduction of the false positive rate for acceptor splice sites and donor splice sites respectively (2), [Islamaj, R. *et al*., submitted].

Our data is a collection of 4000 pre-mRNA human RefSeq sequences. We refer to these sequences as the training sequences. For our experiments, we applied a

*To whom correspondence should be addressed. Tel: (301) 405 2717; Email: rezarta@cs.umd.edu

3-fold cross-validation scheme, and we tested our final splice-site model on the B2Hum data set supplied by the GeneSplicer team (3). This data set is a collection of 1115 pre-mRNA human sequences which do not overlap with our training sequences.

The core of the FGA method is a focused FGA that constructs complex features from simple sequence elements, such as single nucleotides and their position. Optimal features are selected after each generation step in order to keep the number of features manageable, and multiple rounds of feature construction and feature selection are applied in an iterative fashion. The feature types that we consider capture *compositional* and *positional* properties of sequences. A compositional feature is a string of k consecutive nucleotides (*k-mer*), where k ranges from 1 to 6. Compositional features include *upstream, downstream* and *general k-mers*. For each compositional feature, we count the number of times that feature is present in the neighborhood of the splice site. The length of the neighborhood region for the upstream or the downstream k-mer feature type is 80 nt, while that of the general k-mer is 160. The *position-specific k-mer* feature represents the substring appearing at positions $i, i+1, \ldots, i+k-1$ in the sequence. *Conjunctive positional features* are complex features constructed from conjunctions of position-specific 1-mer features. An n-positional feature consists of a conjunction of n nucleotides in n different positions co-occurring in the sequence. This type of feature is intended to capture the correlations between different nucleotides in non-consecutive positions in the sequence. For each positional feature we record the absence or presence of that feature in the neighborhood of the splice site.

For the human RefSeq training sequences, the FGA algorithm selected 3000 features for acceptor splice-site prediction and 1600 features for donor splice-site prediction. The acceptor site model contains 1362 compositional features and 1638 positional features, while the donor site model contains 764 compositional features and 836 positional features. We call these sets of features the acceptor model feature set and the donor model feature set.

The model feature sets then are used as input for the learning algorithm. The learning algorithm we use is C-modified least squares (CMLS), described by Zhang and Oles in (9). CMLS is a max-margin method similar to support vector machines. Relative to standard support vector machines, CMLS has a smoother penalty function which allows calculation of gradients that provide faster convergence (9).

For the splice-site prediction problem, two separate CMLS classifiers are required, one for acceptor and one for donor sites. After the training phase of these classifiers, each feature f_i in the model feature sets is assigned a weight w_i. These weights define the decision boundary of the linear classifier that optimizes the performance. We also use these weights to derive feature ranking.

When the classification model is given a new input sequence (the sequence is in the format [80 nt +AG/GT +80 nt]), initially it checks whether it is a candidate acceptor (AG) or a candidate donor (GT) splice-site sequence. Then, the classifier checks the sequence if it contains any of the features previously identified by the FGA algorithm in the corresponding model feature set. The classifier produces a final score for the input sequence adding the weights of each present feature. This score, assigned by SplicePort and displayed in the output, is best understood in terms of the splice-site classification problem itself.

In Figure 1, we use the B2hum data set supplied by the GeneSplicer team to show the sensitivity and specificity differences for different FGA score thresholds. We also provide a quantitative comparison between the two algorithms. Figure 1A depicts acceptor splice sites and Figure 1B depicts donor splice sites.

SPLICEPORT

This feature generation and classification model is the core of the SplicePort web server (http://www.cs.umd.edu/projects/SplicePort and http://www.spliceport.org). From the SplicePort initial page, the user has two options: splice-site prediction and motif exploration. The splice-site predictor receives the user's input sequence and reports all the predicted constituent splice sites. The motif explorer can be used to investigate acceptor and donor model feature sets identified in the input sequence or the sets of features FGA has discovered in the training sequences. The latter allows the user to browse the entire collection of positional features identifiable during the training phase. We believe our motif exploration is novel and useful. While we illustrate its use on the FGA selected features, we believe this interface is general and can be used to explore other feature types (10–12), and features selected by other learning algorithms (13,14). In Figure 2, we summarize the functionality of SplicePort and we describe its components in greater detail in the following sections.

SPLICE-SITE PREDICTION

Using the splice-site predictor is straightforward. The user inputs a sequence in FASTA format. The sequence can be cut and pasted directly into the window, or uploaded as a FASTA file. The server is case insensitive and accepts either DNA (T) or RNA (U) sequences as input. The length of the submitted sequence determines the time required for prediction (~1 s/kb of submitted sequence). The predictor uses a splice-site neighborhood of 80 nt upstream and 80 nt downstream for a constituent splice-site. After the user submits the input sequence file, the results of splice-site prediction are displayed in a tabular format. Figure 3A shows a sample output. The information listed for each prediction is: donor/acceptor splice site, the location in the sequence, a short subsequence centered at that location and the FGA score. The sensitivity value can be changed by entering a new score threshold. This value by default is 88.5% for donor sites and 88.8% for acceptor sites (corresponding to score = 0). After each change, the new sensitivity and false positive rate values are calculated and displayed to the user, as shown in Figure 3B. Finally, the user can select

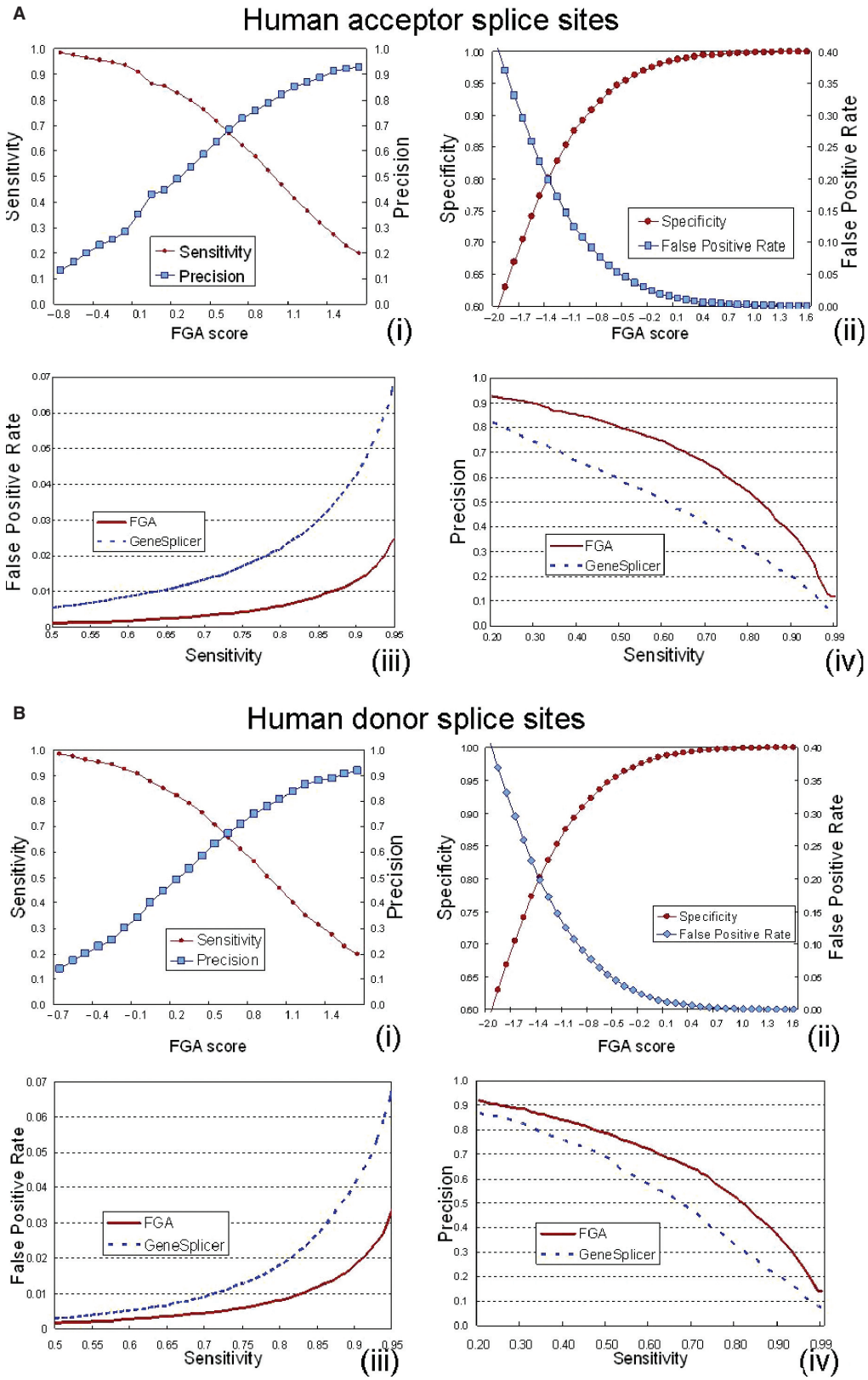

Figure 1. Sensitivity, specificity, false positive rate and precision vary with FGA score. (**A**) Acceptor sites. (**i**) Sensitivity, TP/(TP + FN), and Precision, TP/(TP + FP), vs FGA score. (**ii**) Specificity, TN/(TN + FP), and False Positive Rate, FP/(TN + FP), vs FGA score; (**iii**) FGA results are compared with those of GeneSplicer. False positive rate is shown as a function of sensitivity. These results show that FGA produces fewer false positives for every sensitivity threshold. (**iv**) Precision is shown as a function of sensitivity. These results show that FGA produces higher precision for every sensitivity threshold. These differences are highly statistically significant. (**B**) Donor sites (Graphs are as in A).

one of the predictions to investigate the identified signals, as described in the following section.

BROWSING FEATURES ON WHICH A SELECTED PREDICTION IS BASED

SplicePort allows the user to explore potential splicing signals in the vicinity (160 nt) of any particular splice site (AG or GT) by examining the features that contribute to the score assigned to that potential site. The signals of the acceptor model feature set or the donor model feature set can be listed, browsed and visualized by selecting the *Browse Features* option.

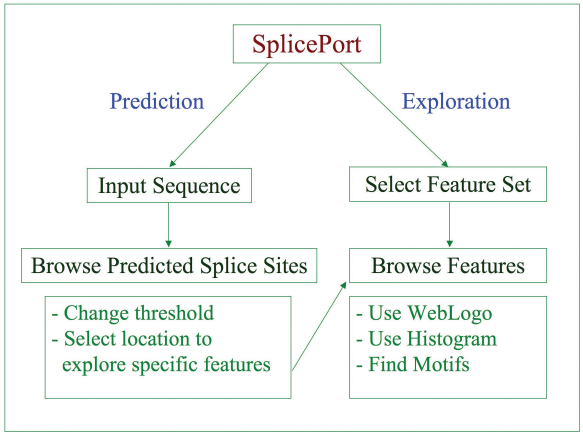

Figure 2. Organization of the SplicePort interactive interface. On the starting page a user chooses between splice-site prediction and motif exploration. After potential splice sites are predicted and scored, the features on which those predictions are based can then be explored.

Features are grouped into compositional features and positional features. Compositional features comprise general, upstream and downstream k-mers. They can all be listed, clustered and sorted by their weight. Positional features comprise position-specific nucleotides, position-specific k-mers and conjunctive positional features in the 160 nt neighborhood. There are a variety of browsing possibilities for this set of features. The user specifies an interval within the 160 nt window by giving the start and the end points. All the positional features that are associated with positions within this interval are listed. They are shown relative to the splice-site location, providing the user with a visual representation of the position of the feature and are ordered by the absolute value of their individual weights. SplicePort supports a rich catalog of visualization tools; the user may further group these features, draw histogram and WebLogo (15) frequency plots, search by motif and set weight threshold.

As an example shown in Figure 4, we used SplicePort to examine exon 7 of the homologous *SMN1* and *SMN2* genes, a well-studied case where a single nucleotide difference at position 6 of the exon accounts for reduced inclusion of this exon in *SMN2* (see (16) for review). SplicePort scores the *SMN1* exon 7 acceptor and donor 1.78 and 0.02, respectively and the single nucleotide change in *SMN2* reduces these numbers to 1.61 and

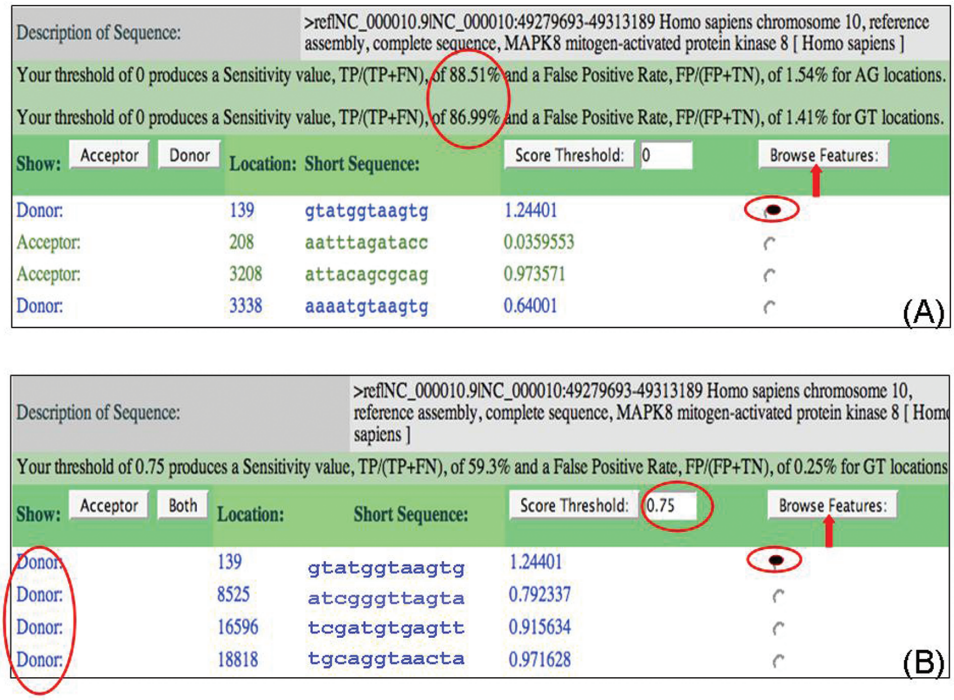

Figure 3. Typical output example of the predicted splice sites. (**A**) For each input sequence SplicePort displays the sensitivity value (circled in the figure). From this screen, the user can select a predicted site (we have selected the donor site at location 139 for illustration) and click on Browse Features, which we show with the arrow, to explore the present features. (**B**) This figure depicts the situation when the user prefers to explore acceptor or donor splice-site locations separately. The user can browse the features that are present in the checked sequence by clicking on Browse Features, which we show with the arrow. The user can change the score threshold, which we have circled on this screen, and list all the sites that score higher than the threshold. The sensitivity and false positive rate values are shown below the FASTA sequence description line.

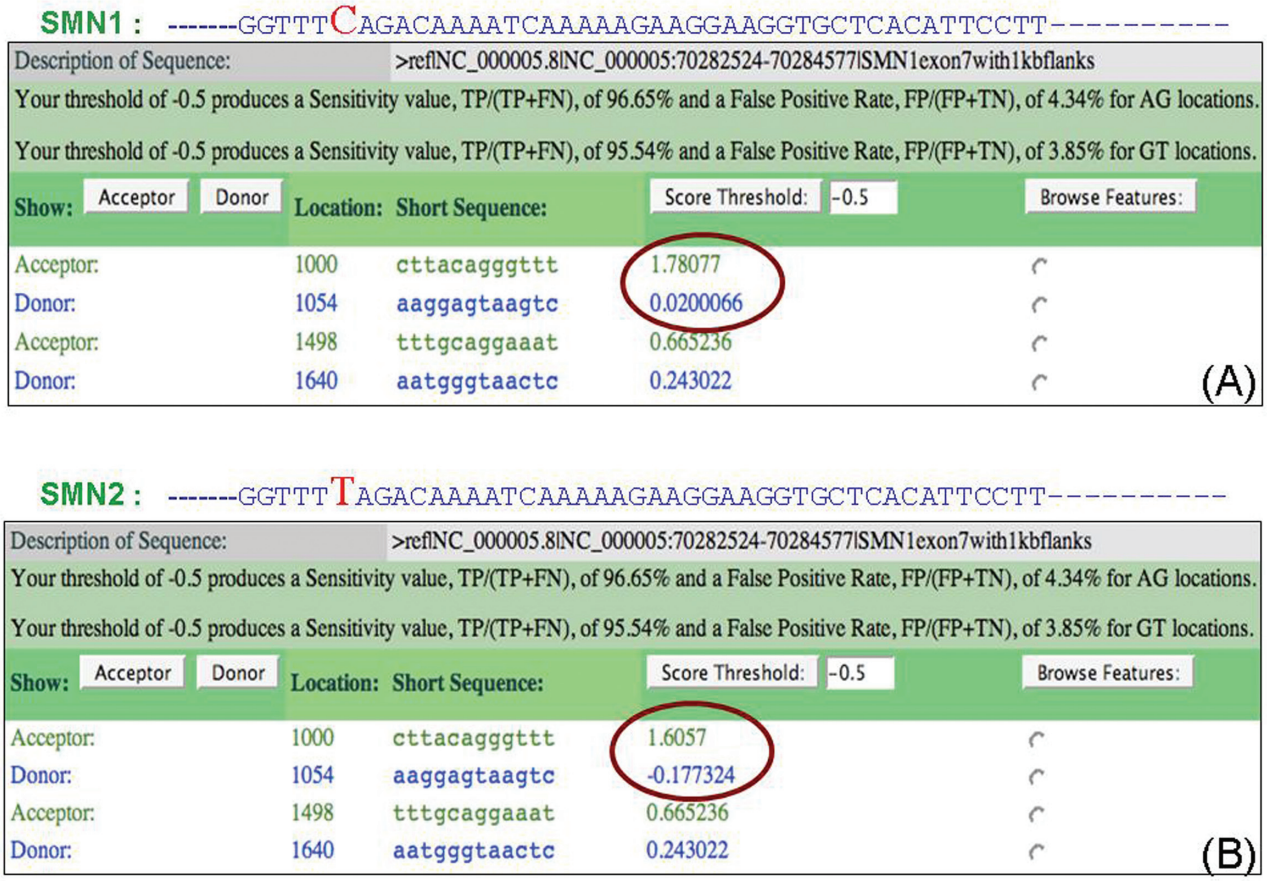

Figure 4. Splice-site prediction output of SplicePort for SMN1 (**A**) and SMN2 (**B**) exon 7 gene sequences with 1kb nucleotides flanking region. The acceptor site of exon 7 is at position 1,000 and the donor site is at position 1,054. We see that the single nucleotide difference at position 6 of the exon reduces the acceptor score from 1.78 to 1.61 and the donor score from 0.02 to −0.18.

−0.18. The feature browser shows that the difference in donor scores is primarily due to the negatively scoring upstream feature TAG (−0.18).

MOTIF EXPLORATION TOOL

Users can explore general features discovered by FGA for human RefSeq sequences using the Motif Exploration Tool. In order to facilitate motif discovery, the Motif Exploration Tool presents a much richer set of features than the sequence-specific feature browser (which presents only those features used to score the submitted sequence). We use a much richer set of features than existing splice-site tools, and focus on these rather than the simple compositional features. Each feature set we considered is the conjunction of a k-mer and a number of arbitrary position-specific nucleotides. We denote a specific set using the notation K-mer + X; for example, 4-mer+2 is the set of 4-mers together with two position-specific nucleotides.

Features of this type may be useful to discover non-adjacent correlations between the different nucleotides in different positions. Each of these sets contains 5000 top ranking features according to the Information Gain criterion.

Figure 5 illustrates a portion of the Motif Explorer. The figure on the top shows how the user selects a feature set and specifies an interval to browse the features. The figure on the bottom shows the results. The features are shown with respect to the splice-site location, and they are ordered according to the absolute value of their weight. The weight of a feature is learned by the classification algorithm during training. These weights can be used to order and group the features. A positively weighted feature is a feature mostly found in splice-site sequences, and a negatively weighted feature is a feature more commonly found in non-splice-site sequences. Figure 6 shows the results of *WebLogo* and *Histogram* functions. The user can view a depiction of the positively and negatively weighted features in the specified interval by generating a WebLogo frequency plot. The histogram allows the user to visualize the role of each nucleotide for each position in the specified interval. We represent this with four different bars, one for each nucleotide, for each position. The height of each bar is the accumulated weight for that position-specific nucleotide and is calculated using the weights of all the features that have that nucleotide at that position.

Because the features generated with the FGA algorithm are position-specific features, we may find the same pattern of nucleotides repeated in a given interval.

Interval Features refer to a set of features which share the same pattern of nucleotides but differ in starting positions. The user can list all the interval features for a specified interval and feature set. SplicePort displays the number of individual features as well as their average weight. To obtain the list of all individual features shown relative to a splice site in their respective locations, the user can use the *Search by Motif* option. This option also facilitates the search for known motifs or partial motifs. The user enters a short sequence and is returned a list of all features in the specified interval that contain that sequence.

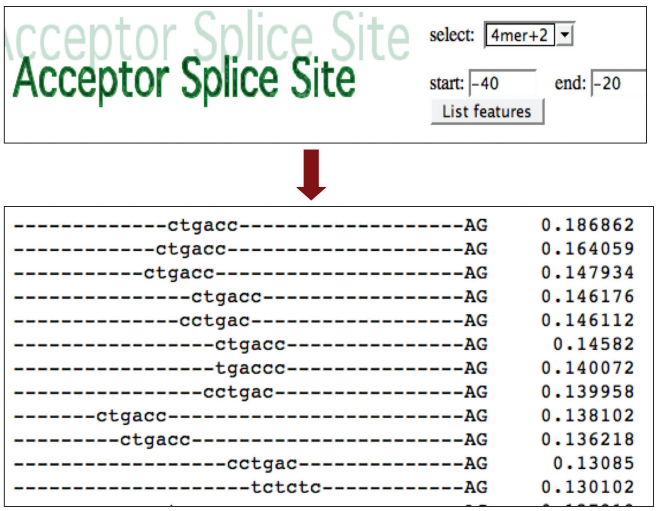

Figure 5. Motif Exploration Tool. This figure shows initially the selection of the feature set 4mer + 2 in the branch site interval. SplicePort outputs the list of features in the specified interval. Each feature is aligned to the splice-site position and has a weight assigned to it by the FGA algorithm. The acceptor splice site is depicted in the output with the capitalized dinucleotide AG.

In addition, for each feature set and specified interval we perform a clustering procedure based on edit distance. We identify similar features and the tool groups them together generating WebLogo frequency plots to represent them. The user can browse these identified clusters and their individual elements by selecting *Identified Motifs*. This option may help the user identify known functional motifs and may guide them in the search for new ones.

An illustrative example inspired by the case of *SMN1* and *SMN2* is a comparison of TAG and CAG among 5-mer features located in the −60 to −30 interval relative to donor sites. Features containing TAG are all negative, with multiple examples of TTTAG. Conversely, CAG shows primarily positive features. This example is shown in Figure 7.

SUMMARY

The SplicePort server is a versatile tool with two main functions. First, the user can perform accurate splice-site prediction on a sequence which they input to the tool, with the flexibility of exploring all the putative splice-site locations, their score, corresponding sensitivity and false positive rate values. Second, the user can explore the motifs for the requested location in the input sequence and browse the complete collection of identified motifs for both acceptor and donor splice sites. This tool can both help a user decide whether there is a splice site in the given sequence, and it can also allow the user to identify elements of functional motifs. An additional benefit of a computational exploration approach such as SplicePort is that it can be readily implemented in other genomes.

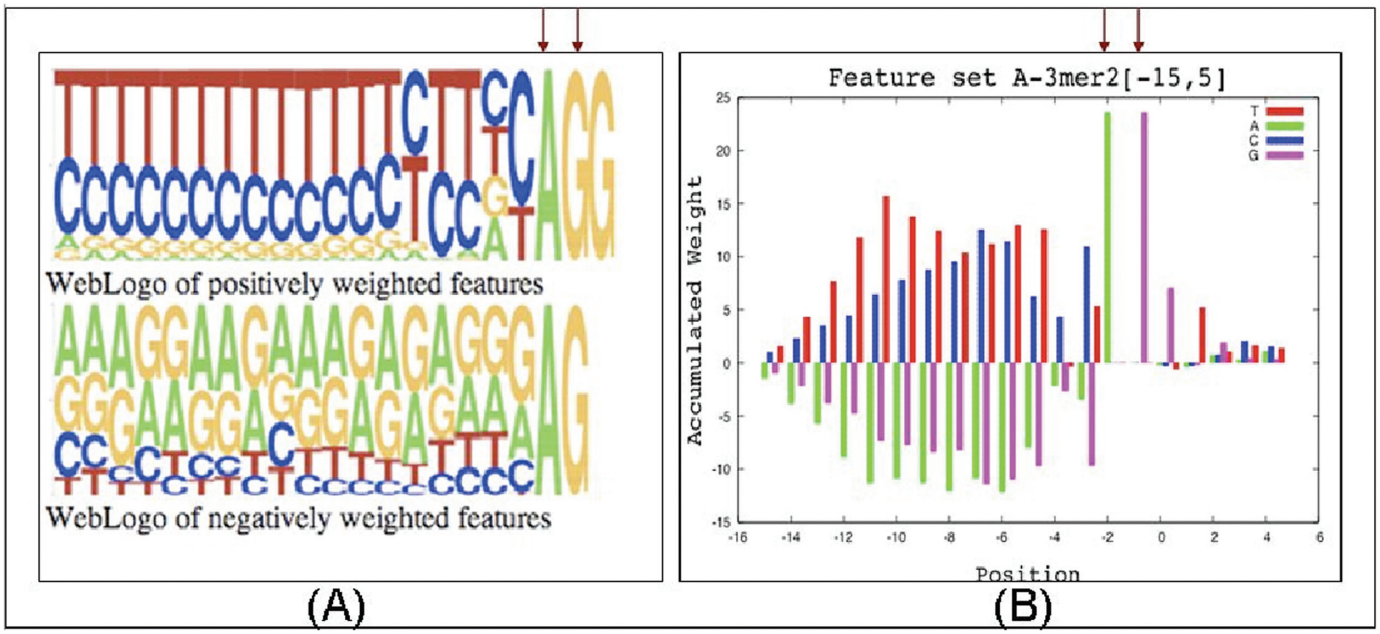

Figure 6. Typical outputs for motif exploration. These are features generated for acceptor splice-site prediction: (**A**) shows WebLogo frequency plots of features when we select the interval [−20,1], and (**B**) shows the histogram generated from accumulated weights of features when we select the interval [−15, 6]. The little arrows denote the location of acceptor splice-site consensus dinucleotide AG

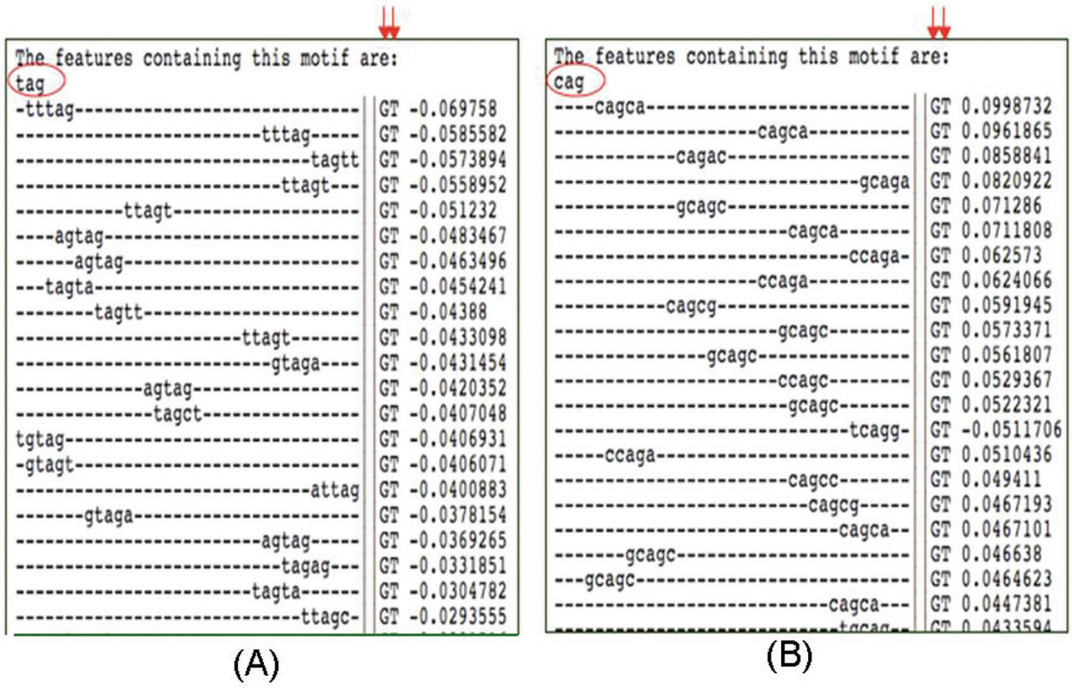

Figure 7. Outputs for 5mer feature set of donor splice-site prediction in the selected interval [−60,−30], using the SMN1 exon 7 example. In (**A**) we list features which contain the motif ''tag''. Note that all these features have a negative weight. In (**B**), we list features which contain the motif ''cag''. Note that these features are mostly positive. The little arrows denote the location of donor splice-site consensus dinucleotide GT.

In summary, SplicePort allows the user to discover useful insight in pre-mRNA splicing signals. This data analysis tool provides the community of researchers investigating pre-mRNA splicing with a powerful and flexible resource for the identification of functional elements. Motif exploration enables researchers to rapidly explore the space of computationally identified signals and effectively pose hypotheses for experimental test and validation.

ACKNOWLEDGEMENTS

This research was supported in part by an appointment to the National Center for Biotechnology Information (NCBI) Scientific Visitors Program sponsored by the National Library of Medicine and administered by the Oak Ridge Institute for Science and Education (RI). This work was supported in part by the National Science Foundation under grant number 0544309 (SMM). Funding to pay the Open Access publication charge was provided by NCBI.

Conflict of interest statement. None declared.

REFERENCES

1. Guigo,R., Flicek,P., Abril,J.F., Reymond,A., Lagarde,J., Denoeud,F., Antonarakis,S., Ashburner,M., Bajic,V.B. *et al.* (2006) EGASP: the human ENCODE Genome Annotation Assessment Project. *Genome Biol.*, **7**, S21–S31.
2. Islamaj,R., Getoor,L. and Wilbur,W.J. (2006) A feature generation algorithm for sequences with application to splice-site prediction. In *Proceedings of European Conference on Principles and Practice of Knowledge Discovery in Databases*, 553–560.
3. Pertea,M., Lin,X. and Salzberg,S. (2001) GeneSplicer: a new computational method for splice site prediction. *Nucleic Acids Res.*, **29**, 1185–1190.
4. Hebsgaard,S.M., Korning,P.G., Tolstrup,N., Engelbrecht,J., Rouze,P. and Brunak,S. (1996) Splice site prediction in *Arabidopsis thaliana* DNA by combining local and global sequence information. *Nucleic Acids Res.*, **24**, 3439–3452.
5. Brunak,S., Engelbrecht,J. and Knudsen,S. (1991) Prediction of human mRNA donor and acceptor sites from the DNA sequence. *J. Mol. Biol.*, **220**, 49–65.
6. Yeo,G. and Burge,C. (2004) Maximum entropy modelling of short sequence motifs with application to RNA splicing signals. *J. Comput. Biol.*, **11**, 377–394.
7. Brendel,V. and Kleffe,J. (1998) Prediction of locally optimal splice sites in plant pre-mRNA with applications to gene identification in *Arabidopsis thaliana* genomic DNA. *Nucleic Acids Res.*, **26**, 4748–4757.
8. Ladd,A. and Cooper,T. (2002) Finding signals that regulate alternative splicing in the post-genomic era. *Genome Res.*, **3**, reviews0008.1–0008.16.
9. Zhang,T. and Oles,F. (2001) Text categorization based on regularized linear classification methods. *Inform. Retriev.*, **4**, 5–31.
10. Fairbrother,W.G., Yeo,G.W., Yeh,R., Goldstein,P., Mawson,M., Sharp,P.A. and Burge,C.B. (2004) RESCUE-ESE identifies candidate exonic splicing enhancers in vertebrate exons. *Nucleic Acids Res.*, **32**, W187–W190.
11. Cartegni,L., Wang,J., Zhu,Z., Zhang,M.Q. and Krainer,A.R. (2003) ESEfinder: a web resource to identify exonic splicing enhancers. *Nucleic Acids Res.*, **31**, 3568–3571.
12. Wang,Z., Rolish,M.E., Yeo,G., Tung,V., Mawson,M. and Burge,C.B. (2004) Systematic identification and analysis of exonic splicing silencers. *Cell*, **119**, 831–845.
13. Goren,A., Ram,O., Amit,M., Keren,H., Lev-Maor,G., Vig,I., Pupko,T. and Ast,G. (2006) Comparative analysis identifies exonic splicing regulatory sequences-the complex definition of enhancers and silencers. *Mol. Cell*, **23**, 769–781.
14. Zhang,X.H., Heller,K.A., Hefter,I., Leslie,C.S. and Chasin,L.A. (2003) Sequence information for the splicing of human pre-mRNA identified by support vector machine classification. *Genome Res.*, **13**, 2637–2650.
15. Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.
16. Cartegni,L, Hastings,ML, Calarco,JA, de Stanchina,E. and Krainer,AR. (2006) Determinants of exon 7 splicing in the muscular atrophy genes SMN1 and SMN2. *Am. J. Hum. Genet.*, **78**, 63–77.