Biochemical and Biophysical Research Communications 431 (2013) 221–224

Contents lists available at SciVerse ScienceDirect

Biochemical and Biophysical Research Communications

journal homepage: www.elsevier.com/locate/ybbrc

A novel computational method for the identification of plant alternative splice sites

Ying Cui, Jiuqiang Han, Dexing Zhong *, Ruiling Liu

The School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, PR China

ARTICLE INFO

Article history:
Received 22 December 2012
Available online 9 January 2013

Keywords:
Alternative splicing sites
Plant
Position weight matrix
Increment of diversity
Support vector machine

ABSTRACT

Alternative splicing (AS) increases protein diversity by generating multiple transcript isoforms from a single gene in higher eukaryotes. Up to 48% of plant genes exhibit alternative splicing, which has proven to be involved in some important plant functions such as the stress response. A hybrid feature extraction approach which combing the position weight matrix (PWM) with the increment of diversity (ID) was proposed to represent the base conservative level (BCL) near splice sites and the similarity level of two datasets, respectively. Using the extracted features, the support vector machine (SVM) was applied to classify alternative and constitutive splice sites. By the proposed algorithm, 80.8% of donor sites and 85.4% of acceptor sites were correctly classified. It is anticipated that the novel computational method is promising for the identification of AS sites in plants.

1. Introduction

In higher eukaryotes, Alternative splicing (AS) creates multiple mRNA transcripts from a single gene by differential selection of splice sites, which is considered to be a major source of proteome diversity and plays a crucial role in the generation of biological complexity [1–3]. Abundant researches have been reported on AS in animals, while the study in plants is still at an early stage. Evidence from global analyses of AS in plants suggests that AS is frequently associated with specific tissue types or environmental conditions such as abiotic stress [5]. AS is involved in some important plant functions, including the stress response, which may impact domestication and trait selection [4]. Additionally, the latest research shows that AS phenomenon affects up to 48% intron-containing genes in the flowering plants [1]. This percentage is significantly higher than previous estimates [4]. However, the majority of plant AS sites has not been functionally characterized. Therefore, there is an urgent need to design an accurate and fast computational method to identify the alternative splicing sites in plants.

As the volume of expressed sequence tags (ESTs) increases, some large-scale studies have employed ESTs to detect AS [6–8]. Although ESTs are useful for studying AS, methods based on them are still far from being accurate due to many problems inherent in ESTs, such as low sequence quality, coverage limitations, RT–PCR bias artifacts and EST fragmentation [9].

Specifically designed microarrays have also been used for the identification of alternative splice variants [10–12]. In [12], *support vector machine (SVM) was employed as classifier to deal with the data acquired from tiling array platforms*. Those methods suffer from several limitations. For example, high economic cost makes it difficult to acquire large data sets in a microarray experiment, and also the accuracy of microarray analysis depends on the consistency of the hybridization data [13]. In addition, high noise levels and limited probe coverage are also problematic [12]. Thus, the quality of microarray-derived results is variable [13].

Several non-EST-based computational approaches have been proposed to predict AS in humans and animals, and fruitful results have been achieved. However, the development an refinement of computational tools to accurately predict AS are still needed. Compared with studies in animals, few studies are about the significance of functional AS in plants at the protein level. Also the complexity and the universality of AS may have been dramatically underestimated. The cis-elements that regulate AS in plants are especially poorly characterized [4]. As a result, it is an urgent problem to develop an accurate algorithm to identify the alternative splicing sites in plants.

In this paper, a computational method is proposed to identify the alternative splice sites in plants based only on genome sequences. Combing position weight matrix (PWM) with increment of diversity (ID), a new hybrid feature extraction approach named PHI (position weight matrix hybridized with increment of diversity) is proposed. The PWM is used to represent the base conservative level near splice sites, and the ID is used to quantitatively describe the similarity of two datasets. Then support vector machine, a nonlinear machine learning algorithm, was adopted to classify alternative and constitutive splice sites. In Fig. 1, we present an example of the classifying hyperplane of SVM in the PHI-axis 3D space. Based on PHI and SVM, 80.8% of donor sites and 85.4% of acceptor sites were correctly classified as alternative or constitutive, respectively. Herein, we verify the suggested

* Corresponding author. Fax: +86 029 82668665 181.
E-mail address: bell@xjtu.edu.cn (D. Zhong).

0006-291X/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.bbrc.2012.12.131

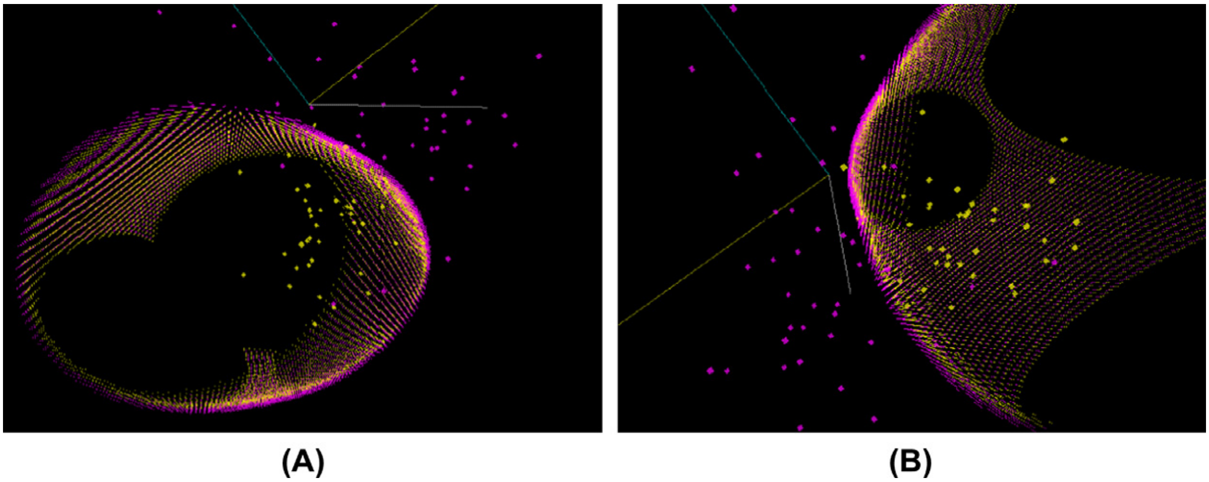

Fig. 1. The SVM classifying hyperplane of the 3′ training data. Pink points are the 3′ alternative sites and yellow ones are the 3′ constitutive sites. All the 80 points are randomly selected from the 3′ training data. In the 3D space, SF, ID1, and ID2 were represented by the yellow, blue, and white axis, respectively. (a) and (b) are images from two different perspectives. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

computational method and report on its identification accuracy using the model organism *Arabidopsis thaliana*.

2. Materials and methods

2.1. Datasets

The *A. thaliana* alternative splice sites were extracted from the AltExtron database at http://www.ebi.ac.uk/asd/index.html [14-16]. All of the sequences from −70 to +70 bp distant from the alternative splice sites obey the GT–AG rule. A total of 589 alternative 5′ splice sites and 589 alternative 3′ splice sites were collected. Constitutive splice sites were also downloaded from the AltExtron database, restricted to those sites whose flanking exons and introns do not show any AS possibilities. A total of 589 constitutive 5′ splice sites and 589 constitutive 3′ splice sites were randomly chosen. In the step of feature extraction, all splicing sites were randomly divided into four independent groups which comprise of the 5′ training, 5′ testing, 3′ training and 3′ testing sets.

2.2. Support vector machine (SVM)

The support vector machine (SVM) is a supervised machine learning algorithm based on the statistical learning theory [25]. SVM is usually used to solve classification and regression problems and has been successfully applied to bioinformatics investigations, such as the identification of plant promoters [17] and the identification of alternative splice sites in humans [18]. The basic thought of SVM is to map the original data into a high-dimensional feature space through a nonlinear mapping function and then to construct a hyperplane as the discriminative surface between the positive and negative data. In this study, the software LIBSVM [19] was employed to fulfill the task. The software was obtained from http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

2.3. Feature extraction

PWMs have been widely used to depict sequence patterns in computational molecular biology and regulatory genomics [22]. These sequence patterns can be formulated as maximum likelihood problems [20]. To describe the base conservative level near splice sites, 3-mer frequencies were selected as parameters for PWM. By aligning the flanking sequences of alternative donor (acceptor) sites in the training sets, the position probability matrix was defined as the following:

$$P_{ib} = f_{ib}/N, \tag{1}$$

where b designates any of the four nucleotides: A, C, G, and T. N is the total number of aligned sequences in the 5′/3′ training set. f_{ib} refers to the absolute frequency of observing nucleotide b in ith position of the N aligned sequences.

PWM was defined as:

$$W_{ib} = \ln(P_{ib}/P_{0b}). \tag{2}$$

In Eq. (2), P_{0b} is the random probability and is equal to 0.25^3, and W_{ib} is PWM value for nucleotide b in ith position. For a sequence with length n, the PWM scoring function (SF) is represented as:

$$SF = \sum_{i=1}^{n} W_{ib}, \tag{3}$$

which represents the strength of the splice site. The site is increasingly likely to be a genuine alternative splice site if it has a larger value of SF.

Based on the theory of the measure of diversity (MD), the ID is a measure of the total uncertainty and information in a system, by which the similarity level of two datasets can be quantitatively described.

In a d-dimensional discrete state space X, the standard diversity measure for diversity source $X:\{n_1, n_2, \ldots, n_d\}$ is defined as [21]:

$$D(X) = D(n_1, n_2, \ldots, n_d) = N\ln N - \sum_{i=1}^{d} n_i \ln n_i, \tag{4}$$

where n_i is the absolute frequency of the i^{th} state, $N = \sum_{i=1}^{d} n_i$.

For the diversity of two sources $X:\{n_1, n_2, \ldots, n_d\}$ and $Y:\{m_1,m_2,\ldots,m_d\}$, ID is defined as:

$$ID(X, Y) = D(X+Y) - D(X) - D(Y), \tag{5}$$

where $D(X+Y)$ is the measure of diversity of the mixed source $X+Y:\{n_1+m_1, n_2+m_2, \ldots, n_d+m_d\}$.The more similar are the two datasets, the smaller score is the ID.

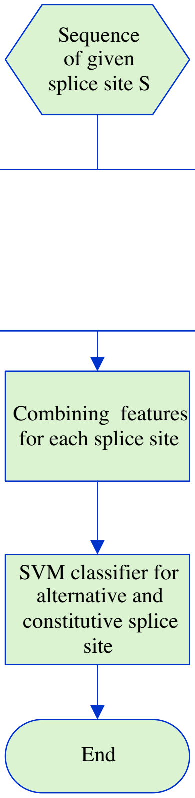
Sequence of given splice site S
SF of S
ID of S
Combining features for each splice site
SVM classifier for alternative and constitutive splice site
End

Fig. 2. Flow chart of the plant AS identification method.

2.4. Performance assessment

Four standard measures (sensitivity, specificity, total accuracy and the Matthew's correlation coefficient) were adopted to evaluate the predictive capability and reliability of our method. True positive (TP) and false negative (FN) are the numbers of positive data (alternative splice sites) that are predicted to be positive and negative, respectively. Analogously, true negative (TN) and false positive (FP) were used to denote the numbers of negative data (constitutive splice sites) that are predicted to be negative and positive, respectively.

The sensitivity (S_n), specificity (S_p), total accuracy (TA) and the Matthew's correlation coefficient (MCC) are defined as the following:

$$S_n = \frac{TP}{TP + FN}, \tag{6}$$

$$S_p = \frac{TN}{TN + FP}, \tag{7}$$

$$TA = \frac{TP + TN}{TP + TN + FN + FP}, \tag{8}$$

Table 2
Performance of the plant AS identification method on the testing set.

	S_n(%)	S_p(%)	TA(%)	MCC
Donor site	79.2	82.3	80.8	0.616
Acceptor site	83.3	87.4	85.5	0.708

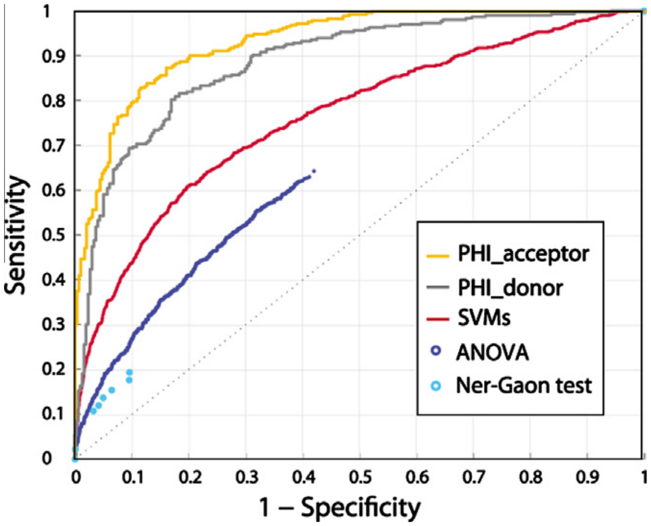

Fig. 3. The receiver operating characteristics (ROC) curves of PHI and other methods. The performance of the PHI method was compared to methods in [12,26] and ANOVA. The yellow curve and the gray curve are the prediction results of acceptor and donor AS sites by our method, respectively. Red, light blue and dark blue are the performance of the methods proposed in [12], [26] and ANOVA. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

$$MCC = \frac{TP \times TN - FN \times FP}{\sqrt{(TP + FN)(TP + FP)(TN + FN)(TN + FP)}}, \tag{9}$$

A flow chart of the proposed method is shown in Fig. 2.

3. Results

In this study, three parameters were extracted by PHI for the SVM classifier.

The 3-mer nucleotide frequencies at the −70 to +70 sites were selected as the PWM parameters for both the 5′ splice sites and 3′ splice sites. We used the 5′(3′) splice site PWM to calculate the SF score of each 5′(3′) splice site sequence in the training set and then set the SF score as the first characteristic parameter for SVM.

For each sequence, one side of the splice site is exon sequence, and the other side is the intron sequence. Intron sequences have different 3-mer compositions from exons [23,24]. Therefore, 3-mer absolute frequencies of −70 to −1 and +1 to +70 sequence fractions were selected as the diversity sources to describe

Table 1
The parameters of SVM.

Parameter	Source of information	SF or ID
P_1	3-mer frequencies at −70 to +70 sites	SF of S[a] (based on the PWM of alternative donor (acceptor) sites in 5′(3′) training set)
P_2	3-mer absolute frequencies in the region from −70 nt to −1 nt	ID between S and alternative donor (acceptor) sites in 5′(3′) training set
P_3	3-mer absolute frequencies in the region from +1 nt to +70 nt	ID between S and alternative donor (acceptor) sites in 5′(3′) training set

[a] S indicates the given splice site sequence.

differences in the sequence compositions. Consequently, two diversity sources were established based on the 5′ and 3′ training sets, and the value of the ID was calculated between the data in the training set and the corresponding discrete source. The IDs of authentic and false AS sites are the other two characteristic parameters for SVM. All calculated values were selected as the parameters for SVM as shown in Table 1. Finally, the three-dimensional vectors were used as inputs to train the SVM classifier. Then a classifying hyperplane was generated to identify the AS sites in testing set. In Fig. 1 and 40 alternative and 40 constitutive splice sites were randomly selected from the 3′ training set to demonstrate how SVM classify the sites based on the three parameters chosen by PHI.

The identification of alternative splice sites was performed on the testing set. For a given splice site, SF and IDs were calculated to establish the feature vector. They were then input to the SVM classifier to discriminate AS from constitutive splice sites. The result is shown in Table 2. The receiver operating characteristics (ROC) curves of PHI and the methods in [12,26] as well as ANOVA are shown in Fig. 3. Two widely used ANOVA-based methods are MIDAS and MADS [27]. Here we choose the first one which can be obtained at the following website:

http://www.affymetrix.com/support/technical/whitepapers/exon_alt_transcript_analy-sis_whitepaper.pdf.

4. Discussion

In this paper, we have presented a method for identifying alternative splice sites in plants using only genomic sequences. Based on the features we extracted as the PWM and IDs, we developed an SVM-based algorithm that can predict alternative splice sites in plant mRNA, and the program was tested with known splice sites. Using this model, we achieved 82.3% sensitivity and 79.2% specificity for donor sites and 87.4% sensitivity and 83.3% specificity for acceptor sites.

Xia et al. [18] used support vector machine to identify the human alternative splice sites and correctly identified approximately 70%+ of the AS sites. Our results reached a higher level of accuracy (although not within a direct comparison) even though plant AS signals are less conserved than those of humans.

The algorithm discussed in this work is highly accurate and easy to apply for other plants. It will be useful in genome annotation for predicting the boundaries between introns and exons, and researches using this tool can elucidate the AS mechanism in plants.

Acknowledgments

This work is supported by grants from National Natural Science Foundation of China (No. 61105021), Ph.D. Program Foundation of the Ministry of Education of China (No. 20110201110010), China Postdoctoral Science Foundation (No. 2011M501442) and the Fundamental Research Funds for the Central Universities.

References

[1] D.L. Black, Mechanisms of alternative pre-messenger RNA splicing, Annu. Rev. Biochem. 72 (2003) 291–336.

[2] A. Kalsotra, T.A. Cooper, Functional consequences of developmentally regulated alternative splicing, Nat. Rev. Genet. 12 (2011) 715–729.

[3] A.S.N. Reddy, M.F. Rogers, D.N. Richardson, M. Hamilton, A. Ben-Hur, Deciphering the plant splicing code experimental and computational approaches for predicting alternative splicing and splicing regulatory elements, Front. Plant Sci. 3 (2012) 18.

[4] W.B. Barbazuk, Y. Fu, K.M. McGinnis, Genome-wide analyses of alternative splicing in plants: opportunities and challenges, Genome Res. 18 (2008) 1381–1392.

[5] S.A. Filichkin, H.D. Priest, S.A. Givan, et al., Genome-wide mapping of alternative splicing in *Arabidopsis thaliana*, Genome Res. 20 (2010) 45–58.

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

[7] C. Grasso, B. Modrek, Y. Xing, C. Lee, Genome-wide detection of alternative splicing in expressed sequences using partial order multiple sequence alignment graphs, Pac. Symp. Biocomp. 9 (2004) 29–41.

[8] P. Bonizzoni, R. Rizzi, G. Pesole, ASPIC: a novel method to predict the exon–intron structure of a gene that is optimally compatible to a set of transcript sequences, BMC Bioinform. 6 (2005) 244.

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

[10] G.K. Hu, S.J. Madore, B. Moldover, T. Jatkoe, D. Balaban, J. Thomas, Y. Wang, Predicting splice variant from DNA chip expression data, Genome Res. 11 (2001) 1237–1245.

[11] J.M. Johnson, J. Castle, P. Garrett-Engele, Z. Kan, P.M. Loerch, C.D. Armour, R. Santos, E.E. Schadt, R. Stoughton, D.D. Shoemaker, Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays, Science 302 (2003) 2141–2144.

[12] J. Eichner, G. Zeller, S. Laubinger, G. Rätsch1, Support vector machines-based identification of alternative splicing in *Arabidopsis thaliana* from whole genome tiling arrays, BMC Bioinform. 12 (2011) 55.

[13] L.D. Burgoon, J.E. Eckel-Pssow, C. Gennings, D.R. Boverhof, J.W. Burt, C.J. Fong, T.R. Zacharewski, Protocols for the assurance of microarray data quality and process control, Nucleic Acids Res. 33 (2005) e172.

[14] G. Koscielny, T.V. Le, C. Gopalakrishnan, V. Kumanduri, et al., ASTD: the alternative splicing and transcript diversity database, Genomics 93 (2009) 213–220.

[15] S. Stamm, J. Riethoven, T.V. Le, C. Gopalakrishnan, V. Kumanduri, Y. Tang, N.L. Barbosa-Morais, T.A. Thanaraj, ASD: a bioinformatics resource on alternative splicing, Nucleic Acids Res. 34 (2006) D46–D55.

[16] T.A. Thanaraj, S. Stamm, F. Clark, J.J. Riethoven, T.V. Le, J. Muilu, ASD: the alternative splicing database, Nucleic Acids Res. 32 (2004) D64–D69.

[17] A.K.M. Azad, S. Shahid, N. Noman, H. Lee, Prediction of plant promoters based on hexamers and random triplet pair analysis, Algorithms Mol. Biol. 6 (2011) 6–19.

[18] H. Xia, J. Bi, Y. Li, Identification of alternative 5′/3′ splice sites based on the mechanism of splice site competition, Nucleic Acids Res. 34 (2006) 6305–6313.

[19] C.C. Chang, C.J. Lin, LIBSVM: a library for support vector machines, <http://www.csie.ntu.edu.tw/~cjlin/libsvm>, 2001.

[20] C. Zhang, M.L. Hastings, A.R. Krainer, M.Q. Zhang, Dual-specificity splice sites function alternatively as 5′ and 3′ splice sites, Proc. Natl. Acad. Sci. 104 (2007) 15028–15033.

[21] R.R. Laxton, The measure of diversity, J. Theor. Biol. 70 (1978) 51–67.

[22] W. Yang, Q. Li, One parameter to describe the mechanism of splice sites competition, Biochem. Biophys. Res. Commun. 368 (2008) 379–381.

[23] L. Zhang, L. Luo, Splice site prediction with quadratic discriminant analysis using diversity measure, Nucleic Acids Res. 31 (2003) 6214–6220.

[24] H. Zhang, X. Hu, Q. Li, The recognition of 27-Class protein folds: approached by increment of diversity based on multi-characteristic parameters, Prot. Pept. Lett. 16 (2009) 1112–1119.

[25] V. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.

[26] H. Ner-Gaon, R. Fluhr, Whole-genome microarray in *Arabidopsis* facilitates global analysis of retained introns, DNA Res 13 (3) (2006) 111–121.

[27] Y. Xing, P. Stoilov, K. Kapur, A. Han, H. Jiang, S. Shen, D.L. Black, W.H. Wong, MADS: a new and improved method for analysis of differential alternative splicing by exon-tiling microarrays. RNA 2008, rna.1070208, <http://rnajournal.cshlp.org/cgi/content/abstract/rna.1070208v1>.