



ISSN: 2321-9122

EJBB 2014; 1 (6): 36-39

Received: 24-06-2014

Accepted: 30-06-2014

Pokkuluri Kiran SreeDept. of CSE,
JNTU Hyderabad, India.**Inampudi Ramesh Babu**Dept. of CSE,
Acharya Nagarjuna University,
Guntur, Andhra Pradesh, India.**SSSN Usha Devi N**Dept. of CS E,
University College of Engineering,
JNTU Kakinada, Andhra Pradesh,
India.

Cellular Automata in Splice Site Prediction

Pokkuluri Kiran Sree, Inampudi Ramesh Babu, SSSN Usha Devi N

ABSTRACT

Splice site prediction is one of the important problems in Bioinformatics. Splicing is the way in which introns are removed from pre-mRNA transcript and exons are joined before translation. The position where the introns are spliced out is called as splice site. Identifying the splice junction plays vital role in understanding the genes. For an efficient study on eukaryotic genes the first step is to predict the splice site accurately. Accurate prediction of splice site will lead to accurate prediction of gene structure. There are three categories of splice site exists; they are acceptor site (AS), donor site (DS) and neither of these. The proposed classifier AIS-SSMACA has to take DNA sequence as input and predict the category (AS/DS/Neither).

Keywords: Salicylic acid, Chickpea, *Fusarium* spp., *Pythium* spp., germination, polyphenol.

1. Introduction

Donor site exists at the start of an intron i.e. 5' towards left. Introns in the donor site frequently start with GT (dinucleotides). Acceptor site exists at the end of an intron i.e. 3' towards right. Introns in the acceptor site frequently end with AG (dinucleotides). The intron/exon borders are called as acceptors (Scanning from left), exon/intron borders are called as donors (Scanning from right) as shown in fig 1.

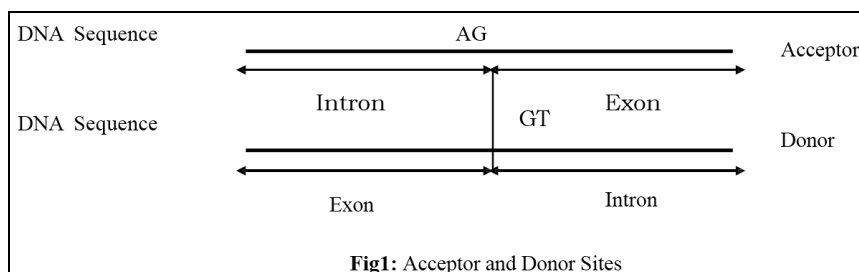


Fig1: Acceptor and Donor Sites

2. Literature Review

Many researchers have proposed various methods for predicting these splicing sites but the search for a good classifier with higher classifier accuracy is needed. We have reviewed the methodologies of the following well known splice site techniques, NNtree[53], Netgene2[152], HSPL[153], NNSplice[154], Splice View[155] and genesplicer[156].

3. Data Collection and Methods

The datasets are extracted from Irvine Primate splice junction database [157] (<http://archive.ics.uci.edu/ml/machine-learning-database>). The data set consist of 3190 DNA sequences each of length 60. Among 3190 sequences, 25% sequences belong to donor site category, 25% sequences belong to acceptor site category and 50% sequences belong to neither of these.

Among 767 donor sites, we have used 191 sequences for constructing AIS-SSMACA tree and 192 for checking the accuracy of the tree. The rest of 384 sequences are used for testing.

Among 768 acceptor sites, we have used 192 sequences for constructing AIS-SSMACA tree and 192 for checking the accuracy of the tree. The rest of 384 sequences are used for testing.

Among 1655 neutral sites (neither acceptor/donor), we have used 413 sequences for constructing AIS-SSMACA tree and 414 for checking the accuracy of the tree. The rest of 828 sequences are used for testing. The window length is fixed as 60.

Correspondence:

SSSN Usha Devi NDept. of CSE,
JNTU Hyderabad, India.

4. AIS-SSMACA

The main aim of the learning algorithm is to encode the DNA in the multiples of three and produce an AIS-SSMACA with n-attractors, k cells and m classless. Since the input is of fixed length that is 60 bp, the n value is fixed as 4, a k value is 3 and an m value is also three. At the end of the execution of the learning algorithm we will have set of basins which represent the classes.

Learning Algorithm

Input: DNA sequence

Output: AIS-SSMACA tree with n attractor basins.

Step 1: Read the input DNA sequence and process the sequence in the multiples of three. (Three neighborhood CA is used)

Step 2: Encode the input in the multiples of three

Step 3: Choose a high fitness rule and apply it on the input to construct an n- attractor, k-cell, 3-class AIS-SSMACA.

Step 4: Store all the basins constructed, repeat steps 1, 2, 3 till n-attractors are stored.

Step 5: Stop.

Testing Algorithm

The main aim of the testing algorithm is to distribute the corresponding input into the generated basins. During this process fitness, diversity of the intermediate node will be calculated for efficient development of the desired tree.

Once the DNA sequence identifies the basin uniquely, we can report the class associated with the basin.

Input: DNA sequence

Output: DNA Class (Acceptor/Donor/Neither)

Step 1: Read the input DNA sequence and process the sequence in the multiples of three.

Step 2: Encode the input in the multiples of three (As shown per discussion in 3.4)

Step 3: Distribute the input into the generated AIS-SSMACA basins till the entire sequence falls into a attractor of the tree as per algorithm 3.6.

Step 4: Report the basin and corresponding class.

Step 5: Stop.

5. Output & Experimental Results of AIS-SSMACA

This section shows the output of the proposed classifier. AIS-SSMACA will take input as a DNA sequence and reports the splice sites in both the stands of the sequence. Input 1 shown below is processed by AIS-SSMACA and identifies donor sites, one in the forward strand and one in the reverse strand. Input 2 is processed by AIS-SSMACA and identifies acceptor site in the forward strand. Input 3 is processed by AIS-SSMACA and identifies the sequence belong to neither donor nor acceptor.

Input 1: CCCAAGGCCAACCGCGAGAAGATGACCCAGGTGAGTGGCCCGCTACCTCTTCTGGTGGCC

Output: # Sequence Sequence_human_Kiran_Splice_123jntuh = 60 bps

Sequence_human_Kiran_Splice_123jntuh, Human Splice Prediction				
Donor Site Prediction				
START	END	SCORE	EXON	INTRON
24	38	0.99	GACCCAGGTGAGTGG	
Donor Site Prediction in Reverse Strand				
START	END	SCORE	EXON	INTRON
53	39	0.72	AGAAGAGGTAGCGGG	
Acceptor Site Prediction				
Nil				
Acceptor Site Prediction in Reverse Strand				
Nil				

Input 2: CTCCCTGATGCCCTCAGAATCTCCCCACAGGCCGCCTGATCTTTGACAACCTGAAGAAAT

Output: # Sequence Sequence_human_Kiran_Splice_83jntuh = 60 bps

Sequence_human_Kiran_Splice_83jntuh, Human Splice Prediction				
Donor Site Prediction				
Nil				
Donor Site Prediction in Reverse Strand				
Nil				
Acceptor Site Prediction				
START	END	SCORE	INTRON	EXON
10	50	0.95	GCCCTCAGAATCTCCCCACAGGCCGCCTG	ATCTTTGACAAC
Acceptor Site Prediction in Reverse Strand				
Nil				

Input 3: CCAGCAGGCTGAGGGCCAGAGCGGCCAGCCCTGGGAGCTGGCACTGGGTCGCTTTTGGGA
 Output: # Sequence Sequence_human_Kiran_Splice_89jntuh = 60 bps

Sequence_human_Kiran_Splice_89jntuh , Human Splice Prediction
Donor Site Prediction
Nil
Donor Site Prediction in Reverse Strand
Nil
Acceptor Site Prediction
Nil
Acceptor Site Prediction in Reverse Strand
Nil

6. Performance of AIS-SSMACA & Discussion

Extensive experiments are conducted to report the superiority of the AIS-SSMACA classifier when compared with the existing approaches like NNtree[53], Netgene2[152], HSPL[153], NNSplice[154], Splice

View[155] and genesplicer[156] is reported in section two. The analysis on the basic parameters of tree building like number of nodes, height of the tree and classification time is reported in table 1.

Table 1: Performance of AIS-SSMACA

Method	Sensitivity	Number of Nodes	Height of the tree	Classification time(ms)
AIS-SSMACA	0.9695	4	3	400
NN Tree	0.9348	5	3	515
C4.5	0.9012	12	12	668

Table 2: Comparison of AIS-SSMACA with other methods

Methods	SeA	SeD	SeN	SeA+ SeD+ SeN
AIS-SSMACA	0.9695	0.9512	0.9620	2.8827
NNtree[53],	0.9348	0.9256	0.9306	2.7910
Netgene2[152],	0.9312	0.8568	0.9263	2.7143
HSPL[153],	0.9494	0.9456	0.9503	2.8453
NNSplice[154],	0.9256	0.9587	0.9006	2.7849
genesplicer[156]	0.9396	0.9562	0.9784	2.8742
SpliceView[155]	0.9489	0.9491	0.9300	2.8280

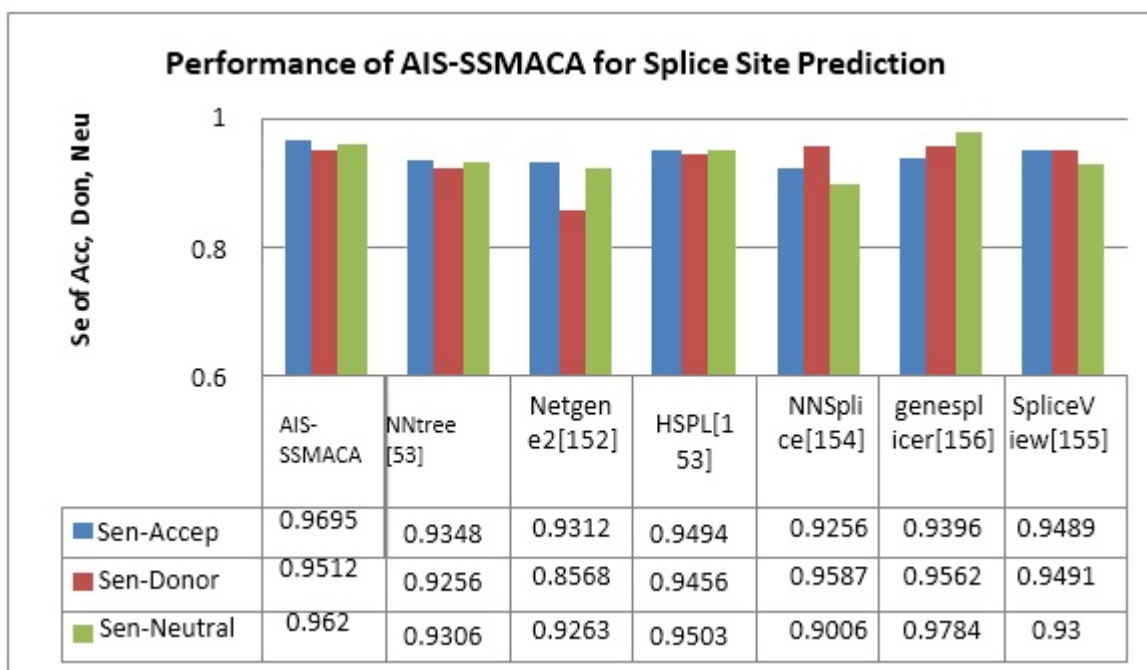


Fig 2: Comparison of AIS-SSMACA with other methods

The most important strength of AIS-SSMACA splice site prediction is predicting the acceptor and donor sites, even the acceptor input do not contain AG and the donor site do not contain GT. Among 796 trained DNA sequences to construct the desired AIS-SSMACA tree the average height of the tree constructed is 3. The number of nodes constructed to take a decision on the class of the DNA sequence is 3. The average time to report the class of the DNA sequence is 0.004 seconds as shown in table 1.

We have three categories of classes to be identified, SeA calculation relates to donor site prediction, SeB relates to acceptor site prediction and SeN relates to neutral prediction. The sensitivity for identifying acceptor class with AIS-SSMACA is high (0.9695) and least for NNSplice (0.9256) due to the increased error rate in NNSplice. The sensitivity for identifying donor is high for genesplicer and least for Netgene2. The sensitivity for identifying neutral prediction is high for AIS-SSMACA and low for NNSplice. In an ideal splice site prediction the value of SeA + SeB + SeN is 3. AIS-SSMACA maintains good balance among SeA, SeB, SeN which produces a value 2.8827, which is highest among the compared methods as shown in figure 2 and table 2. After AIS-SSMACA genesplicer shows good balance among SeA, SeB, SeN, which produces a value 2.8742.

7. Conclusion

We have successfully developed a classifier AIS-SSMACA for predicting splice sites with an accuracy of 96.06%, which is promising for human DNA of lengths 60 bp. It can predict the acceptor and donor sites, even the acceptor input do not contain AG and the donor site do not contain GT. The average numbers of nodes, height of the tree, classification time constructed to predict splice sites are 4, 3 and 400 ms respectively. In future we wish to extend this for

splice site prediction of various species with different lengths

8. Reference

1. Maji, Pradipta, Paul S. "Neural Network Tree for Identification of Splice Junction and Protein Coding Region in DNA." In Scalable Pattern Recognition Algorithms, Springer International Publishing, 2014, 45-66.
2. Pertea, Mihaela, Lin X, Steven LS. "GeneSplicer: a new computational method for splice site prediction." *Nucleic acids research* 2001; 29(5):1185-1190.
3. Softberry. <http://linux1.softberry.com/berry.phtml?topic=spl&group=help&subgroup=gfind>, 01 July, 2014.
4. Berkeley Drosophila Genome Project. http://www.fruitfly.org/seq_tools/splice.html, 01 July, 2014.
5. Institute for Biomedical Studies. http://bioinfo4.itb.cnr.it/~webgene/wwwspliceview_help.html, 01 July, 2014.
6. Pertea, Mihaela, Xiaoying Lin, and Steven L. Salzberg. "GeneSplicer: a new computational method for splice site prediction." *Nucleic acids research* 29, no. 5, 2001:pp.1185-1190.