Research Article

Cellular automata in splice site prediction

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 exist; 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: splicing junction, cellular automata, multiple attractor

Volume 1 Issue 2 - 2014

Pokkuluri Kiran Sree,[1] Inampudi Ramesh Babu,[2] SSSN Usha Devi N[3]

[1]Department of CSE, Jawaharlal Nehru Technological University, India
[2]Department of CSE, Acharya Nagarjuna University, India
[3]Department of CSE, University College of Engineering, India

Correspondence: Pokkuluri Kiran Sree, Department of CSE, Jawaharlal Nehru Technological University, Hyderabad, India, Tel +919493050794, Email profkiransree@gmail.com

Received: April 11, 2014 | **Published:** June 30, 2014

Abbreviations: AS, acceptor site; DS, donor site

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 form left), exon/intron borders are called as donors (Scanning from right) as shown in (Figure 1).

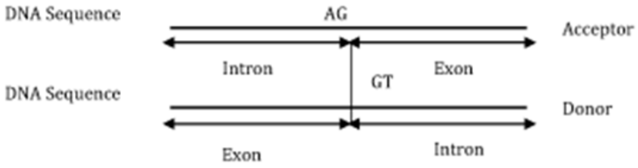


Figure 1 Acceptor and donor sites.

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,[1] Netgene2,[2] HSPL,[3] NNSplice,[4] SpliceView[5] and genesplicer.[2]

Data collection and methods

The datasets are extracted from Irvine Primate Splice junction database[6] (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.

i. 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.

ii. 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.

iii. 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.

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 60bp, 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 5.4)
Step 3: Distribute the input into the generated AIS-SSMACA basins till the entire sequence falls into a attractor of the tree.

Step 4: Report the basin and corresponding class.

Step 5: Stop.

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=60bps

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 CTCCCTGATGCCCTCAGAATCTCCCCACAGGCCGCCTGATCTTTGACAACTTGAAGAAAT

Output:

```
# Sequence Sequence_human_Kiran_Splice_83jntuh=60bps

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 GCCCTCAGAATCTCCCCACAGGCCGCCTGATCTTTGACAAC
Acceptor Site Prediction in Reverse Strand
Nil
```

Input 3 CCAGCAGGCTGAGGGCCAGAGCGGCCAGCCCTGGGAGCTGGCACTGGGTCGCTTTTGGGA

Output:

```
# Sequence Sequence_human_Kiran_Splice_89jntuh=60bps

 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
```

Citation: Sree PK, Babu IR, SSSN Usha Devi N. Cellular automata in splice site prediction. *MOJ Proteomics Bioinform.* 2014;1(2):49–52.
DOI: 10.15406/mojpb.2014.01.00013

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,[1] Netgene2,[2] HSPL,[3] NNSplice,[4] SpliceView[5] and genesplicer[2] 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

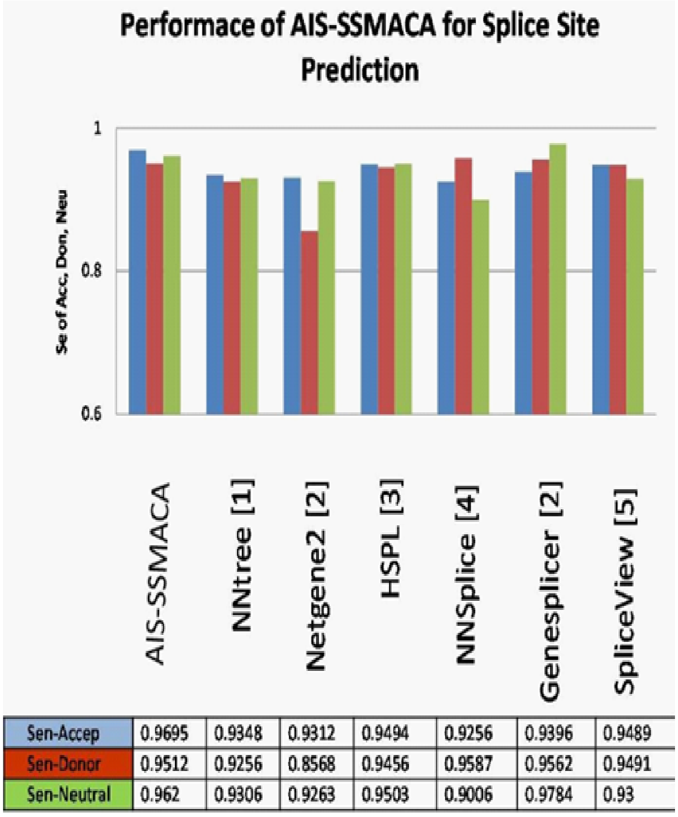
Method	Sensitivity	Number of nodes	Height of the tree	Classification time(ms)
AIS-SSMACA	0.9695	4	3	400
NN Tree	0.9348	5	3	515
C4.5	0.9012	12	12	668

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, Se_A 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 $Se_A+Se_B+Se_N$ is 3. AIS-SSMACA maintains good balance among Se_A, Se_B, Se_N 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 Se_A, Se_B, Se_N, which produces a value 2.8742.

Table 2 Comparison of AIS-SSMACA with other methods

Methods	Se_A	Se_D	Se_N	$Se_A+Se_D+Se_N$
AIS-SSMACA	0.9695	0.9512	0.9620	2.8827
NNtree[1]	0.9348	0.9256	0.9306	2.7910
Netgene2[2]	0.9312	0.8568	0.9263	2.7143
HSPL[3]	0.9494	0.9456	0.9503	2.8453
NNSplice[4]	0.9256	0.9587	0.9006	2.7849
Genesplicer[2]	0.9396	0.9562	0.9784	2.8742
SpliceView[5]	0.9489	0.9491	0.9300	2.8280



Sen-Accep	0.9695	0.9348	0.9312	0.9494	0.9256	0.9396	0.9489
Sen-Donor	0.9512	0.9256	0.8568	0.9456	0.9587	0.9562	0.9491
Sen-Neutral	0.962	0.9306	0.9263	0.9503	0.9006	0.9784	0.93

Figure 2 Comparison of AIS-SSMACA with other methods.

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 60bp. 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 sits are 4, 3 and 400ms respectively. In future we wish to extend this for splice site prediction of various species with different lengths.

Acknowledgements

None.

Conflict of interest

The author declares no conflict of interest.

References

1. Maji P, Sushmita P. Neural network tree for identification of splice junction and protein coding region in DNA. *In: Scalable pattern recognition algorithms*. Switzerland: Springer International Publishing; 2014. p. 45–66.

2. Pertea M, Lin X, Salzberg SL. GeneSplicer: a new computational method for splice site prediction. *Nucleic Acids Res*. 2001;29(5):1185–1190.

3. http://linux1.softberry.com/berry.phtml?topic=spl&group=help&subgroup=gfind

4. http://www.fruitfly.org/seq_tools/splice.html

5. http://bioinfo4.itb.cnr.it/~webgene/wwwspliceview_help.html

6. http://archive.ics.uci.edu/ml/machine-learning-database

Citation: Sree PK, Babu IR, SSSN Usha Devi N. Cellular automata in splice site prediction. *MOJ Proteomics Bioinform.* 2014;1(2):49–52.
DOI: 10.15406/mojpb.2014.01.00013