Splice Site Detection in DNA Sequences using Probabilistic Neural Network

Tripti Nassa
Department of CSE
PEC University of Technology
Chandigarh, India

Shailendra Singh
Department of CSE
PEC University of Technology
Chandigarh, India

Neelam Goel
Department of CSE
PEC University of Technology
Chandigarh, India

ABSTRACT

With numerous of genomes sequenced, gene prediction has become a challenging problem in bioinformatics. Gene prediction helps in identifying physical and mental features of different organisms. A large number of gene prediction tools have been developed in the past two decades. Splice site detection method lies at the heart of ab-initio gene prediction tools and plays an important role in detecting the exon boundaries. In this paper, a method for detecting splice sites by using generalized regression neural network is proposed. The proposed method uses conditional probabilities to preprocess the input which enables it to incorporate the already known sequence features from biological knowledge. The experimental results show that the application of this new architecture to splice site detection has greatly improved the training time and reduces the false positive predictions.

Keywords

Gene prediction, acceptor sites, donor sites, neural networks, conditional probability, consensus sequence

1. INTRODUCTION

With the advancement in technology, DNA sequencing has become easier now than it was in past. The amount of sequencing data increased exponentially in last few years. A large number of new eukaryotic genomes have been sequenced. The enormous data generated from sequencing efforts demands accurate computational tools to extort meaningful information from these data. During the past few decades, Gene prediction in eukaryotes has attracted the attention of many researchers around the globe. Despite many efforts the problem is still not solved satisfactorily. The current gene prediction methods exhibit very low accuracy at transcript level. Eukaryotic genes are organized as alternative segments of exons and introns. Exon forms the coding part which is further converted into protein via the process of translation. Exon/intron boundaries are known as splice sites. The transition from exon to intron is known as donor splice site and the transition from intron to exon is known as acceptor splice site. The splice sites are located in introns; donor splice site is characterized by canonical nucleotide pair GT whereas acceptor splice site is characterized by nucleotide pair AG. Several methods have been developed to detect splice sites in eukaryotes. Splice site detection method includes accurate detection of both donor and acceptor sites. However, the detection of splice sites is often a hard task due to the presence of consensus di-nucleotides at sites other than true splice sites. The accuracy of ab-initio gene prediction methods relies completely on the accuracy of splice site detection methods. Therefore, a splice site detection method is essential to find the location of genes.

There are many methods for splice site prediction, such as hidden Markov model [1], combinatorial methods [2], support vector machine [3], genetic algorithm [4], grammar based algorithms [5] artificial neural network [6-11] and neural network hybrid methods [12-17]. Neural networks have been widely used in splice site detection methods because of their ability to learn and solve many real time problems. Neural network can automatically adjust its internal structure to generate approximate results for the given problem and to find relationship among input and output. [18-19] briefly review some neural network based systems for splice site detection methods.

The method proposed in [6] combines local and global sequence features into a neural network. In a different method pair-wise correlation of di-nucleotides at splice site consensus is used as input to the neural network [7]. The main problem with these methods is their high false positive rate. Moreover, these methods are unable to harness the full potential of biological features available from the sequence. This is because neural network works as a black box. Therefore some methods need to be employed with neural network to extract the specific biological features. This problem is addressed in [12-14]. Here a combined approach based on markov model and neural network is investigated. The method proposed in this paper is based on this hybrid approach. A splice site detection method on similar line is developed by combining inhomogeneous markov chains and neural networks [15]. Though these methods have improved the prediction accuracy but the false predictions are still below the expected level.Thus, there is a need to design a method that will detect the splice sites accurately and reduces the false positive rate.

In this paper, a splice site detection method is proposed that improves the training time and reduces false positive predictions. The proposed method is based on the concept of conditional probabilities of nucleotides which forms the input of the neural network. The rest of the paper is organized as follows. In section II, the proposed method is discussed in detail. Subsequent section describes the dataset used in this study. In addition to it, the experimental results are shown. Finally, the conclusion is provided at the end of the paper.

2. PROPOSED METHOD

Existing neural network based splice site detection methods use orthogonal encoding in which sequence of nucleotides is represented as sequence of bits which is unable to represent features of the nucleotides surrounding the splice sites. An

efficient approach for splice site detection has been developed that combine markov model and neural networks.

2.1 Method Description

In the proposed method, neural network is combined with Markov model to detect the splice sites. As neural network does not take into account the complex relationship among nucleotides in the sequence. Therefore, there is a need to encode the input sequence in order to find the complex relationship among the nucleotides. The method works in two stages. At first stage low order markov chains are used to encode the input sequence which is given as input to the neural network in the second stage.

The splice site detection model composed of three consecutive DNA segments: upstream segment (US), signal segment (SS), and downstream segment (DS) as shown in figure 1. The signal segment consists of nucleotides immediately neighboring the splice site that represent the consensus patterns responsible of splicing mechanism. The upstream and downstream segments adjoining the signal segment on both sides of the signal segment capture the features of the coding and non-coding sequences, respectively, that always surround the splice sites. If the length of the signal segment of sequence at splice site is l_S and the lengths of the upstream and downstream segments are l_U and l_D, appropriately, then the splice site model is represented by a sequence of length $l_U + l_S + l_D$ [12-14].

For acceptor sites the length of input sequence is 190, where length of upstream and downstream is 80 base pairs and length of signal segment is 30, 20 nucleotides are taken from intron and 10 nucleotides are taken from exon. For donor sites the length of input sequence is 176, where length of upstream and downstream is 80 base pairs and length of signal segment is 16, 10 nucleotides are taken from intron and 6 nucleotides are taken from exon. The process starts with the pre-processing phase which includes taking input set of sequences, generating position specific conditional probability (CP) matrices for one order and two order Markov chains.

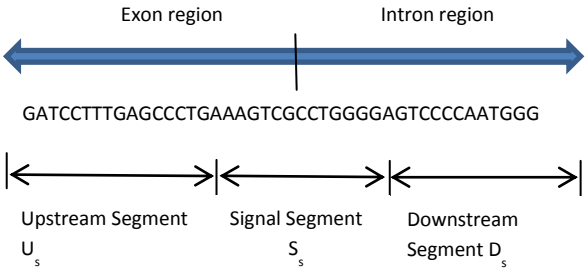

Figure 1. Representation of Splice Site Model

In order to apply the markov model to encode the input sequence, position specific CP matrices are needed. To calculate the CP matrices array of input sequences are used. The first order and second order CP matrices are created using the following formula:

First order:

$$P(s_k|s_{k-1}, \dots, s_0) = P(s_k|s_{k-1})$$

Second order:

$$P(s_k|s_{k-1}, \dots, s_0) = P(s_k|s_{k-1}, s_{k-2})$$

Where, k represents the order of the markov model. Three CP matrices are created for both donor and acceptor sites. These matrices are: first order signal, second order upstream, and second order downstream.

The input sequence is encoded using markov model. The signal segment is encoded using one-order, upstream and downstream segments are encoded by second-order Markov models to capture codon distributions. Each nucleotide in the segment represents a state of Markov model. If the signal segment, upstream segment, and downstream segment models are denoted by MS, MU, and MD, respectively, the emission probabilities are given by:

$$e_i^U(s_i) = P(s_i|s_{i-1}s_{i-2}, M^{U)},$$

$$e_i^S(s_i) = P(s_i|s_{i-1}, M^{S)},$$

$$e_i^D(s_i) = P(s_i|s_{i-1}s_{i-2}, M^{D)}$$

The input sequence is represented as an array of length 190, each element representing emission probability of each nucleotide. After encoding the sequence probabilities are used as input to the neural network. In contrast to the previous approach a probabilistic neural network is employed here for training and testing.

2.2 Neural Network

The neural network architecture used in this study is generalized regression neural network (GRNN). In the previous method a multilayer feed forward network trained using back-propagation is used. One major disadvantage of multilayer feed forward network is that the network architecture is determined by the user. Moreover, the training algorithm can take a large number of iterations to converge to a desired solution. These limitations can be overcome by GRNN. This neural network like other probabilistic neural networks needs only a fraction of the training sample than a back-propagation neural network would need. The GRNN is used to reduce the training time of splice site detection method.

GRNN usually consists of four layers. It has two special layers namely: pattern layer and summation layer. The number of neurons in these two layers is adjusted by the network itself. The other two layers are input and output. The network architectures for donor and acceptor sites are illustrated in figure 2 and figure 3 respectively.

The length of input sequence for acceptor site is 190 bp and for donor site is 176. The number of input nodes in input layer of GRNN is 190. So, 14 zeros are appended at the end of input sequence for donor sites. The input sequence is first encoded using Markov chains. The encoded sequences are then given to the GRNN, which after processing produces two outputs. First output represents score for acceptor site and second represents the score for acceptor site.

The output of neural network is filtered using a threshold value. The sequences with score greater than threshold are selected. The indices of acceptor and donor sites for the selected sequence are given as output.

3. RESULTS AND DISCUSSIONS

3.1 Dataset Description

All the data used to simulate the proposed method is extracted from Genbank release 111.0 (http://www.ncbi.nih.gov/Genbank/). The dataset contains 84 human gene sequences. The total set of 84 sequences contains 455 true donor and 455 true acceptor sites. The dataset is divided into two parts. The first part containing 61 sequences with 305 true

donor and acceptor sites forms the training set. The second part containing 23 sequences having 150 true donor and acceptor site forms the testing set. Additionally, 1000 false donor and acceptor sites with confirmed GT or AG nucleotide present, other than true sites are collected.

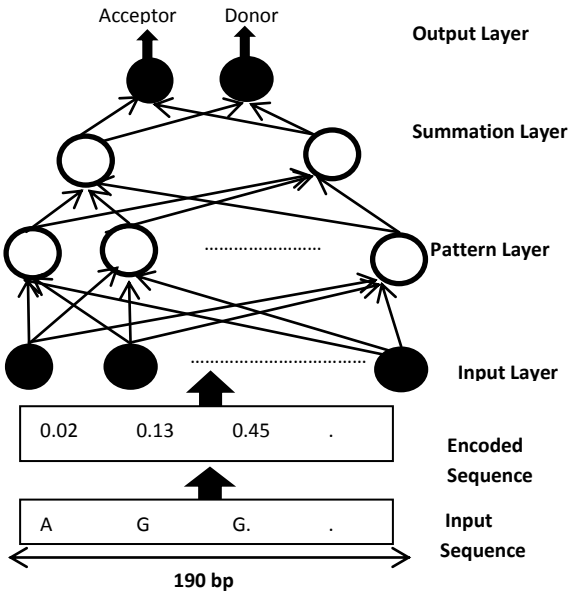

Figure 2. Network architecture for acceptor sites

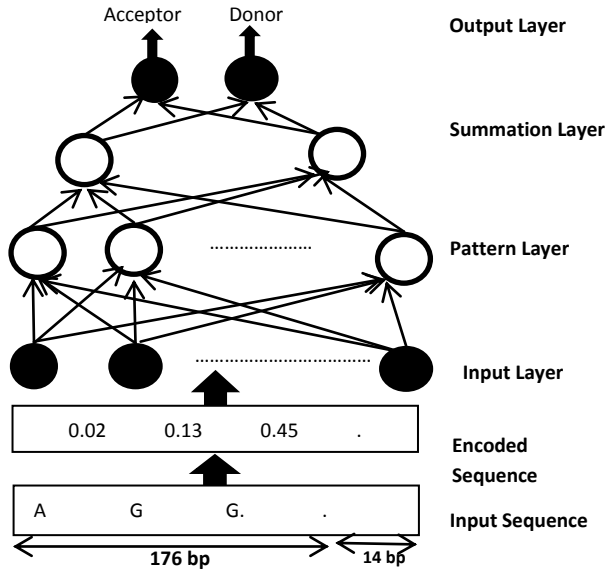

Figure 3. Network architecture for donor sites

3.1 Evaluation Measures

In case of DNA sequences, the false splice sites largely outnumber the true splice site. Due to this imbalance nature, the percentage of correctly predicted splice sites is a bad indicator of the predictive performance. Because if a method assigns high percentage of true splice sites, it will make a lot of incorrect false splice site assignments. Here sensitivity (Sn correctly detected as true splice sites) and specificity (Sp correctly detected as false splice sites) are used to quantify the performance of splice site detection. The sensitivity (Sn) and specificity (Sp) is defined as:

$$Sn = (TP/TP+FN)$$

$$Sp = (TN/TN+FP)$$

Where,TP: The number of correctly predicted splice sites, FN: The number of incorrectly predicted splice sites, TN: The number of correctly predicted false splice sites, and FP: The number of incorrectly predicted false splice sites.

3.2 Results

A two-fold cross validation experiment is used to assess the performance of the proposed method. The training set of 61 sequences is used to train the system. The time needed to train GRNN is very less as compared to multilayer feed-forward network. There is no need to determine the optimized network architecture by applying heuristics because the network adjusts its structure according to the training samples. After training the method is tested on the remaining 23 sequences. The performance of the method is tested on different cut-off levels. The best results obtained from this experiment are shown in table 1.

Table 1. Results of Splice Site detection Method

Splice Sites	TP	Sn (%)	FP	Sp (%)	Cut-off
Donor	128	85	473	94	0.35
Acceptor	116	77	542	95	0.48

The true site in the testing set is 150. Out of 150 true splice sites, 128 are predicted in case of donor site and 116 are predicted in case of acceptor sites. The sensitivity is better in case of donor sites while the specificity is better in case of acceptor. Though the prediction accuracy of the system is not very good but it greatly reduces the training time. Moreover, it simplifies the neural network architecture.

4. CONCLUSION

In this paper a neural network based splice site detection method is proposed. The method uses markov model to preprocess the input sequence. The conditional probabilities obtained at this stage are used as input to the neural network. The proposed method takes into account the features of nucleotides surrounding the splice sites. GRNN is used to train and test the network. The application of this new architecture to splice site detection shows that GRNN is better than multi layer feed forward network. The training time has greatly improved. This method reduces the architectural complexity by using a single neural network for both donor and acceptor sites. The prediction accuracy of this method is still not satisfactory. To further improve the prediction accuracy this method can be combined with knowledge from mRNA sequences.

5. REFERENCES

[1] Zhang, Q. 2009. Splice sites detection by combining Markov and hidden Markov model, In Proceedings of the 2nd International Conference on Biomedical Engineering and Informatics.

[2] Churbanov, A. and Ali, H. 2005. Combinatorial method of splice site prediction. In Proceedings of the IEEE Conference on computational systems bioinformatics conference.

[3] Sonnenburg, S., Schweikert, G., Philips, P., Behr, J., and Ratsch,G. 2007. Accurate splice site prediction using support vector machines. BMC Bioinformatics 8 (Suppl10):S7.

[4] Awadalla, S., Ortiz, J. E., Gopal, S. 2005. Prediction of trans-splicing sites using a genetic algorithm.

[5] Kashiwabara, A.Y., Vieira, D. C. G., Machado-Lima, A., and Durham, A. M. 2007. Splice site prediction using stochastic regular grammars. Genet. Mol. Res. 6(Mar. 2007), 105–115.

[6] Hebsgaard, S. M., Korning, P. G., Tolstrup, N., Engelbrecht, J., Rouzé1, P., and Brunak, S. 1996. Splice site prediction in arabidopsis thaliana pre-mRNA by combining local and global sequence information. Nucleic Acids Research 24(17), 3439–3452.

[7] Hatzigleorgious, A., Mache, N., and Reczko, M. 1996. Functional site prediction on the DNA sequence by artificial neural networks. In Proceedings of the IEEE International Joint Symposia on Intelligence and Systems.

[8] Tolstrup, N., Rouze, P., and Brunak, S. 1997. A branch point consensus from arabidopsis found by non-circular analysis allows for better prediction of acceptor sites. Nucleic Acids Res. 25(1997), 3159-3163.

[9] Reese, M. G., Eeckman, F. H., Kulp, D., Haussler, D. 1997. Improved splice site detection in Genie. In Proceedings of the First Annual International Conference on Computational Molecular Biology.

[10] Cai, T. and Peng, Q. 2005. Predicting splice sites in DNA sequences using neural network based on complementary encoding method. In Proceedings of the International Conference on Neural Networks and Brain.

[11] Johansen, O., Ryen, T., Eftesol, T., Kjosmoen, T., Rnoff, P. 2008. Splice site prediction using artificial neural networks. In Proceedings of the CIBB.

[12] Ho, L. S. and Rajapakse, J. C. 2002. Splice site detection with neural networks/Markov models hybrids. In Proceedings of the 9th International Conference on Neural Information Processing.

[13] Ho, L. S. and Rajapakse, J. C. 2003. Splice site detection with a higher-order markov model implemented on a neural network. Genome Info. 14(2003), 64-72.

[14] Rajapakse, J. C. and Ho, L. S. 200. Markov encoding for detecting signals in genomic sequences. IEEE/ACM Transactions on Computational Biology and Bioinformatics 2(2005), 131-141.

[15] Liu, L., Hu, Y. K., and Yau, S. 2007. Prediction of primate splice site using inhomogeneous markov chain and neural network. DNA Cell Biol. 26(2007), 477-483.

[16] Al-Daoud, E. 2009. Identifying DNA splice sites using patterns statistical properties and fuzzy neural networks. EXCLI Journal 8(2009), 195-202.

[17] Moghimi, F., Shalmani, M. T. M., Sedigh, A. K., and Kia, M. 2012. Two new methods for DNA splice site prediction based on neuro-fuzzy network and clustering. Neural Computing & Applications.

[18] Bajic, V. B., Tang, S., Han, H., Brusic, V., and Hatzigeorgious, A. G. 2002. Artificial neural networks based systems for recognition of genomic signals and regions: a review. Informatica 26 (2002), 389–400.

[19] Nassa, T. and Singh S. 2013. Neural network based systems for splice site detection: a review. IJARCSSE 3(June 2013), 604-608.