Volume 3, Issue 7, July 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Neural Network Based Systems for Splice Site Detection: A Review

Tripti Nassa, Shailendra Singh
Department of Computer Science and Engineering
PEC University of Technology, Chandigarh, India

Abstract— Gene prediction has become an important part of bioinformatics as more and more genome sequences are analysed. The performance of a gene prediction program depends on accurate splice site detection. Detecting splice sites with high accuracy is of prime concern. Many techniques have been used for splice site detection. This paper reviews the techniques that are based on neural network. Various methods that use neural network or its hybrid systems have been discussed along with their performance.

Keywords— Gene prediction, artificial neural network, donor splice site, acceptor splice site, Markov chain, fuzzy logic

I. Introduction

Bioinformatics deals with the application of computational methods to the analysis of biological data. Gene prediction typically refers to the area of bioinformatics that is concerned with algorithmic identification of protein coding genes in a given DNA sequence. As experimental approaches are capital intensive and time consuming, computational approaches for automatic annotation of DNA sequences have gained interest in bioinformatics. Eukaryotic gene contains coding part known as exons and non- coding part known as introns and any two adjacent exons are separated by an intron. Gene in a DNA sequence is converted into mRNA to proteins through the process called transcription and translation. In transcription, exons are connected to each other after splicing and then they are used for translation. The sites at which splicing is done are known as splice sites, the transition from exon to intron is donor splice site and the transition from intron to exon is acceptor splice site. The splice sites are located in introns. The donor site is characterized by canonical nucleotide pair GT and acceptor site is characterized by AG nucleotide pair. The introns are bounded by GT and AG splice sites. The effectiveness of gene prediction program depends on accurate detection of these splice sites. There are many methods for splice site prediction, such as hidden Markov model [1], combinatorial methods [2], support vector machine [3], genetic algorithm [4], positional correlations [5], grammar based algorithms [6] and artificial neural network. Artificial neural network has been usually used in splice site prediction because of its ability to capture and represent complex input and output relationships among data. An artificial neural network is an information processing paradigm that is inspired by the way as brain process information, is composed of large interconnected processing element working in unison to solve a specific problem. In this paper we are briefly reviewing splice site detection methods based on artificial neural network.

II. Splice Site Detection

In eukaryotic gene, splice sites are the boundaries between exon and intron. Identification of protein coding gene in genomic DNA requires accurate splice site detection. Splice site detection programs has two classification problems, discriminating from true and decoy sites for both donor and acceptor splice sites. Algorithms for splice site detection can be classified as (i) Sequence alignment with template sequence that contain known splice sites (ii) Statistical analysis using hidden Markov model (iii) Pattern recognition and pattern matching. In the following section we are reviewing splice site prediction techniques based on artificial neural network.

A. Neural Network Based Techniques

Neural network has been widely used in splice site prediction techniques because of its capability 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. Another advantage of using neural network in molecular biology is that they are fault tolerant as they are able to endure data which is incomplete or contains missing values. Various techniques for splice site detection based on neural network are discussed below:

1) NetGene: One of the first attempts made in splice site detection using artificial neural network was NetGene (1991) [7]. It is a joint prediction scheme where prediction of transition regions between introns and exons regulates a cut off level for splice site assignment. This uses feed-forward neural network that receives input from the windows scanning the DNA sequence. Each window configuration is represented numerically as binary string. The network is trained with error back propagation algorithm. The input sequence of nucleotides is sparsely encoded: A as (1000), C as (0100), G as (0010), T as (0001) to avoid algebraic dependencies between nucleotides in the encoding. The neural network classifies the middle nucleotide in one of the two categories: splicing donor site or non-splicing donor site, or, splicing acceptor

site or non-splicing acceptor site, or, coding or non-coding. In donor splice site detection, the size of input window is 15 i.e. there are 15 nucleotides visible in the input layer, 40 non-linear hidden units and one output unit. The optimal network architecture for predicting acceptor sites embedded in input sequence has a window size of 41 nucleotides and 20 hidden units. The output is compared with a threshold value to classify the input and the cut-off level depends upon the output of coding-non coding network. The number of false positive assigned is lowered when the cut-off assignment level is controlled by the exon signal, by a factor between two and thirty depending on the required level of detection of true sites. This method do not utilise the constraint of a continued open reading frame in the selection of compatible splice sites which may further reduce the number of false positive predictions.

2) NetPlantGene: To further improve the performance of NetGene (1996), the results of neural network are combined with rule based system: NetPlantGene [8]. This system is used to predict intron splice sites in the dicot plant Arabidopsis thaliana. This is a two-step prediction scheme, where a global prediction of the coding potential regulates a cut-off level for a local prediction of splice sites, is refined by rules based on splice site confidence values, prediction scores, coding context and distances between potential splice sites. The first step is prediction step which is equivalent to earlier work in NetGene and the second step is refinement step which is based on rules found by investigating the mistakes of the first step. In the refinement step a number of post-processing steps are designed in order to (i) discard wrong splice site predictions, (ii) choose between two or more nearby equally strong predictions, and (iii) to enhance weak or missing predictions which must be preferred when viewing the prediction non-locally. A number of rules to remove false positive are used that included use of protein coding potential, predicted exon and intron length, scanning acceptor site pairs in T-tract prolongation in 5' exon ends, strength of neighbouring splice sites, exon and intron length distributions and the average GC content. This method is able to detect 57.3% of the true donor sites and 21.1% of the true acceptor sites without any false predictions. The correlation coefficient for donor site is 0.90 and for acceptor site was 0.83.

3) NetGene2: Further improvements in acceptor site prediction were made in Netgene2 [9]. In this approach false positive rate is reduced by finding branch point information. This method trains a hidden Markov model on intron sequence to determine the position of branch nucleotide. The acceptor site prediction is enhanced by adding a new neural network that receives four inputs: the acceptor site score of the NetPlantGene acceptor site network, the log-odds score of the corresponding branch point sequence found by hidden Markov model, the distance between predicted acceptor site and the branch point, the derivative of the coding predicting network from NetPlantGene. Using branch point prediction the correlation coefficient for acceptor site is enhanced to 0.89 that was near to donor site prediction.

4) CLA Based System: A new splice site prediction method (1996) using a different neural network learning algorithm [10] is used in a gene identification program. The splice site recognition is trained using the 'Cascade Learning Algorithm' (CLA), which is able to add neurons to a network to optimize the architecture and the connection strength in a network. For donor site recognition a window of 10 nucleotides is used with 3 nucleotides from the exon and 7 from the intron and a window of 18 nucleotides is used for acceptor site with 3 nucleotides in exon and 15 in the intron. This asymmetric extraction of the windows is used because the most conserved part of the splicing signal lie more into the intron region. The donor classification is correct for 100% of all donor sites and 0.17% of all nucleotides are classified as donor but are not, for acceptor site classification is correct for 83.33% of all acceptor sites and 0.18% of all nucleotides are classified as false positive acceptor sites.

5) NNSplice: Another neural network based splice site prediction method NNSplice is used in gene identification program Genie (1997) [11]. In the first version of genie, a feed forward neural network is implemented that received input from a window of size 10 and 40 for donor site and acceptor sites respectively, where the sequence is encoded using 4 input units for each nucleotide. The networks are trained only on positive and negative examples that have consensus splice site i.e. 'GT' for the donor and 'AG' for the acceptor site. Based on the results of Henderson, to model the pairwise correlations between the adjacent nucleotides, the input representation is changed from 4 bit code per base to a 16-bit code per nucleotide pair. A window of 15 nucleotides is represented as 14 pairs of adjacent nucleotides and each pair is represented by 16 inputs that are all set to 0 except the one in the position representing the letter pair. This network achieves a rate of 98.67% true positives, with only 1.33 false negatives.

6) HNN Simulator: A hierarchical neural network system [12] with back propagation learning algorithm was developed in 1997. In this three types of neural network are defined using the system where each network is trained by the arrangements of bases around the splice sites of DNA sequences. All the three networks are trained to learn both donor and acceptor sites. Results of three networks are compared which showed type 3 network had better ability in predicting splice sites.

7) CEM Based System: All the methods described above uses the orthogonal encoding technique in which each nucleotide is represented by a 4 bit binary string with three bits set to 0 and one bit set to 1. This encoding method makes the computation very complex and neglects the fact that in DNA the nucleotides A and T, G and C are complementary. So a new complementary encoding method (CEM) [13] was introduced where the nucleotide A is represented by 1, T by -1, C by 2 and G by -2. The neural network uses in this method was back propagation with 3 layers and the size of input window and the number of hidden units were changeable. If the middle nucleotide of the input window is a donor site output is 1, if it is an acceptor site then the output is -1, otherwise output is 0. The false positive rate is 92.41% and 87.34% for donor and acceptor site respectively. This complementary encoding method makes much less incorrect splice site detection and reduced the training time of the neural network.

8) Back-propagation and Curve Fitting: In 2009, a splice site detection system was developed using back propagation neural network [14] where the output of neural network is used to find the exact location of the splice sites with the help of a curve fitting of a parabolic function. In this the splice site location is predicted without prior knowledge of 'GT' or

'AG' consensus signals for splice sites. The network uses a sliding window of nucleotides over a gene and four bit code is used for each nucleotide. The neural network produces a score if it finds a splice site in the window. There were 240 input units, 128 hidden layer units and 2 output units. A single network is used for both donor site and acceptor site detection. The network output is 1 when a splice site is in the middle of the sliding window and the farther splice site from the mid-point of the window got the lowest score. A nucleotide get score contribution from 60 outputs corresponding to the sliding window passing over it and all these outputs are accumulated and normalize to find the exact location of the splice site. The output of network is used for curve plotting and the nucleotide closest to the curve maxima is considered as the splice site. This system gives a sensitivity of 0.891, a specificity of 0.816 and a correlation coefficient of 0.552.

B. Neural Network Hybrid Techniques

All the methods described above uses neural network for splice site detection. But it is hard to extract the specific features of DNA sequence using neural network because it works as a black-box. Therefore some methods need to be employed with neural network to extract the specific feature. In the literature, neural network has been combined with many other techniques to detect the splice sites.

1) Neural network-Markov hybrid technique: Accurate splice site detection technique needs to take into account the relationship among nucleotides or features of the surrounding nucleotides. High order Markov model has been used to represent complex relationship among nucleotides but it is difficult to implement practically as it requires the estimation of large number of parameters. On the other hand, neural networks have been widely used for pattern matching and classification but it is unable to find the underlying relationship among the nucleotides. An efficient approach for splice site detection has been developed that combine hidden Markov model and neural network approaches [15, 16, 17]. All the methods described above use orthogonal encoding in which sequence of nucleotides is represented as sequence of bits which was unable to represent features of the nucleotides surrounding the signals. In this approach, inputs of the neural network are encoded using lower-order Markov chains which is refer to as Markov encoding of inputs. Input sequence is represented as Markov Chain where each nucleotide represents a state and there is transition from state i to i+1, each state is characterized by a position-specific probability parameter. The splice site detection model consists three consecutive DNA segments: upstream segment, signal segment, downstream segment. The signal segment consists of nucleotides immediately neighbouring the splice sites. The upstream and downstream segments on the both sides of signal segment represent the features of the coding and noncoding sequences. These three segments are represented as Markov chains and a feed forward multilayer neural network received the Markovian probabilities as inputs. The signal segment is modelled as first-order Markov chain and upstream and downstream segments are represented by second-order Markov model. The emission probabilities from the Markov chains are fed into the neural network that had n input nodes that is equal to number of states in splice site model. Two different networks are used for acceptor and donor site, each network has a single output that gave a score to the input signal site. This approach realize high-order Markov model with the help of neural network. This method gives sensitivity of 0.86 and specificity of 0.85 for acceptor sites, sensitivity of 0.89 and specificity of 0.84 for donor site prediction.

2) Neuro-Fuzzy Techniques: Adaptive network based fuzzy inference system [18] (ANFIS) has been used for splice site detection. This Neuro-Fuzzy inference system outperforms other machine learning methods those have been used for splice site detection. This method gives the classification rate of 0.95 which is better than other methods. Another attempt was made for splice site detection using Adaptive network based fuzzy inference system [19]. But here instead of giving sequence as input to the neural network, the input sequence is processed and clustered into small categories and training of neuro-fuzzy network is carried out based on the similarities and dissimilarities of the selected clusters. Each cluster is given as input to individual network. This method when tested on homo-sapiens gives 80% sensitivity and 50 % sensitivity for donor site. This method has been tested for many organisms (Chicken, Daniorerio, Homo-sapiens, Mouse, Rat) and the experiments show that the prediction rate is same for all.

III. Discussion And Conclusion

Neural network has been widely used for splice site detection. Initially, splice site detection methods were based on only neural networks such as NetGene, NetPlantGene, NetGene2, NNSplice where input DNA sequence is given as input to the neural network. But these methods do not give promising results. To enhance the performance of splice site detection systems various types of input encoding methods has been used. In complementary encoding method, input sequence is encoded using complementary encoding which reduces the training time of neural network, but the results are not much improved. Neural network based systems are combined with other techniques to improve splice site detection. Neural/Markov hybrid systems give promising results. Neuro-Fuzzy based inference systems has been used for splice site detection which has improved the performance than other methods. Table 1 summarizes various techniques of splice site detection based on neural netwok.

References

[1] Zhang, Quanwei , "Splice sites detection by combining Markov and hidden Markov model, Biomedical Engineering and Informatics", 2009. BMEI '09. 2nd International Conference

[2] Churbanov A, Ali H, "Combinatorial method of splice site prediction", Proceedings of the 2005 IEEE computational systems bioinformatics conference, pp 189–190

[3] Soren Sonnenburg, Gabriele Schweikert, Petra Philips, Jonas Behr2 and Gunnar Ratsch, "Accurate splice site prediction using support vector machines", BMC Bioinformatics 2007, 8 (Suppl10);S7

[4] Awadalla S, Ortiz JE, Gopal S , "Prediction of trans-splicing sites using a genetic algorithm", Res Comput Mol Biol 2005 pp 1–7
[5] Watanabe T, Kudo Y, Shimizu T, "Positional correlations in splicing patterns and its application to prediction of splice sites", Genome Inform 2002 13:426–427
[6] Kashiwabara AY, Vieira DCG, Machado-Lima A, Durham AM, "Splice site prediction using stochastic regular grammars", Genetic Mol Res 2007 6:105–115
[7] S. Brunak, J. Engelbrecht, S. Knudsen, "Prediction of human mRNA donor and acceptor sites from the DNA sequence", Journal of Mol. Biol. 220(1991)
[8] Stefan M. Hebsgaard, Peter G. Korning, Niels Tolstrup, Jacob Engelbrecht, Pierre Rouzé1 and Soren Brunak, "Splice site prediction in Arabidopsis thaliana pre-mRNA by combining local and global sequence information", Nucleic Acids Research, 1996, Vol. 24, No. 17 3439–3452
[9] Niels Tolstrup, Pierre Rouze and Soren Brunak, "A branch point consensus from Arabidopsis found by non-circular analysis allows for better prediction of acceptor sites", Nucleic Acids Research, 1997, Vol. 25, No. 15 3159–3163
[10] Hatziglorgiou, N. Mache, M. Reczko, "Functional Site Prediction on the DNA sequence by Artificial Neural Networks", IEEE (1996)
[11] M.G. Reese, F.H. Eeckman, D. Kulp, D. Haussler, "Improved splice site detection in Genie", First Annual International Conference on Computational Molecular Biology (RECOMB), New York, ACM Press (1997) 232-240
[12] H. Ogura, H. Agata, M. Xie, T. Odaka, H. Furutani, "A study of learning splice site of DNA sequence by neural network", Comp. Biol. Med. 27(1997) 67-75
[13] T. Cai, Q. Peng, "Predicting splice sites in DNA sequences using neural network based on complementary encoding method", In Proc. of International Conference on Neural Networks and Brain (2005) 473-476
[14] O. Johansen, T. Ryen, T. Eftesol, T. Kjosmoen, P. Ruoff, "Splice Site Prediction Using Artificial Neural Networks", CIBB 2008, LNBI 5488, pp. 102–113, 2009, Springer-Verlag Berlin Heidelberg 2009
[15] Ho Sy hi, Jagath C. Rajapakse, "Splice site detection with neural networks/Markov models hybrids", Proceedings of the 9th International Conference on Neural Information Processing (ICONIP'OZ) , Vol. 5
[16] Ho Sy hi, Jagath C. Rajapakse, "Splice Site Detection with a Higher-Order Markov Model Implemented on a Neural Network", Genome Informatics 14: 64–72 (2003)
[17] Ho Sy hi, Jagath C. Rajapakse, "Markov Encoding for Detecting Signals in Genomic Sequences", IEEE/ACM Transaction on Computational Biology and Bioinformatics, VOL. 2, NO. 2, April-June 2005
[18] Essam Al-Daoud, "Identifying DNA splice sites using patterns statistical properties and fuzzy neural networks", EXCLI Journal 2009;8: 195-202 ISSN 1611-2156
[19] Fahimeh Moghimi, Mohammad Taghi Manzuri Shalmani, Ali Khaki Sedigh, Mohammad Kia, "Two new methods for DNA splice site prediction based on neuro-fuzzy network and clustering", Neural Comput & Applic, Springer-Verlag London 2012
[20] Vladimir B. Bajic, Suisheng Tang1, Hao Han, Vladimir Brusic, A. G. Hatzigeorgiou, "Artificial Neural Networks Based Systems for Recognition ofGenomic Signals and Regions: A Review", Informatica 26 (2002) 389–400

TABLE I
Neural Network Based Splice Site Detection Technique

Method	Neural Network Architecture	Organism	Methodology
NetGene [7]	Feed Forward	Humans	Orthogonal encoding of input, cut-off level of output is controlled by coding/non-coding network
NetPlantGene [8]	Feed Forward	Arbidopsis Thaliana	Results of neural network are combined with rule based system
NetGene2 [9]	Feed Forward	Humans, Arbidopsis Thaliana	Branch points are predicted to improve acceptor site prediction
CLA based system [10]	Cascade learning	Humans	Asymmetric window extraction for intron and exon portion
NNSplice [11]	Feed Forward	Humans	Models the pairwise correlations between the adjacent nucleotides
Hierarchical Neural Network Simulator [12]	Feed Forward	Humans	Three types of networks are trained using the arrangements of bases around the splice site
CEM based system [13]	Feed Forward	Humans	Complementary encoding of input, A is encoded by 1, T by -1, C by 2 and G by -2

Back propagation and curve fitting [14]	Feed Forward	Arbidopsis Thaliana	Output of neural network is used to find the location of splice site using curve fitting of a parabolic function
NN/Markov hybrid [15]	Feed Forward	Humans	Input of the neural network are encoded using Markov chains
Neuro-Fuzzy inference system [18]	Adaptive network based fuzzy inference system	Humans	Sequence is given input to Adaptive network based fuzzy inference system
Neuro-Fuzzy inference system with clustering [19]	Adaptive network based fuzzy inference system	Chicken, Daniorerio, Homo-sapiens, Mouse, Rat	Input sequence is clustered and neuro-fuzzy network is trained based on similarities and dissimilarities of the selected cluster

