Splice Site Detection Using a Combination of Markov Model and Neural Network

A K. M Abdul Baten [1], Saman K. Halgamuge [1], Nalin Wickramarachchi [2] and Jagath C. Rajapakse [3]

[1] *Mechatronics Research Group, Department of Mechanical and Manufacturing Engineering The University of Melbourne, Parkville-3010, Melbourne, Australia*

[2] *Department of Electrical Engineering, University of Moratuwa, Moratuwa, Katubadda,, Sri Lanka*

[3] *Bioinformatics Research Centre, School of Computer Engineering Nanyang Technological University, Singapore*

Email: a.baten@pgrad.unimelb.edu.au

ABSTRACT: This paper introduces a method which improves the performance of the identification of splice sites in the genomic DNA sequence of eukaryotes. This method combines a low order Markov model in series with a neural network for the predictions of splice sites. The lower order Markov model incorporates the biological knowledge surrounding the splice sites as probabilistic parameters. The Neural network takes the Markov encoded parameters as the inputs and produces the prediction. Two types of neural networks are used for the comparison. This method reduces the computational complexity and shows encouraging accuracy in the predictions of splice sites when applied to several standard splice site dataset.

1 INTRODUCTION

The DNA sequences of most genes create messenger RNA (mRNA), which encodes for protein. In prokaryotes the mRNA is a mere copy of a fragment of the DNA, whereas, in eukaryotes the RNA copy of DNA (the primary transcript or pre-mRNA) contains coding (Exon which code for proteins) and non-coding segments (Intron that do not code for proteins), which should be precisely spliced out to produce the mRNA. The border between introns and exons are termed as splice sites. The splice site in the upstream part of the intron is called the donor splice site and the downstream part is termed as acceptor splice site. The donor splice site usually contains the dinucleotide GT and the acceptor splice site contains dinucleotide AG. The more accurately a splice site can be located, the easier and more reliable it becomes to locate genes-hence protein coding regions in a DNA sequence. So identifying splice site accurately is still a worthwhile problem to be solved.

A number of computational methods have been developed to identify these splice sites, including both stand alone splice site finders and gene finders. The gene finding methods identify splice sites as an essential part of the gene identifying task. Sequencing of many genomes has already been completed, and sound computational model is necessary to find splice sites and hence genes as the amount of training data are also increasing. Computational techniques and algorithms that identify splice site include: neural network approaches [1, 2, 4, 11], probabilistic approaches [3, 9, 12], methods based on discriminant analysis [15]. These methods work based on seeking consensus patterns or features and try to identify the underlying relationships among nucleotides in the splice site and the surrounding region by using sets of training data containing true and false splice sites. Neural networks learn the complex features of neighbourhoods surrounding the consensus di-nucleotides [AG/GT] by learning a complex non-linear transformation. Probabilistic models estimate position specific probabilities of splice sites by computing likelihoods of candidates of signal sequences. The discriminant analysis uses several statistical measures to evaluate the presence of specific nucleotides, allowing recognizing the splice sites without explicitly determining the probability distributions.

Higher-order markov models are always considered as potential models for representing regions of nucleotides in the splice sites; however, their implementation has been practically prohibitive due to the need for large number of training data samples and compute-intensive nature of the training algorithms [3,8]. Neighbouring nucleotides are strongly correlated in the Splice Site (SS) consensus pattern. Neural network approaches take inputs from a neighbourhood window of nucleotides and are capable of learning complex interactions of nucleotides by finding arbitrary complex non-linear mapping.

The method of Loi Sy Ho and J. C. Rajapakse [8], showed for the first time that it is possible to implement a higher order markov model by combining lower order markov models with backpropagation neural network. In this method one first-order and two second-order Markov chains are used in the first stage. The first order Markov chian represents the consensus sequence, and the second order Markov chains are used to model the codon biases around the splice sites. The probabilistic parameters produced by all three Markov models are fed to a feedforward neural network in the second stage. The use of the neural network on top of Markov chain model enables this local interaction of nucleotides to represent higher-order dependencies. In this paper we show that it is possible to achieve almost the same level of accuracy by using a single, first-order Markov model in the first stage (instead of using three different Markov models as

suggested in [8]), which leads to less complexity and computational time.

Several experiments have been performed with this hybrid method, which showed better accuracy and efficiency than that of some other well known splice site detection methods. The proposed method accurately identifies approximately 95% of acceptor and donor sites and falsely predicts approximately 5% of the sites, which outperforms almost all the best-known methods. Sections 2 introduce biologically related computational models of acceptor sites and donor sites. Section 3 explains how a combination of a neural network and low-order markov chains is capable achieving higher-order Markov models of splice sites.

2 SPLICE SITE MODELS

This method is designed in two stages; lower-order Markov chain at the first stage and a three a layer neural network at the second stage The Markov chains aim to model the conserved pattern present of the splice sites and to exhibit the difference in characteristics of coding and non-coding regions before and after the sites. The probabilistic parameter results from the Markov models are then fed to a feed forward neural network, at the second stage, whose outputs are used to make the prediction.

2.1 Markov Chain Models

Segments of genomic sequences are often modelled by Markov chains whose observed state variables are elements drawn from the alphabet Ω_{DNA}, the set of four nucleotides: A, T, G, and C [8]. The Markov chain is defined by a number of states equal to the number of nucleotides in the sequence; each state variable of the model corresponds to a nucleotide in the sequence. Let us define an arbitrary sequence of length l : $\{s_1, s_2, s_3, \ldots\ldots\ldots, s_l\}$ such that $s_i \in \{A, C, G, T\}$, $\forall i \in \{1, \ldots\ldots\ldots, l\}$, then the nucleotide s_i is a realization of the i th state variable of the Markov chain and except from state i to state $i+1$, there is no other transitions from state i to other states. The model consists of states ordered in series. It evolves from state s_i to s_{i+1} and emits symbols from the alphabet Ω_{DNA} ; in which each state is characterized by a position-specific probability parameter.

Suppose the Markov chain, say M , has an order k , the likelihood of a sequence, implied by the model M , is given by

$$P(s_1, s_2, \ldots\ldots s_l | M) = \prod_{i=1}^{l} P_i(s_i), \qquad (1)$$

where s_i is a nucleotide at position i of the sequence and the Markovian probability -

$P_i(s_i) = P(s_i | s_{i-1}, s_{i-2}, \ldots\ldots s_{i-k})$ denotes the conditional appearance of the nucleotide at location i depending on the k predecessors.

A first order Markov model is used to represent the sequences containing the splice sites. Then the Markovian parameters are expressed in-terms of position-specific conditional probabilities -

$$P_i(s_i) = P(s_i | s_{i-1}, M), \qquad (2)$$

where the model is characterized by the set of parameters:

$M = \{P_i(s) | s \in \Omega_{DNA}, i = 1,2,\ldots\ldots\ldots, l\}$ and s is any element of the alphabet Ω_{DNA}.

It would be better to represent the model of the splice site by using higher-order Markov chains to capture all possible interactions among nucleotides, surrounding the splice sites. However, to attain a higher-order Markov model, the set of training sequences must be very large. For an *n*-order Markov model, the training set must cover all possible subsequences of nucleotides of length *n+1* at every sequence position in the splice site model. That is, constructions of a k th order Markov chain requires estimation of at least 4^{k+1} Markovian parameters. This implies that the required number of training samples increases exponentially with the order of the model.

2.2 Neural Networks

Neural network is the computational technique inspired by biological neurons with the ability to adapt or learn, to generalize and to cluster or organize data. Typically, a neural network comprises of many layers of neurons (units or nodes), each of which performs two functions, namely aggregation of its inputs from other neurons or the external environments and generation of an output from aggregated inputs.

In this work a multilayer neural network with fully connected weights is used. We will also use a radial basis function network (RBFN) for comparison purpose. The neural network is trained by a backpropagation algorithm and captures the higher-order dependencies around the splice site. The feedforward multilayer neural network receives its inputs as Markovian probabilities generated by the first order Markov chain M . Suppose, the neural network has n input nodes and if the input to the jth input node is x_j ,

$$x_j = P_i(s_i), \qquad (3)$$

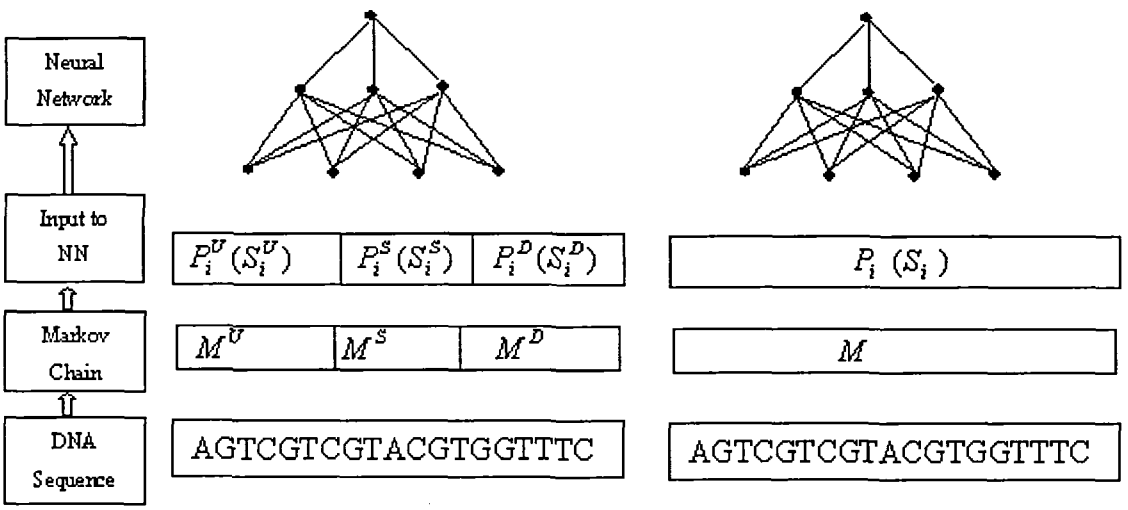
where $P_i(s_i) = P(s_i | s_{i-1}, s_{i-2}, \ldots\ldots s_{i-k})$

The neural network has one hidden layer of m units and one output unit. The network output y predicts whether the input sequence contains an actual splice site or not, where y is given by [8]-

$$y = f\left(\sum_{k=1}^{m} w_k f_k \left(\sum_{j=1}^{n} w_{kj} x_j \right) \right), \qquad (6)$$

where $f_k, k = 1,2,....,\ m$, and f denote the activation functions of the hidden neurons and the output neuron, respectively, $w_k, k = 1,2,....,\ m$ and $w_{kj}, k = 1,2,....,\ m, j = 1,2,.....,\ n$ denote the weights connected to the output neuron and to the hidden layer neurons, respectively. The output activation function is a unipolar sigmoidal and the hidden layer activation functions take the form of hyperbolic tangent sigmoidals.

Where $\hat{P}_i$s are the empirical probability, coefficients a_j are non-negative real numbers satisfying $\sum_{j=0}^{i} a_j = 1$, and the fraction (i/L) accounts for unseen occurrences and guarantees for no zero probability.



(a) Loi-Rajapakse method *(b) Proposed method*

Figure1: Block diagram of the proposed method and Loi-Rajapakse method (a), which consists of two-second order (M^U and M^D), and one first order (M^S) Markov chain to model the splice site. In contrast the proposed method (b), consist of only one first order Markov chain (M).

2.3 Higher-order Markov Model of Splice Sites

The low-order Markov chain provides a probabilistic description of signals. The neural networks receive Markov probabilities and combines non-linearly in order to incorporate more complex and distance interactions among elements in the splice sites. This sections it is shown that, by connecting the outputs of low-order Markov models to the neural network it is possible to achieve a higher-order Markov model.

Schukat-Talamazzini *et. at.* [18], has introduced the interpolated Markov chain for stochastic language modeling.

By the linear interpolation, we have

$$P(s_i|s_1^{i-1},M) \approx a_0\frac{1}{L} + a_1\hat{P}_i(s_i) + a_2P_i(s_i,s_{i-1}) + ...+ a_iP_i(s_1^i), \quad (7)$$

The higher-order conditional dependencies can be approximated by interpolation given a sequence $(s_1, s_2, ...s_i)$:

$$P(s_i|s_1^{i-1}) \approx \frac{\sum_{j=0}^{i-1} a_j g_j(s_{i-j}^{i-1})\hat{P}(s_i|s_{i-j}^{i})}{\sum_{j=0}^{i-1} a_j g_j(s_{i-j}^{i-1})} \quad (8)$$

where g_j is a sigmoid function.

By using the chain rule of probabilities, the likelihood of the sequence s_1^l is given by

$$P(s_1, s_2, ...s_l) = P(s_1)\prod_{i=2}^{i} P(s_i|s_1^{i-1}). \quad (9)$$

By induction of Eq. (8), that is by replacing conditional probabilities with the probabilities conditioned by a less number of elements [8],

$$P(s_1, ... s_l) \approx P(s_1)\prod_{i=2}^{l}\sum_{j=i}^{j-1} b_{ij}\hat{P}(s_i|s_{i-j}^{i-1}), \quad (10)$$

where $\{b_{ij} : i = 2,....l, j = 1,.....i\}$ is a set of linear coefficients. That is, the non-linear relationship amongst

variables in the sequence can be represented by a polynomial of sufficient order.

As supported by [10], a neural network with a single hidden layer, having a sufficient number of hidden neurons, is capable of approximating the continuous multi-variate functions defined on a hypercube $[0,1]^n$, thereby, the input-output relationship represented by any higher-order polynomial. So by application of Eq. (10), the neural network receiving inputs from low-order Markov chains. Whose output is represented in the form [8]

$$y = \sum_{m_1,.....m_l=0;m_1+....m_l=l} c_{m_1,.....m_l} P_1(s_1)^{m_1} .....P_l(s_l)^{m_l} , \quad (11)$$

where $\{m_i; i = 1,2,.....l\}$ are non-negative integers, $\{c_{m_1,.....m_l}; m_i = 1,2,.....l; i = 1,2,.....l\}$ are a set of real value coefficients, and $\{P_i(s_i): i-1,2,.....l\}$ are Markovian probabilities computed from low-order model M. Observing Eq. (10) and Eq. (11) it can be deduced that the neural network output, y, represents a higher-order Markov model, as also pointed in [8]. The higher-order Markov model takes care of all the conditional interactions among all the elements in the input sequence.

3 Experiments

3.1 Dataset

Several experiments have been performed to evaluate the performance of the method with dataset GS1115, provided by Pertea *et. al.*[9], dataset NN269, prepared by Reese *et. al.*[11], HS3D splice site dataset, provided by Rampone *et. al.*[17] and, splice-site detection dataset from the statlog's collection, provided by Michie *et. al.*[6].

In the dataset NN269, together with 1324 confirmed true acceptor sites and 1324 confirmed true sites extracted from 269 human genes, 5552 false acceptor and 4922 false donor sites with the consensus dinucleotides appearing in a neighbourhood of plus and minus 40 nucleotides around a true nucleotide were also collected.

The dataset GS1115 was constructed by using the Exon-Intron database to collect a confirm gene set. After disregarding genes with unknown introns, the dataset consists of 1115 human genes; from which 5733 true acceptor and 5733 true donor sites with confirmed AG or GT dinucleotides present were extracted. Additionally, 650099 false acceptor and 488983 false donor sites with confirmed GT or AG dinucleotides present, which are not annotated as true sites, were collected.

The Homo sapiens splice site dataset (HS3D) is a dataset of Homo Sapiens Exon, Intron and splice regions. From the complete GenBank primate sequences release 123 (8436 entries), then 2955 true acceptor and 2992 donor sites have been extracted as windows of 140 nucleotides around each splice site. Also 287,296 false acceptor sites and 348,370 windows of false donor sites have been selected, by searching canonical AG-GT pairs in non-splicing positions.

The Statlog's DNA dataset is a collection of primate splice junctions gene sequences containing 1770 acceptor and donor sequences and 1416 sequences which are neither acceptor nor donor. There are 2000 training sequences and 1186 test sequences.

3.2 Implementation

The training of the model has been done in two phases: in phase one, the Markov chain's model parameters were estimated and, in phase two, the neural network was trained. First of all, the training sequences were aligned with respect to the consensus dinucleotides; sequences without consensus dinucleotides were discarded to, obtain the maximum likelihood (ML) estimate of the Markov model parameters. The estimates of the $k-$order Markov model, in this case, $\hat{P}_i$, are given by the ratios of the frequencies counted from all partial sequences of $k+1$ elements at i and k elements at $i-1$ positions [8]:

$$\hat{p}_i(s_i) = \frac{\#(s_{i-k}^{i})}{\#(s_{i-k}^{i-1})}, \quad (12)$$

where $k = 1$ for the first order Markov chain and $\#(.)$ presents the observed frequency of its arguments in training dataset. To avoid some frequencies being zeroes due to the non-existence of the corresponding subsequences in the training data set, a constant extra value was added to every counted frequencies. Desired outputs were set to either 0.9 or 0.1 to represent the true or false site at the output.

3.3 Results and Discussion

Several experiments have been performed on the basis of two measures: the sensitivity (S_N) and the specificity (S_P) of the model i.e. the percentage of false sites wrongly predicted as true (%FP) and the percentage of true sites wrongly predicted as false (%FN).

$$sensitivity = \frac{TP}{TP+FN}$$

$$specificity = \frac{TN}{TN+FP}$$

The sensitivity is the correct prediction of true sites and the specificity is the correct prediction of the false sites. Table 1. shows the performance of Loi-Rajapakse method and the proposed method for dataset (GS1115), where the present method showed superior performance for both acceptor and donor splice sites. Here accuracy is defined as :

$$Accuracy = \frac{S_N + S_P}{2}$$

Dataset	Splice site	Loi-Rajapakse method	Proposed Method
		Accuracy	*Accuracy*
GS1115	Acceptor	0.945	**0.954**
	Donor	0.940	**0.958**

Table 1: Comparison of performance between Loi-Rajapakse method and proposed method on dataset GS1115.

Table 1 indicates that the proposed method can accurately identify approximately 95% of the splice sites when applied on non-redundant 1115 human gene dataset (GS115).

Dataset	Splice site	Proposed method	Markov+ RBFN
		Accuracy	*Accuracy*
NN269	Acceptor	**0.970**	0.900
	Donor	**0.960**	0.900
HS3Dataset	Acceptor	**0.945**	0.820
	Donor	**0.960**	0.870
Statlog's DNA Dataset	Acceptor	**0.855**	0.815
	Donor	**0.860**	0.840

Table 2: Comparison of performances between the proposed method with the combination of Markov model and radial basis function networks on three standard splice site dataset.

To evaluate the performance of the current method, it is further applied to three other standard splice site datasets NN269, HS3Dataset and Statlog's DNA dataset. Firstly the performance of the method is compared with the combination of Markov model and radial basis functions networks and secondly with backpropagation neural networks and radial basis function networks. In most cases current showed better accuracy than other methods as shown in table 1 and table 2.

Dataset	Splice site	Proposed method	BPNN	RBFN
		Accuracy	*Accuracy*	*Accuracy*
NN269	Acceptor	**0.970**	0.775	0.775
	Donor	**0.960**	0.895	0.640
HS3Dataset	Acceptor	**0.945**	0.825	0.770
	Donor	**0.960**	0.830	0.825
Statlog's DNA Dataset	Acceptor	**0.855**	0.850	0.790
	Donor	**0.860**	**0.875**	0.795

Table 3: Comparison of performances between the proposed method with the standalone backpropagation neural network and radial basis function networks.

The overall performance of the method is encouraging except the Statlog's DNA dataset, where the performance is not very satisfactory due to the limitations of training samples. Moreover proposed method reduces the computational time and complexity of the original method. The Loi-Rajapakse method divided the splice site into three signal segments namely upstream segment, signal segment and downstream segment. The upstream and downstream segments are modelled by two second order Markov chains and the signal segment by a first order Markov chain. The proposed method models the whole sequence with a single first order Markov chain. The computational complexity increases exponentially with the order of the Markov chain. Hence the proposed model reduces the computational complexity and also the prediction accuracy is increased.

4 CONCLUSION

Accurate identification of splice sites will enhance the performance of gene finding methods. This paper proposes an improved computational method, which will enhance the performance of splice site detection in eukaryotic genes with a higher-order Markov model. Although higher-order Markov models are considered as accurate models to characterize splice sites, their direct implementations were not feasible because of the limitations of estimating the large number of parameters, using limited amount of training data. The use of lower-order Markov models followed by a neural network provides an efficient way of implementing models for the detectionof splice sites.

5 REFERENCES

[1] Hatziggeorgious, A., Mache, N., and Reczko, M., Functional site prediction on the DNA sequence by artificial neural networks, *In Proc. IEEE Int. Joint Symposia in Intelligence and Systems,* IEEE Computer Society, 12-17, 1996.

[2] Brunak, S., Engelbrecht, J., and Knudsen S., Prediction of human mRNA donor and acceptor sites from the DNA sequence, *Journal of Molecular Biology,* 220:49-65,1991.

[3] Burge, C., and Karlin, S., Prediction of complete gene structure in human genomic DNA, *Journal of Molecular Biology,* 268:78-94,1997.

[4] Hebsgaard, S. M., korning P. G., Tolstrup. N., Engelbrecht, J., Rouze, P., and Brunak, S., Splice site prediction in Arabidopsis Thaliana pre-mRNA by combining local and global sequence information, *Nucleic Acids Research,* 24:3439-3452, 1996.

[5] Burset, M., and Guigo, R., Evaluation of Gene Structure Prediction Programs, *Genomic,* 34: 353-367, 1996.

[6] Guigo, R., Agarwal P., Abril., J. F., Burset, M., and Fickett, J. W., An assessment of gene prediction accuracy in large DNA sequences, *Genomic Research,* 10:1631-1642, 2000.

[7] Patterson, D. J., Yashuhara, K., and Ruzzo, W. L., Pre-mRNA secondary structure prediction aids splice site prediction, *Pacific Sym. On Biocomputing, World Scientific,* 223-234, 2002.

[8] Loi, S. H., Rajapakse, J. C., Splice site detection with a higher-order Markov model implemented on a Neural network, Genome Informatics 14: 64-72 (2003).

[9] Pertea, M., Xiao Ying,, and Salzberg, L., GeneSplicer: A new Computational method for splice site detection, *Nucleic Acids Research,* 29:1185, 2001.

[10] Pinkus, A., Approximation theory of MLP in neural networks, *Acta Numerica,* 143-195, 1999.

[11] Reese, M. G., Eeckman, F. H., Kulp, D., and Haussler, D., Improved splice site detection in Genie, *computational Biology,* 4:311-324, 1997.
[12] Sonnenburg, S., New methods for detecting splice junction sites in DNA sequence, Master's Thesis, Humbold University, Germany, 2002.
[13] Thanaraj, T. A., Positional characterisation of false positives from computational prediction of human splice sites, *Nucleic Acids Research,* 28:744-754, 2000.
[14] Yin, M., and Wang, J. T. L., Effective hidden Markov models for detecting splice junction sites in DNA sequences, *Information Sciences,* 139:139-163, 2001.
[15] Zhang, M. Q., Identification pf Protein coding regions in human genome by quadratic discriminal analysis, *Proc. Natl. Acad. Sci, USA,* 565-568, 1997.
[16] Christina L. Z., Michael G., T. Murlitharan, Nair., Virginia R. D. S., On selecting features from splice junctions: An analysis using information theoretic and machine learning approaches, *Genome Informatics,* 14: 73-83 2003.
[17] Pollastro, P., Rampone, S. HS3D-Homo Sapiens Splice Sites Dataset, Nucleic Acids Research, Annual Database Issue, 2003.
[18] E. G. S Talamazzini., Hidden Markov of models - the current key word, Artificial Intelligence , 11(4): 37-38 1997.