Current Bioinformatics, **2013**, *8*, **369-379**

Hidden Markov Model for Splicing Junction Sites Identification in DNA Sequences

Srabanti Maji and Deepak Garg*

Department of Computer Science and Engineering, Thapar University, Patiala-147004, India

Abstract: Identification of coding sequence from genomic DNA sequence is the major step in pursuit of gene identification. In the eukaryotic organism, gene structure consists of promoter, intron, start codon, exons and stop codon, etc. and to identify it, accurate labeling of the mentioned segments is necessary. Splice site is the 'separation' between exons and introns, the predicted accuracy of which is lower than 90% (in general) though the sequences adjacent to the splice sites have a high conservation. As the accuracy of splice site recognition has not yet been satisfactory (adequate), therefore, much attention has been paid to improve the prediction accuracy and improvement in the algorithms used is very essential element. In this manuscript, Hidden Markov Model (HMM) based splice sites predictor is developed and trained using Modified Expectation Maximization (MEM) algorithm. A 12 fold cross validation technique is also applied to check the reproducibility of the results obtained and to further increase the prediction accuracy. The proposed system can able to achieve the accuracy of 98% of true donor site and 93% for true acceptor site in the standard DNA (nucleotide) sequence.

Keywords: Algorithms, coding sequence, cross validation, gene finding, hidden markov model, modified expectation maximization (MEM), splice site.

1. INTRODUCTION

A genomic sequence is a string composed of four different nucleotides, A, T, G and C, which codifies in group of three, called codons that are amino acids that form the proteins and are necessary for all organisms to live. A very large number of computational solutions for the gene identification problem have been reported which are the valuable resources for the human genome program and for the molecular biology community. A gene is a structure that codifies the proteins [1, 2]. In prokaryotes, it is a sequence of codons between a start codon (ATG) and a stop codon (TAA, TAG or TGA) whereas in eukaryotes, the structure is more complex. The coding sequence is usually broken by non-coding sequences, called introns that are removed during the transcription in a process called splicing [3]. The coding sections are called exons. In this manner, the eukaryotic gene begins with first exon, then any number of intron/exon pairs, and ends with a last exon which finishes with a stop codon. This is called an open reading frame (ORF). The eukaryotic genes are composed by a single exon. The boundary between an exon and an intron is called a splice donor site and that between an intron and an exon, a splice acceptor site. The actual gene has the sequences of nucleotides before start codon and after stop codon, known as the untranslated terminal regions (UTRs). However, it is not uncommon in gene recognition to use the term "gene" when referring only to the coding part of it, since that part only determines the protein structure [4].

*Address correspondence to this author at the Department of Computer Science and Engineering, Thapar University, Patiala-147004, India;
Tel: +91-175-2393007/+91-9815599654; Fax: +91-175-2393005;
E-mail: dgarg@thapar.edu

Gene recognition, gene structure prediction or gene finding, all of these three terms consists of determining those parts of a sequence which are coding and constructing the whole gene from its start site to its stop codon. Here, we are concerned with the work related to eukaryotic gene recognition, as it is significant, useful and complex as well. There are two basic approaches to predict the gene structure [5]; first one is homology based approaches that search for similar sequences in databases of known genes and are usually called extrinsic methods. The growing number of sequenced genomes and known genes is increasing the potential of homology based methods. However, it is clear that only genes that are somewhat similar to known genes can be identified in this way. Furthermore, when using homology based techniques, it is very difficult to establish the complete structure of the gene, as the exact bounds of the exons are not easy to determine with certainty. The second approach, usually known as intrinsic approach includes two basic methods: ab initio and de novo [6]. Both are based on obtaining the features that characterize a coding region and/or the functional sites, and using them to find the correct structure of the unknown genes. Ab initio methods use only the information of the genome to be annotated (the target genome), whereas de novo methods add information of one or more related genomes (the informant genomes).

The main function of eukaryotic gene structure predictors is to pin point the locations of all start codons, stop codons, exons and introns in every gene and this step is considered as the rate-limiting step in the gene identification. In predicting splice site (which is the separation between exon and intron), the initial task is finding exons and introns. Splice site junction identification means the identification of donor site (5' boundary containing dinucleotide GT) and acceptor site (3' boundary containing dinucleotide AG) of introns [7-10]. The success in gene prediction largely depend on the

2212-392X/13 $58.00+.00

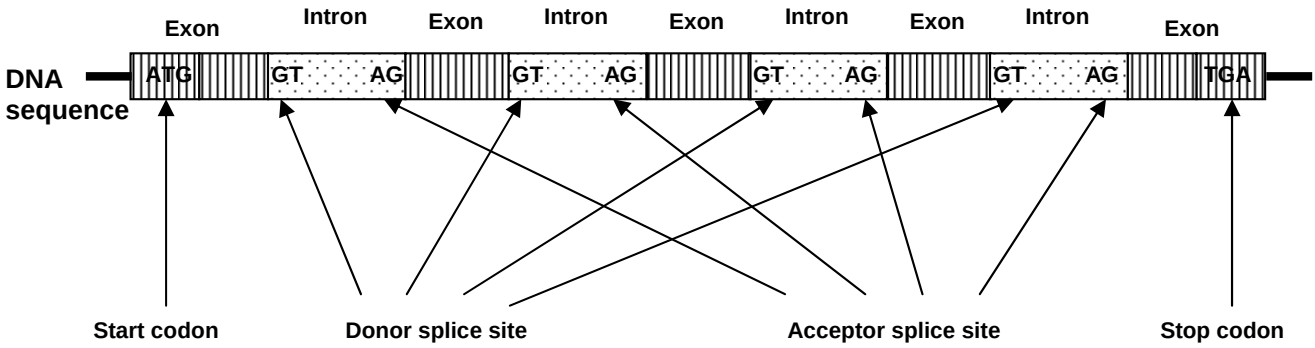

Fig. (1). The splice sites (Donor site and Acceptor site) in eukaryotic DNA sequence.

accuracy in finding the splice site junction, and thus, the removal of the introns from the DNA sequence to get coding regions is possible [7]. Bioinformatics unite the capability and knowledge of researcher from computational and biological areas, and locate a familiar stage for people from these backgrounds to work collectively to decipher gene annotation challenges [11]. Splice site in eukaryotic DNA sequence is shown in Fig. (**1**).

Although methods to predict potential protein coding regions on genomic DNA sequences came into existence since 1980s, the first program to assemble potential DNA coding regions into translatable mRNA sequences were not available until the early 1990s [8]. From the recent past, there are several programs available for biology scientists. GRAIL is the one amongst them, which is widely used today and is available on the BLAST web site for gene structure detection (BLAST: http://www.ncbi.nlm.nih.gov) [8, 9].

Hidden Markov Models (HMM) have been applied successfully in various applications, viz. speech recognitions [10]. An HMM model is a type of process in which some of the details are unknown or hidden and is stochastic in nature. This process uses a number of states and probabilistic state transitions and is usually represented by a graph in which transitions are represented by edges and states by vertices. Individual states are denoted by Y, which are associated with a discrete output probability distribution, P(Y). Transition probability is the probability of going from a certain state to the next state. Thus, the sum of the probabilities of all the transitions from a given states *s* to all other states must be 1. Markov and HMMs are gaining popularity in bioinformatics research for nucleotide sequence analysis [12-16]. For prokaryotes gene identification, Borodovsky *et al.* [17] effectively applied this HMM technique. Eukaryotic promoter detection algorithm using a Markov transition matrix was proposed by Audic and Claverie [18]. A new technique VEIL (Viterbi Exon-Intron Locator) was developed by Salzberg [19] and Henderson *et al.* [20] to identify translational start site and splice sites in eukaryotic mRNA. The HMM based gene predictor GeneScout was developed by Yin *et al.* [21], to detect translational start site and mRNA splicing junction sites. Our proposed technique used in this manuscript differs from Salzberg's and others, in which two different HMM are used; one for 5' and another for 3'splice sites. Every model consists of two elements; one for false sites and another for true sites.

2. RESOURCES AND METHODOLOGY

2.1. Dataset Collection

To build reliable expanded Hidden Markov Model for the detection of human splice sites, high-quality datasets must be used. Splice site dataset is collected from the website http://www.fruitfly.org/sequence/human-datasets.html. There is a collection of 2381 true donor sites and 2381 true acceptor sites from a set of 462 annotated multiple-exon human genes. After removing junk sequence (splice sites that contained base positions not labeled with A, T, C, G but with other symbols) there remained 2379 true donor sites and 2379 true acceptor sites, which were used as the true dataset. Hence, every acceptor site has a conserved AG di-nucleotides and every donor site has a conserved GT di-nucleotides. We also collected a large database of 300,062 false donor sites and 400,314 false acceptor sites from the 462 annotated genes and used it as the 'false dataset'.

Afterwards, we used a 12-fold cross-validation in our dataset to estimate the splice site detection accuracy of all the models. Cross validation is a standard experimental technique in which each model is verified by randomly partitioning the data into several subsets [21, 22]. We tested each subset (testing data) with the parameters trained by the other twelve subsets (training data) under the splice site model. After completing all these operations we took the average of the twelve predictive accuracy measures corresponding to the 12 testing/training data pair. Our proposed HMM system is trained with sequences which contains 2179 true site and 275,055 false sites, tested with 200 true sites and 25,005 false sites, for every time in the cross validation testing.

2.2. Proposed Models

In our proposed models for the identification of acceptor and donor splice sites, the splice site classification problem is subdivided into two – acceptor splice site classification and donor splice site classification. Two different models are constructed for the identification of acceptor splice sites and donor splice sites respectively.

2.3. Notations

For simplicity, we are providing some basic notations which are shown in the Table **1**.

Table 1. Some Basic Notations Related to Acceptor, Donor Hidden Markov Model and Splice Site

Symbol	Description
X	Base in Hidden Markov Model
Y	Various states in the Hidden Markov model
T	Various transitions in the Hidden Markov model
P(Y)	Discrete state probability
P(T)	Transition probability
C_{site}	A candidate sequence
THV	Predefined threshold value
F	Flag variable
L	Length of the candidate site
MOD_t	True Acceptor HMM Unit
MOD_f	False Acceptor HMM Unit
N	The set of sequences that are randomly picked from the positive training data set and negative training data set.
N^t	The sequence collection containing the remaining sequences in the positive training data set, after taking some positive training dataset for M.
N^f	Represent the remaining sequences in the negative (non-coding) training data set, after picking some negative training data set for M.
P	Subset of sequence collection N
MEM	Modified Expectation Maximization (E-M) algorithm.
L_b	The positive lower bound
S_n^{mem}	The sensitivity during the MEM training.
S_p^{mem}	The specificity during the MEM training.
S_{total}	S_{total} represent the total number of states in the Acceptor Model.
$b_i\left(b_i \in \{A,\ G,\ C,\ T\}\right)$	Base at state i, $1 \le i \le S_{total}$.
$tr_i\left(b_i,\ b_{i+1}\right),\ 1 \le i \le S_{total}-1$	The transition from state i to state i+1.
$T_{in}^{(t)}$	Total number of true acceptor sites that have been input into True Acceptor HMM Unit
$T_{in}^{(f)}$	Total number of false acceptor sites that have been input into False Acceptor Unit.
$FLAGHMM_i$	A flag indicating whether C_{site} is a true acceptor site or not.
S_n^{true}	Sensitivity or TPR of the HMM.
S_n^{false}	Specificity of the HMM.
ACC	Accuracy of the Hidden Markov Model.
$f\ \ tr_i^{(t)}(b_i, b_{i+1})$	State transition probabilities in True Acceptor HMM/True Donor HMM Unit.
$f\ \ tr_i^{(f)}(b_i, b_{i+1})$	State transition probabilities in False Acceptor HMM/False Donor HMM Unit.

2.4. Donor Site Hidden Markov Model (HMM) for 5' Splice Site

The nucleotide sequences must pass through this model to move from exon model to intron model. 11 nucleotide bases with GT are included in the conserved sequences which are almost consistent to all the donor sites [2, 23]-[24], an example of which is shown below:

ATGACGTGACC

The di-nucleotides GT are located in position 6 and 7 respectively. The exon-intron boundary occurs between stages 5 and 6, 1-3 is a start codon and so 4-5 are the part of exon and 6-11 are the part of intron. The location of G and T is 6 and 7 respectively in all the true 5' splice sites [25]. There is an 11-base non-donor sequence also present in which the G and T are located at position 6 and 7

respectively as a "false donor site". The motive of our proposed algorithm is to identify whether the given sequence (candidate) is a true donor site or a false donor site.

In our proposed donor HMM for identifying true donor site, 11 states and a set of transitions is used, which is represented as a digraph where vertices depicts the states and edges depicts the transitions. At each state, the model generates a base 'X' in {A, G, C, T} accordance with the state and transition probabilities, with the exception of states 6 and 7. At the state 6, the donor HMM consistently generates base $X = G$, and at state 7, $X = T$. Every state Y is coupled with an output probability distribution, P(Y). We can simply observe that the value of P(Y) is 1 for states 6 and 7. The transition probability of HMM to make a transition is denoted as P(T). At state 5, every base has a constant transition, $P(T) = 1$, to the base **G** at state 6. Similarly, at state 6, the base **G** has a constant transition, $P(T) = 1$, to the base **T** at state 7. The donor site HMM for 5' splice site is shown in Fig. (**2**).

2.5. Acceptor Site Hidden Markov Model (HMM) for 3' Splice Site

In DNA, the acceptor sites are the preserved boundary sequences at 3' splice sites which include 17 nucleotide bases with AG almost consistent to all acceptor sites [2, 26, 27], for example,

CTATCCTTCTCACAGGG

In an acceptor site, nucleotide A and G are located at positions 12 and 13 respectively [28]. There is also a non-acceptor sequence, in which the location of A and G are 12 and 13, which are considered as false acceptor site. Therefore, the proposed algorithm attempts to identify whether the given sequence is true donor site or false donor site. The acceptor HMM for 3' splice site is used to express the basic properties of true acceptor sites.

In our proposed donor HMM for identifying true acceptor site, 17 states and a set of transitions is used, which is represented as a digram where vertices depicts the states and edges depicts the transitions. In a nucleotide sequence, states 1 to 13 belong to an intron and state 14-17 belong to an exon. At each state, the model generates a base 'X' in {A, G, C, T} accordance with the state and transition probabilities, with the exception of states 12 and 13. At the state 12, the acceptor HMM consistently generates base $X = A$, and at state 13, $X = G$. Every state Y is coupled with an output probability distribution, P(Y). We can simply observe that the value of P(Y) is 1 for states 12 and 13. The transition probability of HMM to make a transition is denoted as P(T). At state 11, every base has a constant transition, $P(T) = 1$, to the base A at state 12. Similarly, at state 12, the base **G** has a constant transition, $P(T) = 1$, to the base G at state 13. The acceptor site HMM for 3' splice site is shown in Fig. (**3**).

2.6. Unit Creation for Each Model

The number of false splice sites present is much larger than the number of true splice sites in vertebrate DNA sequence. To identify their difference, for Donor HMM System, we have created two programs – True Donor HMM

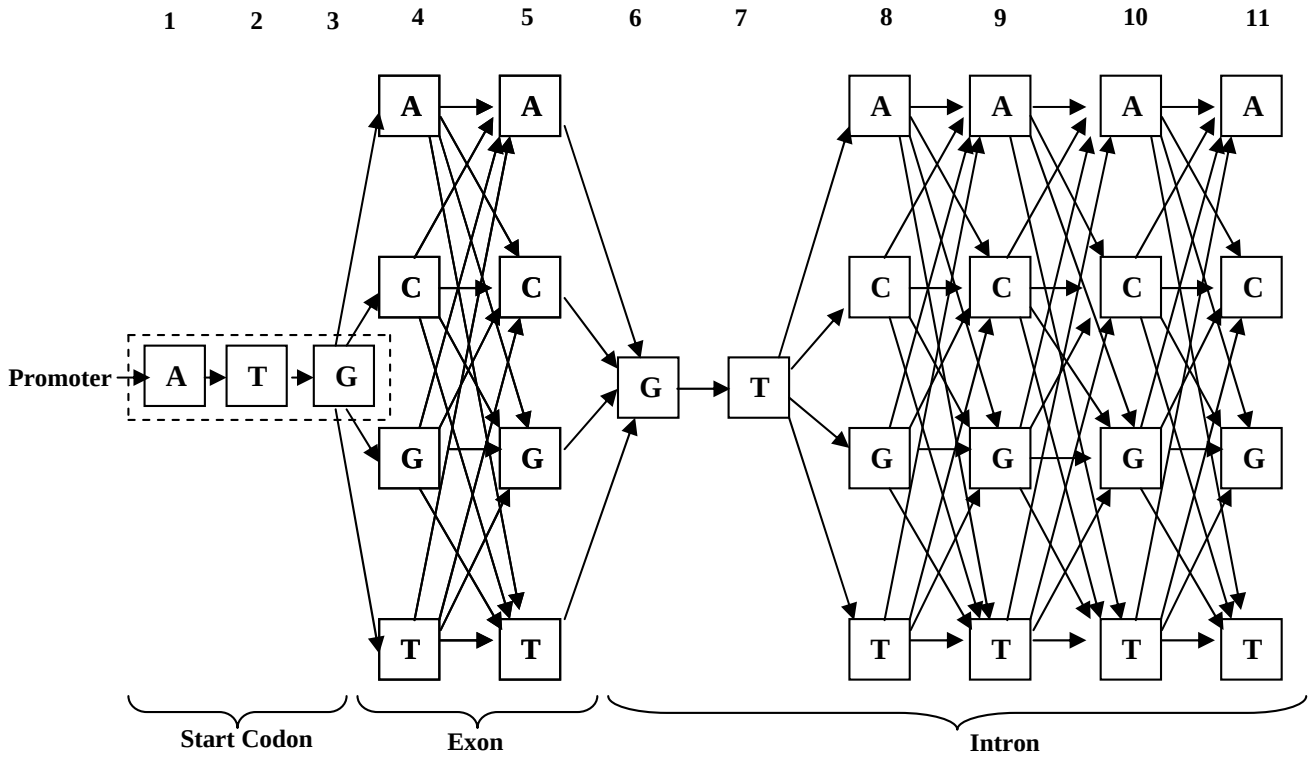

Fig. (2). The Donor site HMM for 5' splice site.

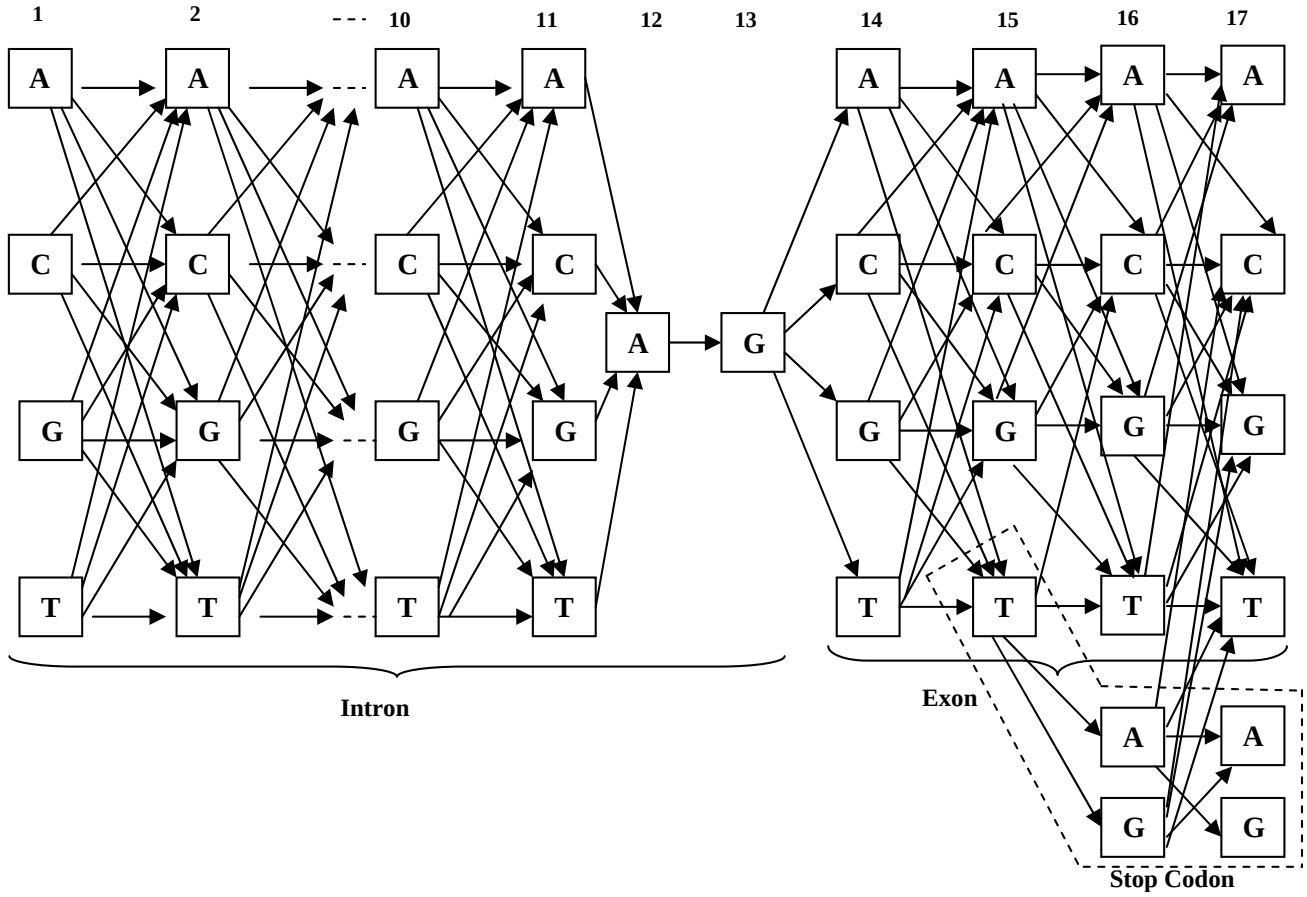

Fig. (3). The Acceptor HMM for 3' splice site.

Unit and False Donor HMM Unit. Similarly, for Acceptor HMM System, another two programs – True Acceptor HMM Unit and False Acceptor HMM Unit are created. The True splice site HMM Unit is the integration of True Donor HMM Unit and True Acceptor HMM Unit; and in a similar manner, False Splice site HMM Unit is the combination of False Donor HMM Unit and False Acceptor HMM Unit. Here, we assume that C_{site} represents the given DNA sequence, and MOD_t and MOD_f denotes True splice site HMM element/Unit and false splice site HMM Unit respectively. For splice site predication, true site and false units are used to classify the given sequence into appropriate categories. We assume that the probability of donor site is $P(X=1|C_{site}, MOD_t)$ when the given sequence is processed by True Donor HMM unit and the probability of the non-donor site as $P(X=0|C_{site}, MOD_f)$ when it is processed by False Donor HMM Unit. The training data for true and false splice sites are used to give training to the true splice site HMM Unit and false splice site HMM Unit respectively. To calculate the result of C_{site}, initially we run True Donor HMM Unit to obtain the probability of being a donor site sequence and then, False Donor HMM Unit to obtain the probability of non-donor sequence. After comparing these values, our given C_{site} is assigned to false donor category or true donor category. Figs. (**4**, **5**) shows the creations of True Splice site HMM unit and False Splice site HMM unit.

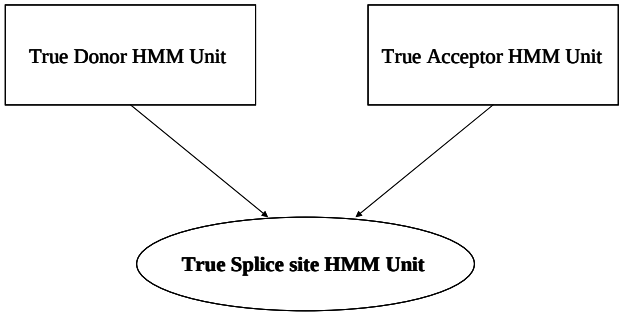

Fig. (4). The True Splice site HMM Unit.

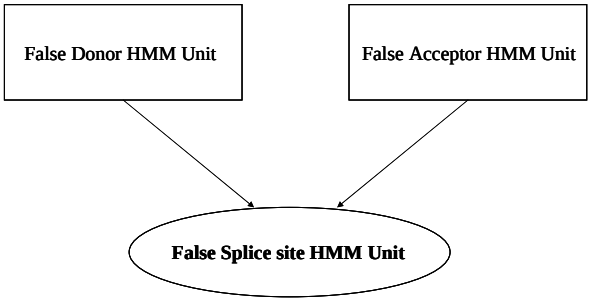

Fig. (5). The False Splice site HMM Unit.

To train these HMM Units, we used modified expectation maximization (MEM) algorithm. In the basic EM algorithm, a set of unaligned sequence and a motif length are provided as input resulting in a probabilistic model for motif [29-32]. Also, each iteration consist of two steps namely expectation step (E-step) and the maximization step (M-step). But, many of pre-trained values, such as those in splice junction models, are fixed and can not be modified by the EM algorithm, whereas, as our dataset contains splicing junction sites of same length which may be aligned to each other, therefore, we developed the proposed MEM algorithm for training a HMM with fixed topology. In this MEM algorithm, we trained the module iteratively to get the maximum value of specificity, i.e. the fraction of correctly classified sites or until positive training data set, N^t and negative training data set N^f become empty with the condition that he value of sensitivity, S_n^{mem} during the training period remains constant.

Assuming that all these sequences may be aligned to each-other; our designed Modified EM (MEM) algorithm works in the following manner: Initially, the value of all the transition probabilities P(T) and state probabilities P(Y) are set to 0 and the HMM Unit topology is constant. Then the first subset of positive training data (e.g. 120 sequences) is given as input to the True Donor HMM Unit; the numbers of the individual bases at each state and from present state to the next are recorded. Afterwards, the prior probabilities for all the states and transitions are calculated in the True Donor HMM Unit. After getting the prior probabilities, we provided another subset of positive training data to the True Donor Unit, and all the subsequent probabilities are re-adjusted. After this, we calculated the differences, *diff*, for all the probabilities between the earlier and subsequent probabilities. If some of the *diff* are larger than a predefined threshold value (THV), set the current posterior probabilities as the new prior probabilities, and the new data set is then run through the True Donor HMM Module again to further refine the probabilities. This training process is repeated until the changes in all probabilities in the True Donor HMM Unit are smaller than the THV. The False Donor HMM Unit is trained using the negative training data in the same way as for the True Donor HMM Unit.

2.7. Algorithms

Three efficient algorithms – Forward, Viterbi and Expectation Maximization (EM) are used for HMM computation. The proposed algorithms can be used mutually for the Donor HMM System and Acceptor HMM System. Initially, Acceptor HMM System and its related units are created, and then, the True Acceptor HMM Unit and False Acceptor HMM Unit are formed accordingly. The algorithms for the Donor HMM System are developed in the similar manner.

2.8. Training Algorithm

In the training algorithm, N represents the set of sequences which are arbitrarily selected from the positive and negatively training data sets, contains about 200 true acceptor sites and 19,000 false acceptor sites. Each sequence in N is labeled as N^t if it is taken from positive training data set and N^f if from the negative training data set and, P is the subset of N. The sum of sequences in N^t and N^f exceeds about twelve times the number of sequences in N.

The algorithm converges in the training phase by advancing iteratively. A few sequences from N^t and N^f are removed by the algorithm at each iteration, and inputs those into True Acceptor HMM Unit and False Acceptor HMM Unit. Then, the algorithm determines the sequences those are located in the subset P. During the MEM training, let S_n^{mem} represents the sensitivity, which is the ratio between the number of true acceptor sites in P and the total number of true acceptor sites in N; and S_p^{mem} represents the specificity, which is the ratio between the number of true acceptor sites in P and the total number of sequences in P. Here, it is important to note that $P \subseteq$ (belongs to) N and the goal of the MEM training is to train the Units repeatedly to get a maximal value of S_p^{mem} until N^t and N^f is emptied, provided S_n^{mem} remains constant. Here we have taken the value of $S_n^{mem} = 0.92$ for the purpose.

Specifically, S_{total} represents the total number of states in the Acceptor Model and $b_i\left(b_i \in \{A, G, C, T\}\right)$ be the base at state i, $1 \le i \le S_{total}$. and $tr_i\ (b_i, b_{i+1}), 1 \le i \le S_{total} - 1$ be the transition from state i to state $i+1$. The topology for the Acceptor HMM System is fixed, and all of the transition probabilities and state probabilities are initialized to random values. Then we selected one twelfth of the sequences from N^t and provided as input into the True Acceptor HMM Unit. At the same time, one twelfth of the sequences from N^f are selected and these are fed as input into False Acceptor HMM Unit. The number of the individual bases b_i and the number of individual transitions from one state to the next state, $tr_i\ (b_i, b_{i+1})$ are recorded at each state. Then we calculated the post probabilities for all the states and transitions in True Acceptor HMM Unit and finally, the False Acceptor HMM Units are computed. Considering $T^{(t)}tr_i\ (b_i, b_{i+1})$ as the total number of transitions from a base b_i at state i to a base b_{i+1} at state $i+1$ in True Acceptor HMM Unit and, $T_{in}^{(t)}$ be the total number of true acceptor sites that have been input into True Acceptor HMM Unit, the state transition probabilities, $f\ tr_i^{(t)}(b_i, b_{i+1})$, in True Acceptor HMM Unit can be calculated from the following equation:

$$f\ tr_i^{(t)}(b_i, b_{i+1}) = \frac{T^{(t)}tr_i(b_i, b_{i+1})}{T_{in}^{(t)}}. \quad (1)$$

Similarly, if $T^{(f)}\ tr_i\ (b_i, b_{i+1})$ is the total number of transitions from a base b_i (at state i) to a base b_{i+1} (at state $i+1$) in False Acceptor HMM Unit and $T_{in}^{(f)}$ is the total number of false acceptor sites that have been input into False Acceptor Unit, then, the state transition probabilities,

$f\ tr_i^{(f)}(b_i, b_{i+1})$ in False Acceptor Unit can be calculated from the following equation:

$$f\ tr_i^{(f)}(b_i, b_{i+1}) = \frac{T^{(f)} tr_i(b_i, b_{i+1})}{T_{in}^{(f)}} \quad (2)$$

Subsequently, all sequences contained in N, which are unlabeled, are considered as input to the True Acceptor and False Acceptor HMM Units. Let $P\left(\text{True} \mid Y, N^{(t)}\right)$ represents the probability of a sequence Y (acceptor sequence) in set N and $P\left(\text{True} \mid Y, N^{(f)}\right)$, the probability of Y (non-acceptor sequence). In order to calculate $P\left(\text{True} \mid Y, N^{(t)}\right)$, the probability of sequence Y must be known by using True Acceptor HMM Unit, which can be computed as follows:

$$p(Y \mid True, N^{(t)}) = \prod_{i=1}^{T_{states}-1} ftr_i^{(t)}(b_i, b_{i+1}),\ b_i \in \{A, G, C, T\}. \quad (3)$$

The proposed MEM algorithm uses Bayesian Theorem (See eq. S1 in supplementary data) for calculating $P\left(\text{True} \mid Y, N^{(t)}\right)$ from $P\left(Y \mid \text{True}, N^{(f)}\right)$,

$$P(True \mid Y, N^{(t)}) = \frac{P(Y \mid True, N^{(t)}) P(True)}{P(Y)} \quad (4)$$

where $P(True) =$ prior probability (assumed to be a constant), $P(Y) =$ product of the individual base probabilities in the sequences (See eq. S2 and S3).

Similarly, equations can be derived for calculating $P\left(\text{False} \mid Y, N^{(f)}\right)$ as follows:

$$P(Y \mid False, N^{(f)}) = \prod_{i=1}^{T_{states}-1} ftr_i^{(f)}(b_i, b_{i+1}),\ b_i \in \{A, G, C, T\}, \quad (5)$$

$$P(False \mid Y, N^{(f)}) = \frac{P(Y \mid False, N^{(f)}) P(False)}{P(Y)} \quad (6)$$

Assuming the probability ratio of sequence Y in the dataset N is represented by *pr*

$$pr = \frac{P(True \mid Y, N^{(t)})}{P(False \mid Y, N^{(f)})}. \quad (7)$$

Once the *pr* is calculated for each sequence in set N, then the sequences in set N is sorted in the descending order according to their respective *pr* values. If the total number of positive sequences in set N is S_{pt}, we select the *pr* value for $S_{pt} * S_n^{mem}$ th positive sequence and use that value as the positive lower bound, denoted by L_b. The sensitivity S_n^{mem} of 200 positive sequences in set N is 0.92, so L_b is the *pr* value of the 184th positive sequence. A sequence $Y \in N$ into set P is assigned by the MEM algorithm if the *pr* value for $Y \geq L_b$. Let $T_{(P+N)}$ be the number of positive sequences in set N that are assigned into set P. Then, sensitivity during the MEM training will be given by

$$S_n^{mem} = \frac{T_{(TP)}}{T_{(P+N)}}. \quad (8)$$

and, let $T_{(pp)}$ be the total number of sequences in N that are assigned into P. Then, by definition, specificity during the MEM training will be given by

$$S_P^{mem} = \frac{T_{(TP)}}{T_{(PP)}}. \quad (9)$$

To increase S_p^{mem}, the entire probabilities are adjusted in the re-estimation procedure hidden in the Donor Model Acceptor System and the new sequences in N^t and N^f are chosen and removed. These sequences are then run through True Acceptor HMM Unit and False Acceptor HMM Unit again and the probabilities are further refined. This process is repeated until the value of S_p^{mem} is maximized or the value of N^t and N^f become zero. Now, the positive lower bound L_b that maximizes S_p^{mem} will be considered as output and used in the detection phase for splicing junction sites. In the training period, MEM algorithm is used, which is depicted in Pseudocode 1.

Input

Untrained HMM site unit (including a true site unit and a false site unit);
Positive training data set, N^t;
Negative training data set, N^f;
MEM testing data set, N;
OUTPUT:
Fully trained HMM site unit and L_b;
ALGORITHM:
max:= false ;
do begin
max := true ;
if N^t is not empty **then begin**
remove one twelfth of the sequences from N^t and input them into the true site unit;
for $i = 1$ **to** $S_{total} - 1$
calculate $ftr_i^t(b_i, b_{i+1})$ as in Equation (1);
end;
if N^f is not empty **then begin**
remove one twelfth of the sequences from N^f and input them into the false site module;
for $i = 1$ **to** $S_{total} - 1$
calculate $ftr_i^f(b_i, b_{i+1})$ as in Equation (2);
end;
for each sequence $Y \in N$ **do begin**
calculate $P\left(\text{True} \mid Y, N^{(t)}\right)$ as in Equation (4);
calculate $P\left(\text{False} \mid Y, N^{(f)}\right)$ as in Equation (6);
calculate *pr* as in Equation (7);
end;
select L_b ;
calculate S_p^{mem} according to L_b ;
if (S_p^{mem} is not maximum) or (either N^t or N^f is non-empty) **then**
max := false ;
end;
while max

Pseudocode 1. The MEM Algorithm in training phase.

2.9. Splice Site Junction Detection Algorithm

The implication of a candidate acceptor site is a 17-base sequence section with the bases A and G at locations 12 and 13 respectively [2]. A section $C_{site,}$ of 17-bases (referred as $b_1, b_2, ..., b_{17}$ respectively) is taken as the input of the site detection algorithm, which is extracted from a genomic DNA sequence Y. The indication whether the C_{site} starting at position i of the genomic DNA sequence Y is a true acceptor site or not is estimated/verified by the output of the site detection algorithm, which is a flag given by $FLAGHMM_i$.

Now, considering that $f\ tr_j^{(t)}(b_j, b_{j+1})$ be the probability of a transition from base b_j to base b_{j+1} $(1 \le j \le 16)$, of C_{site} using True Acceptor HMM Unit, a flag variable F may be defined as 1 if C_{site} belongs to a true site category, otherwise, it will be 0. Let L be the length of the candidate site C_{site} (L is 17 for acceptor sites and 11 for donor sites) and $P\left(C_{site} \mid F=1, N^{(t)}\right)$ be the probability of the candidate site C_{site} with the condition that it is an acceptor site processed by True Acceptor HMM Unit, then

$$P(C_{site} \mid F=1, N^{(t)}) = \prod_{j=1}^{n-1} ftr_j^{(t)}(b_j, b_{j+1}),\ b_i \in \{A, G, C, T\} \quad (10)$$

Therefore, according to Bayesian theorem,

$$P(F=1 \mid C_{site}, N^{(t)}) = \frac{P(C_{site} \mid F=1, N^{(t)}) P(F=1)}{P(C_{site})}. \quad (11)$$

$P(F=1)$ can be treated as a constant [3] while examining a set of sequences to detect true acceptor sites. Then, $P(C_{site})$, the product of the individual base probabilities for $b_1, b_2, ..., b_n$ will be

$$P(C_{site}) = \prod_{j=1}^{n} P(b_j \mid F=1, N^{(t)}),\ b_i \in \{A, G, C, T\}. \quad (12)$$

In the same way as we've followed for True Acceptor HMM Unit, the False Acceptor HMM Unit can be calculated by eq. S4, S5, and S6 with flag variable, F = 0 and $N^{(t)}$ replaced by $N^{ft)}$.

Now, provided the candidate acceptor site C_{site} starting at position i in the DNA sequence Y is given, the proposed algorithm will find the two most likely sets of states through the two HMM Units for C_{site}. Then, the algorithm will calculate $P\left(F=1 \mid C_{site}, N^{(t)}\right)$ and $P\left(F=0 \mid C_{site}, N^{(f)}\right)$. Based on the scoring function, a score *sr* is assigned to the candidate site, as shown below:

$$sr = \frac{P(F=1 \mid C_{site}, N^{(t)})}{P(F=0 \mid C_{site}, N^{(f)})}. \quad (13)$$

After evaluating *sr* and L_b, a flag $FLAGHMM_i$, is assigned to the candidate site C_{site} and calculated. If $sr \ge L_b$, the value of $FLAGHMM_i$ will be 1 and the C_{site} is considered as true acceptor site, and if it is 0, then it is considered as a false acceptor site. The Acceptor Splice site junction classification algorithm is given in Pseudocode 2.

Input

A candidate acceptor site C_{site} of an unlabelled genomic DNA sequence starting at position i

Output

/* $FLAGHMM_i$ is a flag indicating whether C_{site} is a true acceptor site or not. */

$FLAGHMM_i$;

Algorithm

Calculate probability of transition of C_{site} using True Acceptor HMM Unit

$P(F=1 \mid C_{site}, N^{(t)})$ as in equation (11);

by calculating probability of transition of C_{site} using False Acceptor HMM Unit

$P(F=0 \mid C_{site}, N^{(f)})$;

Calculate *sr* as in Equation (13);

Calculate $FLAGHMM_i$;

Pseudocode 2. Acceptor Splice site junction classification algorithm.

3. RESULTS AND DISCUSSION

The classification performance of the models is measured in terms of their sensitivity S_n^{true} (TPR), and specificity S_n^{false} [4, 33]. In the classification performance, TP, TN, FP, and FN stand for true positive rate, true negative rate, false positive rate, and false negative rate respectively [2, 34] (As defined in Table **S1**). The state transition probabilities for the Acceptor and the Donor HMM Systems are shown in Tables **S2-S5**.

Our proposed system increased the differences between the true splice site and false splice sites to the maximum as verified from the results which are shown in Tables **2** and **3**. A 12-fold cross validation technique is applied to identify the splice site prediction accuracy, and the average results for all the 12 test sets are shown. Their classification efficiency was evaluated by various quantitative variables - (i) true positive (TP): the number of correctly classified splice site, (ii) true negative (TN): the number of correctly classified non-splice site, (iii) false positive (FP): the number of incorrectly classified splice site, and, (iv) false negative (FN): the number of incorrectly classified non-splice site. The sensitivity S_n^{true} or True Positive Rate (TPR), defined as the fraction of correctly classified true acceptor (or true donor) sites among the total number of true acceptor (or true donor) sites in the test data, is shown in following equation:

$$\text{TPR or } S_n^{true} = \frac{TP}{TP + FN} \quad (14)$$

Analogously, the specificity S_n^{false} is defined as the fraction of correctly classified false acceptor (or false donor) sites among the total number of false acceptor (or false donor) sites in the test data, i.e.

$$S_n^{false} = \frac{TN}{TN + FP}, \quad (15)$$

The similar calculations are also used for identifying false positive Rate (FPR) as the fraction of incorrectly classified true acceptor (true donor) sites among the total number of false acceptor (or false donor) sites in the test dataset, i.e.

$$FPR = \frac{FP}{TN + FP}, \quad (16)$$

Accuracy (ACC) is a parameter of the test which is the proportion of the candidate site in the given test data those are classified correctly (or accurately) and gives a fair idea that whether the proposed system can classify the true and false splice sites into right categories. Accuracy is calculated by the formula:

$$ACC = \frac{TN + TP}{TN + TP + FN + FP} \quad (17)$$

Acceptor HMM System can correctly identify 95% of the true acceptor sites and 92% of the false acceptor sites in the test data, as shown in Table **2**. Similarly Donor HMM System can able to predict 95% of the true donor sites and 97% of the false donor sites in the test data set, as depicted in Table **3**. Accuracy of the candidate acceptor sites is 92%, and for donor sites value is 97% [5]. Their accuracy performances are shown in Figs. (**6**, **7**) respectively.

Table 2. The Acceptor HMM Performance for 3' Splice Site Prediction

Set	No of True Acceptor	No of False Acceptor	TP	FP	TN	FN	Sensitivity S_n^{true}	Specificity S_n^{false}	FPR	*Accuracy ACC*
1	208	19782	190	1028	18754	18	0.9134	0.9480	0.0519	0.9476
2	200	21531	184	1273	20258	16	0.92	0.9408	0.0591	0.9406
3	209	21001	195	1299	19702	14	0.9330	0.9381	0.0618	0.9380
4	210	18965	197	1301	17664	13	0.9380	0.9313	0.0686	0.9314
5	203	18966	193	1297	17669	10	0.9507	0.9316	0.0683	0.9318
6	200	22000	191	1598	20402	9	0.9550	0.9273	0.0726	0.9276
7	208	21343	199	1587	19756	9	0.9567	0.9256	0.0743	0.9259
8	213	21457	206	1573	19884	7	0.9671	0.9266	0.0733	0.9270
9	206	20876	199	1578	19298	7	0.9660	0.9244	0.0755	0.9248
10	212	18790	206	1485	17305	6	0.9716	0.9209	0.0790	0.9215
11	209	17986	203	1490	16496	6	0.9712	0.9171	0.0828	0.9177
12	209	18003	203	1498	16505	6	0.9712	0.9167	0.0832	0.9174
Average							**0.9512**	**0.9290**	**0.0709**	**0.9293**

Table 3. The Donor HMM Performance for 5' Splice Site Prediction

Set	No of True Donor	No of False Donor	TP	FP	TN	FN	Sensitivity S_n^{true}	Specificity S_n^{false}	FPR	*Accuracy ACC*
1	208	16242	194	200	16042	14	0.9326	0.9876	0.0123	0.9869
2	200	15101	188	199	14902	12	0.94	0.9868	0.0131	0.9862
3	209	16261	196	231	16030	13	0.9377	0.9857	0.0142	0.9851
4	210	13411	198	235	13176	12	0.9428	0.9824	0.0175	0.9818
5	203	12301	192	266	12035	11	0.9458	0.9783	0.0216	0.9778
6	200	15235	190	348	14887	10	0.95	0.9771	0.0228	0.9768
7	208	17221	200	397	16824	8	0.9615	0.9769	0.0230	0.9767
8	213	15815	205	382	15433	8	0.9624	0.9758	0.0241	0.9756
9	206	15863	199	399	15464	7	0.9660	0.9748	0.0251	0.9747
10	212	13123	206	378	12745	6	0.9716	0.9711	0.0288	0.9712
11	209	14331	204	452	13879	5	0.9760	0.9684	0.0315	0.9685
12	209	14354	209	487	13867	5	0.9766	0.9660	0.0339	0.9662
Average							**0.9552**	**0.9776**	**0.0223**	**0.9773**

Table 4. Accuracy of Acceptor and Donor Splice Site Detection Compared for HMM System, NNSplice and GENIO on the Human Test Dataset

	Splice Site Predictor	Sensitivity	Specificity	False Positive Rate (FPR)
Acceptor Site	HMM System (all data)	0.9512	0.9290	0.0709
	NNSplice (all data)	0.6419	0.9483	0.05165
	GENIO (all data)	0.7959	0.9523	0.0476
Donor Site	HMM System (all data)	0.9552	0.9776	0.0223
	NNSplice (all data)	0.7116	0.9367	0.0633
	GENIO (all data)	0.8624	0.9406	0.0594

and for donor sites value is 97% [5]. Their accuracy performances are shown in Figs. (**6**, **7**) respectively.

The result of our proposed HMM System (Acceptor HMM System and Donor HMM System) on the test data were compared with NNSplice (http://www.fruitfly.org/seq_tools/splice.html), GENIO (http://genio.informatik.uni-stuttgart.de/GENIO) using our test data for the comparison. Both of these splice site predictors offer a web page (already mentioned) where the DNA sequences can be submitted for the generating the results of the data, therefore, we submitted our dataset to each of the websites, and used the default parameters to predict the results. Table **4** shows the overall comparison of our proposed HMM System with two of the systems – NNSplice and GENIO.

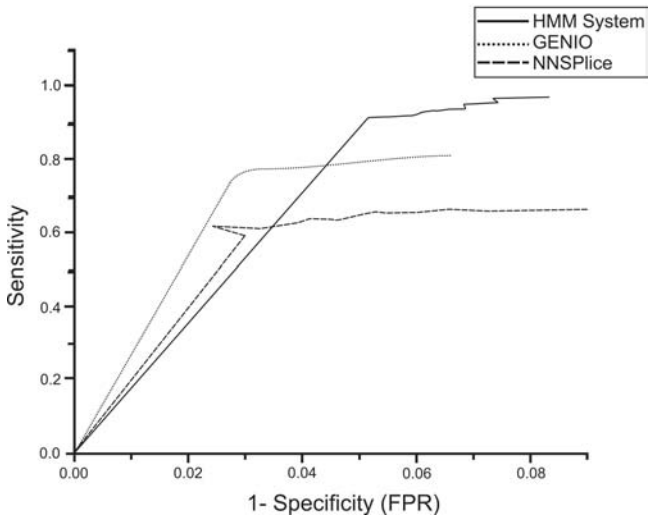

Fig. (6). Receiver Operating characteristic (ROC) curve showing the comparison of performance between HMM System, GENIO and NNSplice Acceptor test dataset.

4. CONCLUSION

A modified hidden markov model (HMM) is developed with new method for the identification of eukaryotic splice sites with a different topology (Donor and Acceptor Model) from the previously reported splicing junction detection mechanism. We have used 12-way cross validation experiment, which proves the method's simplicity and effectiveness. The comparison of our proposed HMM system with other splice site predictors NNSplice and GENIO indicates that HMM system is considerably better in sensitivity. In addition, the system is able to correctly predict 95% of the true donor sites and 97% of the false donor sites in the test data set; 95% of the true acceptor sites and 92% of the false acceptor sites in the test data set. Overall, this system is comparatively better in sensitivity and can correctly detect 97% of the true donor sites and 92% of the true acceptor sites in the standard sequenced data. Hence, this method can be utilized to identify splice sites in the large scale in newly genomes.

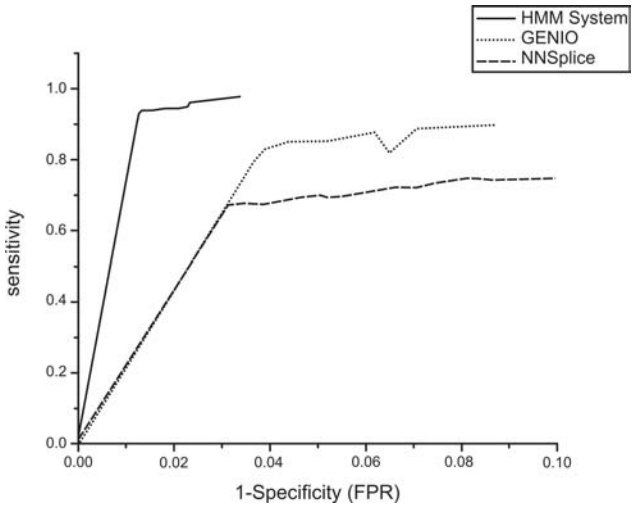

Fig. (7). Receiver Operating characteristic (ROC) curve showing the comparison of performance between HMM System, GENIO and NNSplice Donor test dataset.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflicts of interest.

ACKNOWLEDGEMENTS

The authors acknowledge Prof. Sanghamitra Bandyopadhyay, Machine Intelligence Unit, Indian Statistical Institute; Prof. Ujjwal Moulik, Department of Computer Science and Engineering, Jadavpur University; and Mr. Sabyasachi Hembram, Sr. Staff Engineer, Infinera for their expert, sincere, and valuable guidance in preparing the manuscript.

SUPPLEMENTARY MATERIAL

Supplementary material is available on the publisher's web site along with the published article.

REFERENCES

[1] Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. J Mol Biol 1997; 268(1): 78-94.

[2] Alberts B, Bray D, Lewis J, Raff M, Roberts K, Watson JD. Molecular Biology of the Cell. 3rd ed. Garland Publishing Inc., New York 1994.

[3] Ahmad M, Abdullah A, Buragga K. A Novel Optimized Approach for Gene Identification in DNA Sequences. J Appl Sci 2011; 11: 806-814.

[4] Schellenberg MJ, Ritchie DB, MacMillan AM. Pre-mRNA splicing: a complex picture in higher definition. Trends Biochem Sci 2008; 33(6): 243-246.

[5] Mathe C, Sagot MF, Schiex T, Rouzo P. Current methods of gene prediction, their strengths and weaknesses. Nucleic Acids Res 2002; 30(19): 4103-4117.

[6] Kumar M, Raghava GP. Prediction of nuclear proteins using SVM and HMM models. BMC Bioinformatics 2009; 10: 22.

[7] Pertea M, Lin X, Salzberg SL. GeneSplicer: A new computational method for splice site prediction. Nucleic Acids Res 2001; 29(5): 1185-1190.

[8] Yin MM, Wang JTL. Effective hidden Markov models for detecting splicing junction sites in DNA sequences. Inform Sciences 2001; 139: 139-163.

[9] Larranaga P, Calvo B, Santana R, *et al*. Machine learning in bioinformatics. Brief Bioinform 2006; 7(1): 86-112.

[10] Rajapakse JC, Ho LS. Markov encoding for detecting signals in genomic sequences. IEEE/ACM Trans Comput Biol Bioinform 2005; 2(2): 131-142.

[11] Ghosh Z, Mallick B. Bioinformatics principles and Applications. 2nd ed. Oxford University Press, Delhi 2009.

[12] Brendel V, Kleffe J. Prediction of locally optimal splice sites in plant pre-mRNA with applications to gene identification in Arabidopsis thaliana genomic DNA. Nucleic Acids Res 1998; 26(20): 4748-4757.

[13] Burset M, Guigo R. Evaluation of gene structure prediction programs. Genomics 1996; 34(3): 353-367.

[14] Lopez R, Larsen F, Prydz H. Evaluation of the exon predictions of the GRAIL software. Genomics 1994; 24(1): 133-136.

[15] Rabiner LR. Tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE 1989; 77(2): 257-286.

[16] Lukashin AV, Borodovsky M. GeneMark.hmm: New solutions for gene finding. Nucleic Acids Res 1998; 26(4): 1107-1115.

[17] Borodovsky M, McIninch J. GENMARK: Parallel gene recognition for both DNA strands. Computers and Chemistry 1993; 17(2): 123-133.

[18] Audic S, Claverie JM. Detection of eukaryotic promoters using Markov transition matrices. Comput Chem 1997; 21(4): 223-227.

[19] Sachem SI. A method for identifying splice sites and translational start sites in eukaryotic mRNA. Comput Appl Biosci 1997; 13(4): 365-376.

[20] Henderson J, Salzberg S, Fasman KH. Finding genes in DNA with a Hidden Markov Model. J Comput Biol 1997; 4(2): 127-141.

[21] Yin MM, Wang JTL. GeneScout: A Data Mining System for Predicting Vertebrate Genes in Genomic DNA Sequences. Inform Sciences, Special Issue on Soft Computing Data Mining 2004; 163(1-3): 201-218.

[22] Dong S, Searls DB. Gene Structure Prediction by Linguistic Methods. Genomics 1994; 23(3): 540-551.

[23] Brunak S, Enqelbrecht J, Knudsen S. Prediction of human mRNA donor and acceptor sites from the DNA sequence. J Mol Biol 1991; 220(1): 49-65.

[24] Ohshima Y, Gotoh Y. Signals for the selection of a splice site in pre-mRNA: Computer analysis of splice junction sequences and like sequences. J Mol Biol 1987; 195(2): 247-259.

[25] Chen TM, Lu CC, Li WH. Prediction of splice sites with dependency graphs and their expanded bayesian networks. Bioinformatics 2005; 21(4): 471-482.

[26] Wu S, Green MR. Identification of a human protein that recognizes the 3' splice site during the second step of pre-mRNA splicing. EMBO J 1997; 16(14): 4421-4432.

[27] Reed R. Initial splice-site recognition and pairing during pre-mRNA splicing. Curr Opin in Genet Dev 1996; 6(2): 215-220.

[28] Degroeve S, Saeys Y, De Baets B, Rouze P, Van de Peer Y. SpliceMachine: predicting splice sites from high-dimensional local context representations. Bioinformatics 2004; 21(8): 1332-1338.

[29] Bailey TL, Baker ME, Elkan CP. An artificial intelligence approach to motif discovery in protein sequences: Application to steroid dehydrogenases. J Steroid Biochem 1997; 62(1): 29-44.

[30] McLachlan GJ, Krishnan T. The EM algorithm and extensions. 2nd ed. Hoboken: Wiley-Interscience: John Wiley & Sons 2008.

[31] Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data *via* the EM algorithm. J Roy Stat Soc B Met 1977; 39(1): 1-38.

[32] Moon TK. The expectation-maximization algorithm. Signal Processing Magazine, IEEE 1996; 13(6): 47-60.

[33] Baten AKMA, Halgamuge SK, Chang BCH. Fast splice site detection using information content and feature reduction. BMC Bioinformatics 2008; 9(12): S8.

[34] Yeo G, Burge CB. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J Comput Biol 2004; 11(2-3): 377-394.

Received: June 19, 2012 Revised: October 17, 2012 Accepted: November 5, 2012