Plant Archives Vol. 18, Special Issue (ICAAAS-2018), 2018 pp. 141-147 *ISSN 0972-5210*

IDENTIFICATION OF EXONS IN EUKARYOTIC GENOMIC DNA USING MACHINE LEARNING APPROACH

Noopur Singh[1] and Ravindra Nath[2]

[1]Department of Biotechnology, APJAKT University, Lucknow (U.P.), India.
[2]University Institute of Engineering, Technology, C.S.J.M University, Kanpur (U.P.), India.

Abstract

Today bioinformatics is growing field due to its varied applications in the area of the biology and information science and engineering. It is a combination of biology, mathematics, statistics and information technology. Hence it is very helpful in solving many biological problems. Genome sequencing is becoming popular day by day and is broadly utilized in the gene finding methods for DNA sequences. The gene of eukaryotes is composed of DNA and the DNA sequence is contains four different nucleotides. The DNA sequence contains exon and intron regions that is the coding and non-coding regions respectively. Exon prediction in the gene of eukaryotes by computational methods is not yet solved fairly. The predicted exons are very useful for the research pioneers in the field of biological science. Now-a-days, a lot of work has been carried out in the area of gene prediction or exon prediction. A broad range of significant machine learning approaches are propounded by researchers for exon prediction. There are several machine learning approaches used in solving gene sequence problems. Hidden Markov Model (HMM) is one of the approach that has been used for exon prediction in an anonymous genomic DNA sequence. In this paper, we used HMM for the exon prediction.

Key words : Hidden Markov Model (HMM), Deoxyribonucleic Acid (DNA), coding sequence (CDS), ANN (Artificial Neural Network), SVM (Support Vector Machine), GA (Genetic Algorithm).

Introduction

The biology of DNA is quite complex and interesting. In eukaryotic (*i.e.* multicellular cell either of plants or animals) cell, the nucleus inside the cytoplasm contains the blueprint of life, the DNA. This DNA is composed of four nucleotide bases Adenine (A), Guanine (G), Cytosine (C) and Thymine (T). The Gene that is made up of small segments of DNA is the hereditary material in most of the eukaryotes. The gene is organized into two most important section exons and introns. Exons are the coding region and the introns are non-coding region of DNA. When DNA undergoes into the process of transcription, there occurs a process known as splicing, that removes the introns and join the exons. Basically, the exons are categorized into four classes: 5'exon, internal exon, 3'exon and intronless exon. In CDS, the exon region starts with codon ATG bases and ends with one of the three stop codon, TAA, TAG and TGA bases. At least two exons are there in CDS that makes us to know that there is minimum one intron region and usually intron region starts with GT bases and ends with AG bases (fig. 1). There are several machine learning methods like ANN, GA, SVM, and HMM used for exon prediction.

The term "exon" derives from the expressed region of gene, coined by a biochemist, Walter Gilbert in 1978. The notion cistron i.e., the transcriptional unit containing regions that lost from the mature mRNA called the intron (for the intragenic regions). Introns have a special character that they have two distinct nucleotides at either end. At the 5' end the DNA nucleotides are GT and at the 3' end they are AG. These nucleotides are part of the splicing sites [1].

Splice Site Donor : splicing site at the beginning of an intron, intron 5' left end.

Splice Site Acceptor : splicing site at the end of an intron, intron 3' right end.

The GT/AG mRNA processing rule is applicable for almost all eukaryotic gene. A polypyrimidine (C_nT_n) motif is present upstream of the CAG intron 3' ending. More upstream, the consensus branch site (CTGAC) is a necessary component in the effective splicing of the pre-

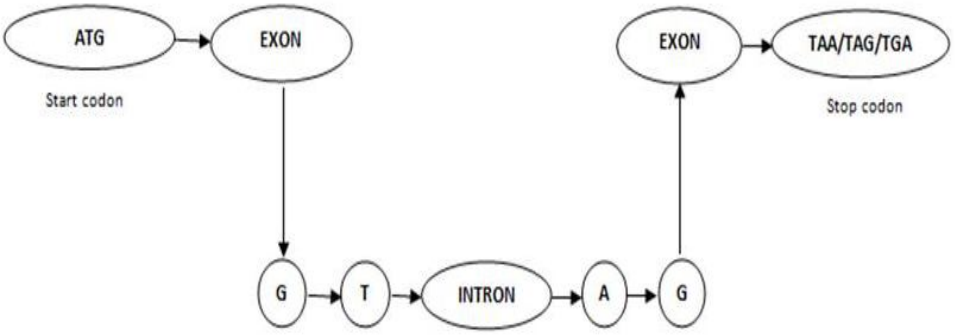

Fig. 1 : Basic structure of HMM model for exon prediction.

mRNA. Once introns are removed *i.e.* spliced out, the mature messenger RNA (mRNA) leaves the nucleus and is translated into protein (Protein synthesis).

Hidden Markov Model (HMM)

HMM is a strong statistical model that is actually an indexed sequence of random variables. Although, this model was developed mostly for speech recognition, pattern recognition, gesture recognition, etc., in 1970s. But nowadays, HMMs are widely applied in the analysis of biological sequences, specially the DNA. HMM is becoming a necessary part of bioinformatics.

Now, consider a process that can be described at any time as being in one of a set of N distinct states as illustrated in the fig. 2, N = 4. At any index of time, system undergoes a change of state (possibly back to the same state) according to a set of probabilities.

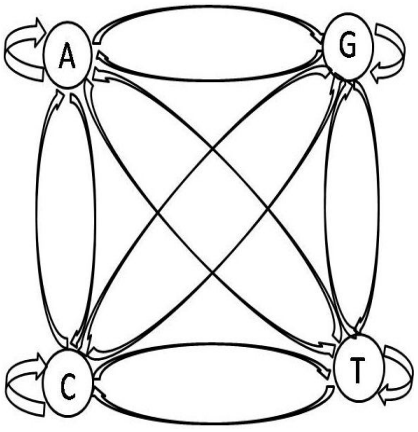

Fig. 2 : Markov chain with four states.

A full probabilistic description of the above system requires specification of current state and all previous states. For the special case of first order Markov process this probabilistic description is truncated to just the current and previous state. [2]

$$P(q_t = S_j | q_{t-1} = S_i,\ q_{t-2} = S_k,\ \ldots) = P(q_t = S_j | q_{t-1} = S) = a_{ij} \quad 1 < i, j < N \quad (1)$$

Where, q_t is actual state at time t and a_{ij} is the transition probability between state i and j. State transition coefficients have the property.

$$\sum_{j=1}^{N} a_{ij} = 1 \quad (2)$$

The above stochastic process could be called an observable Markov Model since the output of process is the set of states at each instant of time, where each state corresponds to a physical event. This model is too restrictive to be applicable to many problems of interest [2]. We can extend the concept of Markov models to include the case where observation is a probabilistic function of the states. Underlying a HMM is a basic Markov process that is not observable, but can be observed through another set of stochastic sources that produce observation.

1. Elements of Hidden Markov Models

In order to characterize an HMM completely, following elements are needed [2].

- The number of states of the model, N
- The number of distinct observation symbols per States M
- The state transition probability distribution

$$A = \{a_{ij}\} \quad a_{ij} = P(q_t = S_j \mid q_{t-1} = S_i) \quad (3)$$

- The observation symbol probability distribution in state

$$j,\ B = b_j(k) = P(V_k \ at\ t \mid q_t = S_j) \quad (4)$$

- The initial state distribution

$$\pi_i = P(q_1 = S_i) \quad (5)$$

- The model parameters notation:

$$\lambda = (A, B, \pi) \quad (6)$$

2. Three Basic Problems of HMM

There are three basic problems that have to be addressed in order to use HMMs in practical applications. Suppose we have a new symbol sequence $O = O_1, O_2, \ldots, O_T$

1. How can we compute the observation probability $P(O|\lambda)$ based on a given HMM?
2. Given a model λ and a sequence of observations $O = O_1, O_2, O_3, \ldots O_T$, what is the most likely state sequence in the model that produced the observations?
3. Given a model λ and a sequence of observations $O = O_1, O_2, O_3, \ldots, O_T$, how should we adjust the model parameters A, B, λ in order to maximize value of $P(O|\lambda)$

3. Solution of first problem

At first, we evaluate the probability of observation sequence using Forward-Backward algorithm. As this is a very efficient procedure even for large values.

4. Forward Procedure

The forward variable $\alpha_t(i)$ defined as

$\alpha_t(i) = P(O_1, O_2, O_3, \ldots, O_t, q_t = S_i|\lambda)$

Where, probability of observation sequence: $O_1, O_2, O_3, \ldots\ldots, O_T$

S = states at given time t and given model (λ),

We can solve $\alpha_t(i)$ inductively as follows:

1. Initialization: $\alpha_t(i) = \pi_i b_i(O_1) \quad 1 \le i \le N$
2. Induction : $\alpha_{t+1}(i) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(O_{t+1})$

$1 \le t \le T\text{-}1$

3. Termination : $P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i)$

$1 \le j \le N$

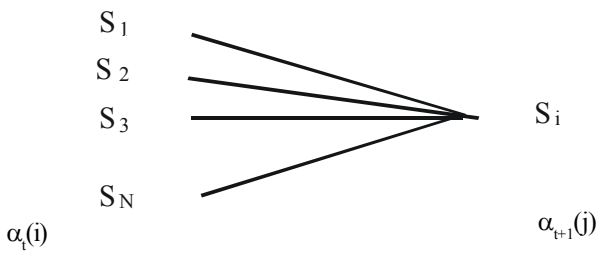

Fig. 3 : Representation of sequences for the computation of forward variable.

Step 1 : Induction step, which is the heart of the forward calculation in the fig. 3, state S_j can be reached at time t+1 from the N possible states S_i, $1 \le i \le N$ at time t. Summing this product $\alpha_t(i) * a_{ij}$ overall all the N possible states the result in the probability of S_j at time t+1 with all accompanying previous partial observations once this is done S_j is known it is easy to see that $\alpha_{t+1}(j)$ is obtained by accounting for observation O_{t+1} in the state j.

Step 2 : By multiplying the summed quantity by the probability $b_j(O_{t+1})$ the computation is performed for all states j, $1 \le j \le N$ for a given t, where t=1 …T-1

Step 3 : Gives the desired calculation of $P(O|\lambda)$ as the sum of the terminal forward variables $\alpha_t(i)$,

$\alpha_t(i) = P(O_1, O_2, O_3, \ldots\ldots, O_t, q_t = S_i|\lambda)$

5. Solution of second problem

We run it for the more than hundred times by changing the values of observation symbol matrix. We get more than hundred output values of $P(O|\lambda)$. Now, we find the maximum value of $P(O|\lambda)$ among them, which will be maximum probability values of $P(O|\lambda)$.

Implementation Details

1. To find exon and intron region in a gene sequence

Initially, we used computational method for finding start and stop codon for an anonymous DNA sequence. At last, all the regions of exon and intron have been fetched. For this we are using computational method. Here, an input (for a large observation sequence), *i.e.* the DNA sequence has been taken and processed. After computer processing an output generated.

2. To find the probability of observation sequence for ATG (Start Codons) and TAA/TAG/TGA (Stop Codon)

Here, using computer language, HMM forward method is implemented and an input file, as an observation sequence (large DNA sequence) has been taken.

For finding both start and stop codons, forward symbol α_t (of forward method) has been calculated. On the basis of previous value of α_t, next value of α_{t+1}, α_{t+2}, α_{t+3}.... are calculated for both start and stop codons individually. For calculation of above values we used state transition matrix ($A = a_{ij}$), observation symbol probability matrix ($B = b_j(k)$) and initial probability (π_i). In this way, we calculated n values of $P(O|\lambda)$ and find the maximum values of $P(O|\lambda)$ both for start codon and stop codon. After comparing these values we found that the similar result are also getting for the problem 2 for both start

and stop codons.

Results and Analysis

In this section, we discuss and analyze the result. At first, an input file as an anonymous large DNA sequence (here a stretch of that sequence has been shown in the input file) has been taken and an output file for the result to be displayed. In input file, we read character by character and by using string matching algorithm we find that the start and stop codons. Now, we fetch string of A, G, C and T that begin from start codons and terminate at end codons. This is the ORF (open reading frame) region that contains both exon and intron regions. In between two exon regions, there is an intron region found. It is starting from GT and ending to AG. We fetch this string and then the region from GT to AG of the sequence *i.e.* the intron region is removed and the rest of the sequence is concatenated. Thus output file contains all the exon regions.

Input File of observation Gene Sequence1:

O[T]=AGCATGTTGAGAAAGGCAAGAAGATTTTTA TTA TGAAGTGT TCCCAGTGCCACACCGTTGA AAAGGGAGGCAAGCACAAGACTG ACCAAAT CTCCATGATCTCTTTGCCGGAAGACAGGTCAG GCCCCTGATACTCTTACACAGCCGCCAATAAGAACA AAGGCATCATCTGACAGAGGATACACTGATGA GTATTTGGAGAATCCCAAGAAGTACATCCCTGGAA CAAAAATGATCTTTGTCGGCATTAAGAAGAA GGAAGAAAGGGCAGACTTAATAGCTTATCT CAAAAAAGCTACTAATGAGTAA

Output File : Figs. 4 and 5 shows the output file containing exon region and exon and intron region, respectively.

Initially, observation sequence probability has been calculated based on the start codons ATG and then for stop codons TAA/TAG/TGA individually. For this, we took an input file an observation sequence and initial probability. In this, we use forward procedure and the forward symbol ($\alpha_t(i)$) has been calculated and on the basis of previous forward symbol ($\alpha_t(i)$) we evaluate the next ($\alpha_{t+1}(i)$) value.

In this, we use forward procedure and the forward symbol ($\alpha_t(i)$) has been calculated and on the basis of previous forward symbol ($\alpha_t(i)$), we evaluate the next ($\alpha_{t+1}(i)$) value. We used recursion method for the calculation and the formula that is being used is given below:

(i) $\alpha_t(i) = \pi_i b_i(O_1)$ $\quad 1 \le i \le N$

Table 1(a) : State Transition Matrix for ATG.

a_{ij}=

	A	T	G	C
A	0.1	0.7	0.1	0.1
T	0.1	0.1	0.7	0.1
G	0.7	0.1	0.1	0.1
C	0.7	0.1	0.1	0.1

Table 1(b) : Symbol Emitting Matrix for ATG.

$b_j(k)$ =

	A	T	G	C
A	0.1	0.7	0.1	0.1
T	0.1	0.1	0.7	0.1
G	0.2	0.3	0.2	0.3
C	0.3	0.2	0.3	0.2

$$\alpha_{t+1}(i) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(O_{t+1}) \quad [1 \le t \le T\text{-}1,\ 2 \le j \le N]$$

(ii) Number of States (N): In gene sequence there are four states we have been taken: A, T, G and C

a) State Transition Matrix A=(a_{ij})

b) Observation Symbol Probability Matrix B = ($b_j(k)$)

c) Initial State Distribution:

π_i = [0.7, 0.1, 0.1, 0.1]

d) Output Matrix for maximum probability of observation sequence for ATG

0.33	0.9	0.33	0.33
0.33	0.33	0.9	0.33
0.7	0.1	0.1	0.1
0.7	0.1	0.1	0.1
0.1	0.7	0.1	0.1
0.1	0.1	0.7	0.1
0.2	0.3	0.2	0.3
0.3	0.2	0.3	0.2

Observation Sequence 2:

O[T]=ATGGAGAACCTGAAGTCTGGAGTGTATCCT CTCAAGGAAGCAAGTGGATGCCCTGGGGCTG ACAGGAATCTTCTGGTGTACTCTTTTTATGAAAAGG GGCCATTGACATTTAGGGATGTGGCCATAGA ATTTTCTCTGGAGGAGTGGCAATGCCTGGACACTGC TCAGCAGGATTTGTATAGAAAAGTGATGTTA GAGAACTACAGAAACCTGGTCTTCTTGGGTA TTGCTGTTTCTAAGCCAGACCTGATCACCTG TCTAGAGCAAGGAAAAGAGCCCTGGAATATG.

Count = 288

Initialization values

value of $P(O|\lambda)_1$ 0.090000

.

.

value of $P(O|\lambda)_{287}$ 3.919639e-129

value of $P(O|\lambda)_{288}$ 3.128430e-128

In this section, observation sequence probability has been calculated for the stop codons TAA/TAG/TGA. For this, we took an input file, an observation sequence and initial probability by applying forward algorithm. The forward symbol ($\alpha_t(i)$) has been calculated and based on previous forward symbol ($\alpha_t(i)$) the next value ($\alpha_{t+1}(i)$) was being evaluated. We use recursion method and using the same formula, that is the final n values of $P(O|\lambda)$.

(a) State Transition Matrix $A=(a_{ij})$

(b) Observation Symbol Probability Matrix $B=(b_j(k))$

Table 2(a) : State Transition Matrix for TAA/TAG/TGA.

a_{ij} =

	A	T	G	C
A	0.9	0.33	0.33	0.33
T	0.4	0.1	0.4	0.1
G	0.9	0.33	0.33	0.33
C	0.7	0.7	0.1	0.1

Table 2(b) : Symbol Emitting Matrix for TAA/TAG/TGA.

$b_j(k)$ =

	A	T	G	C
A	0.1	0.7	0.1	0.1
T	0.1	0.1	0.7	0.1
G	0.2	0.3	0.2	0.3
C	0.3	0.2	0.3	0.2

c) Intial State Distribution:

π_i =[0.33, 0.33, 0.33, 0.9]

d) Output Matrix for maximum probability of observation sequence for TAA/TAG/TGA.

0.9 0.33 0.33 0.33

0.4 0.1 0.4 0.1

0.9 0.33 0.33 0.33

0.1 0.7 0.1 0.1

0.1 0.7 0.1 0.1

0.1 0.1 0.7 0.1

0.2 0.3 0.2 0.3

0.3 0.2 0.3 0.2

Observation sequence 3:

O[T]=ATGGAGAACCTGAAGTCTGGAGTGTATCCTCT
CAAGGAAGCAAGTGGATGCCCTGGGGCTGA
CAGGAATCTTCTGGTGTACTCTTTTTATGAAAAGG
GGCCATTGACATTTAGGGATGTGGCCATAGA
ATTTTCTCTGGAGGAGTGGCAATGCCTGGACACTG
CTCAG CAGGATTTGTATAGAAAAGTGATGTTAG
A G A A C T A C A G A A A C C T G G T C T T C T T G
GGTATTGCTGTTTCTAAGCCAGACCTGATCA
CCTGTCTAGAGCAAGGAAAAGAGCCCTGGAATATG

count = 288

Initialization values

$P(O|\lambda)_1$ 0.033000

.

.

$P(O|\lambda)_{287}$ 1.437170e-130

$P(O|\lambda)_{288}$ 8.535825e-130

1. To find maximum probability for start codon (ATG)

Using observation sequence values of $P(O|\lambda)$ has been calculated and all the values of $P(O|\lambda)$ calculated. Next, maximum value of ATG found. Again like previous step, we took an input file, an output file, an observation sequence and initial probability, forward procedure has been applied.

a) State Transition Matrix $A=(a_{ij})$

b) Observation Symbol Probability Matrix $B=(b_j(k))$

Table 3(a) : State Transition Matrix for ATG.

a_{ij} =

	A	T	G	C
A	0.33	0.9	0.33	0.33
T	0.33	0.33	0.9	0.33
G	0.7	0.1	0.1	0.1
C	0.7	0.1	0.1	0.1

Table 3 (b) : Symbol Emitting Matrix for ATG.

$b_j(k)$ =

	A	T	G	C
A	0.1	0.7	0.1	0.1
T	0.1	0.1	0.7	0.1
G	0.2	0.3	0.2	0.3
C	0.3	0.2	0.3	0.2

c) Intial State Distribution

π_i = [0.9, 0.33, 0.33, 0.33]

d) Output Matrix for maximum probability of ATG

0.33	0.9	0.33	0.33
0.33	0.33	0.9	0.33
0.7	0.1	0.1	0.1
0.7	0.1	0.1	0.1
0.1	0.7	0.1	0.1
0.1	0.1	0.7	0.1
0.2	0.3	0.2	0.3
0.3	0.2	0.3	0.2

Observation sequence 4:

O[T]=ATGGAGAACCTGAAGTCTGGAGTGTAT CCTCTCAAGGAAGCAAGTGGATGCCCTGGGGCTGA CAGGAATCTTCTGGTGTACTCTTTTTATG AAAAGGGGCCATTGACATTTAGGGATGTGGCCA TAGAA TTTTCTCTGGAGGAGTGGCAATGCCT GGACACTGCTCAGCAGGATTTGTATAGA AAAGTGATGTTAGAGAACTACAGAAACC TGGTCTTCTTGGGTATTGCTGTTTCTAA GCCAGACCTGATCACCTGTCTAGA GCAAGGAAAAGAGCCCTGGAATATG

count = 288

T = 288

Length of observation sequence: 288

Value of $P(O|\lambda)_1$ 5.326183e-130

Value of $P(O|\lambda)_2$ 4.355309e-85

Value of $P(O|\lambda)_3$ 1.110692e-85

Value of $P(O|\lambda)_4$ 2.563205e-86

.

.

max value of P(O|λ) value of_initial_codons=5.0509e-84

2. To find maximum probability for stop codons (TAA/TAG/TGA)

In this again, we have taken observation sequence probability has been calculated according to the stop codons TAA/TAG/TGA. Here also, we took an input file, an output file, an observation sequence and initial probability, and all the values of P(O|λ) values been stored.

a) State Transition Matrix A=(a_{ij})

b) Observation Symbol Probability Matrix B = ($b_j(k)$)

Table 4(a) : State Transition Matrix for TAA/TAG/TGA.

a_{ij} =

	A	T	G	C
A	0.9	0.33	0.33	0.33
T	0.4	0.1	0.4	0.1
G	0.9	0.33	0.33	0.33
C	0.1	0.7	0.1	0.1

Table 4(b) : Symbol Emitting Matrix for TAA/TAG/TGA.

$b_j(k)$ =

	A	T	G	C
A	0.1	0.7	0.1	0.1
T	0.1	0.1	0.7	0.1
G	0.2	0.3	0.2	0.3
C	0.3	0.2	0.3	0.2

c) Intial State Distribution :

π_i = [0.33, 0.33, 0.33, 0.9]

Max value of P(O|λ) for stop codon=2.7877e-85

Conclusion

In this paper, we tried to find exon region in the large DNA sequence of eukaryotes using machine learning approach. So, we can use many machine learning approaches like HMM, ANN, SVM and GA. But in this we used only HMM to identify the exon region and removed the intron region. We use computational method for that and found the probability where the exon region must be present. Here, we have taken few DNA sequences, maximum length of each DNA sequence is around thousand characters. In future, this method of HMM is going to be used to find the maximum probability of exon region. Thus, we can identify as well as predict the exon region of any length for genome of any organism.

References

Abo-Zahhad, M., M. A. Ahmed and S. A. Abd-Elrahamn (2012). Genomic analysis and classification of exon and introns sequences usingDNA numericalmapping techniques. *International Journal of Information Technology and Computer Science*, **4(8)** : 22–36.

Ambikairajah, E., Epps J. and Akhtar M (2005). Gene and exon prediction using time-domain algorithms. *IEEE 8th Int. Symp. On Sig. Proc. and its Appl.*, pp. 199-202.

Abbasi, O., A. Rostami and G. Karimian (2011). Identification of exon regions in DNA sequences using cross-correlation and noise suppression by discrete wavelet transform. *BMC Bioinformatics*, **12** : article 430, 2011.

Chris Burge and Samuel Karlin (1997) Prediction of Complete Gene Structures in Human Genomic DNA. *J. Mol. Biol.*, 78-94.

Gilbert, Walter (1978). Why genes in pieces". *Nature*, **271 (5645)** : 501–501. PMID 622185. doi:10.1038/271501a0.

Guangchen Liu and Yihui Luan (2014). Identification of Protein Coding Regions in the Eukaryotic DNA Sequences Based on Marple Algorithm and Wavelet Packets Transform. Hindawi Publishing Corporation Abstract and Applied Analysis Volume 2014, Article ID 402567, 14 pages.

Hamidreza Saberkari (2013). A Novel Fast Algorithm for Exon Prediction in Eukaryotic Genes using Linear Predictive Coding Model and Goertzel Algorithm based on the Z-Curve" *International Journal of Computer Applications* (0975 – 8887), **67(17)** : April 2013.

Jean-Michel Claverie (1997). Computational methods for the identification of genes in vertebrate genomic sequences. *Human Molecular Genetics*, **6(10)** : 1735–1744.

Jonathan E. Allen and Steven L. Salzberg (2005). Integration of multiple sources of evidence for gene prediction. **21(18)** : 3596–3603.

Jonathan H. Badger and Gary J. Olsen (1999). CRITICA : Coding Region Identification Tool Invoking Comparative Analysis. *Molecular Biology and Evolution*, **16(4)** : 512-524.

Katherine, M. Keyes (2015). The mathematical limits of genetic prediction for complex chronic disease. *Journal of Epidemiol Community Health*.

Kevin L. Howe, Tom Chothia and Richard Durbin (2002). A Generic Framework for the Integration of Gene-Prediction Data by Dynamic Programming. pp 1418-1427.

Kwan, H. K., B. Y. M. Kwan and J. Y. Y. Kwan ((2012). Novel methodologies for spectral classification of exon and introns sequences. *Eurasip Journal on Advances in Signal Processing*, **2012(1)** : article 50, 2012.

Luciano Milanesi, Dino D Angelo and Lgor B. Rogozin (1999). GeneBuilder: Interactive in silico prediction of gene structure. pp 612-621.

Mario Stanke and Stephan Waack (2003). Gene prediction with a hidden Markov model and a new intron submodel. **19 (Suppl. 2)** : 215–225.

Marhon, S. A. and Kremer S. C. (2011). Gene prediction based on DNA spectral analysis: a literature review. *Journal of Computational Biology*, **18(4)** : 639–676.

Mena-Chalco, J., H. Carrer, Y. Zana and R. M. Cesar Jr. (2008). Identification of protein coding regions using the modified Gaborwavelet transform. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **5(2)** : 198–206.

Muneer Ahmad, Azween Abdullah and Khalid Buragga (2011). A Novel Optimized Approach for Gene Identification in DNA Sequences. *Journal of Applied Sciences*, **11(5)** : 806-814.

Rao, N. and S. J. Shepherd (2004). Detection of 3-periodicity for small genomic sequences based on AR technique. *International Conference on communications, Circuits and Systems*, ICCCAS, vol. 2, pp. 1032- 1036, June 2004.

Ravindra Nath and Renu Jain (2011). CpG Sequence Identification Using Hidden Markov Models(HMM) and Randomized Search Algorithms presented in the *International Conference on Computer Science and Information Technology* (ICCSIT, 2011) held from 6th Nov -7th Nov, 2011 at VITS, Delhi (achieved scholastic award).

Rabiner, L. R. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceeding of The IEEE*, **77(2)** : 257-268.

Su, Chun-Yi, Subhash Rakheja and Honghai Liu (2012). Intelligent Robotics and Applications: 5th International Conference, ICIRA 2012, Montreal, Canada, October 3-5, 2012, Proceedings, Part II. 10.1007/978-3-642-33515-0.

Suhartati Agoes (2011). A Hidden Markov Model for identification of exons in DNA of genes Plasmodium falciparum. *International Journal of Electric & Computer Sciences* IJECS-IJENS **11(01)** : 2011.

Snyder E. E. and G. D. Stormo (1995). Identification of protein coding regions in genomicDNA. *Journal of Molecular Biology*, **248(1)** : 1–18.

Sleator, R. D. (2010). An overview of the current status of eukaryote gene prediction strategies. *Gene*, **461(1-2)** : 1–4.

Vineet Bafna Daniel H. Huson (2000). The Conserved Exon Method for Gene Finding. pp 3-12.

Vaidyanathan P. P. and Yoon B. -J. (2002). Gene and exon prediction using allpass-based filters. In Proc. *IEEE GENSIPS* (Raleigh, NC, USA).

Xu, S., N. Rao, X. Chen and B. Zhou (2011). Inferring an organismspecific optimal threshold for predicting protein coding regions in eukaryotes based on a bootstrapping algorithm. *Biotechnology Letters*, **33(5)** : 889–896.