iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 **Issue:** XI **Month of publication:** November 2017

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor:6.887
Volume 5 Issue XI November 2017- Available at www.ijraset.com

Splice Site Recognition Using Lower Dimensional LHMM Features and SVM Classifier

Sejal Sahu[1], Sabyasachi Patra[2]
[1, 2] *Department of CSE, IIIT Bhubaneswar, Bhubaneswar, India*

Abstract: ***Recognition of coding region from DNA sequence has gained immense importance in the field of the research of gene identification. Splice sites which are the borders between exons and introns in DNA sequence are found in the eukaryotic organism. At present, there are several algorithms available for splice site recognition with an aim to improve the prediction accuracy. With an objective to further develop an efficient algorithm, Splice site recognition using lower dimensional Linear Hidden Markov Model (LHMM) features have been proposed in this paper. The proposed algorithm of Splice site recognition using lower dimension consists of three stages. Initial step use first order Markov Model (MM1) for feature extraction, in second stage dimension of feature vectors are reduced by using Principal Component Analysis (PCA) and, final or last stage use Support Vector Machine (SVM) with Gaussian kernel for classification. When the results of the proposed algorithm are compared with the existing algorithm of Splice site recognition, it has indicated remarkable performance and accuracy.***
Index Terms: place site, DNA sequence, HMM, SVM, PCA, dimension reduction, feature vector, classification

I. INTRODUCTION

A Bioinformatics is an emerging field which is strengthens by advancement of Computer Sciences, Information Technology, Mathematics, and Bio Technology to assimilate, analyze, evaluate and correlate various kinds of genetic information. In this context, Biological data mining plays an important role to supply data to overcome provoking challenges in the process of research and development, thus enabling various possibilities in this direction [1]. Although the field of bioinformatics is originally aimed to extract information embedded within the three billion bases of human DNA, the field has further evolved to understand its capability and capacity for studying information contents and information flow of biological systems and processes. At present, a huge volume of biological data is available and it grows exponentially.

A. This has precipitated into two problems:

1) Apt information storage and management and,

2) Extraction of advantageous information from these data.

The second problem is one of the major challenges in computational biology, which have necessitated the development of tools and techniques capable of transforming all these heterogeneous data into biological knowledge about the underlying mechanism. These tools and techniques should allow us to go beyond a very explanation of the data and provide knowledge in the form of testable models [2].

Eukaryotic gene classification is a miscellaneous process. It still seems a problematic process to predict the path of the basic fundamental biochemical reactions of gene expression: transcription, splicing, and translation from DNA sequence. One of the foremost and conclusive objectives of any genome sequencing project in bioinformatics is the identification and recognition of all genes, together with the corresponding proteins, their regulation, and functions. In view of the above, the prediction of genes has become one of the most important issues in computational biology.

All living organisms are made up of cells, which are classified as prokaryotes and eukaryotes. The prokaryote cell is simpler and smaller than eukaryote cell. The genomes of most eukaryotes are larger and more complex than those of prokaryotes. The splicing mechanism does not occur in prokaryotic cells. It has been found that the genomes of most eukaryotic cells contain not only functional genes but also large amounts of DNA sequences that do not code for proteins. The presence of large amounts of non-coding sequences is a general property of the genomes of complex eukaryotes. In eukaryotic cells, genes with the coding region called as exons are disrupted by non-coding regions called as an intron. The border moving from exons to introns is called donor site (or 50 boundary), and the border separating introns from exons is called acceptor site (or 30 boundary). Donor sites are described by the occurrence of the consensus dinucleotide GU in mRNA sequence or GT in the DNA sequence and acceptor sites are characterized by the consensus dinucleotide AG in going to 50 to 30 direction; but, the occurrence of such conserved dinucleotide does not develop a sufficient condition for splicing. Hence, the accurate prediction of splice sites is an important issue for eukaryotic

gene prediction, for which limited solutions are present.

The basic aim of this work is to develop a new efficient technique of splice site recognition, which would be capable of predicting the number of splice sites in a DNA sequence by using machine learning procedure. Presently, the approaches like GENSCAN, GENIO, GeneId3, HMMgene, VEIL, NNSplice, GENIO, NetGene2 etc. are well accepted for the purpose.

In order to improve the existing algorithm, a complete analysis of existing algorithms and the tools used in the process sequence data bank has been carried out. The performance of various algorithms is then, compared with existing splice site predictor. In addition to the obvious goal of improving predictive accuracy, multiple additional model properties are considered as mentioned in the following sections.

II. SPLICING

Splicing is the process which takes place in the mechanism of protein synthesis in a eukaryotic cell. In splicing, all introns which are the non-coding regions are removed and remaining exons i.e. coding regions are joined together. The sequence of exons is responsible for the formation of the protein and process can be defined as translation. Figure 1, explains that the gene is transcribed the whole thing becomes an RNA molecule, including the garbage (introns) in between the exons, and then these introns are cut out in a process called splicing. The resulting bits are joined together and translated into a protein.

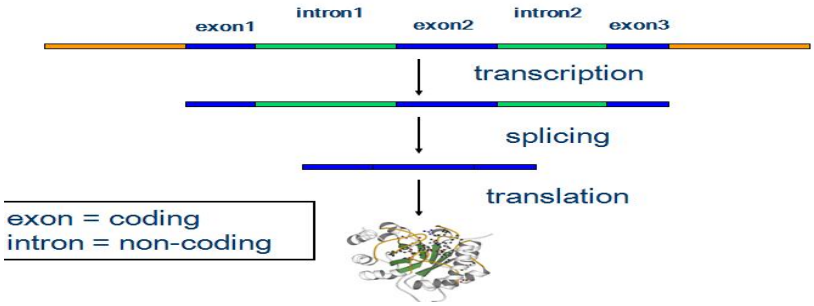


Fig. 1 Mechanism of splicing in eukaryotic

A. Splicing Consenses Sequencing

In nucleotide sequence, introns can be modified without causing any significant alteration in gene function, and the sample contains only highly conserved patterns, those are required for splicing. Basically, introns start with the dinucleotide GU (GT in the original DNA sequence) and end with the dinucleotide AG (in the direction 5' to 3'), and these signals are referred to as donor sites and acceptor sites, respectively. Therefore, intron borders are called splice sites. The occurrence of the preceding dinucleotides downstream and upstream is not sufficient to signal the presence of an intron. Another distinctly important conserved pattern is the branch site, with consensus sequence (C/T) N (C/T) (C/T) (A/G) A (C/T), where A is conserved in all genes, and located 20 - 50 bases upstream of the acceptor site There is also evidence for a pyrimidine-rich region preceding acceptor sites [8].

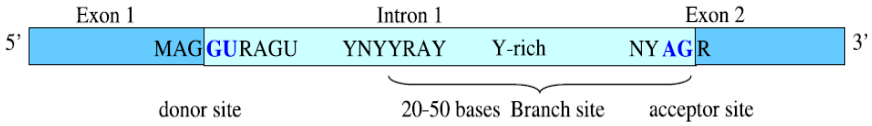


Fig. 2 Splice site consensus region

B. Exons and Introns

An exon is any part of a gene that will become a part of the final mature RNA produced by that gene after introns have been removed by RNA splicing. The term exon refers to both the DNA sequence within a gene and to the corresponding sequence in RNA transcripts. In RNA splicing, introns are removed and exons are covalently joined to one another as part of generating the mature messenger RNA.

The word intron is derived from the term intragenic region, that is, a region of a gene. The term intron refers to both the DNA sequence within a gene and the corresponding sequence in the unprocessed RNA transcript. As part of the RNA processing pathway, introns are removed by RNA splicing either shortly after or concurrent with transcription. Introns are found in the genes of most organisms and many viruses. They can be located in a wide range of genes, including those that generate proteins, ribosomal RNA (rRNA), and transfer RNA (tRNA).

III. LITERATURE SURVEY

This section describes some of the relevant literature of biological research in splice site recognition and its various traditional methods and tools which are available. Once the objectives of Human Genome project is cleared, researchers mainly concentrate their thinking on the vast amount of the biological data that are available and started exploring this data to solve many common problems related to a better understanding of how genes, proteins behave in environments. In the past few years, it has been noticed a high increase in the genomic sequence data for a broad range of organisms [3]. The explanation of data into useful knowledge is the crucial for future biological research and a great challenge as well. The double-helical structure of DNA was discovered by Watson and Crick in 1953 and within the precise period, researchers concluded a detailed understanding of the molecular methodology involved in gene replication and expression. Directly access to the sequence of gene became Possible in 1970's through the innovation of DNA sequencing and cloning [4]. A huge number of experiments has been done since last few years and several methods have been developed for recognizing protein-coding regions in DNA sequences. There are many existing algorithms are present and still to be developed. The basic focus of all gene recognition algorithm is to measure a 'typical ' exonic DNA which is responsible for the building of protein. Bioinformatics is the emerging field with the requirement of managing and extracting a huge amount of information which can be used in solving many common problems, especially related to drug design [4, 5]. Three research communities mainly Biologists, Mathematicians and Computer Scientists joined their hands to solve these interesting problems. Gene identification from large DNA sequence is known to be revelatory task . The main focus of human genome project was the identification of genes in eukaryotic genomes, but a still accurate number of genes in eukaryotic genomes are still unknown and research is going on. In prokaryotes, the computational gene prediction is relatively simple where all the genes are converted into the corresponding mRNA and then into proteins. The process is more complex for eukaryotic cells where the coding DNA sequence is interrupted by random sequences called introns. The mathematical approach in the segment of molecular biology and genomics is gaining a lot of attention and is an interesting research area for many scientists [6, 7, 8]. The methods for gene finding which are being used nowadays are more precise and reliable than the earlier tactics. Hidden Markov Models (HMM) have been applied successfully in various applications, viz. Speech recognitions [11]. 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 S, which is associated with a discrete output probability distribution, P(S). 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 to all other states must be 1. Markov and HMMs are gaining popularity in bioinformatics research for nucleotide sequence analysis [10, 12]. For prokaryotes gene identification, Borodovsky [15] effectively applied this HMM technique. Eukaryotic promoter detection algorithm using a Markov transition matrix was proposed by Audic and Claverie [16]. A new technique VEIL (Viterbi Exon-Intron Locator) was developed by Salzberg [17] and Henderson et al.[18] to identify translational start site and splice sites in eukaryotic mRNA. The HMM-based gene predictor Gene Scout was developed by Yin [19], to detect translational start site and mRNA splicing junction sites.

There are many splice site programs are available through which splice site prediction can be done. List of few programs is described in table 1. Support Vector Machines (SVMs) are the set of related supervised learning methods used for classification and regression . The SVMs have been developed by Vapnik [20] and gained popularity due to many promising features such as better empirical performance. The formulation uses the Structural Risk Minimization (SRM) principle, which has been shown to be superior [22] to traditional Empirical Risk Minimization (ERM) principle, used by conventional neural networks. SVMs were developed to solve the classification problem, but recently they have been extended to solve regression problems [23].

Some of the reasons for utilizing SVMs in Bioinformatics are these have a strong widespread application in machine learning for classification and, they can target relevant data positions automatically. Other applications of SVM in bioinformatics are the identification of human signal peptide cleavage sites, the secondary structure of protein and multi-class protein fold detection. Till date, the most popular techniques in use for splice site recognition are Markov models which need the labor-intensive selection of information resource; SVM, support vector kernels.

Program	Organism	Method
Gene Splicer	Arabidopsis, human	HMM + MDD
NETPLANTGENE	Arabidopsis	NN

(http://www.cbs.dtu.dk/services /NetPGene/)		
NETGENE2 (http://www.cbs.dtu.dk/services /NetGene2/)	Human, C.elegans, Arabidopsis	NN + HMM
SPLICEVIEW (http://l25.itba.mi.cnr.it/ webgene /wwwsplicevi ew.html)	Eukaryotes	Score with consensus
NNSPLICE0.9 (http://www.fruity.org/seqtools= splice.html)	Drosophila, human or other	NN
SPLICEPREDICTOR (http://bioinformatics.iastate.edu/ cgi-bin/sp.cgi)	Arabidopsis, maize	Logitlinear models : (i) score with consensus; (ii) local composition
BCM-SPL (http://www.softberry.com/berry .phtml; http://genomic.sanger.ac.uk/gf/gf. html)	Human, Drosophila, C.elegans, yeast, plant	Linear discriminant

Table 1

Theoretically, ideal gene predictor should have the ability to recognize the exact boundaries of all the attributes common to most eukaryotic protein-coding genes. The specific sequences which are there between the introns and exons can be identified by gene prediction algorithms.

IV. PROPOSED METHOD : SPLICE SITE RECOGNITION USING LOWER DIMENSIONAL LHMM FEATURES

Recognizing the presence of splice site within DNA sequence is the initial step in accurate prediction of gene structure. Biology researchers have extensively studied the laboratory procedures such as PCR on cDNA libraries etc. to identify the accurate gene structure. But, due to the presence of a large number of hidden genes, it is impossible to describe all of them by using experiments only in the lab. Hence, lab experiments are combined with bioinformatics approaches in the modern research. Various computational and bioinformatics approaches have been applied for splice site recognition which will be helpful in gene prediction.

In the process called splicing, the editing of nascent pre-messenger RNA (premRNA) transcript in which introns are removed and exons are joined together. Splicing is carried out in a series of reaction which is catalyzed by the spliceosome. Within the intron, an acceptor site (30 ends of the intron) and a donor site (50 ends of the intron) are essential for splicing. The splice donor site includes invariant sequence GT at the 50 end of the intron with a larger and less preserved region. The splice acceptor site at the 30 end of the intron terminates the intron with nearly invariant AG sequence (Figure 3). The recognition of these acceptor and donor sites (splice sites prediction) is a crucial step in the gene identification process.

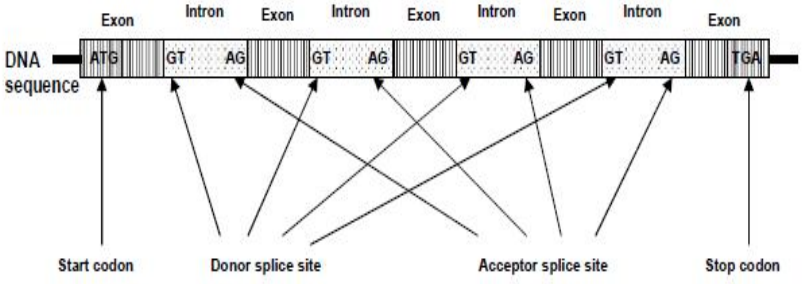


Fig. 3 The splice sites (Donor site and Acceptor site) in eukaryotic DNA sequence

A. Splicing Consenses Sequencing

We have conducted several simulations to evaluate the performance of the proposed algorithm using standard and publicly available splice site datasets.

The first dataset is known as NN269 [27], which consists of 1324 confirmed true acceptor sites, 1324 confirmed true donor sites, 5552 false acceptor sites and 4922 false donor sites collected from 269 human genes. Each of the pseudo acceptor/donor sites also has AG/GT in the splicing junction but is not a real splice site according to the annotation. The window size for an acceptor is 90 nucleotides -70 to +20 with consensus AG at positions -69 and -70. This includes the last 70 nucleotides of the intron and first 20 nucleotides of the succeeding exon. The donor splice sites have a window of 15 nucleotides -7 to +8 with consensus GT at positions +1 and +2. This includes the last 9 bases of the exon and first 6 bases of the succeeding intron. The dataset is available at [28]. This data set is split into a training set and a testing set. The training data set contains 1116 true acceptor, 1116 true donor, 4672 false acceptors, and 4140 false donor sites. The test data set contains 208 true acceptor sites, 208 true donor sites, 881false acceptor sites, and 782 false donor sites. Figure 2 and 3 show the two sample logo [28] of NN269 acceptor and donor sites. They represent the residues enriched and depleted in the sample. In NN269 acceptor dataset, AG is conserved in position 69 and 70 of the sequences, and for donor splice sites, GT is conserved in position 8 and 9 of the sequences.

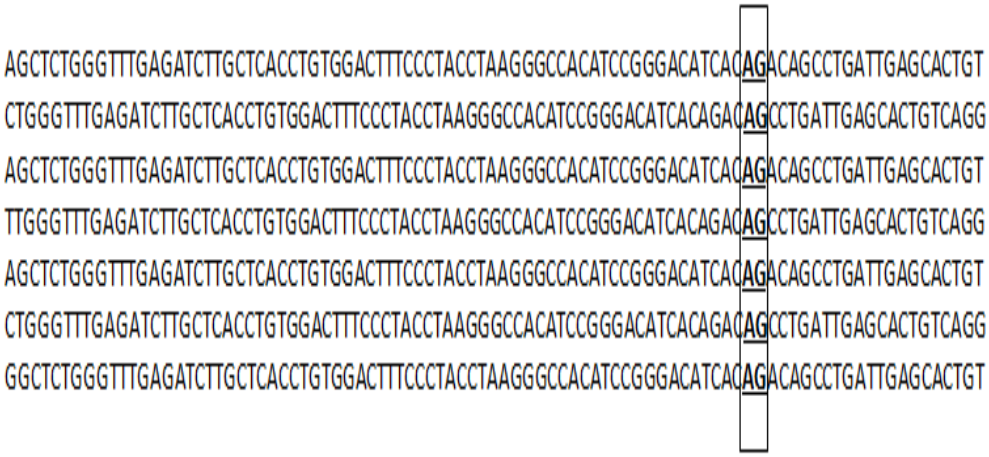


Fig. 4 Sample logo of NN269 acceptor splice site. The conserved dinucleotide AG are located in the position 69 and 70 in the seque nce.

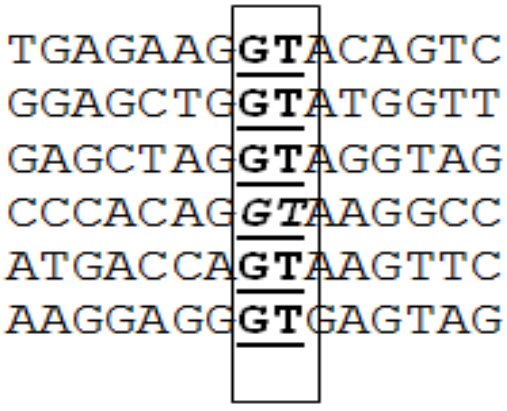


Fig. 5 Sample logo of NN269 acceptor splice site. The conserved dinucleotide GT are located in the position 8 and 9 in the sequence

B. Proposed Method

The basic processing steps are outlined in the following:

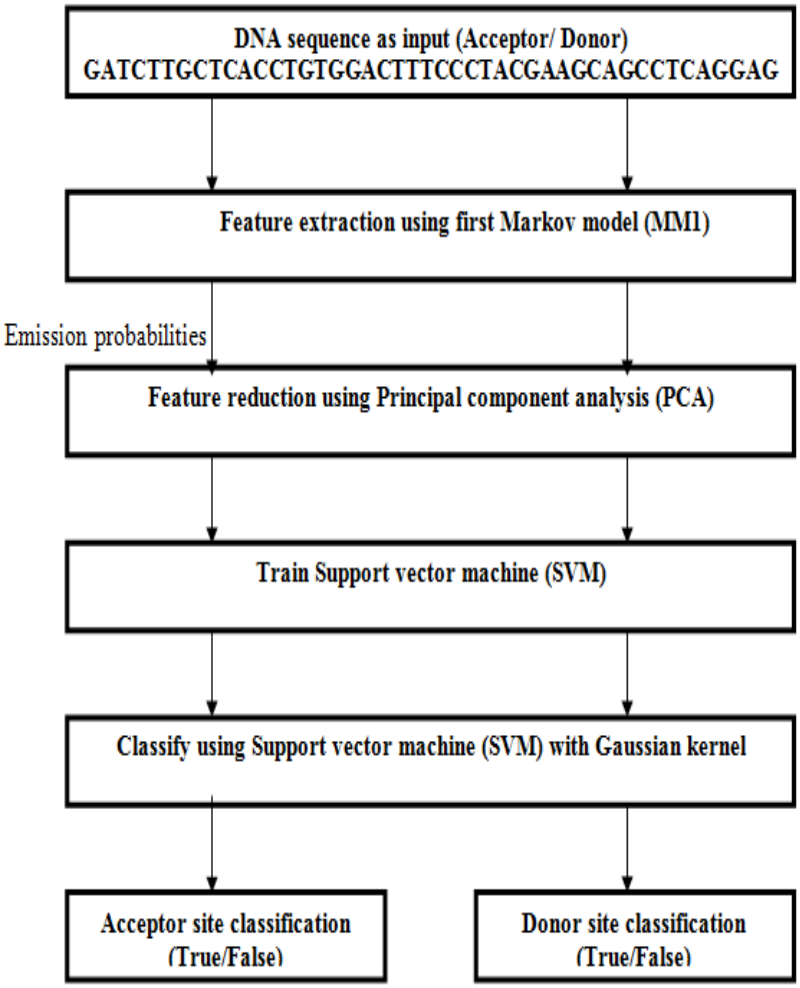


Fig. 6 Proposed model; The input DNA sequence is pre-processed by 1st order Markov model, PCA used for feature reduction. An SVM with Gaussian kernel function takes parameter as its input for Splice site recognition.

Our proposed model consists a number of separate modules and sub-modules that were predicted to capture properties of DNA and s pecially designed to recognize splice site. Splice site corresponds to the acceptor splice site and donor splice site, so splice site can b e divided into two classification modules i.e. acceptor splice site classification and donor splice site classification process. Further, f or the recognition of acceptor splice sites and donor splice sites, two different models are assembled which consist of three sub mod ules. The model includes several important steps, these are (1) appropriate features extraction scheme, (2) feature reduction method, and (3) classification using kernel.

C. Feature Extraction : Markov model pre-processing of splice site data

Markov chain is based on the principle of "memory lessens" property i.e. the next state of the process only depends on the previous state and not the sequence of states. In Markov model, DNA sequence has to decide the number of states and follow the specific type of characteristic called Markov property and the behavior of Markov chains are described by transition probability matrix. Each element of matrix defines the probability of transition of the matrix from one state to another. In Markov model, we required the set of sequences in which probabilities will be predictable. By using this technique, we can simply calculate the probability that the sequence has been produced in conformance with this model.

Every nucleotide in a DNA sequence corresponds to a state in the Markov chain used, where observed state variables are define from the alphabet $\sum$ DNA = {A, G, C, T}. Let us define the length of the sequence be l : {S1,S2,S3,....Sl}, where { Sj $\in$ {A, G, C, T}} , $\forall j$ = {1,2,...,l}, the nucleotide Sj is a consciousness of the jth state variable of Markov chain, and transition of state is only allowed from state j+1 i.e. its adjacent state. Therefore, Markov model is used to obtain ordered series of states. It derives from Sj to Sj+1 and extract symbol from the alphabet $\sum$ DNA, where each state is characterized by specific position of probabilistic parameter. In this proposed method, MM1 is used to define the probabilistic feature set of the given nucleotide sequences.

D. Feature Reduction : Using Principal Component Analysis (PCA)

The feature reduction techniques play a vital role in the field of bioinformatics such as splice site recognition, gene expression analysis to improve the performance in a faster manner and more cost effective performance, also to improve the understanding of the problem as well as a better result. The principal component analysis (PCA) is the well known tool for the feature reduction i.e. transforming the existing input feature into a new lower- dimension space. In PCA, the input feature space is transformed into a lower dimensional feature space using largest eigenvectors of the correlation matrix. Given a set of biological data, PCA finds the linear lower-dimensional representation of the biological data such that the variance of the reconstructed data is preserved. Using feature reduction which is based on PCA limits the feature vectors to the component selected by the PCA which leads to an efficient classification algorithm. So, the key point in our approach is to reduce the dimensionality of the extracted feature using Markov model to get better accuracy result.

E. Classification : Using Support Vector Machine (SVM)

The SVM is a machine learning algorithm which is introduced by Vapnik [41-44]. It is the supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. SVM uses hypothetical space of linear function in high dimensional feature space trained with a learning algorithm based on optimization theory. Find the dual formulation by using the method of Lagrange multipliers.

F. Model Design

The splice site recognition problem consists of two sub problems i.e. Acceptor splice site recognition problem and Donor splice-site recognition problem. We have created two separate models for the recognition of acceptor splice site and donor splice site. For analysis of the model, we have used nn269 dataset. Firstly, we created the model and trained the nn269 donor dataset into it. To compute the classification performance of this model, we have used the test nn269 donor data set. Similar steps have been followed for the acceptor nn269 dataset.

V. RESULTS AND DISCUSSION

A. Implementation Details

We have compared our proposed method splice site recognition using dimension reduction by LHMM with the existing method on the basis of the performance measure. The nn269 splice site dataset was used for the experiment. In existing method, used First Markov model MM1 as Feature extraction and support vector machine SVM with Gaussian kernel for classification. In proposed method, used MM1 as feature extraction, Principal Component analysis for feature reduction and support vector machine SVM with Gaussian kernel for classification.

B. Dataset Parameters

This section defines complete dataset parameter which is used in the experiment performed. Table 2 describes the input dimensions of the data set.

Total number of sequences	Number of True Donor/ Acceptor			Number of False Donor/ Acceptor		
	Sequences used for Training	Sequences used for Testing	Total	Sequences used for Training	Sequences used for Testing	Total
3126 Acceptor	836	279	1115	1541	513	2054
1919 Donor	597	198	795	843	281	1124

Table 2 nn269 dataset parameter used in

C. Evaluation Measure

The proposed method used classification performance is estimated on the basis of confusion matrix refer Table 3.

	Predictive Positive	Predictive Negative
Actual Positive	True Positive (TP)	True Negative (TN)
Actual Negative	False Positive (FP)	False Negative (FN)

Table 3 Confusion Matrix

where True positive(a) is the number of correct predictions that an instance is positive, False negative(d) is the number of incorrect predictions that an instance is negative, False positive(b) is the number of incorrect of predictions that an instance positive, and True negative(c) is the number of correct predictions that an instance is negative.

Several standard terms are defined by using 2 class matrixes:

Accuracy (AC) is the proportion of the total number of tested data that were predicted correctly.

$$AC=(a+d)/(a+b+c+d)$$

The recall or true positive rate (TP) or sensitivity (Sn) is the proportion of positive cases that were correctly identified.

$$TP=a/(a+b)$$

The false positive rate (FP) is the proportion of negatives cases that were incorrectly classified as positive.

$$FP=c/(c+d)$$

The true negative rate (TN) or specificity (Sp) is de fined as the proportion of negatives cases that were classified correctly.

$$TN=d/(c+d)$$

The false negative rate (FN) is the proportion of positives cases that were incorrectly classified as negative

$$FN=b/(a+b)$$

D. Results and Comparision

Refer Table 4 and Table 5, which shows the performance of the Acceptor and donor splice site by using HMM feature and LHMM feature.

Model	Number of Sequences	Number of Sequences Correctly Classified	Number of Sequences Incorrectly Classified
Acceptor using HMM features	792	547	245
Donor using HMM features	479	340	139
Acceptor using LHMM features	792	761	31
Donor using LHMM features	479	395	84

Table 4 Output of the Splice site recognition using HMM and LHMM feature on the basis of correctly and in correctly classified

Model	True Positive (TP)	True Negative (TN)	False Positive (FP)	False Negative (FN)
Acceptor using HMM features	149	513	130	115
Donor using HMM features	170	170	28	111
Acceptor using LHMM features	259	502	11	20
Donor using LHMM features	169	226	55	29

Table 5 Output of the Splice site recognition using HMM and LHMM feature on the basis of TP, TF, FP, FN

This section compares the evaluated output which is obtained by two different approaches of splice site recognition. In the first approach, HMM feature is used and classified by SVM with Gaussian kernel. In the second approach, LHMM features in which HMM features are reduced by PCA and then, classified by SVM with Gaussian Kernel. The comparison is done on the basis of accuracy, specificity and sensitivity.

Splice site	Accuracy	Sensitivity	Specificity
Acceptor Algorithm1	0.69066	0.534	0.758
Donor Algorithm1	0.7098	0.858	0.604
Acceptor Algorithm2	0.960	0.928	0.978
Donor Algorithm2	0.824	0.833	.800

Table 6 Comparison between the models using HMM (Algorithm1) features and LHMM (Algorithm2) features

VI. CONCLUSION

Recognition of DNA splice site sequences is an important issue in the field of biological information processing. In this paper, the core objective was focused on the techniques for splice site recognition. We have used techniques/approaches for splice site recognition, which is a vital part of gene prediction itself for identifying donor splice site and acceptor splice site. In section 4, have discussed a new approach for splice site recognition which uses HMM for feature extraction, PCA for feature reduction and SVM with Gaussian kernel for classification which relinquishes better results with accuracy 0:96% in acceptor model and 0:82% compared with another approach which uses HMM for feature extraction and SVM with Gaussian kernel for classification. Different approaches of feature extraction and reduction can be used to classify the splice site recognition. In literature survey, HMM is used for large dimension for better feature extraction and classification. By reducing the feature dimension can give better results which are proven in proposed experiment.

REFERENCES

[1] Larranaga, Pedro, Machine learning in bioinformatics, Briefings in bioinformatics 7.1 (2006): 86-112.

[2] Akay, Metin, Special issue on bioinformatics part i: advances and challenges, Proceedings of the IEEE 90.11 (2002): 1703-1704.

[3] S.A. Brenner, Genomics: The end of the beginning, Science vol. 287, no. 5461, pp. 2173-2174., 2000.

[4] S. Brenner, Sequences and consequences, Philosophical Transactions of the Royal Society B: Biological Sciences, vol. 365, no. 1537, pp. 207-212, 2010.

[5] H. S. Bodiwala, S. Sabde, P. Gupta, Design and synthesis of caffeoyl-anilides as portmanteau inhibitors of HIV-1 integrase and CCR5, Bioorganic and Medicinal Chemistry, vol. 19, no. 3, pp. 1256-1263, 2011.

[6] J. S. Toor, A. Sharma, R. Kumar, Prediction of drug-resistance in HIV-1 subtype C based on protease sequences from ART naive and first-line treatment failures in North India using genotypic and docking analysis, Antiviral Research, vol. 92, no. 2, pp. 213-218, 2011 .

[7] W. N. vanWieringen, D. Kun, R. Hampel, Survival prediction using gene expression data: A review and comparison, Computational Statistics and Data Analysis, vol. 53, no. 5, pp. 1590-1603, 2009.

[8] G. D. Stormo, Gene-finding approaches for eukaryotes, Genome Research, vol. 10, no. 4, pp. 394-397, 2000

[9] M. Q. Zhang, Computational prediction of eukaryotic protein-coding genes nature, Reviews Genetics, vol. 3, no. 9, pp. 698-709, 2002.

[10] P. Larranaga, B. Calvo, R. Santana et al., Machine learning in bioinformatics, Briefings in Bioinformatics, vol. 7, no. 1, pp. 86-112, 2006.

[11] H. L. Rajapakse, Markov encoding for detecting signals in genomic sequences, IEEE/ACM Trans Comput Biol Bioinform., vol. 2, no. 2, pp. 131-42, 2005.

[12] R. Lopez, F. Larsen, and H. Prydz, Evaluation of the exon predictions of the GRAIL software, Genomics, vol. 24, no. 1, pp. 133-136, 1994.

[13] L. R. Rabiner, Tutorial on hidden Markov models and selected applications in speech recognition, Proceedings of the IEEE, vol. 77, no. 2, pp. 257-286, 1989.

[14] A. V. Lukashin, and M. Borodovsky, GeneMark.hmm, New solutions for gene finding, Nucleic Acids Research, vol. 26, no. 4, pp. 1107-1115, 1998.

[15] M. Borodovsky, and J. McIninch, GENMARK, Parallel gene recognition for both DNA strands, Computers and Chemistry, vol. 17, no. 2, pp. 123-133, 1993.

[16] S. Audic, and J. M. Claverie, Detection of eukaryotic promoters using Markov transition matrices, Computers Chemistry, vol. 21, no. 4, pp. 223-227, 1997.

[17] S. I. Sachem, A method for identifying splice sites and translational start sites in eukaryotic mRNA, Computer Applications in the Biosciences, vol. 13, no. 4, pp. 365-376, 1997.

[18] S. S. Henderson, and K. H. Fasman, Finding genes in DNA with a Hidden Markov Model, Journal of Computational Biology, vol. 4, no. 2, pp. 127-141, 1997.

[19] M. M. Yin and J. T. L. Wang, GeneScout, A Data Mining System for Predicting Vertebrate Genes in Genomic DNA sequences, Information Sciences, Special Issue on Soft Computing Data Mining, vol. 163, no. 1-3, pp. 201-218, 2004

[20] V. Vapnik, The nature of statistical learning theory, Springer, 1995

[21] C. J. C. Burges, A tutorial on support vector machines for pattern recognition, Data Mining and Knowledge Discovery, vol. 2, pp. 121-167, 1998.

[22] J. A. K. Suykens, Support vector machines: A nonlinear modelling and control perspective, European Journal of Control, vol. 7, no. 2-3, pp. 311-327, 2001.

[23] S. Sonnenburg, New methods for detecting splice junction sites in DNA sequence, Master's Thesis, Humbold University: Germany, 2002.

[24] G. Ratsch, S. Sonnenburg, and C. Schafer, Learning interpretable SVMs for biological sequence classification, BMC Bioinformatics, vol. 7, no. Suppl 1: S9, 2006.

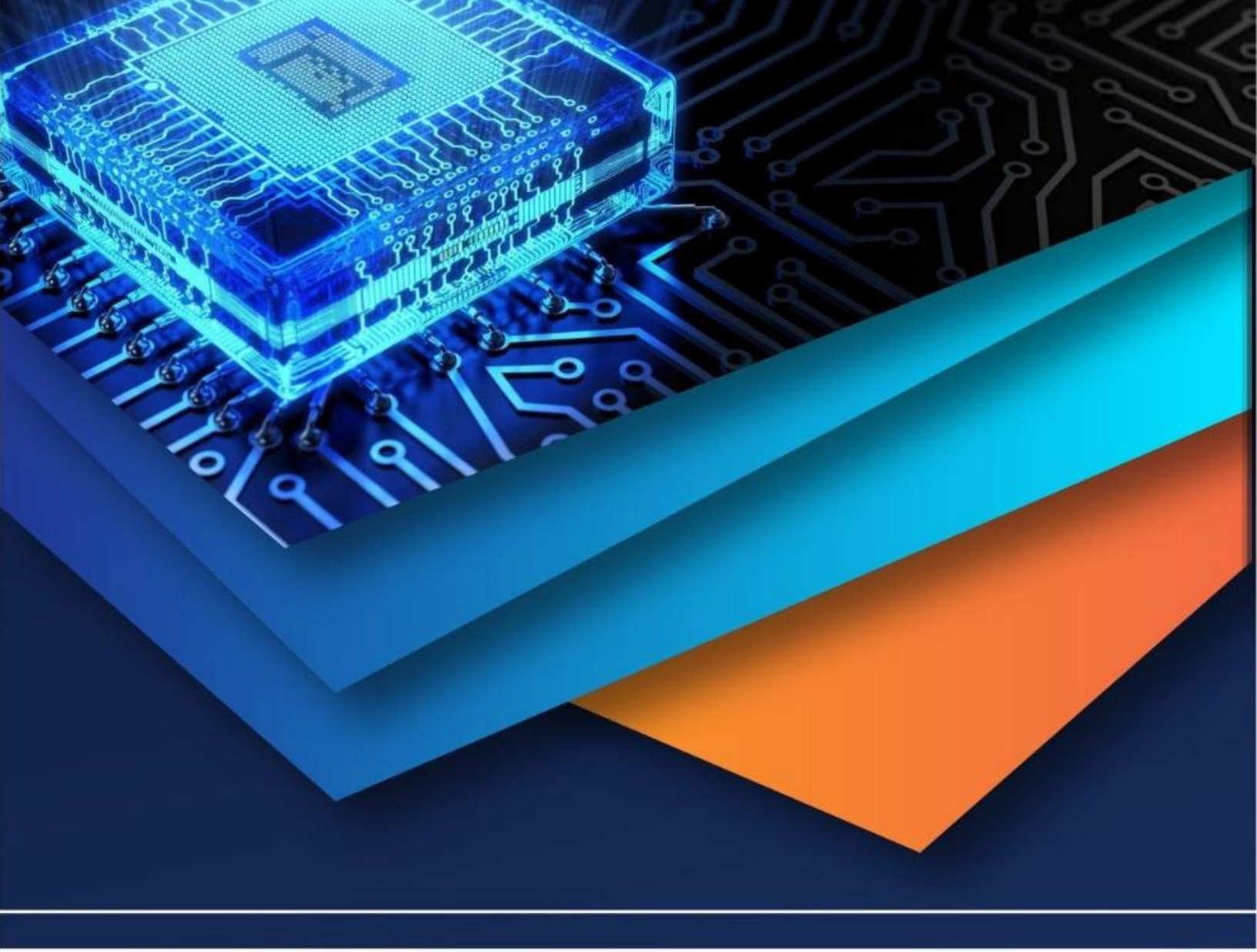
[25] Y.-F. Sun, X.-D. Fan, and Y.-D. Li, Identifying splicing sites in eukaryotic RNA: Support vector machine approach, Computers in Biology and Medicine, vol. 33, no. 1, pp. 17-29,2003.

[26] A. Zien, G. Ratsch, S. Mika, C. Lemmen B. Scholkopf, A.J. Smola, T. Lengauer, and K.-R. Muller, Engineering support vector machine kernel that recognize translation initiation sites in DNA, In Proceedings GCB'99, 1999.

[27] Reese MG, Eeckman F, Kupl D, Haussler D, Improved splice site detection in Genie, Journal of Computational Biology 1997,4(3):311-324.

[28] Vacic VILM, Radivojac P, Two Sample Logo: A Graphical Representation of the Differences between Two Sets of Sequence Alignments, Bioinformatics 2006, 22(12):1536-1537.

[29] Cristianini N, Shawe-Taylor J, An introduction to support vector machine and kernel based learning methods, Cambridge University press, Cambridge; 2000.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY