GJCRT : Global Journal of Computer Science Research & Technology
ISSN (online): 2349- 0608

Splice Site Prediction: A Review Of Progress In Hard Vs Soft Computing Techniques

Dr. Srabanti Maji(Bhunia)
Assistant Professor Dehradun Institute of Technology, Uttarakhand, India
Email: srabantiindia@gmail.com

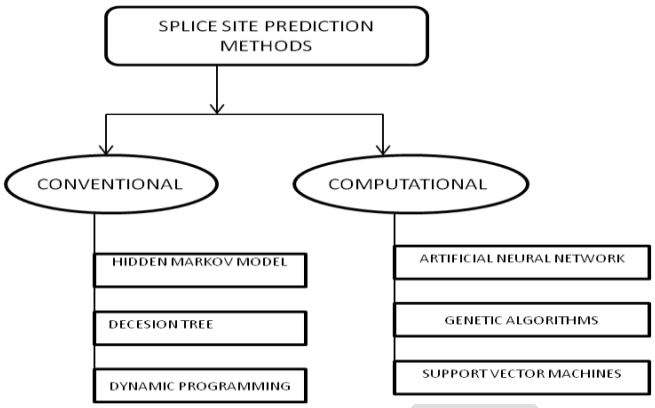
ABSTRACT: Bioinformatics refers to the study of script of life which begins with DNA. DNA holds genetic information and consists of thousands of genes. Genes are typically recognized by the process in which different signals and content recognizers are crafted that predict the likely locations of start, splice, and stop sites along with exonic and intronic regions etc. The performance of gene prediction program depends on accurate splice site prediction. Splice sites flank the boundaries of exons. A method of splice site detection must be based on thorough understanding of eukaryotic splicing process. Many techniques have been used for splice site detection. In this review, we present a general overview of few conventional and soft computing techniques for splice site prediction and implementation details of each method in DNA sequence.

KEYWORDS*:* ANN, DNA, DP, DT, GA , HMM, SVM.

INTRODUCTION

One of the major challenges in bioinformatics is to identify genes from DNA sequence (Baldi and Brunak, 1998; Campbell and Heyer, 2004). In such a scenario, accurate, appropriate and speedy tools to analyze these sequences are required. During the past decade various methods for gene prediction were evolved. Gene prediction method determines the likely locations of start/splice/ stop sites, and exonic/intronic regions etc. One of the important tasks of gene prediction is splice site detection. It is mandatory for splice prediction to distinguish the junctions as ‘intron-exon (IE)’ or ‘exon-intron (EI)’ in a given DNA sequence. Gene prediction methods are discussed in detail (Maji and Garg, 2013). There are various conventional/traditional techniques of splice site prediction like decision trees (Patterson et al., 2002), dynamic programming (Snyder and Stormo, 1993) and hidden Markov model (Ho and Rajapakse, 2003; Hu et al., 2000; Yin and Wang, 2001). The popularity of soft computing techniques has increased these days. Soft computing techniques are good for splice site prediction as they tackle partial truth, imprecision, and robustness. There are number of soft computing techniques like genetic algorithms (Awadalla et al., 2005), neural networks (Ogura et al., 1997), and SVM (Support Vector Machine) as shown in Figure 1.

Figure 1 List of various Conventional/Computational Splicing methods for gene prediction.



RELATED WORK

In 2001 GeneSplicer, a new flexible tool for detecting splice site was used to improve the accuracy as it combines markov modeling techniques with maximum dependence decomposition[MDD] (Pertea et al., 2001). Loi and Rajapakse in 2002 designed a novel machine learning approach for detecting splice site location in DNA sequence. This method was hybrid of neural networks and a Markov model where parameters of the Markov model were given to neural networks. The model is trained using a back propagation algorithm. Lorena et al. (2002) investigated the effects of noisy data on splice junction recognition and proposed a hybridized model which used decision trees and support vector machines. Followed by the hybrid approach by Loi and Rajapakse (2002), and Lorena et al. (2002), a new method called Increment Of Diversity With Quadratic Discriminant Analysis Method [IDQD] was proposed by Zhang and Luo (2003). This came into existence for studying the splicing sites and prediction of exons and introns. In human splice junction identification, Lorena and Carvalho (2003), introduced the concept of multiclass Support Vector Machines [SVM's] with bagging which employed two methods one called one-against-all and other is known as all-against-all. The method proposed earlier in 2002 is now modified by Lorena and Carvalho (2004), as in this work five techniques were studied. In preprocessing three techniques Edited Nearest Neighbor [ENN], repeated Edited Nearest Neighbor [RENN], and ALLkNN (All k Nearest Neighbor) were used. For learning Process, Decision Trees [DT] and SVM were used. During the same year, four other methods were proposed but with different techniques to enhance the splicing process. Firstly NetUTR (Eden and Brunak, 2004) method was presented which is based on neural network training scheme. Then in the second method, new technique based on feature ranking was derived to predict splice site with the help of the estimated distribution of the algorithm (Saeys et al., 2004a). This feature ranking is used to iteratively discard features. Third method introduced SpliceMachine (Degroeve et al., 2004), a new splice site prediction tool which was computationally fast. And the fourth one was based on Feature selection techniques (Saeys et al., 2004b) which are required to minimize the dimensions of dataset. It also enhanced the classification performance. A novel method based on hypernetwork architecture (Jose et al., 2005) was introduced which outperformed leading splice recognition systems for finding DNA splice sites. This method is known as HyperExon system. After HyperExon Sytem a new tool called SpliceScan was invented by Churbanov et al. (2006) which used Bayesian SS sensor based on oligonucleotide counting. Simultaneously, Baten et al. (2006) proposed a new method which works in two stages and in which the 1st order Markov model [MM1] was required for the prior stage followed by the second stage in which SVM having polynomial kernel was used. There were various formulas for implementing SVM so a new method called Support Vector Machines with weighted degree kernel were applied in accurate splice site detection by Sonnenburg et al. (2007). Till then an interactive approach was missing which is overcome by new web-based tool called

SplicePort (Dogan et al., 2007) that enabled interactive feature of browsing and visualization through which user can access a rich catalog of features for detecting splice sites. In search for superior performance than SVM, a novel approach based on discrete wavelet transform was developed by Liu et al. (2008), which used accurate splice sites identification algorithms. After this a hybrid algorithm relevant to state of the art combines several informative and effective input features along with SVM was designed by Baten et al. (2008). To trace out whether the mutation effects on splicing signals, a new Position Weight Matrices by Desmet et al. (2009) was introduced, and was also known as Human Splicing Finder (HSF). Human Splice Finder was based on information theory and was used for finding splicing motifs in any human DNA sequence. In HSF the nucleotides frequencies at each position was calculated using a weight matrix model. In 2010, nearly four more accurate and efficient methods were introduced. The first one was based on comprehensive information (Wang et al., 2010) for identifying splice sites, followed by the second method called length-variable Markov model (Zhang et al., 2010) and third one was related to association analysis (Kerdprasop and Kerdprasop, 2010). Association mining is unsupervised learning task that has been successfully applied to the marketing and business applications. The fourth one was proposed by Malousi et al. (2010) to combine discriminative computational with probabilistic modeling a hybrid method called Splice Identification Technique [SpliceIT]. A hybrid model for machine learning called Neural network tree [NNTree] by Hayashi and Zhao (2011) was used for structural learning. NNTree(s) were more accurate as compared to standard decision trees [DT's]. After this a novel classification model (Li et al.,2012) was constructed in which Homo sapiens splice site dataset was used and it also included SVM with the preselected features. The cross-validation accuracies with enhanced techniques were used for training sets with true and false splice junctions. A state-of-the-art, machine learning approach called averaged one-dependence estimators by Htike and Win (2013) was introduced to handle the problem of recognizing a vital class of genetic sequences known as eukaryotic splice sites. In the same year a Splice Site Prediction in DNA Sequences based on Probabilistic Neural Network was proposed by Nassa et al. (2013b). This method was used for detecting splice junctions by implementing generalized regression neural network. A new effective DNA encoding method for feature extraction which could provide more information of DNA sequence was proposed by Golam Bari et al. (2014). In this encoding method density information of each nucleotide along with its positional information and chemical property was provided. A transfer learning approach based upon K-means for splice junction recognition was proposed by Giannoulis et al. (2014). In this different representations for the secondary structure sequences, based on n-gram graphs was used. Followed by an efficient method, MM2F-SVM (Maji and Garg et al., 2014) was proposed which consists of three stages – First stage, consists of second order Markov Model [MM2] for feature extraction; the second stage proposed principal feature analysis [PFA], for feature selection; and the third stage uses support vector machine (SVM) with Gaussian kernel for final classification. Another predictor called "iSS-PseDNC" was also introduced for identifying splice junctions by Chen et al. (2014).The basic drawback of all these methods is that they are either working on 60bp (base pair) or 140bp and not exceeding that. The further scope of research in this direction is to increase the length of base pairs with less computational cost and time.

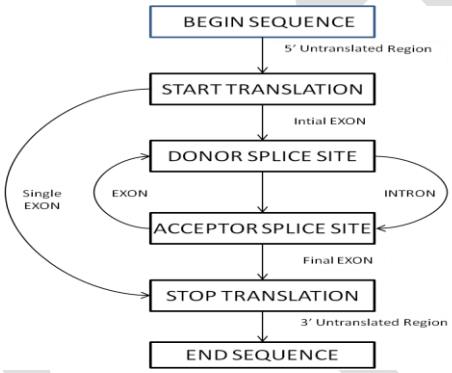
HARD / CONVENTIONAL COMPUTING TECHNIQUES

(a) Hidden Markov Model (HMM)

The statistical model known as Hidden Markov Model [HMM] has finite number of states. In HMM, probability distribution is associated with each state. A good HMM can accurately simulate the source environment and produce observed data for the real world. In this, first stage consists of hidden state in which state probabilities and transition probabilities are used and second stage produces emissions observable at every moment as suggested by Kouemou (2011). The strand of DNA which can be segmented into homogeneous regions is use to identify the specific functions of the DNA. By using HMM it can be determined that which parts of the strand belong to which segments by matching segments to hidden states. The HMM(s) were first discussed by Baum and Petrie (1966) and after that they were applied in a various fields such as speech

recognition (Roberts and Ephraim, 2000), hand-writing identification (Hu et al., 1996), signal processing (Thoraval et al., 1994), bioinformatics (Yada et al., 1999), climatology (Bellone et al., 2000), etc. Maximum biological data has variable length so HMM(s) are more successful because they can naturally accommodate uneven length models of sequence (Birney, 2001; Karplus et al., 2003). They are used for motif finding (Bucher et al., 1996), multiple sequence alignment (Sonnhammer et al., 1998) and identification of protein structure (Di Francesco et al., 1999). Few examples of the HMM based gene prediction tools are Genscan (Burge and Karlin, 1997), and HMMGene (Krogh, 1997). Gene prediction can be done through HMM (Maji and Garg, 2012). The HMMs are the easy to use because very small training sets are required for their implementations. The main drawbacks of HMM(s) is their greater computational cost.

Figure 2 Step by step Implementation of Hidden Markov Model [HMM] on DNA Sequence for Splice site Prediction.



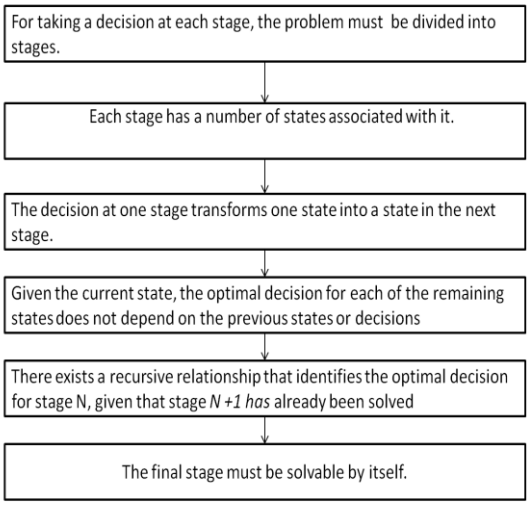
There are several states of HMM in gene identification such as exons, introns, promoter regions, intergenic regions, and 5' and 3' UTRs etc. Two types of probabilities are involved in HMM. First are the emission probabilities for generating observations and second are transition probabilities. Each state has transition probabilities to change or move to the next state. The transition probability to change an exon to an intron usually depends on the local DNA sequence such that it is high only at plausible splice sites. The outcome of HMM is a DNA sequence, which has visibility in any particular state (Maji and Garg, 2013). The term "hidden" in HMM(s) means DNA sequence is directly visible, and the state that is responsible for generating the DNA sequence like exon, intron, is not visible. But each state has different DNA characteristics. DNA emitted by exons must have certain codon bias, some length distributions and an Open Reading Frames [ORF]. Further DNA emitted by intron states has some distinct characteristics. The HMM model described in Figure 2 shows the gene structure for DNA sequence. Recently a new model which uses 3-D skeleton features obtained from an RGB-D, known as human activity detection model by Piyathilaka and Kodagoda (2013) is implemented using Gaussian Mixture Modal (GMM) based on HMM. Precise and multimodal design given by Hidden Markov Models [HMM's] which uses belief propagations is also proposed by Wong et al. (2013).

(b) Dynamic Programming (DP)

Dynamic programming technique is the most powerful method to solve a specific class of problems. Very simple and elegant formulation is required to solve the coding part. The concept is very easy, solve the problem on given input and then store the result for future use, so that same problem cannot be solved again and again. In short we say 'Remember your Past. The basic principle behind DP is to break the given problem into smaller problems (i.e. sub problems) and then after this the smaller problems are further divided in still smaller ones as shown in Figure 3. And after this break down if some over-lapping sub problems exist, then its surely a hint of

DP. If these sub problems are solved optimally then it will give optimal solution to the main problem (referred to as the Optimal Substructure Property).

Figure 3 Steps involved in dynamic programming for problem solving in splice site prediction used in gene identification.



To find the coding region and an optimal path among a list of weighted steps, the dynamic programming algorithm is used. To calculate the performance of given test sequences in all intervals and subintervals, a dynamic programming technique known as GeneParser (Snyder and Stormo , 1995) is used which require signal strength and coding measures to perform this task. Suitable combinations of exons and introns are predicted by DP. Gelfand and Roytberg (1993) reviewed the role of 'vector dynamic programming' for gene prediction which is implemented in CASSANDRA (Gelfand et al., 1996a). Claverie (1997) reviewed the concept dynamic programming for gene finding. To find protein-coding regions i.e. splice site in the DNA sequence GREAT (Gelfand et al., 1996b) program is used. The GenView system (Milanesi et al., 1993) is used to predict splicable Open Reading Frames [ORF] ranked by the strength. GRAIL (Xu et al., 1994c), GRAIL II (Xu et al., 1994a), GeneParser (Snyder and Stromo, 1995), FGENESH (Salamov and Solovyev, 2000), GAP III (Xu et al., 1994b) and recent versions of GeneId (Guigo et al., 1992) also use dynamic programming approach.

(c) *Decision trees (DT)*

A decision tree is a statistical probability model of decisions. It has decision rules in which one choice leads to next choice hence forming branches which are mutually exclusive. To reach a goal and achieve the target, a well established strategy is followed by decision trees. Prediction of splice sites and assignment of protein function can also be done through decision trees. To classify data items, a decision tree sets a list of queries about the features associated with the items. The node contains queries and each possible answer is stored at child nodes. Hence the hierarchy of the questions is encoded as a tree. In the easiest way, yes-or-no queries are asked, and each internal node contains either "yes' or 'no' child. The computational gene finders determine the exact structure of exon-intron in eukaryotic genes. Methods based alignment use similarity sequence whereas Ab-initio gene finders use inherent information. JIGSAW system (Allen et al., 2006) based on decision trees is an integrated method to trace genes and splice sites in the human genome. Decision trees are useful to predict a

strategy which will surely obtain an objective and it is especially used in decision analysis (Breiman et al., 1984). Decision trees segregate coding and non-coding DNA sequences which range from 54 to 162 bp (base pair) in length (Salzberg, 1995). An advantage of decision trees over other techniques is that they perform more functions of feature selection automatically. Decision tree algorithms and routines for splice site prediction are used by MORGAN system which is a tool to identify genes in the DNA sequences. Decision tree are successful methods used for Machine learning approaches (Che et al., 2011). Decision trees are used in one-button data mining approach (Salzberg et al., 1998) where no parameter tuning is needed.

SOFT / COMPUTATIONAL COMPUTING TECHNIQUES

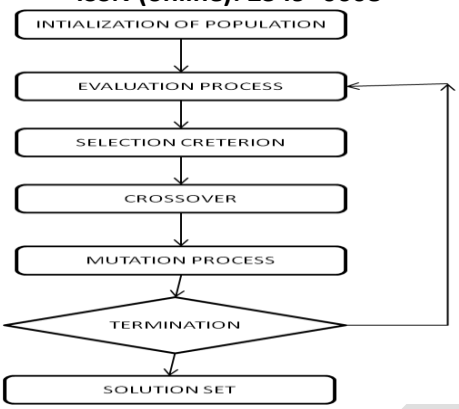
(a) Artificial Neural Network (Ann)

Artificial Neural Networks (ANNs) are widely used in real world applications due to its learning and generalization capabilities. It provides good results mainly where rule based systems could not be used or are much harder to develop. A vehicle control system (Souza et al., 2012) which is able to learn independently was presented. In Pessin et al. (2011) the use of ANN in mobile nodes localization using Wireless Networks was evaluated. Artificial Neural Networks are collection of neurons (units) linked by synapses (weighted links). The input units or hidden units receive signals from the environment and they do not have any contact with the external environment (Nolfi and Floreano, 2000) while the output units transfer signals to the environment. There are two partitions of ANN i.e. architecture and neurodynamics. The architecture includes the number of neurons and interconnection among them while the functional properties of the network are explained using Neurodynamics (Kartalopoulos, 1995). One of the main advantages of ANNs is that they are capable of learning and generalizing. In 1991, neural network was used for the first time in GRAIL (Gene Recognition And Analysis Internet Link) for gene identification. After that in 1996, an advance version of GRAIL came into existence which utilizes a multi agent neural network for tracing coding regions (Ying et al., 1996a). To enhance the prediction results this system uses error detection and correction algorithms.

(b) Genetic Algorithms (Ga)

GA is an evolutionary algorithm which searches for the optimal solution. It is known as a search and optimization technique. The main element in GA is a chromosome. A chromosome represents the best possible candidate solution. An encoding mechanism is used to prepare a chromosome. The collection of candidate solutions refers to the population which is chosen randomly at an initial stage. Then in GA, is uses fitness function to calculate the individual fitness that is to be considered in the population. The GA uses operators like: initialization, evaluations, selection, crossover, mutation, and termination which determine which individuals survive to the next generation.GA provides flexibility, modularity, easiness, efficiency, and robustness. The fundamental GA can be described (Gunnels et al., 1994; Mitchell, 1995) in a very simple way as follows in Figure 4.

Figure 4 ***Flowchart of various operators used in chromosome of DNA sequence using Genetic algorithm process for splice site prediction.***

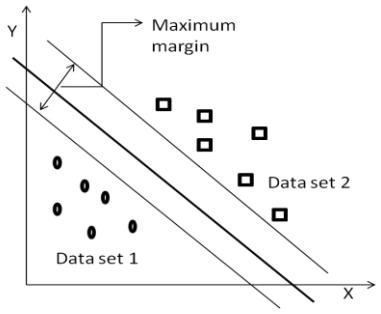


The process of natural Evolution (Chakraborty, 2010; Mitchell, 1998) is based on heuristic search algorithms called Genetic algorithms [GA's]. Genetic algorithm is used find an optimal solution to a problem as in operation research. For the first time genetic algorithm are used for gene prediction in 2011. Genetic algorithms are also used to predict an introns or exons (Perez-Rodriguez and Garcia-Pedrajas, 2011). To limit the search space, weight matrix method (WMM) is used. Positional weight matrix [PWM] and in-frame frequencies are used to calculate the fitness function. As the dataset is not balanced so accuracy cannot be used as the appropriate parameter to compute the performance of the desired system. Genetic algorithms are also used for optimizing the neural networks. A comparison of GA with Simulated Annealing (SA) is done by Gunnels et al. (1994) and states that the GA based method are good solutions for creating superior local maps which can further construct good global maps. *Drosophila melanogaster (*Victor and Alexey, 2003) uses a new approach to identify promoter regions. Dinucleotide frequencies are used to search an optimal division of a promoter region. Genetic algorithms are better when they are applied in very limited fashion.

(c) Support Vector Machine (Svm)

SVM's are considered as non-parametric classifiers which are used for structural risk minimization and statistical learning theory. Binary classification is done through SVM(s). Support Vector Machines are type of sets which uses supervised learning algorithms or in other words we can say that SVM's are learning methods. SVM(s) are used for classification of data as well as calculating the regression rate of data. The basic motive of SVM is to linearly separate hyperplane, which will further maximize the distance between the two classes, and it will act as a classifier as shown in Figure 5. If there exists some problem which cannot be solved by separating the hyperplane linearly then SVM provides two solutions for it. The first one is to introduce soft margin hyperplane which will add penalty function if there is any violation of constraints not leading to optimized results. Second one is to perform non-linear transformations on the original input data. This will lead to generate higher order dimension feature set from given input data set.

Figure 5 Basic SVM showing maximum margins from an optimal hyper plane to partition the datasets for classification .



Support vector machines are categories as linear and nonlinear SVM(s). Nonlinear SVM(s) are designed in such a way that they create a plane in a space and then maps it to the higher dimensional input space. The solution of quadratic programming problem is involved in SVM. The support vector machines [SVM] uses classification method for analyzing the quality of DNA data and then trace an optimal hyperplane which separates two different classes. In 1995, for the first time SVM was developed by Cortes and Vapnik, (1995). SVM is used in large number application fields like telecommunication, finance, etc. (Bhardwaj et al., 2005; Brown et al., 2000; Elish and Elish, 2007; Guyon et al., 2002; Ravi et al., 2008). Deterministic approach, strong mathematical ground, and easy of use are the important for the success of SVM (Hearst, 1998). Markov model can be combined with SVM (Maji and Garg et al., 2014) to get good success rates.

CONCLUSION

Detecting splice site is a tough job due to the presence of consenses di-nucleotide at the sites other than true splice sites. If splice prediction is accurate only then it is assumed that ab-intio gene prediction methods are accurate. The need of hour is to design a method that will accurately detect splice and reduces false positive rates. In this review we discussed conventional and computational methods. Each method has its corresponding advantages and disadvantages. The computational methods are more promising as they provide robustness and they can handle noisy and uncertain data easily. But soft computing techniques are complimentary and not competitive. So the best approach is to combine or hybridize methods to enhance the accuracy of outcome. Like neural network if combined with HMM will give better results for example HMMGene by (Krogh, 1997). The second challenge in splice site prediction is that the training sets used by various splice site prediction method consists of equal number of coding and non coding nucleotide but in DNA sequence intergenic/ non coding regions constitute of 95% of total DNA Sequence and only 2% portion is a coding part. This scenario of equal proportion may lead to biased system. Sensitivity will increase for coding data and classification of non coding data will decrease. This may lead to increase false positive count. So careful selection of training sets is required to achieve higher accuracy rates. There are various splicing tools are available as shown in Table 2.

REFERENCES:

[1] Al-Daoud, E. (2009) 'Identifying DNA splice sites using patterns statistical properties and fuzzy neural Networks', *EXCLI Journal*, Vol. 8, pp.1611-2156.

[2] Allen, J.E., Majoros, W.H., Pertea, M., and Salzberg, S.L. (2006) 'JIGSAW, GeneZilla, and Glimmer HMM: puzzling out the features of human genes in the ENCODE regions',*Genome Biol.*, Vol. 7, No.1, pp.1-13.

[3] Awadalla, S., Ortiz, J.E., and Gopal, S. (2005) 'Prediction of trans-splicing sites using genetic algorithms', *Res Comput Mol Biol*, pp.1-7.

[4] Baldi, P., and Brunak, S. (1998) *Bioinformatics: The Machine Learning Approach*, MIT Press, Cambridge, MA.

[5] Baten, A., Halgamuge, S.K. and Chang, B.C.H. (2008) 'Fast splice site detection using information content And feature reduction', *BMC Bioinformatics*, Vol. 9, No. 12, pp.1-12.

[6] Baten, A., Chang, B.C.H., Halgamuge, S.K., and Li, J. (2006) ' Splice site identification using probabilistic parameters and SVM classification', *BMC Bioinformatics*, Vol. 7, No. 5, pp.1-15.

[7] Baum, L., and Petrie, T. (1966) 'Statistical inference for probabilistic functions of finite state Markov chains' *The Annals of Mathematical Statistics*, Vol. 37, No. 6, pp.1554-1563.

[8] Bellone, E., Hughes, J.P., and Guttor, P. (2000) 'A hidden Markov model for downscaling synoptic atmospheric patterns to precipitation amounts', *Climate research*, Vol. 15, No. 1, pp.1-12.

[9] Bhardwaj, N., Langlois, R., Zhao, G., Lu, H. (2005) ' Kernel-based machine learning protocol for predicting DNA-binding proteins', *Nucleic Acids Res.* , Vol. 33, No. 20, pp.6486-6493.

[10] Birney, E. (2001) 'Hidden Markov Models in biological sequence analysis', *IBM J Res Dev,*Vol. 45, No. 3-4,
pp.449-454.

[11] Breiman, L., Friedman, J.H., Olshen, R.A., and Stone. C.J. (1984) *Classification and Regression Trees*, Chapman and Hall/CRC, Oxford University Press

[12] Brown, M., Grundy, W., Lin, D., Cristianini, N., Sugnet, C., Furey, T., Ares, M., Haussler, D. (2000) 'Knowledge based analysis of microarray gene expression data by using support vector machines', *Proc. Natl. Acad. Sci.USA*, Vol. 97, No. 1, pp. 262-267.

[13] Brunak, S., Engelbrecht, J., and Knudsen, S. (1991) 'Prediction of human mRNA donor and acceptor sites from the DNA sequence', *Journal of Mol. Biol*., pp. 220.

[14] Bucher, P., Karplus, K., Moeri, N., Hofmann, K. (1996) 'A flexible motif search technique based on generalized Profiles', *Comput Biol Chem.I*, Vol. 20, No. 1, pp.3-23.

[15] Burge, C., and Karlin, S. (1997) 'Prediction of complete gene structures in human genomic DNA', *J Mol Biol*, Vol. 268, No. 1, pp.78-94.

[16] Cai, T., and Peng, Q. (2005) 'Predicting splice sites in DNA sequences using neural network based on complementary encoding method', *In Proc. of International Conference on Neural Networks and Brain*, pp. 473-476.

[17] Campbell, A. M. and Heyer, L. J. (2004) *Discovering Genomics, Proteomics & Bioinformatics*, Pearson Education, Singapore.

[18] Chakraborty, R. C. (2010) *Soft computing-introduction* [online]. http://www.myreaders.info/html/soft computing.html.

[19] Che, D., Liu, Q., Rasheed, K., and Tao, X. (2011) 'Decision tree and ensemble learning algorithms with their
applications in bioinformatics', *Adv Exp Med Biol*., Vol. 696, pp.191-9.

[20] Chen,W., Feng, P.M., Lin, H. and Chou, K.C. (2014) 'iSS-PseDNC: Identifying Splicing Sites Using Pseudo Dinucleotide Composition', *BioMed Research International*, Vol. 2014 , pp.1-12.

[21] Churbanov, A., Rogozin, I.B., Deogun J.S. and Ali, H. (2006) ' Method of predicting Splice Sites based on signal interactions', *Biology Direct*, Vol. 1, No. 10, pp.1-20.

[22] Claverie, J.M.(1997) 'Computational methods for the identification of genes in vertebrate genomic sequences'
Hum Mol Gen., Vol. 6, No. 10, pp.1735-1744.

[23] Cortes, C., and Vapnik, V. (1995) 'Support-vector networks', *Mach. Learn*.,Vol. 20, No. 3, pp.273-297.

[24] Desmet, F.O., Hamroun, D., Lalande, M., Collod-Béroud, G., Claustres, M., Béroud, C. (2009) 'Human Splicing Finder: an online bioinformatics tool to predict splicing signals', *Nucleic Acids Res.*, Vol. 37, No. 9, pp.e67.

[25] Degroeve, S., Saeys, Y., Baets, B., Rouzé, P., and Peer, Y.V. (2005) 'SpliceMachine: predicting splice sites from high-dimensional local context representations', *Oxford Journals Science & Mathematics Bioinformatics* Vol. 21, No. 8, pp. 1332-1338.

[26] Di Francesco, V., Munson P.J., Garnier, J. (1999) 'FORESST: fold recognition from secondary structure predictions of proteins', *Bioinformatics*, Vol. 15, No. 2, pp.131-140.

[27] Dogan, R.I., Getoor, L.,Wilbur, W.J., and Stephen, M. Mount, S.M. (2007) 'SplicePort—An interactive splice-site analysis tool', *Nucleic Acids Res.*, Vol. 35, pp.W285–W291.

[28] Eden, E., and Brunak, S. (2004) 'Analysis and recognition of 5' UTR intron splice sites in human pre-mRNA',
Nucleic Acids Research , Vol. 32, No. 3, pp.1131-1142.

[29] Elish, K.O., and Elish, M.O. (2008) 'Predicting defect-prone software modules using support vector machines', *J. Syst. Softw*. , Vol. 81, No. 5, pp. 649-660.

[30] Gelfand, M.S., and Roytberg, M.A. (1993) 'Prediction of the exon-intron structure by a dynamic programming Approach', *Biosystems*, Vol. 30, No. 1-3, pp.173-182.

[31] Gelfand, M.S., Podolsky, L.I., Astakhova, T.V., Roytberg, M.A. (1996a) 'Recognition of genes in human DNA

Sequences', *J Comput Biol,* Vol. 3, No. 2, pp. 223-234.

[32] Gelfand, M.S., Astakhova, T.V., Roytberg, M.A. (1996b) 'An algorithm for highly specific recognition of protein-coding regions', *Genome Inform.,* Vol. 7, pp.82-87.

[33] Giannoulis, G., Krithara, A., Karatsalos, C., Paliouras, G. (2014) 'Splice Site Recognition Using Transfer Learning', *Artificial Intelligence: Methods and Applications,* Vol. 8445, pp. 341-353.

[34] Golam Bari, T. M., Rokeya Reaz, M. and Jeong, B.S. (2014) 'Effective DNA Encoding for Splice Site Prediction Using SVM', *MATCH Commun. Math. Comput. Chem.,* Vol. 71, pp.241-258.

[35]Guigo, R., Knudsen, S., Drake, N., and Smith, T.(1992) 'Prediction of gene structure', *J Mol Biol,* Vol. 226, No. 1, pp.141-157.

[36] Gunnels, J.A., Cull, P., and Holloway, J. (1994) 'Genetic Algorithms and Simulated Annealing for Gene Mapping', *Proceedings of 1st IEEE on Evolutionary Computation,* pp. 385-390.

[37] Guyon, I., Weston, J., Barnhill, S., Vapnik, V. (2002) 'Gene selection for cancer classification using support vector machines', *Mach. Learn.* , Vol. 46, No. (1-3), pp.389-422.

[38] Hatziglorgiou, Mache, N., Reczko, M. (1996) 'Functional Site Prediction on the DNA sequence by Artificial Neural Networks', *IEEE International Joint Symposia on Intelligence and Systems,* pp.12-17.

[39] Hayashi, H., and Zhao, Q. (2011) ' Inducing Compact Nntrees Through Discriminant Multiple Centroid Based Dimensionality Reduction', *International Journal of Innovative Computing, Information and Control,* Vol. 7, No. 5(B), pp. 2971-2985.

[40] Hearst, M. (1998) 'Support vector machines', *IEEE Intell. Syst. Appl.,* Vol.13, No. 4, pp.18-21.

[41] Hebsgaard, S.M., Korning, P.G., Tolstrup, N., Engelbrecht, J., Rouzé, P., and Brunak, S. (1996) 'Splice site prediction in Arabidopsis thaliana pre-mRNA by combining local and global sequence Information', *Nucleic Acids Research,* Vol. 24, No. 17, pp. 3439–3452.

[42] Hi, H.S., and Rajapakse, J.C. (2002) 'Splice site detection with neural networks/Markov models hybrids', *Proceedings of the 9th International Conference on Neural Information Processing (ICONIP'OZ)* , Vol. 5, pp. 2249 – 2253.

[43] Ho, L.S., and Rajapakse, J.C. (2003) 'Splice site detection with higher-order Markov model implemented a neural networks', *Genome Inform,* Vol. 14, pp.64–72.

[44] Htike, Z.Z., and Win, S.L. (2013) 'Classification of Eukaryotic Splice-junction Genetic Sequences Using Averaged One-dependence Estimators with Subsumption Resolution', *Procedia Computer Science 4th International Conference on Computational Systems-Biology and Bioinformatics,* Vol. 23, pp.36–43.

[45] Hu, J., Brown, M., and Turin, W. (1996) 'HMM Based On-Line Handwriting Recognition*', IEEE transactions on pattern analysis and machin*e, Vol. 18, No.10, pp.1039-1045.

[46] Hu, M., Ingram, C., Sirski, M., Pal, C., Swamy, S., and Patten, C. (2000) 'A Hierarchical HMM implementation for Vertebrate Gene Splice Site Prediction', *Technical report, Deptt.Computer Science, Univ. Waterloo.*

[47] Johansen, O., Ryen, T., Eftesol, T., Kjosmoen, T., and Ruoff, P., (2009) 'Splice Site Prediction Using Artificial Neural Networks', *Springer-Verlag Berlin Heidelberg CIBB* , pp. 102–113.

[48] Jose, L., Juareza, S., Colombanob, S., and Kirschnera, D. (2007) 'Identifying DNA splice sites using hypernetworks with artificial molecular evolution', *Biosystems,* Vol. 87, No. 2-3, pp.117-24.

[49] Karplus, K., Karchin, R., and Draper, J., (2003) ' Combining local-structure, fold-recognition, and new fold methods for protein structure prediction', *Proteins,* Vol. 53,No. S6, pp.491-496.

[50] Kartalopoulos, S.V. (1995) *Understanding Neural Networks and Fuzzy Logic: Basic Concepts and Applications.* Wiley- IEEE Press.

[51] Kel, A., Ptitsyn, A., Babenko, V., Meier-Ewert, S., and Lehrach, H. (1998) 'A genetic algorithm for designing gene family-specific oligonucleotide sets used for hybridization: the G protein-coupled receptor protein superfamily ', *Bioinformatics,* Vol. 14, No. 3, pp.259-270.

[52] Kerdprasop, N., and Kerdprasop, K. (2010) 'A High Recall DNA Splice Site Prediction Based on Association Analysis' , *ACS'10 Proceedings of the 10th WSEAS international conference on Applied computer science,* pp. 484-489

[53] Kouemou, G.L. (2011) *History and Theoretical basic of Hidden Markov Models In Hidden Markov Models, Theory and Applications,* Przemyslaw Dymarski (Ed). Published CC BY-NC-SA 3.0 Licence

[54] Krogh A. (1997) 'Two methods for improving performance of an HMM and their application for gene finding',
Proc Intl Conf Intell Syst Mol Biol, Vol. 5, pp.179-186.

[55] Li, J.L., Wang, L.F., Wang, H.Y., Bai, L.Y., Yuan, Z.M. (2012) 'High-accuracy splice site prediction based on sequence component and position features', *Genetics and Molecular Research,* Vol. 11, No. 3, pp. 3432-3451

[56] Liu Q., Wan, S. W., and Sun, Y.F. (2008) 'Identification of Splice Sites Based on Discrete Wavelet Transform and Support Vector Machine', *Bioinformatics and Biomedical Engineering, ICBBE 2008. The 2nd International Conference*, pp.54-57.

[57] Loi, H.S., and Rajapakse, J.C. (2002) ' Splice site detection with neural networks/Markov model hybrids Neural' *Proceedings of the 9th International Conference on Information Processing,* Vol. 5, pp. 2249 – 2253.

[58] Lorena, A.C., and Carvalho, A.C.P.L.F. (2004) 'Evaluation of noise reduction techniques in the splice junction recognition problem', Genetics and Molecular biology', Vol. 27, No. 4, pp.665-672.

[59]Lorena, A.C. and Carvalho, A.C.P.L.F. (2003) 'Human Splice Site Identification with Multiclass Support Vector Machines and Bagging ', *Artificial Neural Networks and Neural Information Processing ICANN/ICONIP,* Vol. 2714, pp.234-241.

[60]Lorena, A.C., Batista, G.E.A.P.A., Carvalho, A.C.P.L.F., and Monard, M. C. (2002) ' Splice Junction Recognition using Machine Learning Techniques.' *In Brazilian Workshop on Bioinformatics, 2002, Gramado, RS, Brazil,* pp.32-39.

[61] Maji, P., and Sushmita, P. (2014) 'Neural network tree for identification of splice junction and protein coding region in DNA. In: Scalable pattern recognition algorithms', *Springer International Publishing, Switzerland,* pp. 45-66.

[62] Maji, S. and Garg, D. (2013) 'Progress in gene prediction: principles and challenges', *Current Bioinformatics,* Vol. 8, No. 2, pp.226-243.

[63] Maji, S., and Garg, D. (2012) 'Gene finding using Hidden Markov Model', *Journal of Applied Sciences,* Vol. 12, No. 15, pp.1518-1525.

[64] Maji, S., and Garg, D. (2013) 'Hidden markov model for splicing junction sites identification in DNA sequences', *Current Bioinformatics,* Vol. 8, No. 3, pp.369-379.

[65] Maji, S. and Garg, D. (2014) 'Hybrid Approach using SVM and MM2 in Splice Site Junction Identification', *Current Bioinformatics,* Vol. 9, No. 1, pp.76-85.

[66] Malousi, A., Chouvarda, I., Koutkias, V., Kouidou, S., and Maglaveras, N. (2010) 'SpliceIT: A hybrid method for splice signal identification based on probabilistic and biological inference', Journal of biomedical Informatics, Vol. 43, No. 2, pp.208-217.

[67] Mitchell, M. (1998) *An Introduction to Genetic Algorithm,* MIT Press, One Rogers Street Cambridge ,USA.

[68] Mitchell M. (1995) 'Genetic Algorithms: An Overview', *Complexity,* Vol. 1, No. 1, pp.31-39.

[69] Moghimi, F., Shalmani, M.T.M., Sedigh, A.K., Kia, M. (2012) 'Two new methods for DNA splice site prediction based on neuro-fuzzy network and clustering', *Neural Comput & Applic, Springer-Verlag London,* Vol. 23, No. 1, pp.407-414.

[70] Milanesi, L., Kolchanov, N.A., Rogozin, and I.B.,(1993) ' GenView: A computing tool for protein-coding regions prediction in nucleotide sequences', *Proceedings of the 2nd International Conference onBioinformatics, Supercomputing, and Complex Genome Analysis*, pp. 573-588.

[71] Nassa, T., Singh, S. and Goel, N. (2013a) ' Neural Network Based Systems for Splice Site Detection: A Review', *International Journal of Advanced Research in Computer Science and Software Engineering,* Vol. 3, No. 7, pp.604-608.

[72] Nassa, T., Singh, S. and Goel, N. (2013b) 'Article: Splice Site Detection in DNA Sequences using Probabilistic Neural Network', *International Journal of Computer Applications,* Vol. 76, No. 4, pp.1-4.

[73] Nolfi, S., and Floreano, D. (2000) *Evolutionary Robotics: The Biology, Intelligence,and Technology of Self-Organizing Machines,* The MIT Press. One Rogers Street Cambridge ,USA.

[74] Ogura, H., Agata, H., Xie, M., (1997) 'A study of learning splice sites of DNA sequence by neural Networks', *Comput Biol Med*, Vol. 27, No.1, pp.67–75.

[75] Palaniappan, K., and Mukherjee, S. (2011) 'Predicting essential genes across microbial genomes: a machine l

learning approach', *In Proceedings of the IEEE International Conference on Machine Learning and Applications*, pp. 189–194.

[76] Patterson, D.J., Yasuhara,K. and Ruzzo,W.L. (2002) 'Pre-mRNA secondary structure prediction aids splice site prediction', *Proceedings of the Pacific Symposium on Biocomputing. Lihue, Hawaii, World Scientific Press*, pp. 223–234.

[77] Perez-Rodriguez, J., and Garcia-Pedrajas, N. 'An evolutionary algorithm for gene structure prediction', *Industrial Engineering and Other Applications of Applied Intelligent Systems II*, Vol.6704, pp. 386–395.

[78] Pertea, M., Lin, X., and Salzberg, S.L. (2001) 'GeneSplicer: a new computational method for splice site prediction', *Nucleic Acids Res.* , Vol. 29, No. 5, pp.1185–1190.

[79] Pessin, G. , Os´orio, F. S. , Ueyama, J., Souza, J. R., Wolf, D. F. , Braun, T.,and Vargas, P. A. , (2011) 'Evaluating the impact of the number of access points in mobile robots localization using artificial neural networks', *Proceedings of the 5th International Conference on Communication System Software and Middleware , COMSWARE Verona, Italy*.

[80] Piyathilaka, L., and Kodagoda, S. (2013) 'Gaussian mixture based HMM for human daily activity recognition using 3D skeleton features', *Industrial Electronics and Applications(ICIEA), 8th IEEE Conference on* , Vol. 19, No. 21, pp.567,572.

[81] Ravi, V., Kurniawan, H., Thai, P.N.K., Kumar, P.R. (2008) 'Soft computing system for bank performance Prediction', *Appl. Soft Comput.*, Vol. 8, No. 1, pp.305-315.

[82] Rebello, S., Maheshwari, U., Safreena, and Dsouza, R. V. (2011) 'Back propagation neural network method for predicting lac gene structure in streptococcus pyogenes M group A streptococcus strains', *International Journal for Biotechnology and Molecular Biology Research*, Vol. 2, pp. 61–72.

[83] Reese, M.G., Eeckman, F.H., Kulp, D., and Haussler, D. (1997) 'Improved splice site detection in Genie', *First Annual International Conference on Computational Molecular Biology (RECOMB)*, New York, ACM Press, pp. 232-240

[84] Roberts, W.J.J., and Ephraim, Y., (2000) 'Hidden Markov modeling of speech using Toeplitz covariance matrices', *Speech communication*, Vol. 31, No. 1, pp.1-14.

Ryden, T., Terasvirta, T. , and Asbrink, S. (1998) 'Stylized Facts of Daily Return Series and the Hidden Markov Model', *Journal of applied econometrics,* Vol. 13, No. 3, pp.217.

[85] Saeys,Y., Degroeve, S., Aeyels, D., Rouzé, P., and Peer, Y.V.(2004a) ' Feature selection for splice site Prediction : A new method using EDA-based feature ranking', *BMC Bioinformatics, BioMed Central Ltd* .,Vol. 5, No.1, p.64.

[86] Saeys, Y.,Sven Degroeve, S., and Peer, Y.V. (2004b) 'Digging into acceptor splice site prediction: an iterative feature selection approach', *Knowledge Discovery in Databases: PKDD 2004, Publisher: Springer Berlin Heidelberg,* pp.386-397.

[87] Salamov, A.A., and Solovyev, V.V. (2000) 'Ab initio gene finding in Drosophila genomic DNA', *Genome Res*., Vol. 10, No. 4, pp.516-522.

[88] Salzberg, S.(1995) ' Locating protein coding region in human DNA using a decision tree algorithm', *J Comput Biol*; Vol. 2, No. 3, pp.473-485.

[89] Salzberg, S., Delcher, A.L., Fasman, K.H., and Henderson J. (1998) 'A decision tree system for finding genes in DNA', *J Comput Biol.,* Vol. 5, No. 4, pp.667- 680.

[90] Snyder, E.E., and Stormo, G.D. (1993) 'Identification of coding regions in genomic DNA sequences: an application of dynamic programming and neural networks', *Nucleic Acids Research* ,Vol. 21, No. 3., pp.607-613.

[90] Snyder, E.E., and Stormo, G.D, (1995) 'Identification of protein coding regions in genomic DNA', *J Mol Biol*, Vol. 248, No. 1, pp.1-18.

[91] Sonnenburg, S., Schweikert, G., Philips, P., Behr, J., and Rätsch, G., (2007) 'Accurate splice site prediction using support vector machines', *BMC Bioinformatics,* Vol. 8, No. 10.

[92] Sonnhammer, E.L.L., Eddy, S.R., Birney, E., Bateman, A., and Durbin, R. (1998) ' Pfam: Multiple sequence alignments and HMM-profiles of protein domains', *Nucleic Acids Res,* Vol. 26, No. 1, pp.320-322.

[93] Souza, J.R., Pessin, G., Shinzato, P.Y., Osorio, F.S., and Wolf, D. F. (2011) 'Vision-based waypoint following using templates and artificial neural networks', *Timely Neural Networks Applications in Engineering Neurocomputing (Elsevier),* Vol. 107, pp.77-86.

[94] Stiglic,G., Kocbek , S., Pernek, I. Kokol , P. (2012) 'Comprehensive Decision Tree Models in Bioinformatics', *American University in Cairo, Egypt*. DOI: 10.1371/journal.pone.0033812

[95] Thoraval, L., Carrault, and G., Bellanger, J. (1994) 'Heart Signal Recognition by Hidden Markov Models: The ECG Case', *Methods of information in medicine. Method in Archive* , Vol. 33, No. 1, pp.10-14.

[96]Tolstrup, N., Rouze, P., and Brunak, S. (1997) 'A branch point consensus from Arabidopsis found by non-circular analysis allows for better prediction of acceptor sites', *Nucleic Acids Research,* Vol. 25, No. 15, pp.3159–3163.

[97] Uberbacher, E. C. and Mural, R. J. (1991) 'Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach', *Proceedings of the National Academy of Sciences of theUnited States of America,* Vol. 88, No. 24, pp.11261–11265.

[98] Victor, G.L., and Alexey, V.K. (2003) 'Recognition of Eukaryotic Promoters Using a Genetic Algorithm Based on Iterative DiscriminantAnalysis', *In Silico Biol,* Vol. 3, No. 1, pp. 81-87.

[99] Wang , K., Lv, j., Feng, W., Wang, X. (2010) 'A Novel Method for Splice Sites Recognition Using Comprehensive Information', *First International Conference on Pervasive Computing Signal Processing and Applications (PCSPA)*, pp.986 – 989.

[100] Wong, K. C., Chan, T. M., Peng, C., Li, Y., and Zhang, Z. (2013) 'DNA motif elucidation using belief Propagation', *Nucleic Acids Research* , Vol. 41, No. 16, pp. e153.

[101]Xu, Y., Einstein, J. R., Mural, R. J., Shah, M., and Uberbacher, E.C. (1994) 'An improved system for exon recognition and gene modeling in human DNA sequences*', Proceedings of the 16th Annual International Conference Intelligent Systems for Molecular Biology*, pp. 376–383,.

[102] Xu Y., Einstein, J.R., Mural, R.J., Shah, M., and Uberbacher, E.C. (1994a) 'An improved system for exon recognition and gene modeling in human DNA sequences', *Proc Second Intl Conf ISMB*, Vol. 2, pp.376-384.

[103] Xu, Y., Mural, R.J., and Uberbacher, E.C. (1994b) 'Constructing gene models from accurately predicted exons: An application of dynamic programming', *Comput Appl Biosci.*, Vol. 10, No. 6, pp.613-623.

[104] Yada, T., Nakao, M. , and Nakai, K. (1999) 'Modeling and predicting transcriptional units of Escherichia coli genes using hidden Markov models', *Bioinformatics*, Vol. 15, No. 12, pp. 987.

[105] Yin, M.M., and Wang, J.T.L. (2001) ' Effective hidden Markov models for detecting splicing junction sites in DNA sequences', *Inform Sciences* , Vol. 139, pp.139-163.

[106] Ying, X.U., Mural, R. J., Einstein, J. R., Shah, M.B., and Uberbacher, E.C. (1996c) 'GRAIL: a multi-agent neural network system for gene identification', *Proceedings of the IEEE*, Vol. 84, No. 10, pp. 1544–1551.

[107] Zhang, L., and Luo, L. (2003) 'Splice site prediction with quadratic discriminant analysis using diversity Measure', *Nucleic Acids Research* , Vol. 31, No. 21, pp.6214-6220.

[108] Zhang, Q., Peng,Q., Zhang, Q., Yan, Y., Li, K., and Li, J. (2010) 'Splice sites prediction of Human genome using length-variable Markov model and feature selection', *Expert Systems withApplications* , Vol. 37, p.2771–2782.