Journal of Theoretical and Applied Information Technology
31st May 2016. Vol.87. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

AN APPROACH FOR IDENTIFYING THE PRESENCE OF FACTOR IX GENE IN DNA SEQUENCES USING POSITION VECTOR ANN

[1]BIPIN NAIR B J, [2]K SYED KHAMARUDHEEN, [3]RANJITHA H S

[1,2,3]Department of Computer Science

[1,2,3]Amrita School of Arts and Sciences, Mysuru Campus

[1,2,3]Amrita Vishwa Vidyapeetham, Amrita University, India

E-mail: [1]bipin.bj.nair@gmail.com, [2]techkamar@gmail.com, [3]rs520159@gmail.com

ABSTRACT

Prediction of presence of genes in DNA sequences is of high importance in the domain of bioinformatics and genetics. Identification of proper gene splice junctions in DNA sequences is a difficult task. In our work, we have focused on a gene called Factor IX, the absence of which causes a deadly disorder called Haemophilia B. This paper proposes an approach for training the system to identify a specific gene by providing relevant protein sequences and identifying exact locations of splice junctions in DNA sequences. The algorithm proposed in this work has yielded promising results with an accuracy of 96.8% in detecting proper splice junctions for the given gene.

Keywords: *Haemophilia B, Factor IX, Position Vector ANN, Gene Prediction, DNA, Protein*

1. INTRODUCTION

The study of blood clotting proteins and genes plays an important role in the field of genetics. The increase in blood related hereditary disorders has expedited the need for human beings to explore new pathways for finding effective cures for these disorders. The numerous variations of blood related genetic disorders, make the task of identification of these disorders challenging.

In our study, we focus on gene "Factor IX", which is highly essential for the clotting of blood. This factor is also called "Christmas Factor". The Factor IX gene belongs to the family of serine proteases, present in the coagulation system of the human body. The genetic information about Factor IX can be found in the X chromosome of the Human genome project [Give ref].The original Factor IX gene can be found with variations in human beings. These variations also pose a challenge in identification as two different sequences of genes can have the same effects.

The absence of this gene is the main cause for a deadly blood related hereditary disorder called Haemophilia B. Haemophilia B is a X linked recessive disorder. Females inherit X chromosome from their parents (XX) and males inherit X chromosome from their mother and Y chromosome from their father (XY). If a mother is found to have an X chromosome with Haemophilia then it can spread to her son. The father cannot pass this disorder to his son. Now, in the case of daughter whose mother has an X chromosome with Haemophilia is most likely to inherit the healthy X chromosome from her father and end up not having Haemophilia. But, daughter may be a carrier, thus affecting the male children of hers. So, cases of Haemophilia B in girls are very rare (See Figure 1).

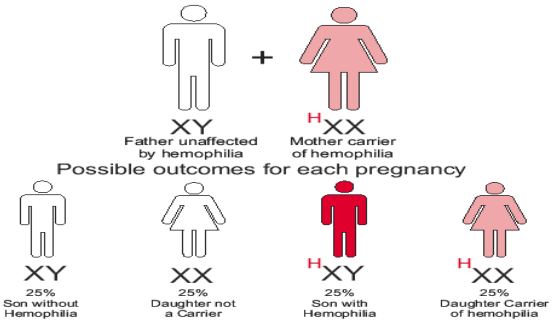

Figure 1: How Haemophilia Is Passed From Parents To Offspring Through Pregnancy [10].

Journal of Theoretical and Applied Information Technology
31st May 2016. Vol.87. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

In most cases, this disorder is transferred from parents to children during pregnancy. But around 1/3rd of the cases observed were a result of spontaneous mutation in gene. Mutations occur as a result of free radical reaction due to inhalation of oxygen. Most are nullified by the death of the mutated cells. But, sometimes these mutated cells remain in the body for a prolonged period which causes harmful effects and ultimately to chronic disorders like Haemophilia B.

The human body performs lot of enzymatic actions for its proper functioning. Whenever a particular protein is required to perform an enzymatic action, the DNA gets converted into RNA and then the required protein is synthesized from it. During this process, every triplet codons of DNA is converted into its corresponding protein codon. So, a human body which has undergone mutation might not be able to process the exact protein which is required at the moment. A protein which is synthesized incorrectly is not useful or may even damage the body. In this case, the mutation caused in body might synthesize a protein of similar nature to Factor IX, barring the actual ability to clot blood.

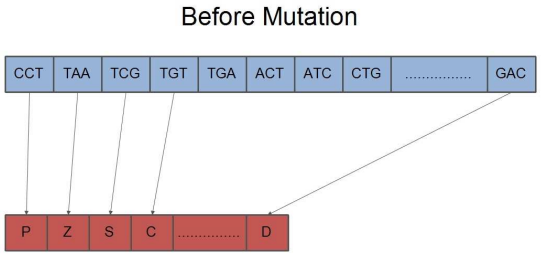

Figure 2: Protein synthesized before mutation of DNA

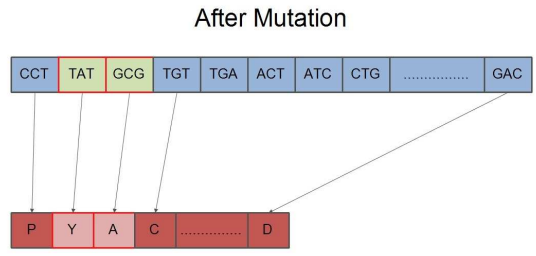

Figure 3: Protein Synthesized After Mutation of DNA

In the Figure 2, the codons in blue color depict the triplets of DNA codon and codons in red color are corresponding codons of proteins synthesized from DNA. But, in Figure 3, the DNA is mutated and the protein which is synthesized from it also shows a mutation from the original protein. The mutation of codons is depicted by using a different color in Figure 3. The newly synthesized protein from the mutated DNA might not work accordingly thus causing disorders like Haemophilia B.

2. OVERVIEW

Haemophilia B, is currently diagnosed clinically by conducting test to evaluate the time taken for blood to clot. A medical history of relatives is also taken into account for the purpose of diagnosis. Medical practitioners and doctors who treat this disorder collect DNA samples of the patients. These DNA sequences collected during the process of consultancy help physicians in identifying mutations which occur in the later stages of the disorder. DNA samples are either collected from a particular chromosome or the entire genome of the patients is also collected for the study.

In this paper, we propose a new approach for the analysis of this disorder. We collected DNA sequence samples from various patients suffering from Haemophilia B. The dataset considered a thousand DNA sequences with various types of mutations of the same disorder in various patients. After thorough analysis of these sequences and experimenting various techniques, we have come up with an approach to predict the absence of Factor IX gene in any DNA sequence thus helping us to detect the presence of this disorder. Our algorithm helps in identifying locations where mutations have occurred.

3. REVIEW OF LITERATURE

Noordewier et.al [6] have mentioned about a method of identifying splice junctions using Feed forward neural networks generated with KBANN. This method mainly concentrates on the knowledge of domain specific inference rules. Here, the domain theory mainly focuses on classifying DNA sequences appropriately. The steps in expression of genes mainly involves in converting a DNA to protein. They coarsely provide a procedure for distinguishing a class of DNA sequences also known as split-junctions. In here there are utilized sequences known as exons and removed sequences

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

also known as introns. This KBANN is a hybrid approach wherein it provides one output for the corresponding four inputs which are supposed to be the DNA's alphabets for all the DNA training sequences. The weights on the links are considered to be frozen i.e. it cannot be changed. In this paper there are no rules with negative examples, which do not cross intron/exon boundaries. This KBANN method result consists of 3 output units, 240 input units, 31 fixed hidden units and 12 tunable hidden units. This KBANN approach uses a training set i.e., the data of 3190 population and a test data set for the rest. The accuracy for the test data set is found to be with a minimal error rate when compared with the training set. This KBANN is effective only for the ANNs to improve the pre-existing knowledge and it's well-suited for the tasks they are intended to learn.

Sommer et.al [5] have mentioned about factor IX gene, which exploits to grow or corroborate multiple methods of mutation uncovering or mutation screening. The latest evaluations have provided intuition promote fields that is genotype-to-phenotype etc. Microinsertions/microdeletions are more common than large insertion/deletion. The repeated microdeletions are unique and they create a perfect hairpin loop from not so perfect sequences. The highly homologous genes are inclined for larger deletions. The insertions and duplications are very infrequent compared to deletions. This paper mainly concentrates on insertion, deletion of the sequences as well as orphan sequences that forms two short complimentary sequences, which forges itself inside single stranded DNA likely during DNA replication or repair. A dataset of a mother and grandmother belonging to a family have been considered to check with. They have used Bayesian approach for the mutation purpose. The inversion mutation of X linked factor VIII gene is mainly because the homologous recombination is solely parental in origin.

4. PROPOSED METHODOLOGY

Identification of proper splice junction is important for locating particular gene in the given DNA sequence. A splice junction is a location which acts as a parameter for cutting the DNA sequence by considering highest probability of finding a gene. In order to find the existence of a gene in the given DNA sequence, the DNA sequence is spliced into smaller DNA sequences called "reeds". They are generated by cutting a DNA sequence at random positions with random length. After splicing, these sequences are stored into a separate file. Whenever a user wants to search for a gene in a given DNA sequence, the file which contains the reeds are taken as input for the search process. This method yields poor results as the given sequence might contain the gene partially or not at all. Use of this approach results in increased running time of the application thus degrading performance. This method of search mostly ends up with negative results as the scoring function might have returned a value which is less than the threshold value for the existence of gene.

4.1 Homology detection using LCCS

Exact string matching techniques are hardly used as the probability of finding the same arrangement of amino acids in two DNA, RNA or protein sequences is negligible. So, we use the concept of similarity in strings. When two sequences have high similarity, they are known to be homologous in nature. The normal approach for finding homology between 2 sequences is to find the longest common subsequence (LCS) and to generate the substitution matrix. The substitution matrix helps us in identifying the mutations between the 2 sequences that is attributed to natural causes. This approach is fine when dealing with smaller sequences. But, DNA sequences which contain codon count in the range of 10,000 to 1, 00,000 the running time of the homology detection function increases. So, in our approach we are using a new approach of finding the Longest Common Continuous Subsequence (LCCS). In this method, we are finding the portion of codons which are continuous and common to both the sequences by providing a threshold value. The algorithm LCCS runs in lesser time when compared to LCS.

4.1.1 LCCS algorithm

Input: Sequence1, Sequence2, Threshold

1. S1=Sequence1.Length
2. S2=Sequence2.Length
3. occupancy=S1/S2
4. if occupancy>=Threshold
5. then
6. LccsLen=S1
7. while LccsLen>=(Threshold*S2)
8. do

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

```
9.      i=0
10.     while (i+LccsLen)<=S1
11.     do
12.      SubSeq=Sequence1[ i : i+LccsLen ]
13.      if SubSeq substring of  Sequnce2
14.      then
15.           return SubSeq
16.      else
17.           Increment i by 1
18.      Decrement LccsLen by 1
19.      done
20. done
21. end if
```

This algorithm returns a NULL value if no proper substring is found or when the threshold value is less than occupancy. This indicates the fact that the two given sequences are not homologous. For the implementation of our work, we have taken 85% as the threshold for checking similarity between sequences.

4.2 Machine Learning

Machine Learning is a crucial part of our application as it determines the splice junction in DNA sequences. Identification of correct splice junction helps us in reducing the search for the gene in locations where they are not likely to be present. For our work, we have collected over 943 protein variations of Factor IX. These protein sequences were collected from GenBank release 210 were used during the phase of training the machine for the identification of proper splice junctions. The main reason behind choosing the protein sequences as a parameter of identification of gene is due to the fact that the DNA gets converted into RNA and then protein is synthesized from it. So, in our training phase we focus on the protein sequences to generate the rules for identifying the best splice junctions.

4.2.1 Position Vector ANN

Position Vector ANN (PVANN) is a model of neural network based on Feed forward neural network. This network initially has 3 input units in input layer and they correspond to length of sequence, starting codon and ending codon. The hidden layer is composed of a layer with 20 units each corresponding to the 20 amino acids present in proteins. There are 2 bias neurons in the hidden layers with initial weights as 0. This is used for the purpose of fine tuning. In the output layer, we have a single unit which tells whether the given input corresponds to a proper splice or not. All nodes of each layer is connected to all nodes of other corresponding layers. Weights for all the nodes in the network are assigned randomly. During the training phase, the weights were reassigned by subtracting the error value from the weight.
We have trained the network by allocating 80% of the samples for training phase and rest 20% for the testing phase. The network is trained using feed forward model and tested with back propagation. The weights of all nodes were frozen once the training completed in order to avoid errors that may come from online training. The final network created by PVANN consist of 3 units in input layer,549 units in hidden layer and 1 unit in output layer.

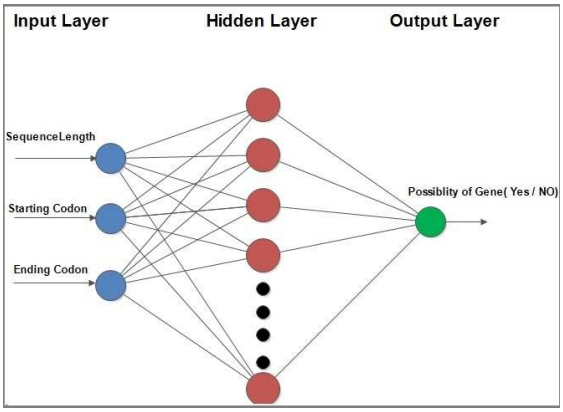

Figure 5: Position Vector ANN Model

Unlike KBANN [6], PVANN does not use knowledge base of domain-specific inference rules for the creation of its Neural Network. Instead, it analyzes the protein sequences during the training phase to formulate rules on its own. This is advantageous as the knowledge base provided was derived from biological analysis is found to be of low accuracy when it comes to the detection of specific genes. The biological base served as a common framework for all genes and it does help to identify proper splice junctions in some cases. But, it doesn't work in most cases like when the gene is present at the beginning of the sequence as most of the rules derived from biological analysis have every subsequence to have a certain number of successor and predecessor codons in order to satisfy the criteria to become a gene. KBANN does not have the capability of identifying the genes when it

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

comes to smaller sequences like that of a Factor IX protein which is of 11 codons. In the above given cases PVANN was found to be successful in identifying proper splice junctions with lesser error rate than that of KBANN.

4.3 Gene Prediction

A single DNA strand of any organism might contain multiple genes embedded in it. According to the Human genome project [Give ref], an average human being is estimated to be having around 2 million genes. So, efficient methods for searching these genes in a given DNA sample is of high importance. Prediction of presence of a particular gene is the core aspect of this work. In our work, we have devised an algorithm to detect the presence of a particular gene the system is trained for. In this case, we have trained the system to detect the presence of Factor IX gene. We can also detect other genes using this approach if the system is trained for any given gene by providing its corresponding protein sequences.

Gene predictor works in two phases. Phase 1 is for training, where the FASTA file containing all the protein sequences of a particular gene is loaded into the trainer. The trainer then generates a Feed forward neural network called PVANN. Once the training is completed then the trained rules are validated with the test dataset. Training process is repeated with different input datasets and test sets until the error rate is minimum and then the final rule set is completed. This rule set will be used later for finding the presence of gene in DNA.

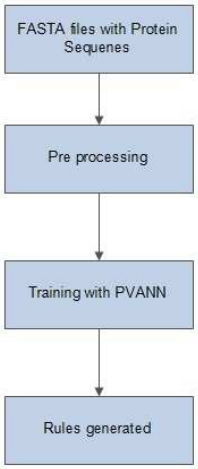

Figure 6: Generation Of Rules Using PVANN

Phase 2 is the detection phase. In this phase, we load the training rules into the application. When a DNA sequence is supplied by the user, most suitable sequences lengths and rules are verified over the DNA sequence to splice the DNA. These spliced DNA sequences are then converted into its corresponding protein sequences. The conversion is done by mapping each DNA codon triplet with its corresponding protein codon.

Table 1: Table For DNA Triplet To Protein Conversion

DNA Triplet	Protein Codon
ATT,ATC,ATA	I
CTT,CTC,CTA,CTG,TTA,TTG	L
GTT,GTC,GTA,GTG	V
TTT,TTC	F
ATG	M
TGT,TGC	C
GCT,GCC,GCA,GCG	A
CCT,CCC,CCA,CCG	P
GGT,GGC,GGA,GGG	G
ACT,ACC,ACA,ACG	T
TCT,TCC,TCA,TCG,AGT,AGC	S
TAT,TAC	Y
TGG	W
CAA,CAG	Q
AAT,AAC	N
CAT,CAC	H
GAA,GAG	E
GAT,GAC	D
AAA,AAG	K
CGT,CGC,CGA,CGG,AGA,AGG	R
TAA,TAG,TGA	Z

Then, the newly obtained protein sequence is cross checked for homogeneity by traversing through all the protein variations of Factor IX. Check for homogeneity is performed with the LCCS method with a threshold of 85%.If LCCS method returns a string which is not null then we conclude that the given DNA sequence has Factor IX gene embedded in it.

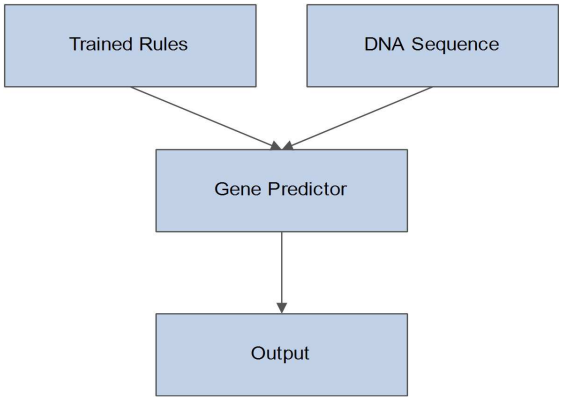

Figure 7: Working Of Gene Predictor

Journal of Theoretical and Applied Information Technology
31st May 2016. Vol.87. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

4.3.1 Gene Predictor Algorithm

Input: Trained Rules, DNA

1. Load trained rules into the application
2. Load DNA into the application
3. Validate the DNA sequence
4. for i=0 to DNA.Length
5. do
6. for each rule in Trained Rules
7. do
8. starcodon=getProtein(DNA[i:i+3])
9. length=rule.getLength()*3
10. endcodon=getProtein(DNA[length:length+3])
11. if ValidRule(length,starcodon,endcodon)
12. then
13. for each protein in FactorIX
14. do
15. Seq1=getProtein(DNA[i:i+length+3])
16. Seq2=protein
17. substr=LCCS(Seq1,Seq2,85)
18. if substr not null
19. then
20. Return "Gene Found"
21. end if
22. //If Gene was not found at all after all iterations
23. Return "Gene Not Found"

5. EXPERIMENT RESULTS

The search for the Factor IX gene was carried out on 4544 DNA samples collected from patients not having Haemophilia B and DNA samples of 1563 patients who were diagnosed with Haemophilia B. The DNA sequences were collected in RAW format as the patients were not ready to disclose their identities or any information that lead back to them. The obtained DNA sequences were not having any unidentified codons or gaps. So, the process of preprocessing was not carried out on any sample.

5.1 TRAINING MODULE

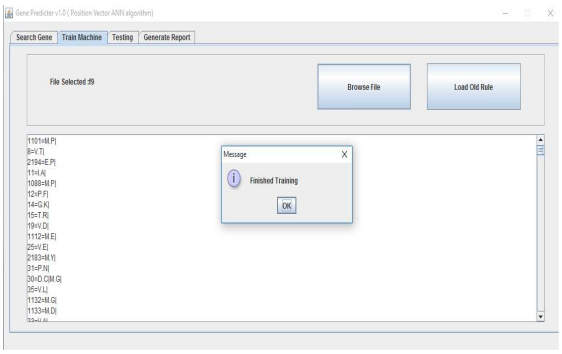

Figure 8: PVANN Training Module

Figure 8 shows the training module which is developed in java. This module performs mainly 2 tasks, it opens a protein file for training the system using PVANN and then it displays the trained rules in the textarea below. It has another button called "Load Old rule" to load the existing rules into the text area.

5.2 GENE PREDICTOR

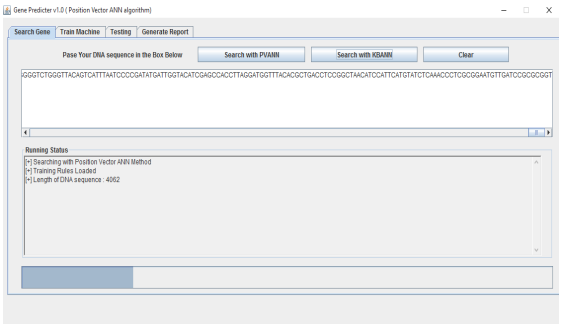

Figure 9: Gene Predictor In Execution

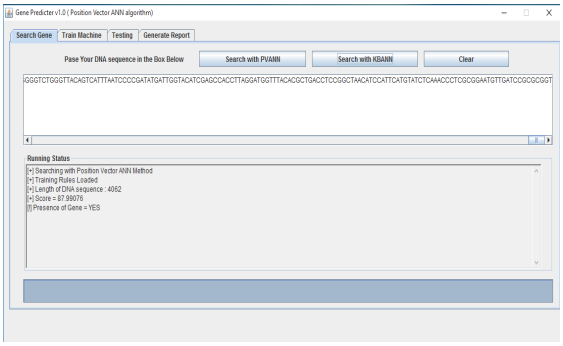

Figure 10: Gene Predictor After Completing Execution

Figure 9 shows the Gene predictor in execution phase. This module takes DNA sequence as input and starts searching for the required gene

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

by splicing the DNA with PVANN method and then after converting to protein, it checks for homogeneity with the help of LCCS algorithm. In Figure 10, the gene predictor has completed its execution and the results are displayed in the textarea below (See Figure 11).

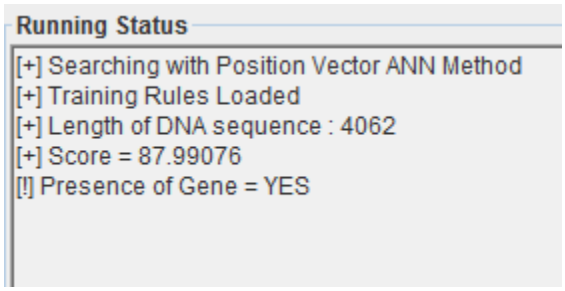

Figure 11: The Result Obtained For The Search Of Gene

5.3 COMPARISON OF PVANN WITH KBANN

Position Vector ANN (PVANN) was compared with Knowledge based ANN (KBANN) with various techniques and we have found PVANN to be efficient in most cases.

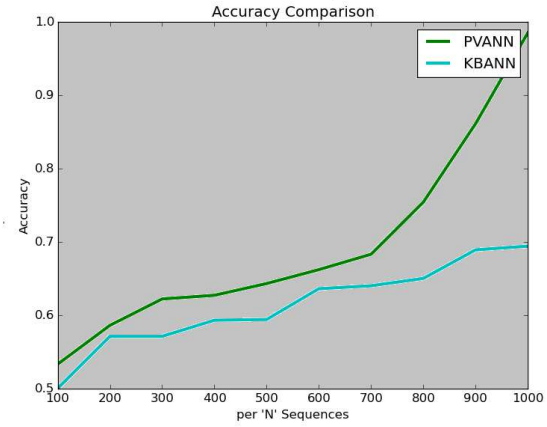

Figure 12: Accuracy Comparison of PVANN with KBANN

The above graph (Figure 12), depicts the comparison of accuracy of splice junctions when it comes to PVANN and KBANN. Accuracy is calculated with the formula

$$\text{Accuracy} = (TP+TN)/(TP+FP+FN)$$

Where, TP stands for True Positive, TN stands for True Negatives, FP stands for False Positive and FN stands for False Negatives.

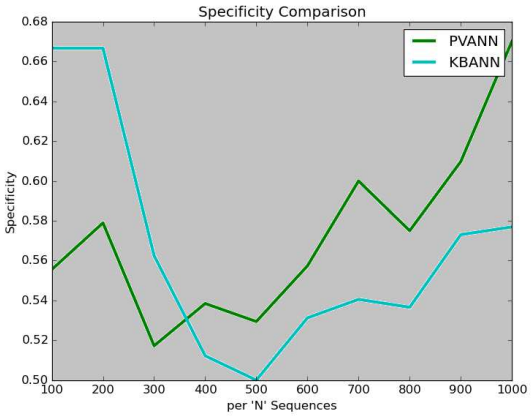

Figure 13: Specificity Comparison of PVANN with KBANN

The above graph (Figure 13), depicts True Positive Rate or Specificity. Here, Specificity is calculated by the formula

$$\text{Specificity} = TN/(TN+FP)$$

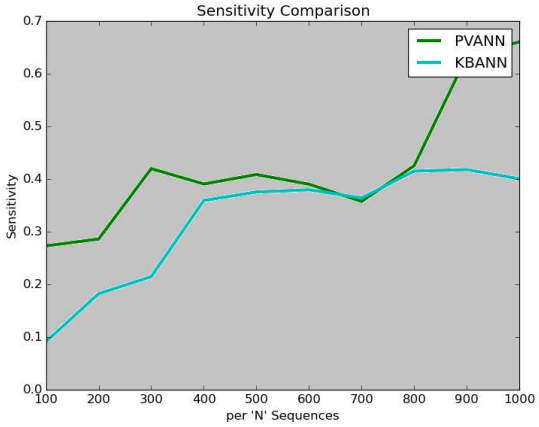

Figure 14: Sensitivity Comparison of PVANN with KBANN

The above graph (Figure 14), depicts True Negative Rate or Specificity. Here, Sensitivity is calculated by the formula

$$\text{Sensitivity} = TP/(TP+FN)$$

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 2: Results Obtained For Gene Prediction With KBANN and PVANN

Method	TP	FP	TN	FN	Accuracy
PVANN	4201	343	871	692	96.8%
KBANN	3998	546	723	840	87.6%

6. CONCLUSION

The proposed algorithm for the prediction of possibility of Factor IX gene in given DNA sequence was found to better when compared to KBANN. Our newly proposed algorithm not only runs faster than its predecessor but also more accurate in identifying splice junctions. This in turn helps us in detecting the presence of Factor IX genes in the given DNA, absence of which causes Haemophilia B. Our algorithm works with an accuracy of 96.8% and this can be improved by enhancements in future.

REFRENCES:

[1] Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., ... & Gocayne, J. D. (2001). The sequence of the human genome. science, 291(5507), 1304-1351.

[2] Armentano, D., Thompson, A. R., Darlington, G., & Woo, S. L. (1990). Expression of human factor IX in rabbit hepatocytes by retrovirus-mediated gene transfer: potential for gene therapy of hemophilia B. Proceedings of the National Academy of Sciences, 87(16), 6141-6145.

[3] Montana, D. J., & Davis, L. (1989, August). Training Feedforward Neural Networks Using Genetic Algorithms. In IJCAI (Vol. 89, pp. 762-767).

[4] Huang, K., Yang, C. B., & Tseng, K. T. (2004, December). Fast algorithms for finding the common subsequence of multiple sequences. In Proceedings of the International Computer Symposium (pp. 1006-1011).

[5] Sommer, S. S., & Ketterling, R. P. (1996). The factor IX gene as a model for analysis of human germline mutations: an update. Human molecular genetics, 5(Supplement 1), 1505-1514.

[6] Towell, G. G., & Shavlik, J. W. (1994). Knowledge-based artificial neural networks. Artificial intelligence, 70(1), 119-165.

[7] Prompramote, S., Chen, Y., & Chen, Y. P. P. (2005). Machine learning in bioinformatics. In Bioinformatics technologies (pp. 117-153). Springer Berlin Heidelberg.

[8] Bailey, T. L., Williams, N., Misleh, C., & Li, W. W. (2006). MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic acids research, 34(suppl 2), W369-W373.

[9] Morgenstern, B., Dress, A., & Werner, T. (1996). Multiple DNA and protein sequence alignment based on segment-to-segment comparison. Proceedings of the National Academy of Sciences, 93(22), 12098-12103.

[10] Transfer of Haemophilia B[Image] Retrieved from http://www.hemophilia-information.com/images/geneMother.gif