International Journal of Engineering & Technology, 7 (1.9) (2018) 167-170

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research Paper

Coding and functional defect region prediction of placental protein in an embryo cell of first trimester using ANN approach

Bipin Nair B J [1] *, Rahul Reghunath [1]

[1] *Department of computer science Amrita School of Arts and Science, Mysuru Campus Amrita Vishwa Vidyapeetham, India*
**Corresponding author E-mail: bipin.bj.nair@gmail.com*

Abstract

The protein coding and functional regions in DNA sequences has become an exciting task in bioinformatics. In particular, the coding region has a 3-base periodicity, which helps for exon identification. Many signal processing tools and techniques have been successfully applied to identify tasks, but still need to be improved in this direction. In our work, we employ ANN classifier to predict coding and functional region of protein in human embryo cell protein in first trimester, and evaluate their performances according to the comparison energy levels of coding region. The obtained from the threshold energy level, results show that in a box plot finally predict the mutation.

Keywords: *Promoter Prediction; ANN; Placenta; DNA; Box Plot.*

1. Introduction

Here we are discussing bioinformatics area in that coding and function region prediction of a protein within a cell. We are introducing a new promising model for energy calculation and identification of the threshold energy of coding regions. S transform is a accurate time-slot analysis. The capability of this function was evaluated by energy calculation studies and the results obtained were compared with the threshold energy. By supporting mRNA degradation and translational repression the small MicroRNAs that are non-coding RNAs work as the main post-transcriptional regulators of gene expression. Many omnipresent and specific miRNAs are expressed by Placenta. Aberrant miRNA expression is associated with pregnancy difficulties such as preeclampsia. Recent studies of placental miRNA have focused on the identification of placental miRNA species. This article reviews the present knowledge about the functions and expression of miRNAs in placenta development and provides further directions for miRNA research.

Mainly our work is useful in the area of gynecology it will help to identify the defects in embryo in the first stage of development itself. Promoter is nothing but a region of DNA that initiate transcription of particular gene. Promoters are located near the transcription sites of genes.in our work we are predicting the coding and functional region of human cellular proteins using ANN classifier .and calculating energy of protein sequence and comparing the threshold energy of the given sequence using various mathematical model .we are identifying the coding region that ie exon from protein structure using simulation method

In normal medical diagnosing identifying the defects in embryo in the early stage is very difficult so we can overcome that using our method as well as consume the time while doing multiple test, doing various such kind of test are costly. Placenta provides many abundant as well as peculiar microRNAs. These microRNAs control trophoblast cell apoptosis, angiogenesis, invasion, and differentiation, demonstrating that miRNAs play important roles during placental growth. So in that development stage any mutation is happening it is very difficult to identify accurately using clinical method. Looking at differential articulation of The miRNA between the placenta of normal and impaired placenta reveals the ability of miRNAs in the placenta, in addition to the dull process. More research is needed to gain additional insight into the actual nature of miRNAs in the field of placental development.Moreover, the possibility of using miRNAs as therapeutic targetsand biomarkers in case of pregnancy-related diseases can be explored. This algorithm says that when every passage of test set is exhibited to a system the system inspects its yield reaction to the information design. In view of wanted yield blunder esteem is calculated..

2. Litrature survey

Researches are reported in the area of bioinformatics for the interactions of protein coding region prediction, some of the related the work are summarized here.

Sitanshu Sekhar[1] Many indication dispensation tools and techniques have been successfully useful to identify tasks, but still need to improve this direction. In this paper, we present another promising model-free time-recurrence sifting technique based on S-change for precise acknowledgment of coding districts. P.Fariselli [2] Describes a neural network-based method for forecasting the contact pattern of proteins using the input chemical and evolutionary information. The neural network data set includes 200 non- decomposable three-dimensional structures of non-homologous proteins.OLOF [3] Based on the neural network method ~ ChloroP! Used to distinguish chloroplast travel peptides and their cleavage locales. Utilizing cross approval, 88% of the sequences of our homology decrease preparing set were accurately named travel peptides or non-travel peptides. This execution level is considerably higher than the openly accessible chloroplast situating indicator PSORT

Bosna [4] the expression of the promoter control gene that is often present in the relevant gene in its DNA sequence. In order to find the gene in a specified sequence, the researcher must find the posi-

tion of the organizer. Therefore, the problem of organizer recognition is very important in biology. So the question is still exposed. Here we used the ANN classifier to forecast the organizer of DNA sequences and to evaluvate the presentation. J. Cheng [5] SCRATCH is a server that predicts the three-dimensional structural and structural features of a protein. The SCRATCH package includes relative solvent accessibility, disulfide bonds, disordered domains, secondary structure, domains, residue contacts, single mutation stability and average, single residues and tertiary predictors. By selecting the desired prediction after providng the amino acid sequence the user submits the sequence to the server. Edgardo [6] over the past few years, the ever-increasing number of known nucleic acid and protein sequences has led to the development of advanced computing tools to search for sequence similarity in macromolecule databases. There are some powerful algorithms for comparing two more sequences, but less sensitive algorithms to identify the relevant proteins are also presented. MORTEN [7] they tells that the several neural network combinations derived implementing various sequence encoding schemes are of desirable performances. The performances of the fresh method is remarkably higher than the others. Through the use of mutual information, we show that the binding of the HLA A * 0204 complex to the peptide shows the higher order sequence associated with the signal.

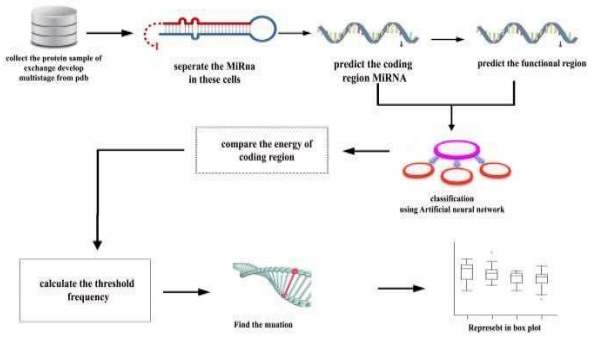
EDWARD [8] Identifying genes in a large area of characteristic DNA is a daunting task that is currently the focus of much research effort. We describe a reliable computational method for locating the protein-coding part of a gene in an anonymous DNA sequence. We use the concept of robot environment perception, our method combines sensor algorithms and neural networks to locate the coding area .Ning Qian [9] A New Technique to Predict 2d structure of Globulin Based on NNN Model. The network model learns the task of predicting the secondary structure of the amino acid sequence from the existing protein structure.T. Hubbard [10] neural networks are cultured to predict sub cellular localization of proteins. For the three possible subcellular locations in prokaryotes, 81% of the predicted accuracy can be achieved. The distribution of reliability indicators, the prediction of 33% can be accurate to 91%. Since subcellular localization limits the possible functions of proteins, this approach should be a useful tool for systematically analyzing genomic data available through servers on the World Wide Web. L.Robert [11] the four nitrogen bases of DNA are spelled out to formulate the protein. A gene usually contains five thousand genes, but usually only a small part is a protein. Computer-based predictive systems are increasingly dependent on exponential growth of major genetic databases. There are several systems that use the usual neural networks and Markov chain porodovsky and McIninch models to locate coding regions (exons) and non-coding regions (introns) within genomic DNA.

Burkhard [12] Introduced a neural network contained in multiple sequence alignments as input, the accuracy can be significantly improved to predict secondary structure. Using location-specific save weights as part of the input can improve performance. The use of the number of insertions and deletions reduces the tendency for over-prediction and improves the overall accuracy. The global increase in amino acid levels is further enhanced mainly by the prediction of structural categories. The final web system has a sustained overall accuracy of 71.6% in 126 unique cross-validation tests of the unique protein chain. Robert Farber [13] neural network technology is used to distinguish the open reading frame (ORF) sequences of introns and exons. For a neural network (essentially a sensor with an S-shaped or "soft-step capable" output) to perform the authentication this method calculates codon frequencies in ORFs of a defined length and uses a codon frequency representation of the DNA fragment. After training, the network is applied to an unrelated "forecast" dataset to assess accuracy. Our previous job accuracy was 98.4%, surpassing the accuracy of other algorithms in the literature. Here, we report the higher accuracy of mutual information calculations. Goel [14] this article's assessment attempts to identify and deal with the most promising methods on neural network classification. To evaluate the accuracy of these methods there are multiple methods. The way that NetGene2 extracts information makes it a complex way of predicting the structure of a gene. Bipin Nair [15] propose an efficient tool to analyze the possibility of getting affected by Non-Small Cell Lung Cancer (NSCLC) by comparing Lung Cancer microRNAs (LC-miRNAs) structures. Here we use global optimal alignment and Target Scan for target comparison and binding location detection. Bipin Nair [16] proposes a methodical study and analysis of seven types of LC-miRNAs: mir-31, mir-21, mir-143, mir-145, mir-155, mir-210, and mir-372. And they aligned the sequence of these miRNAs to observe and analyze matches, mismatch, and gaps with the normal miRNA sequences.

Hatzigeorgiou [17] they focus on using Back-percolation, Cascade-Related and Time Delay neural networks. This method provides a much efficient generalization than the known backpropagation algorithm. Kyoung-jae Kim [18] A feature discretization method based on genetic algorithm (GA) and artificial neural network (ANN) connection weight is proposed to forecast the stock price index.In addition to finding the optimal solution or approximate optimal solution of the connection weight in the learning algorithm, the genetic algorithm also checks for optimal or approximate optimal threshold of feature discretization for dimensionality reduction C. Pijanowski [19] Introduced a version of LTM parametrized in the Traverse Bay basin in Michigan and explored factors such as residential streets, roads, inland lakes, rivers, the Great Lakes lakeshore, recreational facilities, highways, agricultural densities and landscape quality How it affects the urbanization pattern of this coastal basin.

Tomita [20] they presented a sensitive marker ANN to predict the development of allergic diseases. To predict the development of childhood allergic asthma (CAA) and to select for susceptible SNPs, they used parametric reduction (PDM) ANN to analyze 25 SNPs of 17 out of 344 Japanese and screen for ten susceptible SNPs for CAA. Artificial Neural Network (ANN) Artificial neurons are the warm-up components of the body's natural neurons. The algorithm uses a new backpropagation algorithm to improve the existing neural network through computer modeling. The algorithm provides that when each entry in the sample set is presented to the network, the network checks its output response to the input mode. Using predictive coding and function areas based on expected output error values, we can use ANN.

3. Flow diagram



4. Problem formulation

The existing work does not use a classification method. Using Back Propagation algorithm. but it was not accurate because it does not use a classifier. The accuracy of existing system was 87% as well as finding the defects in embryo in first trimester from placental protein is a tedious process from 87% accuracy. System identifying the defects will be very difficult from the clinical perspective wise identifying the defects in the embryo in early stage it will be tedious process.

5. Proposed work

In the proposed system, we use the Back Propagation algorithm as well as a classification method. In the proposed work, we found an accuracy rate of 95% and above. This algorithm is very simple and efficient way to compute accuracy in neural network. In back propagation algorithm adjust the weight to reach the minimal rate of error function. The execution time is also improved in the proposed work.

6. Problem definition

In proposed work we use effective classifier to segregate coding region in a sequence as well as in terms of accuracy it was good because we are proposing an effective method using ANN classifier, it will train the placental protein data set and construct a network with 40 hidden nodes.

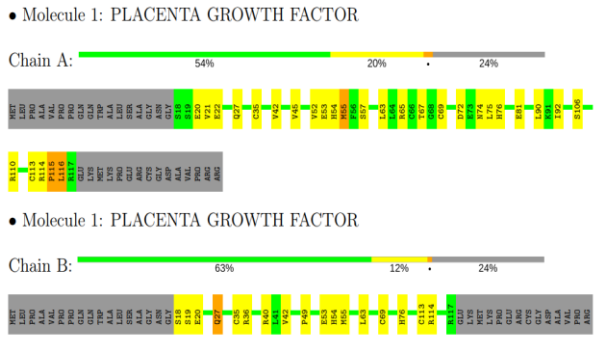
7. Related work

The existing system does not use a classification method. Back Propagation algorithm was used but it was not accurate because it does not use a classifier. The accuracy of existing system was 87%

8. Methodology

Collection of data
The collection of mi RNA placental protein sequence in the testing the proficiency of ANN. Such data were collected from the PDB Repository.
Data set contains a set of 1000 promoter and non- promoter instance of placental protein sequence.
There is two phases – Training phase and Testing Phase
Training Phase: Training will involve establishing a classification model.
Testing Phase: Testing entails the implementation of the classification model.

9. Data set



10. Algorithm

The paper admits the back-propagation algorithm where ANN is simulated and aligned for a forward signal transmission, thus allowing for signal errors to be propagated on the reverse. The back propagation algorithm along with the classification method is used for optimal result.
read the input data
Assign the weights
Assign number of input layer
Assign the hidden layer count
Output layer
Weight <- a node
While check termination
Do
For p=1 to number of sequence
For n=0 to number of features
W[n] =w[n] +rate*(Target-sum[p]-weighted-sum[p]*sequence[p][j]
End return weights

11. Result

Using threshold parameters, the performance of the promoter predictions was evaluated. A pair of equations were integrated to confirm to the result. It is found that the accuracy of the prediction has been improved to a great extent by using the back propagation method along with classification

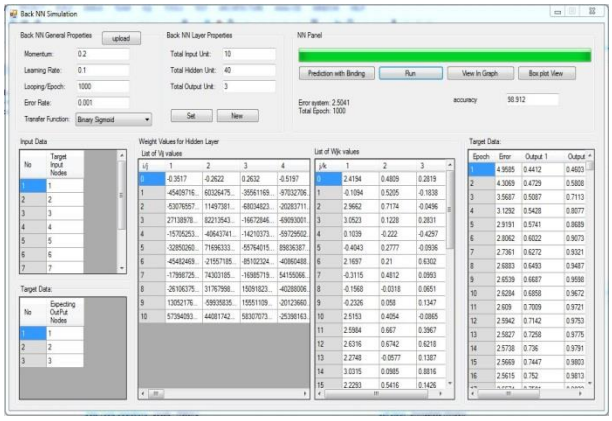


Fig. 10.1: Back Propagation ANN Simulation.

In this phase we will upload the placental protein sequence then we will train the system with 10 input layer 40 hidden layer and 3 output layer. working of ANN based on two functions ,first the set of protein sequence is converts as energy values using mathematical model then pass these values to first single node check the error rate , then pass through multiple node for multiple times. each node is named as 1,2,3etc,first node is taken as i node ,in this output other layer shows normal output layer one to 10 shows the real output as well as when the energy value traverse through each node weight will decide the weight of each node will decide based on actual pass based on time and signal. according to our result 1000 times it will execute that may times input energy will pass through a single way or multiple way ,with this way we will get error rate based on error rate we again change the node and send the input in different path finally we will get the exact path, based on this error approach we will get exact classification of placental protein sequence.

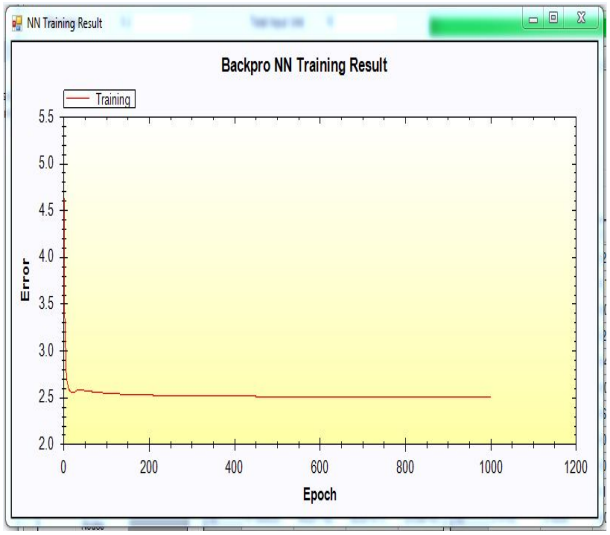


Fig. 10.2: ANN Training Result.

In this approach we are using recursion to decide the best path,so finally we are reducing the error rate which shown in the graph based on that we filter suitable data set and easily with accuracy we can find out the defects in placental protein for first trimester.in this graph x-axis we are taking eroor approach value and y-axis we are taking error rate and ploting the line it says when it will decrease the rate of error that much defects will be less.

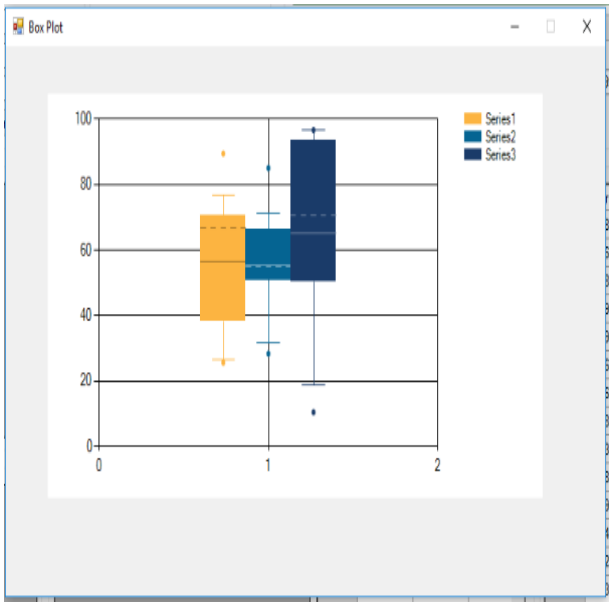


Fig. 10.3: Box Plot.

In this box plot representation it shows that best node and the path which energy traverse is represented in blue box the other two shows the error rate boxes, the blue yellow and light blue box will give the defects in the first trimester placental protein sequence

12. Conclusion

To predict the coding region in a placental protein sequence, we use data sets to achieve accuracy. The result obtained is compared with the previous method and finally using back propagation algorithm we gain the high accuracy which we interpreted with the help of boxplot.in our work around a quarter portion of accuracy we are reaching based on ANN approach

References

[1] Sahu, S. S., & Panda, G. (2011). Identification of protein-coding regions in DNA sequences using a time-frequency filtering approach. Genomics, proteomics & bioinformatics, 9(1), 45-55. https://doi.org/10.1016/S1672-0229(11)60007-7.

[2] Fariselli, P., & Casadio, R. (1999). A neural network based predictor of residue contacts in proteins. Protein Engineering, 12(1), 15-21. https://doi.org/10.1093/protein/12.1.15.

[3] Emanuelsson, O., Nielsen, H., & Von Heijne, G. (1999). ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. Protein Science, 8(5), 978-984. https://doi.org/10.1110/ps.8.5.978.

[4] Karlı, G., & Karadağ, A. Predicting Functional Regions in Genomic DNA Sequences Using Artificial Neural Network.

[5] Cheng, J., Randall, A. Z., Sweredoski, M. J., & Baldi, P. (2005). SCRATCH: a protein structure and structural feature prediction server. Nucleic acids research, 33(suppl_2), W72-W76. https://doi.org/10.1093/nar/gki396.

[6] FerrAn, E. A., Ferrara, P., & Pflugfelder, B. (1993, July). Protein classification using neural networks. In ISMB (pp. 127-135).

[7] Nielsen, M., Lundegaard, C., Worning, P., Lauemøller, S. L., Lamberth, K., Buus, S., & Lund, O. (2003). Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. Protein Science, 12(5), 1007-1017. https://doi.org/10.1110/ps.0239403.

[8] Uberbacher, E. C., & 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, 88(24), 11261-11265. https://doi.org/10.1073/pnas.88.24.11261.

[9] Qian, N., & Sejnowski, T. J. (1988). Predicting the secondary structure of globular proteins using neural network models. Journal of molecular biology, 202(4), 865-884. https://doi.org/10.1016/0022-2836(88)90564-5.

[10] Reinhardt, A., & Hubbard, T. (1998). Using neural networks for prediction of the subcellular location of proteins. Nucleic acids research, 26(9), 2230-2236. https://doi.org/10.1093/nar/26.9.2230.

[11] Roberts, L., Steele, N., Reeves, C., & King, G. J. (1995). Training neural networks to identify coding regions in genomic DNA. https://doi.org/10.1049/cp:19950589.

[12] Rost, B., & Sander, C. (1994). Combining evolutionary information and neural networks to predict protein secondary structure. Proteins: Structure, Function, and Bioinformatics, 19(1), 55-72. https://doi.org/10.1002/prot.340190108.

[13] Farber, R., Lapedes, A., & Sirotkin, K. (1992). Determination of eukaryotic protein coding regions using neural networks and information theory. Journal of molecular biology, 226(2), 471-479. https://doi.org/10.1016/0022-2836(92)90961-I.

[14] Goel, N., Singh, S., & Aseri, T. C. (2016, March). Neural network based splice site prediction methods. In Electrical, Electronics, and Optimization Techniques (ICEEOT), International Conference on (pp. 4282-4287). IEEE. https://doi.org/10.1109/ICEEOT.2016.7755526.

[15] Nair B J,Anju K J & Jeevakumar A(2016). Tobacco Smoking Induced Lung Cancer Prediction by LC-MicroRNAs Secondary Structure Prediction and Target Comparison. https://doi.org/10.1109/IJSIS.1996.565045.

[16] Nair B J,Anju K J (2016). Evaluation of Seven Types of MicroRNA in Cancerous Alveolar Cells Using Customized Hirschberg's Algorithm.

[17] Hatzigeorgiou, A., Mache, N., & Reczko, M. (1996, November). Functional site prediction on the DNA sequence by artificial neural networks. In Intelligence and Systems, 1996. IEEE International Joint Symposia on (pp. 12-17). IEEE.

[18] Kim, K. J., & Han, I. (2000). Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index. Expert systems with Applications, 19(2), 125-132. https://doi.org/10.1016/S0957-4174(00)00027-0.

[19] Pijanowski, B. C., Brown, D. G., Shellito, B. A., & Manik, G. A. (2002). Using neural networks and GIS to forecast land use changes: a land transformation model. Computers, environment and urban systems, 26(6), 553-575. https://doi.org/10.1016/S0198-9715(01)00015-1.

[20] Tomita, Y., Tomida, S., Hasegawa, Y., Suzuki, Y., Shirakawa, T., Kobayashi, T., & Honda, H. (2004). Artificial neural network approach for selection of susceptible single nucleotide polymorphisms and construction of prediction model on childhood allergic asthma. BMC bioinformatics, 5(1), 120. https://doi.org/10.1186/1471-2105-5-120.