International Journal of Engineering Inventions
e-ISSN: 2278-7461, p-ISSN: 2319-6491
Volume 3, Issue 6 (January 2014) PP: 22-27

Predicting Functional Regions in Genomic DNA Sequences Using Artificial Neural Network

Günay Karlı[1], Adem Karadağ[2]

[1]*International Burch University, Faculty of Engineering and IT, IT Department, Sarajevo, Bosnia and Herzegovina*
[2]*Bosna Sema - Educational Institutions, Sarajevo, Bosnia and Herzegovina*

ABSTRACT: *A promoter which frequently appears before its associated gene in DNA sequence governs the expression of genes. In order to locate a gene in a given sequence, researchers have to find the location of a promoter. Thus, the problem of promoter identification is of major importance within biology. So this issue is still open. In this study, we employ ANN (Artificial Neural Network) classifier to predict promoters of DNA sequences, and evaluate their performances. The obtained results show that the classifier competes the existing techniques for identifying promoter regions.*

Keywords: *Promoter prediction, ANN, data mining, bioinformatics, DNA.*

I. Introduction

Cell adaptability to a dynamic environment is an indispensable aspect of the survival of any organism. This entails cells being able to change and respond to the various stimuli as a result of changes in the surrounding. A key aspect underlies this adaptability; this is the protein production, which occurs through two core processes i.e. Translation and Transcription. In a snapshot usually the cells which are the basic units of life facilitate gene mutilation and they get transcribed thereby rendering the newly formed transcripts to get translated. This process may be more intense as exhibited by the capability of cells to biomechanically synthesize the transcripts and proteins thus initiating the adaptability (Kliman & Hoopes, 2010).

This is the point at which the direction and nature of transcription is determined is usually as a result of facilitation by RNA Polymers as various regions exist in the DNA molecule that hold the genetic composition. Such gene regions are known as Promoters (Gabriela & Bocicor, 2012).

The Blueprint; a common name for genetic coding is believed to possess the instructions needed by cells in environmental adaptability. This is further collaborating with research that cites that in addition to the instructions carried by the blueprint, there exists a synthesis point for most of the molecule including the RNA and proteins. The instructions contained in the blueprint are designed in such a manner that they are only readable in two ways as mentioned earlier in this paper i.e. transcription and translation (Clancy, 2008).

The messenger RNA; usually a single strand RiboNucleic Acid molecule, is amalgamated usually from one of the strands contained in DNA through a complementary of the origins. This strand does relate to a gene. Transcription will usually commence with the RNA polymerase being engulfed by a lone point in the DNA molecule known as the Promoter. (Huang, 2003) points out that, efforts to incorporate computerized biology have helped to easily locate the Promoter regions.

This has had a very significant impact on transcription as these regions are numerous in the DNA molecule thus easing the gene expression (Huang, 2003).

This process is facilitated majorly by the Promoter having much consideration of the responsibility it has on the transcription from a DNA strand to an RNA strand. This is further identified as the sequential upstream from the Transcriptional Start Site (TSS). This is well illustrated in the Fig. 1. The entire process commences with the binding of the RNA polymerase to a promoter array in the DNA molecule up to a point where the coding is realized. Coding occurs during the upstream movement around the promoter usually starting at 3’ end of the DNA molecule to the 5’ point in a DNA molecule (Clancy, 2008).

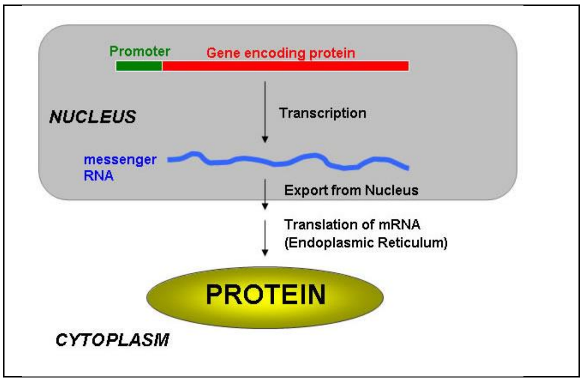

Figure1: Position of the promoter in a dna sequence (Clancy, 2008).

Additionally, efforts to envisage coding promoters has come a long way since the 1980's, laying foundation for further efforts in the development of modifications to help in establishing translatable MRNA sequences from the amassing of coding sequences. Since then several mechanisms have come up to aid in the process. Such are the Gen Viewer (Milanesi, Kolchanov, & Rogozin, 1993), GeneID (Guigo, Knudsen, Drake, & Smith, 1995), GenLang (Dong & Stormo, 1994), GeneParser (Stormo & Snyder, 1993), FGENEH (Solovyev, Salamov, & Lawrence, 1994), SORFIND (Hayden & Hutchinson, 1992), Xpound (Skolnick & Thomas, 1992), GRAIL (Xu, Mural, & Uberbacher, 1994), VEIL (Henderson, Salzberg, & Fasman, 1997), GenScan (Karlin & Burge, 1997), etc. (Hrishikesh, Nitya, & Krishna, 2011) Posits that the Grail and Gen Scan are highly used in learning institutions as well as commercially.

According to (Gordon, Towsey, & Hogan, 2006), the computational methods are more focused on the identification of motifs in a DNA molecule. In addition to the statistical interventions incorporated, other techniques such as using weights matrices in addition to the Markov Models as indicated by (Liu, 2002) (Luo & Yang, 2006) (Premalatha & Aravindan, 2009) and artificial intelligence. This has also entailed the integration of artificial neural network shave which exhibits subtle values (Abeel T. S., 2008) (Zhang, 2009).

In rare cases, then there may be the existence of the upstream of the TSS of DNA array that possess transcriptional characteristics, thus the presence of the promoter may not be a necessity. When promoter prediction is incorporated, a researcher is at ease to constrict down the entirely colossal DNA sequences. This paper acknowledges that through biological interventions, it is made easier to verify the DNA sequence that may either be transcribed or no. This though comes along with economic constraints.

In the recent past, the integration of computerization in the promoter identification and prediction has raised debate. And the results obtained by evaluating the classification model proposed in this paper confirm that applying ANN for promoter sequences recognition is promising.

II. Metarial And Methods

There are two core classes of the promoter prediction, namely '+' and '-'. These classes will denote the existence of promoter prediction in the DNA sequence, having the '+' denoting for a positive indication of promoter location in the DNA sequence and the '-' denoting the absence of promoter locations in the DNA sequence. This research paper proposes to deal with a supervised learning technique in the prediction of promoter regions in the DNA sequence.

2.1. Collection of Data

The research sought to incorporate the E. Cole promoter gene arrays of DNA in the testing the proficiency of ANN. Such data were collected from the UCI Repository (Frank & Asuncion, 2010); this contains a set of 106 promoter and non-promoter instances. The research paper notes that such data is viable in the comparisons of ANN with the models existing in the literature; additionally such information involving the use of the data set is publicly available (Gabriela & Bocicor, 2012).

The 106 DNA arrays are composed of 57 nucleotides each. 53 of the DNA sequences in the data set had a '+' denoting, indicating the presence of promoter location in the DNA array. The research then sought to align the (+) parameter instances separately allowing for transcription. The following data characterize the (+) instances as observed from the experiment. One is that for every occurrence the (+) represents for the promoter positive presence, a name was also given in each instance and a classification of the DNA array was made composing of A, T, G and C stand for Adenine, Thymine, Guanine, Cytosine (Frank & Asuncion, 2010).

2.2. Artificial Neural Network (ANN)

The artificial neuron is an enthused component in the body's natural neurons through computation modeling (Jain, 1996). The artificial Neural Network works in an induced principle where in instances that the body's neurons do receive signals through the naturally induced synapses occurring in the dendrites of the neuron, with a much intense magnitude, the ANN is induced thus releasing signal messages via the axon. In other instances the signal may be sent to other synapses and probably induce other neurons as noted by (Alverez, 2006). Usually the human brain is capacitated with the ability to hold numerous and complex operations, thus hails from the possession of numerous and enabling elements such as the complex neurons that consist of a more than 103 to 104 more neuron affiliations. This compounds the neuron coverage in the brain to approximately 1014 interconnections (Alverez, 2006) and (Kriesel, 2005).

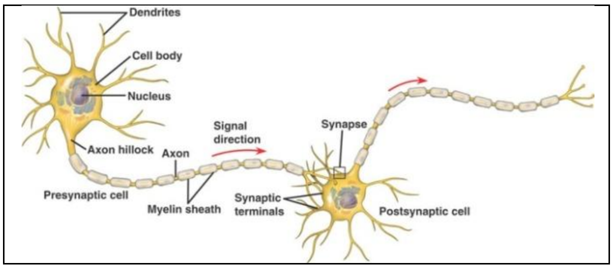

Figure 2: Neural dendrites, axon, and synapse
(Courche, 2013)

In modeling, the density of real neurons is highly exhibited, fundamentally comprising of synapses, which are compounded by the density of the respective signals, in addition this is taken under a mathematical simulation thus helping to evaluate the activation of the neuron. This moreover helps to compute the result of the artificial neuron. According to (Kriesel, 2005) this hails from the property of the ANN that they can integrate artificial neurons in the processing of information.

Usually, getting a precise definition of learning is a difficult task considering that the capability to learn is an essential characteristic of intelligence. From the experiment, it is posited that the ANN description is able to view from the efficient performance of a neuron task owing to updating of network systems. This is evidenced by the literature in (Alverez, 2006) and (Gandhi & Parekh, 2012).

One is able to obtain the desired output from the manipulation of the ANN; this is so by modifying the ANN weights. In such modifications, getting them by hand is a rather complicated and impossible task, giving supportive ground to the incorporation of ANN. In addition, (Gandhi & Parekh, 2012) and (Gershenson, 2008), algorithms may be integrated in the modifications and alignments of ANN weights.

The paper acknowledges the back-propagation algorithm where ANN is aligned in layers and is simulated for a forward signal transmission, thus allowing for signal errors to be propagated on the reverse (Gershenson, 2008). The input area is the location where the neurons impact the networks and therefore initiating the output. Fig. 3 illustrates a three layered neural network having inputs and output.

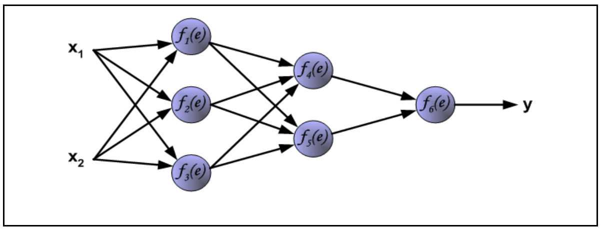

Figure 3: Multi-layer neural network

A neuron posses two units that complement the products of weight coefficients and input signals with the other unit being responsible for the neuron activation function following its capability to decode non-linearity. The units are denoted as Signal e for adder output signal and Signal y for the output signal of non-linearity.

The experiment notes the necessity to obtain a training data set that will comprise of input signals of x1 and x2 with a desired output z. In the network training, modifications of ANN weights are evaluated using the algorithm that will seek to commence with manipulating for both input signals from the training data set. Consequently, the output signals' values are made easier to identify from each neuron in the network (Golda, 2005). (See Fig. 4)

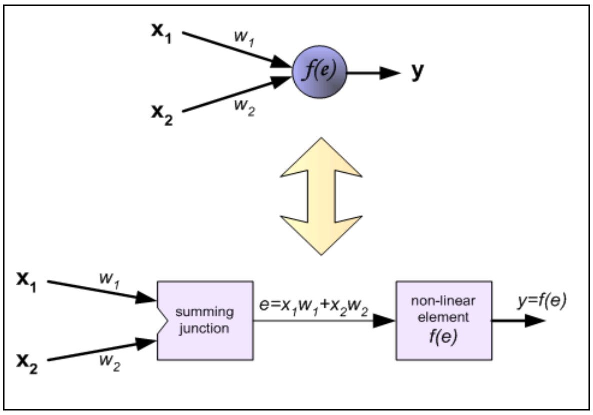

Figure 4: Teaching Process of Multi-Layer NN

The 106 DNA sequences composing the E. coli will feature for having 4 values. These values will stand for the A, T, G and C i.e. Adenine, Thymine, Guanine and Cytosine. Training the ANN and the DNA array with the 57 nucleotides attached to each promoter instance is coupled as an ANN input. The DNA sequence instances present the network output usually a description of either (+) or (-) occurrence.

III. Results And Discussion

The section attempts to evaluate the approach to promoter sequences posed by the experiment with much consideration to other approaches highlighted in the literature. Such endeavors are usually conducted in two phases that showcase a precise learning algorithm, training and testing. Training will involve establishing a classification model; testing entails the implementation of the classification model previously established.

A standard 5-fold cross-validation was integrated into the evaluation of the ANN performance by having the dataset being randomly portioned into 5 subsets. This classification ensures an equal ratio of (+) and (-) promoter locations in the DNA array.

The training occurred on the ANN for a series 5 times engaging only 4 subsets for each training while as retaining the remaining 5 for testing. As a result, 5 models were established during the cross-validation. Additionally, a final prediction performance was carried out on the subsets evaluating the average results from the experiment..

The performance of the promoter predictions was evaluated using the threshold parameters; accuracy (ACC), Mathew's Correlation Coefficient (MCC), sensitivity (SE) and specificity (SP). A couple of equations were integrated to affirm to the results. These were;

$$SE=TP/(TP+FN) \quad (1)$$

$$SP=TN/(TN+FP) \quad (2)$$

$$ACC= (TP+TN) / (TP+TN+FP+FN) \quad (3)$$

$$MCC=((TP*TN)-(FN*FP))/SQRT((TP+FN)*(TN+FP)*(TP+FP)*(TN+FN)) \quad (4)$$

TP is true positive (promoter predicted as promoter)
FN is false negative (promoter predicted as non- promoter)
TN is true negative (non- promoter predicted as non- promoter)
FP is false positive (non- promoter predicted promoter).

The detailed performance of modules in term of SE, SP, ACC and MCC is shown in the following tables.

The following tables attached herewith illustrate the performance of models with regard to the ACC, MCC, SE and SP parameters. In addition to the promoter prediction experiment, also the research put into consideration the testing of various structures of the ANN that were single layered, also investigating the 'logsig' transfer function and "trainrp", "trainscg", "traincgp" learning algorithms.

Table 1: The performance of ANN-1

Hidden Layer / The number of neuron	Transfer Function	Learning Algorithms	MCC	SE	SP	ACC
40-1	logsig	trainrp	0.69	0.75	0.92	0.84
40-1	logsig	traincgp	0.70	0.71	0.97	0.84
40-1	logsig	trainscg	0.66	0.72	0.92	0.82

Table 2: The performance of ANN-2

Hidden Layer / The number of neuron	Transfer Function	Learning Algorithms	MCC	SE	SP	ACC
75-1	logsig	trainrp	0.67	0.75	0.91	0.83
75-1	logsig	traincgp	0.67	0.75	0.91	0.83
75-1	logsig	trainscg	0.64	0.72	0.91	0.82

Table 3: The performance of ANN-3

Hidden Layer / The number of neuron	Transfer Function	Learning Algorithms	MCC	SE	SP	ACC
100-1	logsig	trainrp	0.62	0.75	0.86	0.81
100-1	logsig	traincgp	0.69	0.69	0.97	0.83
100-1	logsig	trainscg	0.65	0.71	0.92	0.82

A significant result is displayed by the overwhelming output of 0.84 (ACC) that had been given out as output by the ANN. They discovered hidden layer was found to possess a total of 40 neurons, 'logsig' transfer function and the trainrp algorithm. The paper in a special manner, recognizes the 'leave-one-out' methodology for evaluating the performance of the ANN that is earlier cited in the literature. This 'leave-one-out' cross-validation is a special case of k-fold cross –validation where k is an equivalent of the number of instances in the data set (34). This technique is widely used in Bioinformatics more so when there is a scarcity of data.

Table 4: The errors of some machine learning algorithms on promoter data set.

System	Errors	Classifier
REX-1	0/106	Inductive L.A
ANN	**0/106**	**One hidden layer**
IREM	2/106	Class-based entropy
KBANN	4/106	A hybrid ML system
BP	8/106	Standard backpropagation
O'Neill	12/106	Ad hoc tech. from the bio.
NearNeigh	13/106	A nearest neighbours
ID3	19/106	Quinlan's decision builder

This illustration clearly indicates that the promoter prediction in this research exceeds the performance of the promoter predictions there in the literature review. Surprisingly, the ANN in this research 'has proved to outperform the existing classifier for promoter prediction. The writer notes that even after the consideration of an occurrence of errors, this result was better than the BP, ID3, KB, NN and the O'Neill.

IV. Conclusions

We pose that promoter prediction and identification is an indispensable package in the Bioinformatics field, considering a digitalized approach. The ANN poses a great stride in this endeavor. Based on the structural and functional aspects of the ANN, there is the impact caused by the transmission of information through the network facilitating changes in the entire network based on the input and output. From these successful results, we note that an integration of ANN for promoter prediction transfers to great and appropriate results thereby providing for ground to endeavor much more in improving promoter prediction and identification

REFERENCES

[1] R. Kliman and L. Hoopes, Essentials of Cell Biology, Nature Education, 2010.

[2] C. Gabriela and M.-I. Bocicor, "Promoter Sequences Prediction Using Relational Association Rule Mining," *Evolutionary Bioinformatics,* vol. 8, pp. 181-196, 2012.

[3] J.-W. Huang, Promoter Prediction in DNA Sequences, Kaohsiung,: National Sun Yat-sen University, 2003.

[4] M. Guigo and R. Burset, "Evaluation of gene structure prediction programs," *Genomics,* vol. 3, no. 34, pp. 353-367, 1996.

[5] L. Milanesi, N. Kolchanov and I. Rogozin, "GenViewer: A computing tool for protein coding regions prediction in nucleotide sequences," in *the 2nd International Congress on Bioinformatics, Supercomputing and Complex Genome Analysis,*, 573-587, 1993.

[6] R. Guigo, S. Knudsen, N. Drake and T. Smith, "Prediction of gene structure," *Journal of Molecular Biology,* no. 226, pp. 141-157, 1995.

[7] S. Dong and G. Stormo, "Gene structure prediction by linguistic methods," *Genomics,* no. 23, pp. 540-551, 1994.

[8] E. Stormo and E. Snyder, "Identification of coding regions in genomic DNA sequences: An application of dynamic programming and neural networks," *Nucleic Acids Research,* no. 21, pp. 607-613, 1993.

[9] V. Solovyev, A. Salamov and C. Lawrence, "Prediction of internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames," *Nucleic Acids Research,* no. 22, pp. 5156-5163, 1994.

[10] G. Hayden and M. Hutchinson, "The prediction of exons through an analysis of spliceable open reading frames," *Nucleic Acids Research,* no. 20, pp. 3453-3462, 1992.

[11] A. Skolnick and M. Thomas, "A probabilistic model for detecting coding regions in DNA sequences," *IMA J. Math. Appl. Med. Biol.,* no. 11, pp. 149-160, 1992.

[12] Y. Xu, R. Mural and E. Uberbacher, "Constructing gene models from accurately predicted exons: An application of dynamic programming," *Comput. Appl. Biosci,* no. 10, pp. 613-623, 1994.

[13] J. Henderson, S. Salzberg and K. Fasman, "Finding genes in DNA with a hidden Markov model," *Journal of Computational Biology,* vol. 2, no. 4, pp. 127-141, 1997.

[14] C. Karlin and C. Burge, "Prediction of complete gene structures in human genomic DNA," *J. Mol. Biol,* no. 268:, pp. 78-94, 1997.

[15] M. Hrishikesh, S. Nitya and M. Krishna, "An ANN-GA model based promoter prediction in Arabidopsis thaliana using tilling microarray data," *Bioinformation,* vol. 6, no. 6, p. 240–243, 2011.

[16] J. Gordon, M. Towsey and J. Hogan, "Improved prediction of bacterial transcription start sites," *Bioinformatics,* vol. 22, no. 2, pp. 142-148, 2006.

[17] R. a. S. D. Liu, "Consensus promoter identification in the human genome utilizing expressed gene markers and gene modelling," *Genome Research,* no. 12, pp. 462-469, 2002.

[18] Q. Luo and W. a. L. P. Yang, "Promoter recognition based on the interpolated Markov chains optimized via simulated annealing and genetic algorithm," *Recognition Letters Pattern,* vol. 9, no. 27, pp. 1031-1036, 2006.

[19] C. Premalatha and C. a. K. K. Aravindan, "On improving the performance of promoter prediction classifier for eukaryotes using fuzzy based distribution balanced stratified method.," in *Proceedings of the International Conference on Advance in Computing, Control, and Telecommunication Technologies IEEE,*, ACT, 2009.

[20] T. S. Y. B. E. R. P. a. V. d. P. Y. Abeel, "Generic eukaryotic core promoter prediction using structural features of DNA," *Genome Research,* vol. 18, no. 2, pp. 310-323, 2008.

[21] Y.-J. Zhang, "A novel promoter prediction method inspiring by biological immune principles.," *Global Congress on Intelligent Systems,* no. 569-573, pp. 569-573, 2009.

[22] S. Clancy, "Nature Education," *DNA transcription,* vol. 1, no. 1, p. 41, 2008.

[23] A. Frank and A. Asuncion, "UCI machine learning repository," 2010. [Online]. Available: http://archive.ics.uci.edu/ml/.

[24] S. Nissen, Neural Networks Made Simple, US: Source Forge, 2005.

[25] D. G. Alverez, Artificial Neural Network, Spain: Edugila, 2006.

[26] D. Kriesel, A Brief Introduction to Neural Networks, US: Snipe, 2005.

[27] A. K. Jain, ANN, Michigan : Michigan State University, 1996.

[28] J. Courche, Terminal Axon Branching Is Regulated by the LKB1-NUAK1 Kinase Pathway via Presynaptic Mitochondrial Capture, US: Cell, 2013.

[29] J. Gandhi and S. Parekh, "Deployment of Neural Network on Multi-Core Architecture," *International Journal of Engineering Research & Technology (IJERT),* pp. 1-5, 2012.

[30] C. Gershenson, ANN for beginner, US: Arxiv, 2008.

[31] D. Rumelhart and J. McClelland, Parallel Distributed Processing, Cambridge: MIT Press, 1986.

[32] R. Rojas, Neural Networks: A Systematic Introduction, Berlin: Springer, 1996.

[33] W. A. Golda, "Principles of training multi-layer neural network using backpropagation," 2005. [Online]. Available: http://galaxy.agh.edu.pl/~vlsi/AI/backp_t_en/backprop.html.

[34] B. Efron, "Estimating the error rate of a prediction rule: improvement on cross-validation," *J. Am. Stat. Assoc.,* no. 78, p. 316–331, 1983.

[35] T. Abeel, Y. Saeys, E. Bonnet and P. Rouzé, "Generic eukaryotic core promoter prediction using structural features of DNA," *Genome Research,* vol. 18, no. 2, pp. 310-323, 2008.

[36] M. Wang, M. Yin and T. Jason, "GeneScout: a data mining system for predicting vertebrate genes in genomic DNA sequences," *Information Sciences,* vol. 163, no. Special issue, pp. 201-218, 2013.

[37] B. Óscar and B. Santiago, "Cnn-promoter, new consensus promoter prediction program," *Revista EIA,* no. 15, pp. 153-164, 2011.