Global Journal of Computer Science and Technology: E Network, Web & Security
Volume 14 Issue 1 Version 1.0 Year 2014
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

On the Investigation of Biological Phenomena through Computational Intelligence

By Jyotsana Pandey & Dr. Bipin Kumar Tripathi
Singhania University, India

Abstract- This paper is largely devoted for building a novel approach which is able to explain biological phenomena like splicing, promoter gene identification, disease and disorder identification, and to acquire and exploit biological data. This paper also presents an overview on the artificial neural network based computational intelligence technique to infer and analyze biological information from wide spectrum of complex problems .Bioinformatics and computational intelligence are new research area which integrates many core subjects such as chemistry, biology, medical science, mathematics, com-puter and information science. Since most of the problems in bioinformatics are inherently hard, ill defined and possesses overlapping boundaries. Neural networks have proved to be effective in solving those problems where conventional com-putation tools failed to provide solution. Our experiments demonstrate the endeavor of biological phenomena as an effec-tive description for many intelligent applications. Having a computational tool to predict genes and other meaningful in-formation is therefore of great value, and can save a lot of expensive and time consuming experiments for biologists. This paper will focus on issues related to design methodology comprising neural network to analyze biological information and investigate them for powerful applications.

Keywords: *splicing, promoter gene, bioinformatics, biological disorder, neural networks.*

GJCST-E Classification : *F.4.1*

Strictly as per the compliance and regulations of:

On the Investigation of Biological Phenomena through Computational Intelligence

Jyotsana Pandey [α] & Dr. Bipin Kumar Tripathi [σ]

Abstract- This paper is largely devoted for building a novel approach which is able to explain biological phenomena like splicing, promoter gene identification, disease and disorder identification, and to acquire and exploit biological data. This paper also presents an overview on the artificial neural network based computational intelligence technique to infer and analyze biological information from wide spectrum of complex problems .Bioinformatics and computational intelligence are new research area which integrates many core subjects such as chemistry, biology, medical science, mathematics, computer and information science. Since most of the problems in bioinformatics are inherently hard, ill defined and possesses overlapping boundaries. Neural networks have proved to be effective in solving those problems where conventional computation tools failed to provide solution. Our experiments demonstrate the endeavor of biological phenomena as an effec-tive description for many intelligent applications. Having a computational tool to predict genes and other meaningful information is therefore of great value, and can save a lot of expensive and time consuming experiments for biologists. This paper will focus on issues related to design methodology comprising neural network to analyze biological information and investigate them for powerful applications.

Keywords: splicing, promoter gene, bioinformatics, biological disorder, neural networks.

I. Introduction

The past few decades have seen a rapid growth in biologi-cal information that is coming in the form of genomes, protein sequences, gene expression, biological disorders data and many other medical diagnosis problems. There is the absolute need of effective and efficient computational tools to store, analyze and interpret the multifaceted data. The conventional techniques [1] of computational biology [2] involve the use of applied mathematics, informatics, statistics and biochemistry to solve biological problems usually on the molecular level. Major research efforts in the field include sequence alignment, gene finding, genome assembly, protein structure alignment, protein structure prediction, and prediction of gene expression, pro-tein-protein interactions and the modeling of evolution. All these problems need to deal with a huge amount of multi-faceted data. For example: there are approximately 26 billion base pairs (bp) representing the various genomes available on the server of the National Center for Biotech-nology Information (NCBI).

The computational biology [2] is concerned with the use of computation to understand biological phenomena and to acquire and exploit biological data, increasingly large-scale data [9]. Methods from computational biology are increa-singly used to augment or leverage traditional laboratory and observation-based biology. These methods have become critical in biology due to recent changes in our ability and determination to acquire massive biological data sets, and due to the ubiquitous, successful biological insights that have come from the exploitation of those data. This transformation from a data-poor to a data-rich field began with DNA sequence data, but is now occurring in many other areas of biology. The bioinformatics involve the crea-tion and advancement of algorithms using techniques in-cluding modern computer science, applied mathematics, statistics, and biochemistry. Hence, in other words, bioin-formatics can be described as the application of computa-tional methods to make biological discoveries [6].

The Computational intelligence [3] is now became a well-established paradigm for solving complex problems dealing with large scale data which are having overlapping, inexact and ill-defined boundaries. Now days, researchers are evolving new theories with a sound biological under-standing in solving problems of molecular and computa-tional biology [9]. They are able to perform a variety of tasks that are difficult or impossible to do with convention-al mathematics, statistics and informatics [12]. To name a few, Tasoulis et al. [10] introduced the application of neural networks, evolutionary algorithms and clustering algo-rithms to DNA microarray experimental data analysis; Liang and Kelemen [11] propose a time lagged recurrent neural network with trajectory learning for identifying and classifying gene functional patterns from the heteroge-neous nonlinear time series microarray experiments.

In this paper we are investigating the method and technique of machine learning through artificial neural networks which proved to be more suitable for genomic and other biological data analysis. The performance of the gene prediction approaches [4] mostly depends on the effectiveness of detecting the splice sites. This paper proposes a system for utilizing an artificial neural network [6] to addresses the problem

Author α: Dept. of Computer Science, Singhania University, Rajasthan, India. e-mail: imjyotsanapandey@gmail.com
Author σ: Dept. of Computer Science & Engineering, HBTI, Kanpur, India. e-mail: abkt.iitk@gmail.com

of splice site detection. ANN takes up it as a two-class problem and classifies a given sequence whether it will be a donor or an acceptor site. Further it predicts the splice form for a given sequence using the scores provided by the single site detectors for every appearing AG and GT dimer. The challenge is to find a splice form that consis-tently combines all predictions. The empirical analysis has further revealed that the results come out more refined if data analyzed in binary format as compared to other format. In the neural network structure, a standard three layer feed forward network of neurons is considered for analysis in which there are two neurons in output corres-ponding to the donor and acceptor splice sites, 128 neurons in hidden layer and 240 units at input end. The 240 input units were used since the orthogonal input scheme uses four inputs each nucleotide in the window.

To provide useful insights for neural network applications in biological information analysis, we structure the rest of the paper as follows: section 2 elaborates the related recent trends in biological information that is coming in the form of genomes [4], protein sequences, gene expression, bio-logical disorders data and medical diagnosis problems. Ar-tificial neural network technique involved in classification and recognition process is presented in section 3. Section 4 presents the empirical evaluation of different biological data analysis and experimental outcome. Finally, section 5 summarizes the paper with the inferences and discussions.

II. Recent Trends in Biological Information

Gene prediction [4] is a very powerful and important task for many ongoing researches in the field of bioinformatics [5]. A gene is a set of instruction which governs the assem-bly and function of all organisms. We know that a gene is a region of DNA that control a certain basic characteristic and ultimately lead to protein synthesis. In the 1990s ge-nomic data started becoming available. Since conventional mathematical models [8] proved to be unworkable in anal-ysis of biological information, bioinformaticians turned to computational intelligence [[9] models for help in tasks such as gene finding and protein structure prediction. The feature selection and class prediction, two learning tasks that are strictly paired in the search of molecular profiles from microarray data, were performed with ANN. The models with ANN have been shown to present a good choice, thus providing analysis and clues for biological information samples. Recently, proteomic data considered potentially rich, but arguably unexploited, for genome an-notation using ANNs which shows favorable performances as compared to conventional mathematical models. The idea of using manifold learning for feature reduction com-bined with an ANN classifier was successful applied in biomedical diagnosis and protein identification.

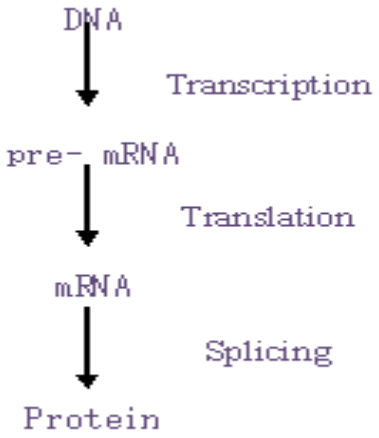

Figure 1 : It shows the importance of splicing which ul-timate cause for making of protein.

In this fig DNA expresses the gene product that it encodes. Figure demonstrates that certain region of the DNA is tran-scribed into RNA in the form of pre-mRNA. Further, the introns of the pre-mRNA are excised, leaving only exon intact to become the mature mRNA by translation. The ribosome then translates the mRNA into a polypeptide chain of amino acids that eventually becomes a protein by splicing [1]. In splice site prediction in E. coli gene DNA se-quences one need to identify the boundaries between exon (the part of DNA sequence retained after splicing) and in-trons (the part of DNA sequence that are spliced out) in given DNA gene sequence. Thus, this problem contains three classes. First is intron–exon (IE) boundary (donors), second is exon-intron (EI) (acceptors) and third class be-longs to neither donors nor acceptors (Neither).

DNA splice sites (Figure 2) are boundaries where splicing occurs and are found between the regions of DNA that code for gene products (exon) and those that do not (intron) [2]. The presence of introns in eukaryotic organisms are believed to be involved in exon shuffling (or alternative splicing) that is responsible for the higher diversity of gene.

Products found in eukaryotic organisms than that of prokaryotic organisms [3]. A typical example of exon shuffling is the generation of antibodies against foreign antigens that may invade the host system. The dinucleotide AG are splice sites that borders the transition from intron to exon (Intron/Exon border) going from 5' to 3', while GT are associated with the transition from exon to intron (Exon/Intron border). The GT dinucleotide is usually referred to as "donor" whereas the AG dinucleotide is known as "acceptor" [4].

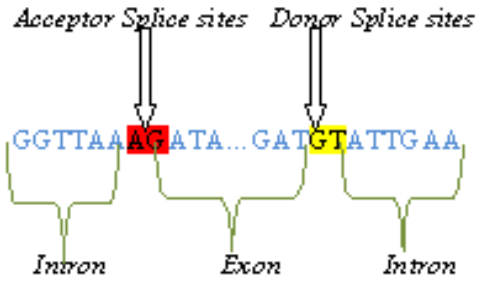

Figure 2 : Schematic representation of the splice site.

III. Neural Network for Biological in-Formation Analysis

Neural networks have several unique characteristics and advantages as tools for the molecular sequence analysis problem. A very important feature of these networks is their adaptive nature, where "learning by example" replaces conventional mathematical techniques which are time-consuming, computation extensive, and weak to noise. A small complexity, robust performance, and quick convergence of artificial neural network (ANN) are vital for its wide applicability. This feature makes such computational models [10] very appealing in application domains where one has little or incomplete understanding of the problem to be solved, but where training data are readily available. Owing to the large number of intercon-nections between their basic processing units, neural networks are error-tolerant, and can deal with noisy data. Neural network [12] architecture encodes information in a distributed fashion. This inherent parallelism makes it easy to optimize the network to deal with a large volume of data and to analyze numerous input parameters. Flexible encoding schemes can be used to combine heterogeneous sequence features for network input. Finally, a multilayer network is capable of capturing and discovering high-order correlations and relationships in input data. The artificial neural networks [13] are "neural" in the sense that they may have been inspired by neuroscience but not necessarily because they are faithful models of biological neural or cognitive phenomena.

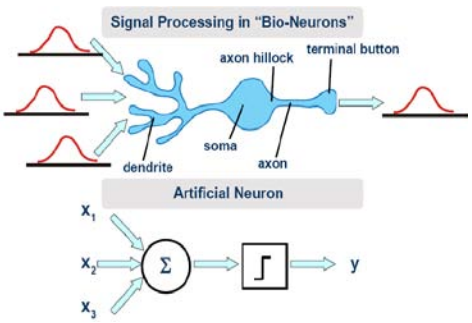

Figure 3 : Real biological neuron to artificial neuron

To enable understanding of neural networks, we start from a theoretical model of a single neuron and then briefly in-troduce a neural network to reveal their structure, training mechanism, operation, and functions. The basic structure of a biological neuron and corresponding artificial neuron is shown in Fig 3 and can be theoretically modeled as (1)

$$Y = f\left(\sum_{i=1}^{n} w_i x_i + b\right), \quad (1)$$

Where X {xi, i=1, 2, . . ., n} represent the inputs to the neu-ron and Y represents the output. Each input is multiplied by its weight wi, a bias b is associated with each neuron and their sum goes through a activation function f. A neural network is characterized by (1) its pattern of connections between the neurons (its architecture), (2) its method of determining the weights on the connections (training or learning, algorithm), and (3) its activation function.

In summary, the applications of ANNs in biological infor-mation processing have to be analyzed individually. ANN has been applied to biological data to deal with the issues that cannot be addressed by traditional algorithms or by other classification techniques. By introducing artificial neural networks, algorithms developed for processing and analysis often become more intelligent than conventional techniques. While neural networks are undoubtedly power-ful tools for classification, clustering and pattern recogni-tion; analysis of the internal weight and bias values for neurons in a network is possible, and a network itself can be represented formulaically, they are sometimes too large to be explained in a way that a human can easily under-stand. Despite this, they are still widely used in situations where a black-box solution is acceptable, and where em-pirical evidence of their accuracy is sufficient for testing and validation.

IV. Design of Learning Machine

It has been widely observed that in comparison to other machine learning approaches [3] neural networks have many positive characteristics for a prospective user. The variety of different network architectures and learning pa-radigms available, coupled with a theoretically limitless number of combinations of layers amounts, connections topologies, transfer functions and neuron amounts, make ANNs incredibly flexible processing tools. They can be applied to data with almost any number of inputs and out-puts, and are well supported in different programming lan-guages and software suites. Through manual modification of weights prior to training, and through imposing custom limitations on their modification during training, existing expert knowledge can be incorporated into their design and construction. Additionally, neural networks based learning machine are usually computationally inexpensive to use after they

have been trained, making them ideal for real-time applications where immediate output is desirable.

The neural networks used in this study (Fig. 3) are of the multi-layer neural network containing neurons of summa-tion aggregation function [13]. They are feed-forward con-nected and have three layers: an input layer, one hidden layer and an output layer. In case of gene prediction prob-lems, the network input is a segment of nucleotides from the nucleotide sequence. The output consists of one unit, giving a real valued output between 0.9 and 0.1. Using a threshold this number is interpreted as a category assign-ment for the nucleotide in the input window. The networks were trained by standard error back propagation learning algorithm on two different tasks: (i) detection of coding nucleotides (versus non-coding nucleotides), and (ii) the prediction of splice sites (defined as the first and last Intron, nucleotide respectively). Thus neural network unarguably possess strong potential for output prediction as can be seen by their widespread use in designing learning machine involving modeling and prediction.

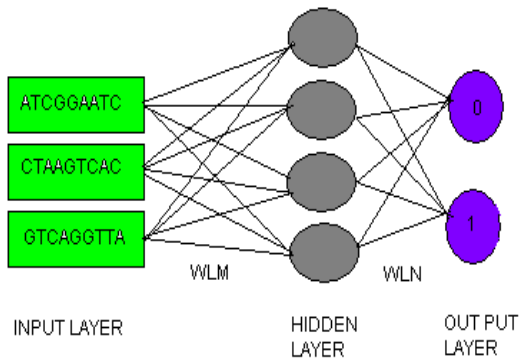

Figure 4 : Learning machine design with artificial neural network

V. Empirical Evaluation of Biological Data Analysis

In order to estimate the strength and effectiveness of bio-logical information from wide spectrum of problems, one needs to analyze them with standard computational intelli-gence technique. I have considered the artificial neural network to prove the motivation and to establish the signi-ficance of work done. Benchmark problems are standard enough to associate the tasks like classification and pattern recognition. They also incorporate the tasks from different fields of importance; few of them are biological engineering, medical and bioinformatics. In this section, I have tho-roughly evaluated kinds of problem to present the impor-tance of neural network in characterizing and analyzing the medical and biological information. In the present investi-gation, we use 5 datasets. Dataset containing primate splice-junction gene sequences and promoter gene se-quences were used in both normalized and binary forms. We observed better error convergence in binary form than the normalized form of dataset. The other datasets Heart Spectf, Bupa Liver Disorder and Protein Localization sites [15] are in numeric forms only, therefore sets normalized in prepro-cessing.

In all the experiments, I have divided whole da-taset into two parts: one is training set and second is test-ing set. Performance is analyzed in terms of parameters which are briefly defined as follows:

Training Accuracy (%)

$$= 100 \times \frac{\text{Number of correct matches}}{\text{Total number of samples in training set}}$$

Testing Accuracy (%)

$$= 100 \times \frac{\text{Number of correct matches}}{\text{Total number of samples in test set}}$$

a) Datasets and Significance

In this paper we have used two genomic datasets and three biological disorders data sets. All these data stets are benchmark and available online for research and academic purposes.

1. E. coli promoter gene sequences (DNA) [11] is acquired to predict the member/non-member of class of sequences with biological promoter activity. The dataset contains non-numeric domain of attributes. The attributes are one of the 'a', 'g','t' and 'c' (a=Adenine, b=Guanine, t=Thymine and c=Cytosine). This dataset have been also used by Harley, C. and Reynolds, R. 1987 in "Analysis of E. Coli Promoter Sequences" Nucleic Acids Research.
2. Primate splice-junctions are the points on DNA sequence at which superfluous DNA is removed during the process of protein creation in higher organisms. The splice-junction gene problem is to identify the boundaries between exons and introns in given DNA gene sequence.
3. Heart SPECTF Data set [14] is based on cardiac single proton emission computed tomography (SPECT) images. Each patient is classified in normal or abnormal categories. Database was used in automated Cardiac SPECT Diagnosis.
4. BUPA liver disorders dataset contains 345 instances that are basically records of 345 males who have taken excessive alcohol consumption. The first 5 attributes are all blood tests which are on thought to be sensitive to liver disorders that might arise from excessive alcohol consumption. The last 2 attributes are different from blood tests. One of them is no of drink having taken in a day and other is selector field.
5. Protein localization sites dataset [15] can be achieve from "Expert System for Predicting Protein

Localization Sites in Gram-Negative Bacteria", Kenta Nakai & Minoru Kanehisa, PROTEINS: Structure, Function, and Genetics 11:95-110, 1991.

b) Learning Machine with Benchmark Datasets

i. *Splice site prediction in E. coli gene DNA sequences*

In E. coli promoter gene sequences (DNA) dataset from University of Wisconsin Biochemistry Department, there are 106 instances with 59 attributes. In these 59 attribute One of {+/-}, indicating the class ("+" = promoter) and second 2-60 remaining 59 fields are the sequence, starting at position filled by one of {a, g, t, c} base pairs. There is no missing attribute Values. In Class attribute there are 53 positive instances and 53 negative instances.

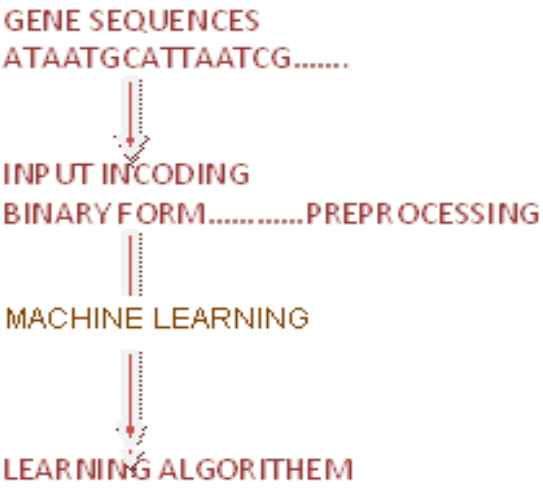

Figure 5 : Processing of biological dataset (binary form) through ANN.

At first step we get DNA sequences to analyze splice sites at second step we used sparse encoding to encode these sequences that is A as (1000), C as (0100), G as (0010) and T as (0001) for preprocessing. It is used to avoid alge-braic dependencies between nucleotides in the encoding also called BIN4 encoding in which each letter coded by four digits with the combination of 0 and 1 to input data. The learning in the neural network is done with error back propagation method and result is presented in Table 1 with discussion in section 6.

ii. *Primate splice-junction gene sequences*

In this dataset all examples taken from Gen bank 64.1 (ftp site: genbank.bio.net), there are three categories "ei", "ie" and "n" for splice sites recognition. Dataset contains 3190 instances including three classes. Class 'EI' as donor con-sists of 767 instances, class 'IE' as accepter consists of 768 instances and rest of as 'N' neither belongs to any class consists of 1655 instances. In this dataset containing pri-mate splice junction gene sequences (DNA), the standard result was 85% and we achieved the accuracy of 81% using binary form of the dataset and 79% of accuracy was achieved when we used the normalized form of dataset. Result is presented in Table 1 with discussion in section 6 for back propagation neural network.

iii. *Heart Spectf Dataset*

This dataset describes diagnosing of cardiac Single Proton Emission Computed Tomography (SPECT) images. This can be achieved from University of Colorado at Denver, Denver, CO 80217, u.s.a.krys.cios@cudenver.edu. Data-base used by Kurgan, L.A., Cios, K.J., Tadeusiewicz, R., Ogiela, M. & Goodenday, L.S. "Knowledge Discovery Ap-proach to Automated Cardiac SPECT Diagnosis" Artificial Intelligence in Medicine, vol. 23:2, pp 149-169, Oct 2001. There are 267 instances as SPECT image sets for patients. Each of the patients is classified into two categories: nor-mal (0) and abnormal (1). Database contains 23 attributes in which 22 spectf image + 1 class. All dataset is divided into training data with 80 instances and testing data with 187 instances. Class 0 consists of 55 instances and class 1 consists of 212 instances. In dataset containing SPECTF heart data the standard result was an accuracy of 87% and we achieve an accuracy of 84% as shown in Table 1.

iv. *BUPA liver disorders data set*

This dataset achieved from "Expert Sytem for Predicting Protein Localization Sites in Gram-Negative Bacteria", Kenta Nakai & Minoru Kanehisa, PROTEINS: Structure, Function, and Genetics 11:95-110, 1991. There are 336 instances each with 8 attributes one of them is name and other are predictive. In this 8 classes are according to the protein location in bacteria. In protein localization dataset containing the standard result was an accuracy of 84% with ad hoc structured probability model; but, we found an ac-curacy of 81% with the artificial neural network.

VI. Inferences and Discussion

Research in bioinformatics is driven by the expe-rimental data. Current biological databases are populated by vast amounts of experimental data. Machine learning has been widely applied to bioinformatics and has gained a lot of success in this research area. At present, with various learning algorithms available in the literature [16], researchers are facing difficulties in choosing the best me-thod that can apply to their data. We performed an empiri-cal study and observed that single learning networks are perfectly usable in splice site prediction, gene prediction, liver disorders and localization site in the same manner. The performance of the learning technique is highly de-pendent on the nature of the training data or on the basis of dataset design. We

conclude that, if dataset is in norma-lized form as well as in binary, then the best results can be achieved.

In the following Table 1, the dataset containing promoter gene sequences (DNA) we achieved the accuracy of 85% using binary form of the dataset and 83% of accu-racy was achieved when we used the normalized form of normal dataset. We can infer that the variation in the results due to the different forms of dataset was because in binary form the variables A-T-G-C are converted into orthogonal vectors. The dataset containing primate splice junction gene sequences (DNA), we achieved the accuracy of 81% using binary form of the dataset and 79% of accuracy was achieved when we used the normalized form of dataset. In the past usage as results of study indicate that machine learning techniques (neural networks, nearest neighbor, contributors' KBANN system) have performed as well/ better than classification based on canonical pattern matching. In dataset containing SPECTF heart data we achieve an accuracy of 84%. In protein localization dataset we found an accuracy of 81% with the artificial neural network. In dataset containing BUPA Liver disorders we found a mediocre accuracy of 77%. We conclude that we achieved the said accuracies with most straight forward and convenient technique which does not possess the complicated computing operations. There may be techniques which may yield little more accuracy for corresponding dataset but our technique is computationally efficient.

We can see the overall analysis at a glance in the Table 1:

Table 1 : Analysis of all datasets with accuracy using back propagation

SNo.	Name of database	No of instances	No of attribute	Training accuracy	Testing accuracy
1	Promoter gene	106	59	85%	83%
2	Primate splice-junction gene sequence	3190	1595	81%	85%
3	SPECTF-heart data	267	23	87%	84%
4	BUPA-liver disorders	345	7	77%	80%
5	Protein Localization Sites.	336	8	81%	84%

VII. Acknowledgements

I want to pay my sincere thanks to all of them who are related directly and indirectly with my work and reviewed & encouraged all the time.

References Références Referencias

1. Krogh A, Brown M, Mian IS, K Sjölander of molecular biology, – Elsevier 1994.
2. Computational systems biology, H Kitano - Nature, 2002.
3. JSR Jang, CT Sun, E Mizutani, -"Neuro-fuzzy and soft computing-a computational approach to learning and machine intelligence Automatic Control", IEEE …, ieeexplore.ieee.org 1997.
4. Pati, NN Ivanova, N Mikhailova, G Ovchinnikova Gene-PRIMP:" A gene prediction improvement pipeline for prokaryotic genomes" Nature …, 2010
5. Pierre Baldi, Søren Brunak Bioinformatics: "The Machine Learning Approach", Second Edition (Adaptive Computation and Machine Learning) on Amazon.com. 2002.
6. Zoheir Ezziane, "Applications of artificial intelligence in bioinformatics" 2006.
7. Mitra S and Hayashi Y, "Bioinformatics with soft computing", IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, Vol. 36, pp. 616-635 2006..
8. Tasoulis D.K., Plagianakos V. P., and Vrahatis M. N. Com-putational Intelligence Algorithms and DNA Microarrays, Studies in Computational Intelligence (SCI) 94, pp. 1-31. 2008.
9. Tasoulis D.K., Plagianakos V. P., and Vrahatis M. N. Com-putational Intelligence Algorithms and DNA Microarrays, Studies in Computational Intelligence(SCI) 94, pp. 1-31. 2008.
10. Arpad Kelemen Ajith Abraham Yuehui Chen (Eds.) "Com-putational Intelligence in Bioinformatics. Studies in Com-putational Intelligence", Springer-Verlag Berlin Heidelberg, 2008.
11. C. and Reynolds, R. "Analysis of E. Coli Promoter Sequences" Nucleic Acids Research, 15: 2343-2361, 1987.
12. Wu CH, McLarty JW. "Neural Networks and Genome Informatics". Methods in computational Biology and Biochemistry, 1 ed. Vol. 1. Elsevier; 2000.
13. Reese MG, Eeckman FH. "Novel Neural Network Prediction. Systems for Human Promoters And. Splice Sites". 1995.
14. Mariofanna G. Milanova3, Tomasz G. Smolinski4, 20. Ogiela, M. & Goodenday, L.S. "Knowledge Discovery Approach to Automated Cardiac SPECT Diagnosis" Artificial Intelligence in Medicine, vol. 23:2, pp 149-169, Oct 2001.
15. Kenta Nakai & Minoru Kanehisa, "Expert Sytem for Predicting Protein Localization Sites in Gram-

Negative Bacteria", PROTEINS: Structure, Function, and Genetics 11: 95-110, 1991.

16. Aik Choon TAN and David GILBERT, “An empirical com-parison of supervised machine learning techniques in bioin-formatics” Bioinformatics Research Centre, Department of Computing Science 12 Lilybank Gardens, University of Glasgow, Glasgow G12 8QQ, UK 2009.

Global Journals Inc. (US) Guidelines Handbook 2014

www.GlobalJournals.org