International Journal for Biotechnology and Molecular Biology Research Vol. 2(4), pp. 61-72, May 2011
Available online http://www.academicjournals.org/IJBMBR
ISSN 2141-2154 ©2011 Academic Journals

Full Length Research Paper

Back propagation neural network method for predicting Lac gene structures in *Streptococcus pyogenes* M Group A *Streptococcus* strains

Santhosh Rebello, Uma Maheshwari*, Safreena and Rashmi Venex DSouza

Department of Bioinformatics, Aloysius Institute of Management and Information Technology, St. Aloysius College (Autonomous), 2nd Cross, Sharada Nagar, Beeri, Kotekar Post, Madoor, Mangalore -575022, Karnataka, India.

Accepted 3 March, 2011

The rapid growth and availability of whole genome sequences of *Streptococcus pyogenes* M Group A *Streptococcus* strains which is a spherical gram-positive bacteria that causes important human diseases ranging from mild superficial skin infections to life-threatening systemic diseases have initiated the need to analyze these sequences. The motivation of this paper is to adopt content based gene prediction method along with the machine learning techniques of Artificial Neural Networks - Back Propagation Network, specific for predicting Lac genes of *S. pyogenes* M Group A *Streptococcus* strains. We first obtained Lac genes from the genome sequences of *S. pyogenes* M Group A *Streptococcus* strains and calculated the mean gene content. The mean gene content had 70 parameters indicating the mean of the percentages of the frequencies of occurrences of 64 possible codons, 4 nucleotides, purines and pyrimidines. We constructed three–layer feed-forward neural network with 70 input units, 20 hidden units and 1 output unit. After being trained in a supervised manner with the Error Back–Propagation Algorithm by mean gene content, the network is examined by testing the algorithm for the mean gene content vector and 99 sample Lac gene vectors to get a range of values for the output that the Lac gene vector falls. The values obtained ranged from 0.9857 to 0.9901 and these ranges of values are used in classifying whether a given sequence is a Lac gene or not. SpyMGASLacGenePred is a tool that has been developed. It accepts a DNA sequence. It finds all possible ORFs in 6 reading frames. It calculates gene content and runs the testing algorithm of the network for all ORFs to confirm whether they are Lac genes or not. For the set of ORFs that the neural network classifies as a Lac gene, the tool determines and displays the position, length, frame information, GC content and translated sequence. The calculated performance measures for evaluation of the developed tool SpyMGASLacGenePred showed that it has a sensitivity of 100% and specificity of 76.9%. Since every Lac gene used for training is taken into consideration by the Back Propagation Neural Network program for testing, the tool has 100% sensitivity. However, if Lac genes of the other strains of *S. pyogenes* which are not used for training is tested, then sensitivity might drop to a certain extent. The tool has a specificity of 76.9% and this indicates that the tool is above an acceptable threshold level to predict the correct Lac genes out of a total of Lac genes. The tool also showed a correlation coefficient of 0.733 which is near +1 and thus can be considered as near perfect prediction. Thus the adopted Back Propagation Algorithm of Artificial Neural Network method has been useful for the development of the SpyMGASLacGenePred tool to identify the Lac gene structures in *S. pyogenes* M Group A *Streptococcus* strains.

Key words: Back Propagation Algorithm, Artificial Neural Network, Lac gene prediction, *Streptococcus pyogenes* M Group A *Streptococcus* strains.

INTRODUCTION

With the development of sequence analysis concepts, gene finding has become more and more important in bioinformatics (Mathe et al., 2002; Li et al., 2003). Gene finding typically refers to the area of computational biology (Xu et al., 1998) and is concerned with algorithmically identifying stretches of sequence, usually genomic DNA, that are biologically functional (Burge et al., 1998). This especially includes protein-coding genes, but may also include other functional elements (Guigo et al., 1992) such as RNA genes and regulatory regions. Gene finding is one of the first and most important steps in understanding the genome of a species once it has been sequenced (Milanesi, 1993; Guigo, 1997). Gene finding tools are mainly based on any of these three methods.

Content based method: Coding regions must contain triplet codons and the non coding regions are not of this nature. Coding regions have specific gene content and gene frequencies. Based on the content and frequency information of the gene, it’s possible to predict genes.

Signal based method: Coding regions have signals, signatures or patterns that accompany them. Based on these signals, it is possible to predict genes.

Homology method: Un-annotated sequence may have a homologous sequence in the database. Based on the homologous sequence identified, it is possible to predict genes.

Streptococcus pyogenes M Group A *Streptococcus* strains are one of the most common and versatile human bacterial pathogens. They cause a variety of diseases ranging from mild and quite frequent non-invasive infections of the upper respiratory tract and skin to severe invasive infections that include necrotizing fasciitis and streptococcal toxic shock syndrome (Facklam, 2002). They are also associated with such life-threatening post streptococcal sequelae as acute rheumatic fever and glomerulonephritis. Vaccines are not available currently to protect against its infection but a specific protective antibody has been shown to persist as long as 45 years after the original infection (Bencivenga et al., 2009). Evaluation of epidemiologic relationships between Group A *Streptococcus* isolates was based on serological typing (Lancefield, 1928; Lancefield et al., 1946; Mora et al., 2005) for detection of a bacterial cell surface M protein which is considered a major virulence factor of these microorganisms for a number of years. Recently, several DNA-based typing methods have been applied to evaluate the diversity of Group A Streptococcus isolates and to elucidate their association with different diseases (Facklam et al., 1999; McGregor et al., 2004; Doktor et al., 2005). A major improvement to serological M typing is Emm Gene Sequencing Analysis (Beall et al., 1997; Teixeira et al., 2001) which along with new typing methods will improve strain differentiation and contribute new insights into the epidemiology and pathogenesis of Group A *Streptococcus* infections. This in turn, will lead to rapid and precise detection of the microorganism so that treatment and prevention of Group A *Streptococcus* diseases will be made possible in an efficient manner (Currie, 2006).

The complete genome sequences of few *S. pyogenes* M Group A *Streptococcus* strains are available in the Genome Database. These genomes available in the Genome Database represent their complete set of DNA. However, only the fragments of these genomes are responsible for the functioning of their cell. These fragments are called genes which form the basic physical and functional units of heredity. Genes are made up of a contiguous set of codons, each of which specifies an amino acid. Genes translate into proteins and these proteins perform most life functions and even make up the majority of cellular structures.

S. pyogenes M Group A *Streptococcus* strains being a bacterial organism, adapt to changes in their surroundings by using regulatory proteins to turn groups of genes on and off in response to various environmental signals. Their DNA is sufficient to encode thousands of proteins but only a fraction of these are made at any one time. The expressions of many of their genes are regulated by means of operons which are a cluster of genes along with an adjacent promoter that controls the transcription of their genes according to the food sources that are available to them.

Lac operon is one such operon required for the transport and metabolism of lactose in bacteria. It consists of three structural genes, a promoter, terminator, regulator, and an operator. The three structural genes are: LacZ, LacY, and LacA. LacZ encodes β-galactosidase (LacZ), an intracellular enzyme that cleaves the disaccharide lactose into glucose and galactose. LacY encodes β-galactoside permease (LacY), a membrane-bound transport protein that pumps lactose into the cell. LacA encodes β-galactoside transacetylase (LacA), an enzyme that transfers an acetyl group from acetyl-CoA to β-galactosides. The regulatory gene LacI produces an mRNA that produces a Lac repressor protein, which can bind to the operator of the Lac operon. In the absence of lactose, the Lac repressor binds to the operator and keeps RNA polymerase from transcribing the Lac genes. When lactose is present, the Lac genes are expressed because allolactose binds to the Lac repressor protein and keeps it from binding to the Lac operator. Allolactose is called an inducer because it turns on, or induces the expression of the Lac genes. When the enzymes encoded by the Lac operon are produced, they break down lactose and allolactose eventually releasing the repressor to stop additional synthesis of Lac mRNA. When both glucose and lactose are available, the genes for lactose metabolism are transcribed at only low levels since glucose is the preferred and most frequently available energy source for them. When the supply of glucose has been exhausted

*Corresponding author. E-mail: ugdreams@gmail.com

Lac genes are transcribed efficiently which allows them to metabolize lactose. Maximal transcription of the Lac operon occurs when glucose is absent and lactose is present. The actions of cyclic AMP and a catabolite activator protein produce this effect. Cyclic AMP is derived from ATP. In the presence of lactose and absence of glucose, cyclic AMP joins with a catabolite activator protein that binds to the Lac promoter and facilitates the transcription of the Lac operon.

With the advent of whole-genome sequencing projects of *S. pyogenes* M Group A *Streptococcus* strains there is considerable use for programs that scan genomic DNA sequences to find genes. Gene prediction has become more and more important as the DNA of *S. pyogenes* M Group A *Streptococcus* strains are sequenced. DNA sequences submitted to databases are often already characterized and mapped when they are submitted. This means that a molecular biologist has already used genetics and biochemical methods to find genes, promoters, exons and other meaningful subsequences in the submitted material. However, the numbers of sequencing projects of *S. pyogenes* strains are increasing, and a lot of DNA sequences have not yet been mapped or characterized.

Having a computational tool to predict genes and other meaningful subsequences is therefore of great value, and can save a lot of expensive and time consuming experiments for biologists. Content based gene prediction is based on the idea that the unknown genes have similar statistical properties to the known ones (Zhang et al., 2000; Zhang et al., 2002) and neural networks have the capability to predict genes efficiently based on the trained statistical properties.

The present work aims at developing a tool named SpyMGASLacGenePred using content based gene prediction method along with the machine learning techniques of neural networks to capture the gene content and to predict and recognize the Lac genes of *S. pyogenes* M Group A *Streptococcus* strains.

MATERIALS AND METHODS

NCBI Genome Database

The National Center for Biotechnology Information (NCBI) is a part of the United States National Library of Medicine (NLM), a branch of the National Institute of Health. The NCBI is located in Bethesda, Maryland and was founded in 1988 through legislation sponsored by Senator Claude Pepper. The NCBI houses genome sequencing data in GenBank. The NCBI Entrez Genome database is a collection of complete large-scale sequencing, assembly, annotation, and mapping projects for cellular organisms.

Retrieval of Lac genes

The Genbank files for complete genome sequence of *S. pyogenes* M Group A *Streptococcus* strains are downloaded from the Genome Database of NCBI. Then 99 Lac genes are traced from downloaded strains of *S. pyogenes* M Group A *Streptococcus* strains namely MGAS 6180, MGAS 10750, MGAS 2096, M1 GAS, MGAS 5005, MGAS 315, MGAS 9429, MGAS 10270 with accessions CP000056, CP000262, CP000261, AE004092, CP000017, AE014074, CP000259 and CP000260 respectively.

Content based gene prediction method

DNA sequences that encode protein are not random chains of available codons for an amino acid, but rather, an ordered list of specific codons that reflect the evolutionary origin of the gene and constraints associated with gene expression. This non random property of coding sequences can be used as an advantage for finding regions in DNA sequences that encode proteins (Ficket, 1982). Each species also has a characteristic pattern of use of synonymous codons. Also there is a strong preference for certain codon pairs within a coding region (Ficket, 1998). Thus genes can be characterized based on its content and thereby, use the gene content information in predicting the genes.

Determination of Lac gene content

Frequency calculation of a codon for a sequence is done by counting the number of occurrences of that codon divided by total number of codons in that sequence. Frequency calculation of a nucleotide for a sequence is done by counting the number of occurrences of that nucleotide divided by total number of nucleotides in that sequence. Frequency calculation of A and T for a sequence is done by counting the number of occurrences of A and T divided by total number of nucleotides in that sequence. Frequency calculation of G and C for a sequence is done by counting the number of occurrences of G and C divided by total number of nucleotides in that sequence.

Thus frequencies of occurrence of all possible 64 codons, 4 nucleotides (A, T, G and C) and chemically similar nucleotides (A, T and G, C) altogether amounting to 70 parameters are calculated for all the 99 Lac genes which have been traced from the Genbank files of *Streptococcus pyogenes* M Group A Streptococcus strains that are downloaded from Genome Database of NCBI. The calculated frequencies for all the 70 parameters of 99 Lac genes are multiplied by 100 and converted into percentages. The mean of the percentages are calculated and the mean vector is used as an input vector for training the network.

Artificial neural network

An artificial neural network is a mathematical model or computational model that simulates the structure and functional aspects of biological neural networks (Bishop, 1995; Haykin, 2001). It consists of an interconnected group of artificial neurons and processes information using a connectionist approach to computation (Hertz et al., 1991).

It is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase (Anderson et al., 1988; Basheer et al., 2000). They are usually used to model complex relationships between inputs and outputs or to find patterns in data (Bishop, 2006; Duda et al., 2000).

Back propagation network

Back propagation is a systematic method that uses gradient-descent based delta learning rule also known as back propagation rule for training multilayer feed forward artificial neural networks

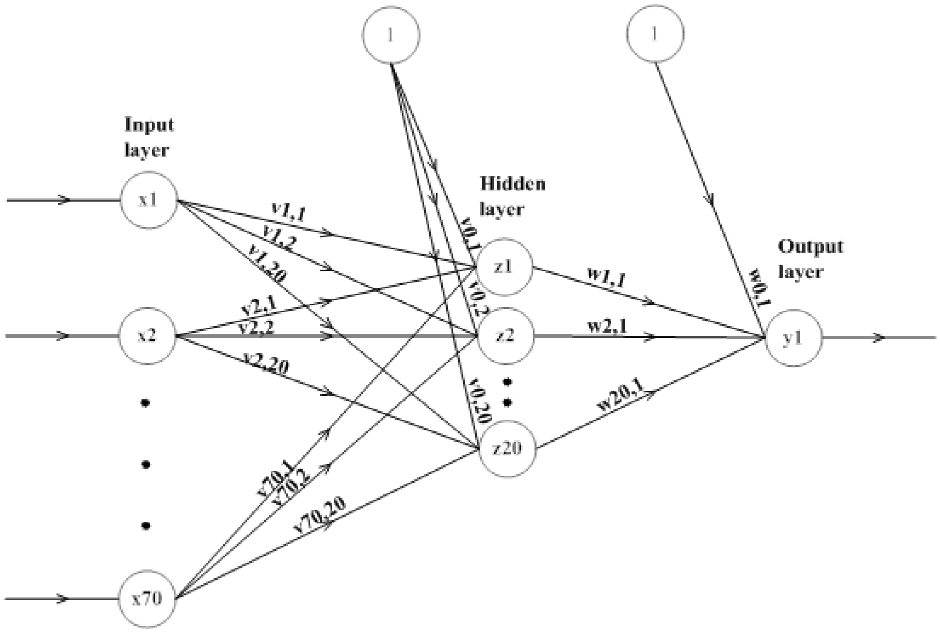


Figure 1. Back propagation network architecture.

(Russell et al., 2003). It also provides a computationally efficient method for changing the weights in the feed forward network with differentiable activation function units to learn a set of input-output patterns. Being a gradient descent method, it minimizes the total squared error of the output computed by the network. The network is trained by a supervised learning method and aims at achieving balance between the ability to respond correctly to the input patterns that are used for training and the ability to provide good responses to the input patterns that are similar.

Back propagation network architecture

The determined gene content of Lac genes are used as input patterns for training the Back Propagation Network. The Back Propagation Network design is a three layer network (Figure 1) with one of each input, hidden and output layers. The input layer has 70 input nodes, out of which 64 correspond to all possible codons, 4 correspond to the nucleotides themselves (i.e. A, T, G and C), and two for chemically similar nulceotides - A, T and G, C. The number of nodes in the hidden layer is randomly set to 20, and the number of output layer nodes is set to 1. Learning rate and initial weights and biases are taken as small random values. The weight matrix and bias matrix for connections between input and the hidden layer is a 70 × 20 matrix and 1 × 20 matrix respectively. The weight and bias matrices for connections between the hidden and output layers is 20 × 1 and 1 × 1 matrices respectively. The activation function used is binary sigmoidal activation function which is given by;

$$f(x) = \frac{1}{1 + e^{-x}}$$

The training algorithm is composed of the following four phases: initialization of weights, feed forward, back propagation of errors, and updation of the weights and biases.

Training algorithm

Initialization of weights (Phase I)

Step 0: Weights, biases and learning rate are initialized to small random values.
Step 1: Steps 2 to 9 are performed when stopping condition is false.
Step 2: Steps 3 to 8 are performed for each training pair.

Feed forward phase (Phase II)

Step 3: Each input unit received input signal x_i and sent it to the hidden unit (i = 1 to n).
Step 4: Each hidden unit z_j (j = 1 to p) summed its weighted input signals to calculate net input:

$$Z_{inj} = v_{0j} + \sum_{i=1}^{n} x_i v_{ij}$$

The output of the hidden unit is calculated by applying its activation functions over z_{inj} (binary sigmoidal activation function)

$$Z_j = f(z_{inj})$$

and sent the output signal form the hidden unit to the input of the output layer units.

Step 5: For each output unit y_k (k = 1 to m) the net input is calculated as $y_{ink} = w_{0k} + \sum_{j=1}^{p} z_j w_{jk}$ and the activation function is applied to compute the output signal, $y_k = f(y_{ink})$.

Back propagation of error (Phase III)

Step 6: Each output unit y_k (k = 1 to m) received a target pattern corresponding to the input training pattern and computed the error correction term $\delta_k = (t_k - y_k)\ f'\ (y_{ink})$ On the basis of the calculated error correction term, the change in weights and bias are updated:

$$\Delta w_{jk} = \alpha\ \delta_k\ z_j$$
$$\Delta w_{0k} = \alpha\ \delta_k$$

Also, δ_k is sent to the hidden layer backwards.

Step 7: Each hidden unit (z_j, j = 1 to p) summed its delta inputs from the output units $\delta_{inj} = \sum_{k=1}^{m} \delta_k\ w_{jk}$ The term δ_{inj} got multiplied with the derivative of f (z_{inj}) to calculate the error term, $\delta_j = \delta_{inj}\ f'\ (z_{inj})$. On the basis of δ_j, changes in weights and bias are updated thus:

$$\Delta v_{ij} = \alpha\ \delta_j\ x_i$$
$$\Delta v_{0j} = \alpha\ \delta_j$$

Weight and bias updation (Phase IV)

Step 8: Each output unit (y_k, k = 1 to m) updates the bias and weights:

$$w_{jk}\ (\text{new}) = w_{jk}\ (\text{old}) + \Delta\ w_{jk}$$
$$w_{0k}\ (\text{new}) = w_{0k}\ (\text{old}) + \Delta\ w_{0k}$$

Each hidden unit (z_j, j = 1 to p) updates its bias and weights:

$$v_{ij}\ (\text{new}) = v_{ij}\ (\text{old}) + \Delta\ v_{ij}$$
$$v_{0j}\ (\text{new}) = v_{0j}\ (\text{old}) + \Delta\ v_{0j}$$

Step 9: Check for the stopping condition. The stopping condition is when the actual output almost equaled the target output.

The training algorithm is written in MATLAB, a high-level technical computing language that uses an interactive environment for algorithm development. The weights and biases obtained from the training algorithm are used for initializing weights and biases in the testing algorithm. The testing algorithm is composed of two phases: initialization of weights and feed forward.

Testing algorithm

Initialization of weights (Phase I)

Step 0: Weights and biases are initialized. The weights and biases are taken from the training algorithm.

Step 1: Steps 2 to 4 are performed for each input vector.

Feed forward phase (Phase II)

Step 2: Set the activation of input unit x_i (i = 1 to n).

Step 3: Calculate the net input to the hidden unit x and its output.

For j = 1 to p,

$$Z_{inj} = v_{0j} + \sum_{i=1}^{n} x_i\ v_{ij}$$

$$Z_j = f\ (z_{inj})$$

Step 4: Compute the output of the output layer unit. For k = 1 to m,

$$y_{ink} = w_{0k} + \sum_{j=1}^{p} z_j\ w_{jk}$$

$$y_k = f\ (y_{ink})$$

Use sigmoidal activation functions for calculating the output.

Testing has been done for the mean vector which is used for training. Then the testing was performed for all the 99 Lac gene vector frequencies in order to get a range of values for the output that the Lac gene vector falls. These ranges of values are used in classifying whether a given sequence is a Lac gene or not.

Tool development

SpyMGASLacGenePred is a tool that has been developed using PERL and CGI. CGI program is stored and executed on the server, in response to a request from a client and PERL is an open source programming language that is used for creating CGI scripts. The tool accepts a DNA sequence and scans the accepted DNA sequence to identify Open Reading Frames (ORFs) in all six reading frames. For all the identified ORFs, it calculates the gene content that is the frequencies of occurrence of all possible 64 codons, 4 nucleotides (A, T, G and C) and chemically similar nucleotides –A, T and G, C altogether making up 70 parameters. It converts the calculated frequencies into percentages by multiplying the frequencies by 100 and these 70 parameters are provided as inputs to the 70 nodes of the input layer of the testing algorithm. Weights and biases are taken from the training algorithm and run on the testing algorithm of Back Propagation Network in order to determine whether the ORFs belong to the Lac gene category or not. If the network predicts ORFs as a Lac gene, the length and GC content of the sequence is determined and displayed along with the start and end positions, lengths of the ORF, frame information, score values and translated sequences of the ORFs. Cross validation tests such as calculation of sensitivity, specificity and correlation coefficient measures have been performed for performance evaluation of the tool.

RESULTS AND DISCUSSION

Retrieved genomes and genes

The genomes of *S. pyogenes* M Group A *Streptococcus* strains namely MGAS 6180, MGAS10750, MGAS 2096, M1 GAS, MGAS 5005, MGAS 315, MGAS 9429, MGAS 10270 with accessions CP000056, CP000262, CP000261, AE004092, CP000017, AE014074, CP000259 and CP000260, respectively (Table 1) have been downloaded from the NCBI Genome Database.

Table 1. Genome information of the *Streptococcus pyogenes* M Group A Streptococcus strains.

S/No.	Strain name	Accession	Length (Nt)	GC Content (%)	Coding (%)	Gene	Protein coding	Structural RNAs
1	MGAS10750	CP000262	1,937,111	38	86	2060	1979	81
2	MGAS2096	CP000261	1,860,355	38	86	1979	1898	81
3	MGAS10270	CP000260	1,928,252	38	86	2067	1986	81
4	MGAS315	AE014074	1,900,521	38	85	1951	1865	86
5	M1 GAS	AE004092	1,852,441	38	83	1810	1696	79
6	MGAS 5005	CP000017	1,838,554	38	86	1950	1865	85
7	MGAS6180	CP000056	1,897,573	38	86	1977	1894	83
8	MGAS9429	CP000259	1,836,467	38	87	1962	1877	85

They have nucleotide sequences ranging from 1,836,467 to 1,937,111 nucleotides. They are circular and double stranded DNAs. The percentage of their coding region ranges from 83 to 87% and the genomes are comprised of 1,696 proteins to 1,986 proteins. There are about 79 to 86 structural RNAs and 38% of GC content is present. 99 Lac genes (Table 2) have been traced from the genomes of *S. pyogenes* M Group A *Streptococcus* strains.

Mean Lac gene content

The mean of the frequencies of occurrence of all possible 64 codons, 4 nucleotides (A, T, G and C) and of chemically similar nucleotides –A, T and G, C altogether 70 parameters of the 99 Lac genes of the the *S. pyogenes* M Group A *Streptococcus* strains (Table 3) represent the mean Lac gene content. This content vector containing 70 mean gene content parameters are multiplied with 100 to convert to mean percentages and are used as input vectors for training algorithm of the Back Propagation Network.

Outcome of training algorithm

The training process of the Back Propagation Network is stopped once it reaches the near target output (0.99). The updated weights and biases are obtained from the training algorithm and are used for initializing weights and biases in the testing algorithm.

Outcome of testing algorithm

The testing done for the percentage of the mean vector which is used for training resulted in the near target output (0.99). Then the testing was performed for all the percentages of the 99 Lac gene mean vector frequencies in order to get a range of score values for the output that the Lac gene vector falls. The values obtained ranged from 0.9857 to 0.9901 and these ranges of values (Figure 2) are used in classifying whether a given sequence is a Lac gene or not.

SpyMGASLacGenePred

SpyMGASLacGenePred (Figure 3) represents the tool that has been developed using PERL and CGI. It accepts a DNA sequence and scans the accepted DNA sequence to identify Open Reading Frames (ORFs) in all six reading frames.

It determines whether all the identified ORFs belong to Lac gene category or not using the back propagation algorithm of the artificial neural network. If the network predicts ORFs as a Lac gene, the length and GC content of the sequence is determined and displayed along with the start and end positions, lengths of the ORF, frame information, score values and translated sequences of the ORFs (Figure 4).

The performance of the tool has been evaluated by the following cross validation tests. The test set included 100 sequences out of which 50 sequences are taken from the training sets which are Lac genes and the remaining 50 sequences are taken from a random set of sequences available in the nucleotide sequence database which are non Lac genes. For the 100 sequences of the test set performance measures of sensitivity, specificity and correlation coefficients are calculated.

Out of 100 test sequences, 50 are evaluated as True Positives (TP) representing Lac genes evaluated as Lac genes, 15 are evaluated as False Positives (FP) representing non Lac genes evaluated as genes, 35 are evaluated as True Negatives (TN) representing non Lac genes evaluated as non Lac genes and 0 evaluated as False Negatives (FN) representing Lac genes evaluated as non Lac genes. 50 True Positives summed up to 0 False Negatives to make up 50 Actual Positive (AP) sequences. 15 False Positives summed up to 35 True Negatives to make up 50 Actual Negative (AN) sequences.

Sensitivity (SN) and specificity (SP) are widely used to evaluate the performance of an algorithm (Burset et al., 1996).

Sensitivity (SN) is the ability to identify as many correct

Table 2. 99 Lac genes of the *Streptococcus pyogenes* M Group A Streptococcus strains with their positions in the genome database.

Lac gene							
S/No.1	**Strain name : MGAS10750**	**Accession: CP000262**					
Lac gene	Lac Z	Lac D1	Lac B1	Lac A1	Lac R1	Lac G	Lac E
Position	1335337..1338843	1445978..1446955	1447465..1447983	1447995..1448420	1451162..1451932	1682308..1683756	1683802..1685499
Lac gene	Lac F	Lac D2	Lac C2	Lac B2	Lac A2	Lac R2	
Position	1685499..1685816	1685840..1686823	1686827..1687756	1687804..1688319	1688354..1688782	1689228..1690001	
S/No. 2	**Strain name : MGAS2096**	**Accession: CP000261**					
Lac gene	Lac Z	Lac D1	Lac B1	Lac A1	Lac R1	Lac G	Lac E
Position	1276707..1280213	1387312..1388289	1388788..1389204	1389219..1389644	1392385..1393155	1613117..1614523	1614596..1616293
Lac gene	Lac F	Lac D2	Lac C2	Lac B2	Lac A2	Lac R2	
Position	1616293..1616610	1616634..1617617	1617621..1618550	1618596..1619111	1619146..1619574	1620020..1620793	
S/No. 3	**Strain name: MGAS10270**	**Accession: CP000260**					
Lac gene	Lac Z	Lac D1	Lac B1	Lac A1	Lac R1	Lac G	Lac E
Position	1352176..1355682	1462445..1463422	1463932..1464450	1464462..1464887	1467629..1468399	1651365..1652771	1652844..1654541
Lac gene	Lac F	Lac D2	Lac C2	Lac B2	Lac A2	Lac R2	
Position	1654541..1654858	1654882..1655865	1655869..1656798	1656844..1657359	1657314..1657822	1658268..1659041	
S/No. 4	**Strain name: MGAS315**	**Accession: AE014074**					
Lac gene	Lac A.1	Lac A.2	Lac B.1	Lac C.1	Lac D.1	Lac D.2	Lac E
Position	1479419..1479844	1671456..1671884	1478889..1479404	1478529..1478879	1477402..1478379	1668944..1669927	1666906..1668603
Lac gene	Lac F	Lac G	Lac R.1	Lac R.2			
Position	1668603..1668920	1665427..1666833	1482586..1483356	1672329..1673102			
S/No. 5	**Strain name : M1GAS**	**Accession : AE004092**					
Lac gene	Lac A.1	Lac A.2	Lac R.2	Lac B.1	Lac B.2	Lac C.1	Lac C.2
Position	1415927..1416352	1605049..1605477	1605923..1606696	1415394..1415912	1604499..1605014	1414870..1415387	1603522..1604451
Lac gene	Lac D.1	Lac D.2	Lac E	Lac F			
Position	1413910..1414887	1602535..1603518	1600497..1602194	1602194..1602511			
S/No. 6	**Strain name : MGAS5005**	**Accession : CP000017**					
Lac gene	Lac Z	Lac D.1	Lac B.1	Lac A.1	Lac R.1	Lac G	Lac E
Position	1258325..1261741	1368872..1369849	1370359..1370874	1370889..1371314	1374039..1374809	1596090..1597523	1597584..1599281
Lac gene	Lac F	Lac D.2	Lac C.2	Lac B.2	Lac A.2	Lac R.2	
Position	11599281..1599598	1599622..1600605	1600609..1601538	1601586..1602101	1602136..1602564	1603011..1603784	

Table 2. Contd.

S/No. 7 Strain name : MGAS6180		**Accession : CP000056**					
Lac gene	Lac A.1	Lac A.2	Lac Z	Lac D.1	Lac B.1	Lac R.1	Lac G
Position	1440544..1440969	1631947..1632375	1327950..1331456	1438527..1439504	1440014..1440532	1443711..1444481	1625919..1627325
Lac gene	Lac E	Lac F	Lac D.2	Lac C.2	Lac B.2	Lac R.2	
Position	1627398..1629095	1629095..1629412	1629436..1630419	1630422..1631351	1631397..1631912	1632821..1633594	
S/No. 8 Strain name : MGAS9429		**Accession : CP000259**					
Lac gene	Lac Z	Lac D1	Lac A1	Lac R1	Lac G	Lac E	Lac F
Position	1252864..1256370	1363420..1364397	1365327..1365752	1368495..1369265	1589232..1590638	1590711..1592408	1592408..1592725
Lac gene	Lac D2	Lac C2	Lac B2	Lac A2	Lac R2		
Position	1592749..1593732	1593736..1594665	1594711..1595226	1595261..1595689	1596135..1596908		

Table 3. Frequency table representing the frequencies and percentages of occurrence of 64 triplet codon, 4 - A, T, G, C bases and 2 - AT, GC bases for the 99 Lac genes of the *S. pyogenes* M Group A *Streptococcus* strains.

	T			**C**			**A**			**G**			
	Triplet codon	**Frequency**	**Percentage**	**Triplet codon**	**Frequency**	**Percentage**	**Triplet codon**	**Frequency**	**Percentage**	**Triplet codon**	**Frequency**	**Percentage**	
T	TTT	0.03	3	TCT	0.01	1	TAT	0.02	2	TGT	0.01	1	T
	TTC	0.01	1	TCC	0	0	TAC	0.02	2	TGC	0	0	C
	TTA	0.03	3	TCA	0.01	1	TAA	0	0	TGA	0	0	A
	TTG	0.02	2	TCG	0	0	TAG	0	0	TGG	0.01	1	G
C	CTT	0.02	2	CCT	0.01	1	CAT	0.01	1	CGT	0.01	1	T
	CTC	0.01	1	CCC	0	0	CAC	0.01	1	CGC	0.01	1	C
	CTA	0.01	1	CCA	0.01	1	CAA	0.02	2	CGA	0	0	A
	CTG	0	0	CCG	0	0	CAG	0.01	1	CGG	0	0	G
A	ATT	0.05	5	ACT	0.02	2	AAT	0.03	3	AGT	0.01	1	T
	ATC	0.02	2	ACC	0.01	1	AAC	0.04	4	AGC	0.01	1	C
	ATA	0.01	1	ACA	0.02	2	AAA	0.06	6	AGA	0.01	1	A
	ATG	0.03	3	ACG	0.01	1	AAG	0.02	2	AGG	0	0	G
G	GTT	0.03	3	GCT	0.04	4	GAU	0.05	5	GGT	0.04	4	T
	GTC	0.01	1	GCC	0.01	1	GAC	0.02	2	GGC	0.01	1	C
	GTA	0.01	1	GCA	0.03	3	GAA	0.06	6	GGA	0.02	2	A
	GTG	0.01	1	GCG	0.01	1	GAG	0.01	1	GGG	0	0	G

Table 3. Contd.

	Individual base frequencies	
Base name	**Frequency**	**Percentage**
T	0.29	29
C	0.17	17
A	0.32	32
G	0.21	21
	AT and GC frequencies	
Purine-Pyrimidine	**Frequency**	**Percentage**
AT	0.62	62
GC	0.38	38

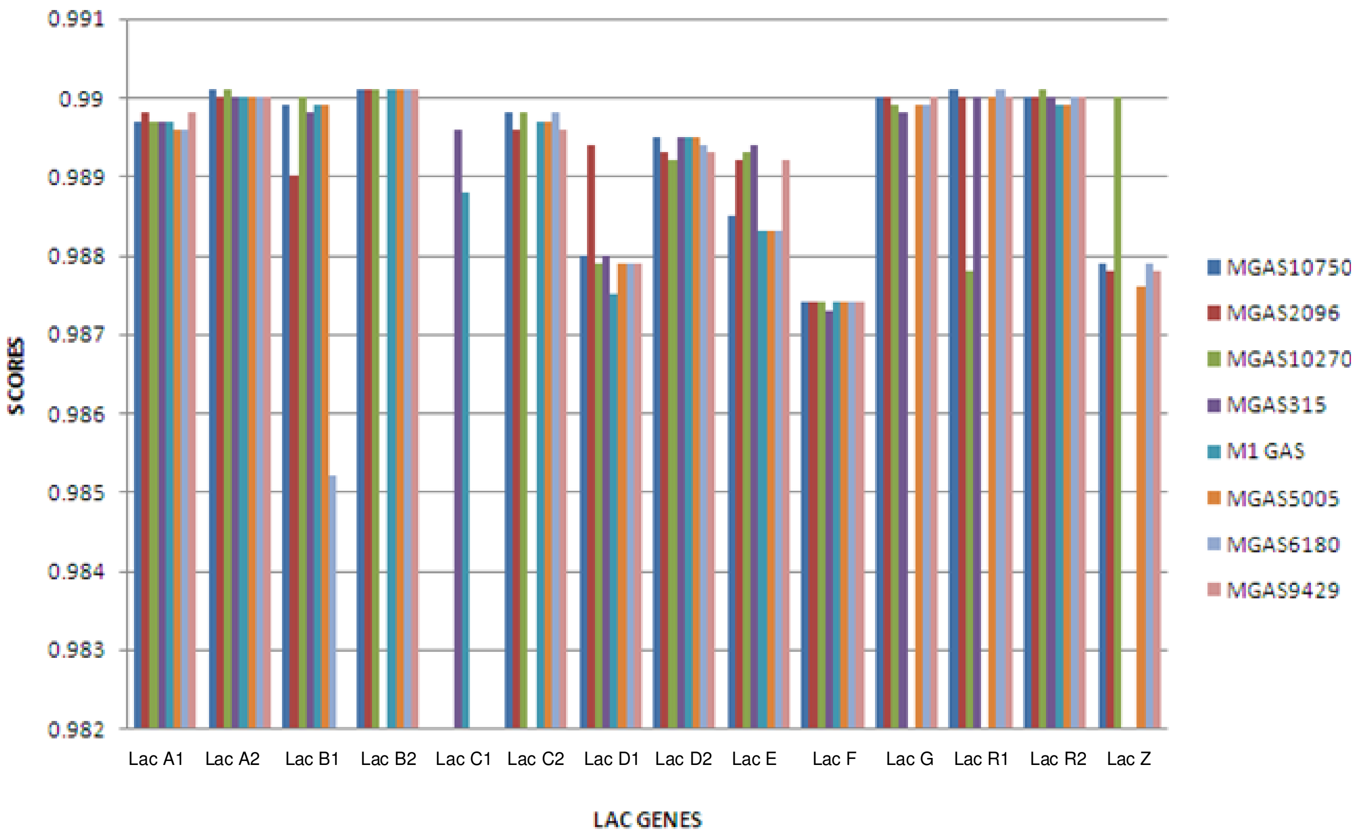


Figure 2. Score values of the testing algorithm for the 99 Lac gene vectors.

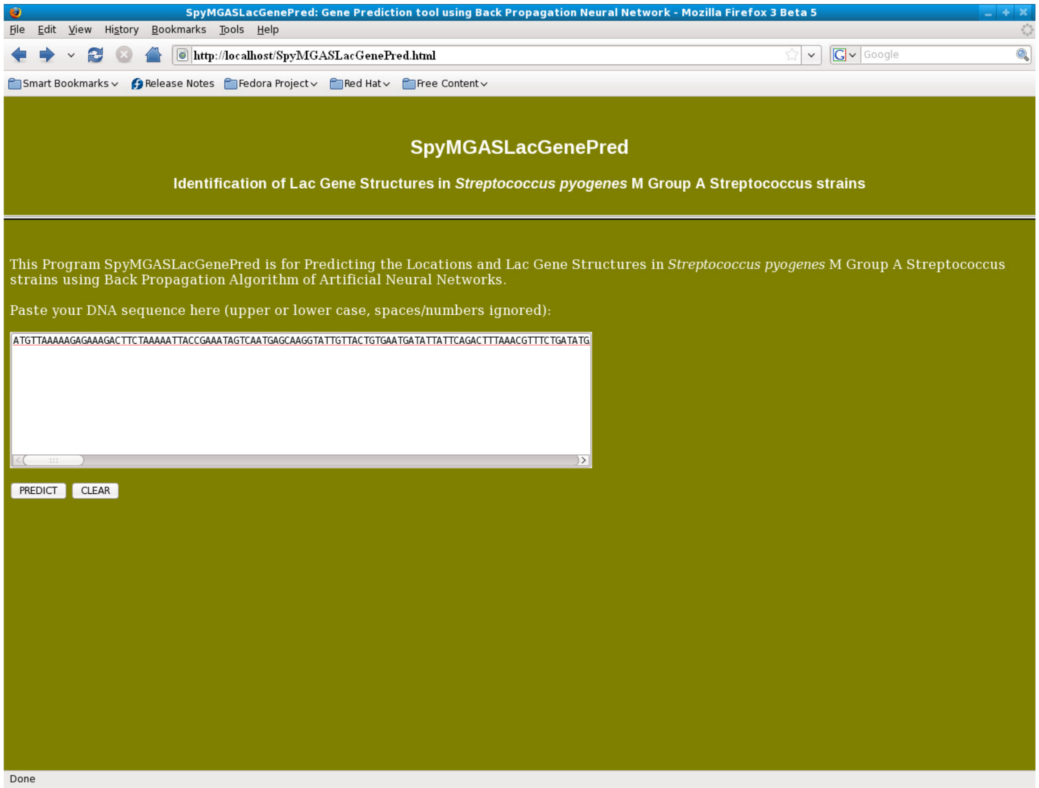
genes as possible and is calculated by the formula:

$$\text{Sensitivity (SN)} = \frac{\text{TP}}{(\text{TP+FN})} = \frac{50}{(50+0)} = 1$$

Specificity (SP) is measure of the proportion of correct genes out of the total genes identified and is calculated by the formula:

$$\text{Specificity (SP)} = \frac{\text{TP}}{(\text{TP+FP})} = \frac{50}{(50+15)} = 0.769$$

The correlation coefficient is used in machine learning as a measure of the quality of binary (two-class) classifications (Baldi et al., 1996). It takes into account true and false positives and negatives and is generally regarded as a balanced measure which can be used even if the classes are of very different sizes. It returns a value between ±1. A coefficient of +1 represents a


SpyMGASLacGenePred: Gene Prediction tool using Back Propagation Neural Network - Mozilla Firefox 3 Beta 5
File Edit View History Bookmarks Tools Help
http://localhost/SpyMGASLacGenePred.html
Google
Smart Bookmarks Release Notes Fedora Project Red Hat Free Content
PREDICT
CLEAR
Done

SpyMGASLacGenePred

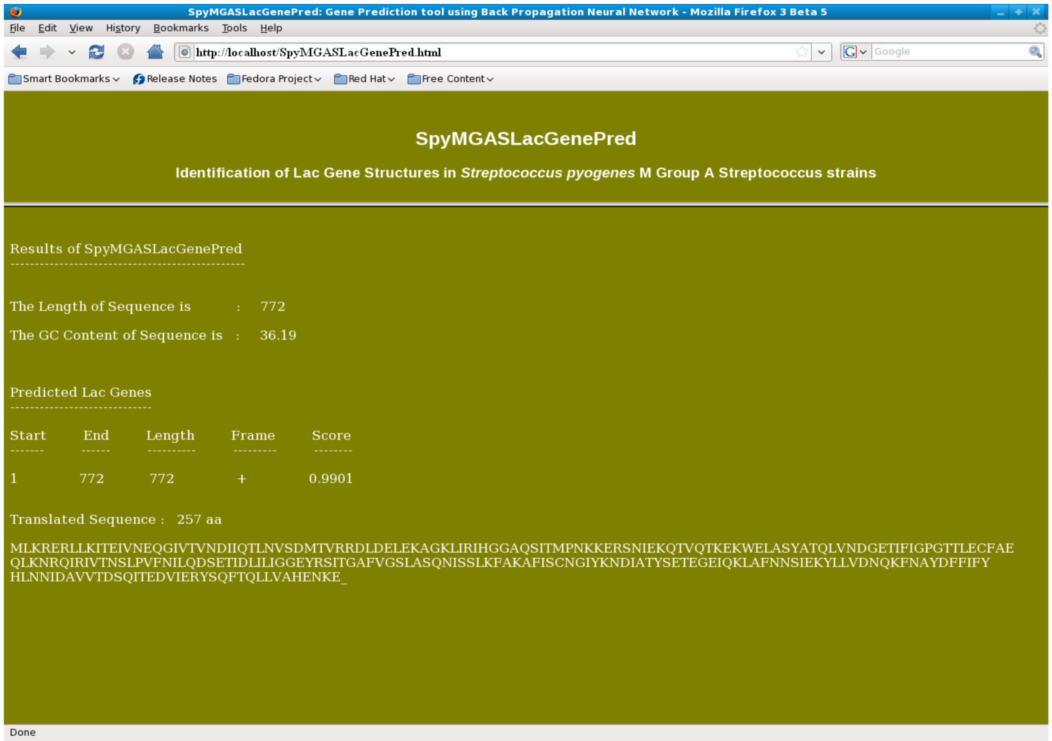
Identification of Lac Gene Structures in *Streptococcus pyogenes* M Group A Streptococcus strains

This Program SpyMGASLacGenePred is for Predicting the Locations and Lac Gene Structures in *Streptococcus pyogenes* M Group A Streptococcus strains using Back Propagation Algorithm of Artificial Neural Networks.

Paste your DNA sequence here (upper or lower case, spaces/numbers ignored):

```
ATGTTAAAAAGAGAAAGACTTCTAAAAATTACCGAAATAGTCAATGAGCAAGGTATTGTTACTGTGAATGATATTATTCAGACTTTAAACGTTTCTGATATG
```

Figure 3. SpyMGASLacGenePred : a Tool to Identify Locations and Lac Gene Structures in *S. pyogenes* M Group A *Streptococcus* strains.


SpyMGASLacGenePred: Gene Prediction tool using Back Propagation Neural Network - Mozilla Firefox 3 Beta 5
File Edit View History Bookmarks Tools Help
http://localhost/SpyMGASLacGenePred.html
Google
Smart Bookmarks Release Notes Fedora Project Red Hat Free Content
Done

SpyMGASLacGenePred

Identification of Lac Gene Structures in *Streptococcus pyogenes* M Group A Streptococcus strains

```
Results of SpyMGASLacGenePred
-------------------------------------------------


The Length of Sequence is        :    772

The GC Content of Sequence is   :    36.19


Predicted Lac Genes
-----------------------------

Start       End       Length       Frame       Score
-------     ------    ----------   ---------   --------

1           772       772            +         0.9901

Translated Sequence :   257 aa

MLKRERLLKITEIVNEQGIVTVNDIIQTLNVSDMTVRRDLDELEKAGKLIRIHGGAQSITMPNKKERSNIEKQTVQTKEKWELASYATQLVNDGETIFIGPGTTLECFAE
QLKNRQIRIVTNSLPVFNILQDSETIDLILIGGEYRSITGAFVGSLASQNISSLKFAKAFISCNGIYKNDIATYSETEGEIQKLAFNNSIEKYLLVDNQKFNAYDFFIFY
HLNNIDAVVTDSQITEDVIERYSQFTQLLVAHENKE_
```

Figure 4. Results of SpyMGASLacGenePred : Identified locations and Lac Gene structures in *S. pyogenes* M Group A *Streptococcus* strains.

perfect prediction, 0 an average random prediction and −1 an inverse prediction.

Correlation coefficient is calculated by the formula:

$$\text{Correlation-Coefficient (CC)} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)\,(TP+FN)\,(TN+FP)\,(TN+FN)}}$$

$$= \frac{50 \times 35 - 15 \times 0}{\sqrt{(50+15)\,(50+0)\,(35+15)\,(35+0)}} = 0.733$$

The calculated performance measures showed that the developed tool SpyMGASLacGenePred has a sensitivity of 100% and a specificity of 76.9%. Since every Lac genes used for training is taken into consideration by the Back Propagation Neural Network program for testing, the tool has 100% sensitivity. However if Lac genes of the other strains of *S. pyogenes* which are not used for training is tested, then sensitivity might drop to a certain extent. The tool has a specificity of 76.9% and this indicates that the tool is above an acceptable threshold level to predict the correct Lac gene out of total Lac genes. The tool also showed a correlation coefficient of 0.733 which is near to +1 and can be considered as near perfect prediction.

Conclusion

A systematic method of back propagation algorithm that uses gradient-descent based delta learning rule also known as back propagation rule for training multilayer feed forward artificial neural networks provided a computationally efficient method for changing the weights in the feed forward network with differentiable activation function units to learn a set of Lac gene input patterns. Being a gradient descent method, it minimized the total squared error of the output computed by the network. The network that is trained by a supervised learning method achieved the balance between the ability to respond correctly to the input Lac gene structures that are used for training and provided good responses to the Lac gene structures that are similar to the trained lac gene structures with a sensitivity of 100%, specificity of 76.9% and correlation coefficient of 0.733. Sensitivity of the tool is perfect, specificity also lies above threshold and correlation coefficient is near +1and specifies the tool to be near perfect prediction. These facts imply that the current back propagation algorithm of artificial neural network method is a useful computer technique for predicting Lac gene structures in *S. pyogenes* M Group A Streptococcus strains.

REFERENCES

Anderson JA, Rosenfeld E (1988). Neurocomputing: Foundations of Research. MIT Press, Cambridge, pp. 4-10.

Baldi P, Brunak S, Chauvin Y, Andersen CAF, Nielsen H (1996). Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics. 16: 412-424.

Basheer IA, Hajmeer M (2000). Artificial neural networks: fundamentals, computing, design and application. J. Microbiol. Meth. 43: 3-31.

Beall B, Facklam R, Hoenes T, Schwartz B (1997). Survey of *emm* gene sequences and T-antigen types from systemic *Streptococcus pyogenes* infection isolates collected in San Francisco, California; Atlanta, Georgia; and Connecticut in 1994 and 1995. J. Clin. Microbiol. 35: 1231-1235.

Bencivenga JF, Johnson DR, Kaplan EL (2009). Determination of group a streptococcal anti-M type-specific antibody in sera of rheumatic fever patients after 45 years. Clinical Infectious Diseases: An Official Publication of the Infect. Dis. Soc. Am., 49(8): 1237–1239.

Bishop CM (1995). Neural Networks for Pattern Recognition. Oxford University Press, Oxford, UK, pp. 14 -18.

Bishop CM (2006). Pattern Recognition and Machine Learning. Springer, New York. pp: 19-29.

Burge C, Karlin S (1998). Finding the genes in genomic DNA. Current Opinion, Struct. Biol., 8: 346-354.

Burset M, Guigo R (1996). Evaluation of Gene Structure Prediction Programs. Genome, 34: 353–367.

Currie BJ (2006). Group A streptococcal infections of the skin: molecular advances but limited therapeutic progress. Curr. Opin. Infect. Dis., 19: 132–138.

Doktor SZ, Beyer JM, Flamm RK, Shortridge VD (2005). Comparison of *emm* typing and ribotyping with three restriction enzymes to characterize clinical isolates of *Streptococcus pyogenes*. J. Clin Microbiol., 43: 150-155.

Duda RO, Hart PE, Stork DG (2000). Pattern Classification. Wiley Interscience, New York. Pp. 34-45.

Facklam R. (2002). What happened to the *streptococci*: overview of taxonomic and nomenclature changes. Clin. Microbiol. Rev. 15: 613-630.

Facklam R, Beall B, Efstratiou A, Fischetti V, Johnson D, Kaplan E, Kriz P, Lovgren M, Martin D, Schwartz B, Totolian A, Bessen D, Hollingshead S, Rubin F, Scott J, Tyrrell G (1999). emm typing and validation of provisional M types for group A streptococci. Emerg. Infect. Dis., 5: 247-253.

Ficket JW (1982). Recognition of protein coding regions in DNA sequences. Nucleic Acid Res., 10: 5303-5318.

Ficket JW (1998). Gene Identification. Bioinformatics, 10: 563-578.

Guigo R (1997). Computational gene identification: an open problem. Comput. Chem., 21: 215-222.

Guigo R, Knudsen S, Drake N, Smith T (1992). Prediction of Gene Structure. J. Mol. Biol. 226: 141-157.

Haykin S (2001). Neural Networks: A Comprehensive Foundation. Second Edition. Tsinghua University Press, pp. 34-48.

Hertz JA, Krogh A, Palmer R (1991). Introduction to the Theory of Neural Computation. Addison-Wesley, Redwood City, pp. 23-34.

Lancefield R.C. (1928). The antigenic complex of *Streptococcus hemolyticus*. J. Exp. Med., 47: 9–10.

Lancefield RC, Dole VP (1946). The properties of T antigen extracted from group A hemolytic Streptococci. J. Exp. Med., 84: 449–71.

Li C, He P, Wang J (2003). Artificial Neural Network Method for Predicting Protein Coding Genes in the Yeast Genome. Internet Electron. J. Mol. Des., 2: 527-538.

Mathe C, Sagot MF, Schiex T, Rouze P (2002). Survey and Summary: Current methods of gene prediction, their strengths and weaknesses. Nucleic Acids Res., 30: 4103-4117.

McGregor KF, Spratt BG, Kalia A, Bennett A, Bilek N, Beall B, Bessen DE (2004). Multilocus sequence typing of *Streptococcus pyogenes* representing most known emm types and distinctions among subpopulation genetic structures. J. Bacteriol., 186: 4285–4294.

Milanesi L, Kolchanov NA, Rogozin IB, Kel IV, Orlov AE, Ponomarenko YL, Vezzoni MP (1993). Genview : A computing tool for protein-encoding regions prediction in nucleotide sequences, Proc. of the Second Int. Conf. on Bioinformatics, Supercomputing and Complex Genomic Analysis, in Lim, H.A., Fickett, J.W., Cantor, C.R. and Robbins, R.J. World Scientific Publishing, Singapore, pp. 573-588.

Mora M, Bensi G, Capo S (2005). Group A *Streptococcus* produce pilus-like structures containing protective antigens and Lancefield T antigens. Proc Natl Acad Sci USA, 102(43): 15641–15646.

Russell S, Norvig P (2003). Artificial Intelligence: A Modern Approach. Prentice Hall, Inc., pp. 4-8.

Teixeira LM, Barros RR, Castro AC, Peralta JM, Carvalho MGS, Talkington DF, Vivoni AM, Facklam RR, Beall B (2001). Genetic and phenotypic features of *Streptococcus pyogenes* strains isolated in Brazil that harbor new emm sequences. J. Clin. Microbiol., 39: 3290–3295.

Xu Y, Uberbacher EC (1998). Computational gene prediction using neural networks and similarity search. Comput. Meth. in Molecular Biology, Elsevier Science, New York, pp. 109-128.

Zhang CT, Wang J (2000). Recognition of protein coding genes in the yeast genome at better than 95% accuracy based on the Z curve. Nucleic Acids Res. 28: 2804-2814.

Zhang CT, Wang J, Zhang R (2002). Using a Euclid distancediscriminant method to find protein coding genes in the yeast genome. Comput. Chem., 26: 195-206.