An Effective Identification of Species from DNA Sequence: A Classification Technique by Integrating DM and ANN

Sathish Kumar S
Research Scholar
Dr MGR Educational and Research Institute University.
Chennai, Tamil Nadu, India
Tel: +919886372152

Dr.N.Duraipandian, M.E., Ph.D
Vice Principal
Velammal Engineering College
Ambattur – Redhills Road, Chennai,
Tamil Nadu, India

***Abstract*— Species classification from DNA sequences remains as an open challenge in the area of bioinformatics, which deals with the collection, processing and analysis of DNA and proteomic sequence. Though incorporation of data mining can guide the process to perform well, poor definition, and heterogeneous nature of gene sequence remains as a barrier. In this paper, an effective classification technique to identify the organism from its gene sequence is proposed. The proposed integrated technique is mainly based on pattern mining and neural network-based classification. In pattern mining, the technique mines nucleotide patterns and their support from selected DNA sequence. The high dimension of the mined dataset is reduced using Multilinear Principal Component Analysis (MPCA). In classification, a well-trained neural network classifies the selected gene sequence and so the organism is identified even from a part of the sequence. The proposed technique is evaluated by performing 10-fold cross validation, a statistical validation measure, and the obtained results prove the efficacy of the technique.**

Keywords- Pattern Generation; DNA Sequence; Pattern Support; Mining; Neural Network.

I. INTRODUCTION

Bioinformatics is a rapidly growing area of computer science [19] that deals with the collection, organization, and analysis of Deoxyribonucleic acid (DNA) and protein sequence [18]. Today it addresses the formal and practical issues that occur in the management and analysis of genomic and proteomic data because it includes the formation and development of databases, algorithms, computational and statistical technique, and hypothesis [1].

Genomic signal processing (GSP) is a relatively new area in bio-informatics that uses traditional digital signal processing techniques to deal with digital signal representations and analysis of genomic data [2] [12]. GSP gains biological knowledge by the analysis, processing, and use of genomic signals and translates the gained biological knowledge into systems-based applications [3]. Integration of signal processing theories and methods with global understanding of functional genomics with significant emphasis on genomic regulation is the main objective of GSP [4].

The whole DNA of a living organism is known as its Genome [5]. Genomic signals carry genomic information to all the processes that take place in an organism [6]. Essentially DNA is a nucleic acid that has two long strands of nucleotides twisted in the form of a double helix and its external backbone is made up of alternating deoxyribose sugar and phosphate molecules. The nitrogenous bases Adenine, Guanine, Cytosine and Thymine are present in the interior portion of the DNA in pairs [13] [9]. DNA and proteins can be mathematically represented as character strings, where each character is a letter of the alphabet [6] [10] [11].

One of the vital tasks in the study of genomes is gene identification [7]. DNA analysis utilizes methods such as clustering [20], data mining [21] [22] [23], gene identification [24] and gene regulatory network modeling [25] [26]. These methods present cutting edge research topics and methodologies for the purpose of facilitating collaboration between researchers and bioinformaticians. Mining bioinformatics data is a rising field at the intersection of bioinformatics and data mining [14]. Some of them belong to the category of data mining that decides whether or not an example not yet noticed is of a predefined type. Increased availability of huge amount of biomedical data and the expectant need to turn such data into useful information and knowledge is the main reason for the recent increased attention in data mining in the biomedical industry.

Large number of research works that incorporate data mining in bioinformatics for different purposes are available in the literature [15] [16] [17]. . A few important such researches are reviewed in section 2. One important research of this type is the identification of species or name of an organism from its gene sequence. Characterization of the unknown environmental isolates with the genomic species is not easy because genomic species are especially heterogeneous and poorly defined [8]. Identifying the species or the organism from its gene sequence is a challenging task. In this paper, we propose a classification technique to effectively classify the species or name of an organism from its DNA sequence. This technique is detailed with mathematical formulations and illustrations in section 3. Section 4 discusses the implementation results and Section 5 concludes the paper.

II. RELEATED WORKS

Plenty of research works deals with the mining knowledge from the genomic sequences. Some of the recent research works are briefly reviewed here. Riccardo Bellazzi et al. [27] have discussed that in the past years, the gene expression data analysis that are aiming at complementing microarray analysis with data and knowledge of various existing sources has grown from being purely data-centric to integrative. Focusing on the evolution of gene expression data mining techniques toward knowledge-based data analysis approaches, they have reported on the overabundance of such techniques. Particularly, latest developments in gene expression-based analysis methods utilized in association and classification studies, phenotyping and reverse engineering of gene networks have been discussed.

The gene expression data sets for ovarian, prostate, and lung cancer was examined by Shital Shah et al. [28]. For genetic expression data analysis, an integrated gene-search algorithm was presented. For making predictions and for data preprocessing (on partitioned data sets) and data mining (decision tree and support vector machines algorithms), a genetic algorithm and correlation-based heuristics was included in the their integrated algorithm. The knowledge, which was obtained by the algorithm, has high classification accuracy with the capability to recognize the most important genes. To further improve the classification accuracy, bagging and stacking algorithms were employed. The results were compared with the literary works. The cost and complexity of cancer detection and classification was eventually condensed by the mapping of genotype information to the phenotype parameters.

Locating motif in bio-sequences, which is a very significant primitive operation in computational biology, was discussed by Hemalatha et al. [29]. Computer memory space requirement and computational complexity are few of the computational requirements that are needed for a motif discovery algorithm. To overcome the intricacy of motif discovery, an alternative solution integrating genetic algorithm and Fuzzy Art machine learning approaches was proposed for eradicating multiple sequence alignment process. The results that were attained by their planned model to discover the motif in terms of speed and length were compared with the enduring technique. By their technique, the length of 11 was found in 18 sec and length of 15 in 24 sec, whereas the existing techniques found length of 11 in 34 sec. When compared to other techniques, the proposed one has outperformed the accepted existing technique. By employing MATLAB, the projected algorithm was put into practice and with large DNA sequence data sets and synthetic data sets, it was tested.

An interactive framework which is based on web for the analysis and visualization of gene expressions and protein structures was described by Ashraf S. Hussein [30]. The formulation of the projected framework encountered various confronts because of the variety of significant analysis and visualization techniques, moreover to the survival of a diversity of biological data types, on which these techniques function. Data incorporated from heterogeneous resources, for instance expert-driven data from text, public domain databases and various large scale experimental data and the lack of standard I/O that makes it difficult to integrate the most recent analysis and visualization are the two main challenges that directed the formulation of the current framework. Hence, the basic novelty in their proposed framework was the integration of the state-of -art techniques of both analysis and visualization for gene expressions and protein structures through a unified workflow. Moreover, a wide range of input data types are supported by it and three dimensional interactive outputs ready for exploration by off-the-shelf monitors and immersive, 3D, stereo display environments can be exported by it using Virtual Reality Modeling Language (VRML).

A stomach cancer detection system, which is on the basis of Artificial Neural Network (ANN) and the Discrete Cosine Transform (DCT), was developed by Ahmad M. Sarhan [31]. By employing DCT the projected system extracted the classification features from stomach microarrays. The extracted characteristics from DCT coefficients were applied to an ANN for further classification (tumor or non-tumor). The microarray images that were employed were acquired from the Stanford Medical Database (SMD). Simulation results has illustrated that a very high success rate was produced by the proposed system.

The challenging issue in microarray technique which was to analyze and interpret the large volume of data was discussed by Valarmathie et al. [32]. This can be made possible by the clustering techniques in data mining. In hierarchical and k-means clustering techniques which are hard clustering, the data is split into definite clusters, where each cluster has exactly one data element so that the result of the clustering may be wrong many times.

The problems that are addressed in hard clustering can be resolved in fuzzy clustering technique. Amid all fuzzy based clustering, fuzzy C-means (FCM) is best suited for microarray gene expression data. The problem that is related with fuzzy C-means was the amount of clusters that are to be generated for the given dataset and that needs to be notified first. By combining the technique with a popular probability related Expectation Maximization (EM) algorithm, it can be solved to model the cluster structure of gene expression data and it has offered the statistical frame work. Determining the accurate number of clusters and its efficient interpretation is the main purpose of the projected hybrid fuzzy C-means technique.

Explorative studies in support of solutions to facilitate the analysis and interpretation of mining results was described by Belmamoune et al. [33]. A solution that was located in the extension of the Gene Expression Management System (GEMS) was described, i.e. an integrative framework for spatio-temporal organization of gene expression patterns of zebra fish to a framework that supports data mining, data analysis and patterns interpretation.

As a proof of principle, the GEMS is provided with data mining functionality which is appropriate to monitor spatio-temporal, thus generating added value to the submission of data for data mining and analysis. On the basis of the availability of domain ontologies, the analysis of the genetic networks was done which vigorously offers the meaning to the

discovered patterns of gene expression data. Grouping of data mining with the already accessible potential of GEMS considerably augments the existing data processing and functional analysis strategies.

III. The Integrated Technique for Species Classification

The proposed species classification technique classifies species based on the given DNA sequence. The DNA sequence is comprised of four basic nucleotides, Adenine (A), Guanine (G), Cytosine (C) and Thymine (T). Every species has a long DNA sequence, which is formed by the four nucleotides.

The DNA sequence defines the attributes, nature and type of the species. The proposed technique is an integration of data mining and artificial intelligence. In the proposed technique, firstly, nucleotide patterns are mined from the sequence. The mined patterns form a nucleotide pattern database with higher dimension. So, secondly, the dimension of the pattern database is reduced by MPCA. Finally, the dimensionality reduced pattern database is used to train the neural network. The technique is described in the further sub sections.

The proposed species classification technique classifies a species based on its DNA sequence. The four basic nucleotides, Adenine (A), Guanine (G), Cytosine (C) and Thymine (T) are the building blocks of the long DNA sequence in every species. A DNA sequence defines the attributes, nature and type of the species. The proposed technique is developed by integrating data mining and artificial intelligence techniques. Firstly, the proposed technique mines the nucleotide patterns from the sequence and forms a high dimensional nucleotide pattern database with the mined patterns.

Secondly, the technique uses MPCA and reduces the dimension of the pattern database. Finally, the dimensionality reduced pattern database is used to train a neural network. The following sub sections elaborately describe this technique.

A. Mining Nucleotide Patterns from DNA sequence

The first and initial stage of the proposed technique mines the nucleotide pattern from the DNA sequence. At this stage, patterns formed by different combinations of nucleotides are mined using a novel mining algorithm. Let be the DNA sequence, which is a combination of four nucleotides A, G, C and T. For instance, a sample DNA sequence is given as CGTCGTGGAA.

From the sequence, the mining algorithm extracts different nucleotide patterns and their support. The algorithm is comprised of two stages, namely, pattern generation and support finding. In pattern generation, patterns with different length are generated whereas in support finding, support values for every generated pattern are determined from the DNA sequence. The basic structure of the algorithm is given as a block diagram in Fig. 1.

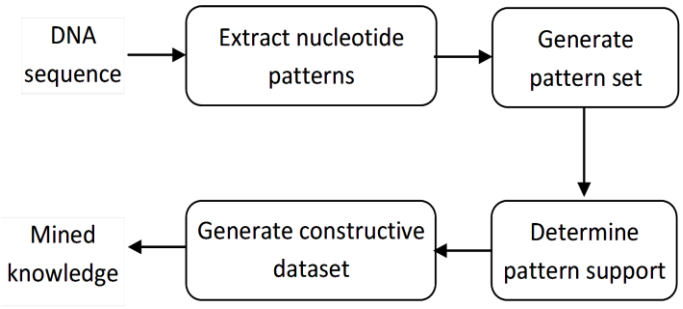


Figure 1. Block diagram of the pattern mining algorithm

1) Pattern generation

In pattern generation, different possible combinations of nucleotide base pairs are generated. As a reference, a base set B is generated with cardinality $|B|=4$, which has the elements $\{A,G,C,T\}$. Let, $\{P_l\}; l=1,2,\Lambda\ ,L_{\max}$, be the pattern set to be generated, where, $L_{\max}$ is the maximum length of a pattern in a pattern set. The pattern set is generated as follows

$$\{P_l\}_{k^{(l)}} = \{P_l\}_{k^{(l)}-1} \cup \{B(a_1)B(a_2)\Lambda\ B(a_l)\} \quad (1)$$

Where, $k^{(l)}=1,2,\Lambda\ ,|B|^l$, $1\le a_1,a_2,\Lambda\ ,a_l \le |B|$ and $\{B(a_1)B(a_2)\Lambda\ B(a_l)\}$ is a set of different combinations of nucleotide bases. Eq. (1) operates with the criterions, $\{P_l\}_{k^{(l)}-1} \subseteq \{P_l\}_{k^{(l)}}$ and $\{B(a_1)B(a_2)\Lambda\ B(a_l)\} \subseteq \{P_l\}$. Eq. (1) formulated for pattern generation is analyzed using two examples.

Example 1: *To generate a two length pattern set* P_2, *i.e.* $l=2$. Here, two indexing variables a_1 and a_2 are generated. At every k i.e. $k=1\ to\ 16$ and for its corresponding a_1 and a_2, the obtained $\{B(a_1)B(a_2)\}$ and $\{P_2\}_{k^{(2)}-1}$ are tabulated in Table II.

Hence, P_2 is obtained as $\{P_2\}=\{AA,AG,AC,AT,GA,GG,GC,GT,CA,CG,CC,CT,TA,TG,TC,TT\}$ by integrating P_2 obtained from every k^{th} iteration. From $P_1,P_2,\mathrm{K}\ ,P_{L_{\max}}$, a consolidated pattern set P, which is the required pattern to be generated, is obtained as $P=P_1\cup P_2\cup\Lambda\ P_{L_{\max}}$. The cardinality of P can be determined as $|P| = \sum_{l=1}^{L_{\max}} |P_l|$.

TABLE I. DIFFERENT COMBINATIONS AND PATTERN SETS GENERATED FOR EVERY a_1, a_2 AND k

k	a_1	a_2	{B(a_1) B(a_2)}	$\{P_2\}_k(2)_{-1}$
1	1	1	AA	{}
2	1	2	AG	{AA}
3	1	3	AC	{AA, AG}
4	1	4	AT	{AA, AG, AC}
5	2	1	GA	{AA, AG, AC, AT}
6	2	2	GG	{AA, AG, AC, AT, GA}
7	2	3	GC	{AA, AG, AC, AT, GA, GG}
8	2	4	GT	{AA, AG, AC, AT, GA, GG,GC}
9	3	1	CA	{AA, AG, AC, AT, GA, GG,GC, GT}
10	3	2	CG	{AA, AG, AC, AT, GA, GG,GC, GT, CA}
11	3	3	CC	{AA, AG, AC, AT, GA, GG,GC, GT, CA, CG}
12	3	4	CT	{AA, AG, AC, AT, GA, GG,GC, GT, CA, CG, CC }
13	4	1	TA	{AA, AG, AC, AT, GA, GG,GC, GT, CA, CG, CC,CT}
14	4	2	TG	{AA, AG, AC, AT, GA, GG,GC, GT, CA, CG, CC,CT,TA}
15	4	3	TC	{AA, AG, AC, AT, GA, GG,GC, GT, CA, CG, CC,CT,TA,TG}
16	4	4	TT	{AA, AG, AC, AT, GA, GG,GC, GT, CA, CG, CC,CT,TA,TG,TC}

2) Determination of Pattern Support

The support, which has to be determined for every extracted pattern, describes the DNA attribute. By performing a window based operation over the sequence g, the support can be determined. Window of sequences are determined for different lengths as follows

$$w_l(j) = g(j, j+1, \Lambda\ , j+l-1) \qquad (2)$$

Once the window of sequences is extracted support is determined for the mined patterns. The pseudo code, which is given below, describes the procedure to determine the support for every pattern.

Initialize C to zero
Read window of sequence w
For every l - length pattern, P_l
  For each element i in P_l, $P_l(i)$
    For each window of sequence $w_l(j)$
      If $P_l(i)$ and $w_l(j)$ are same
        Increment $C_l(i)$
      End if
    End for
  End for
End for
Return C

Figure 2. Pseudo code to determine support for every mined

Figure 3. ained C for each different length pattern hapattern

The obts the support for all the elements that are present in the corresponding pattern set. From the mined pattern and its corresponding support, a constructive dataset is generated.

3) Constructive dataset generation

A raw dataset is generated using the aforesaid mining algorithm. But the dataset is not constructive for further operation. In this stage, a constructive dataset is generated from the mined dataset, which comprises of patterns with different lengths and their support.

To accomplish this, firstly the patterns which have length $l \geq 2$ are taken. From the pattern set, the modified and constructive dataset is generated as given in Table 3.

TABLE II. A GENERAL STRUCTURE OF THE PROPOSED CONSTRUCTIVE DATASET

	A	**G**	**C**	**T**
A	$C_2(1)$	$C_2(2)$	$C_2(3)$	$C_2(4)$
G	$C_2(5)$	$C_2(6)$	$C_2(7)$	$C_2(8)$
C	$C_2(9)$	$C_2(10)$	$C_2(11)$	$C_2(12)$
T	$C_2(13)$	$C_2(14)$	$C_2(15)$	$C_2(16)$
AA	$C_3(1)$	$C_3(2)$	$C_3(3)$	$C_3(4)$
AG	$C_3(5)$	$C_3(6)$	$C_3(7)$	$C_3(8)$
AC	$C_3(9)$	$C_3(10)$	$C_3(11)$	$C_3(12)$
AT	$C_3(13)$	$\Lambda\ \Lambda$	$\Lambda\ \Lambda$	$\Lambda\ \Lambda$
M M				

In the constructive dataset, all the patterns except single length pattern are considered. Hence, the dataset is of size $4^{L_{\max}-1} \times 4$. The generated constructive dataset belongs to a particular gene sequence. Similarly, the constructive dataset for different sequences are generated. Hence, the final dataset $G_{xy}^{(z)}$; $x = 1,2,\Lambda\ ,4^{L_{\max}-1}$, $y = 1,2,3\ and\ 4$ and $z = 1,2,\Lambda\ , N_g$ is obtained, which is subjected to further processing.

B. MPCA-based Dimensionality reduction

In all tensor modes, the multilinear algorithm MPCA captures most of the variation present in the original tensors by seeking those bases in each mode that allow projected tensors and performs dimensionality reduction [35]. Initially, in the process of dimensionality reduction, the distance matrix for every z^{th} matrix is determined as follows, $D^{(z)} = G^{(z)} - \mu$ (3)

Where,

$$\mu_{xy} = \frac{1}{N_G} \sum_{z=0}^{N_G - 1} G_{xy}^{(z)} \quad (4)$$

Using Eq. (3) and by determining the mean matrix μ for $G^{(z)}$ using Eq. (4), the distance matrix can be calculated. Then with mode 2, tensor representations [34] $T_1^{(z)}$ and $T_2^{(z)}$ are given to the obtained distance matrix. A projection matrix Ψ is determined as follows,

$$\Psi = \sum_{z=0}^{N_G - 1} T^{(z)} \left(T^{(z)} \right)^T \quad (5)$$

For both $T_1^{(z)}$ and $T_2^{(z)}$, the projection matrix (Ψ_1 and Ψ_2) are determined using the generalized form of calculation given in Eq. (5). For Ψ_1 and Ψ_2 , the corresponding eigenvectors E_1 and E_2 and the corresponding Eigen values λ_1 and λ_2 are determined by subjecting the projection matrix to a generalized eigenvector problem. The rows of the eigenvector are arranged based on the index of the eigenvalues sorted in the descending order. The modified eigenvector $E_1^{'}$ and $E_2^{'}$ are obtained by transposing the arranged eigenvector. The cumulatively distributed Eigen values for the sorted eigenvalues are generally determined using the following equation.

$$\lambda_x^{'} = \frac{\lambda_x^{cdf}}{\sum_{x=0}^{|\lambda|-1} \lambda_x^{sort}} \quad (6)$$

The sorted Eigen values λ_x^{sort} and the cumulatively distributed Eigen values λ_x^{cdf} of Eq. (6), can be determined as

$$\lambda_x^{cdf} = \lambda_x^{sort} + \lambda_{x-1}^{cdf} \quad (7)$$

Where, $\lambda_0^{cdf} = \lambda_0^{sort}$ at $x = 0$. The new dimension λ_T is calculated from the obtained $\lambda_x^{'}$, using a dimensional threshold D_T . To accomplish this, the indices of all eigenvalues that satisfy the condition $\lambda_x^{'} \geq D_T$ are identified. Then, by extracting the first λ_T rows of $E_1^{'}$ and $E_2^{'}$, the corresponding dimensionality reduced eigenvectors $E_1^{''}$ and $E_2^{''}$ are determined. For the $E_1^{''}$ and $E_2^{''}$, again tensor matrices but $T_1^{(z)}$ and $T_2^{(z)}$ times are determined [35]. The

process followed for projection matrix is repeated for the tensor matrices to obtain $\lambda_{x_1}^{new}$ and $\lambda_{x_2}^{new}$ and E_1^{new} and E_2^{new}. The weight of both the tensor eigen values are determined as $\lambda_x^w = \sqrt{\lambda_{x_1}^{new}\lambda_{x_2}^{new}}$. Then, the dimensionality reduced matrix $G^{(z)}$ of size $N_R \times N_T$, is obtained by using the MPCA projections [35], where, N_R can be determined as $N_R = \lambda_{T_1}.\lambda_{T_2}$.

C. Classification using ANN

For N_G gene sequences, the dimensionality reduced gene patterns and their support are provided by the MPCA. Using ANN, the class of the original sequence can be identified using the dataset. Two classical operations, training and testing are involved in the classification. The neural network is trained using the N_G pattern dataset. Here, the process is performed using multilayer feed forward neural network, depicted in Fig. 3. N_R Input nodes, N_H hidden nodes and an output node are present in the network.

Before performing any task, the ANN must be trained. Once trained, the ANN capably identifies the species by finding the class of the gene sequence. The training phase and classification phase of the ANN are described below.

1) Training Phase

Back Propagation (BP) algorithm is used to train the constructed feed forward network. The step-by-step procedure utilized in the training process is given below.

1. Assign arbitrary weights generated within the interval to links between the input layer and hidden layer as well as hidden layer and output layer.
2. Using Eq. (8), (9) and (10), determine the output of input layer, hidden layer and output layer respectively by inputting constructive dataset to the network.

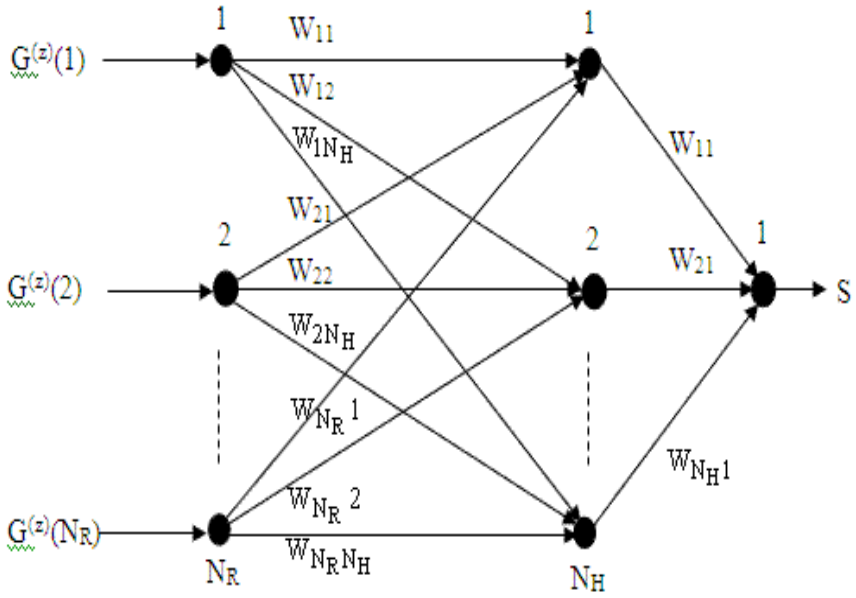


Figure 4. The multilayer feed forward neural network used in the proposed technique

$$h_q^{(1)} = \alpha + \sum_{r=1}^{N_R} W_{rq} G(r);$$
$$r = 1,2,\Lambda\ , N_R,\ q = 1,2,\Lambda\ , N_H \qquad (8)$$

$$h^{(2)} = \frac{1-e^{-h^{(1)}}}{1-e^{-2h^{(1)}}} \qquad (9)$$

$$S = h^{(2)} \qquad (10)$$

where, Eq. (8) is the basis function for the input layer and Eq. (9) and (10) are the activation functions for hidden and output layer, respectively.

1. Determine BP error using

$$e = \frac{1}{N_G} \sum_{p=0}^{N_G - 1} \left(S_T - S_p\right) \qquad (11)$$

where, e is the BP error,
S_T is the target output

2. By adjusting the weights of all the neurons based on the determined BP error, obtain new weights using

$$W^{new} = W^{old} + \Delta W \quad (12)$$

In Eq. (12), the weight to be changed ΔW depends on the rate of network learning γ and the obtained network output S_p for the p^{th} gene sequence and it is determined using the formula $\Delta W = \gamma . S_p . e$.

3. Until the BP error gets minimized to a minimum extent, repeat the process from step 2. The termination criterion for practical cases, is $e < 0.1$.

2) Classification Phase

In the classification phase, the network finds the class of a given or test gene sequence and determines the species to which it belongs. The same processes performed on the training sequence are repeated for the test sequence . Using the mined patterns and their support, the constructive nucleotide dataset is generated. Subsequent to dimensionality reduction of the generated dataset they are tested in a neural network. The neural network decides the class of the species to which the gene sequence belongs.

IV. Implementation Results

The proposed technique is implemented in the working platform of MATLAB (version 7.10) and the technique is evaluated using the DNA sequence of two different organisms, Brucella Suis and Caenorhabditis Elegans (C. Elegans). The evaluation process is performed using 10-fold cross validation test. Here, nucleotide patterns are mined with $L_{\max} = 5$. The nucleotide patterns for $l = 2 \; and \; 3$ and their corresponding support are given in Table III. In Fig. 5, different length patterns and their support are depicted and the constructive dataset that is generated from the pattern set is given in Table IV.

TABLE III. Mined nucleotide patterns from the DNA sequence of Brucella Suis and C.elegans (a) $l = 2$ and (b) $l = 3$ (a part of the pattern is given)

(a)

Species			
S. No	Pattern	*Support*	
		Brucella suis	*C-elegans*
1	aa	169042	168149
2	ag	56284	59645
3	ac	53509	54824
4	at	100894	101778
5	ga	72354	72651
6	gg	45341	45368
7	gc	46001	43023
8	gt	53423	56002
9	ca	67662	69882
10	cg	47630	40344
11	cc	47205	44205
12	ct	57377	58316
13	ta	70670	73713
14	tg	67864	71687
15	tc	73159	70695
16	tt	171584	169717

(b)

Species			
S. No	Pattern	*Support*	
		Brucella suis	*C-elegans*
1	aaa	86090	83349
2	aag	17627	18623
3	aac	18374	19422
4	aat	46951	46755
5	aga	19089	19871
6	agg	11100	10442
7	agc	11318	12264
8	agt	14777	17068
9	aca	17239	17901
10	acg	10577	9453
11	acc	10491	10676
12	act	15202	16794
13	ata	20142	21368
14	atg	15783	16752
15	atc	17406	16451
16	att	47563	47207

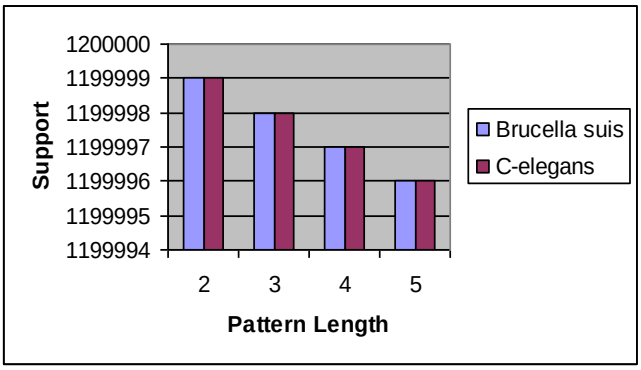


Figure 5. Support obtained for different length patterns

TABLE IV. CONSTRUCTIVE DATASET GENERATED FROM THE MINED NUCLEOTIDE PATTERNS (A) $l = 2$ AND (B) $l = 3$ (A PART OF THE PATTERN IS GIVEN).

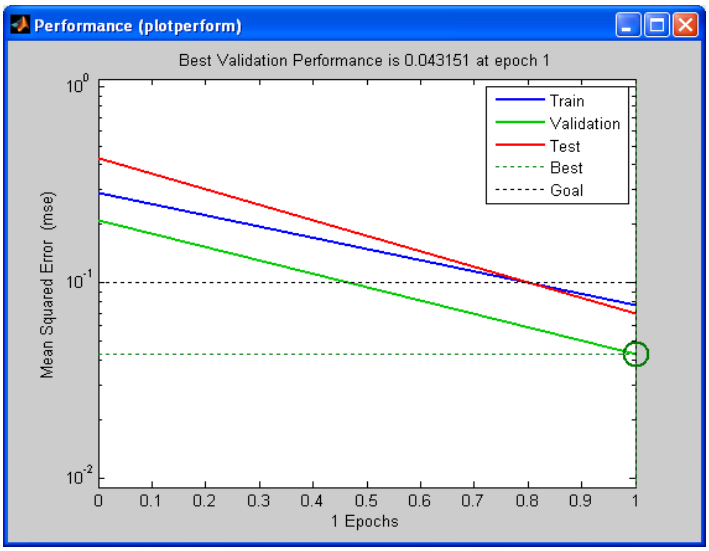
(A)

S.No		Species							
		Brucella suis				*C-elegans*			
		A	***G***	***C***	***T***	***A***	***G***	***C***	***T***
1	**A**	169042	56284	53509	100894	168149	59645	54824	101778
2	**G**	72354	45341	46001	53423	72651	45368	43023	56002
3	**C**	67662	47630	47205	57377	69882	40344	44205	58316
4	**T**	70670	67864	73159	171584	73713	71687	70695	169717

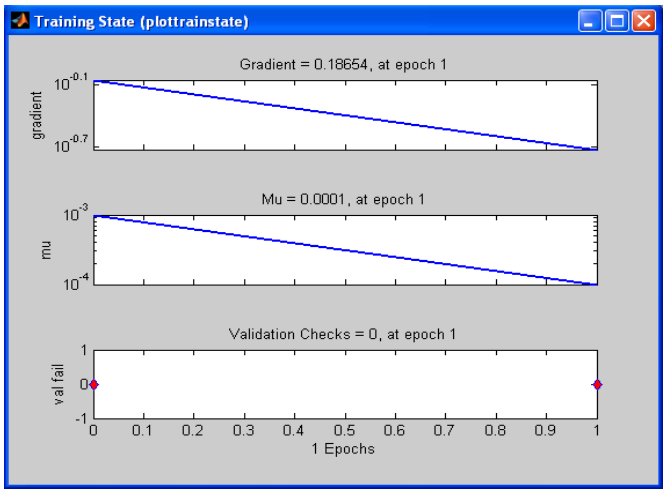
(B)

S. No		Species							
		Brucella suis				*C-elegans*			
		A	***G***	***C***	***T***	***A***	***G***	***C***	***T***
1	**AA**	86090	17627	18374	46951	83349	18623	19422	46755
2	**AG**	19089	11100	11318	14777	19871	10442	12264	17068
3	**AC**	17239	10577	10491	15202	17901	9453	10676	16794
4	**AT**	20142	15783	17406	47563	21368	16752	16451	47207
5	**GA**	31140	13379	10257	17578	31355	13670	10422	17204
6	**GG**	14729	8619	11555	10438	14982	9275	9700	11411
7	**GC**	13025	9601	12066	11309	13656	7494	9646	12227
8	**GT**	12136	12780	10458	18049	12976	12650	10000	20376
9	**CA**	26644	12400	12801	15817	27422	13532	12501	16427
10	**CG**	16283	10873	9652	10822	13747	9848	7594	9155
11	**CC**	15108	11450	9503	11144	15133	9631	9315	10126
12	**CT**	13210	12537	13724	17906	12775	13648	13226	18667
13	**TA**	25167	12878	12077	20548	26023	13819	12479	21392
14	**TG**	22253	14749	13476	17386	24051	15803	13465	18368
15	**TC**	22290	16002	15145	19722	23192	13766	14568	19169
16	**TT**	25182	26764	31571	88066	26594	28637	31018	83467

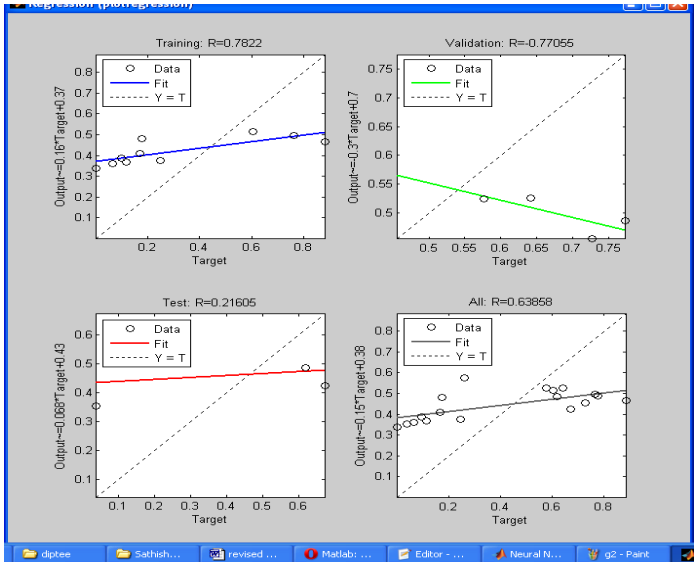
The pattern data and constructive dataset given in Tables III and IV are generated from one of the ten folds of gene sequence of Brucella Suis. Thus, from all the ten folds of gene sequence of both Brucella Suis and C. Elegans, the pattern data have been mined and constructive datasets have been generated. The generated ten folds of data are used to train the neural network. The results obtained from network training are given in Fig. 5.



(a)



(b)



(c)

Figure 6. Performance of training and test results from ANN: (a) Network performance, (b) Training evaluation and (c) Regression analysis.

Once the training process has been completed, the technique is validated using the test sequence. The results obtained from 10-fold cross validation are given in Table VI.

TABLE V. PERFORMANCE EVALUATION USING 10-FOLD CROSS VALIDATION RESULTS

Rounds in cross validation	**Species**			
	Brucella suis		***C-elegans***	
	ANN Output	***Classification Result***	***ANN Output***	***Classification Result***
1	0.2421	TP	0. 5171	TP
2	0.0769	TP	0. 6272	TP
3	0.0828	TP	0. 6361	TP
4	0.2634	TP	0. 8974	TP
5	0.2493	TP	0. 6063	TP
6	0.2613	TP	0. 0141	TN
7	0.5277	TN	0. 9163	TP
8	0.3616	TP	0. 6714	TP
9	0.5849	TN	0. 5103	TP
10	0.2143	TP	0. 5142	TP
Mean Classification Accuracy	80%		90%	

From the results, it can be seen that when a gene sequence is given to the proposed technique it identifies the corresponding species. Here, the technique is evaluated with the DNA sequence of only two genes. The technique is developed in such a way that it can be applied to any kind of DNA sequence. The test results claim that the performance of the technique reaches a satisfactory level.

V. CONCLUSION

In this paper, we have proposed a species identification technique by integrating data mining technique with artificial intelligence. Initially, the nucleotide patterns have been mined effectively. The resultant has been subjected to MPCA-based dimensionality reduction and eventually classified using a well-trained neural network. The implementation results have shown that the proposed technique effectively identifies the organism from its gene sequence and so the species. Moreover, results obtained from 10-fold cross validation have proved that the organism can be identified even from a part of the DNA sequence.

Though the technique has been tested with the DNA sequence of only two organisms, the 10-fold cross validation results have reached a remarkable performance level. From the results, it can be hypothetically analyzed that a technique, which identifies the organism only with a part of gene sequence, have the ability to classify any kind of organism and so the species.

REFERENCES

[1] P. P. Vaidyanathan and Byung-Jun Yoon, "The role of signal-processing concepts in genomics and proteomics", Journal of the Franklin Institute Genomics, Signal Processing, and Statistics Vo. 341, Issue. 1-2, pp. 111-135, January-March 2004.

[2] Achuth Sankar S. Nair, T.Mahalakshmi " Visualization Of Genomic Data Using Inter-Nucleotide Distance Signals", Silico Biology, Issue Volume 6 , 215-222, March ,2006

[3] Edward R. Dougherty, Ilya Shmulevich and Michael L. Bittner, "Genomic Signal Processing: The Salient Issues", EURASIP Journal on Applied Signal Processing, Vol. 1, pp. 146–153, 2004.

[4] Michel Tibayrenc "The species concept in parasites and other pathogens: a pragmatic approach?" TRENDS in Parasitology Vol.22 No.2 February 2006. http://www.th.ird.fr/downloads/TIP.pdf

[5] Swapnoneel Roy, Minhazur Rahman, and Ashok Kumar Thakur "Sorting Primitives and Genome Rearrangement in Bioinformatics: A Unified Perspective", World Academy of Science, Engineering and Technology, Issue 38, 2008.

[6] Mai S. Mabrouk, Nahed H. Solouma, Abou-Bakr M. Youssef, and Yasser M. Kadah "Eukaryotic Gene Prediction by an Investigation of Nonlinear Dynamical Modeling Techniques on EIIP Coded Sequences", International Journal of Biological and Life Sciences 3:4 2007

[7] Jianbo Gao, Yan Qi, Yinhe Cao, and Wen-wen Tung, "Protein Coding Sequence Identification by Simultaneously Characterizing the Periodic and Random Features of DNA Sequences", Journal of Biomedicine and Biotechnology, Vol. 2, pp. 139–146, 2005.

[8] C.L. Winder1, E. Carr2, R. Goodacre1 and R. Seviour2 "The rapid identification of Acinetobacter species using Fourier transform infrared spectroscopy" Journal of Applied Microbiology 96, 328–339,2004

[9] Francielle B. Silva, Sabrina N. Vieira, Luiz R. Goulart Filho, Julien F. C. Boodts , Ana G. Brito-Madurro and João M. Madurro "Electrochemical Investigation of Oligonucleotide-DNA Hybridization on Poly(4-Methoxyphenethylamine)",International Journal of Molecular Sciences, 9, 1173-1188,8-july,2008.

[10] Anne Jensen,Guillaume Calvayrac, Bensu Karahalil, Vilhelm A. Bohr and Tinna Stevnsner "Mammalian 8-Oxoguanine DNA Glycosylase 1 Incises 8-Oxoadenine Opposite Cytosine in Nuclei and Mitochondria, while a Different Glycosylase Incises 8-Oxoadenine Opposite Guanine in Nuclei", The journal of biological chemistry, Vol. 278, 19541-19548, 2003, DOI:10.1074/jbc.M301504200

[11] Jeremy D. Volkening, Stephen J. Spatz "Purification of DNA from the cell-associated herpesvirus Marek's disease virus for 454 pyrosequencing using micrococcal nuclease digestion and polyethylene glycol precipitation", Journal of Virological Methods, Vol. 157, p.p. 55–61. 2009.

[12] Edward R. Dougherty and Aniruddha Datta "Genomic Signal Processing: Diagnosis and Therapy", IEEE Signal Processing Magazine, Vol. 22, No. 1, p.p. 107-112, 2005, DOI: 10.1109/MSP.2005.1407722

[13] TrevorW. Fox,Alex Carreira "A Digital Signal Processing Method for Gene Prediction with Improved Noise Suppression" ,EURASIP Journal on Applied Signal Processing, Vol.1, p.p. 108–114, 2004

[14] A. Bharathi, Dr.A.M.Natarajan, "Cancer Classification of Bioinformatics data using ANOVA" International Journal of Computer Theory and Engineering, Vol. 2, No. 3, June, 2010

[15] Jayanthi Ranjan "Applications of Data Mining Techniques In Pharmaceutical Industry", Journal of Theoretical and Applied Information Technology, 2005 – 2007.

[16] Paul Hooley, Ian J. Chilton, Daron A.Fincham, Alan T.Burns and Michael P.Whitehead "Assigning Level in Data-mining Exercises", Bioscience Education Journal, Volume 9:June 2007

[17] Antony Brownea , Brian D. Hudsonb , David C. Whitleyb ,Martyn G. Fordb , Philip Pictonc "Biological data mining with neural networks:implementation and application of a flexible decision tree extraction algorithm to genomic problem domains" Neurocomputing, Volume 57, March 2004, Pages 275-293

[18] Gerd Pfeiffer, Stefan Baumgart, Jan Schr¨oder, and Manfred Schimmler "A Massively Parallel Architecture for Bioinformatics", Lecture Notes in Computer Science, 2009, Volume 5544/2009, p.p. 994-1003, DOI: 10.1007/978-3-642-01970-8_100

[19] Simon Miles, "Agent-Oriented Data Curation in Bioinformatics", In Workshop on Multi-Agent Systems in Medicine, Computational Biology, and Bioinformatics (MAS*BioMed'05), July 2005, Utrecht, Netherlands
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.124.2202&rep=rep1&type=pdf

[20] Dhaeseleer P, Liang S, Somogyi R, "Genetic network inference: from co-expression clustering to reverse engineering", Bioinformatics, Vol. 16, No. 8, pp.707—726, 2000.

[21] Kirschner M, Pujol G, Radu A, "Oligonucleotide microarray data mining: search for age-dependent gene expression", Biochemical and Biophysical Research Communications, Vol. 298, No. 5, pp. 772—778, 2002.

[22] Ponomarenko J, Merkulova T, Orlova G, Fokin O, Gorshkov E, Ponomarenko M, "Mining DNA sequences to predict sites which mutations cause genetic diseases", Knowl-based Syst, Vol. 15, No. 4, pp.225—233, 2002.

[23] Oliveira and Johnston, "Mining the schistosome DNA sequence database", Trends Parasitol, Vol. 17, No. 10, pp.501—503, 2001.

[24] Fuhrman, Cunningham, Wen, Zweiger, Seilhamer and Somogyi, "The application of Shannon entropy in the identification of putative drug targets", Biosystems, Vol. 55, pp.5—14, 2000.

[25] Arkin, Shen and Ross, "A test case of correlation metric construction of a reaction pathway from measurements", Science, Vol. 277, pp. 1275-1279, 1997.

[26] Cho and Won, "Machine learning in DNA Microarray analysis for cancer classification. In: Yi-Ping Phoebe Chen, in proceedings of the First Asia-Pacific Bioinformatics Conference. Australian Computer Society, pp. 189-198, 2003

[27] Riccardo Bellazzi and Blaz Zupan, "Towards knowledge-based gene expression data mining", Journal of Biomedical Informatics, Vol.40, pp.787-802, 2007

[28] Shital Shah and Andrew Kusiak, "Cancer gene search with data-mining and genetic algorithms", Computers in Biology and Medicine, Vol.37, pp.251-261, 2007

[29] Hemalatha and Vivekanandan, "Genetic Algorithm Based Probabilistic Motif Discovery in Unaligned Biological Sequences", Journal of Computer science, Vol.4, No.8, pp.625-630, 2008

[30] Ashraf S. Hussein, "Analysis and Visualization of Gene Expressions and Protein Structures", Journal of software, Vol.3, No.7, pp.2-11, October 2008

[31] Ahmad M. Sarhan, "Cancer Classification Based on Microarray Gene Expression Data Using DCT and Ann", Journal of Theoretical and Applied Information Technology, Vol.6, No.2, pp.208-216, 2009

[32] Valarmathie, Srinath, Ravichandran and Dinakaran, "Hybrid Fuzzy C-Means Clustering Technique for Gene Expression Data", International Journal of Research and Reviews in Applied Sciences, Vol.1, No.1, pp.33-37, October 2009

[33] Belmamoune, Potikanond and Fons J.Verbeek, "Mining and analysing spatio-temporal patterns of gene expression in an integrative database framework", Journal of Integrative Bioinformatics, Vol.7, No.3, pp.1-10, 2010

[34] Hans knutsson, "A Tensor representation of 3-D structure", 5th IEEE-ASSP and EURASIP Workshop on Multidimensional Signal Processing, The Netherlands, September, 1987.

[35] Haiping Lu Plataniotis, K.N. Venetsanopoulos, A.N., MPCA: Multilinear Principal Component Analysis of Tensor Objects, IEEE Transactions on Neural Networks, Vol.19 No.1, p.p. 18 – 39, 2008, ISSN: 1045-9227, DOI: 10.1109/TNN.2007.901277

[36] Emilio Corchado, Álvaro Herrero, " Neural Visualization of network traffic data for intrusion detection", Applied Soft Computing, Volume 11, Issue 2, March 2011, Pages 2042‑2056

[37] E.W.T. Ngai, Yong Hu, Y.H. Wong, Yijun Chen, Xin Sun**,** "The application of datamining techniques in financial fraud detection: A classification framework and an academic review of literature" Decision Support Systems, Volume 50, Issue 3, February 2011, Pages 559‑569

[38] Arzu Şencan Şahin, İsmail İlke Köse & Reşat Selba, "Comparative analysis of neural network and neuro – fuzzy system for thermodynamic properties of refrigerants" Applied Artificial Intelligence: An International Journal, Volume 26, Issue 7, 2012, DOI: 10.1080/08839514.2012.701427