Unsupervised classification of Eukaryotic DNA sequences using Multi Library Wavelet Networks

Abdesselem DAKHLI[1], Wajdi BELLIL[2], Chokri BEN AMAR[3] and Houssine TLIG[4]

[1] Department of Computer Science, REGIM, University of Gabès 6002 Gabès, Tunisia

[2] Department of Electrical Engineering, REGIM, University of Gafsa 2110 Gafsa, Tunisia

[3] Department of Electrical Engineering, REGIM, University of Sfax 3018 Sfax, Tunisia

[4] National Engineering School of Gabès, Tunisia 6002 Gabès, Tunisia

Abstract

The comparative genomics is the comparative study of the structure and function of genomes of different species. It allows to identifying, classifying and understanding the effects of selection on the organization and evolution of genomes. This new research benefits from the increasing number of sequenced genomes. Genomic sequences allow to classify organisms into different categories and classes which have significant biological knowledge and can justify the evolution and identification of unknown organisms.

Our system consists in three phases. The first phase is called transformation which is composed of three steps; binary codification of DNA sequence, Fourier Transform and Power Spectrum Signal Processing. The second phase is called approximation. This phase is empowered by the use of Multi Library Wavelet Neural Networks (MLWNN).The third phase is called classification which is realized by applying the algorithm of hierarchical classification. The results of this contribution are more interesting in comparison with some others works, in terms of rate classification using Eukaryotic organisms database.

This method permits to avoid the complexity problem of form and structure in different classes of organisms.

Keywords: *Classification, DNA sequence, Beta wavelet networks, Power Spectrum, Fourier Transform, Multi Library Wavelet Neural Networks.*

1. Introduction

A taxonomy of organisms is the science of naming, describing and classifying organisms which includes all plants, animals and microorganisms of the world on the basis of morphological, behavioural, genetic and biochemical observations. Taxonomists identify, describe and arrange species into classifications, including those that are new to scientific study. The biologists have discovered several unknown organisms that can be classified in the taxonomic hierarchy. These discoveries help to understand biological organisms during life time. Computer bodies lying in the DNA sequences explain and redirect the functions of inherited characteristics of different generations of organisms. These sequences can be processed from the raw material by biological methods of DNA sequencing. The DNA sequence is formed by a chain comprising serie of nucleotides. Each nucleotide is composed of three subunits: a phosphate group, a sugar and nucleic bases (A, T, C, G).

Classification of organisms has been studied by several researchers. Sandberg *et al.* [4] proposed a method based on Bayesian approach. The mean accuracy obtained was 85%. Francisca Z. *et al*, used Markov Model to classify proteins of microbes, eukaryotes and Archea. This classification had followed accuracy equal to 83.51%, 82.12% and 66.63% respectively for Eukaryota, Microbes and Archaea [5]. Narasimhan S. *et al.* applied Principal Component Analysis (PCA) to extract features from the genomic sequence to classify organisms. They obtained some effective results [6]. V. Karthika *et al.* in [7] proposed an approach based on an artificial neural network to solve classification problem of eight eukaryotes classes. They used the Frequency Chaos Game Representation (FCGR) to represent genomic sequences. The accuracy obtained was 92.3%. This method has been shown to be capable to recognize the class to which an unknown organism belongs using its genomic sequence. Therefore, the possibility of species classification using Chaos Game Representation (CGR) images of DNA sequences, by using different distance metrics [7] and by using neural networks [8] has been investigated as well.

Initially, each DNA sequence is represented by a vector of input neurons. Then, the networks have been formed by the

learning algorithm. Similarly, it is also used to classify sequences of DNA. This classification was presented in the form of phylogenetic groups.

The network has reached a classification accuracy of 100%. This is used to reduce the research time in a database that contains several sequences of DNA and to assist biologists to organize databases of molecular sequences.

This paper is organized as follow: in section 2 we describe an overview of the proposed approach. Section 3 presents the theory of Beta wavelet. This function will be used at Wavelet Network. Section 4 presents the simulation results of the proposed DNA sequences classification method and section 5 closes with a conclusion and discussion.

2. Proposed approach

This paper presents a new approach of classification of DNA sequence based on wavelet network using Multi Library Wavelet Neural Networks (MLWNN) to approximate f(x) of a DNA sequence. This approach is divided in three stages: transformation of DNA, approximation of the input signal and classification of compact signature DNA sequences using algorithm of hierarchical clustering.

2.1 Transformation of DNA Sequence

2.1.1 Binary Codification of DNA Sequence

2.1.1.1 DNA Sequence Components

The proposed classification of species in class is according to DNA sequence components. This sequence is formed by four basic nucleotides, adenine (A), guanine (G), cytosine (C) and thymine (T), and each organism is identified by its DNA sequence [9].
The representation of multidimensional data is an important question when we have to process data with neural networks in the field of the artificial intelligence. To reduce the complexity and to have a simple data representation we have to extract the characteristics. If the characteristics extracted are suitably chosen, it is possible to use the relevant information for the input data [8].

2.1.1.2 Feature Extraction

Linear feature extraction can be viewed as finding a set of vectors which represent effectively information content of an observation while reducing the dimensionality [10][8]. The method of indicator translates the data into digital format which can be used for DNA signal spectrum analysis. This method uses the binary number and its indicates 1 or 0 for the existence or not of a specific nucleotide at DNA sequence level [1].

Table 1 Binary encoding of nucleotides

Nucleotides	4-bit binary encoding
A	1000
C	0100
G	0010
T	0001

The binary indicator sequence is formed by replacing the individual nucleotides with values either 0 or 1. 1 stands for presence and 0 for absence of a particular nucleotide in specified location in DNA signal [4][16].

For example, if x[n] =[A A A T G TC ...], then using the values from (Tab1), we obtain:

x[n] =[1000 1000 1000 0001 0010 0001 0100 . ..]

2.1.2 Fourier Transform and Power Spectrum Signal Processing

After the genomic data have been converted into these indicator sequences, they can be manipulated with mathematical methods. The discrete Fourier Transform is applied to each indicator sequence x(n) and a new sequence of complex numbers, called $f(x)$, is obtained:

$$f(x) = \sum_{n=0}^{N-1} x(n) e^{-j\pi n/N}, k = 0,1,2,...N-1 \quad (1)$$

It is easier to work with sequence Power Spectrum, rather than original discrete Fourier Transform. The power spectrum Se[k] for frequencies *k= 0, 1, 2, . . . ,N-1* is defined as, $Se[k] = \left| f(x) \right|^2$ (2)

Se[k] has be plotted (Fig1).

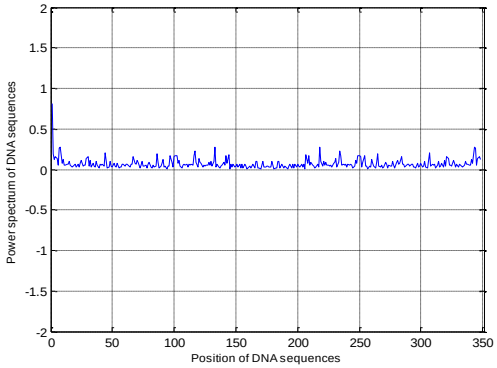

Figure 1:Signal of a DNA sequence using Power Spectrum

2.2 Approximation of DNA Signal

DNA sequences classification is a NP-complete problem. Indeed the alignment beyond two sequences, the problem quickly becomes very complex because the space of

alignment becomes very important. The recent advances of the technologies of sequencing bring today to have a consequent number of DNA sequences. We can be confronted to analyze some million sequences and a first stage for this analysis is to determine if there is a structure of the data in homogeneous groups according to a criterion to be determined.

In this paper, a classifier that classifies the DNA sequences using Fourier Transform, Power Spectrum to process the signal and the application of Beta wavelet networks as a classification model. This classifier solves the classification problems for DNA sequences. Initially, the approach can bring the learning index defined by the 1D wavelet network to develop a compact signature DNA sequences. This signature is formed by the wavelet coefficients and that will be used to match the DNA test sequences with all sequences in the training set. Then, for classification, test DNA sequence is projected onto the wavelet networks of the learning DNA sequences and new coefficients specific to this sequence are calculated (Fig2). Finally, we compare the coefficients of the learning DNA sequences with the coefficients of the test DNA sequences by computing the Correlation Coefficient. In this step we can apply the principle of hierarchical clustering to classify the characteristics of sequences DNAs.

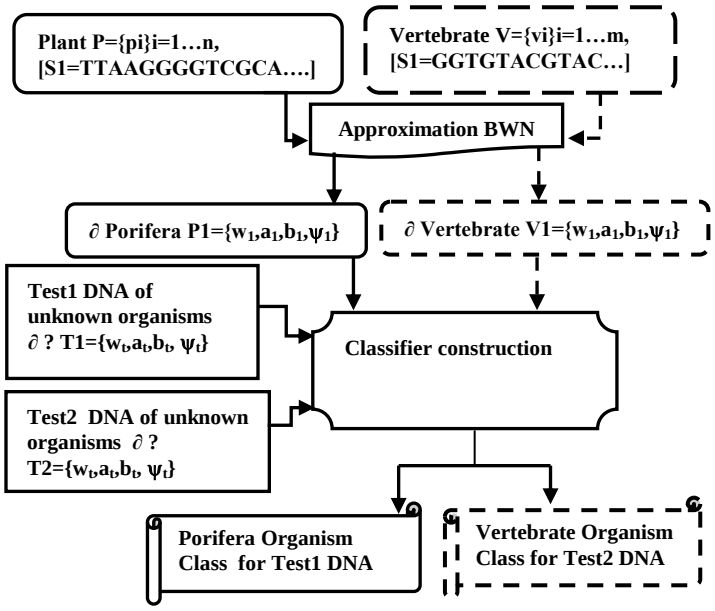

Figure 2: DNA sequence classification by BWN

to approximate f (x) of a DNA sequence must select the optimal wavelet to obtain signal representation with minimal error rate. To solve the approximation problem we use the library wavelet which contains a family wavelet. This library is called Multi Library Wavelet Neural Network Model (MLWNN). In our approach the second phase is to build the library wavelet and to approximate the function f (x) of a DNA sequence. We intend to construct a several wavelets families library for the network construction. Eatch wavelet has different dilations following different inputs. The library size is very important. This size causes difficulty to choose the optimal wavelet to construct the wavelet network.

2.3 Learning Wavelet Network using Multi Library Wavelet Neural Network (MLWNN)

In this section we will show how we can learn a wavelet network using library wavelet.

a) Proposed Learning Algorithm

Step 1: Build a library of candidate wavelet to be choose to construct the wavelet network. This wavelet is used as activation function of network. This step includes the following items:

1) Choose the mother wavelet covering all the support of the signal of DNA sequence to analyze.
2) Build a library that contains wavelets of the discret wavelet transform using dyadic sampling.
3) Choose the lowest frequency wavelet of library. This wavelets allow a coarse approximation of the signal of DNA sequence to be analyzed is introduced the first.
4) Set as a stop learning condition an error Emin between the signal f and the output of the network or a number I of wavelet used for the learning or a number j of neuron in the hidden layer of the network.
5) Each time we choose the next wavelet of the library and iterate the following steps:

Step 2: Compute the dual basis formed by the activation wavelets of the hidden layer of the network and the new selected wavelet.

Step 3: the wavelet is used as an activation function of a new neuron in the hidden layer when it creates a basic orthogonal or bi-orthogonal with the (n-1) activation wavelet of the network; else it will update the (n-1) old weights of network.

Step 4: we compute the output of the network by using the wavelet of hidden layers and the weights of connection which are already calculated.

Step 5: if the error Emin or the number of wavelets used i or the number of neuron j are reached then it's the end of learning, else another wavelet of the library is choose and we return to step2.

b) Creation of the Library Wavelet to build the library of wavelets to join our wavelet network, a sampling on a dyadic grid of dilation and translation parameters is proceded.

2.4 The Hierarchical Ascending Classification

This algorithm includes the following steps:

Step 1: Start the input by preparing a list of DNA sequences signatures and the number of classes that wants to obtain. These signatures are the outputs of the approximation 1D using wavelet networks.

Step 2: Create an empty matrix (Classes_signature) which has to contain the groups of DNA sequences.

Step 3: Starts with as each DNA sequence signature in its own cluster. This procedure starts with n classes (each DNA sequence signature forms a class containing only itself).

Step 4: Compute the similarity between classes. The covariance matrix is used to measure the similarity between the DNA sequences signatures.

Step 5: Research minimum similarity.

Step 6: Find the two classes s1 and s2 with the minimum similarity to each other.

Step 7: Merge the clusters s1 and s2 and replace s1 with the new class. Delete s2 and recalculate all similarities, which have been affected by the merge.

Step 8: Repeat step (6) and (7) until the total number of classes become one.

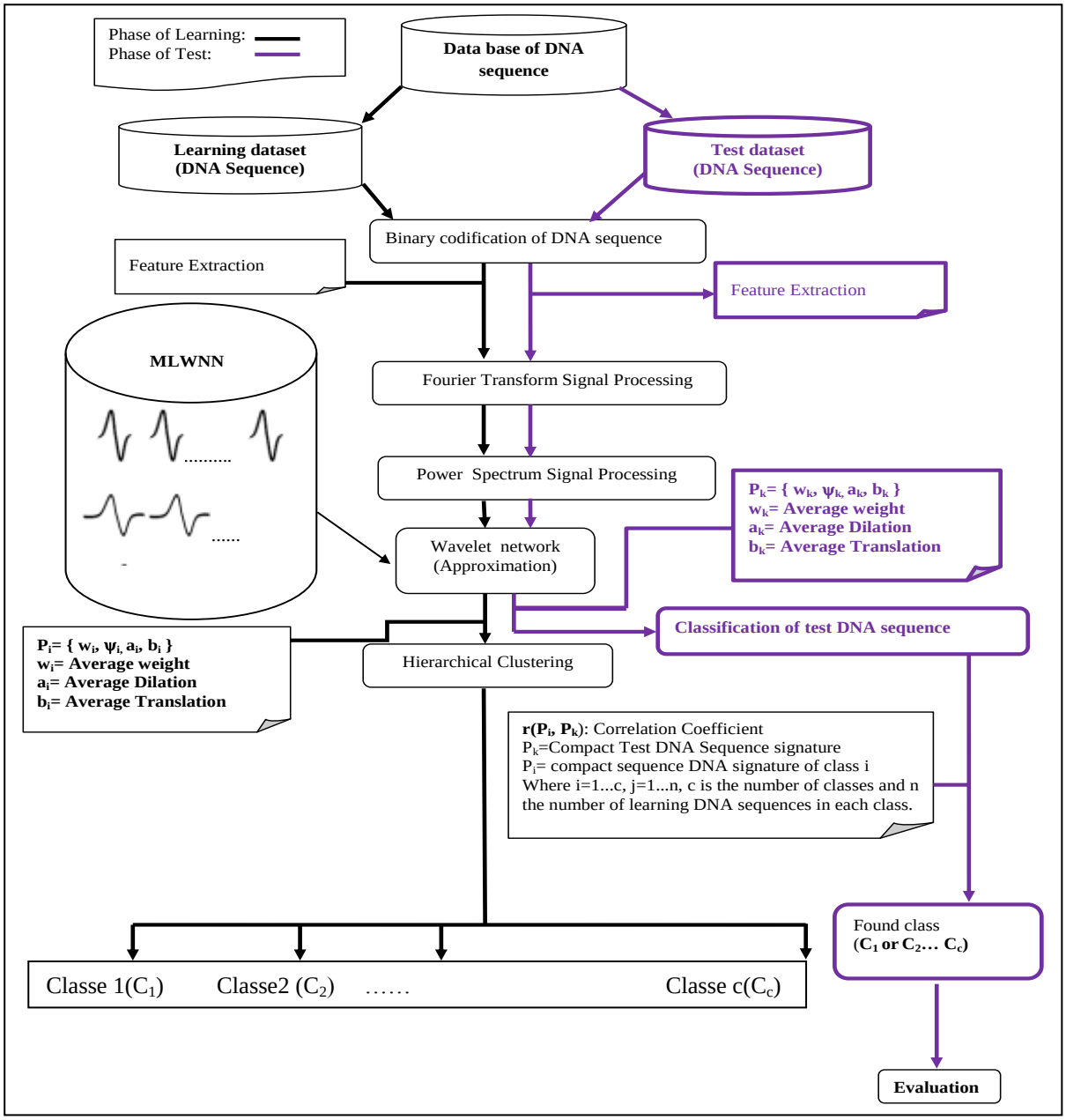

Figure 3 : Proposed Approach

3. The Beta Wavelet Family

The function beta is defined by $\beta(x)=\beta_{x0,x1,p,q}(x)$[17][18][20], x0 and x1 are real parameters. Where x0<x1

$$\beta(x,p,q,x_0,x_1)=\begin{cases}\left(\frac{x-x_0}{x_c-x_0}\right)^p\left(\frac{x-x_1}{x_1-x_c}\right)^q & if\ \ x\in[x_0,x_1]\\ 0 & \text{otherwise}\end{cases} \quad (3)$$

$$\text{Where } x_c=\frac{p\,x_1-x_0}{p+q} \quad (4)$$

We have proved in [20][21[22][23] that all derivatives of Beta function $\in$ L2($\Re$) and are of class C_∞. The general form of the nth derivative of Beta function is:

$$\psi_n(x)=\frac{d^n\beta(x)}{dx^n}=\left[(-1)^n\frac{n!\,p}{(x-x_0)^{n+1}}+\frac{n!q}{(x_1-x)^{n+1}}\right]\beta(x)+P_n(x)P_1(x)\beta(x)+\sum_{i=1}^{n}C_n^i\left[(-1)^n\frac{(n-i)!p}{(x-x_0)^{n+1-i}}+\frac{(n-i)!q}{(x_1-x)^{n+1-i}}\right]\times P_1(x)\beta(x) \quad (5)$$

$$\text{where : } P_1(x)=\frac{P}{x-x_0}-\frac{q}{x_1-x}; \quad (6)$$

$$P_n(x)=(-1)^n\frac{n!\,p}{(x-x_0)^{n+1}}-\frac{n!q}{(x_1-x)^{n+1}} \quad (7)$$

If p = q, for all n $\in$ N and 0 < n < p, the functions

$\Psi n(x) = d^n\beta(x)/dx^n$ are wavelets [21][22][23]. The first, second and third derivatives of Beta wavelet.

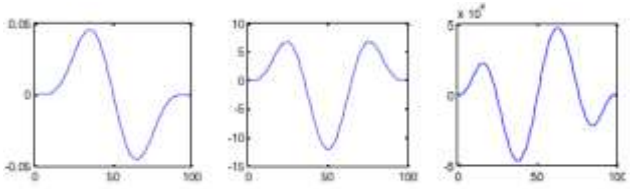

Figure 4 : First, second and third derivatives of Beta function.

3.1 Wavelet Network

The combination of wavelet transform and artificial neuron networks defines the concept of wavelet networks. This network uses the wavelet functions instead of the traditional sigmoid function as the transfer function of each neuron. This type is composed of two layers (input layer and hidden layer).

It has the same structure as architecture radial function. The salaries of weighted outputs are using an adder. Each neuron is connected to the other of the following layer. Wavelet network (Fig5) is defined by pondering a set of wavelets dilated and translated from one mother wavelet with weight values to approximate a given signal f.

$$Y=\sum_{i=1}^{N_w}\omega(a,b)\Psi\left(\frac{x-b}{a}\right)+\sum_{k=0}^{N_i}a_k\,x_k \quad (8)$$

Where y is there the output of the network, (x_1, x_2,…,x_{Ni}) is the vector of the input and Nw is number of wavelets.

It is often useful to consider, besides the decomposition of wavelets cleanly so-called, that the output can have a component refines in relation to the variables, of coefficients ak (k = 0, 1... N_i). (Fig5)

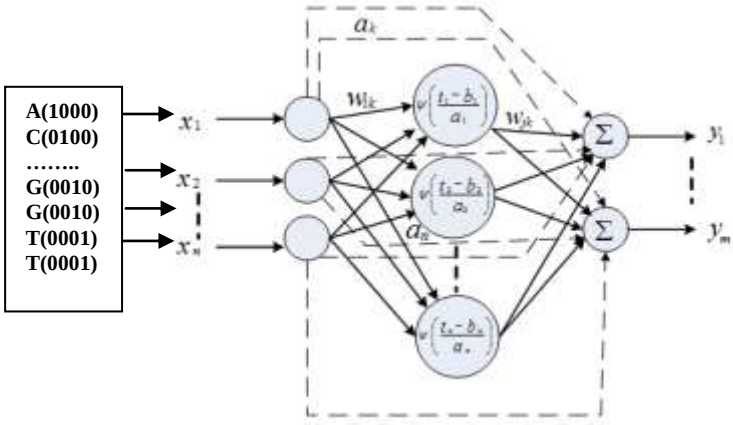

Figure 5 : Wavelet network network

4. Experiment and results

To evaluate the performance of our approach, we have developed different experiments, each consisting of a different subset of test data. The mitochondrial DNA sequences of 1417 Eukaryotic organisms belonging to 8 categories of taxonomical Hierarchy was obtained from NCBI Organelle database [26]. From these 1417 data, 709 are used for training and 708 are used for testing.

In the learning steps and test 300 was taken DNA sequence for each genome. These DNA strands are made to train and

test our network wavelet Beta. And during the test was calculated precision and time of classification developed by our network wavelet beta. The execution, calculations and processing of the results is done by a PC that has the configuration (Intel (R) Core (TM) i3 CPU M370 @ 2, 40 GHz). In this section, we present some experimental results of classification of DNA sequences by using the Fourier transform, Power Spectrum and applying the Beta wavelet networks on approximating three 1-D functions.

Table 2 : Distribution of available data into training and testing set of DNA sequence

Classes	Total	Training	Test
plant	30	15	15
porifera	21	11	10
protostomia	256	128	128
vertebrata	1024	512	512
cnidaria	34	17	17
Fungi	52	26	26
Total	1417	709	708

In the phase of learning our system gets ready to distinguish the various classes by means of the examples of learning of DNA sequences. Our system builds a model for every DNA sequence of learning.

At the beginning, during the phase of approximation our model tried to decompose the input signal for every sequence and at the end it tried to reconstruct the input signal. The estimation of the performance of this phase we measured by the Mean Square Error (MSE). (Tab4) shows that the Mean Square Error(MSE) obtained are low (0.00137498) and the run time increases relatively with size of the DNA sequence.

The result shows that the size of DNA sequence increases the time of the training phase. This time believes according to the size of a DNA sequence. When the size is equal to 1467 the training time equal to 178,080 seconds. To solve the approximation problem we use the library wavelet which contains a family wavelet. This library is called Multi Library Wavelet Neural Network Model (MLWNN). The library contains 6 mother wavelets (Beta1, Beta2, Beta3, Mexican4 hat, Polywog5 and Slog6)(Tab3).

Table 3: Selected mother wavelets and Normalized Root Mean Square Error(NRMSE)

DNA sequence for each Class	Size	Beta1 wavelet	Beta2 wavelet	Beta3 wavelet	Mexhat4 wavelet	Slog5 wavelet	Polywog6 wavelet	*NSRMSE*
Plant(S1)	597	2	4	0	1	2	1	***0.625769***
Porifera(S2)	421	1	4	1	2	0	2	***0.71263***
Protostomia(S3)	635	1	2	1	1	2	3	***0.73648***
Vertebrata(S4)	1467	3	2	2	2	1	0	***0.579626***
Cnidaria(S5)	557	1	3	0	1	1	4	***0.635123***
Fungi(S6)	579	2	2	0	1	0	5	***0.630764***

Table 4: MSE of approximation of the signal for DNA

DNA sequence for each Class (NCBI Organelle database)	Size	MSE (Mean Square Error)	Training Time(sec)
plant	597	0.00357234	69.607
porifera	421	0.00137498	48.297
Protostomia	635	0.010413	63.279
vertebrata	1467	0.109285	178.082
cnidaria	557	0.00444942	65.02
Fungi	579	0.0052393	86.594

Table 5: Confusion matrix

		Predicted class		
		Negative	**Positive**	**Total**
Actual Class	**Negative**	a	b	a+b
	Positive	c	d	c+d

A confusion matrix contains information about actual and predicted classifications done by our classification system.

- **a** is the number of correct predictions that an instance is negative;
- **b** is the number of incorrect predictions that an instance is positive;
- **c** is the number of incorrect predictions that an instance is negative;
- d is the number of correct predictions that an instance is positive;

Table 6: Classification rate of our approach for NCBI Organelle database

	Predicted Classes						
Actual Classes	**1**	**2**	**3**	**4**	**5**	**6**	**Classification Rate (%)**
1 plant	296	0	0	0	0	4	**98,666**
2 porifera	0	285	5	10	0	0	**95**
3 protostomia	0	0	300	0	0	0	**100**
4 vertebrata	0	5	5	280	5	5	**93,333**
5 cnidaria	0	15	1	2	282	0	**94**
6 Fungi	10	0	0	0	0	290	**96,666**
User Accuracy (Recall)%	***96.732***	***93.443***	***96.463***	***95.89***	***98.258***	***96.99***	
Overall accuracy %	**96,2775**						

- **Accuracy**= $(a+d)/(a+b+c+d)$ (9)

Accuracy=(296+285+300+280+282+290)/1800=0,96277778

Table 7: Accuracy for each class provided by PNN and WNN with 64 of DNA sequence.

Class	Accuracy obtained using Probabilistic neural network (PNN)(%)	Accuracy obtained using Wavelet Neural Network (WNN)(%)
plant	73.3	98,666
porifera	80.0	95
protostomia	86.7	100
vertebrata	96.8	93,333
cnidaria	94.12	94
Fungi	65.38	96,666
Average	*82.71*	*96,2775*

The (Tab6) and (Tab7) shows the distribution of the good classifications by class as well as the rate of global classification for all the sequences of the validation phase. These results were obtained using networks wavelets. After the results shown in this table, we can release the following interpretations:

- The sequences of class “protostomia” are perfectly classified. The classes "plant", the "porifera" and "Fungi" present the best rates of classification.
- We can indicate some errors between for example the class “Fungi" and the class "plant" also between the class “Porifera" and the class “vertebrata”. This error can be due to the similarity which exists between its sequences of DNA. These similarities explain the hereditary aspects between the organisms and afterward we can explain and interpret biologically the evolution of the living organisms during time and we can explain the mutual relationships between organisms.
- These evolutions start by the transformation or the change hastens at the level of DNA sequences because of the internal reasons (bad replication of DNA) and reasons extern for example the toxic product and the pollution. This transformation explains biologically the diversity of the living organisms beings in the nature.

(Tab7) shows that the classification using this approach based on the networks of wavelets stupid man has good results. The performance of this classifier is measured by the calculation of accuracy classification. These measures showed the performance of this approach by comparing with the network model of neurons based on the probability (PNN). Our approach had an average accuracy is equal to 96,2775%.

Also the classification was executed during average time equal to 0.266 seconds and we noticed that the speed of classification depends on the length and on the composition of DNA sequences.
Results of comparison have shown that the WNN model performs better than the classical PNN model in the context of training run time and classification rate. We compare the coefficients of the learning DNA sequences with the coefficients of the test DNA sequences by computing the Correlation Coefficient. In this step we can apply the principle of hierarchical clustering to classify the characteristics of sequences DNAs.
Hierarchical Ascending Classification (HAC) has for objective to group objects in a number restricted by homogeneous classes. HAC proceeds to groupings step by step successive.

At each step, the two DNA sequences, or two DNA sequences classes, the closest, according to a criterion of aggregation state, are combined to form a new class.

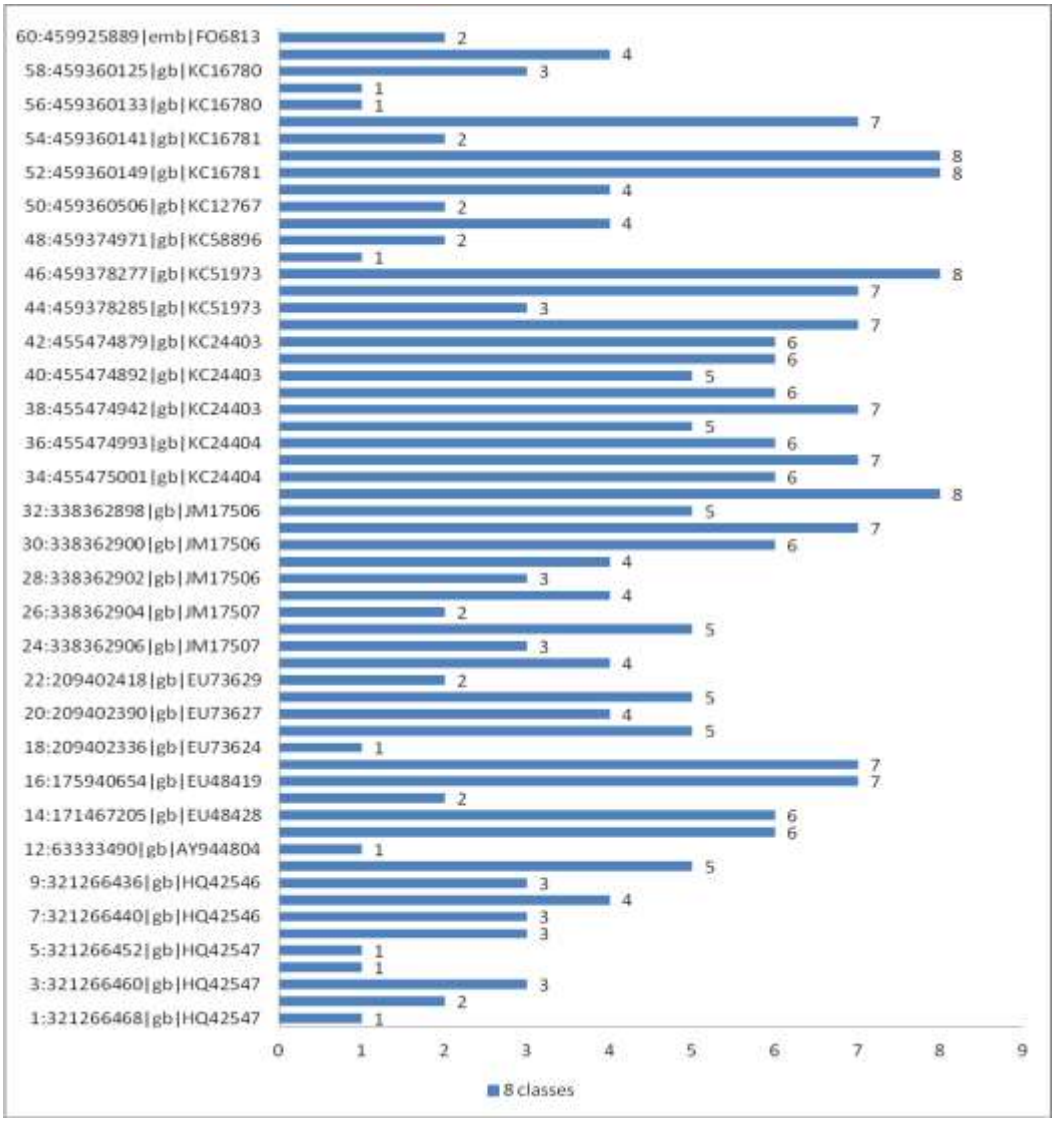

Figure 6: Graphical representation of DNA groups and combinations of groups.

The symbol * indicates that the sequences of class 1 have the same nucleotide at the given position.

The symbol - Genetic Mutations.

Table 8: Alignement of DNA groups of class 1.

			*				*	*			*		*	*			*		*	*	*	*				*	*		
Sequence 1	-	-	A	T	T	A	T	A	T	T	T	-	A	A	T	A	T	T	C	G	G	A	G	C	-	T	T	T	T
Sequence 4	-	-	A	T	T	A	T	A	T	T	T	-	A	A	T	A	T	T	C	G	G	A	G	C	C	T	T	T	T
Sequence 5	-	-	A	T	T	A	T	A	T	T	T	-	A	A	T	A	T	T	C	G	G	A	G	C	C	T	T	T	T
Sequence 12	T	T	A	C	C	G	T	A	G	G	T	G	A	A	-	C	T	G	C	G	G	A	A	G	G	T	T	-	-

(Tab8) shows that the sequences of class DNA have biologic similarity. This resemblance can assure a similar biologic function to the level of the cells of the organisms. The results obtained is very positve and suggests that the proposed techniques can be successfully used in resolution of problem of classification of DNA sequences. This approach allows to calalogue and to characterize any group of living organisms.from one such alignment of DNA sequences we can estimate the place of every body in the family tree to be alive (the tree of life). The similarity of sequences allows to improve in a significant biology way on clusterings of the family of current DNA sequences which are incapable to affect directly the considerable mass of these data.

5. Conclusion

In this paper, we used a new method of training called Library Wavelet Neural Network Model (MLWNN) which is used to construct Wavelet Neural Network (WNN). The WNN is used to approximate the function f (x) of a DNA sequence signal. Our method depends on Binary codification, Fourier Transform and Power Spectrum to process the DNA sequence signal. Applying this hierarchical classification allows us to group the similar DNA sequences according to certain criteria. This classification aims at distributing n sequences of DNA, characterized by p variables X1, X2…Xp in a number m of subgroups which are homogeneous as much as possible where every group is differentiated well from the others.

In our approach we used the Correlation Coefficient or Pearson Correlation Coefficient which is applied to measure of association between two vectors of DNA sequences signal.

Our approach allows us to classify organisms into different categories and classes which have significant biological knowledge and can justify the evolution and identification of unknown organisms. Also they study mutual relations between organisms. This classification will allow the study of living organisms. Classification of organisms is significant not only in the study of evolutionary properties of organisms but also in the study of mutual relationships between organisms and the specific identification of a previously unknown organism.

In this article we noticed that our approach gives rates of classification (96%) better than that given by other approaches proposed by other researchers. Simulation results are demonstrated to validate the generalization ability and efficiency of the proposed Wavelet Neural Network Model.

These results were realized thanks to many capacities listed as follows;

– The capacity of wavelets to decompose the signals supported by the criterion analyzes time – frequency

– The capacity of Library Wavelet Neural Network Model (MLWNN) to construct Wavelet Neural Network (WNN)

– The capacity of Binary Codification, Fourier Transform and Power Spectrum to process the signal of DNA sequences,

– The capacity of model multi-entrances multi-taken out to manage the input which has very long sizes.

– The capacity of the networks of wavelets in approximer of the functions real gives a complex.

Acknowledgment

I would like to to thank Research Group on Intelligent Machines (REGIM).

References

[1] M. Ahmad, A. Abdullah and K. Buragga, “A novel optimized approach for gene identification in DNA sequence” Asian Network for Scientific Information, Journal of Applied Sciences 11 (5), p.806-814, 2011.

[2] S. Brunak, J. Engelbrecht, and S.Knudsen, “Prediction of human mRNA donor and acceptor sites from the dna sequence,” Journal of Molecular Biology, vol. 220, pp. 49-65, Jul. 1991.

[3] A.Vincent Emanuele II, T. Thao Tran, and G. Tong Zhou “a forurier product method for detecting approximate tandom repeats in dna”, Scholl of Electrical and Computer Engineering Georgia Institute of Technology, Atlanta, GA 30332-0250 USA, p. 1390 - 1395, Jul. 2005.

[4] R.Sandberg , G. Winberg, C.I. Branden, A. Kaske , I. Ernberg and Coster, “Capturing Whole - Genome characteristics in short sequences using a naive Bayesian classifer”, Genome Res., Vol. 11, pp. 1404-09, May 2001

[5] F. Zanoguera and M. de Francesco, “Protein classification into domains of life using Markov chain models”, Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference, 0-7695-2194-0/04 , 2004..

[6] S.Narasimhan , S. Sen and Konar, “Species identification based on mitochondrial genomes”, ICCR 2005, International Conference of Cognition and Recognition, Mysore, India, 22-23 Dec. 2005.

[7] K. Vijayan, V. Vrinda Nair and P.Deepa Gopinath “Classification of Organisms using Frequency-Chaos Game Representation of Genomic Sequences and ANN”, 10th National Conference on Technological Trends (NCTT09) 6-7 Nov 2009

[8] W. Cathy, M. Berry, Y.-S. Fung and J. McLarty, “Neural Networks For Molecular Sequence Classification” , Proc Int Conf Intell Syst Mol Biol., p. 429-437, 1993.

[9] S.S. Kumar and N.Duraipandian ,” Artificial Neural Network Based Method for Classification of Gene Expression Data of Human Diseases along with Privacy Preserving”, International Journal of Computers & Technology, Volume 4 No. 2, March-April, 2013, ISSN 2277-3061.

[10] L. Valim de Freitas et A. Paula Barbosa Rodrigues de Freitas , «L'analyse multivariée dans la gestion, l'ingénierie et les sciences" , livre édité par, ISBN 978-953-51-0921-1, parution: 9 Janvier 2013 sous licence CC BY 3.0

[11] S. Narasimhan , S. Sen and Konar, “Species identification based on mitochondrial genomes”, International Conference of Cognition and Recognition, Mysore, India, 22-23 Dec. 2005.

[12] K. Vijayan, Vrinda V. Nair and D. P. Gopinath “Classification of Organisms using Frequency-Chaos Game Representation of Genomic Sequences and ANN”, 10th National Conference on Technological Trends (NCTT09) 6-7 Nov 2009

[13] W. Cathy, M. Berry, Y.S. Fung and J. McLarty, “Neural Networks For Molecular Sequence Classification” , Proc Int Conf Intell Syst Mol Biol., p. 429-437, 1993

[14] S. S. Kumar and N.Duraipandian ,”An Effective Identification of Species from DNA Sequence: A Classification Technique by Integrating DM and ANN”, (IJACSA) International Journal of

Advanced Computer Science and Applications, Vol. 3, No.8,p.104-114, 2012.

[15] L. Valim de Freitas et A. P. Barbosa Rodrigues de Freitas , L'analyse multivariée dans la gestion, l'ingénierie et les sciences , ISBN 978-953-51-0921-1, parution: 2013 sous licence CC BY 3.0

[16] S. bai Arniker and H. Keung Kwan, "Advanced Numerical Representation of DNA Sequences", International Conference on Bioscience, Biochemistry and Bioinformatics IPCBEE vol.3 1,p.1-5, 2012

[17] W. Bellil, C. Ben Amar and A.M. Alimi, "Beta Wavelet Based Image Compression", International Conference on Signal, System and Design, SSD03, Tunisia, vol. 1, pp. 77-82, Mars, 2003

[18] W. Bellil, C. Ben Amar and M.A Alimi, "Synthesis of wavelet filters using wavelet neural networks", Transactions on Engineering, Computation and Technology, vol. 13 . ISSN 1305-5313, pp 108-111, 2006.

[19] S.S. Iyengar, E.C. Cho and V.Phoha, "Foundation of Wavelet Nework and Application", Chapman and Hall/CRC Press, June 2002.

[20] C. Ben Amar, M. Zaied and M. A. Alimi, "Beta wavelets. Synthesis and application to lossy image compression", Journal of Advances in Engineering Software, Elsevier Edition, Vol. 36, N7, pages 459 – 474, 2005.

[21] W. Bellil, C. Ben Amar and M.A Alimi, "Synthesis of wavelet filters using wavelet neural networks", Transactions on Engineering, Computation and Technology, vol. 13 . ISSN 1305-5313, pp 108-111, 2006.

[22] C. Ben Amar, W. Bellil, M.A. Alimi, "Beta Function and its Derivatives: A New Wavelet Family", Transactions on Systems, Signals and Devices, Vol.1, Number 3, p.275-293, 2005-2006.

[23] W. Bellil, C. Ben Amar C. And A.M. Alimi, "Beta wavelets networks for function approximation", International Conference on Adaptative and Natural Computing Algorithms, ICANNGA05, Coimbra Portugal, SpringerWien NewYork, p. 18-21, 2005.

[24] V. Vrinda Nair, P. Lissy Anto and A. Nair, "Naive Bayesian Classification of unknown sequence fragments based on chaos game representation of mitochondrial genomes", Communications of SIWN, vol 7, pp:27-33, May 2009.

[25] Q. Wang and all. "Naive Bayesian Classifier for Rapid *Assignment of rRNA Sequences into the New Bacterial Taxonomy*", applied and environmental mictobiologyGY, Vol. 73, No. 16, p. 5261–5267,Aug. 2007.

[26] www.ncbi.nlm.nih.gov

Abdesselem dakhli continued these academic studies to FSEG Sfax ,Tunisia. He obtained his teacher's cerificate in data processing applied to management in June 2001. He continued his degree of Masters in ISIMG of Gabes, Tunisia in 2008. In 2010, he received his Master degree in Information system in the same Institute. In August 2005 he was assigned to the ISG Gabes, Tunisia. Currently, he is a teacher at ISG. In 2012, now he prepares a Phd thesis of bioinformatic in ENIS. Abdesselem dakhli he also participated in one internationnale conference. His areas of research are: Tomography, Bioinformatics.

Wajdi BELLIL received the B.S. degree in Electrical Engineering from the National Engineering School of Sfax (ENIS) in 2000, the M.S. and PhD degrees in Electrical Engineering from the National Engineering School of Sfax (ENIS), in 2003 and 2009, respectively. He spent five years at the ISET Gafsa, Tunisia, as a technologic assistant and researcher before joining the faculty of Science of Gafsa, Tunisia, as Assistant. He joined the Higher Institute of Applied Sciences and Technology, Gafsa University, where he is currently an assistant professor in the Department of computer science. He was a member of the REsearch Group on Intelligent Machines (REGIM). He is a junior member of IEEE.

Chokri BEN AMAR received the B.S. degree in Electrical Engineering from the National Engineering School of Sfax (ENIS) in 1989, the M.S. and PhD degrees in Computer Engineering from the National Institute of Applied Sciences in Lyon, France, in 1990 and 1994, respectively. He spent one year at the University of "Haute Savoie" (France) as a teaching assistant and researcher before joining the higher School of Sciences and Techniques of Tunis as Assistant Professor in 1995. In 1999, he joined the Sfax University (USS), where he is currently a professor in the Department of Electrical Engineering of the National Engineering School of Sfax (ENIS), He is a senior member of IEEE, and the chair of the IEEE SPS Tunisia Chapter since 2009. He was the chair of the IEEE NGNS'2011 (IEEE Third International Conference on Next Generation Networks and Services) and the Workshop on Intelligent Machines.

Houssine TLIG Received his Ph.D. (Quantitative Methods) degree in 2012 from university of Sfax , Tunisia. He is presently working as Assistant in the National Engineering School of Gabes, Tunisia. His research interest includes Optimization, Fuzzy Sets, Stochastic Systems and Data minin