Different Species and Proteins Classifiers and Protein's Structure Predictors Systems

Roaa I. Mubark, Hesham A. Keshk, and Mohamed I. Eladawy

***Abstract*—** Because the experimental techniques that have been used to determine protein structure such as the x-ray crystallography and Nuclear Magnetic Resonance "NMR" spectroscopy are very expensive and cannot be applied all the time, so the prediction may be the way to get the protein structure. Most of previous works in the field of protein structure prediction given a certain protein sequence works on a database of proteins from different species. In this proposed work we will use a given protein sequence such as hemoglobin, insulin, and albumin to recognize first the species that this sequence belongs to. Knowing the species to which the sequence belongs will give better results in predicting the structure of that sequence, either the 3D or the secondary structures. Knowing the species can even help in the correct recognition of the protein type given the protein sequence. In this paper we used the neural network, Hidden Markov model, and Euclidean distance techniques in the classification and prediction processes.

***Keywords*—**Bioinformatics, Protein prediction, Hidden Markov Model, and Neural network.

I. Introduction

THE recent revolution in genomics and bioinformatics has caught the world by storm. From company boardrooms to political summits, the issues surrounding the human genome, including the analysis of genetic variation, access to genetic information and the privacy of the individual have fueled public debate and extended way beyond the scientific and technical literature [1].

During the past few years, bioinformatics, defined as the computational handling and processing of genetic information, has become one of the most highly visible fields of modern science [2].

One of the most important applications of bioinformatics is the prediction of protein structure. The protein structure prediction has been an active research area during the last few years or so [1].

The technological progress in computational molecular biology during the last decade has contributed significantly to the progress we see today [2]. The major goal of predicting protein structures underpins the correct assumption that three-dimensional structures confer protein function. The linear amino acids sequences must transform to non-linear secondary structures and then to 3D and 4D structures that are responsible for biological functions [3].

Illustrating our paper, may need to define basics in human genome such as DNA, chromosome, RNA, protein, protein structure, hemoglobin, insulin and albumin. DNA code is a sequence of chemicals that form information that control how humans are made and how they work. This encoding information in an ordered sequence of 4 different symbols called "bases", typically denoted A, C, G, and T [3]. These 4 substances are the fundamental "bits" of information in the genetic code, and are called "base pairs" because there is actually 2 substances per "bit" for instance,

C-G-A-T-T-G-C-A-A-C-G-A-T-G-C
| | | | | | | | | | | | | | |
G-C-T-A-A-C-G-T-T-G-C-T-A-C-G

The entirety of human DNA code, called the "human genome", is about 3 million bases in total. Every human being has 2 copies of this code, one copy from each parent, so a human's cell DNA contains a total of around 6 billion bases. These 6 billion odd base pairs are split amongst 46 chromosomes. Each person gets 2 pairs of chromosomes, 23 from each parent, to total 46 chromosomes per human cell [4].

RNA is a more temporary form that is used to process subsequences of DNA messages. RNA is an intermediate form used to execute the portions of DNA that a cell is using. For example, in the synthesis of proteins, DNA is copied to RNA, which is then used to create proteins: DNA->RNA->Proteins.

The structure of DNA and RNA are very similar. They are both ordered sequences of 4 types of substances: ACGT for DNA and ACGU for RNA. Thus RNA uses the same three ACG substances, but uses U (uracil) instead of T (thymine) [5].

The processes that are involved in making proteins from our genes are called transcription and translation and the molecules that are involved in these processes are called DNA, mRNA, tRNA and proteins as shown in Fig. 1. The

This work was supported in part by Helwan University, faculty of engineering, Department of Electronics, Communications, and Computer. Helwan, Egypt.

R. I. Mubark is with Electronics, Communication and Computer Engineering Department, Helwan University, Helwan, Egypt (phone: 202-275-50798; e-mail: roaim79@ yahoo.com).

H. A. Keshk is with Electronics, Communication and Computer Engineering Department, Helwan University, Helwan, Egypt (e-mail: h_keshk@hotmail.com).

M. I. Eladawy is with Electronics, Communication and Computer Engineering Department, Helwan University, Helwan, Egypt (e-mail: mohamed @eladawy.com).

order of information transfer is DNA to mRNA to protein [6], [7].

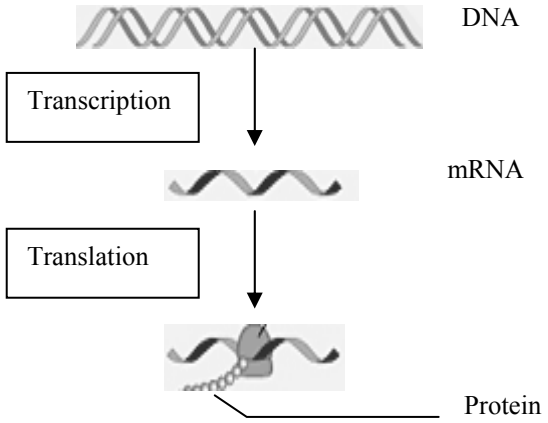


Fig. 1 the transformation of DNA into protein.

All proteins are substances made up of only 20 basic building blocks called amino acids. Proteins are ordered sequences of these 20 amino acids. Proteins have a complex 3D structure. Proteins are chains of 20 different types of amino acids, which in principle can be joined together in any linear order, sometimes called poly-peptide chains. This sequence of amino acids is known as the primary structure, and it can be represented as a string of 20 different symbols.

The length of the protein molecule can vary from few to many thousands of amino acids. For example insulin is a small protein and it consists of 51 amino acids, while titin has 28,000 amino acids. Although the primary structure of a protein is linear, the molecule is not straight, and the sequence of the amino acids affects the folding.

There are two common substructures often seen within folded chains: alpha-helices and beta-strands. They are typically joined by less regular structures, called loops. These three are called secondary structure elements. As the result of the folding, parts of a protein molecule chain come into contact with each other and various attractive or repulsive forces (hydrogen bonds, disulfide bridges, attractions between positive and negative charges, and hydrophobic and hydrophilic forces) between such parts cause the molecule to adopt a fixed relatively stable 3D structure [8].

This is called tertiary structure. In many cases the 3D structure is quite compact. Protein 3D structural domains are often associated with a particular protein function also the structure contains a valuable information to the biologists instead of the meaningless sequence [9]. Because the experimental techniques that used to determine protein structure such as the x-ray crystallography and Nuclear Magnetic Resonance "NMR" spectroscopy are very expensive and can not be applied all the time, so the prediction may be the way to get the protein structure [10].

Hemoglobin is a protein-based component of red blood cells which is primarily responsible for transferring oxygen from the lungs to the rest of the body. Hemoglobin is actually the reason red blood cells appear red, although oxygen-rich blood is noticeably brighter than the depleted blood returning to the heart and lungs. Fresh hemoglobin is produced in the bone marrow as needed. Insulin is a hormone produced by the pancreas that regulates the level of glucose, a simple sugar that provides energy, in the blood. And finally albumin is a protein found in blood plasma and urine, which can be a sign of kidney disease [6].

This paper is organized as follows: Part 2 will introduce the classification of different species based on a certain protein sequence, for example we will use the hemoglobin sequence to differentiate between 13 different species. Part 3 will do approximately the opposite of part 2, where we differentiate between different proteins within a certain species, human for example. Part 4 is dedicated to the prediction of the 3D and secondary structures of different proteins.

II. Different Species Classifier Based on Protein Sequences

The complete genome sequences of some proteins provide an excellent basis for studying the clustering of different species. First we will consider using the hemoglobin to identify the following 13 species: human, horse, wolf, donkey, chicken, clam lucina, Glycera Dibranchiata ,tuna fish, trout fish, hagfish, rice plant, and two different bacteria's; mycobacterium Tuberculosis Trhbn and gutless beard worm Oligobrachia Mashikoi [11]. Insulin will be used to identify the following four species: human, fruit fly, Norway rat and cow. And finally albumin will be considered to identify the following 11 species: human, chicken, snail, sunflower, rape, pea, winged bean, castor bean and two different bacteria's; Finegoldia magna and Streptococcus sp. We down loaded the used protein sequences, hemoglobin, insulin and albumin from the National Center for Biotechnology Information "NCBI" database [6].

We will introduce two different classifiers; one of them based on neural network and the other based on the Euclidean distance technique. Each one of them will be illustrate in details [11].

A. Neural Network

Neural networks are composed of simple elements operating in parallel. These elements are inspired by biological nervous systems. As in nature, the network function is determined largely by the connections between elements. We can train a neural network to perform a particular function by adjusting the values of the connections (weights) between elements. Commonly neural networks are adjusted, or trained, so that a particular input leads to a specific target output [12], [13]. Such a situation is shown below.

There, the network is adjusted, based on a comparison of the output and the target, until the network output matches the target. Typically many such input/target pairs are used, in this supervised learning, to train a network as shown in Fig. 2 [13].

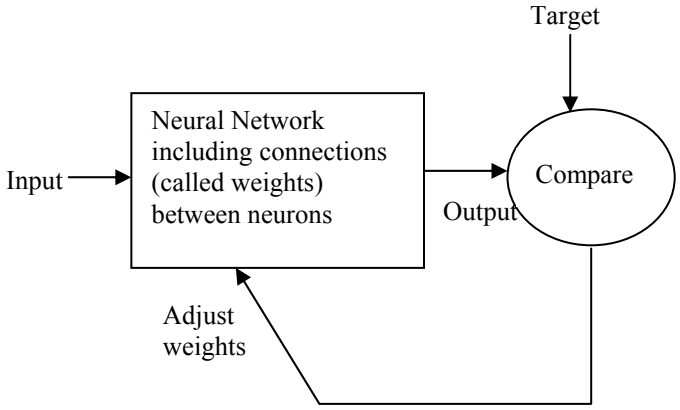


Fig. 2 the Neural Network.

The multi-layer back-propagation networks have been selected to classify different species because; the properly trained of these networks tend to give reasonable answers when presented with inputs that they have never seen. The multi-layer back-propagation network is shown in Fig. 3.

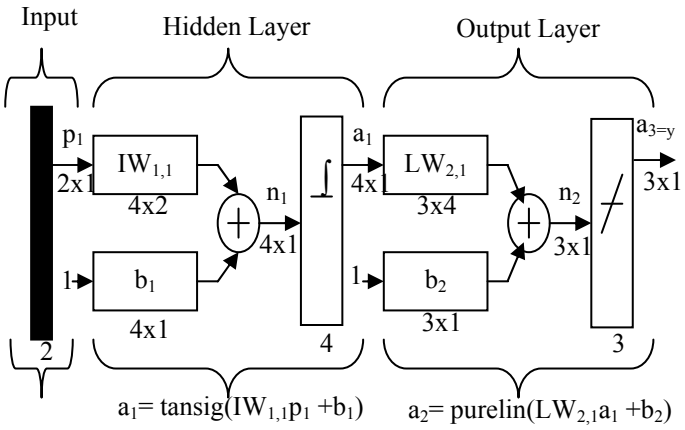


Fig. 3 the multilayer back-propagation network.

The aim of our work is to recognize the type of species from the previous different proteins. We will use the 11 different species based on the albumin sequence as another example for classifications. Using this neural network algorithm based on albumin will make the algorithm that have been done in other work based on hemoglobin more clear [14], [15]. Also we show the modifications to this neural network algorithm if we use other proteins type such as insulin. Using half of the database for training of the neural network and the other half for testing that network. The input of the neural network will be the albumin sequence and the output of that network will be a number related to the type of the specie. The neural network classifier system has the following algorithm:

1) Multi-layers back-propagation network using 3 layers; input, hidden, and output layer.
2) The albumin sequence here is known although that the type of specie is unknown and we want to recognize it so; the albumin sequence will be the input to the neural network and the type of specie will be the output of that network as it will be shown.
3) Because we deal here with different species and each species will have a different length for the albumin sequence and we cannot train the neural network with different inputs lengths. So, the solution was to deal with constant length for the sequences so, we work on the maximum length of those sequences and it was 585 character length, and for the sequences which are less than 585 will be completed by adding letter 'A' to the sequence to reach the length of 585.
4) As we mentioned that albumin sequence length is 585, each letter of that sequence will be converted into binary number- 5 bits for each number as the protein sequence contains 20 amino acids- then 585 x 5 = 2925 bits. So the input layer will be 2925 neurons.
5) The number of output neuron in the output layer will represent the number of different species which will be 11 neurons, one of these 11 outputs will be 1, and the others will be zeros according to the type of the species.
6) Selecting only one hidden layer with about 80 neurons, after many trials.
7) Training half of the database of albumin sequences and the other half will be used in testing that network and write down the results.
8) This classifier system gives 100% of success recognition for the proposed 11 species of albumin database. This network will be repeated for the classification of species using hemoglobin and insulin with little difference in the network construction as shown in table I.

Table I Construction of neural network for classification of different species

Type of protein	Maximum sequence length	No. of neurons in i/p layer	No. of neurons in o/p layer	No. of neurons in one hidden layer	Classifier success recognition
Hemoglobin	330	1650	13	80	100%
Insulin	420	2100	4	10	100%
Albumin	585	2925	11	80	100%

B. Euclidean Distance

Euclidean distance is probably the most commonly chosen type of distance. It simply is the geometric distance in the multidimensional space. In general, the distance d between points $X(x_1,x_2,\ldots..x_n)$ and $Y(y_1,y_2,\ldots\ldots y_n)$ in a Euclidean space is given by:

$$d(X,Y) = |X - Y| = \sqrt{\sum_{i=1}^{n} |x_i - y_i|^2} \qquad (1)$$

The second classifier we introduce in our work is to deal with the Euclidean Distance technique which is based on the DNA sequence of protein and extract about 84 pattern features

of them and store them as a database for each different species. These pattern features will be illustrated as follows:

- Count the number of bases in a nucleotide sequence; means how many times A, C, G, and T are repeated in the DNA sequence. As an example the sequence as 'TAGCTGGCCAAGCGAGCTTG' has A: 4, C: 5, G: 7, and T: 4. This gives 4 features for the sequence.
- Count the number of dimers in a nucleotide sequence; means how many times each couple of bases- AA, AC, AG, AT, CA, CC,..- are repeated in the DNA sequence. As an example the sequence as 'TAGCTGGCCAAGCGAGCTTG' has AA: 1, AC: 0, AG: 3, AT: 0, CA: 1….etc. This gives 16 features for the sequence [16].
- Count the number of standard codons in a nucleotide sequence; means how many times each codon – the codon is a triple of bases as AAA, AAC, AAG…TTT- are repeated in the DNA sequence. As an example the sequence as 'TAGCTGGCCAAGCGAGCTTG' has AAA:0, AAC:0, AAG:1, …GCT:2…etc. This gives 64 features for the sequence [17].

The algorithm of that classifier using albumin goes as follows:

1) The 84 proposed features are collected for the 11 species and stored as a database.
2) For the unknown species we extract its 84 features.
3) Calculate the Euclidean distance between the unknown species and all the 11 species according to the 84 proposed features.
4) The unknown species will be assigned to the specie with the shortest distance.
5) Repeat the previous steps for the classification using the hemoglobin and insulin proteins.
6) This classifier system gives 100% of success recognition for the three classifying systems.

III. Protein Classification Within Human

Most of researchers in the field of protein structure prediction usually use a large database composed of many proteins from many species. We proposed in this work to classify the type of protein within certain species, human, as a first step in this system [11]. The second step will be prediction of the protein structure. In the classification algorithm we proposed a database contains 10 different proteins for human.

These proteins are: Albumin, Globulin, Casein, Hemoglobin, Insulin, Thyroglobulin, Calcitonin, Angiogenin, Myoglobin, and Thymidylate Kinase. The classification algorithm was done by comparison of sequence alignment between the unknown protein and all the 10 proteins in the database. The result of this step is 100%. This means that we were able to classify the unknown protein as one of the known 10 proteins in the database. Now we should be able to apply the proposed prediction algorithm on only one protein. To demonstrate that, we will propose the prediction of the 3D structure for the hemoglobin, insulin and albumin in human.

IV. Protein 3D Structure Prediction

The other aim of this research is to predict the secondary structure and the 3D structure of protein from its DNA sequence with high accuracy. We start by proposing a database contains 26 different structures and sequences of insulin. We have segmented this database into two halves, one half of the database has been used in the training section and the other half in the testing section to find if we have been predicted the structure in proper way or not.

We have been used two prediction techniques in the training section; neural network and Hidden Markov model and we will illustrate them in details in the following sections [14], [15].

A. Neural Network

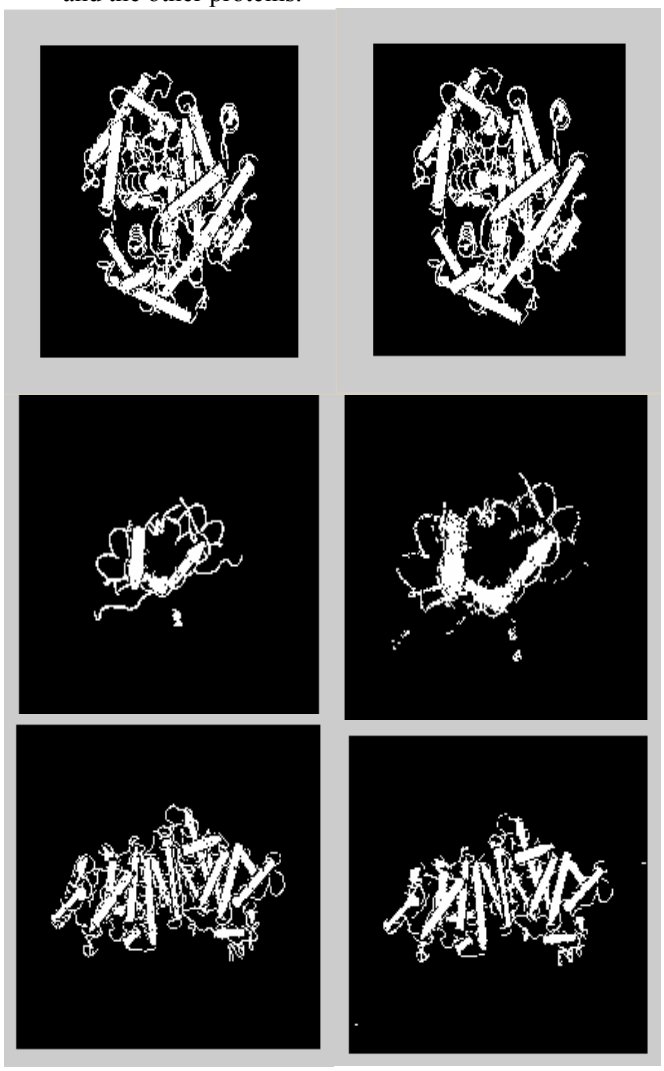
Many researches used neural network techniques in the prediction of protein structures and the best prediction ratio they achieved was almost 77% [6].

We have been introduced this algorithm in previous works but based on hemoglobin [14], [15], here we will introduce this algorithm once more but based on insulin as another example of protein also we will show the modification to this algorithm if we use another protein example such as albumin.

The proposed algorithm, after classifying the given protein as a specific human protein, will go as follows:

1) Three layers backpropagation network has been used; input, hidden, and output layer.
2) The DNA sequence here is known although that the structure is unknown and we want to predict it so; the DNA sequence will be the input to the neural network and the structure will be the output of that network.
3) DNA sequence is a string of ' A, C, G, and T' characters. The length of the DNA sequence of insulin, as an example, is 156 characters, and by representing each character by a binary number; A=00, C=01, G=10, and T=11; and ordering these binary representation in one column to be the input to the neural network. So, the number of neurons in the input layer will be 156x2 =312 neurons.
4) Dealing with the structure as a binary image (dimension 181x200 pixels) and the number of pixels forming that image will be the number of neurons in the output layer which equal to 36200 (181x200).
5) Selecting only one hidden layer with about 200 neurons after many trails.
6) Half of the database (DNA sequences and structures) will be used for training.
7) For testing, enter a DNA sequence that hasn't been used in the training, take the output as the predicted structure and compare it with the original structure of that DNA sequence and calculate the percentage of success of the

predicted structure. Fig. 4 shows an example for the predicted and original structure from the insulin database and the other proteins.



(a) (b)

Fig. 4 (a) The original structure of one of the protein structures in the database, (b) the predicted structure using neural network for hemoglobin, insulin and albumin respectively from top to down.

8) The overall prediction accuracy will be calculated according to the following relations:

$$Q = \frac{\sum_{x=1}^{N} P(x)}{N} \quad (2)$$

Where;
P(x) is the prediction accuracy of each structure.
N is the no. of sequences in the testing part.

$$P(x) = \frac{(XxY) - Er}{(XxY)} x100\% \quad (3)$$

And

$$Er = \|Sp(x,y) - So(x,y)\| \quad (4)$$

Where;
X, Y are the dimensions of the structure and x, y are the index of any pixel.
Er 'Error ratio' is the number of error pixels.
Sp, So are the predicted structure and the original structure respectively. So, the error ratio calculates by count the number of the error pixels in the predicted structure not found in the original one.

9) According to the previous definition we reached to an overall prediction accuracy of insulin equal to 94.1330% which is much better than previous works. Table II shows the difference between the three proteins in neural network algorithm and the overall prediction accuracy.

Table II The difference between the three proteins neural network prediction systems and the overall prediction accuracy.

Type of protein	Database elements	DNA Seq. length	No. of neurons in i/p layer	No. of neurons in o/p layer	No. of neurons in one hidden layer	Overall Prediction Accuracy
Hemoglobin	36	861	1722	36200	200	94.1940%
Insulin	26	156	312	36200	200	94.1330%
Albumin	26	1755	3510	36200	200	94.8132%

B. Hidden Markov Model

Hidden Markov model is one of the powerful prediction tools used in many applications. A Hidden Markov model "HMM" as shown in Fig. 5 is a statistical model in which the system being modeled is assumed to be a Markov process with unknown parameters, and the challenge is to determine the hidden parameters from the observable parameters. The extracted model parameters can then be used to perform further analysis, for example for pattern recognition applications. In a regular Markov model, the state is directly visible to the observer, and therefore the state transition probabilities are the only parameters [18].

In a hidden Markov model, the state is not directly visible, but variables influenced by the state are visible. Each state has a probability distribution over the possible output tokens. Therefore the sequence of tokens generated by a HMM gives

some information about the sequence of states.
Let;

T = the length of the observation sequence
N = the number of states in the model
M = the number of observation symbols
Q = {q_0, q_1, . . . , q_{N-1}} = the states of the Markov process
V = {0, 1, . . . ,M − 1} = set of possible observations
A = the state transition probabilities
B = the observation probability matrix
π = the initial state distribution
O = (O_0,O_1, . . . ,O_{T-1}) = observation sequence.

The observations are always denoted by {0, 1, . . . ,M − 1}, since this simplifies the notation with no loss in generality. Then Oi Є V for i = 0, 1, . . . , T − 1.

A generic hidden Markov model is illustrated in Fig. 5, where the Xi are the hidden states and all other notation is as given above. The Markov process—which is hidden behind the dashed line—is determined by the initial state X_0 and the A matrix. We are only able to observe the Oi, which are related to the actual states of the Markov process by the matrices B and A [19].

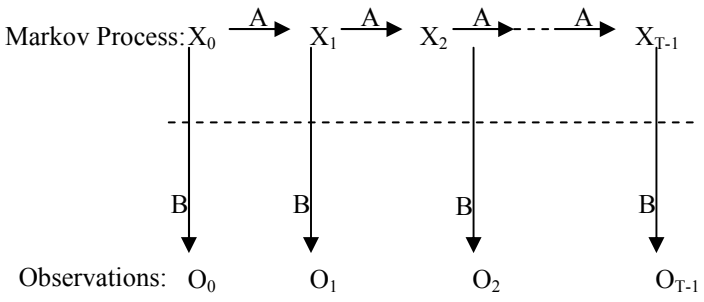


Fig.5 the architecture of HMM.

The matrix A is the state transition probabilities,
A = {aij} is N × N with
aij = P(state qj at t + 1 | state qi at t).

The matrix B is the observation probability matrix,
B = {bj(k)} is an N ×M with
bj(k) = P(observation k at t | state qj at t).

As with A, the matrix B is row stochastic and the probabilities bj(k) are independent of t. The unusual notation bj(k) is standard in the HMM world.

An HMM is defined by A, B and π (and, implicitly, by the dimensions N and M). The HMM is denoted by λ = (A, B, π).

Consider a state generic sequence of length four
X = (x_0, x_1, x_2, x_3)
with corresponding observations
O = (O_0, O_1, O_2, O_3).

Then π_{x0} is the probability of starting in state x_0. Also, $b_{x0}(O_0)$ is the probability of initially observing O_0 and $a_{x0,x1}$ is the probability of transiting from state x_0 to state x_1. Continuing, we see that the probability of the state sequence X is given by:

$$P(X) = \pi_{xo} b_{xo}(O_0) a_{xo,x1} b_{x1}(O_1) a_{x1,x2} b_{x2}(O_2) a_{x2,x3} b_{x3}(O_3) \quad (5)$$

Where;
π is the initial state distribution.
π_{x0} is the probability of starting in state x_0.
A is the state transition probabilities,
A = {aij} is N × N with
aij = P(state qj at t + 1 | state qi at t).
B is the observation probability matrix,
B = {bj(k)} is an N ×M with
bj(k) = P(observation k at t | state qj at t).
N = the number of states in the model
M = the number of observation symbols
O= (O_0, O_1, . . . , O_{T-1}) = observation sequence.

We will start by illustrating the algorithm by using the whole insulin base so; in our work we have 26 DNA sequences and structures for the insulin. We used 13 sequences and structures for the training part and used the remaining 13 sequences and structures in the testing part. Using Hidden Markov Model as a prediction tool in the insulin requires several variables and initializations.

First of all we need to define the main concepts in the proposed HMM as follows:

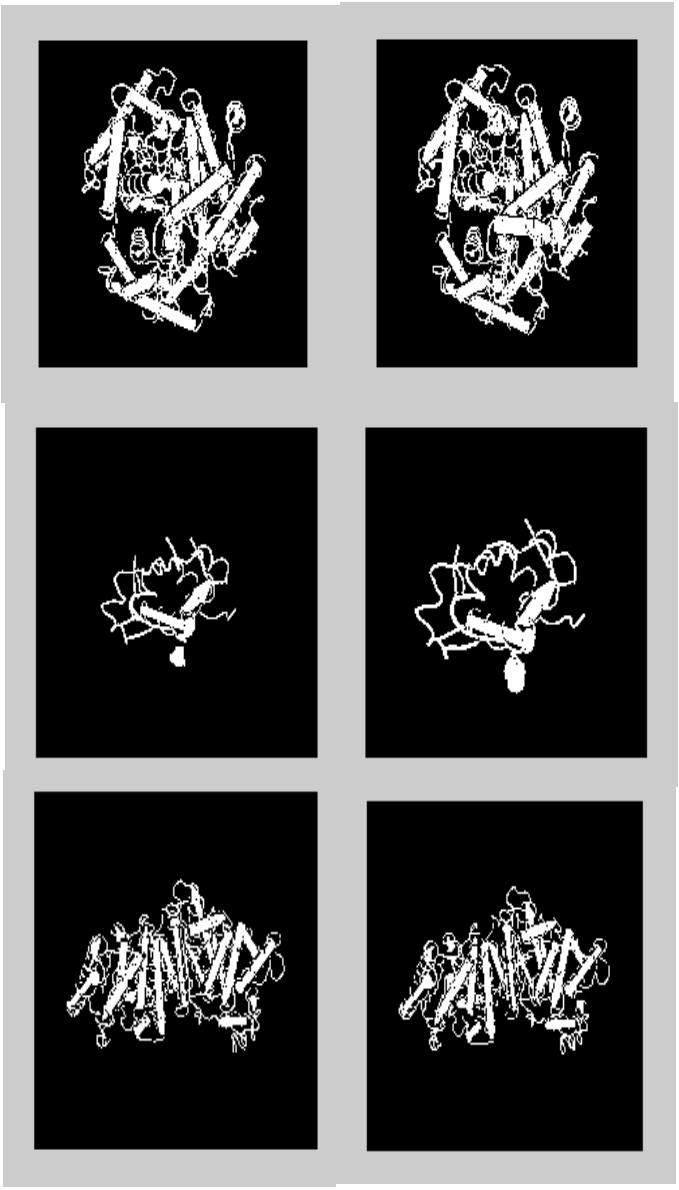
1) In Hidden Markov Model there is a known part called the observations and an unknown part called the states. We want to predict the structure of the protein from the DNA sequence so, the known part here is the DNA sequence, observations, and the unknown part is the protein structure, states.
2) In insulin example we have 13 structures and sequence for the training, so we have 13 states, protein structures, and also 13 observations, DNA sequences.
3) Set the matrix A as state transition matrix in dimension 13x13, which shows the transition between the states, DNA sequence, that ideally would not change or transform to another state or DNA sequence. The ideal initialization for that matrix is an 13x13 matrix with its main diagonal elements equal one, and all other element are zeros as an unity matrix.

$$A = \begin{pmatrix} 1&0&0&0&0&0&0&0&0&0&0&0&0 \\ 0&1&0&0&0&0&0&0&0&0&0&0&0 \\ 0&0&1&0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&1&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&1&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&1&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&1&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&1&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&1&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&1&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0&1&0&0 \\ 0&0&0&0&0&0&0&0&0&0&0&1&0 \\ 0&0&0&0&0&0&0&0&0&0&0&0&1 \end{pmatrix}$$

4) Set the matrix B as the observation matrix in dimension 13x13, which shows the relation between the states as rows, DNA sequences, and the observations as columns, protein structures. The ideal initialization for that matrix is similar to the initialization of matrix A

$$B=\begin{pmatrix}1&0&0&0&0&0&0&0&0&0&0&0&0\\0&1&0&0&0&0&0&0&0&0&0&0&0\\0&0&1&0&0&0&0&0&0&0&0&0&0\\0&0&0&1&0&0&0&0&0&0&0&0&0\\0&0&0&0&1&0&0&0&0&0&0&0&0\\0&0&0&0&0&1&0&0&0&0&0&0&0\\0&0&0&0&0&0&1&0&0&0&0&0&0\\0&0&0&0&0&0&0&1&0&0&0&0&0\\0&0&0&0&0&0&0&0&1&0&0&0&0\\0&0&0&0&0&0&0&0&0&1&0&0&0\\0&0&0&0&0&0&0&0&0&0&1&0&0\\0&0&0&0&0&0&0&0&0&0&0&1&0\\0&0&0&0&0&0&0&0&0&0&0&0&1\end{pmatrix}$$

5) Using the initial values of the matrices A and B training them by using the Baum-Welch algorithm to set the true values of those matrices.
6) Finally we need the observation sequence O, which has number of observation sequences, take four numbers from the 13 DNA sequences, as an example O= (1,1,2,3) that means the first DNA sequence followed by itself again, then followed by the 2nd sequence, then the third one. And if we have A, B, O, and the initial π so; we could compute the sequence of the unknown states, the protein structure, according to the probability in (5). P(x) will get sequence of states, protein structures, but we predict only one protein structure so, we get the average of those structures.
7) But the problem here is to use different 13 DNA sequences that have not been used in the training so, how we can set the observation sequence O by unknown sequence. The solution here was, when we have an unknown sequence we compare it with the 13 sequences that have been used in the training part and get its nearest sequence and use it as the observation sequence O, then we can compute the state probability P(x) and get the unknown protein structure.
8) The obtained overall prediction accuracy for insulin using HMM was 93.6823% of success prediction according to equations (1), (2) & (3), and Fig. 6 shows the original structure and the predicted one of one insulin base as an example and the other proteins.
9) In the previous steps we predicted the 3D structure of insulin represented in the binary form. We also predicted the 3D structure of insulin in the gray level form and in the color form. The percentage of success prediction in the gray level form gives about 95.3532%. Also percentage of success prediction of the colored 3D form gives 86.2206%. Table III shows the overall prediction accuracy of the three proteins for the binary, gray level and colored structures. Fig. 7 & 8 show an example of an original and predicted structure for the gray level and colored form of the three proteins respectively.

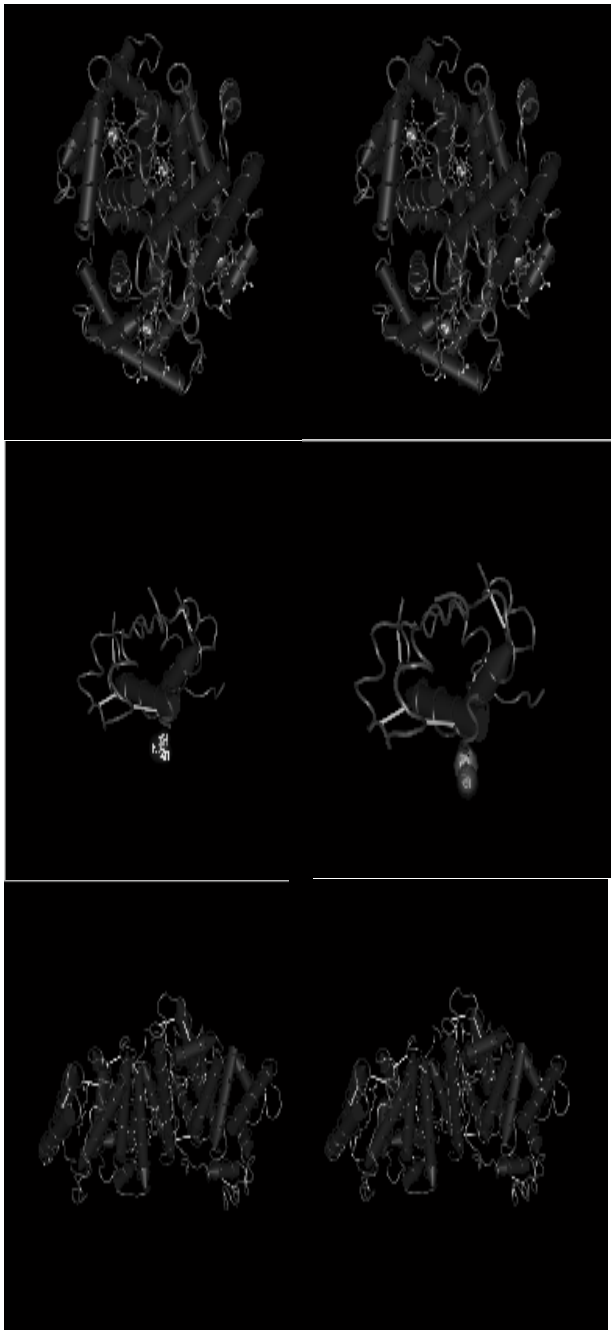


(a) (b)

Fig. 6 (a) The original structures of a protein structures in the database, (b) the predicted structure using Hidden Markov model for hemoglobin, insulin & albumin from top to bottom respectively.

Table III The overall prediction accuracy of the three proteins for different structures.

Type of protein	Binary Structure Prediction Accuracy	Gray-level Structure Prediction Accuracy	Colored Structure Prediction Accuracy
Hemoglobin	91.2190%	86.8198%	59.2865%
Insulin	93.6823%	95.3532%	86.2206%
Albumin	93.7592%	92.6715%	77.5429%

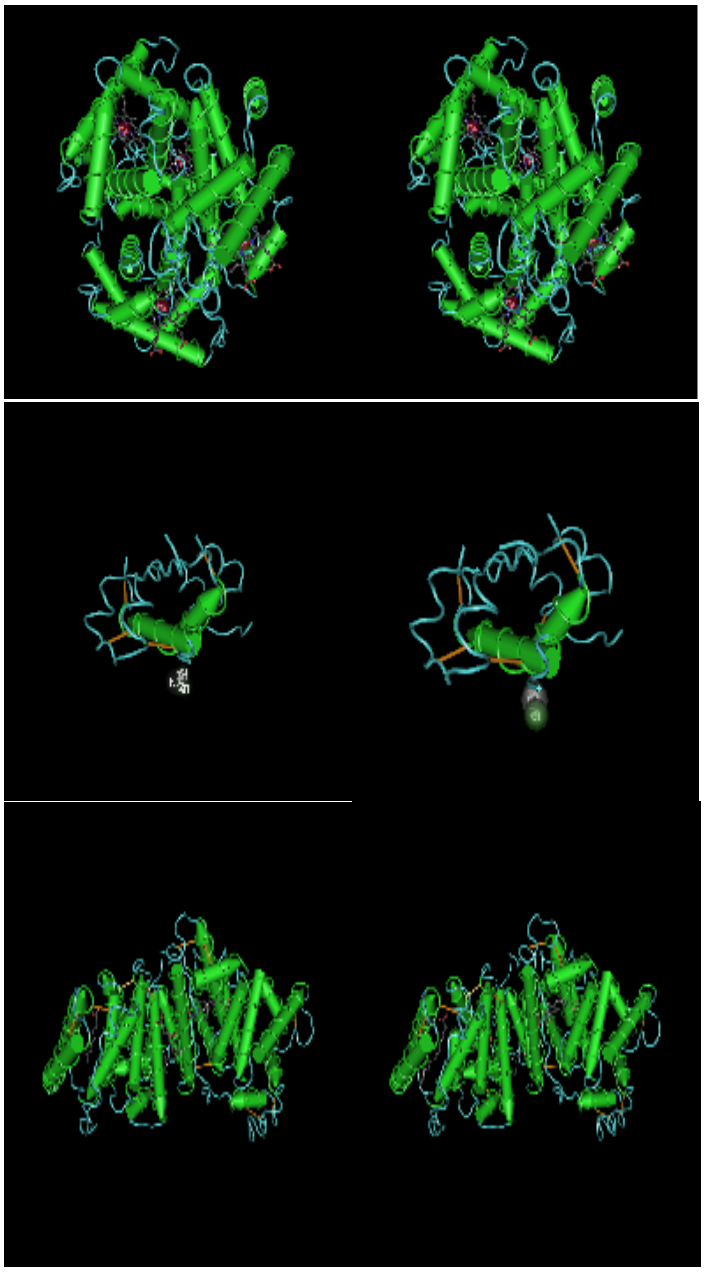


(a) (b)

Fig. 7 (a) The original structures of a protein structures in the database, (b) the predicted structure using Hidden Markov model for hemoglobin, insulin & albumin respectively for gray images.

V. Protein Secondary Structure Prediction

As previously stated, the order in which amino acids occur in proteins is determined by the genetic code. The surrounding chemical environment, which is primarily composed of water (and other solvents) at different concentrations and temperatures, and the amino acid side chains, determine the way in which these are arranged in space relative to each other [20]. In other words, amino acid chains do not fold at random. The basic structures that form are known as sheets (beta-sheets), helices (alpha-helices) and turns or coils as shown in Fig. 9. These are also known as basic secondary structures.



(a) (b)

Fig. 8 (a) The original structures of a protein structures in the database, (b) the predicted structure using Hidden Markov model for hemoglobin, insulin & albumin respectively for colored images.

An alpha helix resembles a ribbon wrapped around a tube, similar to a circular staircase. This structure is very stable but flexible therefore it is often seen in parts of a protein that need to bend or move.

In a beta-sheet, two or more ribbons of amino acids are involved. These lines up to form a pleated like structure similar to folds in fabric. This structure tends to be rigid and less flexible than alpha helices.

Turns are usually related to proline and glycine, which are common and small and are often responsible for sharp bends

and twists in alpha helices and hair-pins in beta sheets. By knowing which spatial geometry neighboring amino acids adopt when they bind together it is possible to determine which secondary structure a protein may have.

Alpha helices and beta sheets are subdivided into classes that extend the basic definitions given here. There are also wide helices, which appear along short segments of proteins. Beta sheets are subdivided into anti-parallel and parallel sheets. Beta stands are also common where a sheet is not formed and turns of various types. Barrels, loops and coils are also found and denote specific regions, which are found in proteins as units [21].

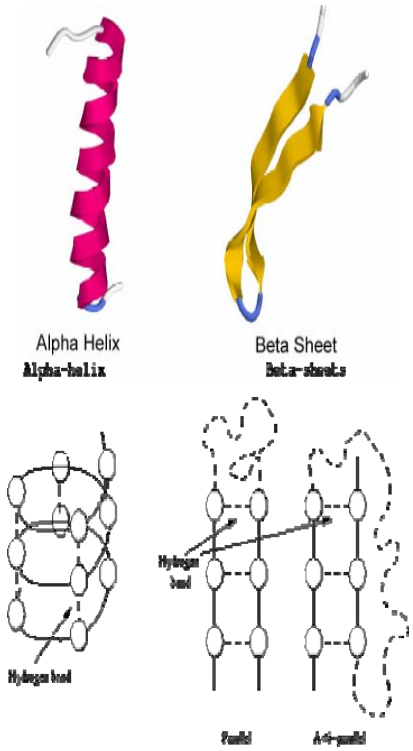


Fig. 9 the classes of secondary structure.

In the following part we will predict the secondary structure of the three types of proteins-hemoglobin, insulin & albumin- but we take hemoglobin as an example. This is done by storing half of the sequences in the database with their secondary structures and using the other half in testing the result of prediction which will be illustrate as follows:

1) Transform the DNA sequence of hemoglobin into the amino acid sequence.
2) Storing 18 amino acid sequences of hemoglobin and their related secondary structures as a database.
3) Using the other half of database, the remaining 18 sequences, in testing by entering one of these sequences and tries to predict its secondary structure.
4) Prediction of the secondary structure for that entering sequence is done by using multiple alignments which gives the secondary structure of the nearest database sequences to the entering sequence.
5) Compare the predicted sequence with the original one character by character and write down the result. Fig. 10 shows the protein sequence, its predicted structure and the original one for hemoglobin where 'h' represent helix & 'c' represent coil.
6) This algorithm gives 99.8% of success prediction which is considered a very high success ratio if it is compared with other researches which gave around 81% of success prediction [21]. This high success ratio because of dealing only with one class of protein-hemoglobin- instead of using many types of proteins [14], [15]. Table IV shows the results of the three different proteins.

Table IV the secondary structure prediction accuracy of the three proteins.

Type of protein	Secondary Structure Prediction Accuracy
Hemoglobin	99.8%
Insulin	93.0%
Albumin	97.3392%

```
Protein Seqence1= mlspadktnvkaawgkvgahageygaealermflsfpttktaf     43
Predicted Structure= nhhhhhhhhhhhhhh  hhhhhhhhhhhhhhh
                                  c          c   cc
Original Structure= nhhhhhhhhhhhhhh  hhhhhhhhhhhhhhh
                                  c          c   cc
Prediction rate= 100%

Protein Seqence2= vlspadktnvkaawgkvgahageygaealermflsfpttktyf     43
Predicted Structure= hhhhhhhhhhhhhhh  hhhhhhhhhhhhhhhhh
                                  c          c   cc
Original Structure= hhhhhhhhhhhhhhh  hhhhhhhhhhhhhhhhh
                                  c          c   cc
Prediction rate= 99.8%
```

Fig.10 the prediction of hemoglobin secondary structure.

VI. Conclusion

The aim of this paper is to present a multi usable system based on bioinformatics which can be used in many applications. The first use of this system is to classify different species based on three different proteins sequences by a fast and easy way. Two different classifiers systems have been introduced to perform this application based on neural network and Euclidean distance techniques. The two techniques gave the same result which is 100% of success classification. Also this system classifies the type of protein within certain specie which is human through 10 different based on sequence alignments. Finally this system can be used to predict the 3D structure the secondary structure of human proteins-such as hemoglobin, insulin and albumin- from its DNA sequence.

Two different techniques have been used to perform the prediction of the 3D structure of the protein, neural network and hidden Markov model. It is found that the neural network technique gave slightly better success prediction than Markov model. The highest obtained success prediction rate was about 94% compared to the 77% obtained in similar works. In addition, a high prediction ratio (99.8%) has been achieved in the prediction of the secondary structure of those proteins compared to 81% from previous works. This work may be applied to other different protein types to make a powerful system for prediction of protein structure.

References

[1] Christos A. Ouzounis, and Alfonso Valencia, "Early bioinformatics: the birth of a discipline—a personal view," *Bioinformatics Journal,* Vol.19, No.17, pp. 2176-2190, 2003.

[2] N.M. Luscombe, D. Greenbaum, and M. Gerstein, *What is Bioinformatics? An Introduction and Overview,* Yearbook of Medical Informatics, 2001, pp. 83-100.

[3] P. Bourne and H. Weissig, *Structural Bioinformatics*, John Wiley & Sons, 2003.

[4] J. Cohen, "Bioinformatics—An introduction for computer scientists," *ACM Computing Surveys*, Vol.36, No.2, pp. 122-158, 2004.

[5] P. G. Wodehouse, "Bioinformatics and pattern recognition come together," *Journal of Pattern Recognition Research*, Vol.1, pp. 37-41, 2006.

[6] www.cnbi.nlm.nih.gov.

[7] P. Cristea, V. Mladenov, R. Tuduce, G. Tsenov, and S. Petrakieva, "Neural networks for prediction of nucleotide sequences by using genomic signals," *WSEAS Transactions on Systems*, Issue 7, Vol. 7,pp.637-644, July 2008.

[8] N. M. Luscombe, D. Greenbaum, and M. Gerstein, "What is bioinformatics? a proposed definition and overview of the field*," Method Inform Med,* Vol.40, pp. 346-358, 2001.

[9] J. Cheng, M. J. Sweredoski, and P. Baldi, "Accurate prediction of protein disordered regions by mining protein structure data, " *Technical report, Springer Science + Business Media*, School of Information and Computer Science, Institute for Genomics and Bioinformatics, University of California Irvine, 2005.

[10] D. Morikis, B. Mallik, and L. Zhang, "Biophysical and bioengineering methods for the study of the complement system at atomic resolution," *Proceedings of the 2006 WSEAS Int. Conf. on Cellular & Molecular Biology, Biophysics & Bioengineering*, Athens, Greece, pp.80-85 , 2006.

[11] R.I. Mubark, H.A. Keshk and M.I. Eladawy, " Different species classifier based on hemoglobin sequences, " *The 4th kuala Lumpur International Conference on Biomedical Engineering, Springer Book Series IFMBE Proceedings,* Vol. 21, pp. 279-281, 2008.

[12] K. Lin, C. Y. Lin, C. Huang, H. Chang, C.Y. Yang,C. Lin, C. Y. Tang, and D.F. Hsu, "Improving prediction accuracy for protein structure classification by neural network using feature combination," *Proceedings of the 5th WSEAS Int. Conf. on Applied Informatics and Communications*, Malta, pp.313-318 , 2005.

[13] Matlab Neural network toolbox.

[14] R.I. Mubark, H.A. Keshk and M.I. Eladawy, "Prediction of hemoglobin structure from DNA sequence through neural network and hidden markov model," *The 7th WSEAS Int.Conf. on Computational Intelligence, Man-machine Systems and Cybernetics (CIMMACS '08)*, Cairo, Egypt, pp.65-76, 2008.

[15] R.I. Mubark, H.A. Keshk and M.I. Eladawy, "Different species classifier and hemoglobin structure predictor based on DNA sequences," *International Journal of Biology and Biomedical Engineering*, Issue 2, Vol. 2, pp.49-58, 2008.

[16] Y. Yamada, and K. Satou, "Prediction of genomic methylation status on CpG islands using DNA sequence features," *WSEAS Transactions on Biology and Biomedicine*, Issue 7, Vol. 5,pp.153-162, July 2008.

[17] T. Al_ibaisi, A. Abu-dalhoum, M. Al-rawi, M. Alfonseca , and A. Ortega, "Network intrusion detection using genetic algorithm to find best DNA signature," *WSEAS Transactions on Systems*, Issue 7, Vol. 7,pp.589-599, July 2008.

[18] P. G. Bagos, T. D. Liakopoulos, and S. J. Hamodrakas, " Finding beta-barrel outer membrane proteins with a markov chain model," *WSEAS Transactions on Biology and Biomedicine*, Issue 2, Vol. 1,pp.186-189, April 2004.

[19] Y. Ephraim and N. Merhav, "Hidden markov processes," *IEEE Trans. Inform. Theory,* Vol. 48, pp. 1518-1569, 2002.

[20] Y. Qi, F. Lin, and K. K. Wong, "High performance computing in protein secondary structure prediction," *WSEAS Transactions on Circuits and Systems*, Issue 3, Vol. 2, pp.619-624, July 2003.

[21] H. Rangwala, K. DeRonne, and G. Karypis, " Protein structure prediction using string kernels, " *Technical Report*, Department of Computer Science & Engineering, University of Minnesota, 2005.

Roaa I. Mubark Was born in Cairo, Egypt on May 1979. She received her B.S. and M.S. degrees in Communications and Electronics Engineering from University of Helwan (Egypt) in 2001 and 2004. She is currently working toward the Ph.D. degree in the area of electronics engineering at Helwan University. Her current research interests are in the area of bioinformatics applications within proteins structure predictions and classifications.

Hesham A. Keshk He received BSc from Department of communication and electronic, Faculty of Engineering, Cairo University (Egypt) in May 1982, M.Sc from Helwan University (Egypt) in 1989, and Ph.D from Kyoto University (Japan) in 1996. Since 1996 he has taught and conducted research in the area of computer engineering at Helwan University.

Mohamed I. Eladawy He graduated from the Department of Electrical Engineering, Faculty of Engineering of Assiut University in May 1974; M.Sc. from Cairo University in May 1979; Ph.D. from Connecticut State University, School of Engineering, in May 1984. He worked as an Instructor at the Faculty of Engineering, Helwan University since 1974. Currently he is a Professor at the Department of Communication and Electronics Engineering and the Vice Dean for Student Affairs in the same faculty. He was working for the general organization for technical and vocational training for 6 years from 1989 to 1995 in Saudi Arabia. Main interest is in signal processing and its medical applications.