Different Species Classifier and Hemoglobin Structure Predictor based on DNA Sequences

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

***Abstract*—** Large-scale analysis studies of genetic sequence data are in progress around the world; one of these studies is to recognize the type of the species that the sequence belongs to. This is very important especially when the source of the sequence is unknown. The complete genome sequence of the hemoglobin provides an excellent basis for studying the clustering of different species. In this paper 13 different species classifier based on hemoglobin sequence will be introduced. Two different classifiers systems also have been used; one of them based on neural network and the other based on extracting 84 pattern feature from the DNA sequence of hemoglobin with the Euclidean distance technique. Also one of the greatest challenges today in bioinformatics is to predict the structure of the protein from the DNA sequence. Protein 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. 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. In this work we will be able to predict the 3D structure of hemoglobin using two techniques; the neural network and hidden Markov model. Also, the prediction of the secondary structure is applied using multiple alignments.

***Keywords*—**Bioinformatics, Classification algorithm, 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, and hemoglobin. 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. Fig. 1 shows DNA and chromosomes [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].

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).

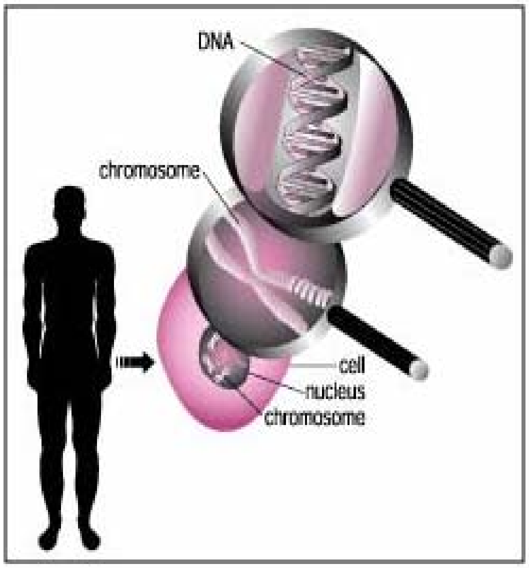
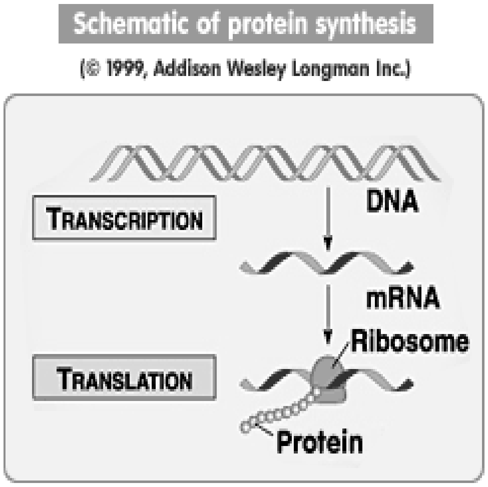


Fig. 1 the DNA and chromosome [4].

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


Schematic of protein synthesis

(© 1999, Addison Wesley Longman Inc.)

Fig. 2 the transformation of DNA into protein [6].

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].

And finally 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 [6].

II. Different Species Classifier

The complete genome sequence of the hemoglobin provides an excellent basis for studying the clustering of different species. Many species has been used in this paper as human, horse, wolf, donkey, chicken, clam lucina, Glycera Dibranchiata ,tuna fish, trout fish, hagfish, rice plant, and two different bacteria's; mycobacterium Tuberculosis Trhbn & gutless beard worm Oligobrachia Mashikoi.. Here in our work we try to recognize the type of species from the previous 13 different species. We down loaded these hemoglobin sequences from National Center for Biotechnology Information “NCBI” website and it was 30 sequences for hemoglobin [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.

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 [11]. 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. 3 [12].

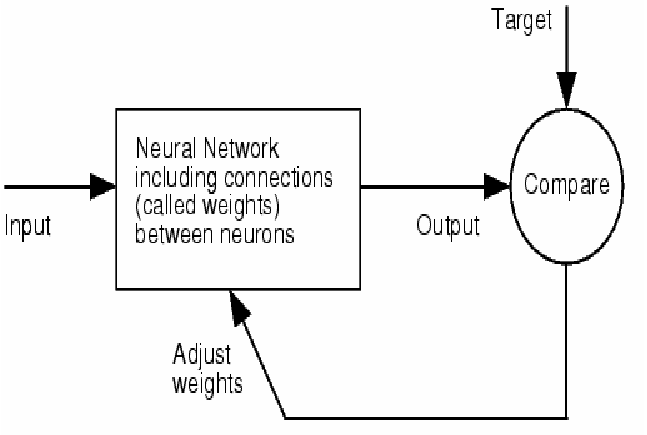


Fig. 3 the Neural Network [12].

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. 4.

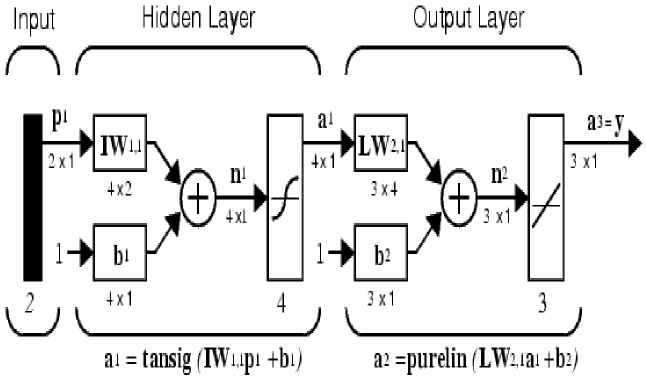


Fig. 4 the multilayer back-propagation network [12].

The aim of our work is to recognize the type of species from the previous 13 different species as mentioned previously. We use 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 hemoglobin 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 hemoglobin sequence here is known although that the type of specie is unknown and we want to recognize it so; the hemoglobin 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 hemoglobin 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 330 character length, and for the sequences which are less than 330 will be completed by adding letter 'A' to the sequence to reach the length of 330.
4) As we mentioned that hemoglobin sequence length is 330, 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 330 x 5 = 1650 bits. So the input layer will be 1650 neurons.
5) The number of output neuron in the output layer will represent the number of different species which will be 13 neurons, one of these 13 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 hemoglobin 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 13 species.

B. Euclidean Distance

Euclidean distance is probably the most commonly chosen type of distance. It simply is the geometric distance in the multidimensional space [13]. 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:

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

The second classifier we introduce in our work is to deal with the Euclidean Distance technique which is based on the DNA sequence of hemoglobin 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 [14].

- 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 [15].

The algorithm of that classifier goes as follows:

1) The 84 proposed features are collected for the 13 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 13 species according to the 84 proposed features.
4) The unknown species will be assigned to the specie with the shortest distance.
5) This classifier system gives 100% of success recognition for the proposed 13 species.

III. Protein Classification

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 [10]. 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. In this paper we applied our prediction algorithm on hemoglobin as an example.

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. The proposed data base contains 36 different structures and sequences of hemoglobin. 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 [12].

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].

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 hemoglobin, as an example, is 861 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 861x2 =1722 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. 5 shows an example for the predicted and original structure from the hemoglobin database.
8) The overall prediction accuracy will be calculated according to the following relations:

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

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

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

And

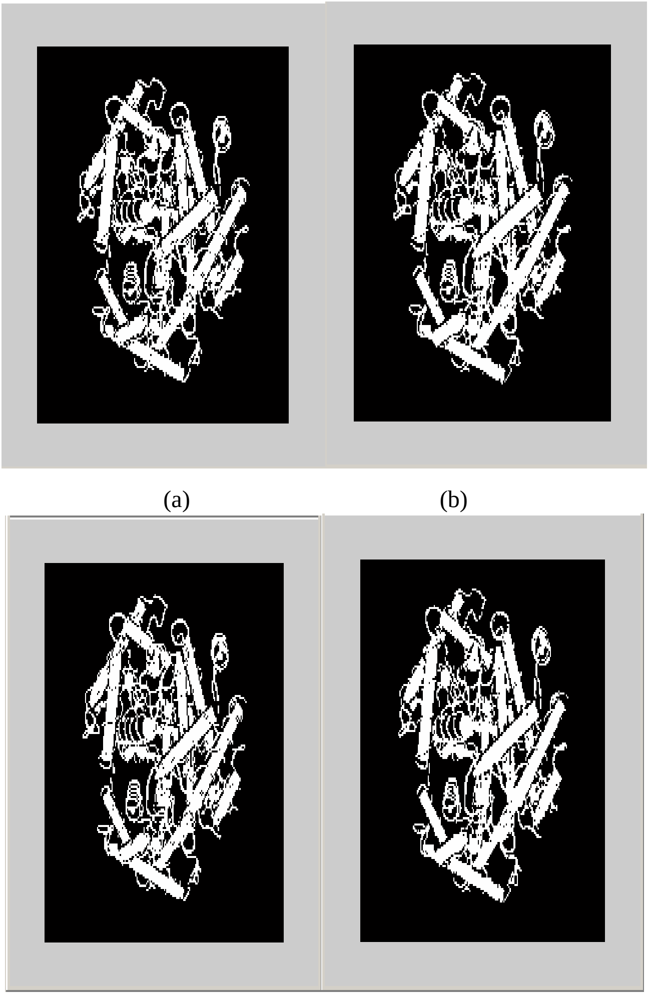
$$Er = \left\| Sp(x, y) - So(x, y) \right\| \qquad (4)$$

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.



(a) (b)

(c) (d)

Fig. 5 (a) & (c) the original structure of one of Hemoglobin structures in the database, (b) & (d) the predicted structure using neural network.

9) According to the previous definition we reached to an overall prediction accuracy equal to 94.1940% which is much better than previous works.

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. 6 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 [16].

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. 6, 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 [17].

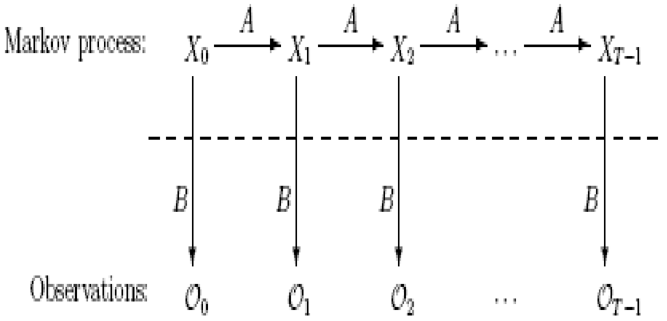


Fig.6 the architecture of HMM.

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

$O = (O_0, O_1, \ldots, O_{T-1})$ = observation sequence.

We will start by illustrating the algorithm by using the whole hemoglobin base so; in our work we have 36 DNA sequences and structures for the Hemoglobin. We used 18 sequences and structures for the training part and used the remaining 18 sequences and structures in the testing part. Using Hidden Markov Model as a prediction tool in the Hemoglobin 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 Hemoglobin example we have 18 structures and sequence for the training, so we have 18 states, protein structures, and also 18 observations, DNA sequences.
3) Set the matrix A as state transition matrix in dimension 18x18, 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 18x18 matrix with its main diagonal elements equal one, and all other element are zeros as an unity matrix.

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

4) Set the matrix B as the observation matrix in dimension 18x18, 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

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

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 18 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 18 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 18 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 using HMM was 91.2190% of success prediction according to (2), (3) & (4), and Fig. 7 shows the original structure and the predicted one of one hemoglobin base as an example.

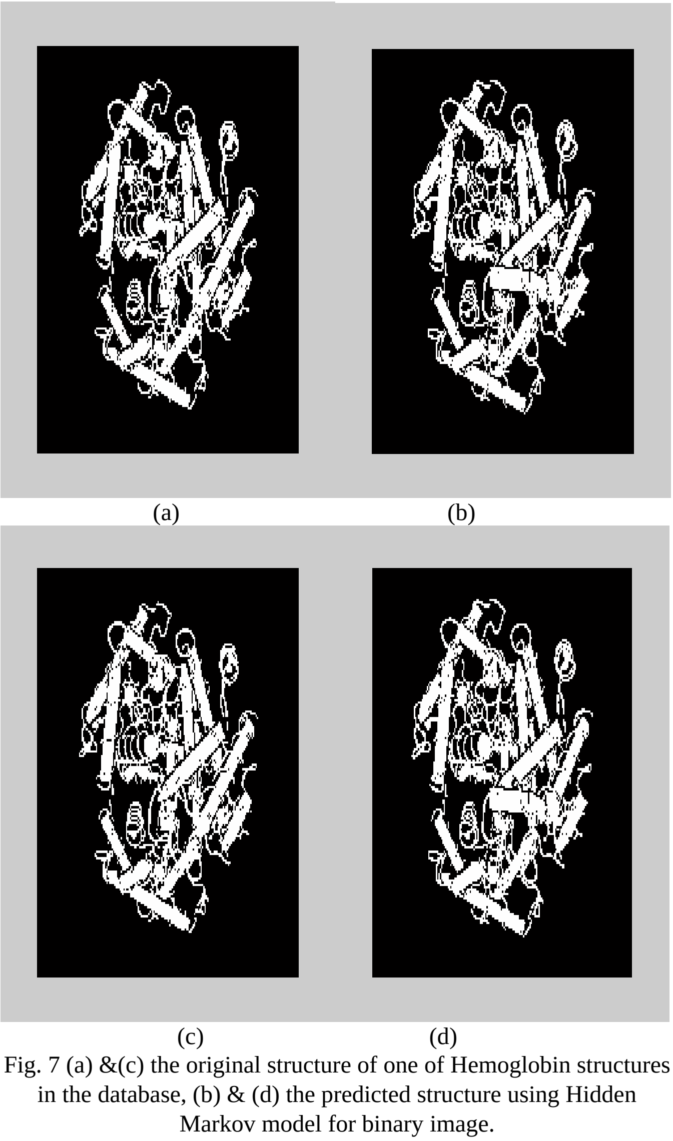


Fig. 7 (a) &(c) the original structure of one of Hemoglobin structures in the database, (b) & (d) the predicted structure using Hidden Markov model for binary image.

9) In the previous steps we predicted the 3D structure of hemoglobin represented in the binary form. We also predicted the 3D structure of hemoglobin in the gray level form and in the color form. The percentage of success prediction in the gray level form gives about 86.8198%. Also percentage of success prediction of the colored 3D form gives 59.2865%. Fig. 8 & 9 show an example of an original and predicted structure for the gray level and colored form respectively.

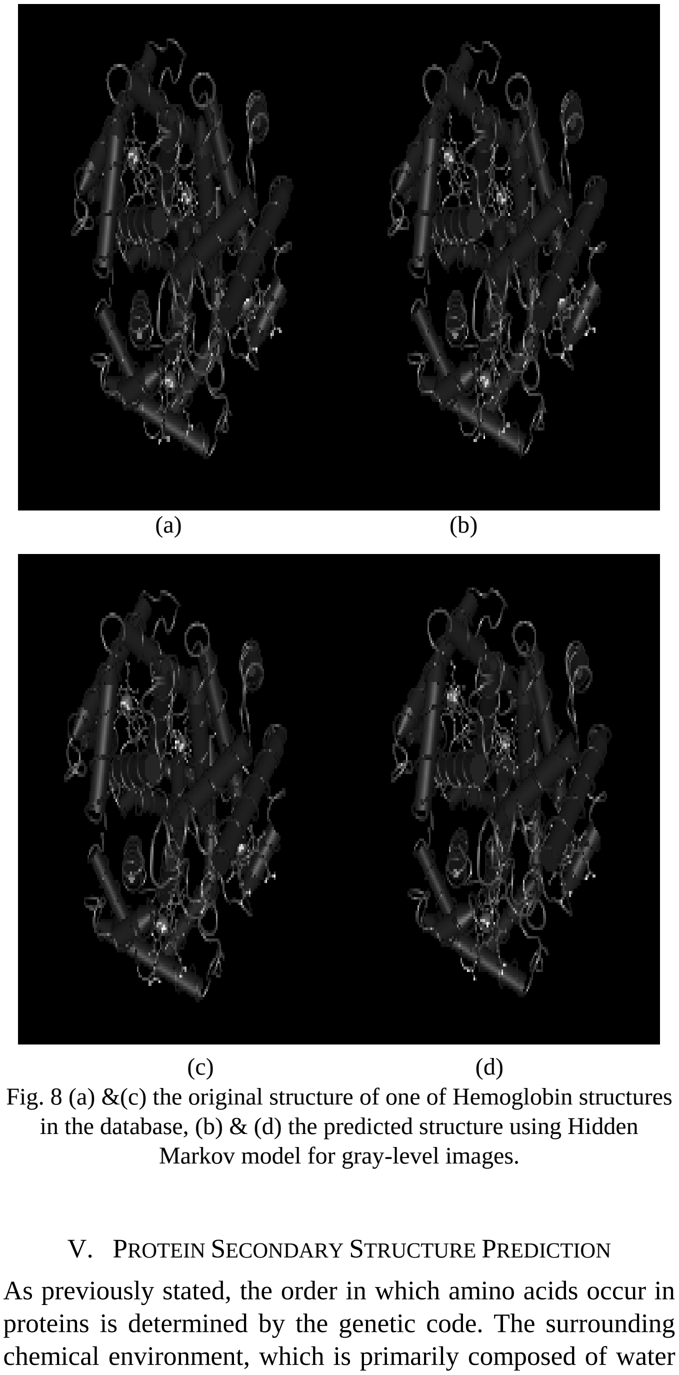
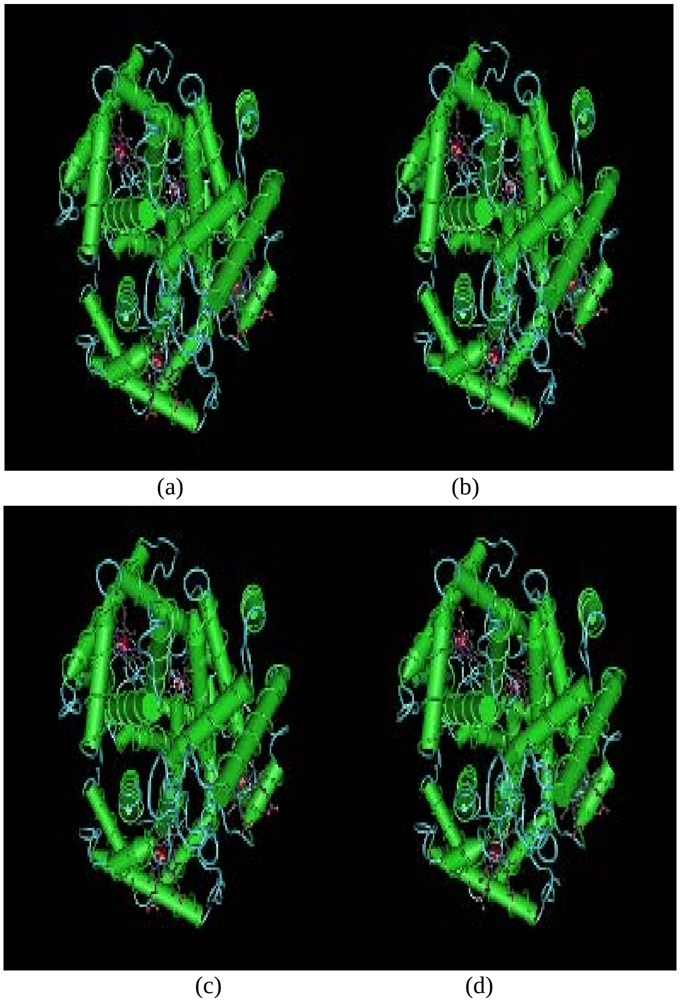


Fig. 8 (a) &(c) the original structure of one of Hemoglobin structures in the database, (b) & (d) the predicted structure using Hidden Markov model for gray-level images.

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 [18]. 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. 10. These are also known as basic secondary structures.


(a) (b)

(c) (d)

Fig. 9 (a) &(c) the original structure of one of Hemoglobin structures in the database, (b) &(d) the predicted structure using Hidden Markov model 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 [19].

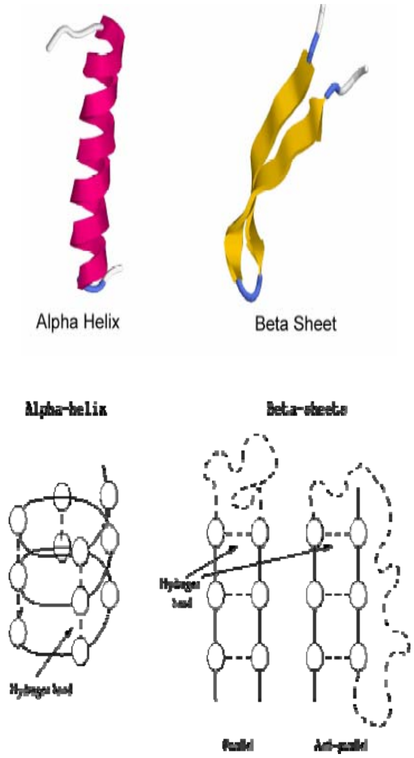
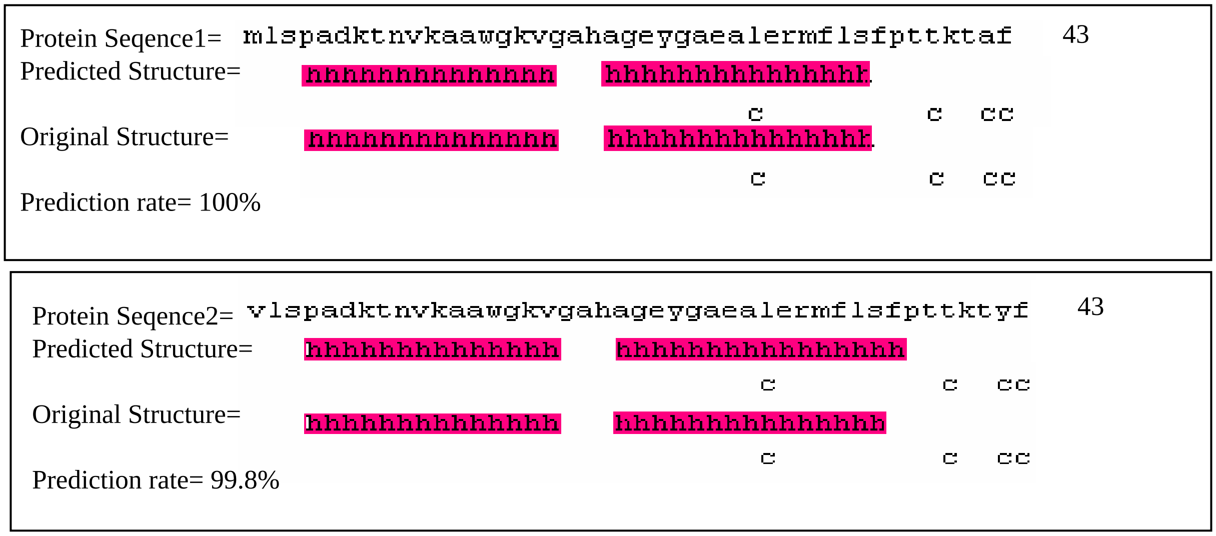


Fig. 10 the classes of secondary structure.

In the following part we will predict the secondary structure of hemoglobin. 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. 11 shows the protein sequence, its predicted structure and the original one 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 [19]. This high success ratio because of dealing only with one class of protein-hemoglobin- instead of using many types of proteins [20].



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

Protein Seqence2= vlspadktnvkaawgkvgahageygaealermflsfpttktyf 43
Predicted Structure= hhhhhhhhhhhhhh hhhhhhhhhhhhhhhh c c cc
Original Structure= hhhhhhhhhhhhhh hhhhhhhhhhhhhhh c c cc
Prediction rate= 99.8%

Fig.11 the prediction of the secondary structure.

VI. Conclusion

The aim of this paper was to present a system that can classify different species based on hemoglobin sequences and also can predict the structure of a specific human protein, hemoglobin, from its DNA sequence by a fast and easy way. Two different classifiers systems have been introduced to perform the classifier system based on neural network and Euclidean distance techniques. The two techniques gave the same result 100% of success classification. This can be applied to other protein types to make a powerful system for classifying different species.

For the predictor system 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 compared to 81% from previous works. This work may be applied to 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] 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.

[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] 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.

[13] Matlab Neural network toolbox.

[14] 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.

[15] 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.

[16] 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.

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

[18] 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.

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

[20] R.I. Mubark, H.A. Keshk and M.I. Eladawy, “Prediction of hemoglobin structure from DNA sequence through neural network and hidden markov model (Accepted for publication),” *The 7th WSEAS Int.Conf. on Computational Intelligence, Man-machine Systems and Cybernetics (CIMMACS '08)*, to be published.

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.