Prediction of Hemoglobin Structure from DNA Sequence through Neural Network and Hidden Markov Model

R. I. MUBARK, H. A. KESHK & M. I. ELADAWY
Electronics, Communication & Computer Engineering Department
Helwan University
Sherif st. No.1, Helwan
EGYPT
roaim79@yahoo.com  http://www.yahoo.com

Abstract: - 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.

Key-Words: - Bioinformatics, Genetic sequences, Hemoglobin, Protein prediction, Neural network, Hidden markov model, Classification algorithm.

1 Introduction
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 application on protein structure, may need to define basics in human genome such as DNA, chromosome, RNA, protein, Hemoglobin & protein structure. DNA code is a sequence of chemicals that form information that control how humans are made and how they work [4]. This encoding information in an ordered sequence of 4 different symbols called "bases", typically denoted A, C, G, and T. 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,


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. 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]. Hemoglobin is a protein-based component of red blood cells which is primarily responsible for transferring oxygen from the lungs to the remainder 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].

The only way to know the function of a protein is to find its structure, and this is always done by experimental methods such as X-ray diffraction and NMR (Nuclear Magnetic Resonance) spectroscopy [7]. Those experimental methods can not be applied all the time besides they also are expensive so, the other way to determine protein structure is through prediction. In this work we try to predict the
structures of hemoglobin through group of structures which have been determined from the experimental methods. We download these structures and its related DNA sequences from the National Center for Biotechnology Information 'NCBI' and it was 36 structures and sequences for human hemoglobin [6]. We used these as a data base allover the proposed work in this paper.

2 Problem Formulation
The aim of this research is to predict the secondary structure and the 3D structure of protein from its DNA sequence with high accuracy.

2.1 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 [8]. 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 was 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.

2.2 Prediction Technique
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. Two prediction techniques have been used in the training section; neural network and Hidden Markov model and we will illustrate them in details in the following sections [9].

3 Problem Solution
The aim of this work is to predict the structure of protein by giving the DNA sequence through two prediction techniques, neural networks, and Hidden Markov Model, HMM.

3.1 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. 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.1 [10]. The multi-layer backpropagation networks have been selected to predict protein structures because; the properly trained of these networks tend to give reasonable answers when presented with inputs that they have never seen.

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.

![Fig.1 Neural network function](image-url)
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. 2 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:

\[ Q = \frac{\sum_{x=1}^{N} P(x)}{N} \]  

(1)

Where;

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

\[ P(x) = \frac{X(x,y) - Er}{X(x,y)} \times 100\% \]  

(2)

And

\[ Er = \frac{\|Sp(x,y) - So(x,y)\|}{X(x,y)} \]  

(3)

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.

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

![Fig. 2 (a) The original structure of one of Hemoglobin structures in the database, (b) the predicted structure using neural network.](image)

3.2 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 figure 3 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.

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. A generic hidden Markov model is illustrated in Fig. 3, where the \( X_i \) are the hidden states and all other notation is 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 \( O_i \), which are related to the actual states of the Markov process by the matrices \( B \) and \( A \) [11].

![Fig.3 The architecture of HMM](image)
So; the probability of the state sequence X is given by:

\[ P(X) = \prod_{t=0}^{T} \pi_{X_t} A_{X_t, X_{t+1}} B_{X_{t+1}, k} \] (4)

Where;

- \( \pi \) is the initial state distribution.
- \( \pi_{x_0} \) is the probability of starting in state \( x_0 \).
- \( A \) is the state transition probabilities, \( A = \{a_{ij}\} \) is \( N \times N \) with \( a_{ij} = P(\text{state q}_j \text{ at } t+1 \mid \text{state q}_i \text{ at } t) \).
- \( B \) is the observation probability matrix, \( B = \{b_{j}(k)\} \) is an \( N \times M \) with \( b_{j}(k) = P(\text{observation k at } t \mid \text{state q}_j \text{ at } t) \).
- \( N \) is the number of states in the model.
- \( M \) is 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.

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 \( A \).

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 \( A, B, O \) & the initial \( \pi \) so; we could compute the sequence of the unknown states, the protein structure, according to the probability equation (4). \( P(x) \) 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 equations (1), (2) & (3), and Fig. 4 shows the original structure and the predicted one of one hemoglobin base as an example.

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. 5 & 6 show an example of an original and predicted structure.
for the gray level and colored form respectively.

(a)                                   (b)

Fig. 4 (a) The original structures of one Hemoglobin structures in the database, (b) the predicted structure using Hidden Markov model.

(a)                                   (b)

Fig. 5 (a) The original structures of one Hemoglobin structures in the database, (b) the predicted structure using Hidden Markov model for gray images.

(a)                                   (b)

Fig. 6 (a) The original structures of one Hemoglobin structures in the database, (b) the predicted structure using Hidden Markov model for colored images.

3.3 Secondary Structure Prediction

The linear ordering of amino acids forms the secondary structure. Secondary structure can be divided into several types, though usually at least three classes: alpha-helix, beta-sheet and coils as shown in Fig. 7. **Alpha-Helix:** It resembles a ribbon wrapped around a tube similar to circular staircase. This structure is very stable but flexible, seen in parts need to move or bend. **Beta-Sheet:** Two or more ribbons of amino acids are involved parallel or anti-parallel. This structure is similar to folds in fabric. It is rigid and less flexible. Coil or turns are often responsible for twists in alpha helix and hair-pins in beta sheets [12].

Fig. 7 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. 8 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 [12]. This high success ratio because of dealing only with one class of protein-hemoglobin- instead of using many types of proteins.
4 Conclusion

The aim of this paper was to present a system that can predict the structure of a specific human protein, hemoglobin, from its DNA sequence by fast and easy way. 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: