Hidden Markov Model and Chapman Kolmogrov for Protein Structures Prediction from Images

Md.Sarwar Kamal[1], Linkon Chowdhury[2], Mohammad Ibrahim Khan[2], Amira S. Ashour[3], João Manuel R.S. Tavares[4], Nilanjan Dey[5]

[1]East West University Bangladesh, sarwar.saubdcoxbazar@gmail.com
[2]Chittagong University of Engineering and Technology, linkoncuet@gmail.com
[2]Chittagong University of Engineering and Technology, muhammad_ikhancuet@yahoo.com
[3]Department of Electronics and Electrical Communications Engineering, Faculty of Engineering, Tanta University, Egypt (email: amirasashour@yahoo.com)
[4]Instituto de Ciência e Inovação em Engenharia Mecânica e Engenharia Industrial, Departamento de Engenharia Mecânica, Faculdade de Engenharia, Universidade do Porto, Porto, Portugal (email: tavares@fe.up.pt)
[5]Department of Information Technology, Techno India College of Technology, West Bengal, 740000, India (email: neelanjan.dey@gmail.com).

Abstract:

Protein structure prediction and analysis are more significant for living organs to perfect asses the living organ functionalities. Several protein structure prediction methods use neural network (NN). However, the Hidden Markov model is more interpretable and effective for more biological data analysis compared to the NN. It employs statistical data analysis to enhance the prediction accuracy. The current work proposed a protein prediction approach from protein images based on Hidden Markov Model and Chapman Kolmogrov equation. Initially, a preprocessing stage was applied for protein images' binarization using Otsu technique in order to convert the protein image into binary matrix. Subsequently, two counting algorithms, namely the Flood fill and Warshall are employed to classify the protein structures. Finally, Hidden Markov model and Chapman Kolmogrov equation are applied on the classified structures for predicting the protein structure. The execution time and algorithmic performances are measured to evaluate the primary, secondary and tertiary protein structure prediction.

Keyword: Hidden Markov Model, Chapman Kolmogrov equation, Flood Fill, Warshall algorithm, Tertiary Structure

1. Introduction

Proteins are the complex biological organic and macromolecules patterns. For body cells, tissues and organs structure and functionalities, portions have a significant role. Proteins consist of amino acids and form a peptide bond by joining different peptide bond. Different protein structures have different functionalities as enzymes or form enzymes' subunits of proteins play a structural or mechanical role. Some proteins are used to transport various structural ligands and some function in immune response. Proteins provide the organism with certain amino acids that are not synthesized by that organism.

An amino acid is macromolecule that contains both an amino group and a carboxylic acid group. An amino acid forms a peptide bond after loses a water molecule. There are 20 different amino acids in nature that form proteins. These 20 are encoded by the universal genetic code. Nine standard amino acids are called "essential" for humans because they cannot be created from other compounds by the human body. Proteins consist of amino acids joined together by peptide bond to form a polypeptide. Resolving the functional variation of protein is challenging issue in molecular biology [1]. Structural knowledge is important for protein functionality analysis. Three-dimensional protein structure prediction is considered the greatest challenge for the structural biologist. Many computational studies and techniques were conducted for structural protein analysis. Such techniques include the evolutionary algorithm [2,3], Monte carlo [4,5] and HP model [6,7]. Genetic algorithm (GA) can be chosen to solve the Protein Structure Problem (PSP). Its performance varies due to the assigned GA parameters. The GA computational steps are involved to find the optimal strategies for large space. Different GA variants [8-17] are used for solving the PSP. This GA variants performance varies due to the different inputs parameters. Some GA approaches have less performance due to the manual tuning of the GA parameters by trail-and-error method.

Furthermore, several biological structures types are preprocessed by using image processing approach. Binarization approach or threshold approach are used for pre-processing [18]. Image thresholding or histogram can be used to

classify the objects into intra-/inter- classes [18]. Multi-level thresholding is another pre-processing algorithm that segment the color image based entropy [19]. The fuzzy C-mean and rough set approaches are clustering and classification approaches [20-21].Otsu is another thresholding method that used the threshold globally and convert a color image into binary image [22]. After pre-processing different types of machine approaches are used for protein structural prediction [23]. Hidden Markov Model (HMM) [24-26], and Support vector Machine (SVM) [27-28] are more accurate match prediction approaches for protein structures. Typically, the HMM is a statistical approach, where the probabilities of certain prediction of two consecutive terms are calculated. It is applied for biological data analysis such as the secondary protein structures, Ribonucleic acid (RNA) secondary prediction, multiple sequence alignment, and Genome/Gene prediction. In addition, when HMM operates on protein data sets under some condition or parameters, it is known as Markov Chain (MC). Since the RNA contains different amino acid structures. Thus, the MC can predict easily certain properties based on certain parameters from large protein structure datasets. Furthermore, the Chapman Kolmogrov can predict any terms of consecutive.

Consequently, in the present work, Hidden Markov Model and Chapman Kolmogrov equation are applied for protein prediction. The protein structure classification is performed by using Chapman Kolmogrov and Hidden Markov Model. Otsu method is employed initially to convert the protein image into binary images as a pre-processing step. A structural prediction approach is proposed to predict different types of protein structures through two phases, namely: RNA structure filter phase and prediction phase. In the filter phase, Otsu method is used as pre-processing to convert the structural images into binary matrix. Image noise is removed in pre-processing phase. Flood fill and Warshall algorithms are used for classification structures. Furthermore, another two machine learning prediction approaches, namely Hidden Markov Model and Chapman Kolmogrov models are used to find the position of the ones and to match with the trained data. The structure of proteins structure is predicted based on one position. The proposed approach classified primary, secondary and tertiary structures by using the proposed machine prediction approaches.

2. Literature Review

Computational protein design is a challenging issue due to their single site mutation and redesign of protein [29-32]. In nature, protein structural design also challenging due to the protein binding affinity/specificity, enzymatic activity [33-34], new folds of protein creation [35] and functionalities [36]. Protein design has been used to find the sequence constraints generation of specific folds or functions [37-39]. These constraints are used to address the physical evaluation. Different types of computational approaches are used to predict the protein physical structure. Mainly, three computational approaches are used for predicting structure of protein: threading or folding reorganization, comparative modeling and ab initio prediction. The comparative modeling or homology modeling requires one or more 3D homologous protein structures. Protein fold recognition handles the non-homologous protein structures. However, all types of protein structures have the similar folds. Several types of machine learning algorithm based predictors have been developed to predict protein folds [40-43] and protein structural classes [44,-46] using features, such as the secondary structure profile, HMM profile and PSSM (position-specific scoring matrices) profile. However, ab initio approach for PSP is the most challenging, which is not previously solved structures. The ab initio prediction of protein structure can illustrate the 3D structure of protein transformation process from sequence from into structure from scratch. To solve PSP, an efficient prediction algorithm along with an optimal energy function [47-49]. A well-defined fitness or energy function to recognize the final goal from the random conformation led to the simplified model based PSP problem [50].
Furthermore, the GA, which is an adaptive heuristic search and optimization algorithm, is used for PSP [51-53]. Nowadays, Eva on-line evaluation [54] is the top performing methods include several approaches based on neural networks, e.g. PSIPRED [55], PROFsec and PHDpsi [56]. Recently, researches applied the SVM (support vector machine) for secondary structure prediction [57-59]. Another effective research methodology for secondary structure prediction is the Hidden Markov Models (HMM) .These models show their ability by allowing an explicit modeling of the data. Asai *et al*. [60] predicted secondary structure using HMMs. Four sub-models were applied separately for particular local structures, namely alpha, beta, coil and turns in pre-clustered phase sequences. The sub-models contain four or five hidden states merged into a single model. Each model achieved 54.7% accuracy

rate. In order to represent specific classes of protein, a collection of HMMs algorithm were used [61-62].These models were constructed a generalized approach that reduce the connectivity and surface loops or turn size [61]. This involved two types of distinct helices position, namely: the N-cap and C-cap positions in helices. An explicit model also designs for amphipatic helices and β-turns. The HMM was applied to perform the secondary structure prediction.

For several amino acid group predictions, semi-HMM model, which is an extension of the HMM was also applied [63-65]. These models allow an explicit consideration for different length of secondary structures. Another machine learning approaches by using homologous sequence was proposed with single sequences with higher accuracy [65-66]. A novel HMM was depicted for protein secondary structures by using prior biological knowledge [67]. By using biological knowledge, HMM is an interesting tool to reveal hidden features of the internal architecture of secondary structures. Novel HMM first analyze in detail the model and predictive potential on single sequences and on multiple sequence information. In this approach, an evaluation data set of 506 sequences and data set of 212 sequences obtained from the EVA Web site [68]. Novel HMM more successfully predict secondary structure than traditional HMM. Different types of biomedical and scientific image processing approach are used for feature or structure selection. SCIFIO (SCientific Image Format Input and Output) is a library that a Bio-Formats to create a domain-independent image I/O framework [78].The goal of SCIFIO is to integrate the scientific formats between Digital image and medicine. DICOM [79], Flexible Image Transport System (FITS) [80] and netCDF [81] are common image I/O framework for scientific and structural data analysis. Several graph theory are approaches are applied in order to identify the protein interaction. Dynamic PPIs Alignment System (DPPIsAS) [82] is an algorithm based system that can be employed to find out the proteins associated interaction by calculating the protein Road Discovery (PRD) in a certain network. Bi-partite graph theory and Protein Road Maintenance (PRM) are used for finding the protein path discovery. Canonical Correlation Analysis (CCA) is also used to find out the degree of correlation among the proteins in a certain network. Furthermore, different types of computational approaches are carried out to classify reliable and unreliable PPIs [83]. Two methods for detecting the reliable PPIs and for evaluating the performance several methods have been reported.

3. Methodology

Different machine learning and structural prediction approaches are used for protein structure prediction, where mathematical and statistical approaches were applied to predict 2D and 3D structure of protein structure. Protein images' noise as well as machine learning complexity affects negatively the protein structure prediction.

In the current work, an approach combining image processing techniques and machine learning for protein structure prediction. The protein structure image is pre-processed and converted into binary image as an array of 0's and 1's.

3.1. Preprocessing

Noise removal, sharpness enhancement and edges blaring are essential before any further processes. Different types of filtering approaches are used for noise removal and sharpness enhancement. Several algorithms [**69**] including Gaussian filtering, Weiner filtering, frequency domain filtering are used for pre-processing. However, such filters suffer from several problems, such as inevitable loss of image, loss of sharpness and ringing effects. Different rank algorithms are more popular rank algorithms are used for image filtering [**70**]. Non-local means filtering is more popular in rank algorithm approach for noise reduction [**71**]. It operates on the average pixels rather than pixels values on their neighbor statistics [**71-72**]. The non-local means filtering calculates the average weighting of the neighborhood pixels. In the current work, the bilateral filtering algorithm is applied. It is a well-known non-local means filtering for pre-processing that performs edges preservation and image enhancement [**73**]. The bilateral filter is used for edge-preservation and noise reduction of images. At each pixel, the intensity value is replaced by the intensity values' weighted average from nearby pixels. The weights depend on the Euclidean distance of the pixels as well as the radiometric differences. The bilateral filter is applied on an image I_{input} and generates filtered image I_{out}, which is formed as a weighted sum function from its neighborhood pixels θ. The generated filtered image is given by [73]:

$$I_{out}(x,y)=\frac{1}{\sum_{i,j\in\theta} w}\sum_{i,j\in\theta} w(x,y,j,i).I(x+j,y+i) \qquad (1)$$

The weight w depends on the geometric distance and color differences between pixels (x,y) and $(x+j,y+i)$. In bilateral filtering, color and spatial coordinates between two pixels are considered to find the pixels similarity for filtering using the following expression:

$$W(x,y,j,i)=\exp\frac{i^2+j^2}{-2\sigma^2}.\exp\frac{(I(x+j,y+i)-I(x,y))^2}{-2\beta^2} \qquad (2)$$

Here, σ^2 and β^2 are the variance of color and spatial pixels coordinates sets. The weight indicates the similarity of color and spatial coordinates pixels. Weight is vary from color intensity and distance of pixels. Comparing the value of two pixels based on content of image patch. The weight of image patch (small square of images) is calculated by:

$$w(x,y,j,i)=\exp\frac{\left|v(x+j,y+i)-v(x,y\right|_2^2}{-2\beta^2} \qquad (3)$$

The square norm of pixel wise pitch difference ensures the average pixels of surrounding contents. Three-dimensional (3D) summation and calculation of weight in equation (3) is adjusted, where the similar patches are searched among the adjacent slices. In the pre-processing phase, sharpens of image is increased by using bilateral filtering algorithm. It reduces the noise and corrects the edges. Afterward, the image is converted to binary for further enhancement for the image quality as follows.

3.2. Binarization

Binarization is a process of separating the image pixels into two groups, namely white pixels that indicate the background pixels and black pixels that indicate the foreground pixels. Image binarization is based on certain threshold value using one of the most common techniques, namely Otsu thresholding method [74]. Otsu thresholding is an iterative process over all threshold values that calculates and measures the pixels' spread as foreground or background. Otsu method applies global thresholding for generation binary matrix. It is employed in the current work is to find out the optimal threshold at which the sum of foreground and background spread is less. Otsu approach generates an intensity bi-modal histogram that has sharp valley between two the peaks representing the foreground and background [**75**]. In histogram graphically represent 256 pixels image intensity in gray scale image and binary image consider two intensities (0 and` 1). In order to deploy the binarization of the image from gray one, let the pixels of a gray image represented in *L* gray levels. The number of pixels at level I is denoted by the total number of pixels $N=n_1+n_2+\ldots\ldots+n_L$. Suppose the pixels are divided into two groups: foreground (P_f) and background (P_b) by a threshold level, where P_f denotes the pixels with level $[1,2,\ldots,t]$ and P_b denotes the pixels with level $[t+1,t+2,\ldots\ldots,L]$. The calculations to separate the foreground and background using the variance is based on the threshold value *t*. For background pixels class P_b, the weight, mean and variance are given respectively by:

$$W_b=\sum_{i=1}^{t}\frac{n_t}{N} \qquad (4)$$

$$\mu_b = \frac{\sum_{i=1}^{t} t * n_t}{\sum_{i=1}^{t} n_t} \tag{5}$$

$$\sigma_b^2 = \frac{\sum_{i=1}^{t} (t - \mu_b)^2 * n_t}{\sum_{i=1}^{t} n_t} \tag{6}$$

In similar procedure, the foreground pixels class P_f has:

$$W_f = \sum_{i=t+1}^{L} \frac{n_t}{N} \tag{7}$$

$$\mu_f = \frac{\sum_{i=t+1}^{L} t * n_t}{\sum_{i=t+1}^{L} n_t} \tag{8}$$

$$\sigma_f^2 = \frac{\sum_{i=t+1}^{L} (t - \mu_f)^2 * n_t}{\sum_{i=t+1}^{L} n_t} \tag{9}$$

The within the class variance is then calculated from the sum of the two variances associated by their weights, which is expressed by:

$$\sigma_w^2 = \sigma_b^2 * w_b + \sigma_f^2 * w_f \tag{10}$$

Finally, the weighted variance (class variance) is compared with the threshold value. All pixels with a level less than the threshold value are considered background and pixels with values greater than the threshold value are considered foreground. The Otsu method convert the gray image into binary image based on threshold value that belongs a matrix of 1 or 0 (binary matrix).

In the present work, Otsu method is applied to allow global thresholding. Global thresholding generally simple and it allows many variations of thresholding for neighbor pixels. But it is failure when gray scale image range varies in locally. In addition, other binarization methods are applied to compare the evaluation the performance of Otsu method. These techniques are: i) the local thresholding [76], which is a binarization process that can convert the colored image into binary image. Local thresholding determine the threshold value with surrounding region. However, it requires more computation and it performs on uniforms neighbors. ii) Optimized thresholding [76] is another approach of binarization by using optimal mathematical approach. It is more accurate than local thresholding though its computational cost is high. The accuracy of this approach depends on mathematical optimization. In order to evaluate the performance of the different binarization methods, some metrics are measured including the Standard deviation, which indicates the error rate of the binarization process. High standard deviation means less accuracy and maximum error. In addition, the signal-to-noise ratio (SNR) is also measured, which indicates the strength of removing noise. Furthermore, the mean is measured to indicate the average numbers of 1's (ones). After the binarization process, the generated binary matrix is processed by hidden markov model and Champman kolmogrov approach. These approaches predict the RNA structure pattern using the 1/0 values.

3.4. Warshall's and Flood Fill Algorithm

The Warshall and flood fill methods are employed to process the binary matrix after the binarization process. Both methods are applied to measure the total number of 1's by counting the number of ones in the binary matrix as they predict the pattern of 1 or 0 for the RNA structure. The binary matrix is spilt into different portions to calculate the number of 1's. Flood fill is a pixel counting process by using 4 connected and 8 connected neighbors. In the present work, 4 connected approach is applied for counting in every portioned portion. Furthermore, the Warshall algorithm is also a counting approach that converts the binary matrix into adjacent matrix. Transitive closure or relation is generated from adjacent matrix from the counting 1's. Simple Boolean operation is performed in Warshall approach that is more efficient than flood fill algorithm. Multiple counting processes are performed in the flood fill algorithm that reduces the system performance. Typically, the Warshall algorithm for counting 1's in binary matrix designs an

adjacent matrix of transitive dependency. It generates the transitive binary matrix to calculate the number of 1's. Different protein structures belongs certain number of 1's. In the present approach, an adjacent matrix A is used by binary data processing, where the generated binary matrix from the binarization procedure is considered as an adjacent matrix. Thus, Warshall's algorithm, which is an efficient algorithm for finding out the effective adjacent matrix of transitive closure relation R is employed. The relation R derived from finite vertices set *S*, where *S* be the finite set $\{v_1, v_2, ..., v_n\}$ and R is a relation on S. The adjacent matrix *A* of R is an *n*n* Boolean matrix, which is defined by:

$$A_{i,j} = \begin{cases} 1, & if\ an\ edges\ from\ v_i\ to\ v_j \\ 0, & if\ no\ edges\ from\ v_i\ to\ v_j \end{cases} \tag{11}$$

An adjacent matrix T of the transitive Closure R is generated using the following algorithmic approach (**Algorithm 1**).

Algorithm 1: Warshall's algorithm

Start

Input: Adjacency matrix A of relation *R* on a set of *n* elements
Output: Adjacency matrix T of the transitive closure of *R*
Initialize the adjacent matrix T:=A
for j=1 to *n*
for i=1 to *n*
if $T_{i,j}$=1 then
Apply Boolean OR operation, $A_i = A_i \vee A_j$
Endif
End
End
T=A

Stop

In the present work, the trained protein structure converts into binary adjacent matrix. Since the test adjacent matrix contains 1 or 0, count the number of 1 in the adjacent matrix. Afterward, the adjacent matrix is compared with the trained datasets for certain protein structure.

In order to count number of 1's in the binary matrix, the flood fill algorithm is used. The flood fill is a linear searching approach that starts from a seed as an interactive flood system. Every pixel is connected by 4 connected way or 8 connected way. It specifies a seed by pointing to the interior of the region to initiate a flood operation. It starts from a seed and flood the whole region until the boundary region meet. From the seed, a search for '1' that is connected with seed in 4 ways (Left, right, up and down) is performed, afterward another seeds are chosen. In the current work, the recursive approach is applied to select the seeds that are called recursively until the whole counting is being completed. Generally, the flood fill recursive approach is basically used in paint system. In flood fill algorithm user select a seed and color the connected neighbor seeds. Flood fill performs in two different categories: 4 connected and 8 connected neighbor. In the current work, 4 connected approach is used for flood fill algorithm. A seed at the (0,0) position in the binary matrix is selected. The connected 4-neighbor position and count the number of 1 are used. Afterward, the next position of 0 or 1 is selected as a seed, and then scanned the 4 connected neighbors and counted the 0's or 1's. This process continues recursively until scanning the whole binary matrix is completed.

3.5 Hidden Markov Model and Chapman Kolmogrov Model

In order to find the positions of the ones from binary matrix, machine learning approaches, such as HMM and Chapman Kolmogrov (CK) are applied. The HMM is a recursive process in which the 1's positions are calculated by

using cumulative distribution. Cumulative distribution probability matrixes are generated for predicting the 1's position. Matrix multiplication operations are performed in HMM process to enhance the system complexity, while Chapman Kolmogrov (CK) has less system complexity. Matrix multiplication operations are not performed in CK operation that enhances system performance. The probabilities of 1's position are calculated directly without consecutive matrix operations.

Generally, the HMM requires consecutive matrix operation for finding certain position of 1, while the CK can directly find out the certain position of 1's. If the position of 1 in training data is similar to trained date the structure of proteins. The predicated positions are chosen randomly. Typically, the most common statistical approach for biological data analysis is applied, namely the Hidden Morkov Model (HMM). Furthermore, the CK approach is efficiently predicts a certain position of 1's superior to the HMM approach that operates on trained data.

It is applied to analyze several biological data such as secondary protein structures, RNA secondary prediction, multiple sequence alignment, Polygenetic analysis and Gene prediction. HMM is based on calculating the probabilities of certain prediction of two consecutive terms. In some cases, HMM is known as Markov Chain (MC), when it operates the whole datasets under some manners. Since the DNA contains billion on nucleotides, MC can easily predict certain properties based on certain factors from large biological datasets. Markov chain need limited memory space as well as time to find specific item from large and multiple biological items. It can effectively handle and analyze the discontinuity in long or very long DNA or protein sequences. The general expressions to define the overall impacts of the Markov Chain are as follows:

$$P_{xy}^{a+b} = \sum P_{xz}^{a} \cdot P_{zy}^{b} \tag{12}$$

Where,

$$P\{H_{a+1}\} = P_{x_0 x_1} \ldots P_{x_{b-1} x_b} \tag{13}$$

$$P\{H_{a+b} = y \mid h_a = x\} = \sum_{z=0}^{\infty} P\{H_{a+b} = y \mid h_{a+1} = z, h_a = x\}.P\{H_{a+1} = z / h_a = x\} \tag{14}$$

$$P\{H_{a+b} = y \mid h_a = x\} = \sum_{z=0}^{\infty} P_{xz}.P_{zy} \tag{15}$$

$$P\{H_{a+b} = y \mid h_a = x\} = P_{xy}^{b} \tag{16}$$

Consequently, the HMM is used to predict specific binary number position by measuring the probability of one's position from the proposed binary matrix. The probability of one's positions are varies for different secondary protein structure. Chapman Kolmogrov equation is the optimal prediction characteristic under certain condition. Chapman Kolmogrov equation finds out the probability from one step to another step of certain events. Transitions matrix and transition equation are design for Chapman Kolmogrov equation. Both HMM and Chapman Kolmogrov equation are stochastic (random) process. In Morkov property, the probability distribution of the current value is conditionally independent of the series of past value. On the other hand, Chapman Kolmogrov equation predicts the discrete value on certain condition. HMM perform cross check between the trained and training datasets that used for protein structures prediction.

3. 6. Proposed Method

Generally, several researchers were interested with employing image processing in the prediction of the protein structures. He *et al*. [85] adapted a matrix-matching algorithm from image processing to offer an effective sequence matching process. The similarity degree of two sequences was calculated using a fast normalized cross-correlation (FNCC) algorithm adjusted from image processing. Chetia and Sarma [86] proposed an approach to predict the protein structure using soft computing techniques. The artificial neural network (ANN), image processing and statistical techniques were carried out to formulate the structure for protein prediction. In the present work, different machine learning approaches are applied for pattern re-organization using the procedure in Fig. 1.

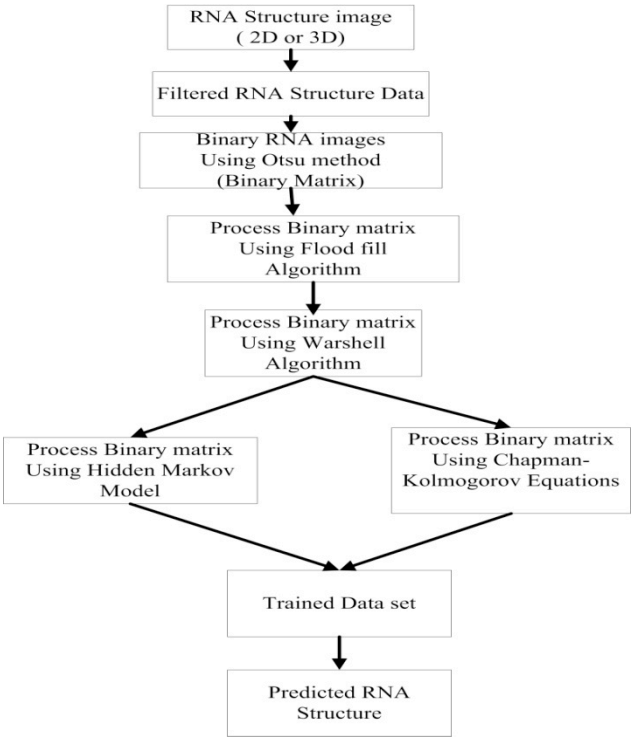


Figure 1: The proposed integrated prediction approach using machine learning approach

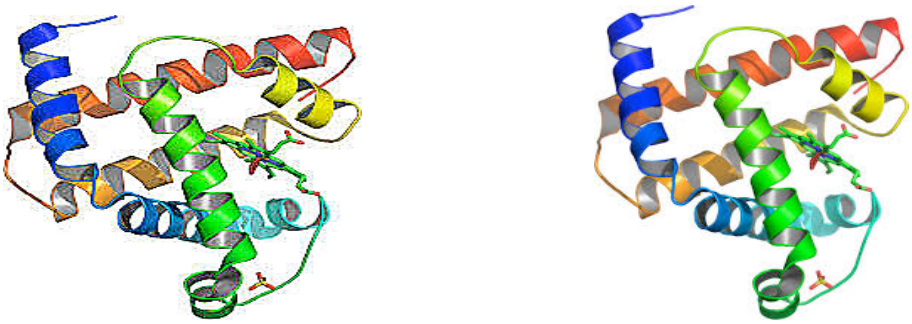
The two-/three-dimensional RNA images are used as input to the proposed system as illustrated in Figure 1. These protein images are processed into binary image by using Otsu method is a binary matrix after the Bilateral filtering pre-processing step. The binary matrix is handled by using Warshall algorithm, Flood fills, Hidden markov model and Chapman Kolmogrov equation for protein prediction. Typically, the flood fill and Warshall algorithms are used to classify the structures by counting the number of one in binary matrix that classifies the structure based on number of one. Markov chain and Chapman Kolmogrov equation are used to estimate the probability of the binary matrix, where the flood fill and Warsheal algorithms are applied to measure the total number of 1's. Thus, these approaches predict the pattern of 1 or 0 for RNA structure prediction. The final output from the Chapman Kolmogrov equation is considered trained data that predict the protein 2D and 3D structure. The obtained output is compared with the tested data to predict the protein structure. In addition, the accuracy rate and system performance among the machine learning approaches are evaluated.

The proposed protein structure prediction approach has been implemented by MathLab and PHP environments. The MathLab handles the protein image and convert into binary image, afterward the PHP environment is designed to analyses the protein structure by using different machine approaches. The classified different protein structure has different positions of 1 and 0 that optimized the classification accuracy. Three datasets are used for the present experiment including primary structure; secondary structure and Tertiary structure. Primary structure is a simple sequence with different types of amino acids for Drosophila melanogaster spices. Different types of primary structures are used for structure prediction. In the current work, the experimental of all 2D and 3D images are generated by using Cytoscape tools [84]. We generate protein image based on different amino acids interactions. In primary structure we select small number of protein instance from database. 2D and 3D structure protein images are generated by using large volume of protein data. Image size varies on different interaction of amino acids. When the interacted amino acid that generated secondary and tertiary image. We measure the image size from Cytoscape tools. Typically, the size (number of the 1's and 0's) of the primary structure images is not more than 20KB. In the current work, 85 different types of primary images are used. In addition, secondary images of size range between 15 KB and 35 KB. Ninety-six different types of secondary structure are collated for protein structure prediction. Furthermore, tertiary protein structure is collected, where the tertiary images have sizes of values between 45 KB and 90 KB. In the present work, 105 different types of tertiary structures are generated from Cytoscape tools for

protein structure prediction. The structure of protein is extracted from the different types of protein images. Typically, the different protein structures have indicated different types of biological and organic functionalities in the living organs. Primary protein structure indicates the linear sequence of the amino acid that formed by peptide bonds. In the current work, the protein folds are extracted in the primary structure. In the secondary protein structure, the linear, unfold and helical features are extracted in the polypeptide chains. Twisted and bending features are extracted from tertiary protein structures.

4. Results Analysis and Discussion

Initially, the Bilateral filtering algorithm is employed as a pre-processing step to attain sharp image as illustrated in Figure 2.



(a) (b)

Figure 2: Bilateral filtering algorithm as a pre-processing: (a) 3D images before pre-processing (b) Increase the sharpness after pre-processing.

Afterward the Otsu method is applied to convert a color image into binary image as illustrated in Figure 3.

(a) (b) (c)

Figure 3: Conversion process from gray image to binary image (a) A secondary protein gray image that contains 256 colors, (b) Binary image that contains two color intensities of values 0 or 1. (c) Histogram that separates foreground and background color based on threshold.

The flood fill and Warshall approaches are used for counting the 1's in the binary matrix. Protein structures are predicted based on the number of 1. The range of 1's for different primary, secondary and tertiary protein structures is assigned from the current experimental results. In addition, there exist a difference in the shape of the protein structures based on the different number of ones in the different protein structures, which is used for the predication and classification. In primary structure, the ranges of numbers of 1 are 102 to 114. In the secondary protein structure, number of 1's is greater than 115. These counting performed by flood fill approach on binary matrix. We classify primary and secondary structure based on these numbers on 1's ranges. All of these prediction algorithms operate on binary matrix. Otsu method is used to generate binary matrix from a protein image. Markov chain and Chapman Kolmogrov equation are used to measure the probability of binary matrix to classify the protein structures.

The three datasets have been used for this experiment including primary structure; secondary structure and Tertiary structure as reported in Table 1.

Table 1: Different protein structure, samples and size for three data sets

Dataset	Number of Samples	Sample size (KB)
Primary Structure	85	≤20
Secondary Structure	96	15 to 35
Tertiary Structure	105	45 to 90

In Table 1, the primary data structure image size is minimum than other two data sets .This size refers total number of 1 and 0 in different protein structures. Secondary structure images size have average image size. In addition, the tertiary (3D) structures have images of maximum size compared to the other datasets. Generally, using the proposed method indicate that every binary image includes different binary matrices which have more similar or different number of 1's and 0's. The binarization process's execution time for the different structure images is measured. Flood fill and Warshall algorithm is used for counting number of 1's. It optimizes the sample classification and unnecessary structures. Finally, the HMM and Chapman Kolmogrov approaches are used to predict the final protein structures.

4.1. Performance analysis of the binarization approaches

The binarization processes are varies based on the threshold values. In the present work, different binarization methods are applied on different protein structures, namely the Otsu method, Local Thresholding (LT) and Optimized Thresholding (OT). The datasets contain different length of structures. Primary structure contains 85, Secondary structure contains 96 and a Tertiary structure contains 105 protein structures; respectively with different sizes, where increased the structures size leads to increased binarization execution time. Different performance parameters including the SNR, mean, standard deviation (SD) and execution time for each structure are measured to compare the different binarization approaches as illustrated in Table 2.

Table 2: Measurement of performance parameters for different protein structures

Protein Structures	Data Size (KB)	Otsu Method			Local Thresholding			Optimized Thresholding		
		Mean	Standard Deviation	SNR	Mean	Standard Deviation	SNR	Mean	Standard Deviation	SNR
Primary	0 to 5	41.35	9.87	41.93	38.87	17.57	38.74	42.18	13.62	40.73
	6 to 10	42.34	9.91	41.45	39.80	17.64	38.34	43.19	13.68	40.35
	11 to 20	42.98	10.45	41.14	40.40	18.60	37.7	43.84	14.42	39.76
Secondary	15 to 20	40.35	11.34	38.23	33.89	17.69	33.23	39.54	13.15	35.83
	21 to 25	41.34	11.78	37.96	34.73	18.38	32.76	40.51	13.66	35.16
	26 to 35	40.98	12.45	36.78	34.42	19.42	32.28	40.16	14.44	34.28
Tertiary	45 to 60	39.67	9.91	36.13	38.08	17.74	30.53	39.27	13.77	32.73
	61 to 75	40.87	10.22	35.34	39.24	18.29	29.94	40.46	14.21	32.14
	76 to 90	41.76	11.14	35.1	40.09	19.94	28.75	41.34	15.48	31.15

Table 2 establishes that the SD of Otsu method is less than the others two methods. The accuracy rate of binary images conversation is high for Otsu method. In the primary structures, the average SD of Otsu, local thresholding

and the optimized thresholding are 10.1, 17.94 and 13.91; respectively. However, the SD is increased for the secondary and tertiary structures by using Otsu method. Optimized thresholding and local thresholding have maximum SD compared to Otsu method for secondary and tertiary structures, which establishes the superiority of Otsu method. Generally, Table 2 reports that the maximum SD is 19.94 for 3D protein structure binarization process which data size belongs 76 to 90 KB, while the minimum SD is 9.87 for primary structures by using Otsu method. In addition, the optimized thresholding achieves average level SD for the secondary structures. Figure 4 illustrates the performance of the different binarization approaches in terms of the different evaluation metrics reported in Table2.

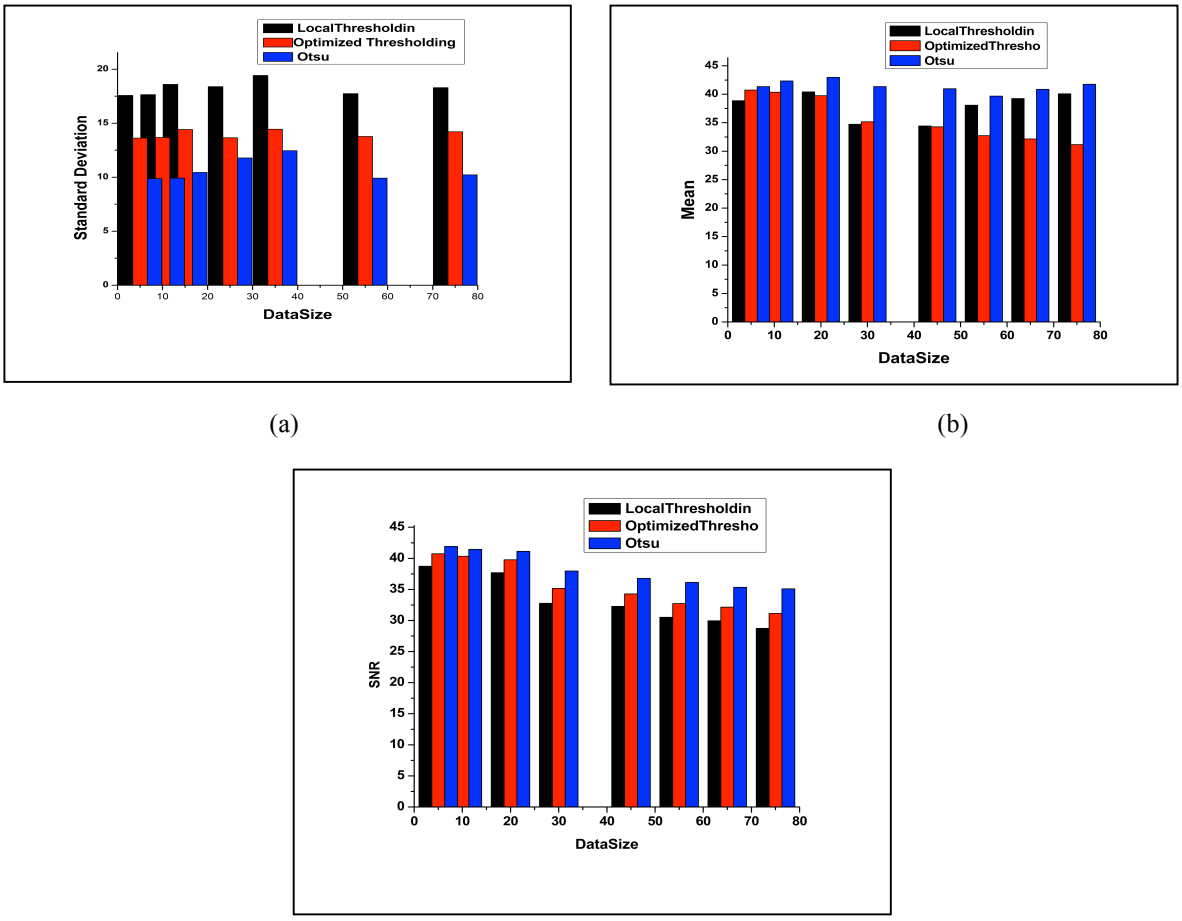


Figure 4: Evaluation of performance analysis for different protein structures binarization. (a) Standard deviation measurement of different protein structures. (b) Mean value for different protein binarization methods. (c) Measurement of SNR for three protein binarization process.

Figure 4 illustrates the performance evaluation for three binarization approaches for different protein structures. Primary structures primary size start from 5 to 20KB, secondary structures belongs 30 to 40 KB and approach and tertiary structures sizes are 55 to 75 KB. Thus, with the data size increase, the SD of Otsu method increases having less values compared to the other two binarization processes as demonstrated in Fig.4(a). In addition, the SNR of the optimized thresholding process is higher than the local thresholding as demonstrated in Fig.4(c). Furthermore, the execution time of the three binarization processes (Otsu method, local thresholding and optimized thresholding) for different protein structures of binary images conversion is measured as given in Table 3.

Table 3: Measurement of execution for different protein structures

Protein Structures	Data Size (KB)	Otsu Method (nano second)	Local Thresholding (nano second)	Optimized Thresholding (nano second)
Primary	0 to 5	3290285.97	7401622.21	5397538.34
	6 to 10	3529203.27	7861026.45	5989254.21
	11 to 20	3901470.34	8621017.65	6436887.45
Secondary	15 to 20	4295122.62	9372982.92	6920413.63
	21 to 25	4545162.43	10548581.54	7321850.55
	26 to 35	4806388.69	10883475.72	7424509.43
Tertiary	45 to 60	4968967.65	10919846.76	7780277.78
	61 to 75	5790848.66	12870506.25	8821850.54
	76 to 90	5966873.66	13464718.87	9338414.66

Table 3 depicts that the execution time is varies for different protein structure images, where the binarization time depends on number of pixels and structures. Thus, with increased number of images and size, the binarization execution time is also increased and vice versa. In addition, the execution time is also depends on the structures of images (primary, secondary and tertiary). Hence, for the primary structure of protein size is 11 to 20 KB, the execution time of Otsu method, local thresholding and optimized thresholding are 3901470.34ns, 8621017.65 and 6436887.45; respectively. The results depicts that the Otsu method is more about twice times faster than the Optimized thresholding approach, where Otsu process is (5397538.34-3290285.97)/ 5397538.34= 39.04% faster than the optimized thresholding for primary thresholding (Figure 4). Otsu method also faster than other two approaches for secondary and tertiary images. Otsu method required less time because it is used global thresholding for binarization. This global thresholding performed on within a class variable. The class variable separates the foreground and background and generates binary images. Figure 5 illustrates the execution time for Otsu method, local thresholding and optimized thresholding for protein structures prediction.

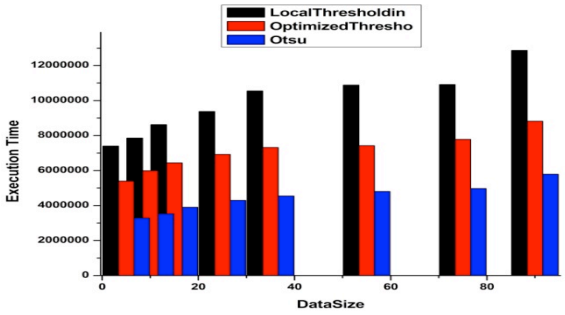


Figure 5: Execution time for Otsu method, local thresholding and optimized thresholding for different protein structures

Figure 5 illustrates the execution time of Otsu method, which is varies for secondary dataset of different images size than other two processes. For tertiary structure binarization execution times are varies due to different images sizes.

4.2. Performance Analysis of Flood fill and Warshall approaches

After the binarization process evaluation, the assessment for both the flood fill and Warshall approaches that used for the binary matrix processing is performed. The execution time for total number of 1's counting is measured, where the execution time increases with the binary size increase and vice versa as illustrated in Table 4. Binary image size depends on structure of protein. Since different protein structures belongs to certain number of 1's. In the

current work, the training dataset is used to generate the total number of 1 in the binary matrix and to classify the structures based on their 1's values.

Table 4: Execution time for flood fill and Warshall approach

Protein Structures	Data Size (KB)	Warshall (nano second)	Flood Fill (nano second)
Primary	0 to 5	4291285.07	6421642.41
	6 to 10	4529203.21	6661726.65
	11 to 20	4905420.33	7421217.55
Secondary	15 to 20	5265182.63	7572782.02
	21 to 25	5555132.63	9448681.24
	26 to 35	5806382.59	9983575.73
Tertiary	45 to 60	5966667.35	10169846.56
	61 to 75	6290542.64	11270526.35
	76 to 90	6366873.06	12464718.47

Table 4 depicts that the execution time of Flood fill and Warshall approaches varies for different protein structure length. The execution or counting time of theses approaches depends on the protein structures and binary matrix size. The execution time of both approaches increases with the increase in the size of binary matrix and vice versa. Warshall approach converts the matrix in adjacency matrix and count number of 1. It compares the number of 1 with trained data set and classifies the structures into primary, secondary or tertiary structures. In addition, the flood fill algorithm is also employed to count the number of 1's. Its counting operation is performed based on 4-connected or 8-connected approach. Flood fill algorithm chooses a position in binary matrix and count the number of 1 in 4-connected neighbor. This process is continued until the whole binary matrix scanning is completed. It is also compared with the trained dataset and is used to classify the protein structures. It required more execution time for similar data operation than Warshall approach. Figure 6 demonstrates the execution time comparison for both approaches.

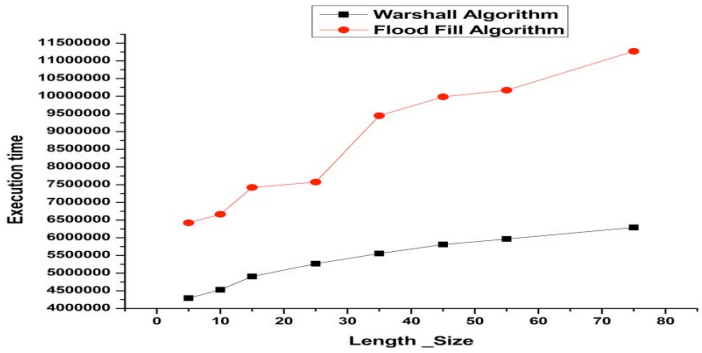


Figure 6: Execution time for Warshall and flood fill algorithm for different protein structures

Figure 6 illustrates the execution time for 1's counting by using Warshall and flood fill algorithms, where the execution time of the Warshall algorithm is less than the flood fill algorithm's execution time. All of the approaches linearly increased due to protein structure data length. In the secondary structure with 26KB to 35KB, the execution time of Warshall algorithm is 5806382.59 ns and flood fill algorithm is 9983575.73 ns. is more about 1.5 times faster than flood fill approach. Warshall approach is (9983575.73 -5806382.59)/ 9983575.73 =71.94% faster than the flood fill algorithm.

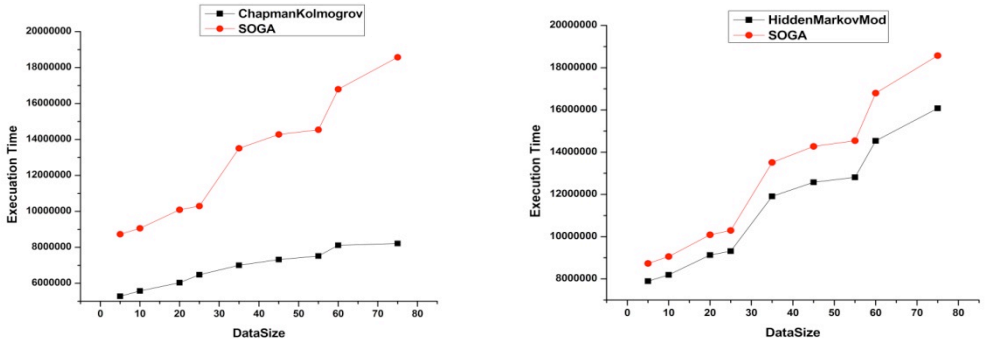
4.3. Performance Evaluation of HMM and Chapman Kolmogrov approaches

In order to evaluate the used machine learning approaches, namely the HMM and Chapman Kolmogrov (CK), the execution time is measured for both of them. In addition, the performance of the proposed approach is compared with the self-organization genetic algorithm (SOGA) [77] technique. Typically, the self-organizing system is a chemical, physical, or biological system without a central control. It is applied to obtain the pattern at global level of a system. In order to create an automated optimized solution, the SOGA has been used for Protein Structure Prediction (PSP). Furthermore, the execution time of SOGA for different protein structures is measured. Typically, the performance rate of the SOGA approach depends on the mutation and crossover process. The complexity of mutation and crossover process is increased with the different structures, which indicates high complexity rate of SOGA. Therefore, the proposed approach outperforms the SOGA as it easily handles the protein structure variation as demonstrated in Table 5.

Table 5: Execution time for Hidden Markov model, Chapman Kolmogrov and SOGA approach

Protein Structures	Data Size (KB)	CK (ns)	HMM (ns)	SOGA (ns)
Primary	0 to 5	5278280.64	7898620.164	8733433.678
	6 to 10	5570919.95	8193923.78	9059948.244
	11 to 20	6033667.01	9128097.587	10092855.87
Secondary	15 to 20	6476174.63	9314521.885	10298983.55
	21 to 25	6999467.11	11905338.36	13511614.17
	26 to 35	7316042.06	12579305.42	14276513.29
Tertiary	45 to 60	7518000.86	12814006.67	14542880.58
	61 to 75	8114800.01	14538978.99	16793084.26
	76 to 90	8213266.25	16079486.83	18572430.52

Table 5 depicts that the execution time of HMM, CK and SOGA approaches are varies for different protein structure length. The HMM and CK's execution time depends on the protein structures and number of prediction positions. The prediction positions indicate the presence 0's or 1's in certain index of binary image. The binary matrix is a two dimensional array. The presence of 0's or 1's is found out in the target position from the binary matrix. This target position is selected randomly. The execution time is increased for HMM and CK, when the size of target positions and protein structures are varies and vice versa. The HMM and CK approach compares the number of 1's with the trained dataset and classifies the structures into primary, secondary or tertiary structures. The SOGA measures the execution time for protein structures for mutation and crossover. The execution time of SOGA is increased with the increase of the data volume. In primary structure with 11KB to 20KB, the execution time of HMM and CK approaches are 9128097.587ns and 6033667.01ns respectively. SOGA approach needs 10092855.87ns for same data sets. CK needs less time than HMM and SOGA. CK and HMM approach are 5.17% and 13.42% less execution time respectively than SOGA. Figure 7 demonstrates the comparison of execution time for the SOGA and the HMM as well as with the CK.



(a) (b)

Figure 7: Execution time for different protein structures (a) Comparison between CK and SOGA approach based on execution time (b) Performance analysis between HMM and SOGA based on execution time.

Figure 7 illustrates that for every dataset the CK estimates less execution time than the SOGA. The CK used less execution time because CK measure probability from first step to another step. On the other hand, the SOGA uses the genetic algorithm, which is linearly search for the position content. Figure 7 (b) indicates the execution time of 1's or 0's prediction, showing that the HMM requires less execution time compared to the SOGA algorithm. All of the approaches are linearly increased due to protein structure data length. Due to various lengths with crossover and mutation approach generates significant number of processing steps that need more time. Furthermore, a comparison between the hybrid processes with HMM & SOGA and CK & SOGA is demonstrated in Figure 8.

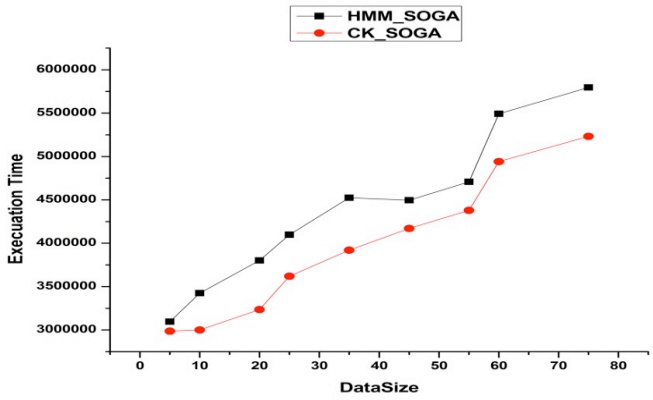


Figure 8: Comparison between HMM & SOGA and CK & SOGA based on execution time

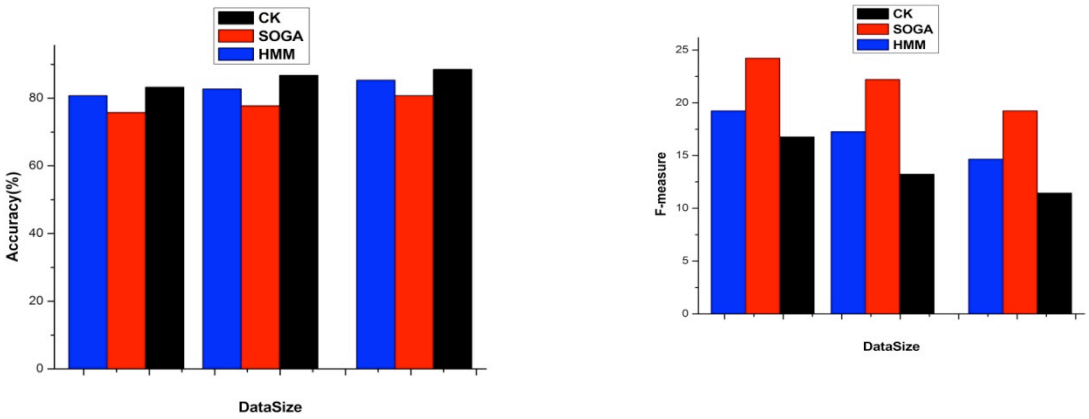
Figure 8 depicts that for every datasets the hybrid approach with CK & SOGA estimates less execution time than HMM & SOGA. Hybrid process with two approaches generates high accuracy than single approaches.

Since the F-measure indicates the false prediction of the data sample and the accuracy rate means the true prediction rate of data rate. In the current work, the F-measure and accuracy rate for different protein structures are measured. Different training sets are depicted of different protein structures based on the trained data. The HMM and CK approach on different structural datasets are operated. It is obvious that, when the datasets is increased and varies, the accuracy rates are also changed. Table 6 reports the execution time, sensitivity and F-measure for HMM, CK and SOGA.

Table 6: Accuracy rate and F-measure for Hidden Markov model, Chapman Kolmogrov and SOGA approach

Protein Structures	Number of Sample	CK (ns)		HMM (ns)		SOGA (ns)	
		F-measure	Accuracy	F-measure	Accuracy	F-measure	Accuracy
Primary	161	11.45%	88.55%	14.67%	85.33%	19.23%	80.77%
Secondary	134	13.24%	86.76%	17.26%	82.74%	22.21%	77.79%
Tertiary	93	16.76%	83.24%	19.24%	80.76%	24.23%	75.77%

Table 6 includes the performance metrics measurements for the 161 primary structures, 134 secondary structures and 93 tertiary structures. It is obvious that the accuracy rate is high for primary structure compared to the secondary and tertiary structures. Consequently, when the approaches are applied on tertiary datasets, the accuracy rate is decreases and F-measure is increased. However, for all the structures, the CK's accuracy rate is high compared to the HMM and SOGA. The HMM accuracy rate is high and the F-measure is low in primary structures. On the other hand, the SOGA accuracy rate is less for every type of the protein structures. Figure 9 demonstrates the accuracy and the F-measure values for the different dataset's size.



(a) (b)

Figure 9: Measurement accuracy rate and F-measure for different protein structures (a) Comparison between CK, HMM and SOGA approach based on accuracy rate (b) Performance analysis between CK, HMM and SOGA based on F-measure.

The bar charts illustrate the comparison between Chapman Kolmogrov, HMM approach and SOGA based on accuracy and F-measure in Figure 9. For every dataset, the CK estimates high accuracy and less F-measure than the SOGA. On the other hand, the SOGA used genetic approach that linearly searches position content. The HMM also achieves less F-measure compared to the SOGA method for every type of the protein structures.

The preceding results established that the HMM is accurately measure the primary structure and rate gradually decrease for secondary and tertiary approaches. However, the accuracy values decreased with the increase in the protein structure dimension as the number of samples for each structure increased. Consequently, it is suggested to apply other optimization algorithms to support the classifier more accurately. Furthermore, in the current proposed approach the count of the ones was used, however, in further works the count of the number of zeros can be used.

Conclusion

The current work proposed Hidden markov model and Chapman Kolmogrov for classifying different types of protein structures. These machine learning approaches are optimal and require less execution time for protein structure prediction. Both HMM and CK approaches operated on binary matrix. The binary matrix is generated by Otsu method that converts protein images into binary matrix. Flood fill and Warshall algorithm were used for counting 1's from the binary matrix for further protein structures classification. The HMM and CK were used then were applied to classify the structures. The simulation results established that the HMM and CK are effective for protein structures analysis based on protein structures images.

The results established that higher accuracy was obtained to classify the primary structures compared to classifying secondary and tertiary structures. Therefore, the accuracy rate decreased while classifying the tertiary datasets. For all structures, the accuracy rate using the CK method is higher compared to the HMM and SOGA. Instead, the SOGA accuracy rate is less for all protein structures.

References:

[1] Manish Kumar An Enhanced Algorithm For Multiple Sequence Alignment Of Protein Sequences Using Genetic Algorithm. EXCLI Journal;14:1232-1255,2015.

[2] Lijun Quan, Qiang Lv, and Yang Zhang. STRUM: structure-based prediction of protein stability changes upon single-point mutation. Bioinformatics, 2016, 1–11

[3] Brender,J.R. and Zhang,Y. (2015) Predicting the effect of mutations on protein-protein binding interactions through structure-based interface profiles. PLoS Comput. Biol., 11, e1004494.

[4] Carnevali P, Toth G, Toubassi G, Meshkat SN. Fast protein structure prediction using monte carlo simulations with modal moves. J Am Chem Soc 2003;125(47):14244-5

[5] Lee B, Kurochkina N, Kang HS. Protein folding by a biased Monte Carlo procedure in the dihedral angle space.FASEB J 1996;10(1):119-25.

[6] Mandal S, Jana ND. Protein structure prediction using 2D HP lattice model based on integer programming approach.In: Proceedings of 2012 International Congress on Informatics, Environment, Energy and Applications.2012 Mar 17-18; Singapore. p. 171-5.

[7] Benitez CM, Lopes HS. Protein structure prediction with the 3D-HP side-chain model using a masterslave parallel genetic algorithm. J Braz Comput Soc 2010;16(1);69-78.

[8] Cui Y, Chen RS, Wong WH. Protein folding simulation with genetic algorithm and supersecondary structure constraints. Proteins 1998;31(3):247-57.

[9] Zhang X, Wang T, Luo H, Yang JY, Deng Y, Tang J, et al. 3D protein structure prediction with genetic tabu search algorithm. BMC Syst Biol 2010;4 Suppl 1:S6.

[10] Hoque MT, Chetty M, Sattar A. Genetic algorithm in ab initio protein structure prediction using low resolution model: a review. In: Sidhu AS, Dillon TS, editors. biomedical data and applications. Heidelberg, Germany:Springer; 2009. p. 317-42.

[11] Dandekar T, Argos P. Applying experimental data to protein fold prediction with the genetic algorithm. Protein Eng 1997;10(8):877-93.

[12] Contreras-Moreira B, Fitzjohn PW, Offman M, Smith GR, Bates PA. Novel use of a genetic algorithm for protein structure prediction: searching template and sequence alignment space. Proteins 2003;53 Suppl 6:424-9.

[13] Goldberg DE. Genetic algorithms in search, optimization and machine learning. Reading (MA): Addison-Wesley Publishing Co.; 1989.

[14] Kaiser CE, Merkle LD, Lamont GB, Gates GH Jr, Pachter R. Case studies in protein structure prediction with realvalued genetic algorithms. In: Proceedings of the 8th SIAM Conference on Parallel Processing for Scientific Computing; 1997 Mar 14-17; Minneapolis, MN.

[15] Day RO, Zydallis JB, Lamont GB, Pachter R. Solving th protein structure prediction problem through a multi objective genetic algorithm. In: Proceeding of the International Conference on Computational Nanoscience and Nanotechnology; 2002 Apr 21-25; San Juan, Puerto Rico. p. 32-5.

[16] Deerman KR, Lamont GB, Pachter R. Linkage-learning genetic algorithm application to the protein structure Vol. 19 • No. 2 • June 2013 www.e-hir.org 147 prediction problem. In: Proceedings of the ACM Symposium on Applied Computing; 2001 Mar 11-14; Las Vegas, NV. p. 333-9.
[17] Schulze-Kremer S, Tiedemann U. Parameterizing genetic algorithms for protein folding simulation. In: Proceedings of the 27th Hawaii International Conference on System Sciences; 1994 Jan 4-7; Wailea, HI. p. 345-54.
[18] Wen-Bing Tao, Jin-Wen Tian, Jian Liu, "Image segmentation by three-level thresholding based on maximum fuzzy entropy and genetic algorithm" 2003 Pattern Recognition Letters 24, pp. 3069–3078.
[19] R.Sukesh Kumar, Abhisek Verma and Jasprit Singh, "Color Image Segmentation and Multi-Level Thresholding by Maximization of Conditional Entropy" 2007 World Academy of Science, Engineering and Technology on International Journal of Computer, Control, Quantum and Information Engineering, Vol: 1, No: 6, pp. 1607-1615.
[20] Debashis Sen and Sankar K. Pal, "Histogram Thresholding Using Fuzzy and Rough Measures of Association Error" 2009 ieee transactions on image processing, VOL. 18, NO. 4, pp. 879-889
[21] G.Padmavathi, M.Muthukumar and Mr. Suresh Kumar Thakur, "Nonlinear Image segmentation using fuzzy c means clustering method with thresholding for underwater images" 2010 IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 3, No 9,pp. 35-40
[22] Ch. Hima Bindu and K. Satya Prasad. An Efficient Medical Image Segmentation Using Conventional OTSU Method. International Journal of Advanced Science and Technology Vol. 38,2012,pp.67-74.
[23] Rost B (2001) Review: protein secondary structure prediction continues to rise. Journal of Structural Biology 134: 204–218.
[24] Aydin Z, Altunbasak Y, Borodovsky M (2006) Protein secondary structure prediction for a single-sequence using hidden semi-markov models. BMC Bioinformatics 7: 178.
[25] Yao XQ, Zhu H, She ZS (2008) A dynamic bayesian network approach to protein secondary structure prediction. BMC Bioinformatics 9: 49.
[26] Malekpour SA, Naghizadeh S, Pezeshk H, Sadeghi M, Eslahchi C (2009) A segmental semi markov model for protein secondary structure prediction. Mathematical Biosciences 221: 130–135.
[27] Guo J, Chen H, Sun Z, Lin Y (2004) A novel method for protein secondary structure prediction using dual-layer svm and profiles. PROTEINS: Structure, Function, and Bioinformatics 54: 738–743.
[28] Nguyen MN, Rajapakse JC (2004) Two-stage multi-class support vector machines to protein secondary structure prediction. In: Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing. pp. 346–357.
[29] Karanicolas J, Kuhlman B (2009) Computational design of affinity and specificity at protein-protein interfaces. Curr Opin Struct Biol 19(4):458–463.
[30] Kortemme T, Joachimiak LA, Bullock AN, Schuler AD, Stoddard BL, Baker D (2004) Computational redesign of protein-protein interaction specificity. Nat Struct Mol Biol 11(4):371–379.
[31] Shifman JM, Mayo SL (2003) Exploring the origins of binding specificity through the computational redesign of calmodulin. Proc Natl Acad Sci U S A 100(23):13274–13279
[32] Lopes A, Busch MSA, Simonson T (2010) Computational design of protein-ligand binding: modifying the specificity of asparaginyl tRNA synthetase. J Comput Chem 31 700(6):1273–1286
[33] Procko E, Hedman R, Hamilton K, Seetharaman J, Fleishman SJ, Su M, Aramini J, Kornhaber G, Hunt JF, Tong L, Montelione GT,Baker D (2013) Computational design of a protein-based enzyme inhibitor. J Mol Biol 425(18):3563–3575 707
[34] Jiang L, Althoff EA, Clemente FR, Doyle L,Rothlisberger D, Zanghellini A, Gallaher JL, Betker JL, Tanaka F, Barbas CF 3rd, Hilvert D, Houk KN, Stoddard BL, Baker D (2008) De novo computational design of retro-aldol enzymes. Science 319(5868):1387–1391 713
[35] Kuhlman B, Dantas G, Ireton GC, Varani G,Stoddard BL, Baker D (2003) Design of a Evolutionary Approach to Protein Design novel globular protein fold with atomic-level accuracy. Science 302(5649):1364–1368
[36] Siegel JB, Smith AL, Poust S, Wargacki AJ,Bar-Even A, Louw C, Shen BW, Eiben CB,Tran HM, Noor E, Gallaher JL, Bale J, Yoshi kuni Y, Gelb MH, Keasling JD, Stoddard BL,Lidstrom ME, Baker D (2015)

Computational protein design enables a novel one-carbon assimilation pathway. Proc Natl Acad Sci U S A 112(12):3704–3709
[37] Ollikainen N, Kortemme T (2013) Computaional protein design quantifies structural constraints on amino acid covariation. PLoS Comput Biol 9(11), e1003313
[38] Fromer M, Linial M (2010) Exposing the co-adaptive potential of protein-protein interfaces through computational sequence design. Bio-informatics 26(18):2266–2272
[39] McLaughlin RN, Poelwijk FJ, Raman A, Gosal WS, Ranganathan R (2012) The spatial architecture of protein function and adaptation.Nature 491(7422):138–142
[40] SHARMA, A., LYONS, J., DEHZANGI, A. & PALIWAL, K. K. 2013. A feature extraction technique using bigram probabilities of position specific scoring matrix for protein fold recognition. Journal of theoretical biology, 320**,** 41-46.
[41] PALIWAL, K. K., SHARMA, A., LYONS, J. & DEHZANGI, A. 2014. A tri-gram based feature extraction technique using linear probabilities of position specific scoring matrix for protein fold recognition. IEEE Transactions on NanoBioscience, 13**,**44-50.
[42] LYONS, J., BISWAS, N., SHARMA, A., DEHZANGI, A. & PALIWAL, K. K. 2014. Protein fold recognition by alignment of amino acid residues using kernelized dynamic time warping. Journal of theoretical biology, 354**,** 137-145.
[43] LYONS, J., DEHZANGI, A., HEFFERNAN, R., YANG, Y., ZHOU, Y., SHARMA, A. & PALIWAL, K. 2015. Advancing the accuracy of protein fold recognition by utilizing profiles from hidden Markov models. IEEE transaction on NanoBioscience, 14**,** 761-772.
[44] DEHZANGI, A., PALIWAL, K., LYONS, J., SHARMA, A. & SATTAR, A. 2013a. Exploring potential discriminatory information embedded in pssm to enhance protein structural class prediction accuracy. Pattern Recognition in Bioinformatics. Springer.
[45] DEHZANGI, A., PALIWAL, K., SHARMA, A., DEHZANGI, O. & SATTAR, A. 2013b. A combination of feature extraction methods with an ensemble of different classifiers for protein structural class prediction problem. Computational Biology and Bioinformatics, IEEE/ACM Transactions on, 10**,** 564-575.
[46] SAINI, H., RAICAR, G., SHARMA, A., LAL, S.,DEHZANGI, A., RAJESHKANNAN, A., LYONS, J., BISWAS, N. & PALIWAL, K. K. 2014. Protein structural class prediction via k-separated bigrams using position specific scoring matrix. J. Adv. Comput.Intell. Intell. Informatics, 8.
[47] COOPER, S., KHATIB, F., TREUILLE, A., BARBERO, J.,LEE, J., BEENEN, M., LEAVER-FAY, A., BAKER, D., POPOVIĆ, Z. & PLAYERS, F. 2010. Predicting protein structures with a multiplayer online game.Nature, 466 756–760.
[48] DAS, R. & BAKER, D. 2008. Macromolecular Modeling with Rosetta. Biochemistry, Annual Reviews, 77**,** 363-382.
[49] IQBAL, S., MISHRA, A. & HOQUE, M. T. 2015.Improved Prediction of Accessible Surface Area Results in Efficient Energy Function Application.Journal of Theoretical Biology, 380**,** 380–391.
[50] HOQUE, M. T. 2015. Genetic Algorithms based Improved Sampling. Tech. Report TR-2015/4.
[51] HOLLAND, J. H. 2001. Adaptation in Natural And Artificial Systems The MIT Press, Cambridge, Massachusetts London, England.
[52] MILAN, T. 2013. Artificial Bee Colony (ABC) Algorithm with Crossover and Mutation. *Appl. Soft Comput.***,** 687-697.
[53] IQBAL, S., KAYKOBAD, M. & RAHMAN, M. S. 2015a. Solving the multi-objective Vehicle Routing Problem with Soft Time Windows with the help of bees. *Swarm and Evolutionary Computation,* 24**,** 50-64.
[54] Koh I, Eyrich V, Marti-Renom M, Przybylski D, Madhusudhan M,Eswar N, Grana O, Pazos F, Valencia A, Sali A, Rost B: EVA: evaluation of protein structure prediction servers. Nucleic Acids Res2003, 31(13):3311-5.
[55] Jones D: Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999, 292(2):195-202.
[56] Przybylski D, Rost B: Alignments grow, secondary structure prediction improves. Proteins 2002, 46(2):197-205.6.
[57] Ward J, McGuffin L, Buxton B, Jones D: Secondary structure prediction with support vector machines. Bioinformatics 2003, 19(13):1650-5.

[58] Kim H, Park H: Protein secondary structure prediction based on an improved support vector machines approach. Protein Eng 2003, 16(8):553-560.
[59] Hu HJ, Pan Y, Harrison R, Tai PC: Improved protein secondary structure prediction using support vector machine with a new encoding scheme and an advanced tertiary classifier.IEEE Trans Nanobioscience 2004, 3(4):265-271.
[60] Asai K, Hayamizu S, Handa K: Prediction of protein secondary structure by the hidden Markov model. Comput Appl Biosci1993, 9(2):141-146.
[61] Stultz CM, White JV, Smith TF: Structural analysis based on state-space modeling. Protein Sci 1993, 2(3):305-314.
[62] White JV, Stultz CM, Smith TF: Protein classification by stochastic modeling and optimal filtering of amino-acid sequences.Math Biosci 1994, 119:35-75.
[63] Zheng W: Clustering of amino acids for protein secondary structure prediction. J Bioinform Comput Biol 2004, 2(2):333-42.
[64] Schmidler SC, Liu JS, Brutlag DL: Bayesian segmentation of protein secondary structure. J Comput Biol 2000, 7(1–2):233-248.
[65] Aydin Y, Altunbasak Z, Borodovsky M: Protein secondary structure prediction with semi-markov HMMs. IEEE International Conference on Acoustics Speech and Signal Processing 2004.
[66] Chu W, Ghahramani Z, Wild DL: A graphical model for protein secondary structure prediction, International Conference on Machine Learning. International Conference on Machine Learning 2004, 161:168.
[67] Aydin Z, Altunbasak Y, Borodovsky M: Protein secondary structure prediction for a single- sequence using hidden semi- Markov models. BMC Bioinformatics 2006, 7:178-178.
[68] Martin J, Gibrat JF, Rodolphe F: Choosing the Optimal Hidden Markov Model for Secondary-Structure Prediction. IEEE Intelligent Systems 2005, 20(6):19-25.
[69] Pratt W.K. "Digital Image Processing: PIKS Scientific inside" (4th ed.) Wiley-Interscience, John Wiley & Sons, Inc.,Los Altos, California, 2007.
[70] Yaroslavsky L.P., Kim V. "Rank Algorithms for Picture Processing" Computer Vision, Graphics and Image Processing,1986, Vol. 35, pp. 234–258.
[71] Buades A., Morel J.M. "A Non-Local Algorithm for Image Denoising" IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005, vol. 2, 20–26 June 2005, pp. 60–65.
[72] Dabov K., Foi A., Katkovnik V., Egiazarian K., "Image denoising by sparse 3D transform-domain collaborative filtering,"IEEE Trans. Image Process., vol. 16, no. 8, pp. 2080– 2095, August 2007.
[73] Tomasi C., Manduchi R. "Bilateral Filtering for Gray and Color Images" Proceedings of the IEEE Sixth International Conference on Computer Vision (ICCV'98), January 1998, pp.839–846.
[74] T. Romen Singh, Sudipta Roy and O. Imocha Singh, "A New Local Adaptive Thresholding Technique in Binarization", Department o Information Technology, School of Technology, Assam University,Silchar 788011, Assam, India.
[75] Nobuyuki Otsu, "A Threshold Selection Method from Gray-Level Histogram". New Afr. 9:62–66,1979.
[76] Gonzales, R. C. & Woods, R. E. 2007. Digital Image Processing, 3rd edn. Prentice Hall, Upper Saddle River, NJ, USA.
[77] Tragante-do-O V, Tinos R. A self-organizing genetic algorithm for protein structure prediction. Learn Nonlinear Model 2010;8(3):135-47.
[78] Mark C. Hiner, Curtis T. Rueden and Kevin W. Eliceiri. SCIFIO: an extensible framework to support scientific image formats. BMC Bioinformatics.(2016) 17:521
[79] Bidgood Jr WD, Horii SC, Prior FW, Van Syckle DE. Understanding and using DICOM, the data interchange standard for biomedical imaging. J Am Med Inform Assoc. 1997;4:199-212.
[80] Pence WD, Chiappetti L, Page CG, Shaw RA, Stobie E. Definition of the flexible image transport system (FITS), version 3.0. Astron. Astrophys.2010;524(Suppl Ser):A42.
[81] Unidata | NetCDF. http://doi.org/10.5065/D6H70CW6. Accessed 29 Nov 2016.
[82] Kamal, S., Xu, S., Nimmy, S.F., Khan, M.I. DGPPIsAS:A Dynamic Global PPIs Alignment System, IJCSNS International Journal of Computer Science and Network Security (2015) 15(2): 29-37.

[83] Peng, Z., Wang, J., Peng, W., Wu, F-X., Pan, Y. Protein–protein interactions: detection, reliability assessment and applications, Brief Bioinform (2016)

[84] Paul Shannon,Andrew Markiel,Owen Ozier,Nitin S. Baliga, Jonathan T. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Research, 13:2498–2504,2003

[85] He, Y., Rackovsky, S., Yin, Y., & Scheraga, H. A. Alternative approach to protein structure prediction based on sequential similarity of physical properties. Proceedings of the National Academy of Sciences, 112(16) (2015). 5029-5032.

[86] Chetia, S., & Sarma, K. K. Protein Structure Prediction using Certain Dimension Reduction Techniques and ANN. In communicated to 3rd International Conference on Computer and Communication Technology, ICCCT-2012, Allahabad (2012, December).