Kerbala Journal for Engineering Sciences

https://kjes.uokerbala.edu.iq

Automate Secondary Protein Structure Prediction based on spatial information

Nahla Ibraheem Jabbar

Department of Chemical Engineering , College of Engineering, University of Babylon, Babylon, Iraq.

E-mail: eng.nahla.ibraheem@uobabylon.edu.iq

Received: 08 November 2023; Revised: 30 November 2023; Accepted: 04 March 2024

Abstract

This study discusses the prediction of secondary structure protein from features of amino acid images. We define a predication algorithm depending on the spatial information of amino acid sequences. Algorithm A and algorithm B are applied to extract spatial information for clustering types of amino acids in three states: helix(H), strand(E), and coil(C). The accuracy of clustering depends on the improvement of algorithm B and the parameters of algorithms A and B. We apply a novel approach to spatial information extraction from primary protein structure to predicate three states of secondary protein structure. Primary protein sequences are divided into sub-images depending on the win window size. The accuracy of prediction is variable due to the size of the windows. Algorithms A and B are applied to achieve accuracy in each type of secondary protein structure. Helix (H) is 91.93%, Sheet(E) is 93.15%, and Coil (C) is 89.0126 %. In algorithms A and B, spatial information provides automated feature extraction to predict the secondary protein structure ion.

Keywords: Amino Acid sequence, Bioinformatics, Image processing, Features extraction and Secondary Protein structure.

1. Introduction

In recent years, bioinformatics has had a standard application in computer science [1]. It derives knowledge from computer analysis of biological data, consisting of information stored in a database [2]. Amino acid sequences represented information to understand the structure and function of the protein. It is one of the biological problems solving concerning information. Each primary protein structure consists of a linear sequence of amino acids. A string of letters arranges the sequences of the primary structure. Either a single-letter or three-letter ode represented each amino acid. Secondary proteins like drug design and novel enzymes are critical in life. In 1974, many primary structure applications were used to predict secondary structure, and the prediction accuracy reached 50% [3]. Statistical analysis in (GOR) method is applied in the secondary structure clustering using the conditional probability of structure [4].

As for clustering it is not limited to predicting the secondary protein structure .Most digital image clustering depend on the pixels of images , valued level of pixels takes extra time for understanding. In addition, the gray level values make images very difficult to interpret. This is a common problem in high dimensionality of data therefore, it is necessary to choose a solution for this problem by processing influence feature extraction with good accuracy. The influential feature is a key component of image assembly. It can reduce the dimensionality of image data [5]. Image feature extraction has positive effects on classification performance. Feature extraction techniques are applied in various digital image processing applications. The most common features extracted from an image are: - texture features, shape features, color, and spatial features [6]. Spatial is a technique for assigning pixel value based on neighbors and surrounding area. A window passes through the image with similar small arrays that work across the entire image through a convolution process[7].

The standard methods are applied in clustering or prediction secondary structure utilized neural network. Qian and Terrence are issued neural networks for prediction structure [8][9]. In 2015, Nazrul Mondal et al [10] improved six types of prediction. The results explain the prediction accuracies and H/C or are: 66.25%, 72.28%, 62.58%, 65.56%, and 70.85%. In 2022 [11], a new challenge in the secondary protein structure clustering combined natural language processing (NLP) and computer vision.

The main aim of our approach is to predict secondary protein structure depending on the spatial features of images. This approach avoids a common problem in other methods of testing and training. A desired sequence of samples was not obtained in the testing problem**.**

The framework of the proposal work relays in section 2. Sections 2.1, 2.2, and 2.3 explain how to get data and preprocessing steps. Section 2.4 is a manipulation with a grey level and neighborhood of center pixel. Algorithms A and B give the relation of spatial information distances around each pixel. The last section 3 has the experiment and result of the prediction of protein sequences depending on automated features prediction secondary.

2. Methodology

The strategy of predicting secondary structure has been discussed in this research. It is summarized in the flowing block diagram.

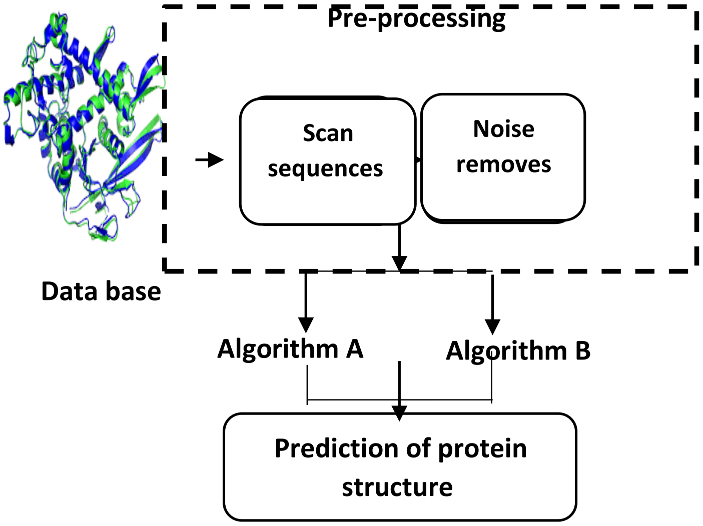

Figure 1. Illustrates the flow diagram of the work.

2.1 Data Base

The database of this work of Amino acids consists of 20 types of protein.

All are collected from a data bank [12]

Protein name

>Acprotease

Sequences

```
GVGTVPMTDYGNDVEYYGQVTIGTPGKFNLNFDTGSSNLWVGSVQCQASGCKGGRKFNPSD
GSTFKATGYDASIGYGDSASGVLGYDTVQVGGIDVTGGPQIQLAQRLGGGGFPGDNDGLLGG
FDTLSITPQSSTNAFQDVSAQGKVIQPVFVVYLAASNISDGDFTMPGWIDNKYGGTLLNTNIDG
EGYWALNVTGATADSTYLGAIFQAILDTGTSLLILPDEAAVGNLVGFAGAQDAALGGFVIACT
SAGFKSIPWSIYSAIFEIITALGNAEDDSGCTSGIGASSLGEAILGDQFLKQQYVVFDRDNGIRLP
VAGVGTVPMTDYGNDVEYYGQVTIGTPGKFNLNFDTGSSNLWVGSVQCQASGCKGGRKFNS
DGSTFKATGYDASIGYGDGSASGVLGYDTVQVGGIDVTGGPQIQLAQRLGGGGFPGDNDGLG
LGFDTLSITPQSSTNAFQDVSAQGKVIQPVFVVYLAASNISDGDFTMPGWIDNKYGGTLLNTNI
DAGEGYWALNVTGATADSTYLGAIFQAILDTGTSLLILPDEAAVGNLVGFAGAQDAALGGFC
TSAGFKSIPWSIYSAIFEIITALGNAEDDSGCTSGGASSLGEAILGDQFLKQQYVVFDRDNGIRG
TVPMTDYGNDVEYYGQVTIGTPGKSFNLNFDTGSSNLWVGSVQCQASGCKGGRDKFNPSDG
STFKATGYDASIGYGDGSASGVLGYDTVQVGGIDVTGGPQIQLAQRLGGGGFPGDNDGLLGL
GFDTLSITPQSNAFQDVSAQGKVIQPVFVVYLAASNISDGDFTMPGWIDNKYGGTLLNTNIDG
EGYWALNVTGATADSTYLGAIFQAILDTGTSLLILPDEAAVGNLVGFAGAQDAALGGFVIACT
SAGFKSIPWSIYSAIFEIITALGNA EDDSGCTSGIGASSLGEAILGDQFLKQQYVVFD RD
NGIRLAPV
```

The sequences above are examples of one type of amino acid; for more details, types of amino acids are explained in [12].

2.2. All sub-sequences are scanned and normalized as images in the same size. The total number of sub-images depends on the number of data collected and the window size.

2.3. Pe-processing is a significant step because sub-images require a noise removal filter [13][14]. Various methods are applied in noise removal from image processing[15]. A hybrid filter was utilized in our research.

2.4 Spatial information features: Each sub-image passes in Algorithm A, and Algorithm B. Amino acid sequences are manipulated with a different window size in sub-images. Spatial feature extraction constructs the relationship between pixels and stores the location of pixels. The research applied two algorithms in spatial information: A and B. The second algorithm, B, modifies algorithm A to a void drawback of creating many clusters. The spatial feature is one of the feature extractions considering the location pixels that represents a new approach in secondary prtein predication.

Algorithm A

Input: Amino acid sequence. Select the window size to scan the sub-image denoise the given sample.

```
            Using filter.
Output: Recognized objects and number of clusters
Begin
Step 1: Estimate the difference between the center pixels in the window and the neighborhood.
Step 2: Select the maximum difference value, and a new cluster is the point
Step 3: Check from the start point that the cluster was not created before letting this point belong to the same   cluster number.
Step 4 Keep the spatial location of pixels in the same region.
Step 5: Scanning by a window all images and repeat steps 1,2,3,4
 End
```

The above algorithm's main problem is that it computed many clusters. Sometimes, a few pixels did not represent a region but pointed to the new cluster, which led to many clusters. Therefore, we need to enhance algorithm B in applying the algorithm after algorithm A.

Algorithm B

```
Input: Input spatial location of pixels of sub-image in the   same clusters and numbers of cluster
Output: Optimal numbers of clusters.
Begin
Step 1: Extract spatial coordinates for all images with the same intensity value.
Step 2: Estimate the region's density by counting the number value.
Step 3: Find the histogram of the relation between cluster numbers and density.
Step 4: Eliminated all clusters of low histograms by Threshold.
End
```

3. Experiment and Results

After we collected the data on amino acid sequences, scanning all amino acid sequences divided the image into sub-images in digital form**.** The median filter is used for noise removal from images and passing windows with different lengths 7, 9,11, and 13**.** The difficulty happens in the choice of length window; when minimization window size, a problem occurs in the damaged information around the center pixel when the expanded window loses the specific information. Fig. (2) explains the window size 3 by 3 with neighbors.

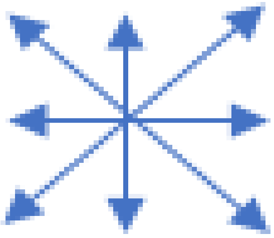

Figure 2. Window size of 3x3.

Several attempts were made to achieve the best results in the prediction states of (Helix, Sheet, and Coil) structure, prediction built on automated spatial features**.** The best accuracy results are obtained in the window 13. Helix equals 91.93% accuracy, 93.15% in Sheet (E), and Coil (C) equals 89.0126 %, as shown in Table 1.

Table 1. Accuracy of three types of secondary

Type	**Accuracy %**
E	93.15
H	91.93
C	89.0126

The number of data sets of amino acid sequences related to the classification and accuracy of helix(H), coil (C) strand (E) as shown in Fig.(3)

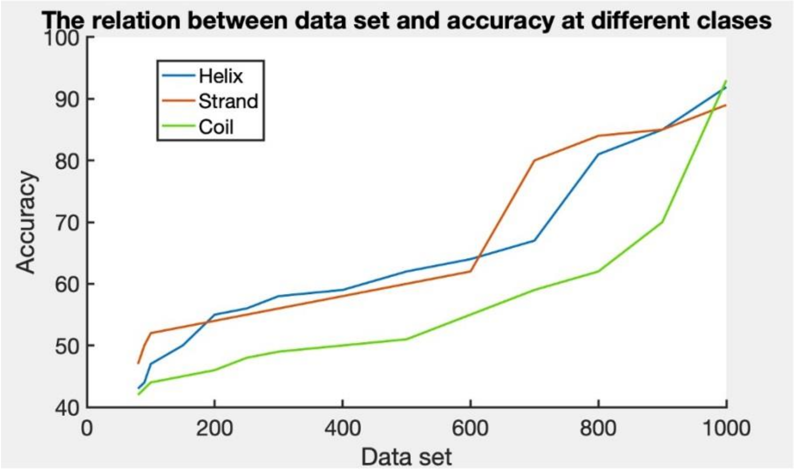

Figure 3. Accuracy of predication with Data set in helix(H), Coil(C) and Strand (E).

4. Conclusion

The major step in the protein structure is the prediction of the secondary protein structure because the tertiary structure function depends on the secondary protein structure. The main idea of the research focuses on spatial information extraction for the prediction of secondary protein structure. The sub-image of amino acid contains many redundant features that affect the performance prediction of the secondary structure. In this approach method are extracted spatial features in algorithm A but need enhancement to minimize the number of clusters by algorithm B. The important conclusions are as follows: -

1. Most methods in predication secondary structure used traditional techniques and orthogonal coding. The drawback of orthogonal coding is that it takes more time and manipulates useless amino acid sequence information. The amino acid sequence was sectioned in images, and spatial features were extracted, which are reconsidered as a new approach to prediction.

2. Faults occur in the research due to the size of the window, but the best accuracy of prediction secondary structure is related to selecting a suitable window length that comes from trial and error**.**

Spatial information provides automated feature extraction for the prediction of secondary protein structure.

3. The experiments proved the performance of spatial features in the predication secondary protein structure, but sometimes prediction fails in other types of protein .

References:

[1] M Younus Wani, NA Ganie, S Rani, S Mehraj, MR Mir, MF Baqual, et al. "Advances and applications of Bioinformatics in various fields of life". International Journal of Fauna and Biological Studies; 5(2): 03-10, 2018.

[2] Thomas Dandekar, Meik kunz. "Bioinformatics An introductory textbook". Springer:.93-

102 p2023.

[3] Chou and Fasman, "Prediction of protein Conformation", Biochemistry, Vol.13, No.2,1974.

[4] Jean Garnier, Jean-François Gibrat and Barry Robson, "GOR method for predicting protein secondary structure from amino acid sequence, Methods in Enzymology", vol. 266, p. 540–553, 1996.

[5] S. Haifeng, C. Guangsheng, W. Hairong, Y. Weiwei, "the improved (2D) 2PCA algorithm and its parallel implementation based on image block", Microprocess. p 170-177 2016.

[6] Muhammad Naim Abdullah› Mohd Afizi W&Mohd Shukran, " Features Extraction Techniques and Approaches for Content-Based Image Retrieval (CBIR) System" .p29-34 2021

[7] Şaban Öztürk, Bayram Akdemir. "Application of Feature Extraction and Classification Methods for Histopathological Image using GLCM, LBP, LBGLCM, GLRLM, and SFTA". International Conference on Computational Intelligence and Data Science 2018.

[8] Ning Qian and T. J. Sejnowski, "Predicting the secondary structure of globular proteins using neural network models", Journal of Molecular Biology, vol. 202, no. 4, p. 865–884, August 1988.

[9] Rahman Ema, Md. Nasim Adnan. "Protein Secondary Structure Prediction based on CNN and Machine Learning Algorithms". International Journal of Advanced Computer Science and Applications”, Vol. 13, No. 11, 2022 .

[10] John-Marc Chandonia and Martin Karplus, "Neural networks for secondary structure and structural class predictions, Protein Science", vol. 4, no. 2, p. 275–285, February 1995.

[11] Haifeng Sui, Wu Qu, Bingru Yan and LiJun Wang, "Improved Protein Secondary Structure Prediction Using a Intelligent HSVM Method with a New Encoding Scheme, "International Journal of Advancements in Computing Technology, " vol. 3, no. 3, pp. 239-250, April 2011.

[12] Dewi Pramudi Ismi, Reza Pulungan & Afiahayati, "Deep learning for protein secondary structure prediction: Pre and post AlphaFold ", Computational and structural biotechnology journal, vol. 20, p. 6721–6286, 2022.

[13] Muna Khalid Jasim, RehanHamdullah Najm, Emran Hassn Kanan, Hamza Esam Alfaar, Mohammed Otair. "Image Noise Removal Techniques: A Comparative Analysis International Journal of Science and Applied Information Technology; " November 2019 volume 8 , No. 6.

[14] Muhammad Aqeel Aslam, Muhammad Asif Munir, and Daxiang Cui1. "Noise Removal from Medical Images Using Hybrid Filters of Technique.” Journal of Physics: Conference Series (2020).

[15] C. M. Maheshan · H. Prasanna Kumar1, "Performance of image preprocessing filters for noise removal in transformer oil images at different temperatures, Research Article," Springer Nature Switzerland AG 2019.

أتمته التنبؤ ببنية البروتين الثانوي بناءً على المعلومات المكانية

الخلاصة **:** تتناول هذه الدراسة التنبؤ البنية الثانوية للبروتين. يعتمد التنبؤ للبنية الثانوية على صفات صور الأحماض الأمينية المسلسلة. تم تطوير وتحديد خوارزمية التنبؤ تعتمد على المعلومات المكانية للصور التسلسل الحامض الاميني. يتم تطبيق الخوارزمية A والخوارزمية B. ان خوارزمية Bتعتبر تطوير للخوارزمية A ولغرض استخراج واستخدام المعلومات المكانية لأنواع تصنيف الأحماض الأمينية في ثلاث حالات من البنية الثانوية للبروتين والحالات هي: حلزونية (H)، وخيط (E)، (C).يتم التصنيف حسب الصفات المستخلصة من الصور وتتم العملية من خلال مرور نوافذ بأحجام مختلفة ويتم حساب التباين الدرجات الرمادية من مركز النافذة ومناطق المجاورة. تتم أيضا من خلال هذه التقنية الى مواقع المكانية. تعتبرهذه الطريقه حديثة ومستخدمة لأول مرة في هذا المجال ونتائج اثبتت الدقة في التصنيف في حلات الثلاث للبنية البروتين. تعتمد دقة التصنيف على خوارزمية التحسين B ومعلمات الخوارزمية A و B. نحن نطبق نهجًا جديدًا لاستخراج المعلومات المكانية من بنية البروتين الأساسي لتنبيه ثلاث حالات من بنية البروتين الثانوية. يتم تقسيم تسلسلات البروتين الأساسي إلى صور فرعية وتتم العملية طبقا الى التقطيع الذي يحدث في تسل الحامض الاميني فتتحول السلسة الحامض الاميني الى صور . ويتم معالجه الأولية للصور لتحسين الصور وتهياه الصور للعملية التصنيف. بعدالمعالجة الأولية يتم تطبيق الخوارزميات A. خوارزمية A غير كافية للحصول على التصنيف المناسب ولذلك يتطلب تطبيق خوارزمية B لتحقيق الدقة في كل نوع من بنية البروتين الثانوية (Helix H) بنسبة 91.93%، والورقة (E) بنسبة 93.15% والملف (C) بنسبة 89.0126%. ان النتائج البحث اضافت تحسن من الناحية الدقة في التصنيف ولا تحتاج الى قاعدة بيانات كبيرة مثل بقية الطرق المستخدمة في الشبكات العصبية.

الكلمات المفتاحية: الأحماض الأمينية ، معالجة صور، التنبؤ البنية الثانوية للبروتين.