ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/358063616

Prediction Secondary Protein Structure from Images of Amino Acid by Using Harr – like Features in Support Vector Machine

Article *in* Webology · January 2022

DOI: 10.14704/WEB/V19I1/WEB19041

CITATIONS
0

READS
28

1 author:

Nahla Ibraheem Jabbar
University of Babylon
13 PUBLICATIONS **74** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Fuzzy C mean clustering View project

All content following this page was uploaded by Nahla Ibraheem Jabbar on 09 May 2022.

The user has requested enhancement of the downloaded file.

Prediction Secondary Protein Structure from Images of Amino Acid by Using Harr - like Features in Support Vector Machine

Nahla Ibraheem Jabbar
Department of Chemical Engineering, University of Babylon, Iraq.
E-mail: eng.nahla.ibraheem@uobabylon.edu.iq

Received August 12, 2021; Accepted November 27, 2021
ISSN: 1735-188X
DOI: 10.14704/WEB/V19I1/WEB19041

Abstract

In this study Haar features are extracted from images of sub-sequences of amino acid and classified by Support Vector Machine (SVM). We apply a novel approach of integration Haar-like features extraction from primary protein structure to predication three states of secondary protein structure. The sequences of primary protein are divided into different slides windows for representation images and then Haar-like feature have been extracted from these images to classify three-category of secondary protein structure helix (H), strand (E) and coil (C). The final prediction results were generated from SVM overall per residue accuracies are: - accuracy of helix(H) reached 83.93%, accuracy of Sheet(E) is 85.15% and accuracy of the Coil (C) is 81.0126 %. Images are scanned from amino acid sequences are specified by the selection window sizes, when the size of window is small the important information of predicting secondary structure relay outside the window. It has taken only local sequence information. When the size of window has been increased the performance has been deteriorated. Haar-like gives a perfect input data of SVM with a huge amount of data, also for improvement of support vector machine are used varying cost parameters.

Keywords

Haar-like Features, Predication Secondary Structures, SVM, Image Processing.

Introduction

Bioinformatic has a topic research in the fields of computer science and information technology. Bioinformatics consider a biological data of database to store biological information for the creation and maintenance information such as nucleotide and amino acid sequences (Shendure Jay *et al*, 2008) in all 20 standard amino acids are a variable in the protein. Proteins are the most abundant biological macromolecules are happen in all types of cells. Protein structure divided in four types of structures:-1-primary protein structure 2-secondary structure 3-tertiary structure 4- quaternary structure. The primary

structure also called a amino acid sequence , secondary structure means an arrangements of adjacent residues coming from hydrogen bonding between backbone groups that the two most common are helices and pleated sheets .The tertiary is a complete three dimensional structure, this structure refers to the spatial information in all amino acid residue in a polypeptide chain. The quaternary structure represented spatial arrangements information and nature of protein (Anil Mandel *et al*, 2012). Primary structure or linear amino acid sequence are used for predicting secondary structure. The important application of secondary proteins structure relies in the fields of drug design and novel enzymes. Many approaches are applied in the prediction of secondary structure protein from primary structure protein in the 1974. It used a single residue statistics to achieved about 50% prediction accuracy (Chou and Fasman, 1974). Statistical Predication represented all methods in statistical analysis the common statistical method is a GOR method **(**Jean Garnier *et al*, 1996) The GOR method implemented conditional probability of immediate neighbours in the amino acid are formed a secondary structure Most methods of secondary are prediction are classified in the Neural Networks. A the first time Qian and Terrence are introduced neural networks for prediction (Ning Qian *et al*, 1988). They implemented with 17 input and three output in 40 hidden. The application of Neural Networks included a various architectures were studied by Chandonia et al targeting better prediction accuracy (John-Marc Chandonia *et al*, 1995). (Haifeng Sui *et al*, 2011). They are developed the accuracy prediction of protein secondary structure by applied hybrid SVM. This approach which gives accuracy better than other types of prediction. In 2015 (Reet Kaur *et al*, 2010) they introduced six types binary classifier by SVM. The results showed the prediction accuracies are for $H/{\sim}H, E/{\sim}E, C/{\sim}C, H/E, E/C$ and H/C are 66.25%, 72.28%, 62.58%, 65.56% and 70.85%. Haar-like features are used in many of applications for object detection. Encoding an input image by Integration of Haar-Like feature and Histogram of Oriented Gradient that obtain vector of visual descriptors (Negri P *et a*l, 2007). The combination Haar- Like features and ROI (Region of Interest) detected in the image car (Stanciulescu B *et al*, 2007; Haselhoff A *et al*, 2007) also other combing have been used in image classification are based in Haar-Like features extraction and SVM (Natalia Larios *et al*, 2010). The proposed method is a predication secondary protein structure from sequences of amino acid by scan image to the sub sequences amnio acid and extraction Haar-like features from these images. We capture basic features of object features that represented primary protein structure that helpful in the predication and used grey-scale to compute the differences between rectangle to extract harr-like feature. A lot of combination features are applied as input data to classify by SVM for recognition three types of secondary protein structure. Finally, the current research trends a new approach in the prediction a secondary protein structure by applying Haar-like features from images of amino acid. Haar-like

feature a good features selection in the Support Vector Machines for supporting huge data input. The results of SVM gives more progress in the accuracies with other methods of predication.

Haar-like Features

Manipulation details of pixel information in digital image processing are need more time for description image, also the intensity values cannot completely represented the structure of image. Haar-like features have been extracted structural information from text images. Viola & Jones produced expanding in the set of basic Harr features. They added a new types of Haar feature in Figure. 1 (Viola. P *et al*, 2001).

Tables and Figures are presented center, as shown in Table 1 and Figure 1, and cited in the manuscript and should appeared before it.

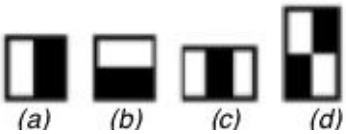

Figure 1 Shows Harr-like features

These rectangles are similar shape, size and horizontal or vertical (Fig. 1). The computation features in two rectangle a and b are differenced sum of variance values between us. The features values of pairs rectangles a and d are computed variance values between diagonal pairs. Features values in c are calculate in the variance outsides in two rectangles subtracted from variance summation of centre rectangles.

Integral Image

The integral image is a recent description pixel (x,y) in the image ,it could be computed by summation values pixels upper and the left of (x,y), as shown in Eq,(1) and Figure.2.:

$$p(x,y) = \sum_{x` \le x, y \le y`}^{x,y} i(x`, y`) \quad (1)$$

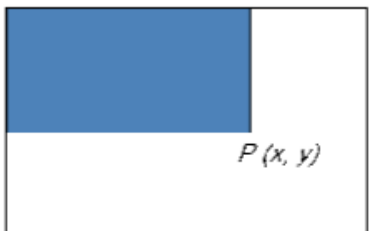

Figure 2 Shown point (x,y) in the integral image

The following representation, the location of the pixel in original image is $i(x,y)$ and $s(x,y)$ is the cumulative column sum. We can calculate the integral image representation of the image using the Eq. (2) and Eq. (3).

$$s(x,y) = s(x,y-1) + i(x,y) \quad (2)$$
$$ii(x,y) = ii(x-1,y) + s(x,y) \quad (3)$$

The process as shown in Figure. 3

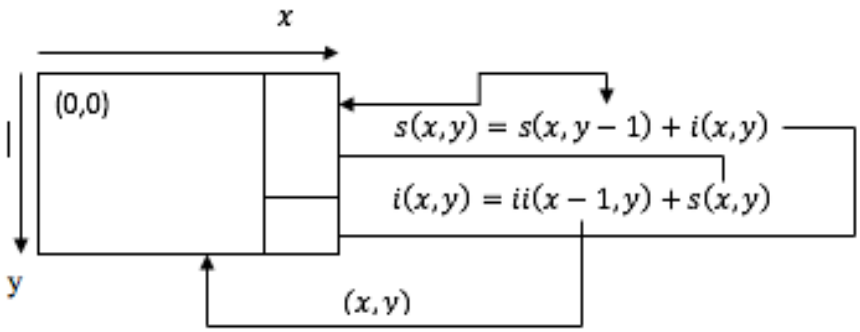

Figure 3 The integral Image Calculation Process

Integral image representation can compute the value of any rectangular sum in a different position or scale. For example an integral features are using only 4 lookups sum inside rectangle D as shown in the Figure. 4, this figures can be computed using Eq (4).

$$P_1 = A, P_2 = A + B, P_3 = A + C, P_4 = A + B + C + D \,, P_1 + P_4 - P_2 - P_3 = A + A + B + C + D - A - B - A - C = D \quad (4)$$

Or by Equation

$$ii(4) + ii(1) - ii(2) - ii(3) \quad (5)$$

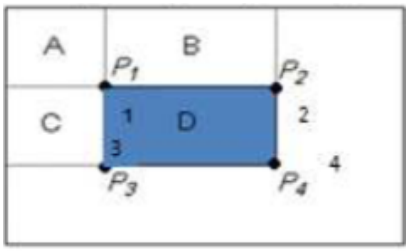

Figure 4 An explanation on how to find the sum of d by using 4 references

These result allow computing two, three and four-rectangular features with 6, 8 and 9 array references.

Support Vector Machine (SVM)

Support vector machine is used in many applications like image classification, face detection, recognition of handwriting and bioinformatics. A binary SVM computes two hyperplanes, the first hyperplane used as the boundary of positive samples and the second in the boundary of negative. The optimization hyperplanes of positive samples and negative samples that relay in minimization computing error and maximization distance between us. All the features in positive and negative samples are combines in dimensional features vectors. The ttrining SVM denote as$\{(x_i, y_i)|x_i \in R^P, 1 \leqq i \leqq m\}$ where x_i represented all features and $y_i \in \{-1,1\}$ is the label of positive sample or negative sample. The object a hyperplane as $w.x - b = 0$, positive samples in the one side of the hyperplane, while the maximized distance from the hyperplane to the close samples. In the Figure. 5 as shown, the estimation of optimization w and b, two hyperplanes $w.x - b = 1$ and $w.x - b = -1$. There is $w.x - b \geq 1$ for positive samples and $w.x - b \leq -1$ for negative samples if the samples are linear separable can be correctly classified while producing maximized distance to the hyperplane of each other from the geometrical computation, the distance between the two hyperplane boundaries is calculated by $2/\| w \|$.

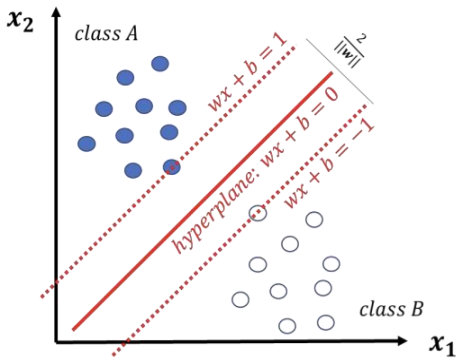

Figure 5 Two classes are separate by SVM

Thus the following task is to minimize the $\| w \|$.The minimization relies in $\| w \|^2/2$.

$$arg\ max \left\{ \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j}^{0,c} \alpha_i \alpha_j y_i\, y_j k(x_i, x_j) \right\} \qquad (6)$$
$$0 \leq \alpha_i \leq c$$

Where α_i is the Lagrange multipliers. K(.) is the a Kernel.

Methodology and Results

For getting a desired results, This work divided into more steps, Starting from collected the data reaching for predication structure by classifier, the flowing diagram explain the outline of the work Figure. 6.

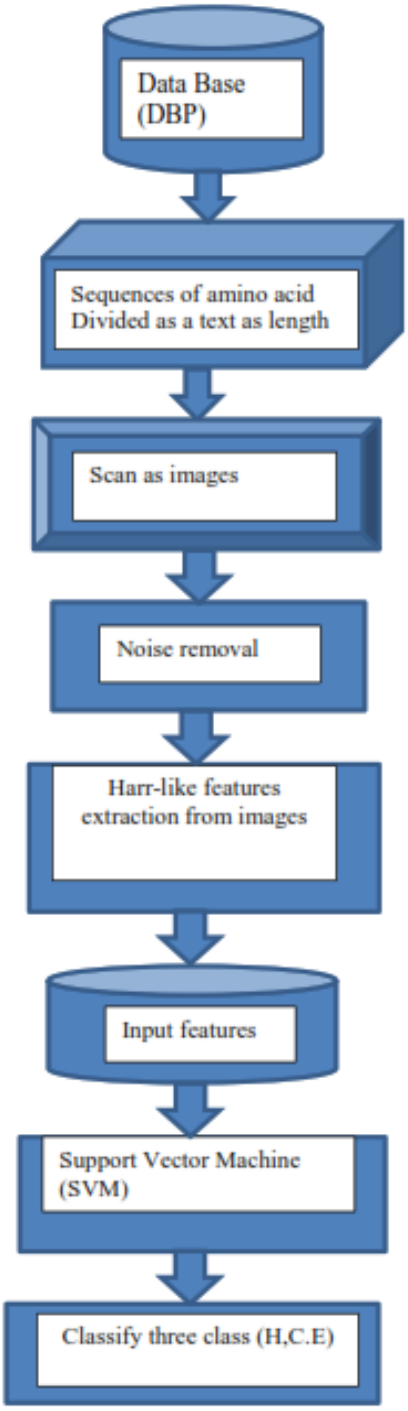

Figure 6 Illustrates the flow diagram of the Work

1. Data Base

There are twenty type of amino acids that can available in the protein (Lorenza Bordoli 2007). The amino acids and their letter codes are given in Table. 1.

Table 1 Explain type of amino acids...

Amino acids	name	Variable
Glycine	Gly	G
Alanine	Ala	A
Serine	Ser	S
Threonine	Thr	T
Cysteine	Cys	C
Valine	Val	V
Isoleucine	Ile	I
Leucine	Leu	L
Prolin	Pro	P
Phenylalanine	Phe	F
Tyrosine	Try	Y
Methionine	Mer	M
Tryptophan	Trp	W
Asparagine	Asn	N
Glutamine	Gln	Q
Histidine	His	D
Aspartic Acid	Asp	D
Glumtamic Acid	Glu	E
Lysine	Lys	K
Arginine	Arg	R

The data are collected from The Protein Data Bank (PDB). Examples of protein structure databases include (in alphabetical order). This method includes on 62 proteins in the database. The sequence of amino acid of protein like these examples

Protein name

>Acprotease

Sequences

GVGTVPMTDYGNDVEYYGQVTIGTPGKSFNLNFDTGSSNLWVGSVQCQASGCK
GGRDKFNPSDGSTFKATGYDASIGYGDGSASGVLGYDTVQVGGIDVTGGPQIQL
AQRLGGGGFPGDNDGLLGLGFDTLSITPQSSTNAFQDVSAQGKVIQPVFVVYLAA
SNISDGDFTMPGWIDNKYGGTLLNTNIDAGEGYWALNVTGATADSTYLGAIFQAI
LDTGTSLLILPDEAAVGNLVGFAGAQDAALGGFVIACTSAGFKSIPWSIYSAIFEIIT
ALGNAEDDSGCTSGIGASSLGEAILGDQFLKQQYVVFDRDNGIRLAPVA
GVGTVPMTDYGNDVEYYGQVTIGTPGKSFNLNFDTGSSNLWVGSVQCQASGCK
GGRDKFNPSDGSTFKATGYDASIGYGDGSASGVLGYDTVQVGGIDVTGGPQIQL
AQRLGGGGFPGDNDGLLGLGFDTLSITPQSSTNAFQDVSAQGKVIQPVFVVYLAA
SNISDGDFTMPGWIDNKYGGTLLNTNIDAGEGYWALNVTGATADSTYLGAIFQAI
LDTGTSLLILPDEAAVGNLVGFAGAQDAALGGFVIACTSAGFKSIPWSIYSAIFEIIT
ALGNAEDDSGCTSGIGASSLGEAILGDQFLKQQYVVFDRDNGIRLAPVA

>Avain polypeptide

GPSQPTYGDDAPVEDLIRFUDNLQQYLNLQYLNVVTRHRY

2. Sequences of amino acid are divided into number of sub-sequences with window length n these are represented slides through sub-sequences. The decision length of window not easy because a short length is chosen that not given more details about structure (local structural features). Through experiments are used different sizes of windows 9,11 and 13. This example shows shifting window with length 7, assume the following sequences of amino acid is:

KLNTDETGACPQACYA

The first window 'TDEPGAC' is the first sub-sequence to scan an image for the residue 'P' $TDEPGAC$, the next window residue 'G', The window moves to 'DEPGAP'.

$DEPGAP$. The window moving to reach in the last group 'CPQACYA' is reach to central residue 'A'$CPQACYA$

3. All sub-sequences are scanned as images and all these images are normalized in the same size. Hence, the total number of the training images is 820 images, while the number of testing images is 770 images, applying the SVM classifier
4. Before the step extraction Haar -like features, it is necessary to pre-processing images by applying noise removal filter.
5. Haar-like features: The manipulation details of pixel information of digital image processing are need more time for description, also intensity values cannot completely represented structure of image for these reasons. We needs efficient features extraction from amino acid sequences images like structural information. Haar-likes structural features are a good idea and a new approach in the bioinformatic. We calculated a Harr feature value from images by passing a slide window like a rectangular region of the image. A window can be defined by two points: the top-left and bottom-right corners. The resize the window in (24, 24) pixels from this size a large number of Haar_ like features are extracted. The number of Haar-like related on the window size also the location of window are affected in generation features. In this methods are used a different types of features. Figure .7(1), (2) and (3) illustrates the feature types. Table.1 describes the number of features in each system

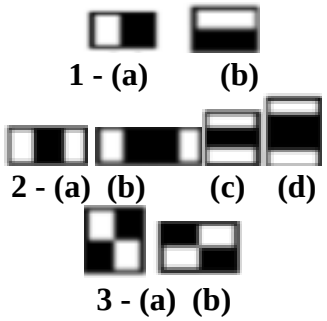

Figure 6 (1,2&3) Explain number of feature

Table 2 Explain numbers and type of features

Features type	Numbers
1a,1b,2a,2c	151600
1a,1b,2a,3a	144940
1a,1b,2a,2c,3a	172520
1a,1b,2a,2c,3a,3b	193910
1a,1b,2a,2c,3a,3b.2b	212999
1a,1b,2a,2c.3a.3b.2b.2d	232734

6. Process of training and classifier

In this work, we use SVM to classify the structural state (Helix, Sheet and Coil) of each residue in a given protein, based on the Haar -features patterns observed during a training phase. In the following approach the results of SVM in the prediction secondary protein structure are given a different accuracies. The optimal accuracy have obtained in the size window 13, The accuracy has been estimated successfully in the Helix which gives of 83.93%. The Second class represented accuracy 85.15% in the Sheet(E) and the last accuracy 81.0126 % in Coil (c). The results are shown on the Table 3.

Table 3 H,E and C with accuracy

Quality index	% Accuracy
H	83.93
E	85.15
C	81.0126

The value of parameter α is fixed in SVM is equal to 0.1 where the cost parameter C has a varying value during working of SVM. A value range of C from 0.1 - 0.6 which leading increasing the efficiency results of SVM. Finally, The number of Haar-like features affected in the accuracy of classification of three states as shown in Figure.8

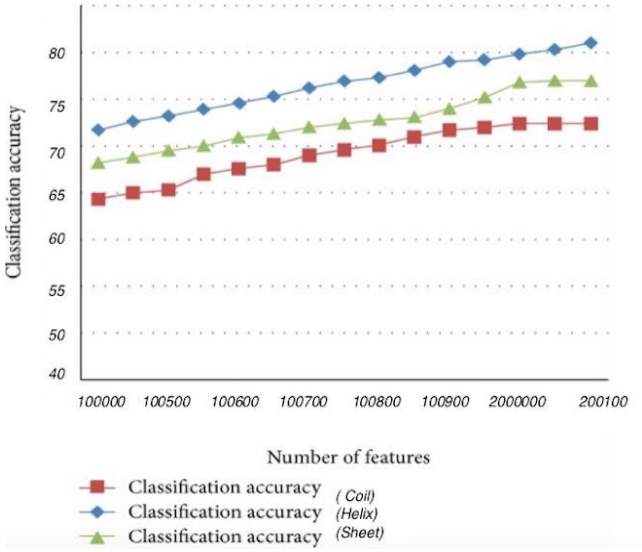

Figure 8 Shows relation number of features and range accuracy

Conclusion

Protein structure prediction is critical major step in the predicting tertiary structure and functions of protein. The important purpose of a paper is a comparison efficiency of Support Vector Machines in the predicting secondary structure with Harr features are extracted from images of amino acid protein sequences. We derived from this work, the following conclusions.

1. Most of researches have been used orthogonal coding for input letters of amino acid sequences. In this paper a new approach are applied Haar-like features for input coding of sequences amino acid which gives more structure details of the sequence that useful in prediction secondary protein structure also it is intelligent way in the dealing images of sub-sequences.
2. Accuracy and time of predication secondary structure by SVM are depends in many factors likes choosing an appropriate window length sliding of sequence and the window type of Harr-like features.
3. The experiments shows Harr-like features in expanding a mount volume training data enhanced the performance considerably.

References

Mandle, A.K., Jain, P., & Shrivastava, S.K. (2012). Protein structure prediction using support vector machine. *International Journal on Soft Computing, 3*(1), 67-78.

Chou, P.Y., & Fasman, G.D. (1974). Prediction of protein conformation. *Biochemistry, 13*(2), 222-245.

Sui, H., Qu, W., Yan, B., & Wang, L. (2011). Improved protein secondary structure prediction using an intelligent HSVM method with a new encoding scheme. *International Journal of Advancements in Computing Technology, 3*(3), 239-250.

Haselhoff, A., Kummert, A., & Schneider, G. (2007). Radar-vision fusion with an application to car-following using an improved adaboost detection algorithm. *In IEEE Intelligent Transportation Systems Conference,* 854-858.

Garnier, J., Gibrat, J.F., & Robson, B. (1996). GOR method for predicting protein secondary structure from amino acid sequence. *Methods in enzymology, 266,* 540-553.

Chandonia, J. M., & Karplus, M. (1995). Neural networks for secondary structure and structural class predictions. Protein Science, 4(2), 275-285.

Bordoli, L. (2007). Protein Structure Bioinformatics Introduction Secondary Structure & Protein Disorder Prediction. *EMBnet course Lausanne*.

Larios, N., Soran, B., Shapiro, L.G., Martínez-Muñoz, G., Lin, J., & Dietterich, T.G. (2010). Haar random forest features and SVM spatial matching kernel for stonefly species identification. *20th International Conference on Pattern Recognition,* 2624-2627.

Negri, P., Clady, X., & Prevost, L. (2007). Benchmarking haar and histograms of oriented gradients features applied to vehicle detection. *In ICINCO-RA*, *1*, 359-364.

Qian, N., & Sejnowski, T.J. (1988). Predicting the secondary structure of globular proteins using neural network models. *Journal of molecular biology*, *202*(4), 865-884.

Kaur, R., Kaur, M., & Kaur, A. (2010). Using Cluster Analysis for Protein Secondary Structure. *International Journal of Computer Application*, *4*(12).

Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nature biotechnology*, *26*(10), 1135-1145.

Stanciulescu. B., Breheret, A., & Moutarde, F. (2007). Introducing New AdaBoost Features for Real-Time Vehicle Detection. *Proceedings of COGIS'07 conference on Cognitive systems with Interactive Sensors, held in Stanford University California,* Nov 2007.

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. *In Proceedings of the IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, *1*, I-I.

View publication stats