VOL.4 NO.2 September 2015 - ISSN 2233 – 1859

UOIuBIH
ORSinBIH
Operations Research Society in Bosnia and Herzegovina

Southeast Europe Journal of Soft Computing

Available online: http://scjournal.ius.edu.ba

IUS Soft Computing Research Group

Protein Secondary Structure Prediction by Using PSSM Pseudo Digital Image of Proteins

Faruk B.Akcesme, Mehmet Can
International University of Sarajevo, Faculty of Engineering and Natural Sciences, HrasnickaCesta 15, Ilidža 71210 Sarajevo, Bosnia and Herzegovina
fakcesme@ius.edu.ba; mcan@ius.edu.ba

Article Info

Article history:
Article received on 17 Jul. 2015
Received in revised form 17 Aug.2015

Keywords:
Protein Secondary Structure Prediction; PSSM, Digital Image Features

Abstract

Protein secondary structure prediction is one of the hot topics of bioinformatics and computational biology. In this article we present a new method to predict secondary structure of proteins. PSSMs of proteins are used to generate pseudo image of proteins. These protein images are used to extract digital image features. Digital image features vectors used for similarity analysis. We believe that PSSM pseudo digital images of proteins could help us to represent protein global intrinsic information in order find globally similar proteins and use these similar proteins during prediction. Highest prediction accuracy for Q3 recorded as 72.1% by using the system. Beside the high accuracy, this method allows us to shorten computational time for predicting secondary structure of proteins.

1. INTRODUCTION

Sequencing technologies are improved and the numbers of protein sequences are increased tremendously in protein databases. Conformational analysis and structural determination of proteins are very costly and time consuming. Biological information carried by amino acid sequences of protein can be analyzed through its structure. This fact increases the need for accurate and reliable methods to predict protein structure. Many attempts have been made to predict secondary structure of protein in the absence of suitable homologous sequences.

Since Anfisen`stheory [1], it is believed that conformation of proteins is determined predominantly by their amino acid sequence. Many efforts have been devoted to predict secondary structure of protein from its amino acid sequences up to date. Prediction of secondary structure has yielded many interesting and promising results. These result increase attraction in predicting secondary structure of proteins from amino acid sequence. Besides knowing biological function of proteins, secondary structure prediction enable to design novel proteins, predicting the effect of point mutations, identifying protein structural classes and predicting epitopes etc.

It is very reliable method to predict secondary structure of protein by comparative modeling with its homologous sequences but in the other hand, it is only possible when very high degree sequence similarity found to the target protein [2].

There are many different ways to describe accuracy of prediction. One of the parameter is the number of amino acid sequences used as a dataset. Way of estimating accuracy is very crucial in order to record fair score. Chou and Fasman`s study on prediction of protein conformation in 1974 which considered as one of the landmark score as 77% accuracy in 1974 coils [3]. Other very famous methods developed by Garnier and his friends recorded the accuracy as 63% [4].

Kabsch and Sander consider more sequences than the dataset used by Chou Fasman and Garnier and found that the overall accuracy of methods were 50% for Chou and

Fasman, 56% for Garnier [5].It is very obvious that the high prediction scores obtained by Chou Fasman and Garnier were highly related to the dataset similarity that they used.

Through multiple alignments secondary structure prediction accuracy was significantly boosted [6].

Most of the prediction methods are designed based on the sequentially local interaction even though formation of secondary structure is rely on more complex chemical interactions[7]. It is very important to consider the possibilities of including long-range interactions when designing a method for protein secondary structure prediction.

Since GOR method, one of the first successful and popular of the secondary structure prediction schemes, scientist uses windows sliding methods to decide about central amino acid conformation. The GOR method analyzes sequences to predict alpha helix, beta sheet, or random coil secondary structure at each position based on 17-amino-acid sequence windows. Method includes four scoring matrices of size 17×20, where the columns correspond to the log-odds score. Scores represent the probability of finding a given amino acid at each position in the 17-residue sequence. The central amino acid, the ninth one, secondary structure conformation reflected by these four matrices probability being in a helical, sheet, coil conformation.

The other method for protein secondary structure prediction is using PSSMs. In many different studies, it is recorded that PSSM base representations that utilize multiple alignment and information about protein families yield highest accuracies in secondary structure [8]. Since the conservation is usually indicative of the formation of repetitive motifs such as the secondary structures, this information was found useful in prediction of proteins. In PSSM each amino acid residues are represented by 20 numbers. These numbers are possible amino acid substitutions that reflect frequencies of substitutions observed at this position in a protein family. The positive numbers indicated that given amino acids substitution occurs more frequently than expect by chance where the negative numbers represent the reverse [9]. The resulting PSSMs are generally generated iteratively. In many applications 3 iterations was used.

Both methods predict the secondary structure based on linear locality, these methods cannot use information about long-range interaction of protein sequences. In order to include global similarity of the proteins rather than linear locality we propose different similarity measure. We created digital image of PSSMs than used sliding window approach to assign secondary structure confirmation of proteins. We believe that generated image features represent long range interaction of proteins. We propose a method to test the hypothesis we propose.

2. METHOD

Method is described into 5 following steps:

1. Generating PSSM for each proteins
2. Creating digital image from PSSM
3. Extracting features from images and representing the proteins with feature vectors
4. Finding most similar feature vectors with the query
5. Sliding windows onto similar proteins and assigning secondary structure

1. Generating PSSM

PSSM has found better alternative to consensus sequence to reveal more reliable similarity. Consensus sequences had previously been used to represent patterns in biological sequences, but had difficulties in the prediction of new occurrences of these patterns. First, a database containing all known sequences (or non-redundant database) is selected. Then, low complexity regions are removed from the nr database. Finally, PSSM profiles are generated by PSIBLAST (Position Specific Iterative-Basic Local Alignment Search Tool) program [10] for each of the proteins with three iterations for each sequence in 25PDB dataset [11]. Here, multiple sequence alignment (MSA) and BLOSUM62 matrix [12] are used in generating PSSMs.

A PSSM profile has L×20 elements, where L is the length of a query sequence. Protein sequences represented by PSSMs then converted into the images. The main purpose of converting PSSMs into the images is to generate features vectors that will be later used for finding similar proteins. Most similar 50 images are selected and similarity analysis has been performed with using these 50 similar proteins for assigning secondary structure to each query.

The other benefits of using images reduce the number of target proteins into data since slicing windows in all vectors (X numbers for each residue) is very time consuming.

- Generating Digital Image from position specific scoring matrixes (PSSM)

Generated PSSM profile (L×20 elements, where L is the length of a query sequence) converted into images. Protein`s sequences are now represented by these images.

Since the input data to (number of images as protein`s) is too large to be processed, it is transformed into a reduced set of features (features vector). We believe features vectors that are composed of 41 elements contain information describing protein important characteristics.

Similarity search through entire protein are performed with using these vectors. It allows us to run similarity analyses in shorter time. It is believed that these feature vectors derived from images are representing protein globally.

- Sliding Window for protein structure prediction

To improve similarity search analysis and due to the limited number of structure samples we use a sliding window for predicting the secondary structure of proteins. Since GOR method.Sliding windows become very popular in the secondary structure prediction and various studies have shown the benefits of the method. This method uses the window as a query rather than the whole sequence. The number of windows is the number of residue of proteins. It is believed that given characteristic of the central AA is determined not only by the AA itself but by the adjacent AAs. Sliding window method use adjacent AAs information since the similarity analyzed within short windows that are 17 AA long. In the other hand optimal sliding window size for protein structure is discussable.

When predicting or analyzing some characteristics of an amino acid Ai, researchers relatively often use a window that is centered at Ai. In other words, a segment composed of Ai-X, Ai-X+1…Ai-1, Ai, Ai+1,…Ai+X, (where X determines the length of window). AAs is used since a given characteristic of the central AA is determined not only by the AA itself, but also by the adjacent AAs.

Generating pseudo digital images of proteins and representing protein by feature vector for similarity search

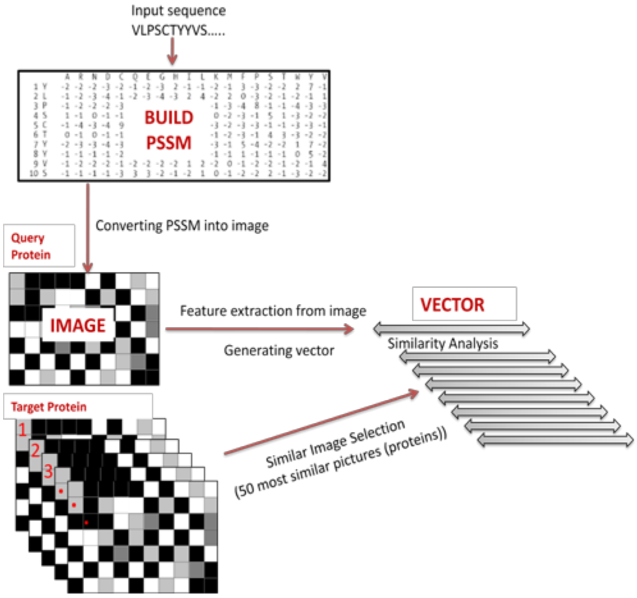


Figure: Systematic representation of generating pseudo digital images of proteins and representing protein by vector for similarity search

5 different strategies have been performed for observing the efficiency of using PSSM pseudo images for selecting similar proteins and representing proteins by features vectors in prediction accuracy.

4. RESULTS AND DISCUSSION

Protein secondary structure prediction by using PSSM pseudo digital images of proteins(Sliding windows onto primary sequence of proteins)

Most similar images to the query images are selected based in feature vectors hamming distance similarity. We tried to reach optimum number of similar images in order to obtain highest prediction score. 150 similar images give the highest score. Windows are slid through the query sequence. Four and more exact matches between two strings of windows are set up as a threshold. Similar strings that meet the criteria are collected as a decision maker for predicting the conformation of query residue. Prediction accuracy calculated in percentage as described in the literature

Prediction accuracy in alpha class reached to 72.1% in average for alpha class in 25PDB.

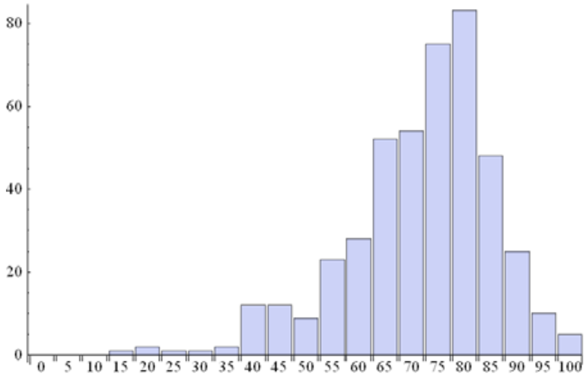


Figure: Prediction accuracy for 25PDB all alpha class is given in horizontal axis and number of proteins is in vertical axis

This is the highest result obtained by the method presented in this report. No previous result are recorder higher than the method which uses the image features to reduce search space, as discussed above we believe that proteins global sequence are represented in the images. This would be the other reason of higher result.

It is important to point here that this method is using different similarity measure then sequence similarity during the similar protein selection then sequence similarity is used to assign secondary structure conformation to each residue in query.

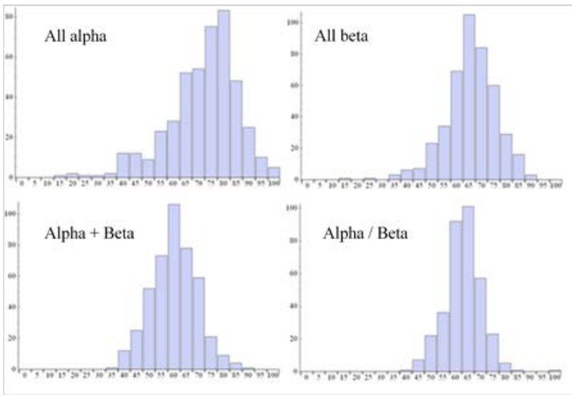


Figure13: Prediction accuracy is given in horizontal axis and number of proteins is in vertical axis in25PDB

Class	Prediction Accuracy
All alpha class	**72.1**
All beta	**65.8**
Alpha + Beta	**60.2**
Alpha / Beta	**63**
Overall	**65.2**

Table5: Prediction accuracies in 25 PDB proteins (protein represented by amino acids)

Protein secondary structure prediction by using PSSM pseudo digital images of proteins (Sliding windows onto PSSMs of the proteins)

After generating PSSM pseudo digital images of protein we propose alternative way to represent our proteins. Rather than using amino acid sequence of proteins we use PSSMs of the proteins. Hamming distance is used as a similarity measure and windows are slid onto PSSMs of the proteins. Efficiency of the methods is shown in the following figures and tables.

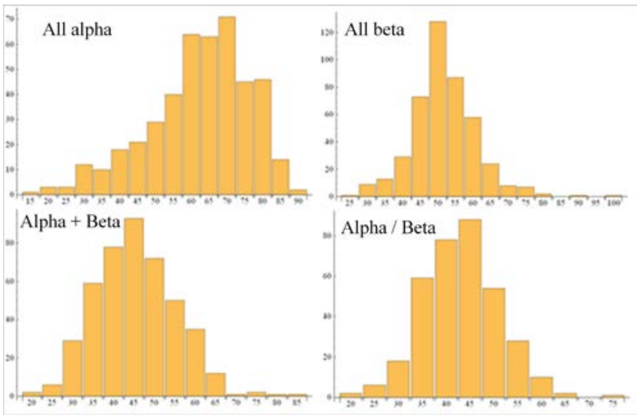


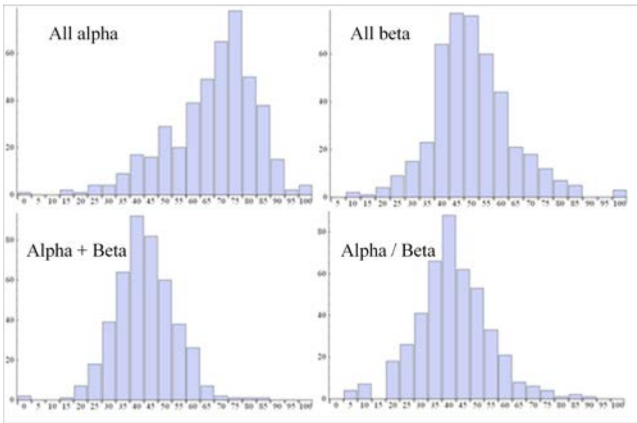
Figure14: Prediction accuracy is given in horizontal axis and number of proteins is in vertical axis in25PDB ((proteins are represented by PSSMs))

Class	Prediction Accuracy
All alpha class	**62.1**
All beta	**51.9**
Alpha + Beta	**45.5**
Alpha / Beta	**43.07**
Overall	**50.6**

Table6: Prediction accuracies in 25 PDB, (proteins are represented by PSSMs)

Protein secondary structure prediction by using PSSM pseudo digital images of proteins (Sliding windows onto physicochemical properties of proteins)

One another way to represent proteins after obtaining similar images (proteins) through the PSSM pseudo digital images of proteins is using physicochemical properties of the proteins. Each residue of the protein sequence is represented by three physicochemical features; net charge, Hydrophobicity, and side chain mass. Prediction accuracy is shown in the following graphs and tables.



Prediction accuracy is given in horizontal axis and number of proteins is in vertical axis in25PDB ((proteins are represented by physicochemical properties.

Class	Prediction Accuracy
All alpha class	**66.9**
All beta	**50.2**
Alpha + Beta	**42.7**
Alpha / Beta	**41.3**
Overall	**50.2**

Table7: Prediction accuracies in 25 PDB, (proteins are represented by physicochemical properties)

Protein secondary structure prediction by using PSSM pseudo digital images of proteins (Using BLOSUM matrix as a similarity measure)

We use PSSM pseudo digital images of proteins to select most similar proteins in the method described above. We use hamming distance measure for assigning secondary structure conformations. (In progress report I other similarity measure also discussed). We consider BLOSUM matrix as an alternative similarity measure since it is believed that BLOSUM as a substitution matrix biologically meaningful information beside quantitative similarity. The scores don`t give better results but using substitution matrix will be analyzed more in order to reveal

Class	Prediction Accuracy
All alpha class	**67.1**
All beta	**64.6**
Alpha + Beta	**58.9**
Alpha / Beta	**59.8**
Overall	**62.6**

Table8: Prediction accuracies in 25 PDB with BLOSUM62 as a similarity measure, (proteins are represented by polypeptide chain

Result and Discussion

We obtain highest prediction accuracy 72.1% (overall) by using the system described above. Beside overall accuracy presented methods predict very few proteins less then %50accuracy. This makes our method superior to other methods. This system works best with sliding windows directly on protein polypeptide chain. It is a new approach for predicting PSSMs pseudo digital images of proteins and our result are promising in two aspects:

A) PSSM pseudo digital images of proteins could help us to represent protein global intrinsic information in order find globally similar proteins and use these similar proteins during prediction.

B) It allows us to narrow search space. The limited computation time could be overcome with this strategy.

Using PSSMs of protein and digital image features in protein secondary structure prediction allow us higher prediction accuracy. We believe that our statistical analysis on primary-secondary structure relations on some specific proteins will allow us to modify our algorithms. Further studies are needed to analyze benefits of image processing features into proteins secondary structure prediction.

REFERENCES

1. Anfinsen, C. B., & Haber, E. (1961). Studies on the reduction and re-formation of protein disulfide bonds. *J BiolChem*, *236*(5), 1361-1363.
2. Zhang, Z. (2002). An Overview of Protein Structure Prediction: From Homology to Ab Initio. *Bioc218*, 1-10.
3. Chou, P. Y., &Fasman, G. D. (1978). Empirical predictions of protein conformation. *Annual review of biochemistry*, *47*(1), 251-276.
4. Garnier, J., Osguthorpe, D. J., & Robson, B. (1978). Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *Journal of molecular biology*, *120*(1), 97-120.
5. Kabsch, W., & Sander, C. (1983). How good are predictions of protein secondary structure?. *FEBS letters*, *155*(2), 179-182.\
6. Levin, J. M., Pascarella, S., Argos, P., &Garnier, J. (1993). Quantification of secondary structure prediction improvement using multiple alignments.*Protein Engineering*, *6*(8), 849-854.
7. Kihara, D. (2005). The effect of long-range interactions on the secondary structure formation of proteins. *Protein Science*, *14*(8), 1955-1963.
8. Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *Journal of molecular biology*, *292*(2), 195-202.
9. Ng, P. C., &Henikoff, S. (2001). Predicting deleterious amino acid substitutions. *Genome research*, *11*(5), 863-874.
10. Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., &Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*,*25*(17), 3389-3402.
11. Kurgan, L., Cios, K., & Chen, K. (2008). SCPRED: accurate prediction of protein structural class for sequences of twilight-zone similarity with predicting sequences. *BMC bioinformatics*, *9*(1), 1.
12. Henikoff, S., &Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, *89*(22), 10915-10919.