International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 – 5161

Available at http://inpressco.com/category/ijcet

Research Article

An Artificial Neural Network Classifier for the Prediction of Protein Structural Classes

Li Ningbo* **and Huo Hua**

†Laboratory of Intelligent Computing & Application Technology for Big Data, Henan University of Science & Technology, 263# Kaiyuan Road, Luolong District, Luoyang City, Henan, 471003, P.R. China

Accepted 08 May 2017, Available online 10 May 2017, **Vol.7, No.3 (June 2017)**

Abstract

As there are quite a few difficulties for us to predict a protein structural class directly from its primary sequence, the protein structural prediction based on the predicted secondary structure will undoubtedly be the first choice we would like to take. Protein structural classes are generally defined as four classes: α, β, α/β, $\alpha+\beta$. The protein secondary structure describes the local structural conformation of the polypeptide backbone, and it can be obtained fairly accurately from the primary sequence, all of these very features make the protein secondary prediction a critical way to predict the structural class. We constructed a more balanced PSIPRED (a neural network predictor with psi-blast, original method first proposed by Rost & Sander in 1994) algorithm to predict the protein secondary structure. Finally the features about Chaos Game Representation of the predicted secondary structure sequence were selected as the input of neural network classifier. As a result, the predictor has got an overall accuracy score of 71.2% on 40% identity dataset of astral on Structural Classification of Proteins database. Such situation proved that the predictor via secondary structure prediction is an effective approach to classify the structural classes.

Keywords: *protein structural classes, protein secondary structure, neural network, sequence analysis, balanced classifier, chaos game representation.*

1. Introduction

Proteins play a vital role in all living processes as material assumer of a range of functions to sustain lives(Rithvik & Rao, 2015). The structure of a protein determines the function it perform(Huang, Chen, & Lü, 2006). The protein structure is divided into four levels, namely primary, secondary, tertiary and quaternary structure (Huang *et al*., 2006; Koswatta, Samaraweera, & Sumanasinghe, 2011). The primary structure is the amino-acid residue sequence of a polypeptide chain. The secondary structure describes the local structural conformation of a polypeptide backbone. The tertiary structure (structural class) is defined into four categories (Hutchinson, Morris, & Thornton, 1976): all-α, all-β, α/β, α+β. Among them, all-α class is composed almost by the structure α, all-β almost by β, α/β and α+β are composed with both α and β. It is noteworthy that α/β class has many parallel strands, and α+β has strands anti-parallel. Another difference between them is that α/β include folds in which α-helices and β-strands that are largely segregated, while class α+β has these two secondary structures interspersed (L. Kurgan, Cios, & Chen, 2008).

The structural class of a protein is largely determined by its primary sequence(C. Chen, Chen, Zou, & Cai, 2008), however, we can hardly obtain a satisfied result directly from it. Fortunately, the folding conformation is as conservative as secondary structure, even more so.

This is due to the degenerate nature of the sequence-structure relationship (K. C. Chou & Zhang, 1995). Thus, we can classify the structural classes based on the prediction of protein secondary structure. The early methods who achieved a good performance are based on statistical algorithm. The least hamming distance method (P. Y. Chou, 1989) and the least Euclidean distance method (Nakashima, Nishikawa, & Ooi, 1986) are based the amino acid composition. Besides the widely used analytical methods, many machine learning algorithms have been proposed to classify structural classes. Such as neural networks (NN), support vector machines (SVM), hidden Markov model (HMM), and so on. One of the critical factors that influence the classifier is the selection of features. Among the many algorithms that used the secondary structure as the input feature, some are different in classifiers, and some are different in the ways of extracting features. Lukasz *et al*. propose an algorithm based on the secondary structure, and extracted the feature of secondary structure through analyzing the secondary structure statistically.

*Corresponding author's Phone: +86 13403797370
DOI: http://Dx.Doi.Org/10.14741/Ijcet/22774106/7.3.2017.30

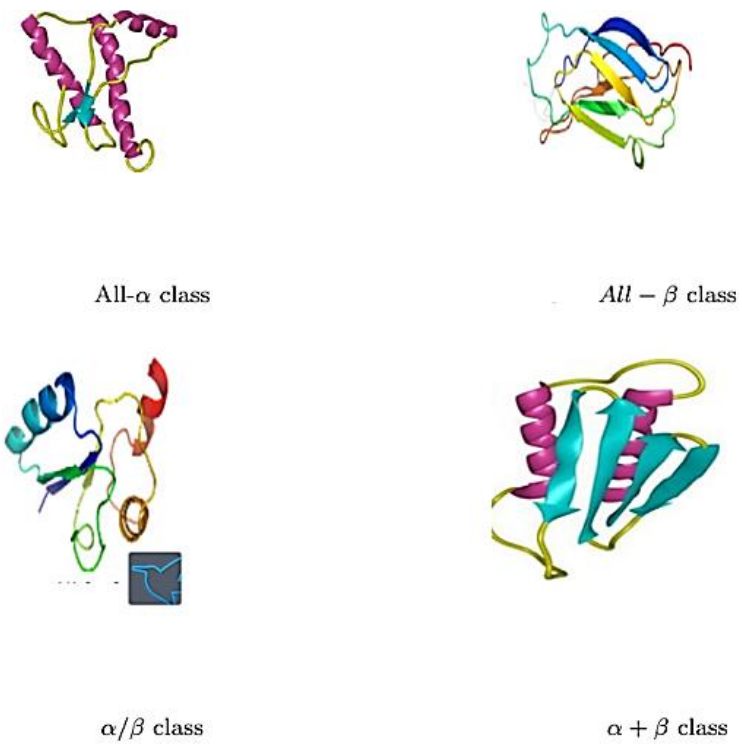

Fig.1 Four representations of structural classes

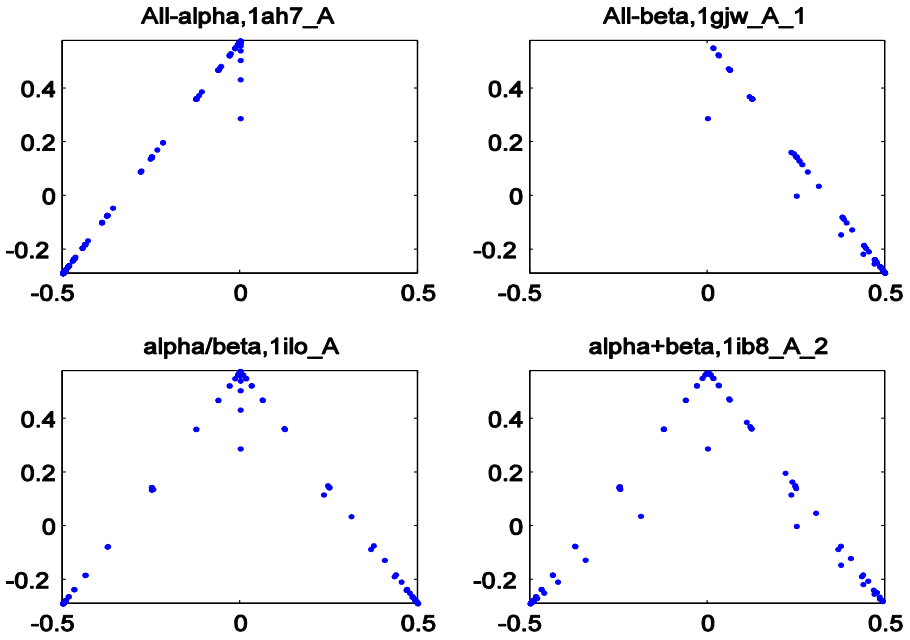

Fig.2 The chaos game representation (CGR) map of each structural class; These points represent traces of the CGR generation processing

They implemented a classifier with calculating the length of each sequence fragments (L. A. Kurgan, Zhang, Zhang, Shen, & Ruan, 2008). In addition, some papers constructed classifier through extracting features from protein secondary structure sequence while without the neural network(Olyaee & Yaghobi, 2010; Yang, Yu, & Anh, 2008). Of course, the situation we discuss later is based on the assumption that the structure of a protein is unknown. In other words, we can only construct a classifier from its primary sequence.

We would like to extract features from secondary structure predicted already; however, the length of secondary structure sequence is variable and what the classifier need are just the feature vectors whose dimensions is fixed. It is Chaos Game Representation (CGR) that allows us to transform a variable-

dimensions feature into a fixed dimensions vector. Yang and Olyaee have already selected the CGR of secondary structure as the feature for classifier. Now, we also choose it as the input of our neural network classifier.

Suppose that we have already got a secondary structure sequence like this: CCCHHHHHHHHHHHHH HHHHHHHCCCCCEEEEEEEECCCCCCCEEEEECCC. Since the protein secondary structure was defined into three classes, we provide an equilateral triangle, representing α, β and coil structure, respectively. First we set the center point as the initial point, then get the meddle point of the connecting line between initial point and the corresponding vertex and replace the initial point with the new point, followed by the processing until all positions on sequence have been processed (Jeffrey, 1990). In that paper, the CGR graph is used to represent the DNA sequence. Finally, we will get the original CGR map. We finally choose the projection of CGR on its three edges as the feature vector. With these feature vectors as input, we used a two layer BP neural networks as classifiers and achieved an acceptable performance.

2. Materials and methods

2.1 Data set

A data set which we call Astral40 was introduced in this experiment with a sequence identity threshold value of 40%. It is very important to ensure objectivity of the prediction by finding a data set with non-correlation sequences. Astral40 is a non-redundant data set with 570 sequences whose secondary structures are clear. More than 100 samples per class make it easier to introduce a cross-validation procedure in our experiment. Astral40 contains 129 All-α, 124 All-β, 158 α/β and 159 α+β proteins. Thus it is suitable for us to develop an algorithm to test the predictor.

Despite the existence of other data sets, we have chosen this widely used (L. Chen *et al*., 2009; Mizianty & Kurgan, 2009; Zhang, Ding, & Wang, 2011) one which makes it easier for us to compare with other papers. Each sequence in the Astral40 has a clear secondary structure sequence as well as its the structural class

2.2 Protein secondary structure prediction with PSIPRED.

The predictor of structural classes in this paper is based on the prediction of protein secondary structure. We selected the CGR of protein secondary structure predicted as the input of the next neural network classifier. Hence, we adopted PSIPRED algorithm using a neural network classifier with psi-blast to predictor the protein secondary structure(Jones, 1999).

Our predictor adopted a similar encoding method to the original PSIPRED; that is to say, using Position-Specific Scoring Matrices (PSSM) in the input encoding. PSSM are generated in multiple sequence alignment through a program named psi-blast. We introduced a sliding window with a size of 15. For each amino acid residue in the sliding window, PSIPRED adopted 21 units to describe them. It includes 20 units of the PSSM score normalized by using the standard sigmoid function:$1/(1+e^{-x})$, and an extra unit to locate its position in the sequence(Jones, 1999). Now we get an input code of 15*21 units. We finally adopted three parallel PSIPRED classifiers to predict the protein secondary structure. In view of the limitation of space, we will not specify the training process and its parameters.

2.3 Protein structural classes prediction based on secondary structure

The structural classes predictor in this paper was performed with a neural network classifier witch extracted the feature of the protein secondary structure as input. With this training and testing procedure, the protein has be divided into four categories by the predictor.

2.3.1 Feature extraction and encoding

Given that a sequence of protein secondary structure has an variable length, but the neural network need a fixed dimensions feature vector as its input, we need extract a fixed length feature vector from these very sequences. In the introduction we have already introduced that we can select the CGR of a sequence as its fixed dimensions feature. More importantly, CGR is closely related to the order of sequences, and different orders generate different CGR maps, so we can see them as fingerprints of sequences.

As shown in Fig.2, all these dots in CGR are distributed inside an equilateral triangle. We set the center point of the triangle as the origin point, and constructed three coordinate systems which are perpendicular to the edges. We used the projection histograms of these dots in the three coordinate systems as a classifier input feature, using a sampling interval of 0.025; the merge of the histograms is shown in Fig.3.

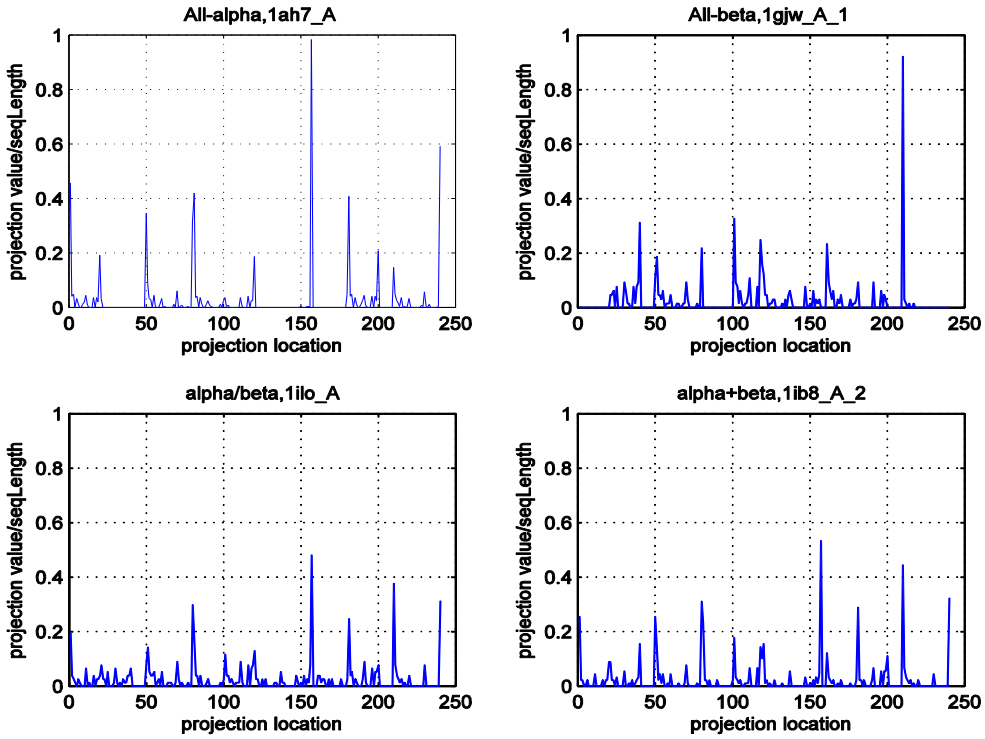

Fig.3 The Projection of CGR, because we set the sampling interval as 0.025, we will get a 1/0.25*6 namely a 240 dimensions vector

2.3.2Neural network in structural classes prediction

Artificial neural network (ANN) is a nonlinear and adaptive information processing system composed of a large number of processing units interconnected. Given that we are all familiar with neural networks, we will not go into the details about them. What we need to account for is the activation functions we chose.

We select the *tanh* function as the activation function for the first layer and the function *sigmoid* as in the secondary layer.

$$\tanh = \frac{1-e^{-x}}{1+e^{-x}} \quad (1)$$

Note that, the function who is an odd routine maps the output value into -1 to 1; but the sigmoid one maps its output from 0 to 1.

$$sigmoid = \frac{1}{1+e^{-x}} \quad (2)$$

As we got 240 dimensions feature vectors from the secondary structures, the input nodes number of the final network was set to 240; doubtlessly, we set the output layer node number to 4. Finally, we choose 80 as the number of hidden layer nodes according to the experience.

2.4 Training and testing processes

Using a 10-fold cross-validated strategy, we divided all these 570 sequences into 10 groups. For each group, we took it as the test s*et al*ternately, and other groups as the training set and the validation set. The validation set do help to avoid the over fitting occurs. This training procedure has a momentum of 0.9 and an initial learning rate of 0.01. We selected a stochastic gradient descent algorithm as the training mode in this paper, and utilized the validate set to avoid the over fitting.

Now, the roughly prediction process are described in the Figure 4.

4. Results and analysis

4.1 Accuracy score of the protein secondary structure prediction

The accuracy of protein secondary structure prediction directly affects the later classification of structural classes. The predictor with three parallel classifiers performs better than the one with a single classifier. Hence, we only display the multiple classifiers result here,

As this figure is just the summary of the above descriptions, and it is so clear that we need not to give more additional elaborate.

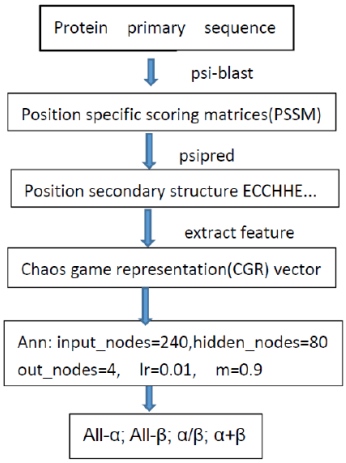

Fig.4 A general prediction process from the primary sequences

3. Performance measurement

The Q4 score was introduced into this paper to evaluate the performance of these predictors. The Q4 score for a prediction is defined as the ratio of correct predicted proteins number relative to the total number of protein present in the appoint classes.

$$Q_i = \frac{A_{ii}}{\sum_{j=1}^{4} A_{ij}} \times 100\%, \ for\ (i, j) = \text{All-}\alpha, \text{All-}\beta, \alpha/\beta, \alpha+\beta. \quad (3)$$

Most authors use the overall 4-state accuracy which suggests the percentage of all correctly predicted proteins:

$$Q_{total} = \frac{\sum A_i}{N} \times 100\%, \ for\ i = \text{All-}\alpha, \text{All-}\beta, \alpha/\beta, \alpha+\beta. \quad (4)$$

Where N is the number of all test proteins. You can also call it recall rate or sensitivity(Yu *et al.*, 2013).

We have not suggested the Matthew correlation coefficient (MCC) or the F-measure because of the balanced performance of the prediction. In these very scenes, there are only tiny differences in the number of samples of different classes. The overall correct rate is quite objective.

4. Results and analysis

4.1 Accuracy score of the protein secondary structure prediction

The accuracy of protein secondary structure prediction directly affects the later classification of structural classes. The predictor with three parallel classifiers performs better than the one with a single classifier. Hence, we only display the multiple classifiers result here. In table 1, we find that the predictor have achieved a fairly good accuracy, and the classification base on the prediction of protein secondary structure is reliable

Table 1 Q3 accuracy of the protein secondary structure

fold	1	2	3	4	5	6	7	8	9	10	overall
α	81.7	77.2	77.6	81.3	79.4	79.0	79.1	82.3	75.2	76.2	78.9
β	69.7	68.2	64.1	65.7	67.2	64.2	67.4	65.9	68.0	67.8	66.8
coil	78.8	76.5	79.0	76.5	76.8	75.9	79.1	78.1	77.4	78.2	77.6
overall	78.1	74.9	75.1	76.3	76.1	74.2	76.6	76.7	74.4	75.3	75.8

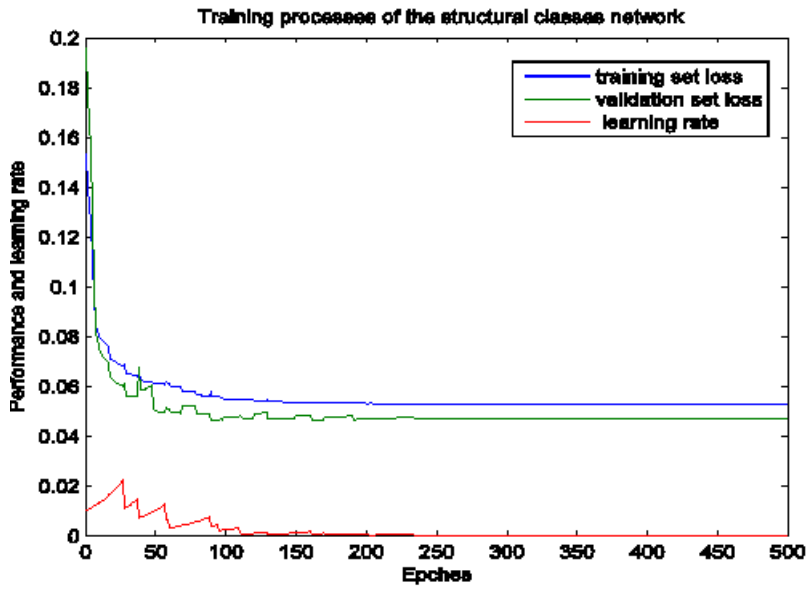

Fig.5 An general prediction process from the primary sequences

4.2 The performance of the training process

We will evaluate the performance of the training process before discussing the exact experiment result. As shown above , we find that training process rapidly converged, this situation benefited from the appropriate encoding mechanism . The adaptive learning rate makes the program converge quickly, and avoids the over-fitting to a certain extent.

Various methods have been discussed to avoid over fitting, for example, the dropout in depth learning (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014), and the early stop mechanism in the neural network (Caruana, Lawrence, & Giles, 2001). The adaptive learning rate we adopted is also an early stopping mechanism to a certain extent.

4.3 Comparisons with related methods

For the structural classes prediction problem, there are several methods related, they has also adopted the CGR or the secondary structure prediction. Yang and Peng *et al*. donated a method named AAD-CGR (amino acid to DNA CGR) which adopt the CGR but gave up the secondary structure information. They obtain an overall accuracy of 65.2% on the 40% identity database. In addition to this comparison, this paper also compares the other methods mentioned in that paper (Yang *et al*., 2009).

Table 2 Q4 accuracy of the related methods

Method	Q_α	Q_β	$Q_{\alpha/\beta}$	$Q_{\alpha+\beta}$	$Q_{overall}$
This paper	80.6	71.0	77.8	57.2	71.2
AAD-CGR	63.2	67.7	63.1	66.5	65.2
20-CGR	44.4	54.4	51.5	63.2	54.4

Table 2 shows the Q4 accuracy scores of each method involved in this paper. This paper achieved an overall Q4 score of 71.08% with the CGR of predicted secondary structure. We take note that the method 20-CGR who introduced the amino acid information into predictors performed unsatisfactorily. Reasons for this kind of situation may lie in many aspects, for instance, it may be that the dimension of CGR is too large. With the AAD-CGR, Yang reversed the amino acid into DNA, finally ADD-CGR algorithm achieve a relative better result of 65.2%. The reason for their lower classification accuracy may be that they ignored the secondary structure of the protein. At the same time, we also considered the evolutionary information with PSSM. So we obtain a satisfactory result of 71.2%. The result of 10 fold cross-validate experiment displays below:

Table 3 Q4 accuracy of each fold

fold	1	2	3	4	5	6	7	8	9	10	overall
alpha	77.8	85.0	75.0	65.0	83.4	84.6	100.0	100.0	81.8	72.7	80.6
beta	75.0	60.0	53.8	71.4	73.3	83.3	90.9	58.9	45.4	81.2	71.0
alpha	60.0	78.6	88.2	64.1	71.4	71.4	66.7	75.0	77.3	85.7	77.8
alpha	58.3	53.1	60.0	72.0	50.0	50.0	63.6	62.5	61.5	55.6	57.2
overall	68.4	68.4	70.2	71.9	68.4	70.2	77.2	71.9	68.4	77.2	71.2

Conclusions

We could conclude that the neural network classifier with CGRs as its input vectors performs effectively. We introduced the evolutionary information to the prediction with introduction of PSSM. The CGR can transform the variable length sequence to a fixed length feature vector. However, the accuracy of the protein secondary structure has great influence in the final classifier. If we use the correctly predicted secondary structure as the feature, we will obtain accuracy beyond 80%.

Nevertheless we will also search for the global information which actually determines the prediction of the secondary structure. Many scholars have done a lot of work before, and have obtained considerable results (Ceroni & Frasconi, 2004; Ni & Niranjan, 2010). It is noteworthy that there is still missing remote information in our predictor, and exploring suitable remote information is also the future work we will consider. Only the accuracy of secondary structure prediction is enhanced, will the classification of structural classes will be improved.

Acknowledgments

The authors would like to thank all reviewers for their suggestions to improve the manuscript. This research is supported by the National Natural Science Foundation of China Under the Grant 61672210.

References

Caruana, R., Lawrence, S., & Giles, L. (2001). *Overfitting in neural nets: backpropagation, conjugate gradient, and early stopping.* Paper presented at the International Conference on Neural Information Processing Systems.

Ceroni, A., & Frasconi, P. (2004). *On the role of long-range dependencies in learning protein secondary structure.* Paper presented at the IEEE International Joint Conference on Neural Networks, 2004. Proceedings.

Chen, C., Chen, L. X., Zou, X. Y., & Cai, P. X. (2008). Predicting protein structural class based on multi-features fusion. *Journal of Theoretical Biology, 253*(2), 388-392.

Chen, L., Lu, L., Feng, K., Li, W., Song, J., Zheng, L., . . . Lu, W. (2009). Multiple classifier integration for the prediction of protein structural classes. *Journal of Computational Chemistry, 30*(14), 2248-2254. Chou, K. C., & Zhang, C. T. (1995). Prediction of protein structural classes. *Critical Reviews in Biochemistry and Molecular Biology, 30*(4), 275.

Chou, P. Y. (1989). *Prediction of Protein Structural Classes from Amino Acid Compositions*.

Huang, W., Chen, M., & Lü, Z. (2006). Energy optimization for off-lattice protein folding. *Physical Review E, 74*(74), 041907.

Hutchinson, E. G., Morris, A. L., & Thornton, J. M. (1976). Structural patterns in globular proteins. *Nature, 261*(5561), 552.

Jeffrey, H. J. (1990). Chaos game representation of gene structure. *Nucleic Acids Research, 18*(8), 2163. Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology, 292*(2), 195-202.

Koswatta, T. J., Samaraweera, P., & Sumanasinghe, V. A. (2011). Simple comparison between specific protein secondary structure prediction tools. *23*(1).

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), 1-15.

Kurgan, L. A., Zhang, T., Zhang, H., Shen, S., & Ruan, J. (2008). Secondary structure-based assignment of the protein structural classes. *Amino Acids, 35*(3), 551.

Mizianty, M. J., & Kurgan, L. (2009). Modular prediction of protein structural classes from sequences of twilight-zone identity with predicting sequences. *BMC Bioinformatics, 10*(1), 414.

Nakashima, H., Nishikawa, K., & Ooi, T. (1986). The Folding Type of a Protein Is Relevant to the Amino Acid Composition. *Journal of biochemistry, 99*(1), 153-162. Ni, Y., & Niranjan, M. (2010). *Exploiting Long-Range Dependencies in Protein β-Sheet Secondary Structure Prediction.* Paper presented at the Iapr International Conference on Pattern Recognition in Bioinformatics.

Olyaee, M., & Yaghobi, M. (2010). *Improved Protein Structural Class Prediction Based on Chaos Game Representation*: IEEE.

Rithvik, M., & Rao, G. N. (2015). A Comparative Study of Methodologies of Protein Secondary Structure. 37-45.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research, 15*(1), 1929-1958.

Yang, J. Y., Peng, Z. L., Yu, Z. G., Zhang, R. J., Anh, V., & Wang, D. (2009). Prediction of protein structural classes by recurrence quantification analysis based on chaos game representation. *Journal of Theoretical Biology, 257*(4), 618-626.

Yang, J. Y., Yu, Z. G., & Anh, V. (2008). Protein Structure Classification Based on Chaos Game Representation and Multifractal Analysis.

Yu, D. J., Hu, J., Tang, Z. M., Shen, H. B., Yang, J., & Yang, J. Y. (2013). Improving protein-ATP binding residues prediction by boosting SVMs with random under-sampling. *Neurocomputing, 104*, 180-190.

Zhang, S., Ding, S., & Wang, T. (2011). High-accuracy prediction of protein structural class for low-similarity sequences based on predicted secondary structure. *Biochimie, 93*(4), 710-714.