# Auto Covariance combined with Artificial Neural Network for predicting Protein-Protein interactions

Juanjuan Li[1, a], Yuehui Chen[2, b]

[1]Shandong Provincial Key Laboratory of Network Based Intelligent Computing University of Jinan, 250022, Jinan, China

[2]Shandong Provincial Key Laboratory of Network Based Intelligent Computing University of Jinan, 250022, Jinan, China

**Keywords:** predicting PPIs, auto covariance, ANN

**Abstract.** Proteins play biological function through the interactions in organisms. Proteins are major components of organisms, and they are of great significance. As an increasing number of high-throughput biological experiments are carried out, a large amount of biological data is produced. Bioinformatics is developed to study the relative data which turns out to be difficult to study using biological methods. The paper mainly studies how to apply the intelligent calculation methods to protein- protein interactions (PPIs) prediction. We proposed an approach, by combining auto covariance with artificial neural network classifier, to predict PPIs. Experiments show that our method performs better than related works with a 5% higher accuracy.

## Introduction

The original study of PPIs starts from biological experiments. Among these methods Yeast two-hybrid (Y2H) [1] is commonly used. It is a biological method aiming at the yeast. MS-PCI [2], short for Mass spectrometry protein complex identification, is another popular use. It takes a viewpoint of the protein molecular and atomic micro, based on which forecast is carried on. Also, we have Protein chip technology [3], which solidifies some proteins already known to us on a chip and then use the chip to predict the interactions of proteins. Biological experiments used for PPIs prediction problems have many advantages. These experiments are easy to manipulate, and the results turn out to be intuitive and reliable. However, such experiments for high throughput data are impossible; it is considerably time-consuming. Intelligent calculation methods are introduced to deal with such problems. Intelligent calculation combines computer techniques with biological ones and benefits from development of computer science. It is a notable solution to biological problems with high throughput calculation [4].

## Technical Introduction

**Auto covariance.** In statistics, auto covariance refers to the covariance of a specific time sequence or a continuous signal *Xt*. That is, the covariance between signals and their neighbor time signals. Auto covariance (AC) here is used as a feature extraction method. As AC derives from signal simulation, we tend to believe that some fluctuations of signals will be tested [5] when we replace the protein sequences with numerical sequences using appropriate physicochemical properties. A further analysis of the numerical sequence signals, concerning the difference between protein-protein interactions sequence and non-interacting sequence pairs will lead to a more effective method than any other methods [6]. Actually, in our experiments, AC applied to PPIs interactions is proved to be effective.

**The artificial neural network (ANN).** The neural network is an algorithm simulating the process of human neurons. The nonlinear input data is transformed to a single output node through a non-linear transformation. The neural network has advantages of both robustness and fault-tolerant. At the same time it performs effectively in learning and adapting to uncertain systems. The typical structure of neural network has three layers. these are the input layer, the hidden layer and the output layer. To use neural network we need to set the number of neurons of the input layer,

Published by Atlantis Press, Paris, France.

the hidden layer and the output layer, respectively. The neural network classification process has two steps: first, training neural network parameters by continuously inputting training data; second, inputting test data to get the results of the test data using the optimization neural network. Artificial neural network (ANN) classification has the advantages of speed, potential super speed, and good fault-tolerance ability. It is suitable for problems without good solving rules. Compared to classifiers using clustering, support vector machine (SVM) and nuclear nearest neighbor for example, ANN has better classification results.

## Research programs

**Dataset.** The dataset used in this paper is Human dataset. The complexity of both human biological structure and interactions between proteins that cooperate to provide functions makes it difficulty to predict the dataset. The human dataset contains 914 positive protein pairs and 941 negative protein pairs, with 1882 protein pairs as a total. This dataset is used to examine the effectiveness of the method we choose [7].

**Feature extraction.** There are three main steps in the whole process to extract features of the PPIs sequence pairs and form a one-dimensional vector as the neural network input. Firstly, replacing the protein interaction sequences with corresponding numerical sequences using appropriate choice of the physicochemical properties of the amino acid descriptors; Secondly, calculating auto covariance of the numerical sequence; Thirdly, connecting the feature values of PPIs sequence pairs to obtain an one-dimensional vector which can be used as an input to ANN to be trained. After consideration, we choose three types of physicochemical properties. They are transfer free energy (TFE), amino acid composition (AAC) and CC in regression analysis (CC). The AC is calculated as follows:

$$R(i,i+k)=\frac{1}{N-k}\sum_{i=1}^{N-k}(X_i-u)*(X_{i+k}-u) \quad (1)$$

Where R(i,i+k) is the AC value, N is the length of the protein sequence. We have k as the amino acid descriptors interval, with value ranges [1, lg], where lg defines the maximum interval. $X_i$ represents the amino acid physicochemical properties value, and $X_{i+k}$ is the physicochemical properties of amino acid that has an offset of k from the current one. Also, u represents the average value of the sequence corresponds to the physicochemical properties of one protein sequence, and it can be calculated as follows:

$$u=\frac{1}{N}\sum_{n=1}^{N}x_n \quad (2)$$

Conforming to the chosen interval, the protein sequence can be converted into different dimensions. We choose a maximum interval of 20, and it follows that the protein sequence is converted to be 20-dimensional. Therefore, our neural network input is 40-dimensional.

**Classify.** We need to establish three base classifiers to study three types of physicochemical properties. The classifier chosen is artificial neural network of three layers. After several attempts, the final choice of artificial neural network structures with three types of physicochemical properties are: for TFE, the structure is 40-7-1 (input layer - hidden layer - output layer), for AAC, the structure is 40-8-1, for CC in regression analysis, the structure is 40-8-1.

## Analysis of Results

Repeated experiments has been established to prove the effectiveness of selected methods; the following tables show the results of three physicochemical properties and the final best result integrated.

Table 1 result of three physicochemical properties and integrated

| PCP | TFE | AAC | CC | Integrate |
|---|---|---|---|---|
| result | 75.0 | 76.5957 | 74.4681 | 83.5106 |

Table 2 lists the results of different methods, where AC (3) present calculating AC of three

physicochemical properties:

Table 2 results of different method used

| Data | classifier | feature | code | result |
|---|---|---|---|---|
| Human | SVM | KMC+KNN+BIO[8] | Link | 73.10 |
| Human | KNN | ACC[9] | Summation | 73.90 |
| Human | PNN | AC(11) | Link | 78.37 |
| Human | ANN | AC(3) | Link | 83.51 |

Table 3 lists the results of different methods using the same feature combined with different classifiers:

Table 3 results of different method I used

| Data | classifier | feature | code | result |
|---|---|---|---|---|
| Human | KNN | AC(3) | Link | 69.15 |
| Human | PNN | AC(3) | Link | 78.37 |
| Human | ANN | AC(3) | Link | 83.51 |

## Summary

This paper uses ANN as a classifier to predict protein-protein interactions. We choose three physicochemical properties among a variety of physicochemical properties of proteins to calculate combined auto covariance, which turned out to be an effective means. Furthermore, feature extraction method combining with artificial neural network achieves high precision. The advantage of integration is notable, as we see a improved accuracy. Therefore, the method we used is effective.

## Acknowledgements

This research was partially supported by the Natural Science Foundation of China (61070130), the Key Project of Natural Science Foundation of Shandong Province (ZR2011FZ001), the Key Subject Research Foundation of Shandong Province and the Shandong Provincial Key Laboratory of Network Based Intelligent Computing.

## References

[1] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori and Y. Sakaki, A comprehensive two-hybrid analysis to explore the yeast protein interactome, Proceedings of the National Academy of Sciences USA 98(8) (2001)4569-4574.

[2] Y. Ho, A. Gruhler, A. Heibut, et al, Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectromety, Nature 415(2002) 180-183.

[3] H. Zhu, M. Bilgin, R. Bangham, D. Hall, A. Casamayor, et al, Global analysis of protein activities using proteome chips, Science 293 (2001)2101-2105.

[4] Zhoujun Li, Yiming Chen,Study of protein interactions in a review of calculation methods. Journal of Computer Research and Development. 45(12)(2008)2129—2137.

[5] Min. Zhu, Yongqing. Zhang, Menglong. Li, Dawei. Zhou, Huang Jun，Based on the integrated learning approach for predicting protein-protein interactions. JOURNAL OF SICHUAN UNIVERSITY(ENGINEERING SCIENCE EDITION). 43(3)(2011).

[6] Y. Guo, L. Yu. Z. Wen and M. Li, Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences, Nucleic Acids Research. 36(9) (2008) 3025-3030.

[7]Guo Y, Li M, Pu X, et a1.PRED-PPI:a server for predicting protein-protein interactions based on sequence data with probability assignment[J].BMC Research Notes,3(1) (2010)145-151.

[8] Jie. Song, Prediction of protein-protein interaction using kernel nearest neighbor algorithm, Application Research of Computers. 26(11)(2009).

[9] Zhengrong. Zhou, Xiaofeng. Song, Minghao. Wang, Using a combination of classifiers for predicting protein-protein interactions, ACTA ELECTRONICA SINICA. 38(6)(2010).