

# MAHALANOBIS DISTANCE WITH RADIAL BASIS FUNCTION NETWORK ON PROTEIN SECONDARY STRUCTURES

T. İbriği<sup>1,2</sup>, M. E. Brandt<sup>2</sup>, G. Wang<sup>2</sup>, M. Açıkkar<sup>1</sup>

<sup>1</sup> Department of Electrical-Electronics Engineering, Çukurova University, Adana, Turkey

<sup>2</sup> Center for Computational Biomedicine, University of Texas Health Science Center, Houston, Texas, 77030, USA

**Abstract-** In this paper, the radial basis function (RBF) network method with the Mahalanobis distance was applied to predict the content of protein secondary structure elements. A study of the Mahalanobis-RBF with different window sizes on the dataset developed by Qian-Sejnowski is given. The RBF network predicts each position in turn based on a local window of residues, by sliding this window along the length of the sequence. Comparison of Gaussian-RBF and Mahalanobis-RBF on the Qian dataset shows that the Mahalanobis distance in using RBF gives better results in the prediction of secondary structure for local sequence structural state.

**Keywords – Mahalanobis Distance, Gaussian-RBF Neural Networks, Protein Secondary Structure**

## I. INTRODUCTION

Secondary structure prediction is an important element in understanding how the amino acid sequence of a protein determines the native state. The principles governing protein structure are complex and not yet well understood.

Early attempts to predict secondary structure focused on development of mapping from local windows of residues in the sequence to the structural state of the central residue in the window. Qian and Sejnowski established superior results in 1988 [1]. They used 104-protein sets of known proteins obtained from Brookhaven National Laboratory in California, USA, and extracted from several testing sets of known proteins without structural or sequence homology. They used all the sets except the test set to train the network to predict the secondary structure (helix (H), strand (S), or coil (C)), and used the sigmoid equation for the transfer function. The general rules for taking protein structures from existing databases and applying them to sequences of unknown structures currently appear to be the most practical starting point for protein structure prediction. Published results in the literature that neural networks have produced the most accurate secondary structure prediction with respect to the more conventional methods over the last fifteen years. The current best methods reach accuracies of about 80% with multiple nonhomologous sequences and 70% for single sequence prediction [2,3].

## II. MATERIALS AND METHODS

### A. Data Sets

The data set of Qian and Sejnowski classifies known structures as  $\alpha$ -helix (H), and  $\beta$ -strand (S) [1]. Residues which are neither H nor S are classified as coil (C). The set contains 104 globular proteins, 21,630 amino acid with 24.5%  $\alpha$ -helix, 20.5%  $\beta$ -strand and 55 % coil.

### B. Structure of Network

As in most of the existing methods, the secondary structure of the  $i^{\text{th}}$  position of amino acid chain  $R_i$  is predicted from the window of amino acids,  $R_{i-n}, \dots, R_i, R_{i+1}, \dots, R_{i+n}$  where

$$\text{winsize} = 2*n+1. \quad (1)$$

Each pattern presented to the network comprises

$$N_{\text{inp}} = \text{winsize} * M \quad (2)$$

inputs for an input vector of size  $n$ ,  $M$  is the encoding which we use 24 values. The first 20 values, which are either 1 or 0, which are used to represent the identity of each protein residue in the sequence string, encode amino acids. The string has a vector of 20 values among which 19 have a value of 0, and one has a value of 1. The input vector also has two values for the relative position in the protein sequence, two values for the relative size of the chain which are  $L/L_{\text{max}}$ ,  $1-(L/L_{\text{max}})$  and  $i/L$  and  $1-(i/L)$ , respectively, where  $L$  is length of protein amino acid sequence,  $L_{\text{max}}$  is the maximum length of protein sequence in the dataset, and  $i$  is the position of the residue in the protein chain [2]. The network architecture is a fully connected

$N_{\text{inp}} - m - N_{\text{out}}$  network, where  $N_{\text{inp}}$  is the input vector size,  $h$  is the number of hidden layer nodes which were calculated by

$$h = (N_{\text{inp}} * N_{\text{out}})^{1/2} \quad (3)$$

and  $N_{\text{out}}$  is the output vector size that is three for secondary structures which are helix, strand and coil [4].

### C. Measurement of Accuracy

The accuracy was measured by the  $Q_3$  standard

$$Q_3 = \frac{H + S + C}{N} \quad (4)$$

$H$ ,  $S$ , and  $C$  are the correctly predicted helix, strand and coil, respectively, divided by the total number of predicted residues,  $N$ .

#### D. Method

Prediction of protein secondary structure is tested by using RBF network that traditionally has only a single hidden layer, and techniques from statistics, such as forward selection and ridge regression, as strategies for controlling model complexity [5,6].

$$r^2 = (x - m_x)' c_x^{-1} (x - m_x) \quad (5)$$

is called the Mahalanobis distance from the input vector  $x$  to the mean vector  $m_x$ , where  $c_x$  is the covariance matrix for  $x$ . It can be shown that the surfaces on which  $r$  is constant are ellipsoids that are centered about the mean  $m_x$ .

The basis functions are usually local, that they respond most strongly to the inputs nearest to center  $r$ , in the matrix determined by the radius. The function is given as

$$f(x) = \sum_{j=1}^n w_j r_j(x) \quad (6)$$

where  $f(x)$  function is a transfer function,  $x$  are point of residue,  $w_i$  are weight parameters, and  $r(x)$  is the Mahalanobis distance.

### III. RESULTS AND DISCUSSION

We have found an increase in accuracy of secondary structure prediction with RBF with Mahalanobis distance, compared with Gaussian-RBF. We reached the high value ( $Q_3$ ) that is 72.67 % at window size 13 on the Qian data by using Mahalanobis distance (Fig. 1). The  $Q_3$  value decreased 2% from window size 13 to 21. The highest value of Mahalanobis is 9.47 % higher than Qian's study [1]. When the Mahalanobis-RBF and Gaussian-RBF were compared with each other, the  $Q_3$  of Mahalanobis was found to be higher than the  $Q_3$  of Gaussian-RBF.

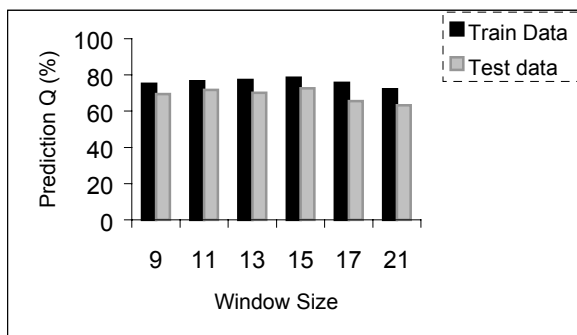


Fig. 1. Prediction Accuracy of Mahalanobis RBF on Qian-Sejnowski Dataset

The value is 17.80 % higher than the Gaussian-RBF (Fig. 2) [7]. But the relative size of the chain, and the

relative position in the protein sequence were not included in the Gaussian-RBF. They may affect the accuracy of  $Q_3$ . The all windows sizes are not a critical issue for Gaussian-RBF, but Mahalanobis-RBF gives different results for different window sizes.

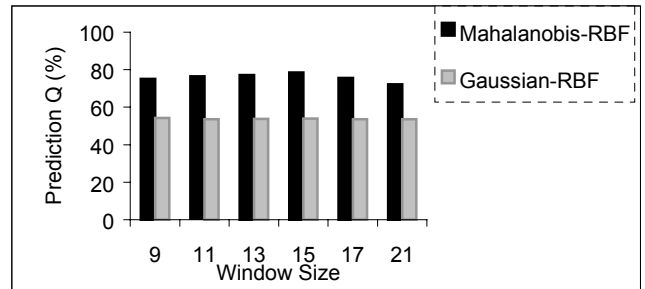


Fig. 2. Comparing Gaussian-RBF and Mahalanobis-RBF on Qian Dataset

### V. CONCLUSION

The prediction of secondary structure of proteins using radial basis functions (RBF) with Mahalanobis distance and different windows sizes was studied, and compared with the use of the Gaussian-RBF. RBF networks have been applied successfully to many everyday engineering problems. In this study we found that the RBF network with Mahalanobis distance is suitable to predict the secondary structure of a protein. The Mahalanobis-RBF also shows measurable improvements in accuracy  $Q_3$ .

### ACKNOWLEDGMENT

The Scientific and Technical Research Council of Turkey (TUBITAK) supported this research with grant number EEEAG-100E037.

### REFERENCES

- [1] Qian, N., Sejnowski, T. J. "Predicting the Secondary Structure of Globular Proteins Using Neural Network Models", *J. Mol. Biol.* Vol:202,pp: 865-884 , 1988.
- [2] Petersen, T. N., Lundegaard, C., Nielsen, M., Bohr, H., Bohr J., Brunak, S., Gippert, G. P. and Lund, O. "Prediction of Protein Secondary Structure at % 80 Accuracy", *Proteins: Structure, Funct., and Genetics*; 41:17-20 (2000)
- [3] Rost, B., Sander, C. "Prediction of protein secondary structure at better than 70% accuracy", *Journal of Molecular Biology*, Vol: 232, pp: 584-599, 1993.
- [4] Masters, T. *Practical Neural Network Recipes in C++*, Academic Press, Boston (1994)
- [5] Bishop, C. M. *Neural Networks for Pattern Recognition*, *Oxford University Press*, 1996.
- [6] Broomhead, D, S., Lowe, D. "Multivariable functional interpolation and adaptive networks" *Complex Systems* Vol.2 pp:321-355, 1988.
- [7] Ibrikci, T.,Guler M., Acikkar, M., "Assessment of Radial Basis Function Network on Protein Secondary Structures", (#606) EMBC-2001