Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

E. coli Promoter Prediction Using Feed-Forward Neural Networks

Fan Zhang, Michael D. Kuo, Adrian Brunkhors*

[1]Department of Radiology and [2]Center for Translational Medical Systems, University of California, San Diego
** To whom correspondence should be addressed*

Abstract

E. coli promoter recognition is an area of great interest in bioinformatics. In this paper, we describe the implementation of a feed forward neural network to predict the E. coli promoter. According to the sequence conservation, some sequences with 60 bases are selected as positive samples and some corresponding non-promoters from E. coli coding areas are selected as negative samples, and a classifier based on feed forward neural network is trained. Results show that feed forward neural networks can extract the statistical characteristics of promoters more effectively, and that coding with four dimensions for nucleic acid data is superior to two dimensions. Another result demonstrated here is that the number of hidden layers seems to have no significant effect on E. coli promoter prediction precision. The research results in this paper can provide reference for promoter recognition research.

1. Introduction

Currently, there are more than 142 organisms that have been sequenced and annotated. However, a large number of promoter regions in these organisms, which are important transcriptional control points, have not been determined. To date, an additional four more sequencing tasks for E. coli strains have been completed but there still exists many vague promoter regions that need to be further explored and annotated. In this paper we describe the implementation of a feed forward neural network to predict E. coli promoter regions.

2. Feed-Forward Neural Network

Neural Networks have several unique advantages and characteristics as research tools for the molecular sequence analysis problem. A very important feature of these networks is their broad adaptive nature, where "learning by example" replaces conventional "programming on a case-by-case" basis for solving problems[1-4]. For the promoter prediction problem, this feature makes neural networks very appealing.

The proposed new solution uses a feed forward neural network, which has three layers: input layer, hidden layer, and output layer, trained using a back propagation supervised training algorithm. The input is used as activation for the input layer and is propagated to the output layer. The received output is then compared to the desired output and an error value is calculated for each node in the output layer. The weights on edges going into the output layer are adjusted by a small amount relative to the error value. This error is propagated backwards through the network to correct edge weights at all levels.

3. Design/Methods

The positive training sets consist of 471 promoter patterns obtained from promoter databases with experimentally determined and validated transcriptional start sites, each of which is 60 bps long spanning from -55 to +5.[5] Comparing with [7] where only a set of 80 known promoter sequences combined with different numbers of random sequences was trained, more training sets in this paper are used.

The negative training sets are 365 promoter patterns from the coding region of E.coli K12 data available in the NCBI database, each of which is also 60 bps long. [6]

For a neural network to function properly, the input data has to be transformed into binary or numerical data. Two kinds of encoding schemes for nucleic acids are tested. First we use a four dimensional encoding scheme; that is to say, A=1000, C=0100, G=0010,

T=0001. Then we also apply a two dimensions encoding scheme: A=00, C=01, G=10, T=11.

In the four dimensional encoding scheme, the input layer has 240 nodes, the hidden layer 80 nodes, and the output one node.

In the two dimensional encoding scheme, the input layer has 120 nodes, the hidden layer 80 nodes, and the output one node.

For positive training sets, the output is +1, and for negative training sets, the output is -1. If the output is greater than or equal to zero, we say the input is a promoter sequence. And if the output is less than zero, we say the input is not a promoter sequence.

4. Results

Training is performed using the back propagation algorithm for the two kinds of encoding schemes and various hidden layer numbers.

For the first encoding scheme, a learning rate of 0.5, a momentum constant of 0.95 maximum number of epochs to train of 5000 and sum-squared error goal of 0.1 are used. When the hidden layer number is changed, the network training parameters are kept unchanged.

For the second encoding scheme, a learning rate of 0.25, a momentum constant of 0.95 maximum number of epochs to train of 5000 and sum-squared error goal of 0.1 are used. When the hidden layer number is changed, the network training parameters are kept unchanged.

The four dimensional encoding scheme training is shown in figure 1, and the two dimensional encoding scheme training in figure 2.

The four dimensional encoding scheme error rate is shown in table 1, and two dimensional encoding scheme error rate in table 2. Comparing figure 1 and figure 2, and table 1 and table 2, we can conclude that the four dimensional encoding scheme training is superior to two dimensions.

For the four dimensional encoding scheme, the error rate responding to the hidden layer number is shown in figure 3. And for the two dimensional encoding scheme, the error rate responding to the hidden layer number is shown in figure 4. Comparing figure 3 and figure 4, it can be seen that the number of hidden layers seems to have no significant effect on E. coli promoter prediction precision.

5. Conclusion

Neural networks are useful for prediction problems. Their ability to capture and model information from non-linear systems and generalize information from learned data makes them suitable for E coil promoter prediction.

The feed forward neural network described in this paper is capable of determining the position of E. coli promoters. Although other techniques may prove accurate at the same task, the neural network seems to be a suitable and sufficiently accurate choice. The use of a neural network in this manner makes it possible to automatically detect the location of E. coli promoters. Thus, algorithm can more easily determine multiple promoters in different sequences.

The results in this paper also showed that for nucleic acid data, coding with four dimensions is superior to two.

Another result demonstrated here is that the number of hidden layers seems to have no significant effect on E. coli promoter prediction precision.

ACKNOWLEDGMENT

This work was supported in part by UC-San Diego.

REFERENCES

[1] DEMELER B, ZHOU Guang-wen. Neural netw ork optimization for E. coli promoter prediction[J]. Nucleic Acids Re search, l99l, l9(7): l593- 1599.

[2] O'NEIL M C. Training back-propagation neural networks to define and detect DNA binding sites[J]. Nucleic Acids Research, l99l, l9(2): 3l3-3l8.

[3] ANDERS G P, PIERRE B, SOREN B, et a1. Characterization of prokaryotic and eukaryotic promoters using hidden markov models[A]. Proceeding of the Fourth International Conference on Intelligent Systems for Molecular Biology[C]. St. Louis: AAAI Press, l996. 182-l91.

[4] RAN Ling hua, RUAN Xiao gang. An Approach Based on Support Vector M achine for E. coli Promoter Recognition, Journal of Beijing university of technology, vol. 30 No. 4 Dec. 2004 p 432-p436

[5] Ruti Hershberg, Gill BEjerano, Alberto Santos-Zavaleta and Hanah Margalit, PromEC: An updated database of Escherichia coli mRNA promoters with experimentally identified transcriptional start sites, Nucleic Acids Research , 2001, Vol. 29, No.1, P277

[6] NCBI homepage, http://www.ncbi.nlm.nih.gov/

[7] B. Demeler and G. Zhou. Neural network optimization for E. coli promoter prediction. Nucleic acids research, 19:1593, 1991

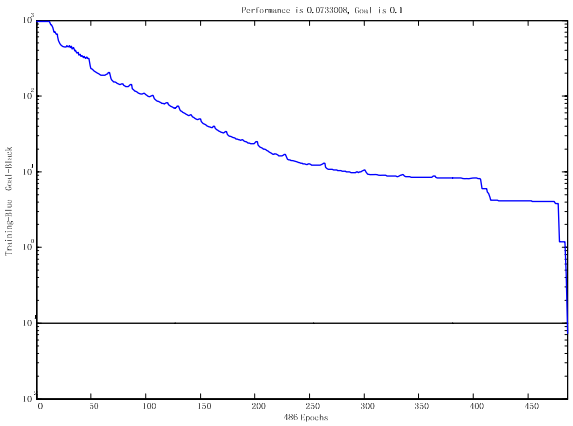

Figure 1. four dimension encoding scheme training

TABLE 1 FOUR DIMENSIONS ENCODING SCHEME ERROR RATE

	Training set		Testing set		Training + testing set		
sequence	P	NP	P	NP	P	NP	ALL
Total No.	471	365	100	100	571	465	1036
Error No.	24	12	10	3	34	15	49
Error rate%	FN 5	FP 3.3	FN 10	FP 3	FN 6	FP 3	4.7

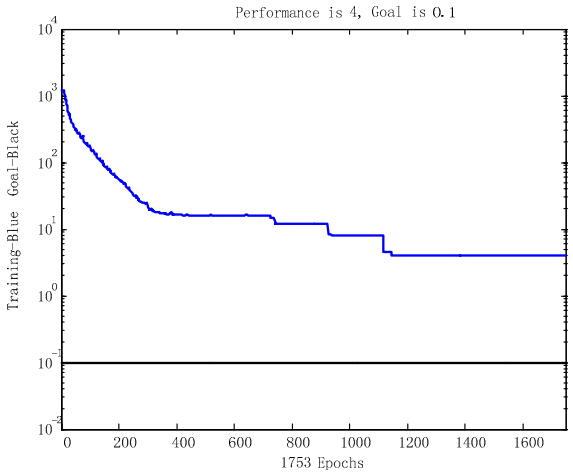

Figure 2. two dimension encoding scheme training

TABLE 2 TWO DIMENSIONS ENCODING SCHEME ERROR RATE

	Training set		Testing set		Training + testing set		
sequence	P	NP	P	NP	P	NP	ALL
Total No.	471	365	100	100	571	465	1036
Error No.	30	14	11	4	41	25	66
Error rate %	FN 6.4	FP 3.8	FN 11	FP 4	FN 7.2	FP 5.4	6.4

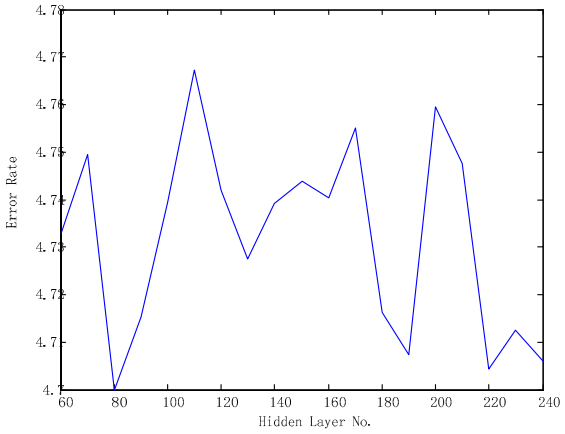

Figure 3. For four dimensions encoding scheme, response of Error rate to hidden layer number

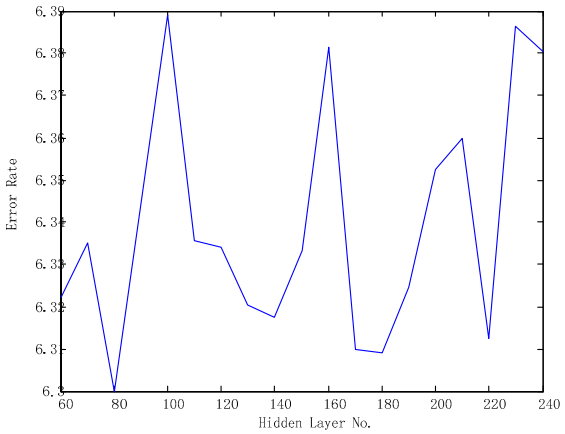

Figure 4. For two dimensions encoding scheme, response of Error rate to hidden layer number