เป็นหนังสือภาษาอังกฤษ

รายงานการวิจัย

การสกัดลำดับเบสที่เด่นสำหรับการทำนายโปรโมเตอร์โดยใช้ค่าน้ำหนักสูงสุดจากโครงข่ายประสาทเทียม

Bases Feature Extraction for Promoter Prediction Using Maximum Weighting from Neural Networks

สถานวิจัยจีโนมและชีวสารสนเทศ
หลักสูตรเทคโนโลยีชีวภาพโมเลกุลและชีวสารสนเทศ

ได้รับทุนอุดหนุนการวิจัยจากเงินกองทุนวิจัยคณะวิทยาศาสตร์
ประเภทพัฒนานักวิจัย ประจำปี 2550

Bases Feature Extraction and Binary Representation Technique for Promoter Prediction Using Maximum Weighting from Neural Networks

Unitsa Sangket [1], Wiphada Wettayaprasit [2], Alisa Nugkaew [1], Amornrat Phongdara[1], Wilaiwan Chotigeat[1]

[1] Center for Genomics and Bioinformatics Research, Faculty of Science, Prince of Songkla University, Songkhla, 90112, Thailand, E-mail: usangket@yahoo.com, E-mail: joy_alisa@yahoo.com
[2] Artificial Intelligence Research Laboratory, Department of Computer Science, Faculty of Science, Prince of Songkla University, Songkhla, 90112, Thailand, E-mail: wwettayaprasit@yahoo.com

***Abstract*—This paper presents a technique of promoter prediction from feature extraction in base sequences and binary representation using maximum weighting from neural networks. The study had tested with the benchmark data set which were the *Drosophila melanogaster* promoter sequences from EPD, the *D. melanogaster* gene sequences from Genbank, and the *Escherichia coli* promoter gene sequences from UCI. The experimental results for promoter prediction received high accuracy comparing with other methods. The training time was also decreased.**

***Keywords*—Promoter Prediction, Neural Networks, feature extraction**

I. INTRODUCTION

Promoters are responsible regions functional for the regulation and initiation of DNA transcription [1] using RNA polymeraseII. The promoter located upstream of the transcription start site (TSS), which informs the enzyme RNA polymerase where to begin the transcription. The promoter sites typically have a complex structure consisting of multiple functional binding sites (BSs) for proteins, called transcription factors (TFs) involved in the transcription initiation process [2]. The structure of promoter is show in Figure 1.

Computational prediction of promoters from the nucleotide sequence is one of the biggest challenges problems in sequence analysis today [3]. Neural Networks is an important knowledge in data mining because data can be separated into groups with high accuracy [4].

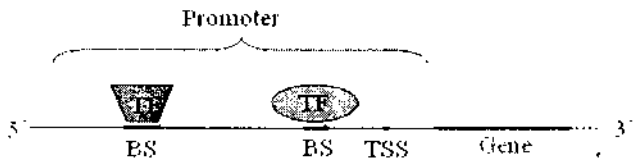

Figure 1. The structure of promoter.

The structure of neural networks imitate the human brain structure that composes of three layers which are input layer, hidden layer, and output layer. The hidden layer can be more than one layer, which is called Multi-Layer Perceptron (MLP). The layer of neural networks composes of multi-neurons connected as networks with the parallel structure [5,6]. The structure of each neuron of neural networks is shown in Figure 2 [7] .

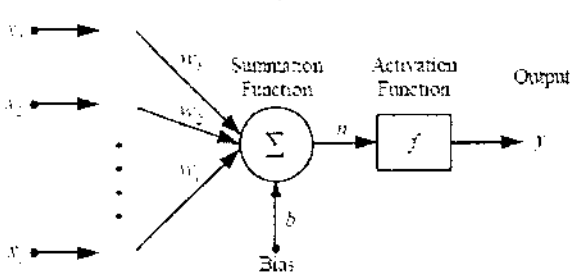

Figure 2. A neuron

Let variable i be the number of node, variable x be the input data, variable w be the weight value, variable b be the bias value, and variable n be the output of summation function. The summation function is shown in (1).

$$n = \sum_{i=1}^{z} x_i w_i + b \qquad (1)$$

Let variable y be the output of activation function. The activation function in (2) is called sigmoid activation function.

$$y = \frac{1}{1+e^{-n}} \qquad (2)$$

Neural Networks is used to one of the attractive fields of computational biology, and especially in the area of computational DNA sequence analysis, to predict the promoter sites. The example of promoter prediction algorithms using neural networks is Promoter Prediction using Time-delay Neural Networks (NNPP) [8].

This paper presents promoter prediction from feature extraction in base sequences and binary representation using maximum weighting from neural networks. The result of the study will be in Section 3 and the last Section will be the conclusion.

II. METHODOLOGY

The promoter prediction from feature extraction using maximum weighting from neural networks has 3 parts as shows in Figure 3. Part 1 is train neural networks by normal data representation. Part 2 is feature extraction of base sequences. In this process, we will prune neural networks using maximum feature weighting [4]. And part 3 is train neural networks by binary data representation.

In part 1, we will train neural networks using backpropagation learning. The multilayer perceptron neural networks architecture has three layers as follows: input layer, hidden layer, and output layer. Let x be the input bases with n bases, h be the hidden nodes with m nodes, and c be the output attributes for class of base sequences with 1 node, which have been represented by 1 if it is a promoter and 0 if it is a non-promoter as shows in Figure 4. The bases representation of each nucleotide as digits are Adenine (A) = 1, Cytosine (C) = 2, Guanine (G) = 3, Thymine (T) = 4, and no nucleotide (N) = 5.

In part 2, we will prune neural networks from part 1 using maximum feature weighting which is considered by the accepted weight values (*Accept_w*) of each layer of neural networks between output layer to hidden layer and between hidden layer to input layer since the larger weight values will have more significance than the smaller weight values. This means that the system will find the maximum weight (*Max_w*) of weight link and calculate the weight that would be accepted (*Accept_w*) from accepted weight percentage (δ) as specified by the user. In this step, the hidden node and input node that are not accepted will be pruned from neural networks [4].

In part 3, we will consider only input node and hidden node that are suggested from part 2. The process of this part is similar to part 1 except the input base representations of each nucleotide. In addition, the base representations of this part as digits are A = 10000, C = 01000, G = 00100, T = 00010 and no nucleotide (N) = 00001 as shows in Figure 5.

Part 1. Train NN by normal data representation
Part 2. Feature extraction of base sequences
Part 3. Train NN by binary data representation

Figure 3. The promoter prediction from feature extraction using maximum weighting from neural networks method

III. EXPERIMENT RESULTS

The benchmark data sets are The Drosophila promoter sequences and The E. coli promoter sequences.

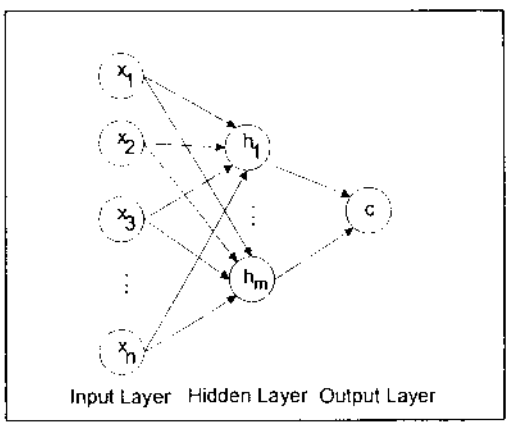

Figure 4. Structure of neural networks of part 1

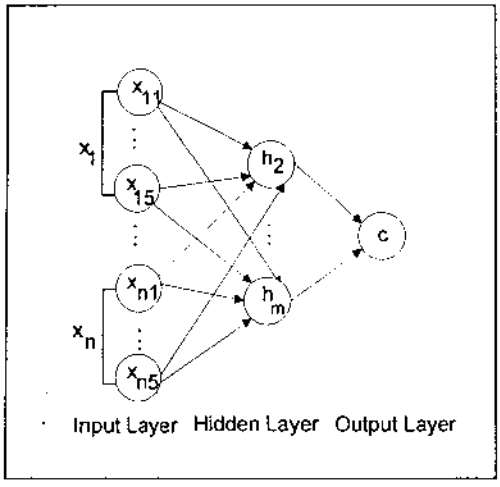

Figure 5. Structure of neural networks of part 3

1.) The *D. melanogaster* promoter sequences (Promoter-Drosophila) contains 4,651 DNA sequences, with 1,926 samples of *D. melanogaster* promoter sequences from EPD, hereafter called the "promoter sequences", and 2,725 samples of *D. melanogaster* gene sequences from Genblank, hereafter called the "non-promoter sequences". Let the range of promoter sequence be from -40 (upstream) to +10 (downstream) relative to the TSS which is defined as position +1, so their lengths are all 51 bases (attributes). A DNA sequence also consists of four types of nucleotides: A, C, G, and T. However some bases have no nucleotide (N).

2.) The *E. coli* promoter sequences from UCI (Promoter-Ecoli) [9] contains 106 DNA sequences, with 53 samples promoter of sequences and 53 non-promoter sequences. The lengths are all 57 bases (attributes). A DNA sequence consists of four types of nucleotides: A, C, G, and T.

Both data sets of experiment have data dividing for training set by 70% and for testing set by 30%. Let the percentage of accepted weight $\delta = 20$. The experiment results shows in Figure 6. The normal method means that the neural networks is trained by normal data representation. The binary method means that the neural networks is trained by binary data representation. The feature extraction and normal method means that there are feature extraction of base sequences and normal method. The feature extraction and binary method means that there are feature extraction and binary method. The study from both data sets show that the feature extraction and binary method gives higher accuracy 93.8 % than the other method.The time used for promoter prediction Figure 7. shows that the feature extraction and binary method used much less time than binary method.

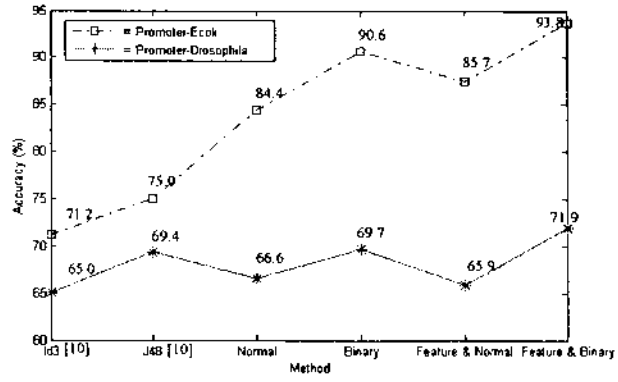

Figure 6. Accuracy Comparison

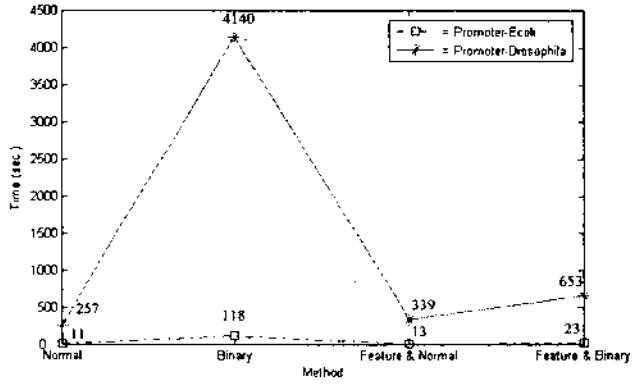

Figure 7. Time Comparison

IV. CONCLUSION

This paper presents promoter prediction from feature extraction in base sequences and binary representation using maximum weighting from neural networks that has 3 parts. Part 1 is train neural networks by normal data representation. Part 2 is feature extraction of base sequences. In this process, we will prune neural networks using maximum feature weighting. And part 3 is train neural networks by binary data representation. The benchmark data sets are The Drosophila promoter sequences and The E. coli promoter sequences. The results of promoter prediction using feature extraction and binary method gives higher accuracy than the other method. In addition, the time used for promoter prediction using feature extraction and binary method used much less time than binary method.

ACKNOWLEDGMENT

This work was supported by Faculty of Science, Prince of Songkla University.

REFERENCES

[1] Bajic,V. B., Seah, S. H., Chong, A., Zhang, G., Koh, J. L. Y., and Brusic, V. Dragon Promoter Finder: recognition of vertebrate RNA polymerase II promoters, *Bioinformatics*, 18, 1 (Jan. 2002), 198-199.

[2] Reese, M. G. *Computational prediction of gene structure and regulation in the genome of Drosophila melanogaster*, Ph.D. Thesis, University Hohenheim, Stuttgart, Germany, 2000.

[3] Pedersen, A. G., Baldi, P., Chauvin, Y., and Brunak, S. The biology of eukaryotic promoter prediction -a review, *Computers & Chemistry*, 23, 2-3 (Jun. 1999), 191-207.

[4] Wettayaprasit, W. and Sungket, U. Linguistic Knowledge Extraction from Neural Network Using Maximum Weight and Frequency Dat Representation. In *Proceedings of the 2006 IEEE International (CIS'06)*, Jun. 7-9, 2006.

[5] Ramírez M. C. V., Velho H. F. C. de and Ferreira N. J.,"Artificial neural network technique for rainfall forecasting applied to the São Paulo region," Journal of Hydrology, Vol. 301, pp. 146-162, 2005.

[6] Roiger R. J., and Geatz M. W., Data mining a tutorial-based primer, Pearson Education, Inc., 2003.

[7] Wettayaprasit, W. and Nanakorn, P. Feature Extraction and Interval Filtering Technique for Time-series Forecasting Using Neural Networks. In *Proceedings of the 2006 IEEE International (CIS'06)*, Jun. 7-9, 2006.

[8] Shamir, R. Promoter Analysis, [Online], Available: www.cs.tau.ac.il/~rshamir/ge/04/scribes/lec09.pdf

[9] Mertz, J. and Murphy, P.M. UCI repository of machine learning databases, [Online], Available: ftp://ftp.ics.uci.edu/pub/machine-learning-databases.

[10] Witten, I. H. and Frank, E. WEKA software, [Online], Available: http://www.cs.waikato.ac.nz/~ml.