RBF NEURAL NETWORK BASED HUMAN GENOME TSS IDENTIFICATION*

Chen Jie, Peng Zhihong, Cao Lijun, Gao Tingting

College of Information Science and Technology,
Beijing Institute of Technology, Beijing, 100081 , P .R. China

Abstract: Identification of functional motifs in a DNA sequence is fundamentally a statistical pattern recognition problem. This paper introduces a new algorithm for recognition of functional transcription start sites (TSSs) in human genome sequences, in which RBF neural network is adopted, and an improved heuristical method for 5-tuple feature viable construction is proposed and is implemented in two RBFPromoter and ImpRBFPromoter packages developed in Visual C++ 6.0. The algorithm is evaluated on several different test sequences sets. Compared with several other promoter recognition programs, this algorithm is proved to be more flexible with stronger learning ability and higher accuracy. *Copyright © 2005 IFAC*

Keywords: Promoter recognition, Human genome, Transcription start site, RBF neural network.

1. INTRODUCTION

Today, science is advanced by new observations and technologies. Human Genome Project has led to a massive outpouring of genomic data, which in turn fueled the rapid developments of high-throughput biotechnologies. A new field of computational molecular biology, saying bioinformatics, is witnessing a revolution brought out by biological science, medical research and information science. Gene finding is one of the most important research fields in bioinformatics, i.e. prediction of gene location and gene products from experimentally uncharacterized DNA sequences (Roderic Guigo, 1997; James W. Fickett, 1996) according to biological meanings. Promoter is a key DNA region that controls and regulates transcription. Computational prediction of eukaryotic promoters from the nucleotide sequences is one of the most attractive fields in sequence analysis nowadays, but it is also a very difficult one since the transcriptional process is incomplete(Anders Gorm Pedersen, et al, 1999). There have been many computational approaches to this problem, such as Markov chain model, Linear discriminant analysis, Quadratic discriminant analysis and simple neural networks, which can be divided into two classes of general promoter recognition methods and specific promoter recognition methods. The general method is to identify TSS (Transcription Start Site) and/or core promoter elements for all genes in a genome. The specific methods focus on identifying specific regulatory elements, e.g. TF sites that are shared by a particular set of transcriptional related genes (Uwe Ohler and Heinrich Niemann,2001; James W. Fickett, et al, 1997; Tao Jiang,et al, 2002).

*Supported by the National Natural Science Foundation of China (60374069).

Notably, compared with the feed forward network, RBF network is a widely used network model. As its name implies, this network makes use of radial basis functions. RBF neural network is designed to perform nonlinear mapping from the input space to the hidden-unit space and linear mapping from the hidden-unit space to the output space. Problems can be solved by transforming into a high dimensional space in a nonlinear manner. The structure of a RBF neural network indicates that a complex pattern classification problem cast in high dimensional space is more likely to be linearly separable than in a low dimensional space. It is promising to apply RBF neural network to promoter recognition.

Although promoter recognition is a typical statistic pattern recognition problem, one difficulty lies on how to obtain available feature variables which represent biological meanings as well as statistical importance in order to improve the prediction accuracy. k-tuple frequency measure is one of the most widely used methods. A global 6-tuple frequency measure has been used for promoter recognition as a "content" measure in the sense of Staden (Hutchinson, G.B., 1996). But it is suggested that this content approach should not consider all of the positional information that is crucial for the recognition of promoter. Also, a pure "signal" approach has no meaning in this case because of the large variation in the signal positions. However , a "mixed" approach, using position-specific windows, had been proposed by Michael Q. Zhang(Michael Q. Zhang, 1997) and proved to be powerful in the QDA algorithm in a program CorePromoter developed by Michael Q. Zhang.

This paper focuses on promoter recognition in human genome with the aim of improving True Positive prediction numbers significantly. In order to explore the truth of promoter recognition, two different methods are implemented in two programs RBFPromoter and ImpRBFPromoter developed in VC++6.0, where RBF neural network is adopted, the later is a particularly creative one with an improved feature variable method. In order to compare with that of Michael Q.Zhang, k is set to be 5 in both RBFPromoter and ImpRBFPromoter, the same as CorePromoter's in promoter recognition. By using several different test sets, evaluation results demonstrate that RBFPromoter and ImpRBFPromoter are valid and efficient.

2. RBF NEUTRAL NETWORK BASED HUMAN GENOME TSS IDENTIFICATION SYSTEM

2.1 Structure of the system

The system consists of four blocks conceptually shown in Fig1. Firstly, the input sequence is pre-processed to satisfy the format requirements of the

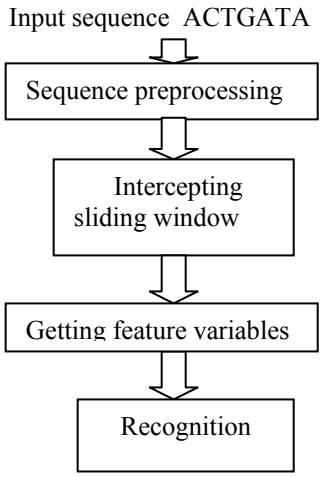

Fig.1 Schematic Representation of the system structure

system, such as changing minuscule letters of sequences to capital characters. Then, a fixed-length window slides along the input sequence to get feature variables based on 5-tuple frequencies. Finally, recognition block gives out recognition results. RBFPromoter is different from ImpRBFPromoter in the third block.

2.2 Feature Variables of RBFPromoter

In RBFPromoter, feature variable of an overlapping window shown in Fig 2 is calculated as that in CorePromoter :

The 5-tuple value of a 5-tuple ***s*** in a window ***w*** can be obtained as following:

$$x(s) = \frac{f_w(s)}{f_w(s) + f_b(s)} \quad (1)$$

where $f_w(s)$ is the signal frequency of a 5-tuple ***s*** in a window ***w***, $f_b(s)$ is the background frequency as in Eq.(2).

$$f_b(s) = \frac{1}{2}(f_L(s) + f_R(s)) \quad (2)$$

where L and R indicate the left and the right nearest neighbour of non-overlapping windows.

Therefore, feature variable of a position-specific window ***w*** is defined as the mean of all 5-tuple values in this window.

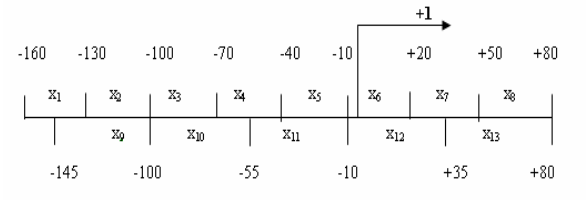

Fig.2 Over-lapping window feature variables (Michael Q. ZhangI, 1997)

2.3 Feature Variables of ImpRBFPromoter

Although feature variables obtained in the overlapping window above are effective, surprising computing data increases the executing time to predict TSS even for one sequence. Based on statistic analysis of all 5-tuple frequencies in each position of 120 sequences in a training set , it is found that 5-tuples appear in different regions with diverse frequencies. For promoter regions (240bp sequence samples in -160～+80) and non-promoter regions (1360bp sequence samples outside of promoter regions), some 5-tuples concentrate on appearing in promoter regions whereas seldom in non-promoter regions. On the contrary, some 5-tuples have extensively higher frequencies in non-promoter regions than in promoter regions.

In promoter regions of sequences in the training set, the 5-tuples whose appearance frequencies are in the top 20 are GGCGG, GGGCG, GCGGG, GGGGC, GGAGG, GCGGC, GGGAG, GCGCG, GCCGC,GGCTG,AGGGG, GCTGC, CGGGG, GAGGG, GCTAG, GGGGG, CCGCC, GGGCC, CTGGG, CGCGG. However, as TATA box is proven to be a typical signal of promoter, frequencies of all 5-tuples containing TATAX (X represents any of A, T, C, G) are not in the Top 20, actually in No.32, which may be caused by the statistical character of the training set.

In non-promoter regions of sequences in the training set, 5-tuples whose appearance frequencies are in the top 20 are TTTTT, GGAGG, TTTCT, GGGAG, CCTCC, GGGGG, CTGGG, GAGGG, ATTTT, AGAAA, GGCTG, TGGGG, CCTGG, TTTTA, GGGGC, CTCCC, CCCAG, GAGGA, CAAAA, TGGGA. Consequently, it assumes that different 5-tuples play different roles in distinguishing promoter region from non-promoter region. If 5-tuples could be chosen according to some relative function scores, an improved feature variables method could be developed based on over-lapping window as following:

①Building frequency table of all 5-tuples in each position of the 120 training sequences.

Firstly, 8520 short sequences of the length 240bp are selected, which include 120 true samples in the position between 640 and 880 and 8400 false samples in the position between 0 and 540 and between 980 and 1360, respectively, with the position of the first letter in the sequence marked 0.

If there is no repetitive element, a 5-tuple is a short sequence in which the character of every position can be any one of A, T, C and G. There are 4^5=1024 possible combinations. However, it is necessary to consider the affection of the repetitive element N. But it is impossible to judge the base pair elements of the 5-tuple NNNNN, so NNNNN is omitted without consideration. Thus it is essential to compute 5^5-1=3124 5-tuples frequencies in each position of the 8520 training samples.

② Getting statistic frequencies of all 5-tuples in promoter regions and non-promoter regions according to the frequency table;

③ Choosing a relative score function as following

$$relfunc(i) = \frac{fretra(i) - frefalse(i)}{1 + fretra(i) + frefalse(i)} \qquad (3)$$

where ***i*** represents 5-tuple, fretra(***i***) is the frequency of 5-tuple ***i*** in promoter regions and frefalse(***i***) is the frequency of 5-tuple ***i*** in non-promoter regions. The absolute value of the function for each 5-tuple is thus computed. Higher absolute value of the function shows more effect of the corresponding 5-tuple on distinguishing promoter region than on non-promoter region. As there are many choices of the relative function, the above function is chosen from several experiments to get the best recognition accuracy.

④ Ranking 5-tuples decreasingly according to the relative score absolute value of the function and choosing the first n 5-tuples as the most important short sequences. Here, n is chosen to be 300.

⑤ Eq. (1) is then applied. What is different is that only the ***n*** 5-tuples chosen in the fourth step are considered in the sequence window, the other short sequences are omitted.

⑥ Introducing the score of CpG islands

CpG islands are unmethylated regions of the genome that are associated with the 5' ends of most housekeeping genes and many regulated genes. The absence of methylation slows CpG decay, and so CpG islands can be detected. In fact, about 80% of CpG islands are common in man and mouse DNA sequences. Generally, CpG islands overlap the promoter and extend about 1000 base pairs downstream into the transcription unit. Identification of potential CpG islands during sequence analysis helps to define the extreme 5' ends of genes, something that is notoriously difficult with cDNA based approaches. Probably because they are associated with genes, CpG islands tend to be unique sequences and are therefore very useful in genome mapping projects.

There are mainly two rules to judge whether or not there are CpG islands in sequence windows, of which the first is whether the content of G+C is more than 50% and the second is whether CGratio which is defined in (4) is more than 60%.

$$CGratio = \frac{Obs}{Exp} \qquad (4)$$

where *Obs* is the frequency of CG.

and

$$Exp = \frac{appearing\ \ numbers\ \ of\ \ C \times appearing\ \ numbers\ \ of\ \ G}{window\ \ length}$$

If there are CpG islands in a sequence window, then the score of that window is added a constant value 0.1, which is only an experimental value. Indeed it is better to build up an independent CpG islands model to get that constant value.

2.2 Construction of a RBF neural network

Radial basis function (RBF) neural network is a feed forward network, but different from the conventional feed forward network in connections between the hidden and output layers, which is shown in Fig3. For an RBF network, a node in the hidden layer represents a unique prototype and an output node represents a unique category.

According to Fig 2, the input of the RBF neural network are feature variables of 13 dimensions in both RBFpromoter and impRBFpromoter. So there are 13 nodes in the input layer. The output layer has only one node, in which +1 represents true TSS and -1 represents false TSS. In this paper, Gaussian radius function is chose as the radial basis function. Let X_k be the input vector, m be the node number of the hidden layer, $MX_{(j)}$ be the connection weight of the jth hidden node, which is also the clustering center of the jth hidden node. The distributed parameter σ_j^2 represents data distributing condition in the hidden layer. Let $\boldsymbol{W_j}$ be the connection weight between the jth hidden node and the output node, the output of the jth hidden node be O_j and the output of network be O, then

$$O_j = e^{-\|X_k - MX_j\|^2 / 2\sigma_k^2} \qquad (5)$$

$$O = \sum_{j=1}^{m} W_j O_j \qquad (6)$$

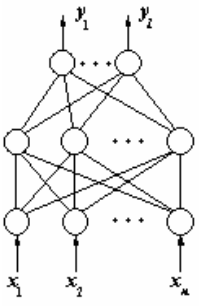

Fig3 strucuture of a RBF neural network

To improve the training efficiency of the hidden layer, an automatic clustering method is adopted as following.

Firstly, a clustering radius R_0 is set to be the product of the average of the minimum distance among training categories and a constant α as in (7).

$$R_0 = \alpha \frac{1}{P} \sum_{i=1}^{P} \min_{i \neq j} (\|x_i - x_j\|) \qquad (7)$$

where P is the number of training samples.

Then repetitive clustering is adopted. If a given training sample is in a region with the existing cluster radius R_0, then it is included in that cluster and the clustering center is adjusted simultaneously. But if the sample is not included in any existing region, a new cluster is then produced. The center of the new cluster is that training sample. Therefore, little α andR_0 always mean more clustering numbers.

Finally, it is necessary to get the width of the hidden node. If the distances of the clusters' centers are different, each node should select different width. The node which is much farther away from other centers should be given higher width, otherwise lower. In order to achieve this, it is essential to find out the distance of minimum clustering centers among different clusters.

Automatic clustering algorithm is very effective to construct hidden node numbers because it needs compute only once for all training samples to finish all clustering.

After the number and center of hidden nodes are obtained, LMS algorithm is adopted to solve this linear optimization of the linear equation units in order to train the weight between the hidden layer and the output layer. LMS is explained as followings: $\Delta W_{ki} = \eta\ \delta_k O_i$ in which η is the learning rate, O_i is the response output of the ith hidden node, ΔW_{ki} is the increment of the weight between the ith hidden node and the kth output node, δ_k is the error of the node k, $\delta_k = T_k - O_k$,where T_k is the ideal output value of the kth output node, O_k is the actual response output of the kth output node. In order to make sure the algorithm's constringency, let $\eta = p/(t+1)$, where $0<p<1$ and t is the iterative times.

3. TRAINING SET

In construction of the algorithm, training set is of specially crucial because all of the statistical data are extracted from the set and thus the set has a quite impact on the prediction accuracy of the algorithm. The choice of the training set is thus an important job,

manual sequencing error must be deleted, the range of any sequence in the training set should be representative enough, the number of sequences in the training set should be large enough to meet the requirements of the statistic, redundancy or homology presented within the data set must be reduced. As a result, firstly, a complete human promoter sequence set can be downloaded from EPD (Eukaryotic Promoter Database) at http://www.epd.isb-sib.ch, including 1796 sequences with the range from -499 to +100. Secondly, BLAST searching (GenBank, release 134) is adopted for each sequence in the original complete set to extend qualified of the 1796 sequences from -800 to 800, of which 153 sequences can be extended. Thirdly, RepeatMasker are adopted to those153 sequences to mark repeated segments. Finally, a sequence set is obtained, in which 120 consist of training set and 33 consist of verifying set.

4. RESULTS

In order to compare the system results thoroughly, a test set is considered which consists of 93 sequences, constructed by the following steps:

①choosing 30 sequences of the length 1600bp randomly in the training set, which is 32.258% of the whole test set;

②choosing 33 sequences of the length 1600bp randomly in the verifying set, which is 35.484% of the whole test set;

③choosing 30 sequences of the length 600bp randomly in the downloaded EPD sequences, which is 32.258% of the whole test set. All of test set sequences are also run by the Internet in the web server of CorePromoter (http://rulai.cshl.org/tools/genefinder/CPROMOTER/human.htm).The standard evaluation rules , such as the specificity, sensitivity and correlation coefficient (*CC*) are used to evaluate the prediction results. Only the highest score can be regarded as the right prediction result, the strict evaluation result is shown in Tab.1.

Tab.1 Strict evaluation result of the test set

	RBFPromoter	ImpRBFPromoter	CorePromoter
S_n	0.3226	0.3333	0.0322
S_p	0.3226	0.3333	0.0322
CC	0.3176	0.3176	0.0251

where $S_n = TP/(TP + FN)$ $S_p = TP/(TP + FP)$

$$CC = \frac{[(TP)(TN) - (FP)(FN)]}{\sqrt{(TP + FP)(TN + FN)(TP + FN)(TN + FP)}}$$

TP and *FP* are the numbers of true and false promoter predictions, respectively, *TN* and *FN* the numbers of true and false "non-promoter" predictions, respectively.

Aiming at evaluating the algorithm more completely, the common test set used by James W. Fickett and Artemis G. Hatzigeorgiou in the Reference (James. W. Fickett ,et al, 1997) is adopted to test CorePromoter, RBFPromoter and ImpRBFPromoter.

The evaluation results are shown in Tab2 . Of all 18 mammalian sequences in which the transcription initiation site have been experimentally mapped, two sequences have no identification marks and searching results of one sequence do not match with the reference. Thus only 15 sequences containing 19 promoters in a total of 27277bp are chosen. None of them matches a sequence in EPD (either at the level of identity or at the level of clear homology). The predicted TSS, explicit or implicit, was counted as correct if it was within 200bp 5', or 100bp 3', of any experimentally mapped TSS.

Tab 2 Test results of the common test set

Seq.	Audic	Autogene	GeneID	NNPP
Se	4/19 21%	6/19 32%	9/19 47%	11/19 58%
Sp	27fp 1/1010bp	48fp 1/568bp	40fp 1/682bp	67fp 1/407bp

Seq.	P'Fin-d	TATA	TSSG	TSSW
Se	5/19 26%	6/19 32%	5/19 26%	8/19 42%
Sp	25fp 1/1091bp	44fp 1/620bp	21fp 1/1299bp	36fp 1/758bp

Seq	CorePromoter	RBFPromoter	ImpRBFPromoter
Se	9/19 47%	14/19 74%	12/19 64%
Sp	57fp 1/479bp	119fp 1/229bp	122fp 1/224bp

For each program the Se (sensitivity, as the number and percentage of promoters correctly detected) and Sp (specificity, as number of false positives and number of base pairs per false positive) are given. Higher Sensitivity and smaller Specificity mean better recognition accuracy.

5. DISCUSSION

It is suggested by scientists that combining several models' results could improve prediction accuracy. But if artificial neural network could play a better role in predicting TSS, high accuracy of each single model and reliable training set should be crucial. It could be seen that even the famous CorePromoter prediction results for different test sets are diverse greatly, which reveals poor performance of promoter recognition as a result of relying extremely on the quality and quantity of the training set. This means that biological properties hiding behind original sequences are not reflected by the algorithm effectively. Another important problem is different signals in sequences. For example, although CpG islands are strongly associated with TSS, a factor that gives rise to experimental procedures for isolating promoters, it is still difficult to make full use of such a signal in the algorithm. Building up a special independent model for CpG islands would be helpful to improve the accuracy of the RBF neural network. At the mean time, it should be remembered that because of the limited sample size and the possibly skewed nature of the sample, results should be taken as provisional.

6. CONLUSIONS

A novel method of promoter recognition in the human genome is proposed in this paper where RBF neural network and an improved feature variable method are adopted and implemented in RBFPromoter and ImpRBFPromoter developed in VC++ 6.0. And the method is tested by different test sets. Test results reveal that compared with CorePromoter, RBFPromoter and ImpRBFPromoter have more flexible learning ability and higher TP predictions.

REFERENCES

Anders Gorm Pedersen, Pierre Baldi et al. 1999,The biology of eukaryotic promoter prediction — a review. Computers & Chemistry , 23：191-207

Bian Zhaoqi et.al., 2000， Pattern recognition, Beijing:Tsinghua University Press,

Chris Burge and Samuel Karlin. 1997a, Prediction of complete gene structures in human genomic DNA, J.Mol.Biol., 268,78-94

Christpher Burge 1997b, Identification of genes in human genomic DNA, Stanford University,

Dasgupta, N.; Lin, S.; Carin, L.. 2002, Sequential modeling for identifying CpG island locations in human genome. Signal Processing Letters. IEEE, 9 (12): 407～409

Hutchinson,G .B. 1996,The prediction if vertebrate promoter regions using differential hexamer frequency analysis,Corp.Appl.Bio.Sci. 12:391-398

James W. Fickett, 1996,The gene identification problem: an overview for developers Computers Chem, 20(1):103-118

James W. Fickett and Artemis G. Hatzigeorgiou. 1997, Eukaryotic Promoter Recognition enome Rearch, 1997, 7:861-878

Lampariello, F., Sciandrone, M.. 2001, Efficient training of RBF neural networks for pattern recognition. Neural Networks, IEEE Transactions on, 12 (5):1235 ～124

Larsen,F., Gundersen,G.,Lopez,R.and Prydz,H. 1992,CpG islands as gene markers in the human genome. Genomics, 13:1095～1107

Michael Q. Zhang, 1997,Identification of Human Gene Core Promoters in Silico ,Genome Research, 8:319-326

Moises Burset and Roderic Guigo. 1996,Evaluation of Gene Structure Prediction Programs,Genomics, 34(3):353-367

Roderic Guigo, 1997,Computational gene identification : an open problem. Computers Chem, 21(4), 215-222

Sardo, L.; Kittler, J. 1996,Complexity analysis of RBF networks for pattern recognition 。 Computer Vision and Pattern Recognition,. Proceedings CVPR '96, 1996 IEEE Computer Society Conference on,:574～579

S.F.Altschul,W.Gish,W.Miller,E.W.Myers and D.J.Lipman, 1990,Basic local alignment search tool, J.Mol.Biol., 215:403-410

S.F.Altschul et al. 1997,Gapped BLAST and PSI-BLAST: a new generation of protein database search program,Nucleic Acids Res., 25:3389-3402

Tao Jiang， YingXu， Michael Q.Zhang. 2002, Current Topics in Computational Molecular Biology. Beijing: Tsinghua University press，

Uwe Ohler and Heinrich Niemann. 2001, Identification and analysis of eukaryotic promoters: recent computational approaches. Trends in Genetics, 17(2), 56-60

Ying Xu et al., 1996, GRAIL:A Multi-Agent Neural Network System for Gene Identification, Proceedings of the IEEE, 84(10):1544～1552