Journal of Zhejiang University SCIENCE
ISSN 1009-3095
http://www.zju.edu.cn/jzus
E-mail: jzus@zju.edu.cn

A hybrid neural network system for prediction and recognition of promoter regions in human genome*

CHEN Chuan-bo (陈传波), LI Tao (李 滔)

(*School of Computer Science & Technology, Huazhong University of Science & Technology, Wuhan 430074, China*)

E-mail: chuanboc@163.com; ljrlt@public.wh.hb.cn

Received Oct. 8, 2004; revision accepted Mar. 7, 2005

Abstract: This paper proposes a high specificity and sensitivity algorithm called PromPredictor for recognizing promoter regions in the human genome. PromPredictor extracts compositional features and CpG islands information from genomic sequence, feeding these features as input for a hybrid neural network system (HNN) and then applies the HNN for prediction. It combines a novel promoter recognition model, coding theory, feature selection and dimensionality reduction with machine learning algorithm. Evaluation on Human chromosome 22 was ~66% in sensitivity and ~48% in specificity. Comparison with two other systems revealed that our method had superior sensitivity and specificity in predicting promoter regions. PromPredictor is written in MATLAB and requires Matlab to run. PromPredictor is freely available at http://www.whtelecom.com/Prompredictor.htm.

Key words: Hybrid neural network, Promoter prediction, Compositional features, CpG islands
doi:10.1631/jzus.2005.B0401 **Document code:** A **CLC number:** Q78

INTRODUCTION

The publication and preliminary analysis of the human genome sequence (Lander *et al.*, 2001; Venter *et al*., 2001) marks a significant milestone in the field of molecular biology. One of the main goals of the Human Genome Project is the characterization, annotation−recognition and categorization of genes from human genome to serve as a periodic table for biomedical research (Lander, 1996). In the past few years, many efforts have been devoted to gene annotations. The National Center for Biotechnology Information (NCBI), Ensembl and Golden Path, for instance, provided the initial annotations, but the whole process of annotation is expected to go on for many years, and the current gene annotations only refer to protein-coding regions, relatively few tools have been developed to identify the regulatory regions required for the correct transcriptional activity of the genome. This task is particularly difficult in the case of eukaryotic organisms in which regulatory regions represent a small percentage, overwhelmed by presumably non-functional DNA. So prediction and characterization of regulatory regions is still a challenging problem. Here, we focus on detecting promoters, which are in the class of regulatory regions.

A promoter is the region of genomic sequence proximal to the transcription start site (TSS) responsible for the initiation of transcription. In spite of the fact that characterizing regulation of gene expression is an important aspect of understanding gene function, for most human genes, promoters have not been defined or studied. Therefore, reliable recognition and characterization of promoters is a high priority goal in human genome study. Knowledge of promoters will be useful in elucidating regulation and expression mechanisms of genes and may shed light on the function of novel and uncharacterized genes.

A well-established measure for promoter prediction accuracy scores a TSS prediction as positive if

* Project (No. 2001AA231071) supported by the Hi-Tech Research and Development Program (863) of China

it is within the range of 200 bp upstream to 100 bp downstream of the true TSS (Fickett and Hatzigeorgiou, 1997). Several research groups have developed methods for in silico promoter prediction, including knowledge-based methods, comparative genome analysis as well as methods based on statistical-compositional properties of DNA sequences (Fickett and Hatzigeorgiou, 1997; Ohler and Niemann, 2001). For most methods, the false-positive rate is roughly estimated at one per kilobase. In another aspect, the ratio of true prediction to false prediction is a small percent, with the exception of one method, PromoterInspestor, which shows predicted accuracy of 43% (Scherf *et al*., 2000). In recent years, many efforts have been devoted to improve promoter predicted accuracy by using CpG islands association (Ioshikhes and Zhang, 2000; Davuluri *et al*., 2001; Hannenhalli and Levy, 2001; Ponger and Mouchiroud, 2002), combination with exon/intron/3′-UTR predictions (Bajic *et al*., 2002; 2003) and consensus promoter identification that combines several existing methods (Liu and David, 2002).

Motivated by these methods, we developed a new hybrid neural network system−the PromPredictor for human genome promoter recognition. It is a combination of a novel promoter recognition model, coding theory, feature selection and dimensionality reduction with machine learning algorithm. The method is based on the statistical concept of pentamer distributions in specific functional regions of DNA and selected the most significant pentamer vocabularies from training sequences by an unsupervised learning technique, in addition to CpG islands features.

FEATURE EXTRACTION FROM DNA SEQUENCE

From a one-dimensional point of view, a DNA sequence contains characters from the 4-letter nucleic acid alphabet α={A, C, T, G}. An important issue in applying neural networks to promoter classification is how to encode DNA sequences, i.e., how to represent the DNA sequences as the input of the neural networks. In fact, sequences may not be the best representation at all. Good input representations make it easier for the neural networks to recognize underlying regularities. Therefore, good input representations are crucial to the success of neural network learning (Hirsh and Noordewier, 1994).

Compositional features

It is well known that genomes are characterized by species-specific compositional features, and that coding and non-coding DNA are distinguishable in terms of their pentamer and hexamer distributions (Claverie *et al*., 1990). In promoter regions except core promoter elements such as TATA boxes, CAAT boxes and transcription initiation sites (INR), there exists a couple of other individual elements or sequence properties that are associated with promoter sequences. Among these are higher CpG content−CpG islands (Shago and Giguere, 1996), secondary structure elements like the HIV-1 TAR regions (Bohjanen *et al*., 1997), cruciform DNA structures (Wang *et al*., 1998), or simple direct repeats (Bell *et al*., 1997). Three-dimensional structures, such as curved DNA (Kim *et al*., 1995), also influence promoter function. Most of these elements can be detected by computer-assisted sequence analysis (Chetouani *et al*., 1997; Schuster *et al*., 1997; Nakaya *et al*., 1995; Nielsen *et al*., 1995), but none of them is really promoter specific and can be found frequently outside of promoters. The secret of promoter function lies in the combination of several promoter elements that must cooperate in transcriptional activation, while none of them can achieve alone. This also summarizes the main problem of promoter recognition. It is necessary to compile several individual weak signals into a composite signal which then indicates a potential promoter. As an attempt, we use pentamer frequency coding method to capture core elements as well as weak signals.

The pentamer encoding method extracts various patterns of five consecutive nucleic acids in a DNA sequence and counts the number of occurrences of the extracted pentamer. For instance, given a DNA sequence CGAATCG, the pentamer encoding method gives the following results: 1 for CGAAT (indicating CGAAT occurs once), 1 for GAATC and 1 for AATCG. For each DNA sequence, there are 4^5=1024 possible pentamers in total.

If all the 1024 pentamers were chosen as input features of the neural network, it would require many weight parameters and training data, which makes it

difficult to train the neural network−a phenomenon called "curse of dimensionality". Different methods have been proposed to solve the problem by careful feature selection and by scaling of the input dimensionality (Chuzhanova *et al*., 1998). What we are proposing here is to select relevant features by employing a distance measure to calculate the relevance of each feature (Dash and Liu, 1997).

Let X be a feature and x be its value. Let $p(x|Class=1)$ and $p(x|Class=0)$ denote the class conditional density functions for feature X, where *Class*_1 represents the positive class and *Class*_0 is the negative class. Let $D(X)$ denote the distance function between $p(x|Class=1)$ and $p(x|Class=0)$, defined as (Bassat, 1982)

$$D(X)=\int \left| p(x\,|\,Class=1)-p(x\,|\,Class=0)\right| \mathrm{d}x \quad (1)$$

The distance measure prefers feature X to feature Y if $D(X)>D(Y)$. Intuitively, this means that it is easier to distinguish between *Class*_1 and *Class*_0 by observing feature X than by observing feature Y. That is, X appears often in *Class*_1 but seldom in *Class*_0 or vice versa. In our work, each feature X is a pentamer. Let c denote the occurrence number of the feature X in a sequence S. Let l denote the total number of pentamers in S and $len(S)$ represents the length of S. We have $l=len(S)-4$. Define the feature value x for the pentamer X with respect to the sequence S as

$$x=\frac{c}{len(S)-4} \quad (2)$$

Since a promoter sequence may be short, random pairings can have a large effect on the result. $D(X)$ in Eq.(1) can be approximated by the Mahalonobis distance (Solovyev and Makarova, 1993) as

$$D(X)=\frac{(m_1-m_0)^2}{d_1^2+d_0^2} \quad (3)$$

where m_1 and d_1 (m_0 and d_0, respectively) are the mean value and the standard deviation of the feature X in the positive (negative, respectively) training data set. Intuitively, in Eq.(3), the larger the numerator is (or the smaller the denominator is), the larger the interclass distance is, and therefore the easier to separate *Class*_1 from *Class*_0 (and vice versa). The mean value m and the standard deviation d of the feature X in a set X of sequences are defined as

$$m=\frac{1}{N}\sum_{i=1}^{N}x_i \quad (4)$$

$$d=\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i-m)^2} \quad (5)$$

where x_i is the value of the feature X with respect to sequence $S_i \in X$, and N is the total number of sequences in X.

Let $X_1, X_2, \ldots X_{Na}$, $N_a<1024$, be the top N_a features (pentamers) with the largest $D(X)$ values. Intuitively, these N_a features occur more frequently in the positive training data set and less frequently in the negative training data set. For each DNA sequence S (whether it is a training or an unlabeled test sequence), we examine the N_a features in S, calculate their values as defined in Eq.(2), and use the N_a feature values as input feature values to the HNN for the sequence S.

To compensate for the possible loss of information due to ignoring the other pentamers, a linear correlation coefficient (*LCC*) between the values of the 1024 pentamers with respect to the DNA sequence S and the mean value of the 1024 pentamers in the positive training data set is calculated and used as another input feature value for S. Specifically, the *LCC* of S is defined as

$$LCC(S)=\frac{1024\sum_{i=1}^{1024}x_i\overline{x_i}-\sum_{i=1}^{1024}x_i\sum_{i=1}^{1024}\overline{x_i}}{\sqrt{1024\sum_{i=1}^{1024}x_i^2-\left(\sum_{i}^{1024}x_i\right)^2}\sqrt{1024\sum_{i=1}^{1024}\overline{x_i}^2-\left(\sum_{i}^{1024}\overline{x_i}\right)^2}} \quad (6)$$

where $\overline{x_i}$ is the mean value of the ith pentamer, $1<i<1024$, in the positive training dataset, and x_i is the feature value of the ith pentamer with respect to S as defined in Eq.(2).

CpG islands features

In the human genome, many genes were recognized and validated successfully (Lander *et al*., 2001; Venter *et al*., 2001) by using the so-called CpG islands as gene markers. CpG islands are unmethylated segments of DNA longer than 200 bp, with a G+C content of at least 50%, and the number of CpG di-

nucleotides being at least 60% of what could be expected from the G+C content of the segment (Bird *et al*., 1987; Gardiner and Frommer, 1987; Larsen *et al*., 1992; Cross and Bird, 1995). CpG islands are found around a gene that starts in approximately half of mammalian promoters (Larsen *et al*., 1992; Cross and Bird, 1995) and are estimated to be associated with ~60% of human promoters (Cross *et al*., 1999). For this reason, Pedersen *et al.*(1999) suggested that CpG islands could represent a good global signal to locate promoters across genomes. At least in mammalian genomes, CpG islands are good indicators of gene presence. In our prediction system, we use two CpG island features−G+C content and ratio of expected to observed CG dinucleotides (i.e. Obs/Exp). Let *len* represent the length of one segment of a DNA sequence, the G+C content (*GC_con*) and Obs/Exp (*o*/*e*) (Gardiner and Frommer, 1987) are defined as

$$GC\_con = \frac{\text{number of C's} + \text{number of G's}}{len} \quad (7)$$

$$o/e = \frac{\text{number of CG's} \times len}{(\text{number of C's} \times \text{number of G's})} \quad (8)$$

ARCHITECTURE OF THE PREDICTION SYSTEM

The conceptual structure of our system is depicted in Figs.1 and 2. The overall system shown in Fig.1 comprises a collection of four basic classifiers: promoter_classifier, exon_classifier, intron_classifier and 3′-UTR_classifier. Each of the classifiers is a modified BP neural network and has the same structure. The basic classifier of promoter is shown in Fig.2. Each classifier is trained by different training sets and the parameters for each classifier are optimized independently.

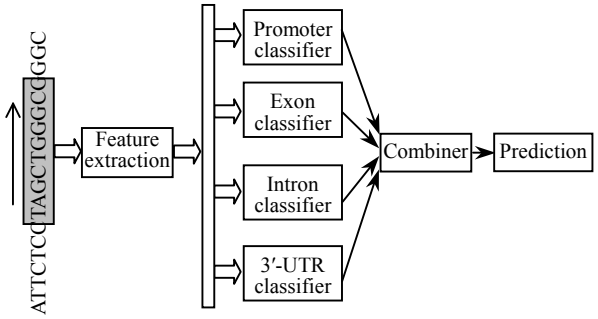

Fig.1 Overall structure of the PromPredictor

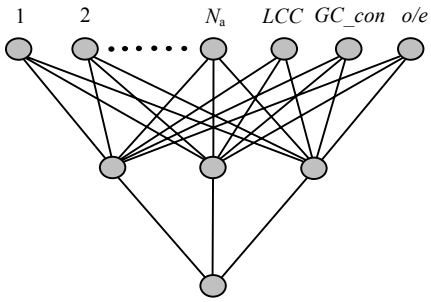

Fig.2 Promoter_classifier

An unknown sequence is partitioned into windows 250 bp long, shifted by 1 bp. For each sliding data window, we compute the feature values following procedures in the previous section and these feature values are used as input of the neural network.

The prediction system assigns the sequence to the class promoter if three classifiers−exon, intron, 3′-UTR decide that the sequence is not an exon, intron, 3′-UTR respectively, and only promoter_classifier decides that the sequence belongs to this class.

Sequence training and test set

From the vertebrate section of the Eukaryotic Promoter Database (EPD), V.79.0 (Cavin *et al*., 1998), promoter sequences from 200 bp upstream to 50 bp downstream of the TSS were taken. Vertebrate exon and vertebrate intron sequences of different location, covering a total of 5×10^6 bp in each set were randomly extracted from GenBank. Vertebrate 3′-UTR sequences with a total of 3×10^6 bp were extracted from the UTR database (Pesole *et al*., 2002). Sequences training set for the four basic classifiers (promoter, exon, intron and 3′-UTR) were created by randomly extracting non-overlapping sequences of 250 bp from the four example sets mentioned above. Redundant sequences were deleted by the program CLEANUP (Grillo *et al*., 1996) which resulted in sets consisting of 1837 sequences from promoter regions, 4500 exon sequences, 6500 intron sequences and 5000 3′-UTR sequences. In these four sets, 2/3 of the sequences were used for training, and the rest were used for validation.

System training and parameter optimization

According to the definitions above, the number of N_a, the neuron number of hidden layers and the training algorithm must be determined for each classifier. Furthermore, an optimal assignment threshold must be calculated.

The training for each classifier is independent. For example, the positive training set for exon_ classifier is exon, and the negative training set includes intron, promoter and 3′-UTR. We tested the exon_classifier's performance with different parameters and different training algorithms and recorded the optimized parameters based on accuracy and computer time. The training algorithms include Gradient descent algorithms, Conjugate gradient (CGB) algorithms (Powell, 1977), Quasi-Newton algorithm, One Step Sccant (OSS) algorithm (Battiti, 1992), Resilient backpropagation (RPROP) algorithm (Riedmiller and Braun, 1993) and Levenberg-Marquardt (LM) algorithm (Hagan and Menhaj, 1994). Table 1 summarizes the default threshold, corresponding training algorithm and the optimized parameters for four basic classifiers. Table 2 shows the validation results. From this table, we can see that the combination of four classifiers can improve the sensitivity in predicting promoter regions obviously.

EXPERIMENTS AND RESULTS

In order to compare the performance of our system with two other promoter recognition systems: PromoterInspestor (Scherf *et al*., 2000) and Dragon Promoter Finder (DPF) (Bajic *et al*., 2002; 2003), we used the same evaluation set in PromoterInspestor and DPF. The first set (SET 1) consisted of six Genbank genomic sequences with a total length of 1.38 Mb and 35 known TSSs on these sequences. The second set (SET 2) consisted of publicly available sequence for human chromosome 22−the 35 Mb sequence with 377 known genes. The annotation data (Rel. 2.3) for human chromosome 22 were produced by the Chromosome 22 Gene Annotation Group at the Sanger Centre and were obtained from the world wide web (http://www.sanger.ac.uk/HGP/Chr22/).

For the former dataset a promoter region was counted as true positive (TP), if a transcription start site (TSS) was located within or up to 200 bp downstream of the predicted promoter region. Or otherwise the promoter region was counted as false positive (FP). For the latter dataset we used the same method as that used by Scherf *et al*.(2000) with PromoterInspector: all the predictions located in the range −2000~+500 around the 5′ extremity of a known gene were considered as a true positive promoter region (TP). While the FP predictions were considered as those that fall on the annotated part of the human chromosome 22 covered by known genes, but not sufficiently close to the 5′ end of these genes, thus not representing the TP predictions.

The main results and comparisons are summarized in Tables 3−6. Results are given with respect

Table 1 Parameters and default threshold for four basic classifiers

Basic classifiers	N_a	Neuron numbers in hidden layer	Training algorithm	Default threshold
Promoter_classifier	900	3	Levenberg-Marquardt algorithm	0.9368
Exon_classifier	800	2	Resilient backpropagation	0.9470
Intron_classifier	800	2	Resilient backpropagation	0.9353
3′-UTR _classifier	800	2	Resilient backpropagation	0.9170

Table 2 Results of the four classifiers and PromPredictor on validation sequences

Sets	$N^{\#}$	Number of predicted sequences above threshold[a]				
		Promoter_classifier (threshold=0.8)	Exon_classifier (threshold=0.8)	Intron_classifier (threshold=0.8)	3′-UTR_classifier (threshold=0.8)	PromPredictor[b]
Promoter	612	484	20	74	81	422
Exon	1500	27	1485	207	214	11
Intron	2166	85	678	2144	912	19
3′-UTR	1666	277	452	523	1489	156
Sensitivity: S_e (%)		55.4[c]				69.4[d]

[#]Number of sequences; [a]One class of classifier assigns an unknown sequence to this class when the predicted value is above threshold; [b]PromPredictor assigns an unknown sequence to the class promoter if three classifiers−exon, intron, 3′-UTR decide that the sequence is not an exon, intron, 3′-UTR respectively, and only promoter_classifier decides that the sequence belongs to this class; [c]55.4%=484/(484+27+85+277); [d]69.4%=422/(422+11+19+156)

to several criteria related to the maximum allowed distance between the predicted promoter region and the real TSS. In these experiments, PromPredictor, DPF and PromoterInspector were used with their default parameter settings.

Table 3 The results of promoter prediction by PromPredictor on SET 1

Sequence accession number	Length (bp)	Number of TSS	*TP*	*FP*	Coverage* (%)
AC002397	227538	17	6	9	35
L44140	219447	11	8	26	73
D87675	301692	1	1	2	100
AF017257	101569	1	1	1	100
AF146793	204625	4	1	4	25
AC002368	324816	1	1	1	100

*The coverage describes the percentage of true promoters in a sequence that has been predicted by PromPredictor

Table 4 Comparison of three prediction systems on SET 1

Program	*TP*	*FP*	S_e (%)[a]	S_p (%)[b]	*CC* (%)[c]
PromoterInspector	15	55[d]	42.8	21.4	30.3
DPF(S_e=0.37)	16	79	45.7	16.8	27.7
PromPredictor	18	43	51.4	29.5	38.9

[a]Sensitivity: $S_e=TP/(TP+FN)$; [b]Specificity: $S_p=TP/(TP+FP)$;
[c]Correlation Coefficient (*CC*):

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

FN: false negative; *TN*: true negative
[d]The number of FP predictions on SET 1 made by PromoterInspector is 55 by our criteria because strand-nonspecific predictions were counted twice (once for each DNA strand)

Table 5 The results of promoter prediction by PromPredictor on SET 2

SET 2 performance	PromPredictor
TP	248
FP	274
Total number of human gene	377
Total number of predictions*	869
The average length of predicted promoter region	568

*Represents the total number of predictions in the two-strand search and with strand-specific counting of predictions

Table 6 Comparison of three prediction systems on SET 2

Program	S_e (%)	S_p (%)
PromoterInspector	45	33
DPF	64	33
PromPredictor	66	48

CONCLUSION

In this paper we presented a hybrid neural network approach−PromPredictor for predicting promoter regions in large genomic sequences. It is based on a new promoter model with statistical-compositional features and CpG information and integrates multi-classifier via HNN. The prediction accuracy of PromPredictor, achieved on a large and diverse evaluation-set, shows our novel method is promising for modeling biological systems in general, which does not require any specific knowledge about a particular promoter to make a prediction and thus has a big advantage especially when nothing is known about the promoter to be predicted.

Future research should include combination of other groups of signals that may characterize some aspects of gene structure, such as translation initiation site, splice sites or polyA site.

References

Bajic, V.B., Seah, S.H., Chong, A., Zhang, G., Koh, J.L.Y., Brusic, V., 2002. Dragon Promoter Finder: recognition of vertebrate RNA Polymerase II promoters. *Bioinformatics*, **18**:198-199.

Bajic, V.B., Seah, S.H., Chong, A., Krishnan, S.P.T., Koh, J.L.Y., Brusic, V., 2003. Computer model for recognition of functional transcription start sites in RNA polymerase II promoter of vertebrates. *Journal of Molecular Graphic and Modeling*, **21**:323-332.

Bassat, M.B., 1982. Use of Distance Measures, Information Measures and Error Bounds in Feature Evaluation. *In*: Krishnaiah, P.R., Kanal, L.N. (Eds.), Classification, Pattern Recognition and Reduction of Dimensionality: Handbook of Statistics. Volume 2, North-Holland Publishing Company, Amsterdam, p.773-791.

Battiti, R., 1992. First and second order methods for learning: Between steepest descent and Newton's method. *Neural Computation*, **4**(2):141-166.

Bell, P.J.L., Higgins, V.J., Dawes, I.W., Bissinger, P.H., 1997. Tandemly repeated 147 bp elements cause structural and functional variation in divergent MAL promoters of Saccharomyces cerevisiae. *Yeast*, **13**:1135-1144.

Bird, A.P., Taggart, M.H., Nicholls, R.D., Higgs, D.R., 1987. Non-methylated CpG-rich islands at the human α-globin locus: Implications for evolution of the α-globin pseudogene. *EMBO J*, **6**:999-1004.

Bohjanen, P.R., Liu, Y., GarciaBlanco, M.A., 1997. TAR RNA decoys inhibit Tat-activated HIV-1 transcription after preinitiation complex formation. *Nucleic Acids Res*., **25**:4481-4486.

Cavin, P.R., Junier, T., Bucher, P., 1998. The Eukaryotic Promoter Database EPD. *Nucleic Acids Res*., **26**:353-357.

Chetouani, F., Monestié, P., Thébault, P., Gaspin, C., Michot, B., 1997. ESSA: an integrated and interactive computer tool for analyzing RNA secondary structure. *Nucleic Acids Res*., **25**:3514-3522.

Chuzhanova, N.A., Jones, A.J., Margetts, S., 1998. Feature selection for genetic sequence classification. *Bioinformatics*, **14**:139-143.

Claverie, J.M., Sauvaget, I., Bougueleret, L., 1990. K-tuple frequency analysis from intron/exon discrimination to Tcell epitope mapping. *Methods Enzimol*., **183**:237-252.

Cross, S.H., Bird, A.P., 1995. CpG islands and genes. *Curr. Opin. Genet. Dev*., **5**:309-314.

Cross, S.H., Clark, V.H., Bird, A.P., 1999. Isolation of CpG islands from large genomic clones. *Nucleic Acids Res*., **27**:2099-2107.

Dash, M., Liu, H., 1997. Feature selection for classification. *Intelligent Data Analysis*, **3**:1-6.

Davuluri, R.V., Grosse, I., Zhang, M.Q., 2001. Computational identification of promoters and first exons in the human genome. *Nature Genetics*, **29**:412-417.

Fickett, J.W., Hatzigeorgiou, A.G., 1997. Eukaryotic promoter recognition. *Genome Res*., **7**:861-878.

Gardiner, G.M., Frommer, M., 1987. CpG islands in vertebrate genomes. *J. Mol. Biol*., **196**:261-282.

Grillo, G., Attimonelli, M., Liuni, S., Pesole, G., 1996. CLEANUP: a fast computer program for removing redundancies from nucleotide sequence databases. *Comput. Applic. Biosci*., **12**:1-8.

Hagan, M.T., Menhaj, M., 1994. Training feedforward networks with the Marquardt algorithm. *IEEE Transactions on Neural Networks*, **5**(6):989-993.

Hannenhalli, S., Levy, S., 2001. Promoter prediction in the human genome. *Bioinformatics*, **17**:90-96.

Hirsh, H., Noordewier, M., 1994. Using Background Knowledge to Improve Inductive Learning of DNA Sequences. Proceedings of the Tenth Annual Conference on Artificial Intelligence for Applications. San Antonio, p.351-357.

Ioshikhes, I.P., Zhang, M.Q., 2000. Large-scale human promoter mapping using CpG islands. *Nature Genetics*, **26**:61-63.

Kim, J., Klooster, S., Shapiro, D.J., 1995. Intrinsically bent DNA in a eukaryotic transcription factor recognition sequence potentiates transcription activation. *J Biol. Chem*., **270**:1282-1288.

Lander, E.S., 1996. The new genomics: global views of biology. *Science*, **274**:536-539.

Lander, E.S. Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., Fitzhugh, W., 2001. Initial sequencing and analysis of the human genome. *Nature*, **409**:860-921.

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

Liu, R.X., David, J., 2002. Consensus promoter identification in the human genome utilizing expressed gene markers and gene modeling. *Genome Res*., **3**:462-469.

Nakaya, A., Yamamoto, K., Yonezawa, A., 1995. RNA secondary structure prediction using highly parallel computers. *Comp Appl Biosci*., **11**:685-692.

Nielsen, D.A., Novoradovsky, A., Goldman, D., 1995. SSCP primer design based on single-strand DNA structure predicted by a DNA folding program. *Nucleic Acids Res*., **23**:2287-2291.

Ohler, U., Niemann, H., 2001. Identification and analysis of eukaryotic promoters: recent computational approaches. *TRENDS Genet*., **17**:56-60.

Pedersen, A.G., Baldi, P., Chauvin, Y., Brunak, S., 1999. The biology of eukaryotic promoter prediction–A review. *Comput. Chem*., **23**:191-207.

Pesole, G., Liuni, S., Grillo, G., Licciulli, F., Mignone, F., Gissi, C., Saccone, C., 2002. UTRdb and UTRsite: specialized database of sequences and functional elements of 5' and 3' untranslated regions of eukaryotic mRNAs. Update 2002. *Nucleic Acids Res*., **30**:335-340.

Ponger, L., Mouchiroud, D., 2002. CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences. *Bioinformatics*, **18**: 631-633.

Powell, M.J.D., 1977. Restart procedures for the conjugate gradient method. *Mathematical Programming*, **12**: 241-254.

Riedmiller, M., Braun, H., 1993. A Direct Adaptive Method for Faster Backpropagation Learning: The RPROP Algorithm. Proceedings of the IEEE International Conference on Neural Networks, San Francisco.

Scherf, M., Klingenhoff, A., Werner, T., 2000. Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach. *J. Mol. Biol*., **297**:599-606.

Schuster, P., Stadler, P.F., Renner, A., 1997. RNA structures and folding: from conventional to new issues in structure predictions. *Curr. Opin. Struct. Biol*., **7**:229-235.

Shago, M., Giguere, V., 1996. Isolation of a novel retinoic acid-responsive gene by selection of genomic fragments derived from CpG-island enriched DNA. *Mol. Cell Biol*., **16**:4337-4348.

Solovyev, V.V., Makarova, K.S., 1993. A novel method of protein sequence classification based on oligopeptide frequency analysis and its application to search for functional sites and to domain localization. *Computer Applications in the Biosciences*, **9**(1):17-24.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., 2001. The sequence of the human genome. *Science*, **291**:1304-1351.

Wang, W.D., Chi, T.H., Xue, Y.T., Zhou, S., Kuo, A., 1998. Architectural DNA binding by a high-mobility-group/kinesin-like subunit in mammalian SWI/SNF-related complexes. *Proc. Natl. Acad. Sci. USA*, **95**:492-498.