USING SIGNAL PROCESSING TECHNIQUES IN PROMOTER PREDICTION

ZHANG, XUEJUAN
(B.Eng (Hons),BTU)

A THESIS SUBMITTED
FOR THE DEGREE OF MASTER OF ENGINEERING
DEPARTMENT OF ELECTRICAL&COMPUTER ENGINEERING
NATIONAL UNIVERSITY OF SINGAPORE

2005

Acknowledgements

I would first like to thank my supervisor, Professor Vladimir B. Bajic, for accepting me as his project student even though my prior research was not in the field of Bioinformatics. Prof Bajic's persistent encouragement, enthusiasm, and enlightenment make the course of work a rewarding and pleasant one.

I would also like to extend my thanks to many of my good friends and colleagues in i2r, for their help and encouragement in the course of my work.

Table of Contents

Summary

Promoter prediction is currently an important problem in the field of bioinformatics, since the problem of gene discovery, gene annotation, regulatory elements identification and transcriptional control is related to promoters. Digital Signal Processing (DSP) techniques have not been largely used for promoter prediction. Our project is to develop a new promoter prediction system based on DSP techniques. Systematic simulation studies are done regarding the suitability of possible DSP techniques such as Correlation Coefficient (CC) and the domain transforms of Discrete Fourier Transform (DFT), Discrete Cosine Transform (DCT), and Discrete Wavelet Transform (DWT). From these experiments, it is concluded that CC is not a feature that is generalized well for accurate classification. More suitable features, which include the coefficients of DFT, DCT, and DWT transforms of the original signal, were adopted and experiments were performed to select the optimal combination of features and classifier model for different promoter groups split by the GC-content. The performance of different combinations was systematically evaluated. Several general conclusions are made: the capability to recognize promoters reduces with the reduction of GC-content of the data; there are no significant differences in the prediction performance when any of the three transform is applied; and the best performance is achieved by combining all the three transforms. We finally draw the conclusion that the application of domain transforms is promising in predicting promoters. A system is developed, which incorporates signal pre-processing, feature extraction, system optimization, and promoter

recognition with performance assessment. The prediction system was applied to the plus strand of human chromosome 22 (NCBI Built 35). Performance evaluation was done for several gene categories that are taken from gene annotation. Comparison was made with the results for the six different categories of genes. We have examined how to combine possible features extracted under domain transforms in DSP field with biological features of promoters and non-promoters such as the number of CpG dinucleotides, GC-content and the number of different combinations of mono-.di-, tri- nucleotides. This slightly shows that the use of the three domain transforms for predicting human promoters should be combined with more of other appropriate biological features to achieve better prediction results. In the process of development of prediction system it is useful to reduce the number of features. The reduction of features has to be done on a case-to-case basis. Moreover, with the suitability of the DSP techniques such as the three domain transforms of DFT, DCT, and DWT to provide good features that work efficiently with biological features to enhance promoter prediction, future studies that involve applying other DSP techniques might also be done to further contribute to promoter prediction.

List of Tables

List of Figures

List of Symbols

$f(t)$	A signal in time domain
$F(s)$	The Continuous Fourier Transform of the signal $f(t)$
F^{-1}	The Inverse Continuous Fourier Transform
f_i	The i-th point of discrete form of signal $f(t)$
F_n	The n-point Discrete Fourier Transform
$w_{n,i}$	The exponential function item in the matrix W of n*i
$A(i)$	A signal in the spatial domain
$B(k_1)$	A signal in the frequency domain after Discrete Cosine Transform

List of Abbreviations

CC Correlation Coefficient

DCT Discrete Cosine Transform

DFT Discrete Fourier Transform

DSP Digital Signal Processing

DWT Discrete Wavelet Transform

EIIP Electron–ion Interaction Potential

FN False Negative

FP False Positive

PPV Positive Predictive Value

SE Sensitivity

SVM Support Vector Machine

TN True Negative

TP True Positive

TSS Transcription Start Site

Chapter 1

Introduction

Bioinformatics is a new field which combines information technology and biological science. Bioinformatics uses various disciplines such as statistics, pattern recognition, data mining, machine learning, artificial intelligence, biology, medicine and chemistry. The aim of bioinformatics is to try to use computational techniques to narrow down the candidate tests that need to be done by wet lab experiment that is time consuming and expensive. Bioinformaticians intend to use data processing techniques to extract useful and valuable knowledge from biological data and aid wet lab experiment more effectively.

Earlier, only small amounts of biological experimental data were available for studies. However, with the advent of genomics and proteomics, greater amounts of experimental data have become available. This makes the use of bioinformatics necessary. Proper application of bioinformatics techniques may lead to extraction of useful information effectively from gene and protein data.

Promoter prediction is one of the important problems in the field of bioinformatics. Promoters could be related to the problems of gene discovery, gene annotation, regulatory element identification and transcriptional control. Promoter is defined as the region that contains necessary DNA elements to initiate transcription of a gene. In general, promoters are located in the immediate upstream region of the gene, containing TSS (Transcription Start Site). Thus, accurately locating TSS becomes the critical step in promoter prediction. Currently, mapping EST (expressed sequence tags) fragments or most 5'-part of cDNA to genomic DNA is an effective method to identify TSS along genomic DNA. However, due to lack of complete sets of 5' end cDNA sequences, it becomes difficult to identify accurate promoter regions. Many different promoter prediction systems have been developed in the past but none of them have given a satisfactory solution.

Signal processing techniques have not been largely used in bioinformatics. For promoter prediction, digital signal processing is not broadly used as a core methodology. The goal of this project is to develop a new promoter prediction system based on digital signal processing techniques. For this purpose, we will use Discrete Fourier Transform (DFT), Discrete Cosine Transform (DCT), and Discrete Wavelet Transform (DWT).

1.1 Biological Background

Prediction of eukaryotic promoters by computational means is one of the most challenging problems in biological sequence analysis today. In the section that follows, I will present biological background necessary to understand the problem that will be studied.

A genome is the set of complete genetic information inherited from the parents. A genome comprises all the genes and is contained in nuclei of cells of a eukaryotic organism. The genome is physically present in the form of DNA (Deoxyribose Nucleic Acid), which is a polymer. The basic unit of the DNA is a nucleotide comprising sugar-phosphate backbone and one of the four bases: A (adenine), C (cytosine), G (guanine) and T (thymine). Only 2-5% of the human DNA sequences are coding sequences, which contain information used for synthesis of proteins, while the other parts represent non-coding sequences. Promoter belongs to non-coding sequences [Levitsky *et al.* , 2001].

The production of proteins involves two stages, namely transcription and translation. In transcription, a gene is copied base by base into RNA, specifically A to U, C to G, T to A and G to C. mRNA refers to "messager RNA". DNA is transcribed into RNA which is later converted into messager RNA during the RNA processing. In translation, a polypeptide (protein) is synthesized under the direction of mRNA.

Gene expression is the process when the information contained in a gene is converted into a cellular product. Gene expression process can be controlled at many levels, most significant level being the gene transcription level. The transcription is achieved through enzymes called RNA polymerases, which bind to the promoter region of the DNA.

1.1.1 Gene Transcription

Gene transcription is regulated. Transcription regulation may involve the DNA regions of promoters, enhancers, locus control regions (LCRs), and scaffold/matrix attachment regions (S/Mars). In eukaryotes, gene transcription process, or formation of primary

transcript from the DNA is done by recruiting RNA polymerase. RNA polymerase II is recruited for genes that encode for mRNA (messager RNA) [Latchman, 1998].

The transcription activity involves many proteins such as TFs (transcription factors), TAFs (transcription accessory factors), GTFs (general transcription factors) and the complexes of these proteins, as well as the RNA polymerase. GTFs include proteins that may already be multiprotein complexes themselves. Such GTFs may include TFIIA, TFIIB, TFIID, TFIIE, TFIIF, TFIIH, and among these TFIID includes TBP (TATA box binding protein). All of these form TIC (transcription initiation complex). TIC is a necessary substance in transcription initiation [Fickett and Hatzigeorgiou, 1997].

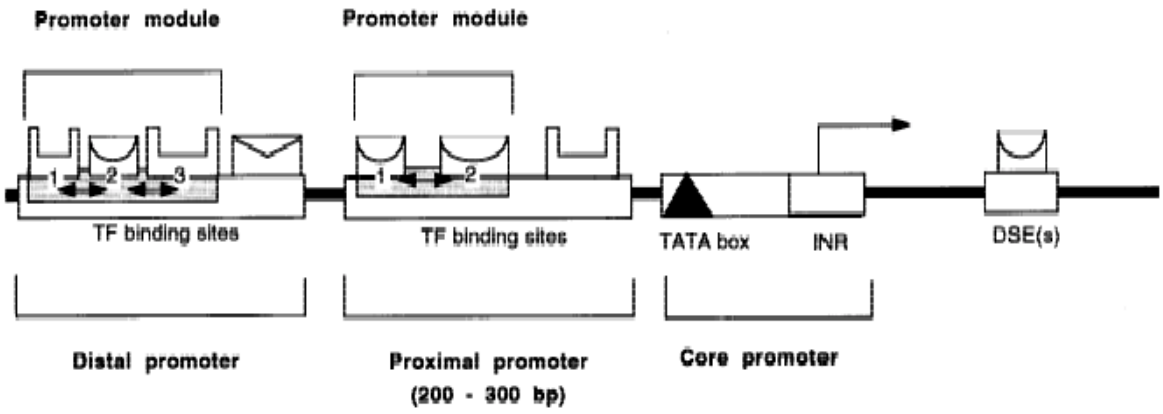


Figure 1.1 Promoter Structure: the schematic of a pol II promoter. [Werner, 2003]

A promoter could be structurally divided into three parts on the DNA: core promoter, proximal promoter and distal promoter.

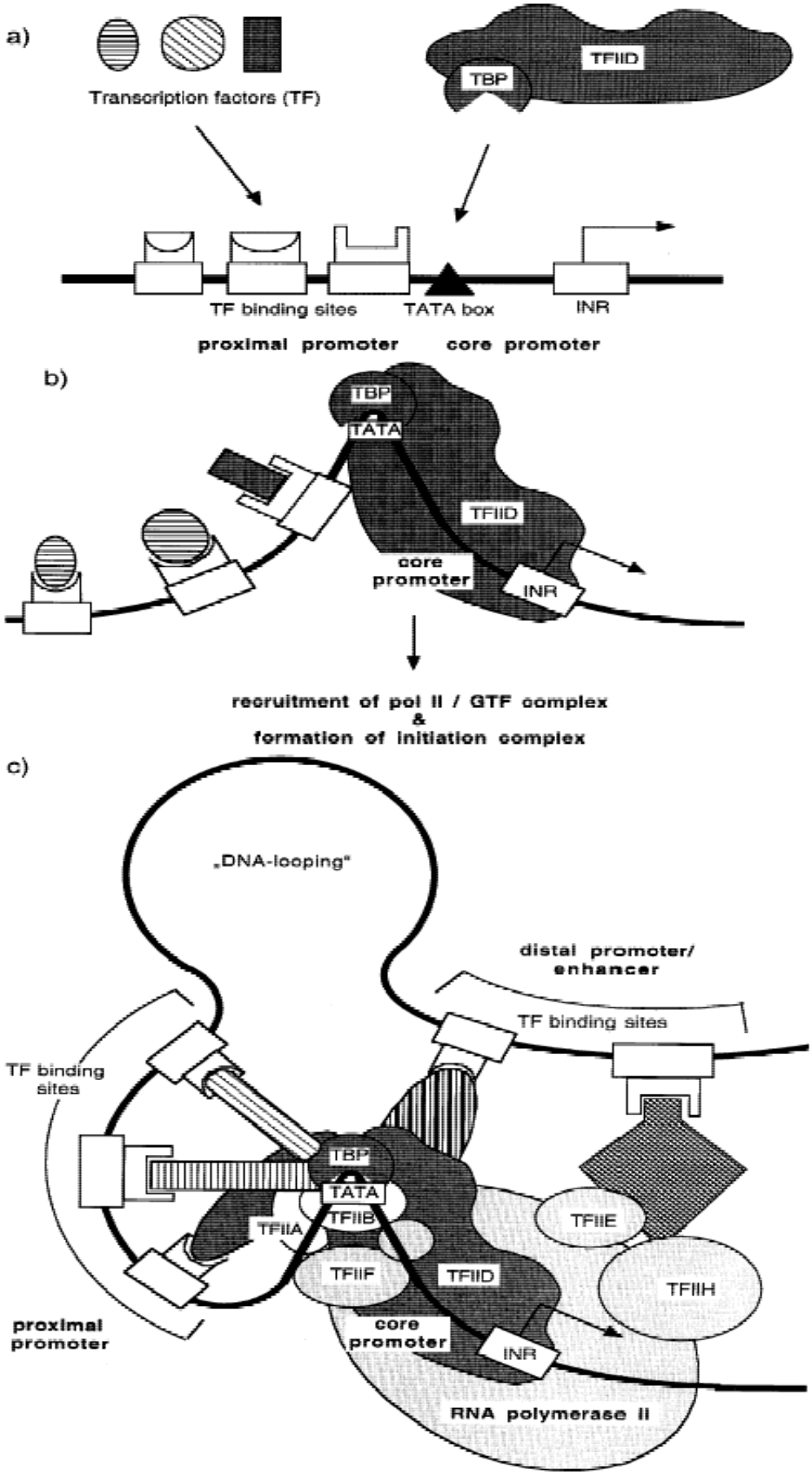

a)
Transcription factors (TF)
TFIID
TBP
TF binding sites
TATA box
INR
proximal promoter
core promoter
b)
TBP
TATA
TFIID
core promoter
INR
recruitment of pol II / GTF complex
&
formation of initiation complex
c)
„DNA-looping"
distal promoter/ enhancer
TF binding sites
TF binding sites
TBP
TATA
TFIIA
TFIIB
TFIIE
TFIIF
TFIID
TFIIH
proximal promoter
core promoter
INR
RNA polymerase II

Figure 1.2 Schematic of gene transcription initiation process [Werner, 1999]:
a) Proximal and core promoter
b) TFs and TFIID complex get bound to the promoter
c) Formation of transcription initiation complex (TIC) following recruitment of RNA Polymerase II

In Figure 1.2a, proximal promoter is 200-300 base pairs long, locating immediately upstream of the core promoter. CCAAT box is mostly located in proximal promoter. Core promoter is the region of the promoter that is sufficient to determine the precise TSS. Core promoter always contains the TSS. Core promoter usually contains some of the three promoter boxes: TATA box, Inr (initiator), and DPE (downstream promoter element). TATA box is usually located 30 base pairs upstream of TSS and determines the upper bound of core promoter. Initiator overlaps with the TSS. These two elements may not be present concurrently in Core promoter. DPE has similar function as the TATA box and is located downstream of Initiator [Werner, 1999, Pedersen *et al*., 1999]. When TATA box is missing in core promoter, DPE takes the role of TATA box. TATA box, Inr and DPE may combine differently to render different functions.

In Figure 1.2b, GTFs and other TFs around core promoter recruit RNA polymerase II to form TIC in the next step.

In Figure 1.2c, TFs get attached to their binding sites in the proximal promoter, distal promoter, enhancer, and silencer regions. (Enhancer and silencer are the regions that are both far away from the TSS. Enhancer increases transcription, while silencer decreases transcription.) In this step, polymerase II requires TFIID to look for TSS along the DNA for transcription initiation. TAFs are associated with TFIID. TBP of the TFIID complex has affinity for TATA box and must bind to it. This way TFIID identifies the exact

location of the TSS in the core promoter. Polymerase II gets complexed with other GTFs and then was recruited to form TIC assembly in the core promoter region. TIC then initiates the transcription.

Transcription can be broadly classified as basal and activated transcription. Basal transcription involves the minimal promoter. The minimal promoter binds a bare minimum number of proteins required to initiate a transcription. Minimal promoter includes core promoter and a few more upstream and downstream regions located close to the TATA box or the TSS [Werner, 1999]. Core promoter and a few other sites upstream and downstream of core promoter are necessary for basal transcription and their combination represents the minimal promoter. Activated transcription may involve certain additional TFs for regulation.

1.1.2 Promoter Basics

Promoter contains the starting point of transcription, the TSS. During initiation of transcription, the TFs bind to specific binding sites of promoters first, and then RNA polymerases are able to recognize the complex between TFs and DNA and bind to the promoter [Scherf *et al.* , 2000].

Currently, there is no computer tool to accurately predict different types of promoters in the genome. The false reporting rate is usually high. The reason lies in the variability of different promoters. Due to high complexity of different organisms, promoters in different cells or tissue are different in structure and characteristics. Eukaryotic promoters may contain regions of TATA-box, CAAT-box, initiator, GC-Box and other transcription

binding sites. Not all of these need to be present in a promoter at the same time, and they may be present in different combinations and their location relative to TSS may also be different in different promoters.

Figure 1.3 is given below to show the modular functional organization of binding sites in promoter hierarchy.

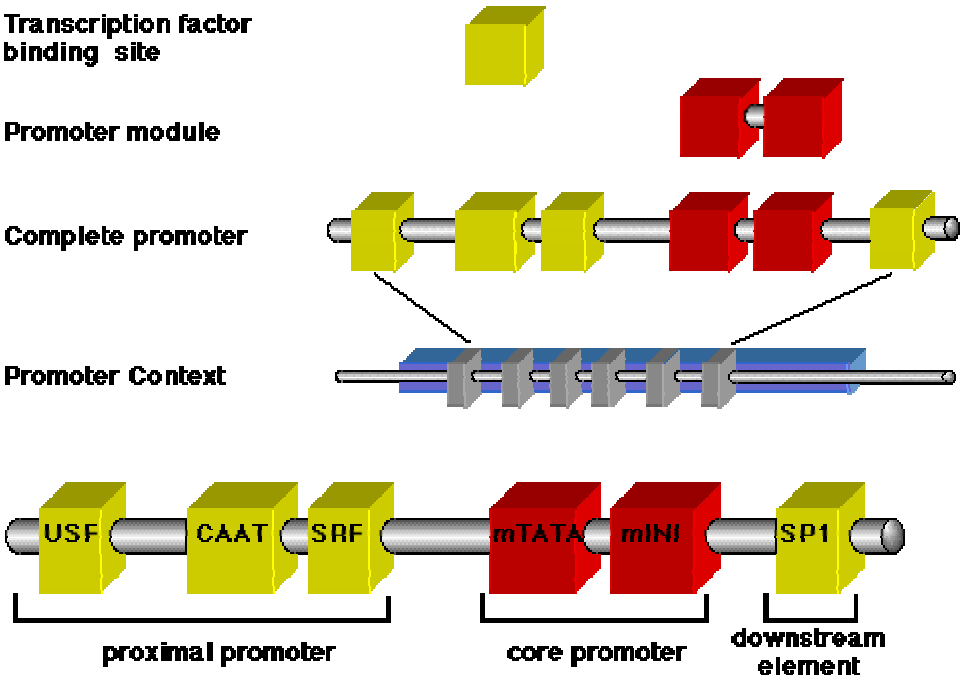


Figure 1.3 Modular organization: modular functional organisation of binding sites in promoter hierarchy. [http://www.genomatix.de/genomics_tutorials/promoter_hierarchy/promoter_hierarchy.html].

Figure 1.4 is shown below to describe the modular organization of promoter elements.

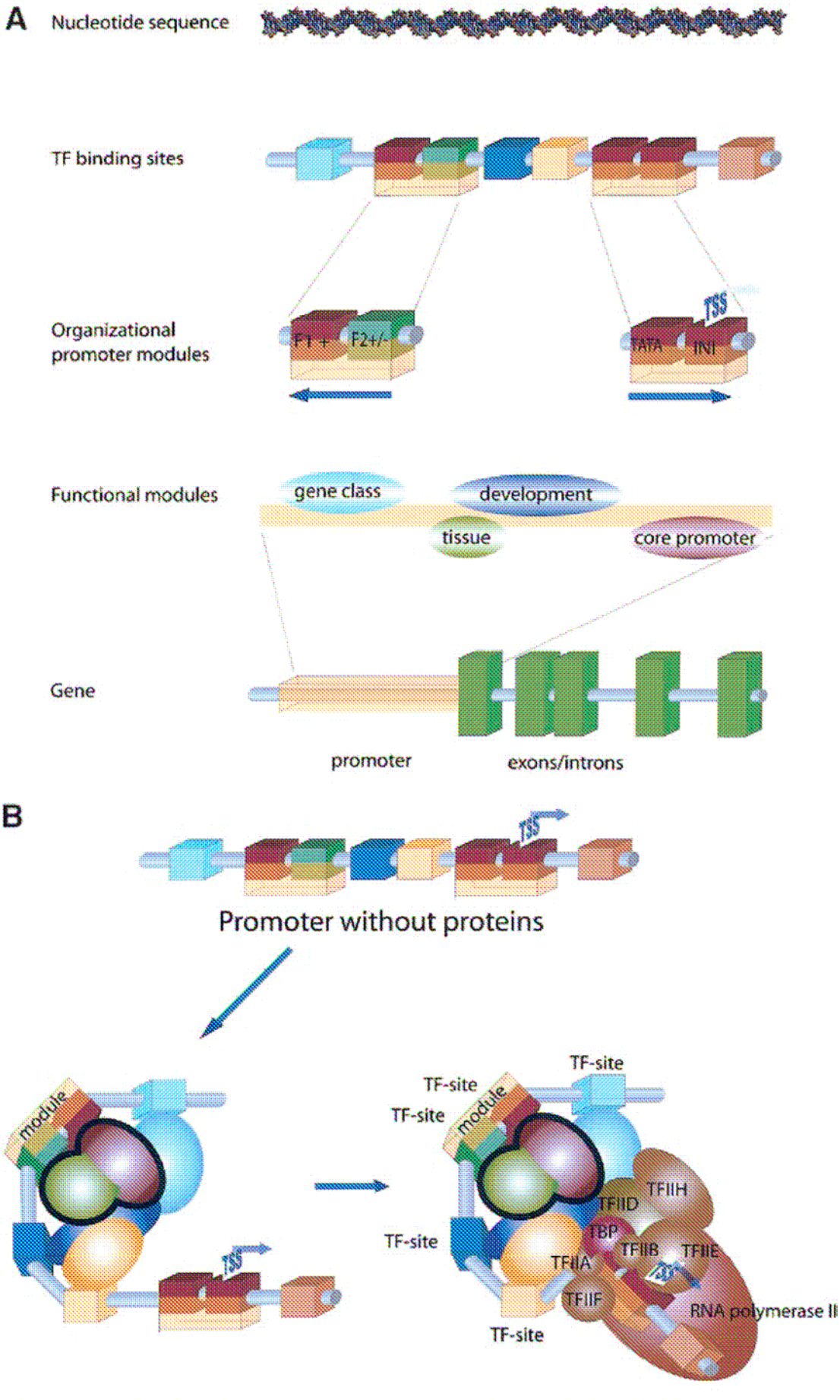

A
Nucleotide sequence
TF binding sites
Organizational promoter modules
F1+/-
F2+/-
TSS
TATA
INR
Functional modules
gene class
development
tissue
core promoter
Gene
promoter
exons/introns
B
TSS
Promoter without proteins
module
TSS
TF-site
TF-site
TF-site
module
TF-site
TFIIH
TFIID
TBP
TFIIB
TFIIE
TFIIA
TFIIF
RNA polymerase II
TF-site
Activation complex
Initiation complex

Figure 1.4 Modular organization of promoter elements [Werner *et al*., 2003]
A) **Promoters in higher eukaryotes are organized hierarchically and elements that control a specific pattern of expression may also be found in other promoters expressed under similar circumstances.**
B) **Active promoters have a unique 3-dimensional structure. Changing the order or spacing of important transcription factor binding sites can change the overall structure of the promoter and thus effect transcription.**

1.2 Existing promoter prediction solutions

The programs mentioned here include those that attempt promoter prediction or localization in anonymous genomic sequences without need of any gene annotation information or other means to limit the actual search space for promoter finding [Werner, 2003].

PromoterInspector [Scherf et al., 2000] is indicated in [Bajic et al., 2004] as the first program to discover eukaryotic polymerase II promoter regions in mammalian genomic sequences with efficiency. It is reported with experiments in [Scherf et al., 2000] to have 43% of predictions as true positives, and the program can predict correctly 43% of the annotated TSS. This program focuses on the genomic context of promoters, not their exact location on the sequences. PromoterInspector is not heuristics based, but relies on content analysis of promoter features represented by IUPAC (International Union of Pure and Applied Chemistry) words, the libraries of which are extracted from training sequences by an unsupervised learning approach. The program compares word frequencies between four functional regions of genes: promoters, exons, introns and 3'UTR [Scherf *et al.* , 2000], which form the four models used. Promoter models are derived with those segments from the EPD data, using regions of [-500, +500] relative to the reference location of TSS (+1). (Under the location of TSS, which is pre-defined as

+1, the range of [#start, #end] is described by the two numbers #start and #end, which are the relative locations of the start and end position of the sequence relative to the location of TSS.) Non-promoter models are derived 100-bp segments from sequences collected randomly from the GenBank database (totalling 1Mbp for each non-promoter group). The system uses a searching window of 100bp that slides along the DNA strand, shifting 4bp ahead each time as a step. The four sensors compete, and the promoter sensor signal must be stronger than signals from the other three sensors. The system predicts a promoter on the occurrence of a minimum of 24 successive positive predictions [Scherf *et al.* , 2000]. The system can scan 100kb in less than 1 minute on a workstation. The system is available at http: genomatix.gsf.de/cgi-bin/promoterinspector/promoterinspector.pl. [Scherf *et al.* , 2000].

Dragon Promoter Finder (DPF) [Bajic *et al*., 2003] is a program that predicts strand-specific TSS and is a general TSS-finding program, not specialized to any particular vertebrate promoter groups. It can successfully recognize both CpG island-related and non related promoters. The system groups the input sequences according to their CpG content first. Then sensors of promoter, exon, intronic sequence are applied to the data. Finally, an ANN is applied to predict the TSS [Bajic *et al*., 2002]. The system uses five different promoter models to enhance its predictive capabilities, and allows several levels of sensitivity, which is chosen by the user. The system was tested on whole human chromosome 22 and it showed a consistent satisfactory performance.

The consistency of Dragon Promoter Finder (DPF) predictions shows that it provides reliable identification of a wider promoter group and does not favour a specific promoter type (such as CpG island- related promoters). Owing to strand-specific predictions,

PromoterInspector cannot achieve a PPV (Positive Predictive Value) greater than 0.5, because it produces one FP prediction for each TP prediction. Furthermore, PromoterInspector can not pinpoint the TSS but only indicates a region that might overlap or be in proximity with the promoter region.

Eponine [Down, and Hubbard, 2002] uses Relevance Vector Machine based on TATA-box motif as its recognition tool, and has better performance when giving predictions for a particular category of genomic sequences of high GC-content. As indicated in [Bajic *et al*., 2004], this program performs with a sensitivity=40.07% and PPV=66.97% using the whole human genome. This is different from the report in [Down, and Hubbard, 2002] that sensitivity=53.5% and PPV=72.73%.

FirstExonFinder [Davuluri *et al*., 2001] is a program to predict TSS, by calculating the value of discriminatory functions. It can also predict the first splice site (intron1). This program has been applied to the 15kb upstream sequences of known genes on chromosomes 21 and 22. The search is restricted with the prior knowledge of the approximate position of the gene start and the strand orientation. As indicated in [Bajic *et al*., 2004], the concepts of CpG island and GC-content are incorporated into the algorithm. It is a program with general purpose and can predict diverse sets of promoters. This program has accurate predictions for CpG-island-related promoters but does not perform well in non-CpG-island-related promoter prediction.

ConPro [Liu and States, 2002] is a system that includes five promoter prediction programs: TSSG, TSSW, Proscan, PromFD, NNPP. Each of these program has high FP prediction rate individually. With these five programs working together, this system is

reported to give 14000 promoters predicted in the genome, and among these 6400 predictions are well-characterized genes. Only a maximum of 1.5 kb upstream sequence of TSS are searched for promoter recognition with this system to reduce false predictions. Because the first introns generally are several kilobytes long, the TP predictions are relatively low in number.

NNPP2.2 [Reese, 2001, 2000] is a program in which promoter prediction is based on artificial neural networks. The system is trained on the TATA-box, the Initiator and allows variable lengths between them, giving the predicted TSS as output [Reese and Eeckman, 1995]. NNPP2.2 makes recognition of TATA box, the initiator and the part in between these two elements in the promoter region. The system uses three time-delay ANNs, with one to predict the TATA box, one to predict the Initiator and one to combine these two outputs and give prediction regarding the spatial distance between the TATA-box and the Initiator. However, according to [Bajic *et al*., 2004], this program does not give satisfactory performance on the whole human genome, producing predictions close to or worse than random guessing, For application in large-scale analyses or even in short DNA segments analyses, NNPP2.2 does not show good performance when considering the cost of obtaining one TP prediction. In the system presented in [Reese and Eeckman, 1995], a neural network is trained to recognize promoter elements. After the neural network is trained, the weights that add the lowest predictive value to the overall prediction in the ANN are pruned. Then the ANN is retrained until the predefined minimum of error level is reached. Finally, by studying the remaining weights of the pruned ANN, the importance of specific positions in the promoter element and the importance of the various promoter elements can be found out.

Promoter2.0 [Knudsen, 1999] also uses ANNs to do promoter prediction, based on conserved sequences and conserved distances between them, giving the predicted TSS as output. The first ANN uses a small window of DNA sequence as input. The system is based on ANNs and was trained to recognize four specific signals most commonly present in eukaryotic promoters-TATA box, Initiator (Inr), GC-box, and CCAAT-box, and their mutual distances. The weights of the neural networks are optimized to give the best separation of promoter and non promoters, by using genetic algorithm [Knudsen, 1999]. For a test set of vertebrate promoter and non promoter sequences, the algorithm was able to give a prediction with correlation coefficient of 0.63. All the five known TSS on the plus strand of the complete adenovirus genome were within 161 bp of 35 predicted TSS. On standardized test set consisting of human genomic DNA, the system gives better performance than other software. But DPF makes 21 times fewer FP predictions than this system with the same level of TP prediction [Bajic *et al*., 2002].

CpGpromoter [Ioshikhes & Zhang, 2000] is a program to do a large-scale human promoter prediction based on results of discriminant analysis between the promoter-related CpG islands and non-related ones. CpG islands are an important signature of 5' region of many mammalian genes. In the DNA range of [-500, +1500] around a TSS (+1) that containing a CpG island inside, the mapping of human promoters can be implemented efficiently with a resolution of 2kb. As indicated in [Gardiner-Garden and Frommer, 1987], CpG islands have a length of more than 200 bps, a high GC-content (more than 50%), and a high frequency of CpG dinucleotides (at least 0.6 of their expected frequency).

CpGProD [Ponger and Mouchiroud, 2002] is a system to predict mammalian promoter regions that are CpG islands related in large genomic sequences. CpG-islands-related promoters count for approximately half of all the genes, and CpGProD is exclusively restricted for identification of this class of promoters. However, as indicated in [Bajic *et al*., 2004], CpGProD finds TSS with greatest accuracy, with low sensitivity (37%) and CpGProD requires the use of RepeatMasker. This program uses different parameters to do promoter prediction for the two different spices of human and mouse accordingly.

Dragon Gene Start Finder [Bajic and Seah, 2003a; Bajic and Seah, 2003b] is an advanced system for recognition of gene starts in mammalian genomes. The system makes predictions of gene start location by combining information about CpG islands, TSSs, and signals downstream of the predicted TSSs. The system aims at predicting a region that contains the gene start or is in its proximity. Evaluation on human chromosomes 4, 21, and 22 resulted in SE (Sensitivity) of over 65% and in a PPV of 78%. The system makes on average one prediction per 177,000 nucleotides on the human genome, as judged by the results on chromosome 21. Comparison of abilities to predict TSS with the two other systems on human chromosomes 4, 21, and 22 reveals that our system has superior accuracy and overall provides the most confident predictions. This system studies the statistical properties of promoter regions, with Artificial Neural Network applied as part of its design, GC-content used in its algorithm, and concept of CpG island combined with predictions of DragonPF [Bajic *et al*., 2003]. As indicated in [Bajic *et al*., 2004], the sensitivity and PPV are approximately equal in the design of DragonGSF. On three whole chromosomes of human chromosomes of 4, 22, 21, this system achieves a PPV=78%, but on the human genome, it only achieves a PPV=62.98%. RepeatMasker has no benefits when applied to DragonGSF. The system makes approximately 0.6 FP predictions for

every TP prediction. It will cover about 65% of all promoters, with the preference to the CpG-island-related ones.

McPromoter [Ohler *et al*., 2000; Ohler *et al*., 2002] is a program that locates eukaryotic polymerase II TSSs in genomic DNA based on statistics study of promoters versus non-promoters and the different physical properties of promoter regions, with Artificial Neural Network and interpolated Markov model as its recognition technology basis. It consists of a model for promoter sequences and a mixture model for non-promoter sequences, containing submodels for coding and non-coding sequences. A sliding window of 300 bps long is searching over the sequence, with the step of 10 bp. At every position, the difference between the log likelihood of the promoter and the non-promoter model is computed. The resulting plot describes the regulatory potential over the sequence and is smoothed by a median and hysteresis filter to eliminate single false predictions and reduce the high number of neighbouring minima that are due to noise. The program then makes a prediction for each local minimum below a pre-specified threshold. As indicated in [Bajic *et al*., 2004], its performance on the human genome has improved compared with its reported one in [Ohler *et al*., 2002] on chromosome 22, from sensitiviy=52.8% and PPV=62.6% to sensitivity=57.92% and PPV=74.13%, though the two criteria are different. The use of RepeatMasker results in evident improvement of McPromoter accuracy. Its performance is good but its unsatisfactory speed prevents it from applying to large-scale promoter prediction.

Here I also give a summary for those programs that are less famous but worth mentioning. rVISTA [Loots and Ovcharenko, 2004] is a program that combines TFBS database search with a comparative sequence analysis. The human and mouse gene sequences are

aligned and potential TFBS were predicted, and the human –mouse sequence conservation of a DNA region spanning a TFBS was assessed. TraFaC (Transcription Factor Binding Site Comparison) [Jegga *et al*., 2002] is a program that has been built to find out regulatory regions by implementing sequences comparison. Levitsky and his colleagues built a system [Levitsky and Katokhin, 2003] to calculate different characteristics of genomic DNA, among which they found out the potential to form nucleosomal complexes, which may be an important feature in tissue-specific expressed promoters. This system however is only good to assess properties of such promoters after location of promoters have been made. Signal Scan [Prestridge, 1991] is a program that finds promoter elements in the input sequence, by doing a specific, consensus and matrix searches in the SIGNAL SCAN database. The database is composed of specific sequence elements derived from biochemical characterization and elements from derived consensus sequences. In another program developed by Audic and Claverie [Audic and Claverie, 1998], Markov Models are developed to do a sequence comparison and Bayesian method is applied to separate promoters from non promoters. PromFind [Hutchinson, 1996] is a system using the idea to give score to the input sequences according to their differential hexamer measure. This system works with two other programs named SorFind and RepFind to generate a feature table to assess the predicted promoter regions. PromoterScan [Prestridge, 1995] is a program that evaluates based on combined scores from the features of the TATA box weight matrix and the density of TFBSs, giving the output of a TSS or the core promoter shown by a window of 250bp long, in which case TSS can be decided with the end position of the window. Also, this system can be used to give a further analysis, e.g. aligning the predicted promoter to EPD to search similar promoter and a number of TFBSs that are common to both the predicted promoters and their corresponding promoters in EPD. TSSW/TSSG/TSSP (W-TFD, G-TransFac, P-

Plant) [Wingender, 1994, Prestridge, 1995, Ghosh, 1993] is a program that uses the idea of Linear Discriminant function based on the combinational sores with TATA box, Triplet preferences around TSS, Hexamer frequencies in consecutive upstream 100-bp regions, and TFBSs, giving the output of predicted promoters and their transcriptional elements. TSSG and TSSW were accessed at the site http://dot.imgen.bcm.tmc.edu:9331/gene-finder/gf.html. TSSG correctly predicted 7 (29%) of the true promoters and predicted 25 false positives (1/1325 bp). TSSW correctly predicted 10 (42%) of the true promoters and gave 42 false positives (1/789 bp)" [Fickett and Hatzigeorgiou, 1997]. CorePromoter [Zhang, 1998] is a program that gives predictions of TSS by a quadratic discriminate analysis based on Pentamers within a window of 30bp and 45bp sliding in a region with the length of 240 bps. GSF suite is composed of three programs called PatSearch [Pesole *et al*., 2000], MatInspector [Quandt *et al*., 1995, Cartharius, 2005.], and ConsInspector [Frech et al.,1993, Frech et al., 1997, Quandt et al.,1995b, Wingender et al., 1995]. PatSearch separates core and whole site, allocating weights for important bases, and allows mismatches. MatInspector applied core and matrix cut offs by organism classes. ConsInspector can create new matrices. FastM [Klingenhoff, 1999, Lavorgna *et al*., 1998.] is a program that gives sequences as input. It searches for TFBSs which are clustered in groups. This method is creative in the sense that it builds DNA unit models. These models are built using various individual elements, such as TFBSs, and promoters.

1.3 Contribution of Thesis

Promoter prediction is currently a hot problem in the field of Bioinformatics. DSP techniques have not been largely applied in studying this topic. The project is to explain the suitability of DSP techniques to enhance promoter prediction.

Several valuable findings have been made based on systematic simulation studies and experiments. Instead of using the feature of CC (Correlation Coefficient), the more appropriate features of the coefficients of DFT, DCT, DWT transform of the original signal, are adopted. The process of how to select the optimal combination of features and classifier model for each of the 22 groups split by GC-content is presented in this thesis. The performances of different combinations of features are evaluated. Findings are also made, that the capability to recognize promoters degrades with the reduction of GC-content of the data; there are no significant differences in the prediction performance when any of the three transform is applied; and the best performance is achieved by combining all the three transforms.

A promoter prediction system based on Support Vector Machine (SVM) is developed. Results of the application of the system to the human chromosome 22 (NCBI built 35) are given in the thesis, as well as performance analysis of the six annotated categories of genes.

Future work can be extended based on the achievements here. I have examined how to combine possible features extracted under domain transforms in DSP field with biological features of promoters and non-promoters. The biological features adopted here include the number of CpG dinucleotides, GC-content and the number of different combinations of mono-.di-, tri- nucleotides. Other probable DSP techniques should be explored to combine with more of other appropriate biological features and physical properties of promoter regions to achieve even better prediction performance.

1.4 Thesis Organisation

This thesis is organised as follows:

Chapter 1: The concept of promoter and necessary biological background is introduced together with the promoter prediction problem. A review of the available techniques follows and an account of the thesis outline and contribution is given.

Chapter 2: The signal model for the promoter sequence and non-promoter sequence in this thesis is formulated and the comparison of their respective statistical characteristics is made by the means of signal mean, correlation coefficient of specific sequence with the mean signal, and the distribution of this correlation coefficient. The finding of the experiments is discussed.

Chapter 3: The more general features, which are coefficients of DFT, DCT, DWT transform of the original signal, are adopted and simulation studies are presented on selecting the best combination of features and the optimal classifier model. Comparison is made systematically and the performance is analysed.

Chapter 4: The optimal model for each GC group is explored and a comparison of their performances is made. The final prediction system is applied to human chromosome 22 (NCBI built 35). The score indicating probability of possible promoter position on the chromosome sequence is reported.

Chapter 5: The conclusion and findings of this thesis are given.

Chapter 2

Signal Model and the Effectiveness of Transforms

In this chapter, we develop the signal model and show how the biological prediction problem can be considered as a signal processing problem. However, we find the features obtained from the CC with reference to the mean signal to be not effective. In Chapter 3, we study the effectiveness of the features obtained from the transform domain coefficients of signals.

2.1 Signal Model

The promoter sequence is assumed to be the sequence, which contains a TSS. Its counterpart, the non-promoter sequence is assumed not to contain a TSS. We define the promoter sequence as positive data, and the non-promoter sequence as negative data which facilitates classification between the promoter and non-promoter sequence.

Dataset:

The sequence of 2500bp nucleotide, with the positive data from the [-2000, +500] relative to the TSS (+1) is used. The range of [#start, #end] is defined by the two numbers #start and #end, which are the locations of the start and end position of the sequence relative to the reference location of TSS (+1). The negative data is from the DNA range of [5001, 7500] relative to the TSS (+1). We choose the range of [5001, 7500] since the sequence

in this range is regarded to be distant enough relative to TSS (+1), hence can be used as "negative" data; whereas the sequence with range of [-2000, +500] relative to the TSS (+1) is used as the "positive" data.

In another experiment, truncated data, which is from the [-500, +500] relative to the TSS (+1) is also generated and applied. Each base pair (a, c, g, t) is represented by the respective value of the EIIP (Electron–ion Interaction Potential) [Veljkovic and Slavic, 1972], with a=0.1260, c=0.1340, g=0.0806, t=0.1335. By assigning these numbers to the base pairs, the nucleotide sequence is converted to a sequence of numbers. Thus, the problem can be solved in digital signal processing domain. The sequence which contains 'N' will not be processed here. (e.g. "aacggt" is converted to "0.1260, 0.1260, 0.1340, 0.0806, 0.0806, 0.1335".)

Figure 2.1 below depicts the mean sequences of promoter (positive) and non-promoter (negative) sequences in the "reviewed" data set.

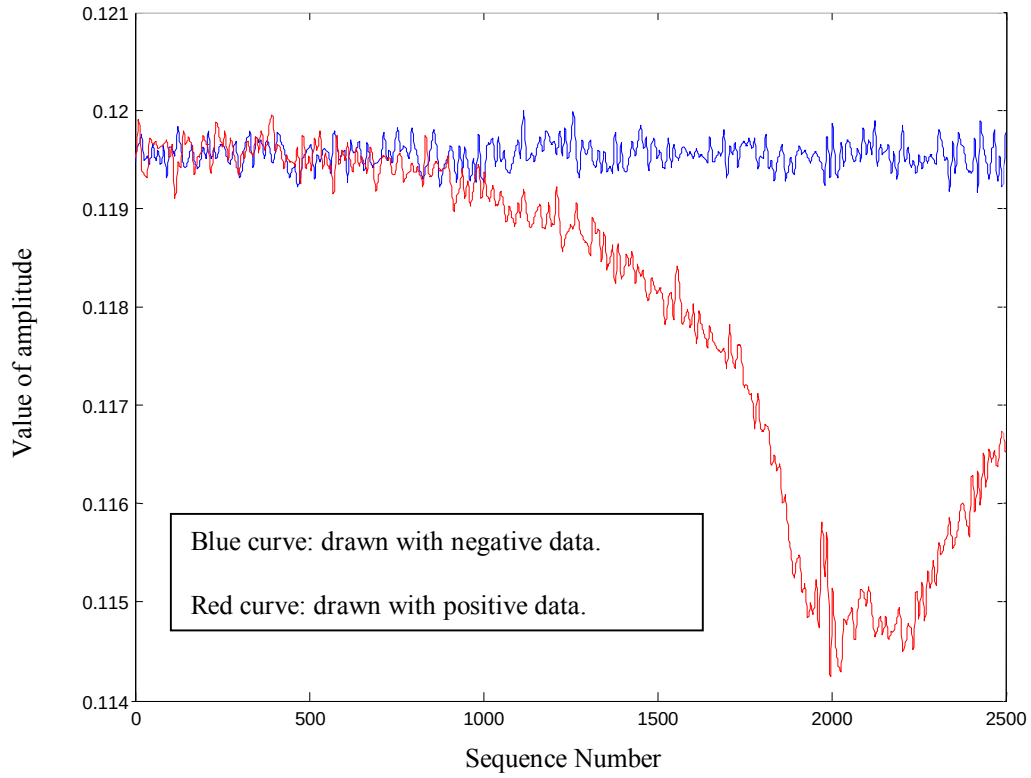


Figure 2.1 The mean signal of the original positive and negative data

Figure 2.1 shows the mean signal of the negative and positive sequences that are converted from the original nucleotide sequences using the value of EIIP. The x axis is the length of the sequence, from 0 to 2500. The y axis is the value of the mean signal's amplitude at each position of the sequence. Clearly, the two curves are quite different from each other. The negative sequence (blue curve) resembles a random while the positive sequence shows the lowest at approximately 2000 as shown in Figure 2.1, which is most likely to be the location of TSS. This is consistent with the fact that we use the positive sequence of 2500bp nucleotide, with the range in [-2000, +500] relative to the TSS (+1).

The data used in the experiments contains three data sets: the "predicted", the "provisional" data, and the "reviewed" data. The number of negative and positive sequences in each data is given in Table 2.1. In our experiment, only the "reviewed" data is used.

	Predicted	provisional	reviewed
Negative	2440	4696	3243
Positive	2428	4655	3219

Table 2.1 Three sets of negative and positive data used in experiments

The "predicted" data in the first column represents TSS data that is obtained by FIE2 from LocusLink's Evidence Viewer (EV) page where one of the sequences that was aligned against the human genomic sequence to determine the TSS was a predicted RefSeq. A predicted RefSeq record has not been subjected to individual review. The transcript may represent an ab initio prediction or may be partially supported by other transcript data; in both cases, the protein is predicted. Support for the transcript may include the existence of cDNA clones, ESTs, or homology [Maglott *et al*., 2000; Pruitt *et al*., 2000; Pruitt and Maglott, 2000].

The "provisional" data in the second column represents TSS data that is obtained where one of the sequences that were aligned against the human genomic sequence to determine the TSS was a provisional RefSeq. A provisional RefSeq record has not yet been subject to individual review and is thought to be well supported and to represent a valid transcript and protein. The initial sequence-to-gene name associations are established by outside collaborators or NCBI staff. This is the default status code applied to some genomes for which there is no clear information about the method used to define the sequence [Maglott *et al*., 2000; Pruitt *et al*., 2000; Pruitt and Maglott, 2000].

The "reviewed" data in the third column represents TSS data that is obtained where one of the sequences that were aligned against the human genomic sequence to determine the TSS was a reviewed RefSeq. A reviewed RefSeq record has been the reviewed by NCBI staff or by a collaborator. The NCBI review process includes reviewing available sequence data and frequently also includes a review of the literature and other sources of information. Some RefSeq records may incorporate expanded sequence and annotation information including additional publications and features, as deemed relevant. More detailed descriptions of the review process are provided for the separate NCBI projects which supply these records [Maglott *et al*., 2000; Pruitt *et al*., 2000; Pruitt and Maglott, 2000].

The correlation coefficient (a number between 0 and 1) is a good indicator in statistics which shows the correlation between two variables. The CC between the two variables increases as the strength of the relationship increases.

We calculate the CC between individual sequence and the mean sequence of the reconstructed positive data as follows. The individual sequence x is composed of n points x_1, x_2, … , x_n and the mean sequence y of the reconstructed positive data is composed of n points y_1, y_2, … , y_n.

The mean of x and y are: $\bar{x} = \frac{1}{n}\sum_{i=1\ldots n} x_i$, $\bar{y} = \frac{1}{n}\sum_{i=1\ldots n} y_i$ respectively.

The standard deviations x and y are: $\delta_x = \frac{1}{n}\sqrt{\sum_{i=1\ldots n}(x_i - \bar{x})^2}$, $\delta_y = \frac{1}{n}\sqrt{\sum_{i=1\ldots n}(y_i - \bar{y})^2}$ respectively.

The covariance between x and y is: $\delta_{xy} = \frac{1}{n}\sum_{i=1\ldots n}(x_i - \bar{x})(y_i - \bar{y})$.

The correlation coefficient between x and y is: $CC = \frac{\delta_{xy}}{\delta_x \delta_y}$.

The distribution of correlation coefficients between the individual sequence and the mean sequence of the reconstructed positive data is presented in Appendix B.

2.2 Transformation applied to the signal

The digital signal obtained after conversion from nucleotide sequence with EIIP is decomposed by DFT, DCT, and DWT transformations.

We compare the performance of these different transformations in pre-processing the signals before they are classified as promoters and non promoters. Each specific transform is applied to a pre-selected data segment from database.

2.2.1 Discrete Fourier Transform

The Fourier Transform (FT) is a powerful tool for signal analysis. In Digital Signal Processing (DSP), we may work between the spatial domain and the frequency domain while proceeding through a problem. This ability is quite useful, since one can work in either the spatial or frequency domain with the FT, and different information is provided from different angle.

Continuous Fourier Transform:

$$F\{f(t)\} = F(s) = \int_{-\infty}^{\infty} f(t)e^{-j2\pi st}dt$$

$f(t)$ is the signal in time domain, and $F(s)$ is the Continuous Fourier Transform of the signal $f(t)$.

Inverse Continuous Fourier Transform:

$$F^{-1}\{F(s)\} = \int_{-\infty}^{\infty} F(s)e^{j2\pi st}ds$$

F^{-1} is the Inverse Continuous Fourier Transform of the signal $F(s)$. For any function, its Fourier Transform function is unique, and vice versa.

Discrete Fourier Transform (DFT):

$$F_n = \frac{1}{\sqrt{N}} \sum_{i=0}^{N-1} f_i e^{-j2\pi \frac{n}{N} i}$$

Invert DFT:

$$f_i = \frac{1}{\sqrt{N}} \sum_{i=0}^{N-1} F_n e^{j2\pi \frac{i}{N} n}$$

f_i is the i-th point of discrete form of signal $f(t)$. F_n is the n-point Discrete Fourier Transform of signal f_i.

The practical implementation of the FT to a signal is often realized by the means of FFT, which is developed based on the DFT. With the sampling rule, the DFT can be viewed as essentially equivalent to CFT (Continuous Fourier Transform). The continuous transform can be firstly employed when formulating a solution to a signal processing problem, and then the discrete transform can be implemented with that solution.

Fast Fourier Transform (FFT):

The number of multiplication and addition operations require to implement DFT or IDFT is on the order of N^2. FFT reduces the required number of operations to the order of $N \log_2(N)$. In FFT, N is usually a power of 2, hence producing the highest efficiency and the simplest implementation result.

So later in Chapter 3, we apply another dataset in which the length of the sequence is 1024bp, that is N=1024, the 10th power of 2.

$$\begin{bmatrix} F_0 \\ \cdot \\ \cdot \\ \cdot \\ F_{N-1} \end{bmatrix} = \begin{bmatrix} W_{0,0} & \cdots & W_{0,N-1} \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ W_{N-1,0} & \cdots & W_{N-1,N-1} \end{bmatrix} \begin{bmatrix} f_0 \\ \cdot \\ \cdot \\ \cdot \\ f_{N-1} \end{bmatrix}$$

F=W*f

$$w_{n,i} = \frac{1}{\sqrt{N}} e^{-j2\pi\frac{ni}{N}}$$

$w_{n,i}$ is the exponential function item in the matrix W of n*i. Since the exponential function is periodic in the product of n and i, there is considerable symmetry in the matrix W. The matrix can be factorized into a product of N-by-N matrices that contain repeated values, including many zeros and ones. If $N = 2^p$, matrix W can be factorized into p number of such matrices. The total number of operations required to implement p of those factorized matrix products is substantially less than that required for the original matrix equation. Thus, the speed of calculation is greatly improved.

The factor by which the FFT reduces the computational workload compared to the original workload is:

$$\frac{N^2}{N\log_2(N)} = \frac{N}{\log_2(N)}$$

This value increases with N. For N=1024, the FFT is approximately 100 times as efficient as the direct implementation, so that the speed of computation is greatly enhanced. This is good when we later process nucleotide sequences with the length of 1024bp in Chapter 3.

2.2.2 Discrete Cosine Transform

The Discrete Cosine Transform (DCT) separates the signal into parts (or spectral sub-bands) of different importance, which is reflected by the signal's amplitude value. The DCT is similar to the discrete Fourier transform in the functionality that it transforms a signal from the time domain to the frequency domain. With an input signal $A(i)$ in the spatial domain, the signal in the frequency domain after Discrete Cosine Transform is:

$$B(k_1) = \sum_{i=0}^{N_1-1} 2 \cdot A(i) \cdot \cos\left[\frac{\pi \cdot k_1}{2 \cdot N_1} \cdot (2 \cdot i + 1)\right]$$

To retain only the low frequency component of the original signal, a low-pass filter can be applied. Similarly, a high-pass filter can be applied if high frequency component is needed.

2.2.3 Discrete Wavelet Transform

Conventional Fourier transforms provide only the frequency information, since temporal information is lost in the transformation process. WT is different from conventional Fourier analysis in the sense that it can also discover the signal's "local" periodicities. Unlike the Fourier transform, whose basis functions are sinusoids, wavelet transforms are based on small waves, called wavelets, of varying frequency and limited duration.

Multi resolution theory was born in the mid 1980's, and the scaling function of wavelets was first used and the own family of wavelets can be constructed. Multi-resolution theory is concerned with the representation and analysis of signals at more than one resolution. The appeal of such an approach is that obvious features that might go undetected at one resolution may be easy to be clear at another. Multi resolution theory incorporates and unifies techniques from a variety of disciplines, including sub-band coding from signal processing, quadrature mirror filtering from digital speech recognition, and pyramidal image/signal processing. Although the imaging/signal community's interest in multi resolution analysis was limited until the late 1980s, there has been enormous new findings with this subject today [Gonzalez and Woods, 2004].

Similarly, the generalized wavelet series expansion in wavelet domain is the counterpart of Fourier series expansion in Fourier domain. The discrete wavelet transform is the counterpart of discrete Fourier transform, and the continuous wavelet transform is the counterpart of integral Fourier transform, respectively. Usually the discrete wavelet transform is implemented as fast wavelet transform with computational efficiency.

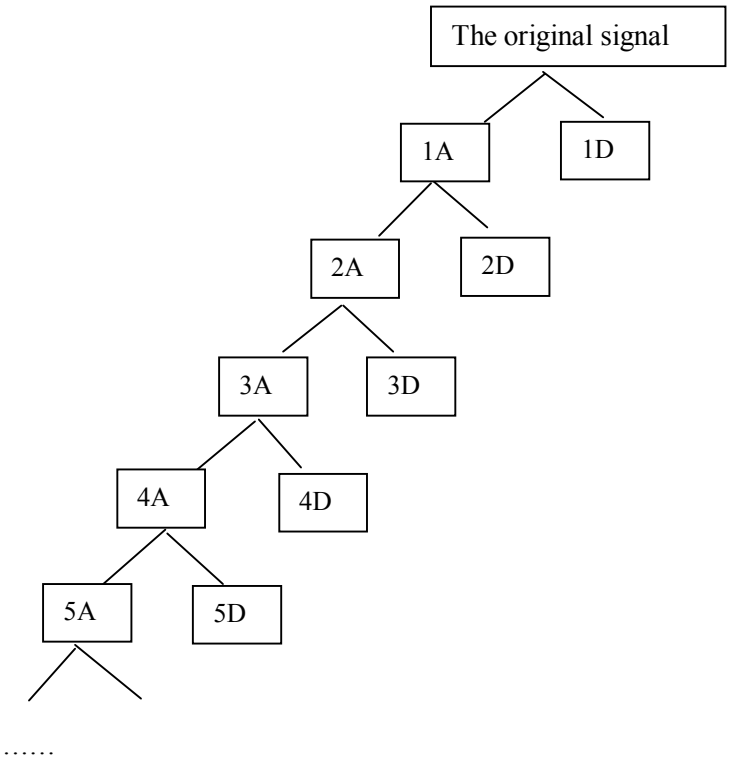


Figure 2.2 The wavelet decomposition is implemented at different levels

As shown in Figure 2.2, the original signal is decomposed from level 1 to level 5 with DWT. Respectively, the 1A, 2A, 3A, 4A, 5A…is the approximate part of the original signal while the 1D, 2D, 3D, 4D, 5D…is the detailed part in each level. The signal can be reconstructed by these different parts of the original signal specifically to observe the signal with different resolutions. This process is similar to applying a low-pass filter or a high-pass filter to the original signal accordingly to observe the low or high frequency part of it. In level one, the original signal is decomposed into the 1A (approximate) and 1D (detailed) part, with 1A being the low frequency content and 1D being the high frequency content of the signal. Then in level 2, the 1A part is decomposed into 2A and 2D part, with 2A being the low frequency content and 2D being the high frequency content of 1A. The 1D part is no longer included in the decomposition in level 2, 3, 4 and 5. Similarly,

the wavelet transforms implemented in deeper levels decomposes only the approximated part of the signal in the previous level.

In Appendix B, Figure B.1, B.2 and B.3 are the plots obtained when the original signal is decomposed respectively at level 1, 2, and 7.

2.3 Simulation studies on the feature of CC.

Table 2.2 gives details of the "reviewed" data that is first split into 22 groups by their GC-content (Grouping of 22 groups by GC-content is made from Group1-- "GC rich", with G+C>80%, and Group22-- "GC poor", with G+C<40%, holding a step of 2% decrease of GC content of groups in between) through preprocessing in the experiment. The so-called GC-content is an important sequence feature and is tightly correlated with different aspects of sequence biological activities. GC-content can be defined as (#G + #C)/Sequence Length, where #G and #C are the total number of G and C nucleotides in the original nucleotide sequence, respectively. In the context of transcription activation, human promoters are known to be characterized by a higher GC-content than the bulk genomic sequences, although there are a smaller proportion of promoters that are GC-poor [Zhang *et al*., 2004].

	G+C content	# of Negatives	# of Positives
Group1	G+C>80%	0	15
Group2	78%<G+C<=80%	0	24
Group3	76%<G+C<=78%	1	49
Group4	74%<G+C<=76%	2	105
Group5	72%<G+C<=74%	3	144
Group6	70%<G+C<=72%	6	162
Group7	68%<G+C<=70%	6	193
Group8	66%<G+C<=68%	26	228

Group9	64%<G+C<=66%	33	261
Group10	62%<G+C<=64%	45	231
Group11	60%<G+C<=62%	71	259
Group12	58%<G+C<=60%	92	235
Group13	56%<G+C<=58%	113	216
Group14	54%<G+C<=56%	158	191
Group15	52%<G+C<=54%	132	173
Group16	50%<G+C<=52%	211	137
Group17	48%<G+C<=50%	232	97
Group18	46%<G+C<=48%	236	95
Group19	44%<G+C<=46%	261	72
Group20	42%<G+C<=44%	335	58
Group21	40%<G+C<=42%	320	63
Group22	G+C<40%	960	211
sum		3243	3219

Table 2.2 Grouping of the "reviewed" data set into 22 parts by GC content

More figures and tables obtained from the experiments are attached in Appendix B for reference.

2.4 Performance of the feature of CC

	G+C content (1024componet)	#of Neg	#of Posi	#of(N=P) Train	#of Test	Confusion matrix	TP rate	FP rate	PPV	F-measure
Group1	G+C>80%	0	17							
Group2	78%<G+C<=80%	1	28							
Group3	76%<G+C<=78%	1	96							
Group4	74%<G+C<=76%	6	198	3/3	3/195	44 151 0 3	0.2256	0	1	0.3682
Group5	72%<G+C<=74%	8	312	4/4	4/308	19 289 0 4	0.0617	0	1	0.1162
Group6	70%<G+C<=72%	31	484	15/15	16/469	7 462 0 16	0.0149	0	1	0.0294
Group7	68%<G+C<=70%	42	587	21/21	21/566	162 404 7 14	0.2862	0.3333	0.9586	0.4408
Group8	66%<G+C<=68%	78	798	39/39	39/759	116 643 3 36	0.1528	0.0769	0.9748	0.2642
Group9	64%<G+C<=66%	135	841	67/67	68/774	8 766 1 67	0.0103	0.0147	0.8889	0.0204
Group10	62%<G+C<=64%	211	1032	105/105	106/927	48 879 2 104	0.0518	0.0189	0.96	0.0983
Group11	60%<G+C<=62%	305	989	152/152	153/837	79 758 4 149	0.0944	0.0261	0.9518	0.1717
Group12	58%<G+C<=60%	368	1091	184/184	184/907	25 882 1 183	0.0276	0.0054	0.9615	0.0536
Group13	56%<G+C<=58%	441	1019	220/220	221/779	13 786 1 220	0.0163	0.0045	0.9286	0.032
Group14	54%<G+C<=56%	541	1092	270/270	271/822	8 814 1 270	0.0097	0.0037	0.8889	0.0193
Group15	52%<G+C<=54%	701	952	350/350	351/602	53 549 10 341	0.088	0.0285	0.8413	0.1594
Group16	50%<G+C<=52%	869	855	427/427	442/428	12 416 2 440	0.028	0.0045	0.8571	0.0543
Group17	48%<G+C<=50%	1055	772	386/386	669/386	2 384 0 669	0.0052	0	1	0.0103
Group18	46%<G+C<=48%	1073	556	278/278	795/278	256 22 698 97	0.9209	0.878	0.2683	0.4156
Group19	44%<G+C<=46%	1239	485	242/242	997/243	2 241 0 997	0.0082	0	1	0.0163
Group20	42%<G+C<=44%	1285	420	210/210	1075/210	0 210 0 1075	0	0	NaN	0
Group21	40%<G+C<=42%	1338	339	169/169	1169/170	141 29 982 187	0.8294	0.84	0.1256	0.2181
Group22	G+C<40%	4273	1038	519/519	3754/519	17 502 74 3680	0.0328	0.0197	0.1868	0.0557
Overall		14001	14001	3661/3661	10338/10199	1012 9187 1786 8552	0.0992	0.1728	0.3617	0.1557

Table 2.3 Performance of the feature of CC

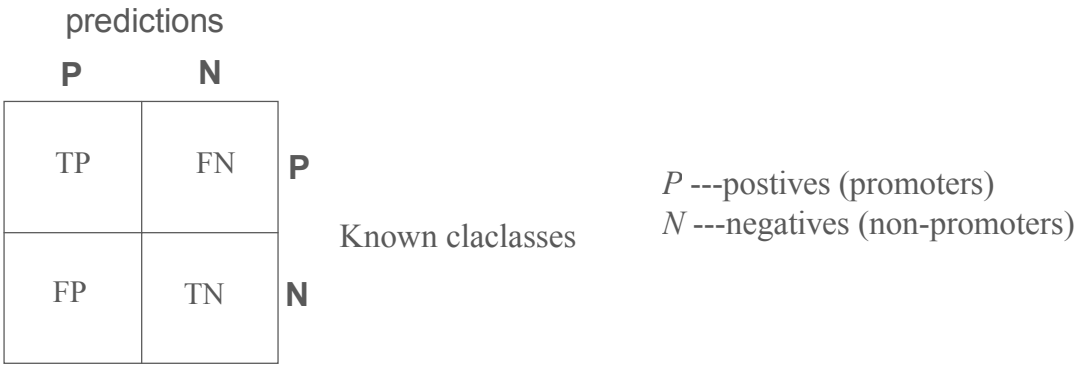
Table 2.3 presents the experiment result obtained with classifier NaïveBayes. The NaïveBayes classifier makes predictions using Bayes' Theorem and derives the probability from the underlying evidence. As for the dataset used here, the length of the sequence is 1024bp. The sequence is with the range of [-512, +512] relative to TSS.

14001 positive sequences and 14001 negative sequences are used. 50% of the minimum of the positive data and the negative data in each group is used for training, and the rest are used as test data. Since the number of negative sequence contained in Group 1, 2, and 3 is only 0,1,and 1 respectively, these 3 groups are not included in the experiment.

The terms we used here in data analysis:

The rates of True and False Positives have to be taken into consideration. A confusion matrix is used for checking the accuracy of a classification.

One way is the representation in a confusion matrix.



TP---True Positive FP---False Positives

TN---True Negatives FN---False Negatives

TP (True Positive) ---correct classifications. $TP \quad rate = \frac{TP}{TP + FN} \times 100\%$, it is defined as TP over whole positives. TP rate is also called “Se” or “Recall”.

FP (False Positive) ---if the sample is incorrectly predicted as positive, while actually should be negative. $FP \quad rate = \frac{FP}{FP + TN} \times 100\%$, it is defined as FP over whole negatives.

$\text{PPV - Precision} = \frac{TP}{TP + FP} \times 100\%$, it is defined as TP over whole predictions.

We will also use the *F-measure* [Van Rijsbergen, 1979] which combines recall and precision in a single efficiency measure (it is the *harmonic mean* of precision and recall).

$$\text{F-measure} = \frac{2 \times recall \times precision}{recall + precision} = \frac{2TP}{2TP + FP + FN}$$

As shown in Table 2.3, the Se and PPV obtained with the feature of CC is only 9.92% and 36.17%, which shows that it is not a proper feature to separate the positive and negative data compared to findings of later chapters. (At the end of the next chapter, the best combined result produces Se = 71.31% and PPV = 71.22% will be shown.) So we move forward to select the features of the coefficients of the signals after domain transforms. Their performance in promoter and non-promoter classification will be described in the next Chapter.

Chapter 3

Feature Combination and Model Selection

In this Chapter, we evaluate the suitability of three domain transforms, DFT, DCT and DWT for recognition of human promoter sequences.

3.1 Raw Data

Here the human promoter sequences are collected using human genome built 35 from the NCBI site and two tools, PromoSer [Halees and Weng, 2004] and FIE2.1 [Chong *et al*., 2003]. In total, 14,001 promoter sequences with the length of 1024bp are used. They are the gene segments covering the range [-512, +512] relative to TSS (+1). The same number of 'non-promoter' sequences is selected by extracting segments of length 1024bp from randomly chosen chromosomal positions. The number of sequences extracted from one chromosome is proportional to the chromosome length. Thus, we obtain a set of sequences that have very low probability of containing any significant proportion of transcriptional regulatory segments [Zhang *et al*., 2004].

3.2 Training and testing set

	GC-content	# of non-promoters	# of promoters
Group1	G+C>80%	0	17
Group2	78%<GC<=80%	1	28
Group3	76%<GC<=78%	1	96

Group4	74%<GC<=76%	6	198
Group5	72%<GC<=74%	8	312
Group6	70%<GC<=72%	31	484
Group7	68%<GC<=70%	42	587
Group8	66%<GC<=68%	78	798
Group9	64%<GC<=66%	135	841
Group10	62%<GC<=64%	211	1032
Group11	60%<GC<=62%	305	989
Group12	58%<GC<=60%	368	1091
Group13	56%<GC<=58%	441	1019
Group14	54%<GC<=56%	541	1092
Group15	52%<GC<=54%	701	952
Group16	50%<GC<=52%	869	855
Group17	48%<GC<=50%	1055	772
Group18	46%<GC<=48%	1073	556
Group19	44%<GC<=46%	1239	485
Group20	42%<GC<=44%	1285	420
Group21	40%<GC<=42%	1338	339
Group22	GC<40%	4273	1038
Sum		14001	14001

Table 3.1 Data in 22 groups split by GC-content [Zhang *et al*., 2004].

Details of the data are given in Table 3.1. To eliminate the influence of the GC-content in my analysis, we divided all sequences into 22 groups by their GC-content first. Next, sequences in each group are randomly ordered and further divided into training and testing data. For the training data, we use the same number of promoter and non-promoter sequences but the number is different for different groups. For the four groups with the highest GC-content, it is not practical to make such a split since only small proportion of non-promoters is available. The information is summarized in Table 3.1. After that, for each of the data groups the same protocol of feature generation has been applied.

3.3 Features and Classification

3.3.1 Algorithm

The promoters and non-promoters are divided into 22 disjoint groups based on their GC-content. We examine three well-known domain transforms, DFT, DCT, DWT, for generating features to be used in the classification algorithm. The number of single nucleotides, di-nucleotides and tri-nucleotides in the sequences is initially determined and these account for first 84 features (4 for single nucleotides; 16 for di-nucleotides; 64 for tri-nucleotides). In addition to these, we add features of signals' coefficients under individual transforms (1024 from DFT, 512 from DCT, and 256 from DWT). Different transform methods such as DFT, DCT, and DWT, and their combinations are tried. DWT is based on two levels of decomposition and only the low resolution.

Once the feature vectors have been generated for the sequences in the group, the feature selection process is applied to select the most prominent features for classification. The top 30 features determined based on the Mahalanobis distance between the promoter and non-promoter data is used.

In statistics, Mahalanobis distance is a distance measure invented by P. C. Mahalanobis in 1936. Mahalanobis distance is the distance between two points scaled by the statistical variation in each component of the point.

The statistical distance or Mahalanobis distance between two points $x = (x_1,...,x_p)$ and $y = (y_1,...,y_p)_t$ is defined as: $d_s(x,y) = \sqrt{(x-y)^t S^{-1}(x-y)}$, p is the number of dimension of the space, and S is the covariance matrix. And the norm of x is defined as: $d_s(x,0) = \sqrt{x^t S^{-1} x}$.

Mahalanobis distance takes into account correlations, which means that there are associations between the variables. Feature vectors whose elements are quantities having different ranges and amounts of variation can be compared using Mahalanobis distance.

In each group, the data is first divided into training and test sets after random ordering of positive and negative data before it is further divided into two sets. Standard linear discriminant analysis (LDA) is then applied to separate promoter (positive) and non-promoter (negative) sequences. The models for each group are trained on the training dataset and applied to the test dataset.

.

3.3.2 Feature description

Previous (described in Chapter 2) results showed that the CC of each sequence with the mean of the reconstructed positive signal from the transformation domain, which is obtained respectively by DFT, DCT, and DWT, is not a proper feature to discriminate promoters and non-promoters.

Here we systematically examine the effects of using the coefficients of the signal under the three domain transforms: DFT, DCT, and DWT. There will be 7 kinds of combinations of features. The first 84 features are the number of different combinations of mono-, di-, tri- nucleotides. These 84 features' detail is described below and they will be accompanied with the combination of the 3 kind of transforms.

Features (there are 1877 features in the order described as below):
1-84: Nucleotides combination
1. a

2. c
3. g
4. t
5. aa
6. ac
7. ag
8. at
9. ca
10. cc
11. cg
12. ct
13. ga
14. gc
15. gg
16. gt
17. ta
18. tc
19. tg
20. tt
21. aaa
22. aac
23. aag
24. aat
25. aca
26. acc
27. acg
28. act
29. aga
30. agc
31. agg
32. agt
33. ata
34. atc
35. atg
36. att
37. aa
38. ac
39. ag
40. cat
41. cca
42. ccc
43. ccg
44. cct
45. cga
46. cgc
47. cgg
48. cgt
49. cta
50. ctc
51. ctg
52. ctt
53. gaa
54. gac
55. gag
56. gat
57. gca
58. gcc
59. gcg
60. gct

61. gga
62. ggc
63. ggg
64. ggt
65. gta
66. gtc
67. gtg
68. gtt
69. taa
70. tac
71. tag
72. tat
73. tca
74. tcc
75. tcg
76. tct
77. tga
78. tgc
79. tgg
80. tgt
81. tta
82. ttc
83. ttg
84. ttt
85-596: DCT coefficients: 512
597-1620: DFT coefficients: 1024
1621-1876:DWT coefficients: 256

For the 7 combinations of features, their number of features is as follows:

1. 84+256(DWT)=340

2. 84+512(DCT)=596

3. 84+1024(DFT)=1108

4. 84+256(DWT)+512(DCT)=852

5. 84+256(DWT)+1024(DFT)=1364

6. 84+512(DCT)+1024(DFT)=1620

7. 84+256(DWT)+512(DCT)+1024(DFT)=1876

3.3.3 Experiments

As for the dataset used here, the length of individual sequences is 1024bp. The sequence is with the range of [-512, +512] relative to TSS. The data is firstly divided into 22 groups according to their GC-content. Feature selection is the most important issue

later. To find two or a few features that can best separate the promoter and non-promoter data is preferred.

The experiment is carried out in Matlab, C, WEKA environments with different focus. Matlab is used to quickly test the performance of different ideas. Based on the performance, we can find out what features or what combination of features does work and what does not. The feature extraction algorithm is developed in C language, by which the speed is comparatively satisfactory; speed of the data processing is measured and improved. The tool named WEKA [Ian and Eibe, 2005] is a kind of commonly used software to do classification. Based on the results, what features are effective to give the best separation result can be quickly found out.

The terms we used here in data analysis includes those used in Chapter 2: Confusion matrix, TP rate, FP rate, PPV and F-measure. Their detailed definitions are given in Chapter 2.

	G+C content (1024componet)	# of Neg	# of Posi	#of(N=P) Train	#of Test	Confusion matrix	TP rate	FP rate	PPV	F-measure
Group1	G+C>80%	0	17	0/0	0/17					
Group2	78%<G+C<=80%	1	28	0/0	1/28					
Group3	76%<G+C<=78%	1	96	0/0	1/96					
Group4	74%<G+C<=76%	6	198	3/3	3/195	2 1 119 76	0.390	0.333	0.987	0.559
Group5	72%<G+C<=74%	8	312	4/4	4/308	0 4 54 254	0.825	1	0.984	0.898
Group6	70%<G+C<=72%	31	484	15/15	16/469	5 11 166 303	0.646	0.688	0.965	0.774
Group7	68%<G+C<=70%	42	587	21/21	21/566	11 10 214 352	0.622	0.476	0.972	0.759
Group8	66%<G+C<=68%	78	798	39/39	39/759	16 23 134 625	0.823	0.59	0.965	0.888
Group9	64%<G+C<=66%	135	841	67/67	68/774	48 20 197 577	0.745	0.294	0.966	0.842
Group10	62%<G+C<=64%	211	1032	100/100	111/932	84 27 218 714	0.766	0.243	0.964	0.854
Group11	60%<G+C<=62%	305	989	100/100	205/889	157 48 210 679	0.764	0.234	0.934	0.84
Group12	58%<G+C<=60%	368	1091	100/100	268/991	177 91 250 741	0.748	0.34	0.891	0.813
Group13	56%<G+C<=58%	441	1019	100/100	341/919	247 94 224 695	0.756	0.276	0.881	0.814
Group14	54%<G+C<=56%	541	1092	100/100	441/992	324 117 270 722	0.728	0.265	0.861	0.789
Group15	52%<G+C<=54%	701	952	100/100	601/852	456 145 221 631	0.741	0.241	0.813	0.775
Group16	50%<G+C<=52%	869	855	100/100	769/755	553 216 186 569	0.754	0.281	0.725	0.739
Group17	48%<G+C<=50%	1055	772	100/100	955/672	657 298 153 519	0.772	0.312	0.635	0.697
Group18	46%<G+C<=48%	1073	556	100/100	973/456	747 226 130 326	0.715	0.232	0.591	0.647
Group19	44%<G+C<=46%	1239	485	100/100	1139/385	869 270 128 257	0.668	0.237	0.488	0.564
Group20	42%<G+C<=44%	1285	420	100/100	1185/320	822 363 146 174	0.544	0.306	0.324	0.406
Group21	40%<G+C<=42%	1338	339	100/100	1238/239	742 496 81 158	0.661	0.401	0.242	0.354
Group22	G+C<40%	4273	1038	100/100	4173/938	2059 2114 344 594	0.633	0.507	0.219	0.326
Overall		14001	14001	1449/1449	12552/12552	7976 4574 3445 8966	0.722	0.364	0.622	0.668

Table 3.2 Prediction result on training/test dataset

Table 3.2 is the details of data in this experiment. The number of the training samples is the number of half of the minimum of the negative and positive samples in Group 1 to 9, and is 100 in Group 10 to 22, respectively. Table 3.2 also gives the experiment record with classifier NaïveBayes.

3.3.4 Discussion on the design of a classifier

In a classification problem, the error rate measures the overall performance of the classifier. The error rate is the proportion of errors made over a whole set of samples. The error is defined as such a sample when it is incorrectly labelled by prediction. Similarly, when a sample is predicted as actually it should be, it is defined as a success.

The error rate on the training data is not a reliable predictor of the true error rate on new data, whose label is unknown and is defined as "test set". To predict the performance of a classifier, it is necessary to assess the error rate of a classifier on the test set, which does not play a part in the formation of the classifier.

The "training" data is used by one or more learning schemes to come up with the classifiers. The "validation" data is used to optimize parameters of those classifiers, or to select a particular one to make the relatively best performance for the system. The "test" data is used to calculate the error rate of the final, optimized scheme. Each of the three sets of "training data", "validation data" and "test data" must be chosen independently. In Chapter 4, we will specify what data we use respectively for these three data sets. For the experiment in this chapter, we use only the "training data" and the "validation data" and show the optimized system generated by them.

For simulated studies of experiments here, if the training sample set is large enough, a classifier will be well schemed; if the test sample set is large enough, the error estimate

will be done accurately. So when the data is sufficient enough, a large sample set can be used for training, and another independent large sample set for testing.

There is a dilemma for the circumstance when the data is not sufficient enough: to get a good classifier, we want to use as much of the data as possible for training; to get a good error estimate we want to use as much of it as possible for testing.

As shown in Table 3.3 below, practically in our experiment, 14001 positive sequences and 14001 negative sequences are used. 50% of the minimum of the positive data and the negative data in each group is used for training. The data is presented as: "number of non-promoter sequences" / "number of promoter sequences" in each group, e.g. "3/195 in the Test set of Group 4". It Is shown in Table 3.3 that the number of promoters (P) and non-promoters (N) is taken to be the same in the training set of individual groups.

	Training set (N = P)	Test set (N/P)
Group1	0/0	0/17
Group2	0/0	1/28
Group3	0/0	1/96
Group4	3/3	3/195
Group5	4/4	4/308
Group6	15/15	16/469
Group7	21/21	21/566
Group8	39/39	39/759
Group9	67/67	68/774
Group10	105/105	106/927
Group11	152/152	153/837
Group12	184/184	184/907
Group13	220/220	221/799
Group14	270/270	271/822
Group15	350/350	351/602
Group16	427/427	442/428
Group17	386/386	669/386
Group18	278/278	795/278
Group19	242/242	997/243

Group20	210/210	1075/210
Group21	169/169	1169/170
Group22	519/519	3754/519
Over all	3661/3661	10340/10340

Table 3.3 Training and test data set

3.4 Results

We performed seven experiments with features obtained using DFT, DCT, DWT and their possible combinations, based on the same setup.

DWT		DCT		DFT	
Se	PPV	Se	PPV	Se	PPV
1	**1**	1	1	1	1
1	**0.966**	1	0.966	1	0.966
1	**0.989**	1	0.989	1	0.989
0.559	0.973	0.554	0.982	**0.631**	**0.984**
0.461	**0.993**	0.458	0.979	0.471	0.98
0.55	**0.996**	0.507	0.979	0.397	0.989
0.701	**0.988**	0.597	0.991	0.454	0.977
0.777	**0.982**	0.76	0.981	0.747	0.978
0.788	**0.985**	0.767	0.98	0.753	0.981
0.765	**0.977**	0.761	0.978	0.736	0.977
0.753	**0.957**	0.749	0.951	0.759	0.955
0.713	**0.96**	0.711	0.954	0.713	0.957
0.73	**0.959**	0.72	0.955	0.726	0.951
0.72	**0.949**	0.707	0.945	0.708	0.951
0.716	**0.929**	0.709	0.928	0.714	0.921
0.72	**0.873**	0.738	0.849	0.755	0.812
0.635	**0.814**	0.65	0.78	0.655	0.76
0.662	0.584	0.683	0.585	**0.698**	**0.586**
0.613	0.458	0.646	0.452	**0.712**	**0.425**
0.586	0.28	0.643	0.292	**0.69**	**0.28**
0.541	0.197	**0.6**	**0.213**	0.606	0.202
0.651	0.189	**0.68**	**0.184**	0.64	0.185

Table 3.4 Performance under different transform [Zhang *et al*., 2004].

The results obtained for each of the 22 groups are summarized in Table 3.4 for each of the basic domain transforms. DWT, DCT and DFT result in Se of 0.7, 0.692, 0.68, and

PPV of 0.722, 0.706 and 0.7, respectively. For each group, we attempt to select the best performing transform. The selected cases are highlighted in Table 3.4 (shaded and in bold numbers). The best combined result produces Se = 0.7131 and PPV = 0.7122. Other combinations produce similar, but inferior results [Zhang *et al*., 2004].

3.5 Discussion and conclusion

The experiments show certain consistent patterns. For example, the ability to separate promoters from non-promoters reduces significantly with the reduction of GC-content. In the three top ranked GC-groups, we observe sensitivity of 1 and PPV of over 0.96, while in the lowest GC-content groups these degrade to 0.6 and 0.19, respectively. This can be explained by the specific properties of regulatory regions in promoters with higher GC-content. These regions include most of the house-keeping genes. While those with the lower GC-content may predominantly be tissue-specific and could account for greater variability in their promoter content.

Another important observation is that the use of three different domain transforms does not result in dramatically different performance in classification of promoters and non-promoters. This suggests that all three domain transforms could provide useful information that could be integrated with information from biological features to predict promoters. Since information from biological features is not 'correlated' with that obtained via domain transforms, the classification performance should be significantly improved.

Thirdly, we observe that combining results from the three domain transforms does improve the classification performance to some extent. The best performance with Se = 71.31% and PPV = 71.22% was achieved by combining all three transforms. This greatly increased efficiency in prediction, compared to results of Se=9.92% and PPV=36.17% by use of feature of CC in Chapter 2. When no feature selection is done and the whole set of features are used, no significant change has been observed. This suggests that there is significant correlation between the information obtained from the three domain transforms.

In conclusion, the use of domain transforms for predicting human promoters is promising and should be combined with more of other physical, statistical or biological features of promoter regions to achieve better performance results. Also, the reduction of features has to be done on a case to case basis.

Chapter 4

Finding Starting Position of a Gene by Promoter Prediction System

In this Chapter, we first describe the details of the prediction system developed in C language module by module. We use Visual C++6.0 compiler. We aim to find the most probable position at which the TSS is located along DNA. The concept used is introduced and the efficiency of my scheme is discussed. Finally the prediction results are given when the system is applied to human chromosome 22 (NCBI, built 35). Based on the results obtained, the conclusion about the effectiveness of the features extracted is drawn and the SVM classifier models are finalized.

4．1 System description

4.1.1 Training the system

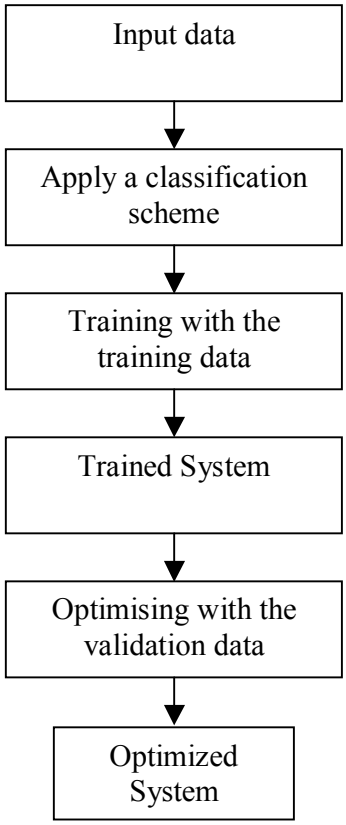


Figure 4.1 The depiction of the system structure relevant for 'training' and 'optimization'

Figure 4.1 is the simplified structure of the system during training and optimization before it is applied to do promoter prediction on human chromosomes. Models are trained with the training data and optimized with the validation data. The performance evaluation is made with different combination of features, different parameters and different kernel functions tried in the classifier. The data set comprises of 14001 positive sequences and 14001 negative sequences. 1449 positive and 1449 negative sequences are used as training data and 12552 positive and 12552 negative sequences are used as validation data, respectively. The test data is not included in this stage, and is applied in the stage of prediction shown in Figure 4.2 below. The models for each of the 22 groups of data are optimized one by one. The optimization is done with the most proper combination of features, parameters and kernel functions when the system gives the best overall performance on the validation data. Details will be presented in later

part of this chapter.

4.1.2 Predict the TSS position along human chromosome

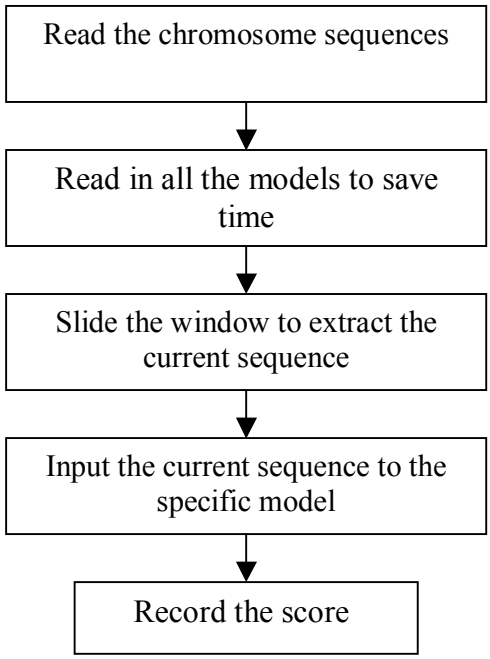


Figure 4.2 Depiction of the final prediction system

Figure 4.2 is depiction of the final prediction system which has already been trained based on the training data and optimized based on the validation data in the previous stage. The steps of how the system does promoter predictions on one half of human chromosome 22 are given below.

1. Open the file named "Homo_sapiens.NCBI35.dec.dna.chromosome.22.fa", which contains the chromosome 22 sequence data;

2. Extract a sequence of length 1024bp from the chromosome sequence by using a sliding window and by neglecting any sequence containing "N";

3. Calculate the first 84 features of the combination of number of single nucleotide, di-nucleotides, and tri-nucleotides. Based on the sequence's GC content (GC content= (#G + #C)/Sequence Length, where #G and #C are the total number of G and C nucleotide), divide the sequence into 22 groups.

4. Based on the sequence's group number, decide the 'combination' of DCT/FT/DWT features to calculate all the features needed and select the model needed in promoter prediction(the model is trained and optimized in previous experiment group by group);

5. Write all the calculated features into input files to be classified later by the trained system;

6. Classify this sequence with the model which is already saved for each Group;

7. Read the "value of decision function" from the file named "prediction.txt". If it is above or equal to zero, print this value and the order of this sequence into the file named "final_report.txt". If it is negative, print the value to the file named "nega.txt".

8. Move the sliding window by a defined step along the chromosome to extract the next sequence until the window reaches the last 1024bps to the end of the chromosome;

9. Draw the distribution plot of the scores with their corresponding positions along the human chromosome.

4.2 SVM used in classification

The idea of Support Vector Machine (SVM) is to separate data from different categories by a hyperplane, after mapping the data into a sufficiently high dimension with an appropriate non-linear mapping function [Duda *et al*., 2001]. SVM is trained to obtain the largest margin to separate the different classes of data. The larger the margin that

can be found, the better generalization ability of the classifier can be obtained in the future.

The support vectors are the training samples that are the most difficult to classify, and they decide the optimal separating hyperplane. Training an SVM involves finding the optimal hyperplane, which is the one with the maximum genomic margin over it. So support vectors are the training samples that are most informative for the classification task. When SVM is applied in classification problems, generalization control is obtained by maximizing the margin, or to minimize the weight vector correspondingly. The support vectors obtained as the solution can be sparse. These support vectors lie on the boundary and in this way summarize the information required to separate the data [Gunn, 1998].

To train an SVM, the commonly used method is the perceptron learning rule. The perceptron learning rule is to update the weight vector by an amount proportional to any misclassified patterns that are randomly selected. There is a simple method of training SVM, conceptually based on a small modification to this perceptron training rule. An SVM can be trained by choosing the current worst pattern in classification. During the training period, in most cases, such a pattern is one on the wrong side of the current decision boundary----the farthest side from that boundary. At the end of the training period, such a pattern will be one of the support vectors.

But this method is still only suitable for small number of data, since for each update, a search through the entire training set needs to be done to find the worst-classified pattern. For instance, if there are ‘n’ points in the training set, an SVM can be trained on the ‘n-1’ points of them, and the single remaining point can be test on. There will be an

error corresponding to each support vector. Thus, the optimal hyperplane that will separate the data is needed, so that the expected number of support vectors is small, and then the expected error rate will be lower.

Support vector machine tends to be less likely to suffer the problem of over fitting than some other methods. The complexity of the trained classifier is characterized by the number of support vectors rather than the dimensionality of the transformed space [Duda *et al*., 2001]. Due to this advantage I finally choose SVM classifier for our promoter prediction system.

4.3 Tuning the model

4.3.1 The features applied

Kernel functions are used in SVM to construct a mapping from the input feature space into a high dimensional feature space. The idea of the kernel function used to transform input data is to enable operations performed in the input space to be performed in a mapped new space. That is to say operation is not necessarily to be done in the input space, which is the potentially high dimensional feature space.

This provides a promising solution to problems in which potentially high dimensionality is involved. However, the computation is still critically dependent upon the number of training samples. Also, for the purpose of providing a good data distribution for a high dimensional problem, a large training set will generally be required [Gunn, 1998].

Here in my experiment, the 7 combinational features of the data are systematically applied (defined in Chapter 3). The features are numbered as described below.

Features No.1-84: the nucleotides combination features.

These first 84 features are the number of single nucleotide, di-nucleotides and tri-nucleotides in the original nucleotide sequence. Features from No.1-4 are the number of single nucleotides of 'a', 'c', 'g', and 't'; features from No.5-20 are the number of di-nucleotides of 16 kinds of combination with two nucleotides of 'a', 'c', 'g', and 't'; and features from No.21-84 are the number of tri-nucleotides of 64 kinds of combination of with three nucleotides of 'a', 'c', 'g', and 't'.

Features No.85-596: DCT coefficients

Features from No. 85-596 are the 512 coefficients of the signal after DCT transform.

Features No.597-1620: DFT coefficients

Features from No. 597-1620 are the 1024 coefficients of the signal after DFT transform.

Features No.1621-1876: DWT coefficients

Features from No. 1621-1876 are the 256 coefficients of the signal after DWT transform.

The 84 Nucleotides combination features will be accompanied with the combination of the 3 kinds of transformations (DFT, DCT, and DWT) to be 7 combinational different sets of features. The seven tables are given below to show the parameters and the performances:

1. Features**:** 84+256(DWT)=340

	Op_c (F-	Se	PPV

	measure)		
Group4	0.000001	91.17	98.87
Group5	0	68.27	99.09
Group6	0.01	45.66	99.04
Group7	0	83.81	99.29
Group8	0.000001	98.28	96.41
Group9	0.000001	98.11	94.25
Group10	0.0001	85.85	97.13
Group11	0.000001	90.72	91.67
Group12	0.0001	77.71	97.14
Group13	0.01	72.98	96.71
Group14	1	73.96	93.11
Group15	1	72.48	76.95
Group16	0.0001	61.54	69.14
Group17	1	71.95	50.43
Group18	0.0001	73.08	34.91
Group19	0.0001	72.89	27.94
Group20	0.01	73.01	23.71
Group21	0.0001	71.9	22.36
Group22	0.0001	64.81	16.88
overall		79.700249	71.409142

Table 4.1 The parameters with the 1st set of features

2. Features: 84+512(DCT) =596

	Op_c (F-measure)	Se	PPV
Group4	0.000001	91.17	98.87
Group5	0	68.27	99.09
Group6	1000	45.66	99.36
Group7	0.000001	83.93	99.29
Group8	0.000001	98.28	96.41
Group9	0.000001	98.11	94.25
Group10	0.0001	85.85	97.13
Group11	0.000001	90.62	91.66
Group12	0.0001	77.71	97.14
Group13	0.0001	72.72	96.87
Group14	0.01	74.3	90.69
Group15	0.0001	64.53	88.28
Group16	0.0001	61.54	69.71
Group17	0.0001	61.38	53.93
Group18	0.0001	73.08	35.19
Group19	0.0001	72.89	28.07

Group20	0.0001	77.3	23.08
Group21	0.0001	71.24	22.34
Group22	0.0001	65.48	17.1
overall		79.292351	71.658806

Table 4.2 The parameters with the 2nd set of features

3. Features**:** 84+1024(DFT)=1108

	Op_c (F-measure)	Se	PPV
Group4	0	91.43	98.88
Group5	0	68.27	99.09
Group6	1000	46.39	99.37
Group7	0.000001	84.17	99.29
Group8	0.000001	98.28	96.41
Group9	0.000001	98.11	94.25
Group10	0.0001	85.95	97.13
Group11	0.000001	90.62	91.66
Group12	0.0001	77.71	97.26
Group13	0.0001	72.72	96.87
Group14	0.0001	68.33	97.4
Group15	0.0001	64.53	88.28
Group16	0.0001	61.9	70.12
Group17	0.0001	62.2	54.06
Group18	0.0001	73.08	35.28
Group19	0.0001	72.29	27.97
Group20	0.0001	77.3	23.08
Group21	0.0001	71.24	22.15
Group22	0.0001	65.26	17.01
overall		79.169037	71.716431

Table 4.3 The parameters with the 3rd set of features

4. Features: 84+256(DWT)+512(DCT)=852

	Op_c (F-measure)	Se	PPV
Group4	0.000001	91.17	98.87
Group5	0	68.27	99.09
Group6	1000	45.66	99.36
Group7	0.000001	83.93	99.29
Group8	0.000001	98.28	96.41
Group9	0.000001	98.11	94.25
Group10	0.0001	85.85	97.13
Group11	0.000001	90.62	91.66

Group12	0.0001	77.71	97.14
Group13	0.0001	72.72	96.87
Group14	0.01	74.3	90.69
Group15	0.0001	64.53	88.28
Group16	0.0001	61.54	69.71
Group17	0.0001	61.38	53.93
Group18	0.0001	73.08	35.09
Group19	0.0001	72.89	28.07
Group20	0.0001	77.3	23.08
Group21	0.0001	71.24	22.34
Group22	0.0001	65.48	17.1
overall		79.210144	71.695145

Table 4.4 The parameters with the 4th set of features

5. Features: 84+256(DWT)+1024(DFT)=1364

	Op_c (F-measure)	Se	PPV
Group4	0	91.43	98.88
Group5	0.000001	68.27	99.09
Group6	1000	46.39	99.37
Group7	0.000001	84.17	99.29
Group8	0.000001	98.28	96.41
Group9	0.000001	98.11	94.25
Group10	0.0001	85.95	97.13
Group11	0.000001	90.62	91.66
Group12	0.0001	77.71	97.26
Group13	0.0001	72.72	96.87
Group14	0.0001	68.33	97.4
Group15	0.0001	64.53	88.28
Group16	0.0001	61.54	70
Group17	0.0001	62.2	54.06
Group18	0.0001	73.08	35.28
Group19	0.0001	72.29	27.97
Group20	0.0001	77.3	23.08
Group21	0.0001	71.24	22.15
Group22	0.0001	65.26	17.01
overall		79.03624	71.771904

Table 4.5 The parameters with the 5th set of features

6. Features: 84+512(DCT)+1024(DFT)=1620

	Op_c (F-measure)	Se	PPV
Group4	0	91.43	98.88

Group5	0.000001	68.27	99.09
Group6	0.01	46.69	99.37
Group7	0.000001	84.17	99.29
Group8	0.000001	98.28	96.41
Group9	0.000001	98.22	94.26
Group10	0.0001	85.95	97.03
Group11	0.000001	90.52	91.65
Group12	0.0001	77.71	97.26
Group13	0.0001	72.85	96.88
Group14	0.0001	68.49	97.41
Group15	0.0001	64.53	88.28
Group16	0.0001	61.17	69.01
Group17	0.0001	62.6	54.42
Group18	0.0001	73.08	35.37
Group19	0.0001	72.29	27.97
Group20	0.0001	77.3	23.08
Group21	0.0001	71.24	22.15
Group22	0.0001	65.48	17.13
overall		79.083664	71.821159

Table 4.6 The parameters with the 6th set of features

7. Features: 84+256(DWT)+512(DCT)+1024(DFT)=1876

	Op_c (F-measure)	Se	PPV
Group4	0	91.43	98.88
Group5	0.000001	68.27	99.09
Group6	0.01	46.69	99.37
Group7	0.000001	84.17	99.29
Group8	0.000001	98.28	96.41
Group9	0.000001	98.22	94.26
Group10	0.0001	85.95	97.03
Group11	0.000001	90.52	91.65
Group12	0.0001	77.71	97.26
Group13	0.0001	72.85	96.88
Group14	0.0001	68.49	97.41
Group15	0.0001	64.53	88.28
Group16	0.0001	61.17	69.29
Group17	0.0001	62.6	54.42
Group18	0.0001	73.08	35.37
Group19	0.0001	72.29	27.97
Group20	0.0001	77.3	23.08
Group21	0.0001	71.24	22.15
Group22	0.0001	65.48	17.12

overall		79.083664	71.821159

Table 4.7 The parameters with the 7th set of features

Table 4.1-7 summarizes the results of the experiments using the 7 sets of features, which are automatically generated by the system. The first column is the optimization of parameter C to produce the best F-measure, and the second and third column is the SE and PPV value under this optimal setting. The parameter C is the trade-off between training error and margin (defaulted as being defined as [arg.x*x]^-1). Here I adopt one of the most popular measures called the F-measure: $F_\beta = \frac{PR}{(1-\beta)P + \beta R}$. The trade-offs between competing objectives is controlled by the variable β . When β =0.5, the F-measure is the harmonic mean of precision and recall [Fisher *et al*., 2004]. During the experiment, I set $\beta = 0.5$, to stress the equal importance of precision and recall. It can be observed from the tables, that for different group, the optimal choice of combination of features may be different to obtain the best performance. For the final settings of the system, the optimal model should be selected group by group respectively to obtain the optimal overall recall and precision level for the whole data set.

4.3.2 Find the appropriate kernel

An SVM is largely characterized by the choice of its kernel function. Thus SVM connect the problem they are designed for to a large body of existing research on kernel-based methods [Wong, 2004]. I focus to tune the models by finding the optimal kernel function as well as the optimal important parameters to optimize the models' performance in classifying each group of data, respectively.

4.3.3 Tuning the models

Careful tuning is required to achieve the best performance of the system in recognition of TSS in a large-scale promoter search. The general goal of tuning is to maximize the level of true positives versus false positives over whole data set and at the same time maintaining a satisfying sensitivity level. Different models are trained and each is tuned for the best performance in each group respectively. That is to say, I aim at producing the highest PPV.

The tuning process can thus be considered as an optimization process with two goals-to maximize sensitivity and maximize the positive predictive value.

To develop a set of optimal models of the system, I need to tune a large number of parameters. However, to optimize the parameters usually involves going towards multiple competing objectives. And what balances to be set between precision and specificity (recall) in the system also should be considered. Since this two values will not be high or low simultaneously. That is to say that it is difficult to obtain good values for both precision and recall concurrently. Here I adapt one of the most popular measures called the F-measure: $F_\beta = \frac{PR}{(1-\beta)P + \beta R}$. The trade-offs between competing objectives is controlled by the variable β. When β =0.5, the F-measure is the harmonic mean of precision and recall. By setting the value of β, the relative importance of precision and recall to the system can be given in advance [Fisher *et al.*, 2004]. In my project, I set $\beta = 0.5$, to stress the equal importance of precision and recall.

Select the proper kernel:

The kernels applied in the experiments include: linear kernel, polynomial kernel----(s a*b+c)^d, radial basis function kernel-----exp(-gamma ||a-b||^2), sigmoid kernel----tanh(s a*b + c).

During the experiment, I set $\beta = 0.5$, to stress the equal importance of precision and recall. The parameter C is the trade-off between training error and margin (defaulted as being defined as [arg.x*x]^-1), and this parameter C has the same effect as it is defined in all kinds of kernel functions.

Only the tables obtained by radial basis function kernel and polynomial kernel are given below as two examples of our experiment results. Other tables with other kernel functions are attached in Appendix B.

	c=1.000000		c=1.000000		c=1.000000		c=1.000000		c=1.000000	
	g=0.000010		g=0.000100		g=1.000000		g=10.000000		g=100.000000	
Group #	Se	PPV	Se	PPV	Se	PPV	Se	PPV	Se	PPV
	TP	FN	TP	FN	TP	FN	TP	FN	TP	FN
	FP	TN	FP	TN	FP	TN	FP	TN	FP	TN
Group4	87.01	98.82	78.96	98.7	0	0	0	0	0	0
	0	4	0	4	0	4	0	4	0	4
	50	335	81	304	385	0	385	0	385	0
Group5	66.39	99.07	57.83	100	0	0	0	0	0	0
	2	3	5	0	0	5	0	5	0	5
	161	318	202	277	479	0	479	0	479	0
Group6	42.71	99.66	60.68	99.76	0	0	0	0	0	0
	12	1	12	1	0	13	0	13	0	13
	389	290	267	412	679	0	679	0	679	0
Group7	85.97	99.31	88.13	99.46	0	0	0	0	0	0
	13	5	14	4	0	18	0	18	0	18
	117	717	99	735	834	0	834	0	834	0
Group8	88.37	99.03	86.01	99.01	0	0	0	0	0	0
	37	8	37	8	1	44	1	44	1	44
	108	821	130	799	929	0	929	0	929	0
Group9	87.2	98.46	86.57	98.33	0	0	0	0	0	0
	64	13	63	14	0	77	0	77	0	77
	122	831	128	825	953	0	953	0	953	0
Group10	84.88	96.99	84.39	97.08	0	0	0	0	0	0
	100	27	101	26	0	127	0	127	0	127
	155	870	160	865	1025	0	1025	0	1025	0

Group11	77.84	98.18	74.64	98.37	0	0	0	0	0	0
	127	14	129	12	0	141	0	141	0	141
	215	755	246	724	970	0	970	0	970	0
Group12	76.56	97.22	74.79	97.42	0	0	0	0	0	0
	153	21	155	19	1	173	1	173	1	173
	225	735	242	718	960	0	960	0	960	0
Group13	72.06	97.35	69.06	97.42	0	0	0	0	0	0
	199	15	200	14	0	214	0	214	0	214
	214	552	237	529	766	0	766	0	766	0
Group14	67	96.88	65.84	98.02	0	0	0	0	0	0
	259	13	264	8	0	272	0	272	0	272
	199	404	206	397	603	0	603	0	603	0
Group15	60.86	91.71	62.39	89.47	0	0	0	0	0	0
	351	18	345	24	0	369	0	369	0	369
	128	199	123	204	327	0	327	0	327	0
Group16	47.62	80.25	53.11	71.43	0	0	0	0	0	0
	552	32	526	58	0	584	0	584	0	584
	143	130	128	145	273	0	273	0	273	0
Group17	48.37	66.11	58.94	52.35	0	0	0	0	0	0
	756	61	685	132	0	817	0	817	0	817
	127	119	101	145	246	0	246	0	246	0
Group18	60.44	28.5	72.53	34.29	0.55	0.11	0.55	0.11	0.55	0.11
	631	276	654	253	0	907	0	907	0	907
	72	110	50	132	181	1	181	1	181	1
Group19	80.72	18.38	66.27	23.35	0	0	0	0	0	0
	428	595	662	361	0	1023	0	1023	0	1023
	32	134	56	110	166	0	166	0	166	0
Group20	84.05	17.28	74.23	20.27	0	0	0	0	0	0
	435	656	615	476	0	1091	0	1091	0	1091
	26	137	42	121	163	0	163	0	163	0
Group21	72.55	17.1	72.55	18.85	0	0	0	0	0	0
	603	538	663	478	0	1141	0	1141	0	1141
	42	111	42	111	153	0	153	0	153	0
Group22	85.3	13.53	71.71	17.6	0	0	0	0	0	0
	1487	2448	2427	1508	2	3933	2	3933	2	3933
	66	383	127	322	449	0	449	0	449	0

Table 4.8a Experiment result when an RBF kernel is applied

	Op_c(F-measure)	Op_g(F-measure)	Se	PPV
Group4	1	0.00001	87.01	98.82
Group5	1	0.00001	66.39	99.07
Group6	1	0.0001	60.68	99.76
Group7	1	0.0001	88.13	99.46
Group8	1	0.00001	88.37	99.03
Group9	1	0.00001	87.2	98.46
Group10	1	0.00001	84.88	96.99
Group11	1	0.00001	77.84	98.18
Group12	1	0.00001	76.56	97.22
Group13	1	0.00001	72.06	97.35

Group14	1	0.00001	67	96.88
Group15	1	0.0001	62.39	89.47
Group16	1	0.0001	53.11	71.43
Group17	1	0.00001	48.37	66.11
Group18	1	0.0001	72.53	34.29
Group19	1	0.0001	66.27	23.35
Group20	1	0.0001	74.23	20.27
Group21	1	0.0001	72.55	18.85
Group22	1	0.0001	71.71	17.6
overall			76.190475	70.617195

Table 4.8b Optimal parameters for each group of data

Table 4.8 is the result obtained with the radial basis function kernel. The parameter g is the parameter gamma in radial basis function kernel -----exp (-gamma ||a-b||^2). Table 4.8a records SE, PPV and confusion matrix (TP, FN; FP, TN) for the data of each group, under the parameter combinations of C and g. The optimal parameter combinations of C and g for each group of data are indicted in Table 4.8b. The overall SE and PPV is 76.19% and 70.62%, respectively.

	c=0		c=0		c=0.000001		c=0.000001		c=0.0001		c=0.0001		c=0.01		c=0.01		c=1		c=1	
	d=2		d=3		d=2		d=3		d=2		d=3		d=2		d=3		d=2		d=3	
Group #	Se	PPV	Se	PPV	Se	PPV	Se	PPV	Se	PPV	Se	PPV	Se	PPV	Se	PPV	Se	PPV	Se	PPV
	TP	FN	TP	FN	TP	FN	TP	FN	TP	FN	TP	FN	TP	FN	TP	FN	TP	FN	TP	FN
	FP	TN	FP	TN	FP	TN	FP	TN	FP	TN	FP	TN	FP	TN	FP	TN	FP	TN	FP	TN
Group4	92.21	98.89	0	0	69.35	98.52	69.35	98.52	69.35	98.52	69.35	98.52	69.35	98.52	69.35	98.52	69.35	98.52	69.35	98.52
	0	4	0	4	0	4	0	4	0	4	0	4	0	4	0	4	0	4	0	4
	30	355	385	0	118	267	118	267	118	267	118	267	118	267	118	267	118	267	118	267
Group5	69.1	99.1	0	0	56.78	99.27	56.78	99.27	56.78	99.27	56.78	99.27	56.78	99.27	56.78	99.27	56.78	99.27	56.78	99.27
	2	3	0	5	3	2	3	2	3	2	3	2	3	2	3	2	3	2	3	2
	148	331	479	0	207	272	207	272	207	272	207	272	207	272	207	272	207	272	207	272
Group6	32.84	99.55	0	0	46.98	99.38	46.98	99.69	46.98	99.38	46.98	99.69	46.98	99.38	46.98	99.69	46.98	99.38	46.98	99.69
	12	1	0	13	11	2	12	1	11	2	12	1	11	2	12	1	11	2	12	1
	456	223	679	0	360	319	360	319	360	319	360	319	360	319	360	319	360	319	360	319
Group7	84.89	99.3	0	0	52.88	98.22	55.52	98.51	52.88	98.22	55.52	98.51	52.88	98.22	55.52	98.51	52.88	98.22	55.52	98.51
	13	5	0	18	10	8	11	7	10	8	11	7	10	8	11	7	10	8	11	7
	126	708	834	0	393	441	371	463	393	441	371	463	393	441	371	463	393	441	371	463
Group8	80.3	99.07	0	0	81.92	98.58	82.13	98.71	81.92	98.58	82.13	98.71	81.92	98.58	82.13	98.71	81.92	98.58	82.13	98.71
	38	7	0	45	34	11	35	10	34	11	35	10	34	11	35	10	34	11	35	10
	183	746	929	0	168	761	166	763	168	761	166	763	168	761	166	763	168	761	166	763
Group9	87.62	98.35	0	0	77.02	97.87	77.12	97.74	77.02	97.87	77.12	97.74	77.02	97.87	77.12	97.74	77.02	97.87	77.12	97.74
	63	14	0	77	61	16	60	17	61	16	60	17	61	16	60	17	61	16	60	17
	118	835	953	0	219	734	218	735	219	734	218	735	219	734	218	735	219	734	218	735
Group10	84.49	97.19	0	0	77.27	95.88	77.76	95.79	77.27	95.88	77.76	95.79	77.27	95.88	77.76	95.79	77.27	95.88	77.76	95.79
	102	25	0	127	93	34	92	35	93	34	92	35	93	34	92	35	93	34	92	35
	159	866	1025	0	233	792	228	797	233	792	228	797	233	792	228	797	233	792	228	797
Group11	78.56	98.07	0	0	74.33	95.75	74.85	95.78	74.33	95.75	74.85	95.78	74.33	95.75	74.85	95.78	74.33	95.75	74.85	95.78
	126	15	0	141	109	32	109	32	109	32	109	32	109	32	109	32	109	32	109	32
	208	762	970	0	249	721	244	726	249	721	244	726	249	721	244	726	249	721	244	726
Group12	76.88	97.23	0	0	74.06	95.31	74.17	95.57	74.06	95.31	74.17	95.57	74.06	95.31	74.17	95.57	74.06	95.31	74.17	95.57
	153	21	0	174	139	35	141	33	139	35	141	33	139	35	141	33	139	35	141	33
	222	738	960	0	249	711	248	712	249	711	248	712	249	711	248	712	249	711	248	712
Group13	72.06	97.18	0	0	73.24	91.97	74.02	92.05	73.24	91.97	74.02	92.05	73.24	91.97	74.02	92.05	73.24	91.97	74.02	92.05
	198	16	0	214	165	49	165	49	165	49	165	49	165	49	165	49	165	49	165	49
	214	552	766	0	205	561	199	567	205	561	199	567	205	561	199	567	205	561	199	567
Group14	66.83	96.88	0	0	73.96	86.94	74.46	87.35	73.96	86.94	74.46	87.35	73.96	86.94	74.46	87.35	73.96	86.94	74.46	87.35
	259	13	0	272	205	67	207	65	205	67	207	65	205	67	207	65	205	67	207	65
	200	403	603	0	157	446	154	449	157	446	154	449	157	446	154	449	157	446	154	449
Group15	60.55	90.41	0	0	76.15	66.76	75.23	67.21	76.15	66.76	75.23	67.21	76.15	66.76	75.23	67.21	76.15	66.76	75.23	67.21
	348	21	0	369	245	124	249	120	245	124	249	120	245	124	249	120	245	124	249	120

	129	198	327	0	78	249	81	246	78	249	81	246	78	249	81	246	78	249	81	246
Group16	49.45	72.19	0	0	66.67	54.65	65.2	52.35	66.67	54.65	65.2	52.35	66.67	54.65	65.2	52.35	66.67	54.65	65.2	52.35
	532	52	0	584	433	151	422	162	433	151	422	162	433	151	422	162	433	151	422	162
	138	135	273	0	91	182	95	178	91	182	95	178	91	182	95	178	91	182	95	178
Group17	54.88	45.76	0	0	70.73	41.73	71.54	42.51	70.73	41.73	71.54	42.51	70.73	41.73	71.54	42.51	70.73	41.73	71.54	42.51
	657	160	0	817	574	243	579	238	574	243	579	238	574	243	579	238	574	243	579	238
	111	135	246	0	72	174	70	176	72	174	70	176	72	174	70	176	72	174	70	176
Group18	51.1	24.54	0	0	68.13	33.7	66.48	33.33	68.13	33.7	66.48	33.33	68.13	33.7	66.48	33.33	68.13	33.7	66.48	33.33
	621	286	0	907	663	244	665	242	663	244	665	242	663	244	665	242	663	244	665	242
	89	93	182	0	58	124	61	121	58	124	61	121	58	124	61	121	58	124	61	121
Group19	83.73	17.75	0	0	58.43	24.13	60.84	24.88	58.43	24.13	60.84	24.88	58.43	24.13	60.84	24.88	58.43	24.13	60.84	24.88
	379	644	0	1023	718	305	718	305	718	305	718	305	718	305	718	305	718	305	718	305
	27	139	166	0	69	97	65	101	69	97	65	101	69	97	65	101	69	97	65	101
Group20	67.48	17.08	0	0	68.71	21.92	69.33	21.61	68.71	21.92	69.33	21.61	68.71	21.92	69.33	21.61	68.71	21.92	69.33	21.61
	557	534	0	1091	692	399	681	410	692	399	681	410	692	399	681	410	692	399	681	410
	53	110	163	0	51	112	50	113	51	112	50	113	51	112	50	113	51	112	50	113
Group21	66.01	16.78	0	0	66.01	18.4	65.36	18.62	66.01	18.4	65.36	18.62	66.01	18.4	65.36	18.62	66.01	18.4	65.36	18.62
	640	501	0	1141	693	448	704	437	693	448	704	437	693	448	704	437	693	448	704	437
	52	101	153	0	52	101	53	100	52	101	53	100	52	101	53	100	52	101	53	100
Group22	78.62	13.48	0	0	62.36	14.72	64.59	15.27	62.36	14.72	64.59	15.27	62.36	14.72	64.59	15.27	62.36	14.72	64.59	15.27
	1670	2265	0	3935	2313	1622	2326	1609	2313	1622	2326	1609	2313	1622	2326	1609	2313	1622	2326	1609
	96	353	449	0	169	280	159	290	169	280	159	290	169	280	159	290	169	280	159	290

Table 4.9a Experiment result when a polynomial kernel is applied

	Op_c(F-measure)	Op_g(F-measure)	Se	PPV
Group4	0	2	92.21	98.89
Group5	0	2	69.1	99.1
Group6	0.000001	3	46.98	99.69
Group7	0	2	84.89	99.3
Group8	0.000001	3	82.13	98.71
Group9	0	2	87.62	98.35
Group10	0	2	84.49	97.19
Group11	0	2	78.56	98.07
Group12	0	2	76.88	97.23
Group13	0	2	72.06	97.18
Group14	0.01	3	74.46	87.35
Group15	0	2	60.55	90.41
Group16	1	2	66.67	54.65
Group17	0.01	3	71.54	42.51
Group18	1	2	68.13	33.7
Group19	0.01	3	60.84	24.88
Group20	1	2	68.71	21.92
Group21	0.000001	3	65.36	18.62
Group22	0.01	3	64.59	15.27
overall			75.516983	68.962234

Table 4.9b Optimal parameters for each group of data

Table 4.9 is the result obtained with the polynomial kernel. The parameter d is the parameter in polynomial kernel ----- (s a*b+c)^d. The parameter d and s in polynomial kernel is not a very effective one to decide the performance of the model, since the change of these two parameters does not make evident changes in the performance of prediction using the model. Table 4.9a records SE, PPV and confusion matrix (TP, FN; FP, TN) for the data of each group, under the parameter combinations of C and d. The optimal parameter combinations of C and d for each group of data are indicted in the Table 4.9b. The overall SE and PPV is 75.52% and 68.96%, respectively.

Summary:

Most of the 22 groups of our data can use linear kernel as the optimal model, with a few using polynomial kernel of power 2 or 3. The parameter C, which is the trade-off between

training error and margin, is a very important parameter. The change of C has a big influence on the performance of prediction.

Future directions include: A technique for choosing the kernel function by computational means and how to design a kernel function to get a good generalization performance of SVM [Gunn, 1998].

As shown in Table 4.10, the data used in the experiment comprises of 14001 positive sequences and 14001 negative sequences. 3040 positive and 3040 negative sequences are used as training data and 10961 positive and 10961 negative sequences are used as test data, respectively. For each of the 22 groups, 50% of the minimum of the number of the positive and negative data is extracted as the training set. The performance in the measurements of TP, FP, SE and PPV from each group of Group1 to Group22 are shown in Table 4.10. The Overall Se and PPV are calculated with the two formulas as shown below.

	Training set	Test set	TP	FP	Se	PPV
Group1	0/0	0/57	57	0	1	1
Group2	0/0	1/120	120	1	1	0.992
Group3	0/0	3/242	242	3	1	0.988
Group4	4/4	4/385	355	4	0.922	0.989
Group5	4/4	5/479	331	3	0.691	0.991
Group6	13/13	13/679	412	1	0.607	0.998
Group7	17/17	18/834	735	4	0.881	0.995
Group8	45/45	45/929	913	34	0.983	0.964
Group9	76/76	77/953	936	57	0.982	0.943
Group10	127/127	127/1025	881	26	0.860	0.971
Group11	140/140	141/970	880	80	0.907	0.917
Group12	173/173	174/960	746	21	0.777	0.973
Group13	214/214	214/766	559	19	0.730	0.967
Group14	271/271	272/603	446	33	0.740	0.931
Group15	327/327	369/327	211	28	0.645	0.883
Group16	272/272	584/273	169	72	0.619	0.701
Group17	246/246	817/246	154	129	0.626	0.544
Group18	181/181	907/182	133	243	0.731	0.354
Group19	166/166	1023/166	121	310	0.729	0.281
Group20	163/163	1091/163	119	383	0.730	0.237
Group21	152/152	1141/153	110	382	0.719	0.224
Group22	449/449	3935/449	294	1422	0.655	0.171
Overall	3040/3040	10961/10961	8924	3255	0.814	0.733

Table 4.10 Performance result

Overall Se=TP/ALL POSITIVE=8924/10961=81.42%

Overall PPV=TP/(TP+FP)=8924/(8924+3255)=73.27%

Based on the results in Table 4.10, the SE and PPV of selected groups are summarized in Table 4.11. The 3rd column of "Se Total" is calculated with the TP number of selected groups divided by the total number of positives of all the 22 groups.

	Se	PPV	Se Total
Group1	1.0000	1.0000	0.0052
Group1→Group2	1.0000	0.9944	0.0161
Group1→Group3	1.0000	0.9905	0.0382
Group1→Group4	0.9627	0.9898	0.0706
Group1→Group5	0.8613	0.9901	0.1008
Group1→Group6	0.7732	0.9922	0.1384
Group1→Group7	0.8054	0.9929	0.2055
Group1→Group8	0.8497	0.9844	0.2888
Group1→Group9	0.8767	0.9746	0.3741
Group1→Group10	0.8736	0.9740	0.4545
Group1→Group11	0.8785	0.9649	0.5348
Group1→Group12	0.8657	0.9658	0.6029
Group1→Group13	0.8533	0.9659	0.6539
Group1→Group14	0.8457	0.9638	0.6946
Group1→Group15	0.8387	0.9614	0.7138
Group1→Group16	0.8324	0.9539	0.7292
Group1→Group17	0.8273	0.9405	0.7433
Group1→Group18	0.8255	0.9161	0.7554
Group1→Group19	0.8240	0.8872	0.7664
Group1→Group20	0.8225	0.8545	0.7773
Group1→Group21	0.8210	0.8248	0.7873
Group1→Group22	0.8142	0.7327	0.8142

Table 4.11 Performance when using different part of the dataset

The performance in the measurements of SE and PPV can be observed from Figure 4.3. The SE and PPV curves are drawn with the two columns of SE and PPV results from the 4th row to the 22nd row shown in Table 4.11. The y coordinates on the points with x coordinate of 4 are the SE and PPV value of the corresponding the experiment using the data from group 1 to group 4. The y coordinates on the points with x coordinate of 15 are the SE and PPV value of the corresponding experiment using the data from group 1 to group 15. The curve of SE does not change consistently: it drops sharply when the data of Group 5 and 6 are added; it rises when the data of Group 7, 8, 9, and 10 are added; and it gradually drops when the data from Group 11 to 22 are added. While the PPV value drops consistently when more data from those groups that are GC poor are included. So a finding can be made that promoters with higher GC-content can be predicted more accurately.

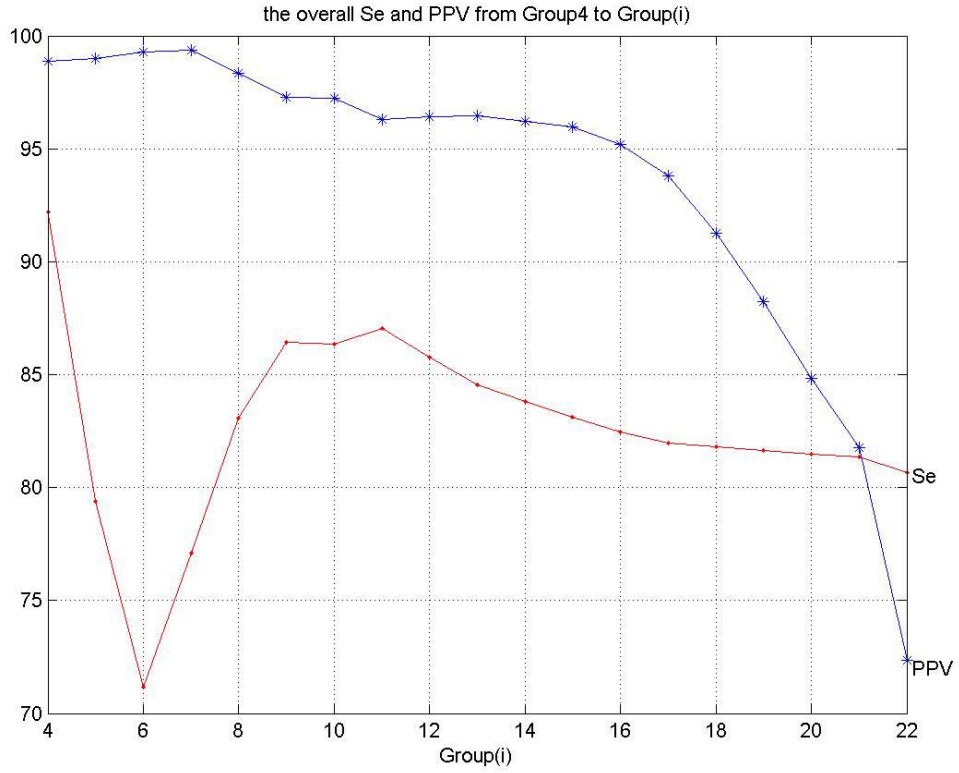


Figure 4.3 SE and PPV obtained with data from Group 4 to Group (i) (i=4 to 22)

4.3.4 Transductive versus Inductive SVM:

Traditional inductive SVM is popular in data mining, while transductive SVM is developed and expected to be more advanced to inductive SVM. The transductive training is different from inductive training in that the testing set can be used as an additional source of information for deciding margins besides the training set. That is to say, transductive SVMs take into account a particular test set and try to minimize misclassifications of just those particular examples in training procedure. In transduction, one estimates the classification function at points within the data set using information from both of the training and the test set data. This is contrast to the training procedure of Inductive SVMs. Thus, it is often expected that transductive SVM can be more powerful due to its ability to improve the SVM's generalization performance, especially in cases

such as when the training data are inadequate and when the training and test set sub samples are quite deviated from each other [Chen *et al*., 2003b].

However, for our data, the performance of prediction is not dramatically enhanced when replacing the inductive SVM with the transductive SVM. The tables given below are obtained from the experiment results using inductive and transductive SVM, respectively.

Inductive SVM:

Here the traditional inductive SVM with rbf kernel is applied.

	g=0.000010		g=0.000100		g=1.000000		g=10.000000		g=100.000000	
Group #	Se	PPV	Se	PPV	Se	PPV	Se	PPV	Se	PPV
	TP	FN	TP	FN	TP	FN	TP	FN	TP	FN
	FP	TN	FP	TN	FP	TN	FP	TN	FP	TN
Group4	87.01	98.82	78.96	98.7	0	0	0	0	0	0
	0	4	0	4	0	4	0	4	0	4
	50	335	81	304	385	0	385	0	385	0
Group5	66.39	99.07	57.83	100	0	0	0	0	0	0
	2	3	5	0	0	5	0	5	0	5
	161	318	202	277	479	0	479	0	479	0
Group6	42.71	99.66	60.68	99.76	0	0	0	0	0	0
	12	1	12	1	0	13	0	13	0	13
	389	290	267	412	679	0	679	0	679	0
Group7	85.97	99.31	88.13	99.46	0	0	0	0	0	0
	13	5	14	4	0	18	0	18	0	18
	117	717	99	735	834	0	834	0	834	0
Group8	88.37	99.03	86.01	99.01	0	0	0	0	0	0
	37	8	37	8	1	44	1	44	1	44
	108	821	130	799	929	0	929	0	929	0
Group9	87.2	98.46	86.57	98.33	0	0	0	0	0	0
	64	13	63	14	0	77	0	77	0	77
	122	831	128	825	953	0	953	0	953	0
Group10	84.88	96.99	84.39	97.08	0	0	0	0	0	0
	100	27	101	26	0	127	0	127	0	127
	155	870	160	865	1025	0	1025	0	1025	0
Group11	77.84	98.18	74.64	98.37	0	0	0	0	0	0
	127	14	129	12	0	141	0	141	0	141
	215	755	246	724	970	0	970	0	970	0
Group12	76.56	97.22	74.79	97.42	0	0	0	0	0	0
	153	21	155	19	1	173	1	173	1	173

	225	735	242	718	960	0	960	0	960	0
Group13	72.06	97.35	69.06	97.42	0	0	0	0	0	0
	199	15	200	14	0	214	0	214	0	214
	214	552	237	529	766	0	766	0	766	0
Group14	67	96.88	65.84	98.02	0	0	0	0	0	0
	259	13	264	8	0	272	0	272	0	272
	199	404	206	397	603	0	603	0	603	0
Group15	60.86	91.71	62.39	89.47	0	0	0	0	0	0
	351	18	345	24	0	369	0	369	0	369
	128	199	123	204	327	0	327	0	327	0
Group16	47.62	80.25	53.11	71.43	0	0	0	0	0	0
	552	32	526	58	0	584	0	584	0	584
	143	130	128	145	273	0	273	0	273	0
Group17	48.37	66.11	58.94	52.35	0	0	0	0	0	0
	756	61	685	132	0	817	0	817	0	817
	127	119	101	145	246	0	246	0	246	0
Group18	60.44	28.5	72.53	34.29	0.55	0.11	0.55	0.11	0.55	0.11
	631	276	654	253	0	907	0	907	0	907
	72	110	50	132	181	1	181	1	181	1
Group19	80.72	18.38	66.27	23.35	0	0	0	0	0	0
	428	595	662	361	0	1023	0	1023	0	1023
	32	134	56	110	166	0	166	0	166	0
Group20	84.05	17.28	74.23	20.27	0	0	0	0	0	0
	435	656	615	476	0	1091	0	1091	0	1091
	26	137	42	121	163	0	163	0	163	0
Group21	72.55	17.1	72.55	18.85	0	0	0	0	0	0
	603	538	663	478	0	1141	0	1141	0	1141
	42	111	42	111	153	0	153	0	153	0
Group22	85.3	13.53	71.71	17.6	0	0	0	0	0	0
	1487	2448	2427	1508	2	3933	2	3933	2	3933
	66	383	127	322	449	0	449	0	449	0

Table 4.12a Experiment results with inductive SVM

	Op_c (F-measure)	Se	PPV
Group4	0.00001	87.01	98.82
Group5	0.00001	66.39	99.07
Group6	0.0001	60.68	99.76
Group7	0.0001	88.13	99.46
Group8	0.00001	88.37	99.03
Group9	0.00001	87.2	98.46
Group10	0.00001	84.88	96.99
Group11	0.00001	77.84	98.18
Group12	0.00001	76.56	97.22
Group13	0.00001	72.06	97.35
Group14	0.00001	67	96.88
Group15	0.0001	62.39	89.47
Group16	0.0001	53.11	71.43
Group17	0.00001	48.37	66.11

Group18	0.0001	72.53	34.29
Group19	0.0001	66.27	23.35
Group20	0.0001	74.23	20.27
Group21	0.0001	72.55	18.85
Group22	0.0001	71.71	17.6
overall		76.190475	70.617195

Table 4.12b Optimal parameters for each group of data

Transductive SVM:

Here the transductive SVM with rbf kernel is applied.

	g=0.000010		g=0.000100		g=1.000000		g=10.000000	
Group #	Se	PPV	Se	PPV	Se	PPV	Se	PPV
	TP	FN	TP	FN	TP	FN	TP	FN
	FP	TN	FP	TN	FP	TN	FP	TN
Group4	52.21	99.01	50.39	99.49	100	99	100	98.97
	2	2	3	1	0	4	0	4
	184	201	191	194	0	385	0	385
Group5	50.31	99.18	50.31	99.59	49.5	97.9	49.48	97.93
	3	2	4	1	0	5	0	5
	238	241	238	241	242	237	242	237
Group6	51.4	98.87	50.66	99.42	49	96.2	49.04	96.24
	9	4	11	2	0	13	0	13
	330	349	335	344	346	333	346	333
Group7	50	98.58	50.6	99.06	48.9	95.8	48.92	95.77
	12	6	14	4	0	18	0	18
	417	417	412	422	426	408	426	408
Group8	52.1	98.57	51.88	98.97	47.7	91	47.69	90.97
	38	7	40	5	1	44	1	44
	445	484	447	482	486	443	486	443
Group9	55.19	98.5	55.3	98.69	46	85.1	45.96	85.05
	69	8	70	7	0	77	0	77
	427	526	426	527	515	438	515	438
Group10	58.83	97.57	57.17	98.16	43.8	78	43.8	77.95
	112	15	116	11	0	127	0	127
	422	603	439	586	576	449	576	449
Group11	63.51	98.09	58.25	99.12	99.8	87.3	99.79	87.29
	129	12	136	5	0	141	0	141
	354	616	405	565	2	968	2	968
Group12	70.42	98.54	59.9	98.12	41	69.5	41.04	69.49
	164	10	163	11	1	173	1	173
	284	676	385	575	566	394	566	394
Group13	71.15	97.67	65.27	98.43	36	56.3	36.03	56.33
	201	13	206	8	0	214	0	214
	221	545	266	500	490	276	490	276
Group14	66.83	97.11	68.66	96.96	100	68.9	100	68.91

	260	12	259	13	0	272	0	272
	200	403	189	414	0	603	0	603
Group15	66.06	77.42	73.09	77.35	0	0	0	0
	306	63	299	70	21	348	21	348
	111	216	88	239	327	0	327	0
Group16	69.23	45.22	76.19	52.26	100	31.9	100	31.86
	355	229	394	190	0	584	0	584
	84	189	65	208	0	273	0	273
Group17	69.51	32.76	77.64	36.24	0	0	0	0
	466	351	481	336	285	532	285	532
	75	171	55	191	246	0	246	0
Group18	60.44	19.96	84.07	28.02	0.55	0.18	0.55	0.18
	466	441	514	393	363	544	363	544
	72	110	29	153	181	1	181	1
Group19	63.86	17.55	70.48	19.6	100	14	100	13.96
	525	498	543	480	0	1023	0	1023
	60	106	49	117	0	166	0	166
Group20	58.28	15.15	63.19	16.43	0	0	0	0
	559	532	567	524	464	627	464	627
	68	95	60	103	163	0	163	0
Group21	67.97	16.12	69.28	16.33	0	0	0	0
	600	541	598	543	494	647	494	647
	49	104	47	106	153	0	153	0
Group22	64.37	13.08	73.27	15.02	0	0	0	0
	2014	1921	2073	1862	1743	2192	1743	2192
	160	289	120	329	449	0	449	0

Table 4.13a Experiment results with transductive SVM

	Op_c (F-measure)	Se	PPV
Group4	1	100	98.97
Group5	0.0001	50.31	99.59
Group6	0.00001	51.4	98.87
Group7	0.0001	50.6	99.06
Group8	0.00001	52.1	98.57
Group9	0.0001	55.3	98.69
Group10	0.00001	58.83	97.57
Group11	1	99.79	87.29
Group12	0.00001	70.42	98.54
Group13	0.00001	71.15	97.67
Group14	10	100	68.91
Group15	0.0001	73.09	77.35
Group16	0.0001	76.19	52.26
Group17	0.0001	77.64	36.24
Group18	0.0001	84.07	28.02
Group19	0.0001	70.48	19.6
Group20	0.0001	63.19	16.43
Group21	0.0001	69.28	16.33

Group22	0.0001	73.27	15.02
overall		68.763046	59.785568

Table 4.13b Optimal parameters for each group of data

As shown in Table 4.12 and Table 4.13, the overall SE and PPV is 76.19% and 70.62 with inductive SVM, and the overall SE and PPV is 68.76% and 59.79% with transductive SVM. So a conclusion can be drawn that the transductive SVM is not superior to traditional inductive SVM in our experiments. So the selection of SVM should be a case by case issue.

4.4 Results

As a classifier, SVM first embeds its data into a suitable space and then learns a decision function to separate the data with a hyperplane that has the maximum margin from a small number of critical boundary samples from each class. A support vector machine's decision function for a test sample is a linear combination of kernels computed at the training data points [Wong, 2004].

I applied the prediction system to the human chromosome 22 (Built 35), and gave the final report for each position on the long sequence. The final report contains the scores of the decision function at each position extracted by the sliding window along the chromosome.

The speed of the system is such that I can process 240000 sequences per hour, with each sequence having the length of 1024 nucleotide. These sequences do not include those that contain 'N' or 'n'. Unlike sequences containing only 'a', 'c', 't', 'g', sequences containing 'N' and 'n' can not be transformed into digital signals by EIIP means. So when I slid the window, I only grasped the sequences which contain only 'a', 'c', 't', 'g'. I moved the

window by step of 10bp, in order to maintain a properly high resolution in recording the possibility score of the positions on the chromosome sequence.

The scores of the positions were recorded into two files, one of which is for promoters and another for non-promoters. The final statistical analysis such as distribution plot of these scores was also generated. Based on this, the range of the scores of the positive candidates along the chromosome could be observed. The threshold to classify the positive and negative data is zero, since the all of the positive data have a score above zero, while the entire negative have a score below zero.

I also analysed the prediction results under different thresholds, which filter the predictions. Only predictions with scores that are above the threshold were retained. Another process was to group the predictions into different clusters according to different cluster distances and replaced the each cluster of data with their means in each cluster. Thus, the predictions could become more compact.

Then I fixed the threshold, and tuned the distance of the predicted positions obtained under this threshold. Later, I fixed the distance and tune the threshold under the distance. Then I compared the newly obtained predictions with the reference, and calculated TP, FP, and plot PPV and SE. Based on the plot of PPV and SE, I could select the optimal thresholds and distances. There were 28 different clustering distances, which were 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, 5000, 5500, 6000, 6500, 7000, 7500, 8000, 8500, 9000, 9500, 10000, and there were 93 different thresholds, which were from 0,0.5 to 46, evenly distributed with 0.5 as interval.

Since the reference file of the chromosome 22 contains five categories of genes: coding genes, non-coding genes, pseudo genes, partial genes, IGLV/J, I designed the assessment

under six categories (the above five categories of different genes plus one category of all these genes). I drew the plot at different thresholds of scores and different cluster distances; each point had two values of PPV and SE respectively. I obtained different predictions under different thresholds of the predicted scores. The shapes of the curves in different categories were different.

As described in [Collins *et al.,* 2003], the gene categories are defined as:

A complete protein Coding gene had exact sequence identity to human cDNAs or ESTs across its entire length, and a predicted ORF of at least 300 bases.

A Partial gene had sequence similarity to cDNA, EST or peptide sequence but did not comply with the complete gene criteria.

Non-coding RNA genes included small RNAs, and published(or complete) genes which did not contain an ORF of at least 300 bases.

A pseudogene had similarity to a known gene or protein but had evidence of disrupted function。

IGL V/J indicated IGLV and J gene segments, which is the immunoglobulin joining and variable regions, including pseudogenes.

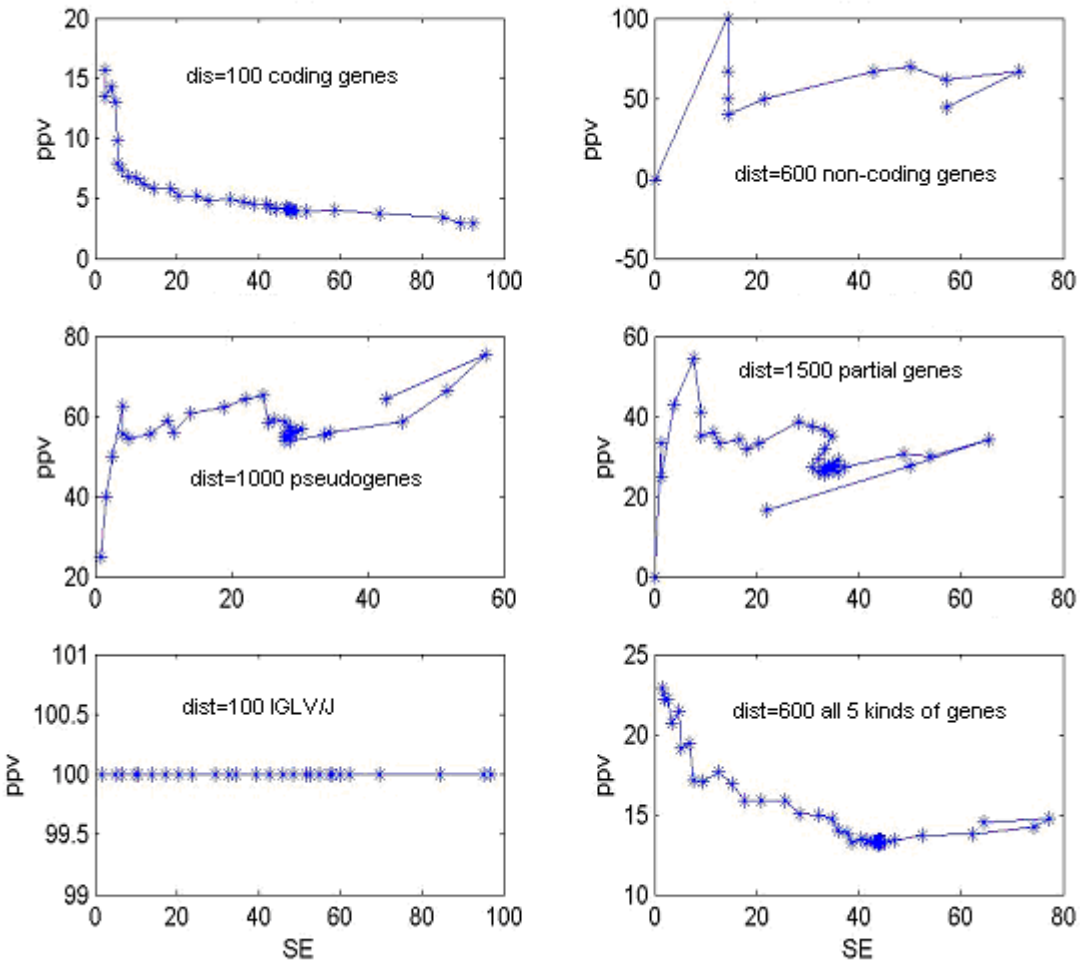


Figure 4.4 Results on the data of Group1-22

Figure 4.4 is drawn with the 6 sets of thresholds and distances of each category on the data of all the 22 Groups. The optimal points are those with the most appropriate "distance" and "threshold" to generate the best performance in SE and PPV. For each of the six categories of genes, there is one optimal point corresponding in each plot.

	Optimal threshold	Optimal distance	SE	PPV
Coding Genes	0.5	100	89.55	2.97
Non-Coding Genes	0.5	600	71.43	66.67
Pseudo Genes	0.5	1000	57.38	75.27
Partial Genes	1	1500	65.38	34.23
IGL V/J	0	100	96.72	100
All of the 5 kinds of Genes	0.5	600	77.28	14.76

Table 4.14 Optimal points on the curves in the six categories of Group 1-22

They are the point with SE=89.55% and PPV=2.97% in category of coding genes with cluster distance 100; SE=71.43% and PPV=66.67% in category of Non-coding genes under cluster distance of 600; SE =57.38% and PPV=75.27% in the category of pseudo genes with cluster distance of 1000; SE=65.38% and PPV=34.23% in the category of partial genes under cluster distance of 1500; SE=96.72% and PPV=100% in the category of IGLV/J genes under cluster distance of 100; SE=77.28% and PPV=14.76% in the category of all the five kinds of genes under cluster distance of 600. The performance obtained for the category of coding-gene is least satisfactory. This means that the coding-gene is the most difficult category of data to predict by our system.

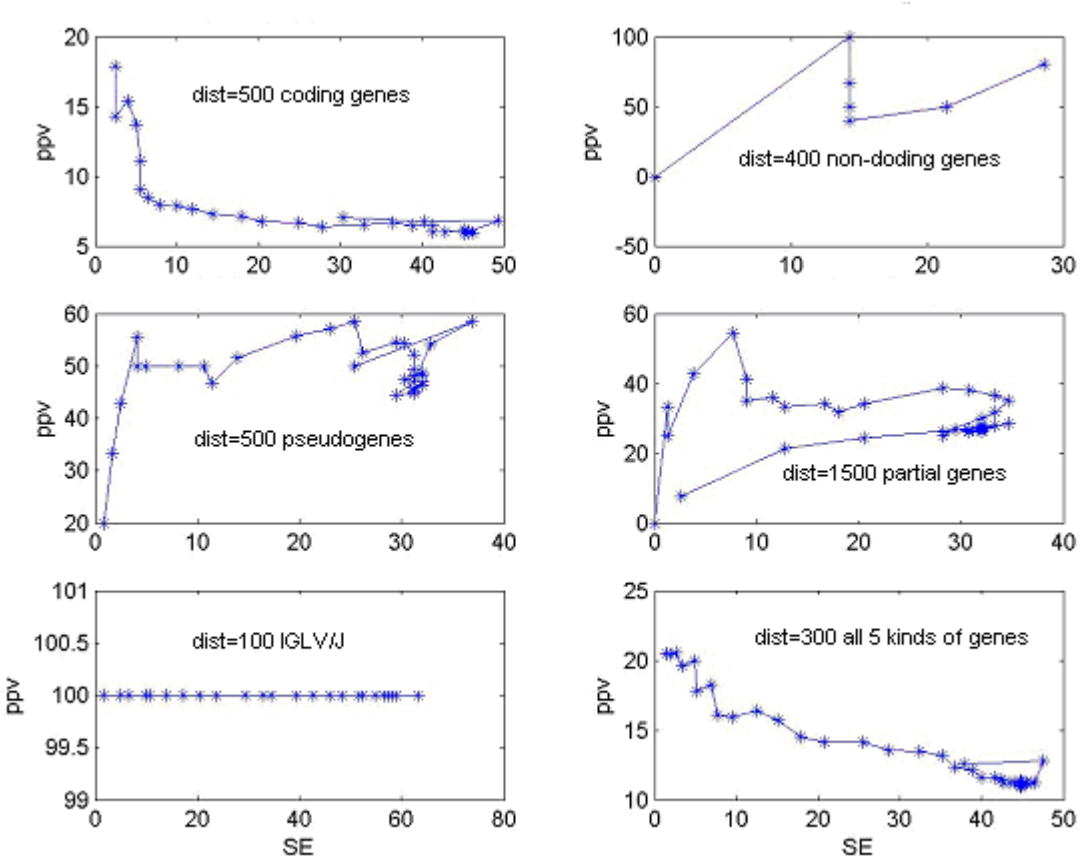


Figure 4.5 Results on the data of Group1-16

The curves obtained with the predictions of Group 1 to 16 are displayed in Figure 4.5. Compared with the curves of Group 1 to 22, the shapes of the curves of Group 1 to 16 are similar, but the value of SE is bigger for each category of genes. This may due to the fact that data that have higher GC-content are easier to be predicted.

	Optimal threshold	Optimal distance	SE	PPV
Coding Genes	1	500	49.25	6.83
Non-Coding Genes	0	400	28.57	80
Pseudo Genes	0.5	500	36.89	58.44
Partial Genes	38	1500	34.62	35.06
IGL V/J	0.5	100	63.11	100
All of the 5 kinds of Genes	0.5	300	47.49	12.78

Table 4.15 Optimal points on the curves in the six categories in Group 1-16

In this figure, the optimal points are the one with SE=49.25% and PPV=6.83% in category of coding genes with cluster distance 500; SE=28.57% and PPV=80% in category of Non-coding genes under cluster distance of 400; SE =36.89% and PPV=58.44% in the category of pseudo genes with cluster distance of 500; SE=34.62% and PPV=35.06% in the category of partial genes under cluster distance of 1500; SE=63.11% and PPV=100% in the category of IGLV/J genes under cluster distance of 100; SE=47.49% and PPV=12.78% in the category of all the five kinds of genes under cluster distance of 300.

Chapter 5

Conclusions

This thesis examines the capability of using some possible Digital Signal Processing (DSP) techniques for promoter prediction. Systematic simulation studies for features extracted under different domain transforms were carried out. Based on the experiments, we observed that DSP techniques can provide complementary information that can be combined with biological features of promoters and non-promoters to enhance promoter prediction**.**

In Chapter 2, we define the signal model for the promoter prediction problem. Specific techniques based on the three domain transforms: DFT, DCT and DWT are studied for possible applications in prediction systems. Using simulations, we compared the promoters and non-promoters based on statistical characteristics including the signal mean, correlation coefficient of specific sequence with the mean signal, and the distribution of the correlation coefficient. From the experiments, it can be concluded that CC is not a good feature to effectively distinguish between promoters and non-promoters.

In Chapter 3, we study the use of the DFT, DCT, DWT transform coefficients of the original signal as features. Based on experiments, we are able to select an optimal combination of features and define a classifier model. The performance of different

combinations is systematically evaluated with commonly used measures such as SE and PPV value. Based on the results, we observed that the ability to recognize promoters degrades with the reduction of GC-content. It is also found out that there is no significant difference in the prediction performance when any transform is used. Also, the best performance is achieved by combining all the three transforms. In all, the application of domain transforms in predicting promoters is promising and thus should be combined with other features obtained from the physical or statistical properties of promoter regions for better prediction.

In Chapter 4**,** we present the implementation of the promoter prediction system. The system includes signal pre-processing, feature extraction, system optimization, and promoter recognition with performance analysis. By system optimization, the model with optimal parameters is determined for different groups of sequence with different GC-content. The final prediction system is applied to human chromosome 22 (NCBI built 35). Performance evaluation is done with the prediction results under different thresholds which filter the predicted position and with different distances which cluster the results. Comparison is made with the results for the respective six different categories of genes.

Bibliography

Audic, S., and Claverie, J. M., 1998. Proc. Nat. Acad. Sci. USA, 95 (17), 10026-10031.

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,Vol.18(1),pp.198-199.

Bajic, V. B., Seah, S. H., Chong, A., Krishnan, S. P. T., Koh, J. L. Y., and Brusic,V. , 2003. Computer model for recognition of functional transcription start sites in polymerase II promoters of vertebrates, Journal of Molecular Graphics & Modeling, 21 (5): 323-332.

Bajic. V. B., and Seah, S. H., 2003a. Dragon Gene Start Finder identifies approximate locations of the 5' ends of genes, Nucleic Acids Research, 31(13):3560-3563, 2003.

Bajic, V. B., and Seah, S. H., 2003b. Dragon Gene Start Finder: An Advanced System for Finding Approximate Locations of the Start of Gene Transcriptional Units, Genome Research, 13:1923-1929, 2003.

Bajic, V. B., Tan, S. L., Suzuki, Y., and Sugano, S., 2004. Promoter prediction analysis on the whole human genome, Nature Biotechnology, 22(11):1467-73.

Cartharius, K., 2005. MatInspector: Analysing Promoters for Transcription Factor Binding Sites in Bioinformatics - The DNA Tools, in press, DNA Press.

Chen, Y., Wang, G., and Dong, S., 2003. Learning with progressive transductive support vector machine, Source Pattern Recognition Letters archive Pages: 1845 - 1855,Volume 24 , Issue 12 , (August 2003) ,Elsevier Science Inc. New York, NY, USA.

Chong, A., Zhang, G., and Bajic, V. B., "FIE2: A program for the extraction of genomic DNA sequences around the start and translation initiation site of human genes," Nucleic Acids Res., vol. 31, no. 13, pp. 3546-53, 1 Jul 2003.

Collins, J. E., Goward, M. E., Cole, C. G., Smink, L. J., Huckle, E. J., Knowles, S., Bye, J. M., Beare, D. M., Dunham, I., 2003. Reevaluating human gene annotation: a second-generation analysis of chromosome 22. Genome Res. 2003 Jan;13(1):27-36.

Veljković V, Slavić I. Simple general-model pseudopotential. Phys. Rev. Let., 29, 105 (1972).

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

Down, T. A., and Hubbard, T. J. P., 2002. Computational Detection and Location of Transcription Start Sites in Mammalian Genomic DNA Genome Res. 12:458-461.

Duda, R. O., Hart, P. E., and Stork, D. G., 2001. Pattern Classification (2nd ed.),Wiley Interscience.

Fickett, J.W., and Hatzigeorgiou, A.G., 1997. Eukaryotic promoter recognition. Genome Research. Sep;7(9):861-78, Review.

Fisher, M. J., Fieldsend, J. E., and Everson, R. M., 2004.Multi-Objective Optimisation for Information Access Tasks, ACM Transactions on Information Systems.

Frech, K., Herrmann, G., and Werner, T., 1993. Computer-assisted prediction, classification, and delimitation of protein binding sites in nucleic acids.Nucleic Acids Res. 21, 1655-1664..

Frech, K., Dietze, P., and Werner, T., 1997. ConsInspector 3.0: new library and enhanced functionality.Comp. Appl. Biosci. 13, 109-110.

Gardiner-Garden, M., and Frommer, M., 1987. CpG islands in vertebrate genomes. J. Mol. Biol. 196, 261-282.

Ghosh, 1993. Status of the transcription factors database (TFD). Nucleic Acids Res 21, 3117-8 (1993)

Gonzalez, R. C., and Woods, R. E., 2004. Digital Image Processing, 2nd Edition, Prentice Hall.

Gunn, S. R., 1998. Technical Report "Support Vector Machines for Classification and Regression".

Halees, A. S., and Weng, Z., "PromoSer: improvements to the algorithm, visualization and accessibility," Nucleic Acids Res., vol. 32, Web Server issue, pp. 191-4, 1 Jul 2004.

Hutchinson, G. B., 1996. The Prediction of Vertebrate Promoter Regions Using Differential Hexamer Frequency Analysis. Comp. Appl. Biosci. (1996) In Press.

Ian, H. W. and Eibe, F., 2005. Data Mining: Practical machine learning tools and techniques, 2nd Edition, Morgan Kaufmann, San Francisco.

Ioshikhes, I., and Zhang, M. Q., 2000. Large-scale human promoter mapping using CpG islands. Nature Genetics 26, 61-63 (2000).

Jegga, A. G., Sherwood, S. P.,1 Carman, J. W., Pinski, A. T., Phillips, J. L., Pestian, J. P., and Aronow, B. J. ,2002. Detection and Visualization of Compositionally Similar cis-Regulatory Element Clusters in Orthologous and Coordinately Controlled Genes.Genome Research 12: 1408-1417, September 2002.

Klingenhoff, A., Frech, K., Quandt, K., and Werner, T., 1999. Functional promoter modules can be detected by formal models independent of overall nucleotide sequence similarity.Bioinformatics 15, 180-186.

Knudsen, S., 1999. 'Promoter2.0: for the recognition of PolII promoter sequences', Bioinfomatics, Vol 15 no.5 1999 p.356-361.

Lavorgna, G., Boncinelli, E., Wagner, A., and Werner, T., 1998. Detection of potential target genes in silico?Trends Genet. 14, 375-376.

Latchman, D. S.,1998. Gene Regulation — A Eukaryotic Perspective. Stanley Thornes Ltd, 3rd edition.

Levitsky, V. G., Podkolodnaya, O. A., Kolchanov, N. A., and Podkolodny, N. L., 2001. 'Nucleosome formation potential of eukaryotic DNA: calculation and promoters analysis'. Bioinformatics, Vol.17(11), pp. 998-1010.

Levitsky, V. G., and Katokhin, A. V., 2003. "Recognition of eukaryotic promoters using a genetic algorithm based on iterative discriminant analysis", In Silico Biology, 3:0008.

Liu, R., and States, D. J., 2002. Consensus promoter identification in the human genome utilizing expressed gene markers and gene modeling, Genome Res. 12 (2002) 462–469.

Loots, G., and Ovcharenko, I., 2004. rVista 2.0: Evolutionary analysis of transcription factor binding sites. Nucleic Acids Research, 32(Web Server Issue), W217-W221.

Mache, N., Reczko, M., and Hartzigeorgiou, A., 1996. Multistate Time-Delay Neural Networks for POL-II Promoter Sequences, Proc. 10th Conf. Intelligent Systems for Molecular Biology , (ISMB 96), St. Louis.

Maglott, D. R., Katz, K. S., Sicotte, H., and Pruitt, K. D., 2000. NCBI's LocusLink and RefSeq. Nucleic Acids Res., 28, 126–128.

Ohler, U., Stemmer, G., Harbeck, S., and Niemann, H., 2000. Stochastic segment models of eukaryotic promoter regions. Proc. Pac. Symp. Biocomput. 5, 380–391 (2000).

Ohler, U., Liao, G. C., Niemann, H., and Rubin, G. M., 2002. Computational analysis of core promoters in the Drosophila genome. Genome Biol. 3(12), RESEARCH0087. Epub 2002 Dec 20 (2002).

Pedersen, A. G., Baldi, P., Chauvin, Y., and Brunak, S., 1999. The biology of eukaryotic promoter prediction--a review. Comput. Chem. 23(3-4):191-207 .

Pesole, G., Liuni, S., and D'Souza, M., 2000. PatSearch: a pattern matcher software that finds functional elements in nucleotide and protein sequences and assesses their statistical significance.Bioinformatics. 2000 May;16(5):439-50.

Ponger, L., and Mouchiroud, D., 2002. CpGProD: Identifying CpG islands associated with transcription start sites in large genomic mammalian sequences. Bioinformatics 18, 631–633 (2002).

Prestridge, D.S., 1991. SIGNAL SCAN: A computer program that scans DNA sequences for eukaryotic transcriptional elements. CABIOS 7, 203-206.

Prestridge, D. S., 1995. Predicting Pol II promoter sequences using transcription factor binding sites. J. Mol. Biol. 249:923-932.

Pruitt, K. D., Katz, K. S., Sicotte, H., and Maglott, D. R., 2000. Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. Trends Genet., 16, 44–47.

Pruitt, K. D., and Maglott, D. R., 2000. RefSeq and LocusLink: NCBI gene-centered resources. Nucleic Acids Res., 29, 137–140.

Quandt, K., Frech, K., Karas, H., Wingender, E., and Werner, T., 1995. MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data.Nucleic Acids Res. 23, 4878-4884.

Quandt, K., Frech, K., Herrmann, G., and Werner, T., 1995. A consensus match scoring system that is correlated with biological functionality.in Bioinformatics: From Nucleic Acids and Proteins to Cell Metabolism (Eds. D. Schomburg, U. Lessel), 47-57, GBF Monographs Volume 18, VCH Publishers, Inc., NewYork, NY.

Reese, M.G., and Eeckman, F.H., 1995. "Novel Neural Network Algorithms for Improved Eukaryotic Promoter Site Recognition",Accepted talk for The seventh international Genome sequencing and analysis conference, Hyatt Regency, Hilton Head Island, South Carolina September 16-20.

Reese, M. G., 2000. ``Computational prediction of gene structure and regulation in the genome of Drosophila melanogaster", PhD Thesis (PDF), UC Berkeley/University of Hohenheim.

Reese, M. G., 2001. ``Application of a time-delay neural network to promoter annotation in the Drosophila melanogaster genome", Comput Chem 26(1),51-6.

Scherf, M., Klingenhoff, A., and Werner, T., 2000. 'Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach.' J Mol Biol., 297(3):599-606.

Scherf, M., Klingenhoff, A., Frech, K., Quandt, K., Schneider, R., Grote, K., Frisch, M., Gailus-Durner, V., Seidel, A., Brack-Werner, R., and Werner, T., 2001. First pass annotation of promoters on human chromosome 22, Genome Research, Vol.11(3), pp.333-340.

Van Rijsbergen, C. J., 1979. *Information Retrieval*, London: Butterworth.

Werner, T., 1999. Models for prediction and recognition of eukaryotic promoters. Mammalian Genome 10:168-175.

Werner T., 2003. The state of the art of mammalian promoter recognition Briefings in Bioinformatics,. 2003 Mar;4(1):22-30.

Werner, T., Fessele, S., Maier, H., and Nelson, P. J., 2003. Computer modeling of promoter organization as a tool to study transcriptional coregulation. FASEB Journal 17, 1228-1237.

Wingender, E., 1994. Recognition of regulatory regions in genomic sequences. J Biotechnol. 1994 Jun. 30;35(2-3):273-80.

Wingender, E., Knüppel, R., Dietze, P., Karas, H., Frech, K., Quandt, K., and Werner,

T., 1995. The TRANSFAC database and ConsInd program as tools for describing and understanding regulatory functions of the genome. SAMS, 18-19, 823-826.

Wong, L.(editor), 2004. The Practical Bioinformatician, World Scientific, New Jersey.

Zhang, M. Q., 1998. Identification of Human Gene Core Promoters in Silico. Genome Res 8(1):319-326.

Zhang, X., Bajic, V. B., and Kassim, A., 2004. Digital signal processing for potential promoter prediction, IEEE Biocas.

Appendix A

List of Publications

Zhang, X., Bajic, V. B., and Kassim, A., 2004. Digital signal processing for potential promoter prediction, IEEE Biocas.

Appendix B

1. The CC distribution plot and reconstructed mean signal respectively at level 1 ,2 and 7

Level --1:

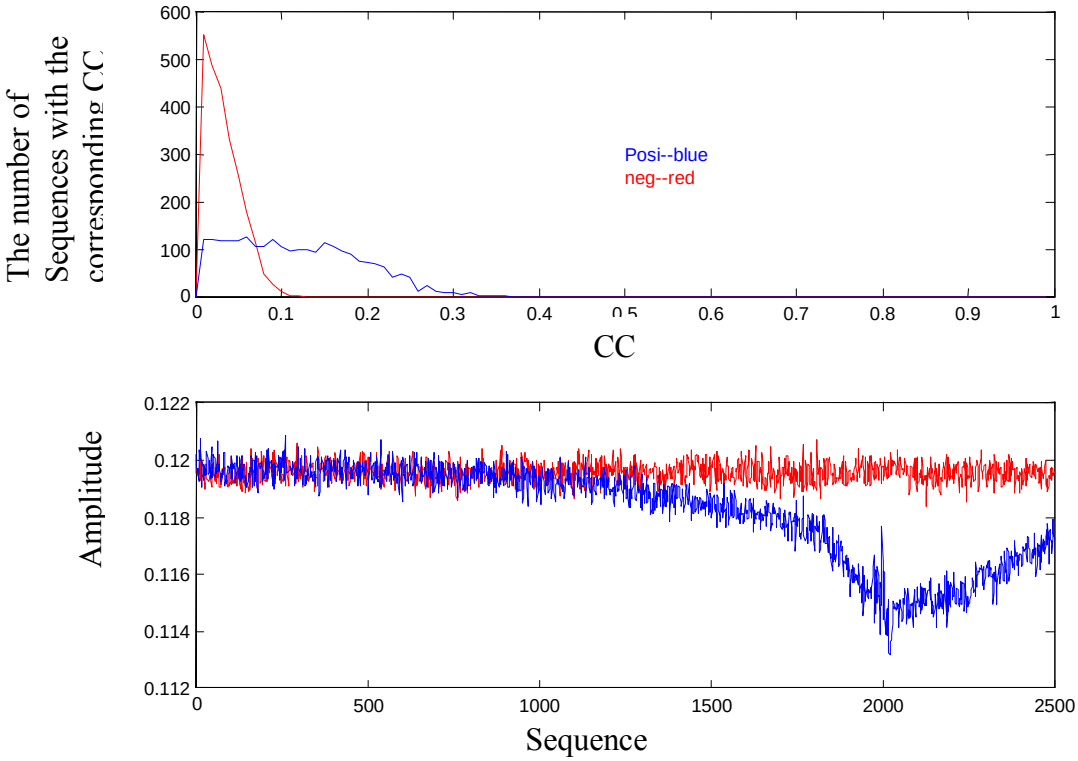


Figure B.1 The CC distribution plot and reconstructed mean signal at level 1

Level—2

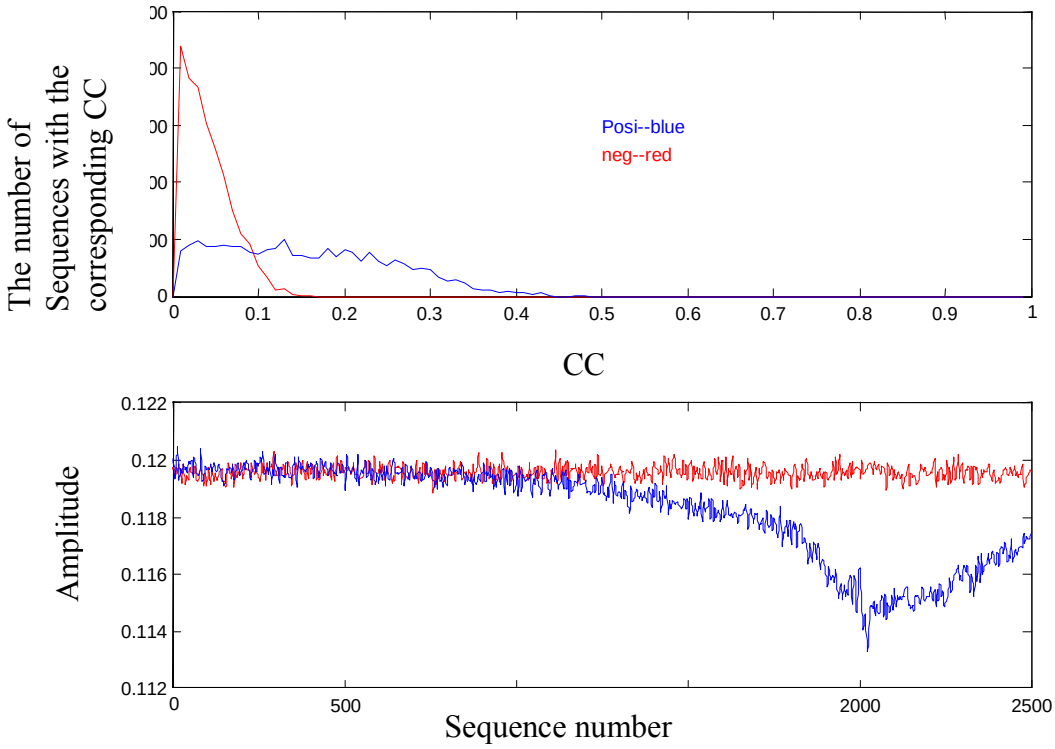


Figure B.2 The CC distribution plot and reconstructed mean signal at level 2

Level—7

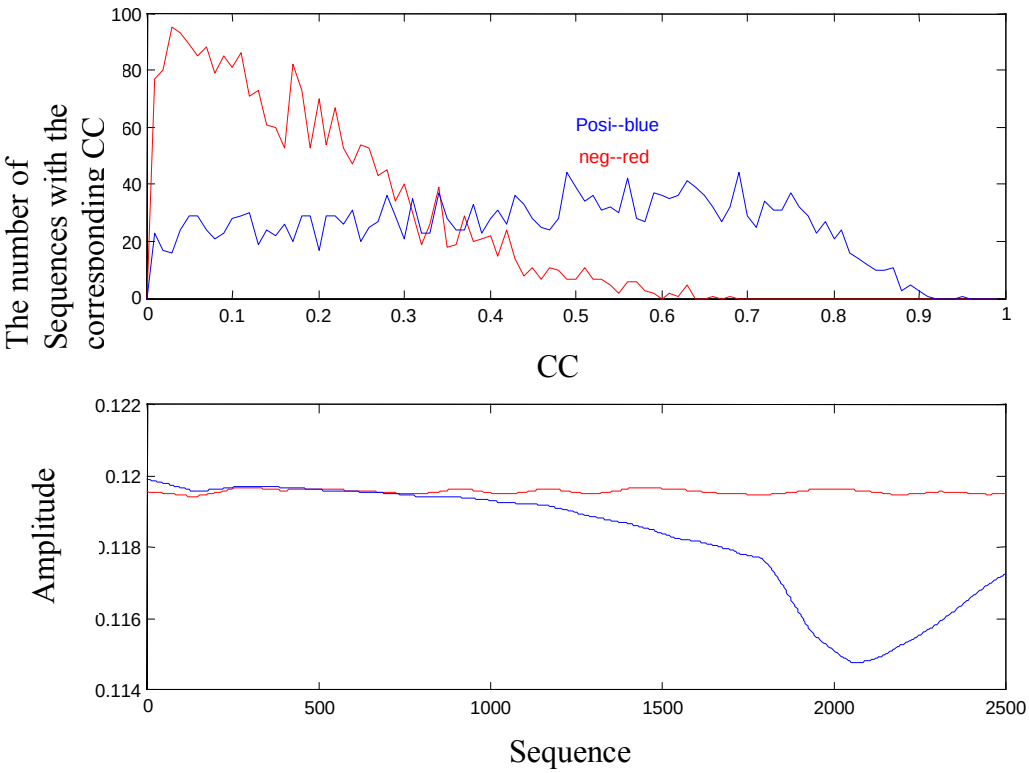


Figure B.3 The CC distribution plot and reconstructed mean signal at level 7

Figure B.1, B.2 and B.3 are respectively the plots obtained when the original signal is decomposed respectively at level 1, 2, and 7. After we decompose the signal in different levels by DWT, signal reconstruction can be made with the specific 'approximate' or 'detailed' part at any level. By filtering away the 'detailed' part, the signal can be de-noised effectively. The mean sequence of the reconstructed positive signal is regarded as a reference sequence, and the CC value of the individual reconstructed signal with this reference sequence can be found. By this CC value, the relativity of an individual (positive or negative) signal with the reference sequence can be found out, thus the difference of positive and negative signal indicated by CC is expected.

In Figure B.1, B.2 and B.3, the upper plot is the CC distribution plot, the value of x varies from 0 to1, the y axis is the number of sequences with the same specific CC value shown on the x axis. To find out the most appropriate level at which the biggest difference occurs to separate the positive and negative sequence effectively, we scheme to have the least overlap of the positive and negative data when we use a threshold to make classification.

The lower plot in Figure B.1, B.2 and B.3 is the reconstructed mean signal of the positive and negative data, using only the approximate part of the original signal in level 1, 2, and 7 at which the signal is decomposed respectively. The x axis is the length of the sequence, from 0 to 2500. The y axis is the value of the mean signal's amplitude at each position of the sequence. Based on the figure, we can observe that when the reconstruction is done with the approximate part of the signal, the shape of the signal's curves will be smooth and the curves reconstructed in level 7 will be smoother than those reconstructed in level 1and 2.

2. The CC distribution

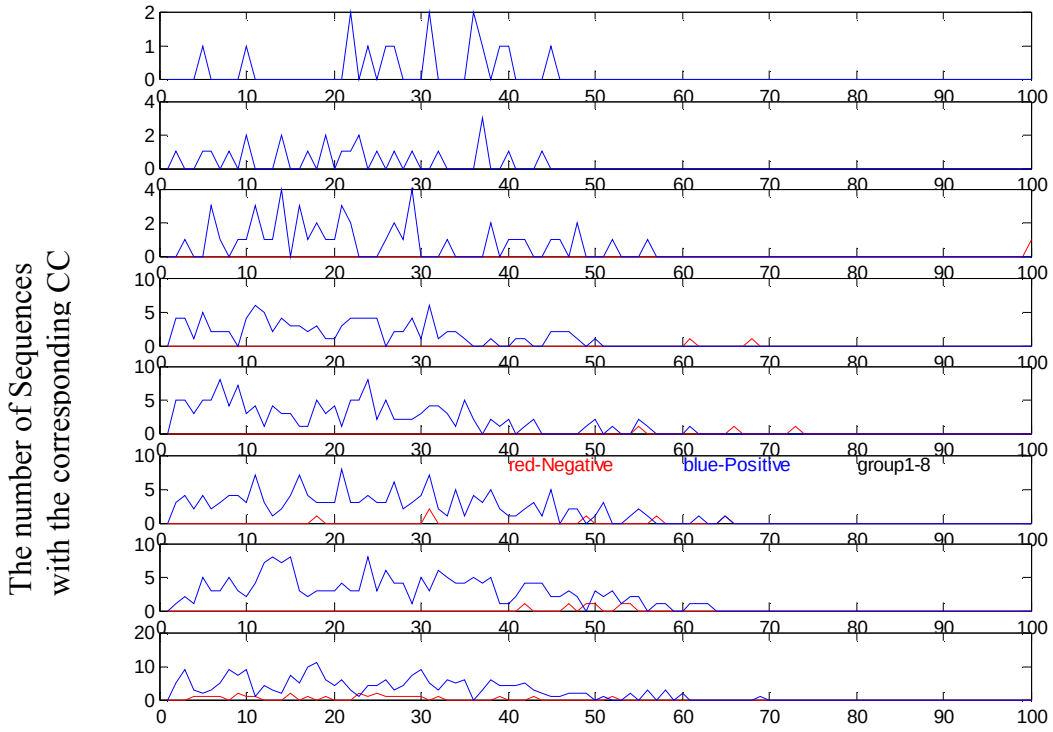


Figure B.4 The CC distribution plot in group 1-8

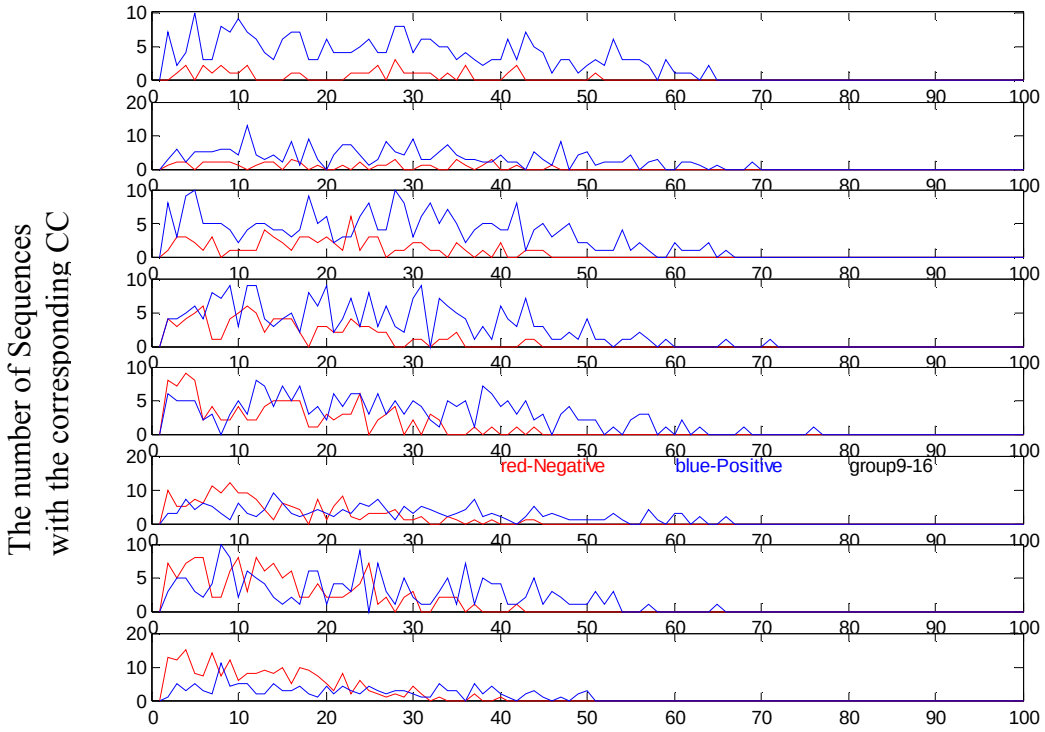


Figure B.5 The CC distribution plot in group 9-16

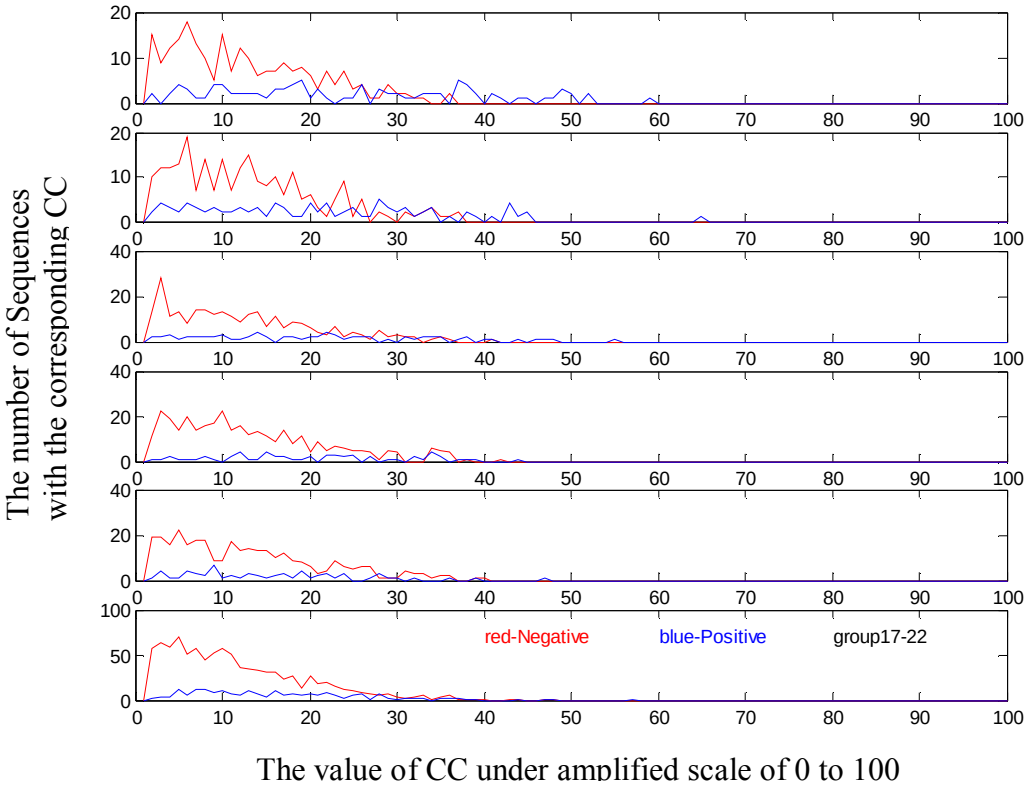


Figure B.6 The CC distribution plot in group 17-22

Figure B.4, B.5, and B.6 is the distribution plot of CC in group 1-7, 8-16, 17-22 respectively. The CC is calculated with the input signal and the mean of all positive data in each of the 22 groups; the y axis is the number of sequences under the same CC value shown along x axis. The curves in red are the result obtained with all the negative sequences, while those in blue are with all the positive sequences. The x axis is previously from 0-1 (CC's range) and is scaled to 0-100 for the purpose of observation more clearly.

From Figure B.4, B.5, and B.6, we can find out that the difference of positive and negative data by plotting the distribution of the feature CC. The difficulty to separate the positive and negative data differs from group to group. In short, the groups that are GC rich are easier to separate than groups that are GC poor.

3. Classification with the feature of CC under different thresholds

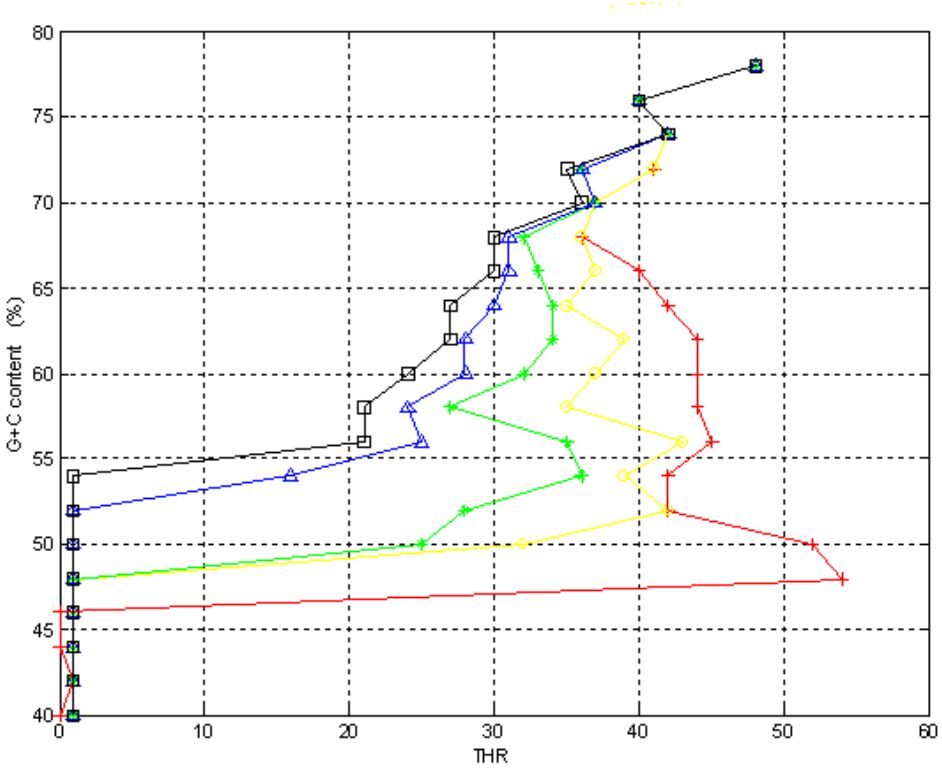


Figure B.7 The threshold versus GC content

In Figure B.7 the 5 curves are plotted under 5 different TP/FP ratios, which are 14%, 10%, 7%, 4%, and 2% from left to right. TP means true positive prediction, which is the prediction that is correctly made. FP means false positive prediction. Each of the 19 points on the curve is drawn with one group of data from group 4 to 22, into which the input data was firstly divided using the criterion of GC content. Every point is plotted with two coordinates, one being the threshold of its group to give best separation of positive and negative data, the other being the GC content of its group. "THR" means "threshold", which is the label of x axis.

From Figure B.7, it is clear that the FP rate decreases while as the threshold is increased. The experimental results of TP, FP, FN, TN, Se, Sp, and PPV are recorded in Table B.1 to B.5 shown below. Each table is under different condition of FP/TP ratio, which is 2%, 4%, 7%, 10%, and 14%, respectively. The "reviewed" data are used in the experiment and since there are no negative data in Group 1 and 2, only results from group 3 to 22 are summarized.

Table B.1 to B.5 is given below to record the result of the experiment group by group under the above five different FP/TP ratios.

	TP	FP	FN	TN	Se(%)	Sp(%)	PPV(%)
Group1							
Group2							
Group3	49	1	0	0	100.0000	0	98
Group4	105	2	0	0	100.0000	0	98.1308
Group5	144	3	0	0	100.0000	0	97.9592
Group6	162	6	0	0	100.0000	0	96.4286
Group7	7	0	186	6	3.6269	100	100
Group8	11	0	217	26	4.8246	100	100
Group9	56	1	205	32	21.4559	96.9697	98.2456
Group10	50	1	181	44	21.6450	97.7778	98.0392
Group11	35	0	224	71	13.5135	100	100
Group12	22	0	213	92	9.3617	100	100
Group13	31	0	185	113	14.3519	100	100
Group14	31	0	160	158	16.2304	100	100
Group15	25	0	148	132	14.4509	100	100
Group16	13	0	124	211	9.4891	100	100
Group17	26	0	71	232	26.8041	100	100
Group18	12	0	83	236	12.6316	100	100
Group19	5	0	67	261	6.9444	100	100
Group20	1	0	57	335	1.7241	100	100
Group21	1	0	62	320	1.5873	100	100
Group22	1	0	210	960	0.4739	100	100
Overall	787	14	2393	3229	24.7484	99.5683	98.2522

Table B.1 Experiment result with FP/TP=2%

	TP	FP	FN	TN	Se(%)	Sp(%)	PPV(%)
Group1							
Group2							
Group3	49	1	0	0	100	0	98
Group4	105	2	0	0	100	0	98.1308
Group5	144	3	0	0	100	0	97.9592
Group6	162	6	0	0	100	0	96.4286
Group7	193	6	0	0	100	0	96.9849
Group8	76	4	152	22	33.3333	84.6154	95
Group9	76	4	185	29	29.1188	87.8788	95
Group10	58	2	173	43	25.1082	95.5556	96.6667
Group11	57	3	202	68	22.0077	95.7746	95
Group12	54	2	181	90	22.9787	97.8261	96.4286
Group13	78	4	138	109	36.1111	96.4602	95.122
Group14	41	2	150	156	21.466	98.7342	95.3488
Group15	48	2	125	130	27.7457	98.4848	96
Group16	21	1	116	210	15.3285	99.5261	95.4545
Group17	26	0	71	232	26.8041	100	100
Group18	12	0	83	236	12.6316	100	100
Group19	5	0	67	261	6.9444	100	100
Group20	1	0	57	335	1.7241	100	100
Group21	1	0	62	320	1.5873	100	100
Group22	1	0	210	960	0.4739	100	100
Overall	1208	42	1972	3201	37.9874	98.7049	96.6400

Table B.2 Experiment result with FP/TP=4%

	TP	FP	FN	TN	Se(%)	Sp(%)	PPV(%)
Group1							
Group2							
Group3	49	1	0	0	100	0	98
Group4	105	2	0	0	100	0	98.1308
Group5	144	3	0	0	100	0	97.9592
Group6	162	6	0	0	100	0	96.4286
Group7	193	6	0	0	100	0	96.9849
Group8	201	22	27	4	88.1579	15.3846	90.1345
Group9	217	24	44	9	83.1418	27.2727	90.0415
Group10	68	7	163	38	29.4372	84.4444	90.6667
Group11	102	11	157	60	39.3822	84.507	90.2655
Group12	99	10	136	82	42.1277	89.1304	90.8257
Group13	85	9	131	104	39.3519	92.0354	90.4255
Group14	69	7	122	151	36.1257	95.5696	90.7895
Group15	54	6	119	126	31.2139	95.4545	90
Group16	38	4	99	207	27.7372	98.1043	90.4762

Group17	32	3	65	229	32.9897	98.7069	91.4286
Group18	12	0	83	236	12.6316	100	100
Group19	10	1	62	260	13.8889	99.6169	90.9091
Group20	1	0	57	335	1.7241	100	100
Group21	1	0	62	320	1.5873	100	100
Group22	1	0	210	960	0.4739	100	100
Overall	1643	122	1537	3121	51.6667	96.2381	93.0878

Table B.3 Experiment result with FP/TP=7%

	TP	FP	FN	TN	Se(%)	Sp(%)	PPV(%)
Group1							
Group2							
Group3	49	1	0	0	100	0	98
Group4	105	2	0	0	100	0	98.1308
Group5	144	3	0	0	100	0	97.9592
Group6	162	6	0	0	100	0	96.4286
Group7	193	6	0	0	100	0	96.9849
Group8	228	26	0	0	100	0	89.7638
Group9	261	33	0	0	100	0	88.7755
Group10	205	36	26	9	88.7446	20	85.0622
Group11	151	24	108	47	58.3012	66.1972	86.2857
Group12	113	18	122	74	48.0851	80.4348	86.2595
Group13	98	15	118	98	45.3704	86.7257	86.7257
Group14	82	11	109	147	42.9319	93.038	88.172
Group15	59	8	114	124	34.104	93.9394	88.0597
Group16	39	6	98	205	28.4672	97.1564	86.6667
Group17	34	6	63	226	35.0515	97.4138	85
Group18	12	2	83	234	12.6316	99.1525	85.7143
Group19	10	1	62	260	13.8889	99.6169	90.9091
Group20	1	0	57	335	1.7241	100	100
Group21	1	0	62	320	1.5873	100	100
Group22	1	0	210	960	0.4739	100	100
Overall	1948	204	1232	3039	61.2579	93.7095	90.5204

Table B.4 Experiment result with FP/TP=10%

	TP	FP	FN	TN	Se(%)	Sp(%)	PPV(%)
Group1							
Group2							
Group3	49	1	0	0	100	0	98
Group4	105	2	0	0	100	0	98.1308
Group5	144	3	0	0	100	0	97.9592
Group6	162	6	0	0	100	0	96.4286
Group7	193	6	0	0	100	0	96.9849

Group8	228	26	0	0	100	0	89.7638
Group9	261	33	0	0	100	0	88.7755
Group10	231	45	0	0	100	0	83.6957
Group11	190	47	69	24	73.3591	33.8028	80.1688
Group12	156	38	79	54	66.383	58.6957	80.4124
Group13	122	29	94	84	56.4815	74.3363	80.7947
Group14	108	27	83	131	56.5445	82.9114	80
Group15	77	16	96	116	44.5087	87.8788	82.7957
Group16	44	11	93	200	32.1168	94.7867	80
Group17	36	8	61	224	37.1134	96.5517	81.8182
Group18	13	3	82	233	13.6842	98.7288	81.25
Group19	10	2	62	259	13.8889	99.2337	83.3333
Group20	1	0	57	335	1.7241	100	100
Group21	1	0	62	320	1.5873	100	100
Group22	1	0	210	960	0.4739	100	100
Overall	2132	303	1048	2940	67.0440	90.6568	87.5565

Table B.5 Experiment result with FP/TP=14%

Table B.1 to B.5 is the summary of result of the experiment group by group under different FP and TP rate. We can find out that the value of Se drops when GC content decreases. PPV remains satisfyingly high under different GC content. The overall Se is 24.7484% in Table B.1, and 37.9874% in Table B.2, 51.6667% in Table B.3, 61.2579% in Table B.4, and 67.0440% in Table B.5, respectively. The overall PPV is 98.2522% in Table B.1, and 96.6400% in Table B.2, 93.0878% in Table B.3, 90.5204% in Table B.4, and 87.5565% in Table B.5, respectively. Se is calculated with TP over all positive, which is the sum of TP and FN. Se is also called “recall” rate. Sp is calculated with TN over all negative, which is the sum of TN and FP. PPV is calculated with TP over all prediction, which is the sum of TP and FP. PPV is also called “precision” rate. Se and PPV are two of the most commonly used performance criterions in promoter prediction.

Figure B.8 shown below is drawn with the all the positive and negative data using the two features of #CpG and GC content.

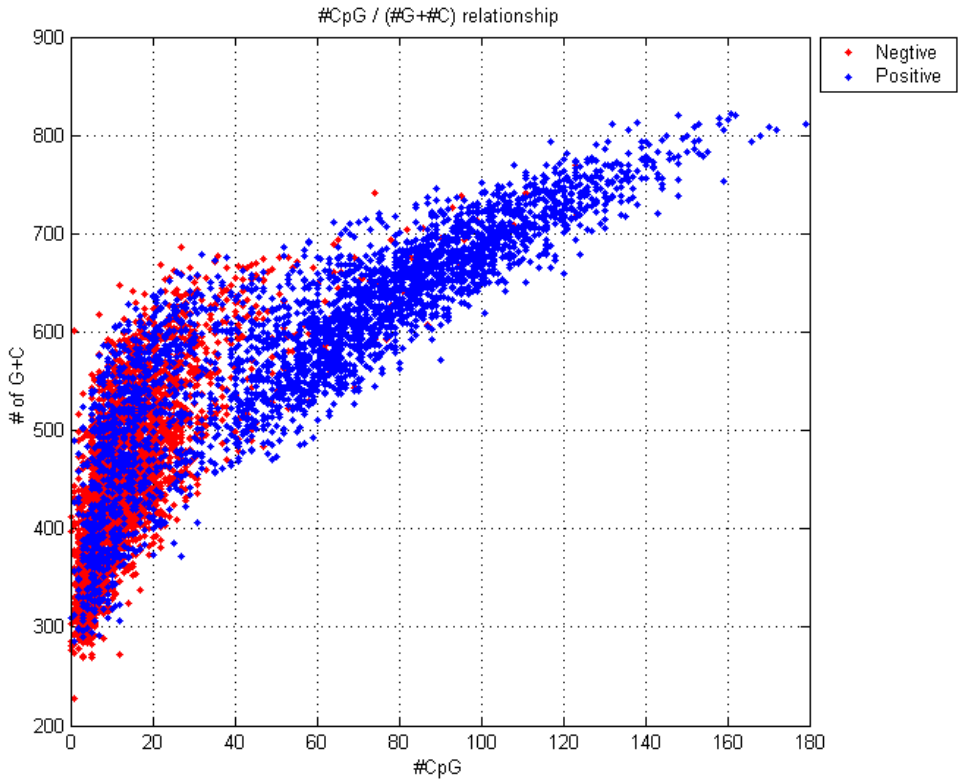


Figure B.8 The data under feature of #CpG and GC content

Each point in Figure B.8 is represented with two the coordinates of #CpG and GC content. #CpG is the number of CG di- nucleotides in the sequence. GC content is the sum of the number of G and C single nucleotides in the sequence over the sequence length. It can be found out that using the two features of #CpG and GC content, the positive data is separable from the negative data, which is a good indication that the adoption these two features is very promising in promoter prediction.

4. Combinational features with CC and #CpG

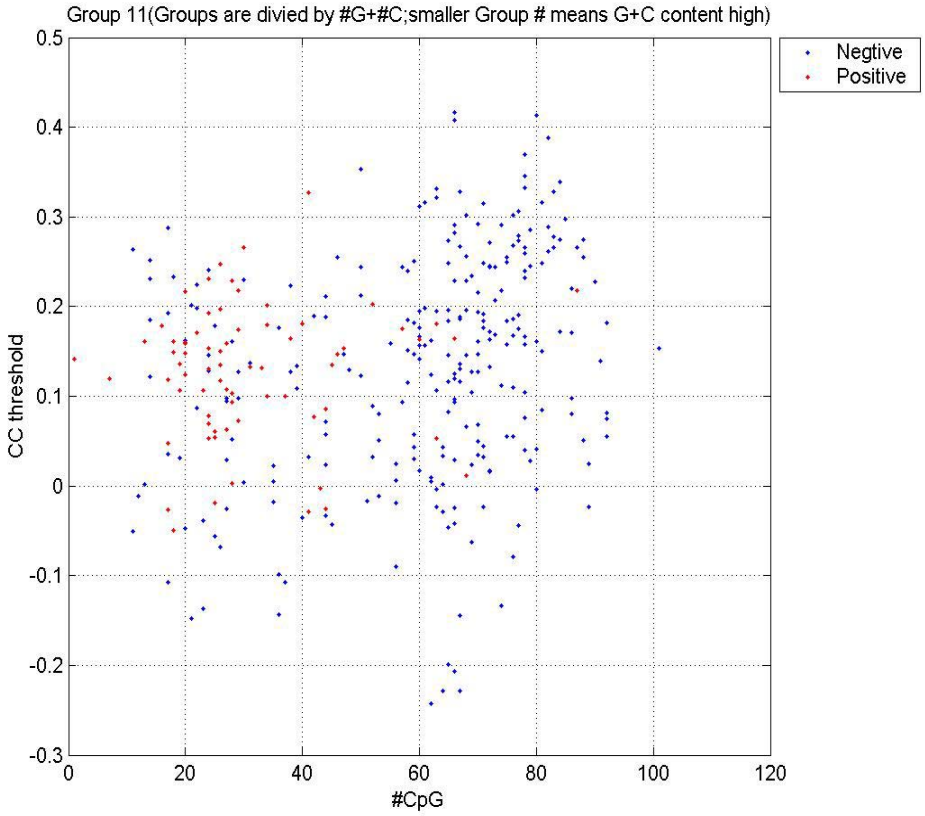


Figure B.9 Data represented by features of CC and #CpG (at level 2)

Figure B.9 is the positive and negative data in Group 11 under the features of CC and #CpG. The signal is decomposed in different levels, from level 1 to level 7; the CC is calculated using individual signal and the mean sequence of the reconstructed positive signal in respective level. Figure B.9 is obtained with signals decomposed and reconstructed at level 2 with DWT. Here the detailed (high frequency) part of the original signal is filtered. We can observe how the positive and negative samples are to be separable, using the two features of CC and #CpG.

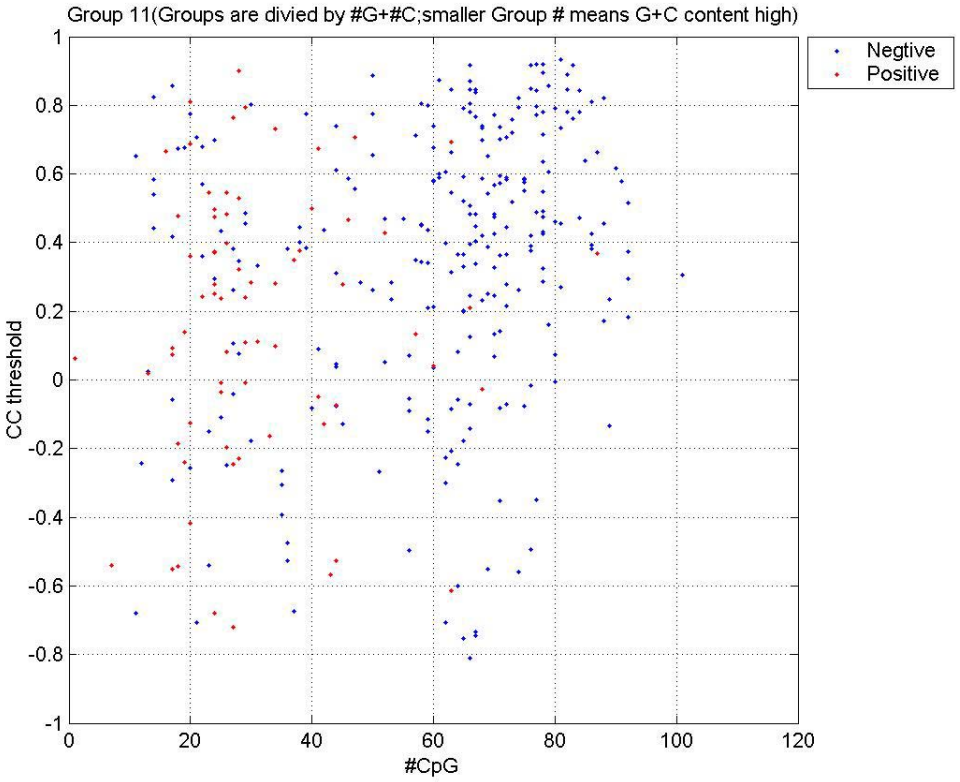


Figure B.10 Data represented by features of CC and #CpG (at level 7)

Figure B.10 shown below is different from Figure B.9 in that the signal is decomposed in level 7 rather in level 2 with DWT. The most proper level to decompose the signal is expected to be found out by comparison of the experimental results shown in these two figures. Here the detailed (high frequency) part of the original signal is filtered before reconstruction. From this figure, we can see that it is not as good as the figure obtained at level 2. So this means that the 'high frequency' part of the original signal that is filtered should not be too much. Comparatively, the resolution at level 2 is more appropriate for this task of separation.

5. Combinational features with CC and GC content

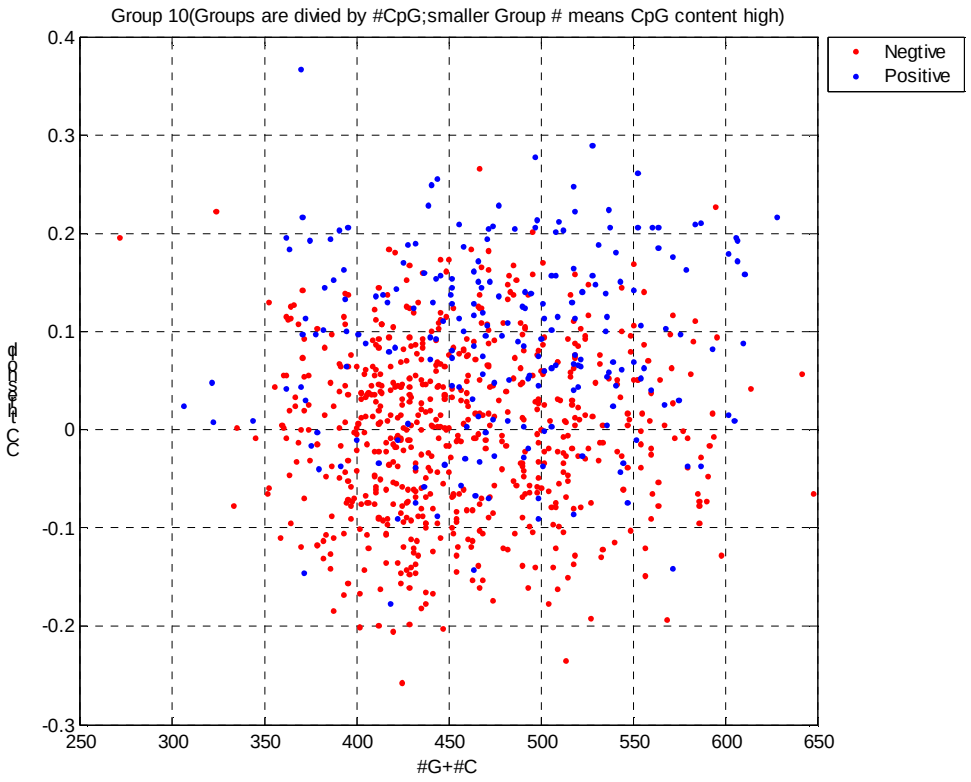


Figure B.11 Data represented by features of CC and GC content

Figure B.11 is the positive and negative data in Group 10 under the features of CC and GC content. The signal is decomposed in different levels, from level 1 to level 7, respectively. The CC is calculated using the individual signal and the mean sequence of the reconstructed positive signal at a level where the DWT is implemented. This plot is from the signal decomposed and reconstructed (after low pass filter) at level 2. Here we can observe whether the positive and negative samples are separable, using the two features of CC and GC content. We can observe that these two features are not as good as the previous two features of CC and #CpG for classification.

6. Combinational features with #CpG and GC content

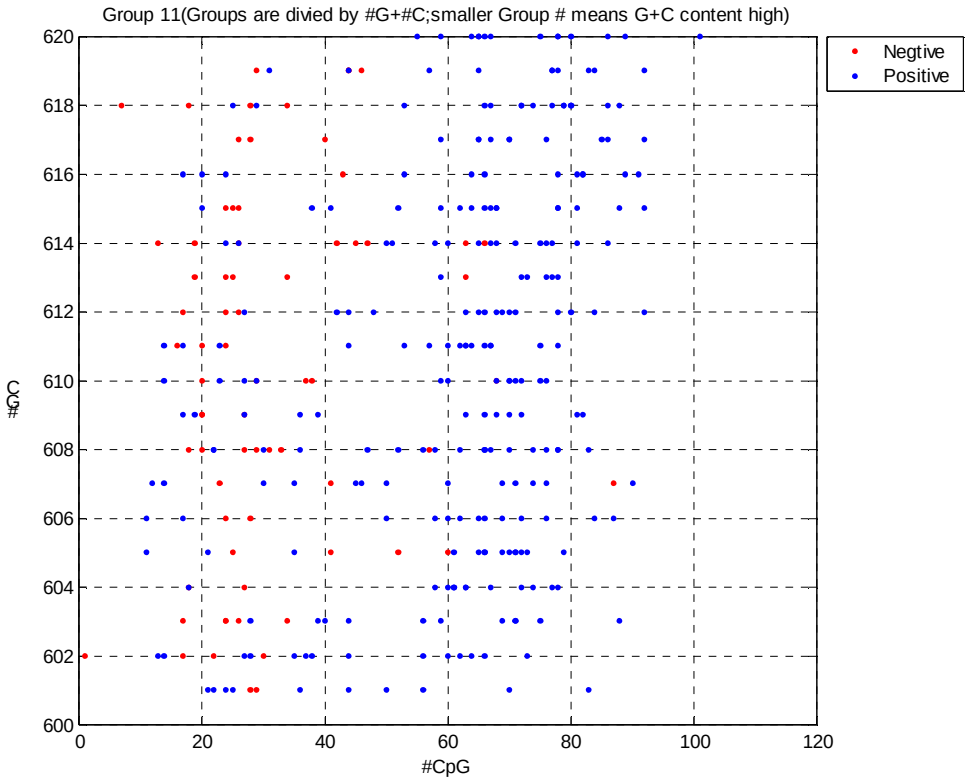


Figure B.12 Data represented by features of GC content and #CpG

Figure B.12 is the positive and negative data in Group 11 under the two features of GC content and #CpG. We can find out the ability of the two features in separating the data from Figure B.12. So Figure B.12 is given to facilitate the comparison of the three kinds of combinations shown in Figure B.10, Figure B.11, and Figure B.12.

7. Set threshold

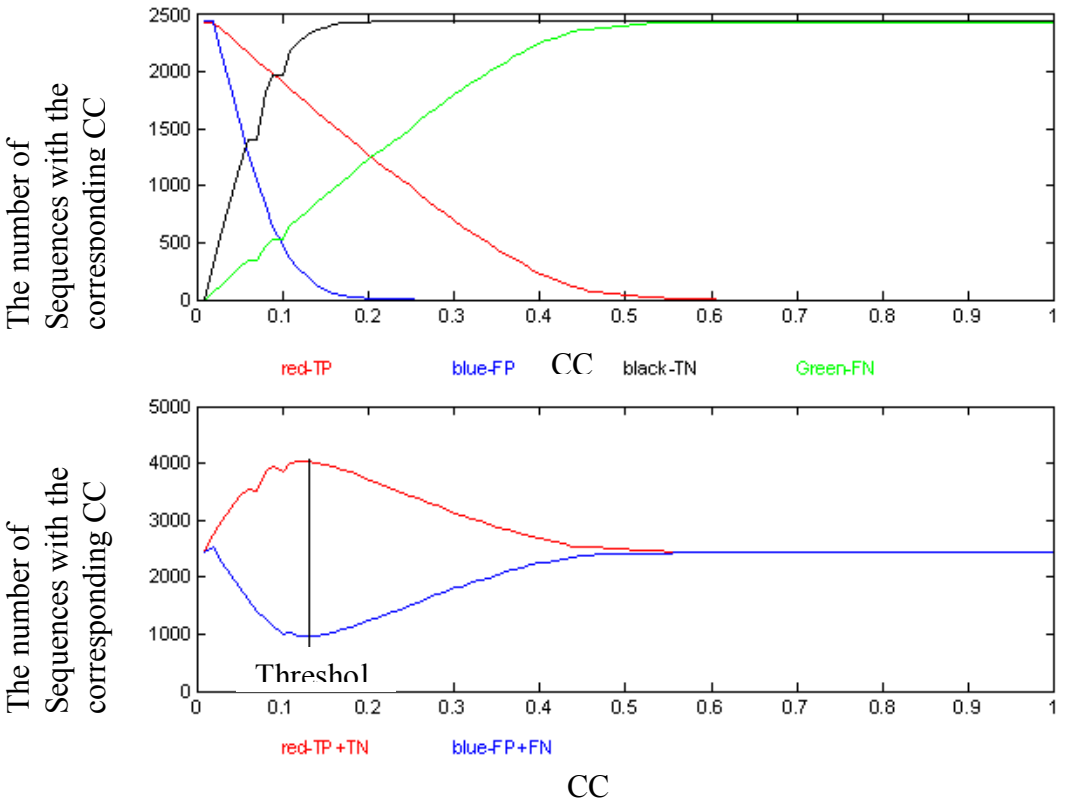


Figure B.13 The curves of TP, FP, TN, and FN under different thresholds

Figure B.13 give us the description of the process of how to decide the threshold based on the criterion, which is to maximize the correct predictions and minimize the incorrect predictions. The y axis is the number of sequences under the corresponding thresholds indicated by x axis. In the upper plot in Figure B.13, the curves of TP, FP, TN and FN are plotted under different thresholds from 0 to 1. In the lower plot in Figure B.13, the curves of correct predictions (TP+TN) and incorrect predictions (FP+FN) are plotted. Based on the lower plot, we can select the optimal threshold to make (TP+TN) maximum and (FP+FN) minimum concurrently. In Figure B.13, this optimal threshold should be at the CC value of 0.12.

Appendix C

1.Linear kernel

		c=0		c=0.000001		c=0.0001		c=0.01		c=1		c=1000
Group #	Se	PPV	Se	PPV	Se	PPV	Se	PPV	Se	PPV	Se	PPV
	TP	FN	TP	FN	TP	FN	TP	FN	TP	FN	TP	FN
	FP	TN	FP	TN	FP	TN	FP	TN	FP	TN	FP	TN
Group4	91.17	98.87	91.17	98.87	88.05	98.83	69.09	98.52	69.09	98.52	69.09	98.52
	0	4	0	4	0	4	0	4	0	4	0	4
	34	351	34	351	46	339	119	266	119	266	119	266
Group5	68.27	99.09	68.27	99.09	59.71	99.31	55.74	99.26	55.74	99.26	55.74	99.26
	2	3	2	3	3	2	3	2	3	2	3	2
	152	327	152	327	193	286	212	267	212	267	212	267
Group6	34.9	99.58	34.9	99.58	45.07	99.67	45.66	99.04	45.66	99.04	45.66	99.04
	12	1	12	1	12	1	10	3	10	3	10	3
	442	237	442	237	373	306	369	310	369	310	369	310
Group7	83.81	99.29	83.81	99.29	82.97	99.14	49.76	98.34	49.76	98.34	49.76	98.34
	13	5	13	5	12	6	11	7	11	7	11	7
	135	699	135	699	142	692	419	415	419	415	419	415
Group8	93.97	98.31	98.28	96.41	88.91	99.16	81.16	98.56	81.49	98.57	81.49	98.57
	30	15	11	34	38	7	34	11	34	11	34	11
	56	873	16	913	103	826	175	754	172	757	172	757
Group9	87.2	98.23	98.11	94.25	88.46	98.48	79.64	97.81	65.16	95.83	65.58	96.15
	62	15	20	57	64	13	60	17	50	27	52	25
	122	831	18	935	110	843	194	759	332	621	328	625
Group10	84.1	97.51	83.02	97.59	85.85	97.13	78.34	96.05	71.51	95.19	47.9	93.52
	105	22	106	21	101	26	94	33	90	37	93	34
	163	862	174	851	145	880	222	803	292	733	534	491
Group11	76.91	98.16	90.72	91.67	78.97	98.21	74.33	96.91	75.36	95.31	63.51	96.4
	127	14	61	80	127	14	118	23	105	36	118	23
	224	746	90	880	204	766	249	721	239	731	354	616
Group12	75.63	97.58	71.25	98.42	77.71	97.14	76.46	95.95	76.25	94.45	70.31	90.6
	156	18	163	11	152	22	143	31	131	43	104	70
	234	726	276	684	214	746	226	734	228	732	285	675
Group13	70.76	97.31	65.4	98.04	72.85	96.88	72.98	96.71	72.72	95.38	59.4	79.68
	199	15	204	10	196	18	195	19	187	27	98	116
	224	542	265	501	208	558	207	559	209	557	311	455
Group14	66.33	97.56	62.35	98.17	68.16	97.39	73.47	93.86	73.96	93.11	33.5	76.52
	262	10	265	7	261	11	243	29	239	33	210	62
	203	400	227	376	192	411	160	443	157	446	401	202
Group15	60.86	90.05	73.09	68.68	64.53	88.28	68.81	78.4	72.48	76.95	71.25	60.36
	347	22	260	109	341	28	307	62	298	71	216	153
	128	199	88	239	116	211	102	225	90	237	94	233
Group16	54.95	55.35	53.48	53.09	61.54	69.14	69.96	54.42	70.7	54.99	78.39	35.37
	463	121	455	129	509	75	424	160	426	158	193	391

	123	150	127	146	105	168	82	191	80	193	59	214
Group17	58.13	36.02	86.18	29.08	61.38	53.17	64.23	48.92	71.95	50.43	70.33	33.14
	563	254	300	517	684	133	652	165	643	174	468	349
	103	143	34	212	95	151	88	158	69	177	73	173
Group18	35.71	24.25	58.24	21.37	73.08	34.91	71.43	32.91	71.43	32.26	58.79	27.86
	704	203	517	390	659	248	642	265	634	273	630	277
	117	65	76	106	49	133	52	130	52	130	75	107
Group19	68.67	18.97	68.67	18.97	72.89	27.94	63.86	27.89	62.65	26	80.12	17.14
	536	487	536	487	711	312	749	274	727	296	380	643
	52	114	52	114	45	121	60	106	62	104	33	133
Group20	66.87	16.2	53.37	16.48	76.69	22.94	73.01	23.71	71.17	22.75	49.69	17.46
	527	564	650	441	671	420	708	383	697	394	708	383
	54	109	76	87	38	125	44	119	47	116	82	81
Group21	57.52	15.91	64.05	16.17	71.9	22.36	62.75	21.82	64.71	22.86	47.06	14.04
	676	465	633	508	759	382	797	344	807	334	700	441
	65	88	55	98	43	110	57	96	54	99	81	72
Group22	77.95	12.81	75.06	12.82	64.81	16.88	65.92	16.4	76.39	10.13	22.94	8.65
	1553	2382	1644	2291	2502	1433	2426	1509	892	3043	2847	1088
	99	350	112	337	158	291	153	296	106	343	346	103

Table A.1a Use linear kernel

	Op_c (F-measure)	Se	PPV
Group4	0.000001	91.17	98.87
Group5	0	68.27	99.09
Group6	0.01	45.66	99.04
Group7	0	83.81	99.29
Group8	0.000001	98.28	96.41
Group9	0.000001	98.11	94.25
Group10	0.0001	85.85	97.13
Group11	0.000001	90.72	91.67
Group12	0.0001	77.71	97.14
Group13	0.01	72.98	96.71
Group14	1	73.96	93.11
Group15	1	72.48	76.95
Group16	0.0001	61.54	69.14
Group17	1	71.95	50.43
Group18	0.0001	73.08	34.91
Group19	0.0001	72.89	27.94
Group20	0.01	73.01	23.71
Group21	0.0001	71.9	22.36
Group22	0.0001	64.81	16.88
over all		79.700249	71.409142

Table A.1b Use linear kernel

2. Sigmoid kernel

	r=-2		r=-1		r=0		r=1		r=2	
	s=0.000010									
Group #	Se	PPV	Se	PPV	Se	PPV	Se	PPV	Se	PPV
	TP	FN	TP	FN	TP	FN	TP	FN	TP	FN
	FP	TN	FP	TN	FP	TN	FP	TN	FP	TN
Group4	84.94	98.79	84.94	98.79	84.94	98.79	84.94	98.79	84.94	98.79
	0	4	0	4	0	4	0	4	0	4
	58	327	58	327	58	327	58	327	58	327
Group5	65.34	99.05	65.34	99.05	65.34	99.05	65.14	99.05	64.93	99.04
	2	3	2	3	2	3	2	3	2	3
	166	313	166	313	166	313	167	312	168	311
Group6	52.28	99.44	52.43	99.44	52.43	99.44	52.28	99.44	51.4	99.43
	11	2	11	2	11	2	11	2	11	2
	324	355	323	356	323	356	324	355	330	349
Group7	77.34	99.38	77.34	99.38	77.34	99.38	77.34	99.38	76.02	99.37
	14	4	14	4	14	4	14	4	14	4
	189	645	189	645	189	645	189	645	200	634
Group8	51.35	98.96	40.9	99.22	39.07	99.18	38.75	99.17	38.97	99.18
	40	5	42	3	42	3	42	3	42	3
	452	477	549	380	566	363	569	360	567	362
Group9	94.86	95.86	94.86	95.86	94.86	95.86	94.86	95.86	94.86	95.86
	38	39	38	39	38	39	38	39	38	39
	49	904	49	904	49	904	49	904	49	904
Group10	81.07	97.76	80.98	97.76	80.98	97.76	80.98	97.76	80.88	97.87
	108	19	108	19	108	19	108	19	109	18
	194	831	195	830	195	830	195	830	196	829
Group11	68.14	97.93	66.49	97.88	66.29	97.87	66.29	97.87	66.29	97.87
	127	14	127	14	127	14	127	14	127	14
	309	661	325	645	327	643	327	643	327	643
Group12	72.6	98.45	72.6	98.45	72.5	98.44	72.5	98.44	72.6	98.45
	163	11	163	11	163	11	163	11	163	11
	263	697	263	697	264	696	264	696	263	697
Group13	65.8	97.86	64.88	97.83	64.75	97.83	64.75	97.83	64.75	97.83
	203	11	203	11	203	11	203	11	203	11
	262	504	269	497	270	496	270	496	270	496
Group14	64.84	98.24	64.84	97.99	64.84	97.99	64.84	97.99	65.01	98
	265	7	264	8	264	8	264	8	264	8
	212	391	212	391	212	391	212	391	211	392
Group15	74.01	67.22	65.44	79.26	74.31	66.39	74.31	66.39	74.31	65.85
	251	118	313	56	246	123	246	123	243	126
	85	242	113	214	84	243	84	243	84	243
Group16	67.03	48.8	65.57	48.51	65.57	48.64	65.57	48.64	65.57	48.51
	392	192	394	190	395	189	395	189	394	190

	90	183	94	179	94	179	94	179	94	179
Group17	45.53	39.44	41.87	38.72	41.46	38.64	41.46	38.64	41.06	38.4
	645	172	654	163	655	162	655	162	655	162
	134	112	143	103	144	102	144	102	145	101
Group18	12.64	24.21	12.64	25.56	12.64	26.14	12.64	26.14	12.64	27.06
	835	72	840	67	842	65	842	65	845	62
	159	23	159	23	159	23	159	23	159	23
Group19	53.61	20.55	53.61	20.84	53.61	20.89	53.61	20.89	53.61	20.89
	679	344	685	338	686	337	686	337	686	337
	77	89	77	89	77	89	77	89	77	89
Group20	31.9	14.9	30.67	14.62	30.67	14.66	30.67	14.66	30.67	14.79
	794	297	799	292	800	291	800	291	803	288
	111	52	113	50	113	50	113	50	113	50
Group21	50.33	15.16	50.33	15.37	50.33	15.37	50.33	15.37	50.33	15.37
	710	431	717	424	717	424	717	424	717	424
	76	77	76	77	76	77	76	77	76	77
Group22	8.02	8.91	7.35	8.44	7.35	8.59	7.35	8.59	7.35	8.59
	3567	368	3577	358	3584	351	3584	351	3584	351
	413	36	416	33	416	33	416	33	416	33

Table A.2a Use sigmoid kernel

	Op_r(F-measure)|	Op_s(F-measure)	Se	PPV
Group4	2	0.00001	84.94	98.79
Group5	-2	0.00001	65.34	99.05
Group6	-1	0.00001	52.43	99.44
Group7	-2	0.00001	77.34	99.38
Group8	-2	0.00001	51.35	98.96
Group9	2	0.00001	94.86	95.86
Group10	-2	0.00001	81.07	97.76
Group11	-2	0.00001	68.14	97.93
Group12	-2	0.00001	72.6	98.45
Group13	-2	0.00001	65.8	97.86
Group14	2	0.00001	65.01	98
Group15	-1	0.00001	65.44	79.26
Group16	-2	0.00001	67.03	48.8
Group17	-2	0.00001	45.53	39.44
Group18	2	0.00001	12.64	27.06
Group19	0	0.00001	53.61	20.89
Group20	-2	0.00001	31.9	14.9
over all			68.209259	84.580841

Table A.2b Use sigmoid kernel

Group 1-22

1. Coding-gene, at the distance of 100bp

Threshold	reference	j	TP	HitTP	FP	Se	PPV
0	201	27702	186	664	6062	92.5373	2.97695
0.5	201	26151	180	609	5864	89.5522	2.97816
1	201	21691	171	509	4874	85.0746	3.3895
1.5	201	15876	140	373	3582	69.6517	3.76142
2	201	12244	118	302	2830	58.7065	4.00271
2.5	201	10657	104	263	2541	51.7413	3.93195
3	201	10071	99	251	2433	49.2537	3.90995
3.5	201	9816	97	243	2383	48.2587	3.91129
4	201	9720	97	242	2367	48.2587	3.93669
4.5	201	9702	97	242	2366	48.2587	3.93829
5	201	9685	97	242	2366	48.2587	3.93829
5.5	201	9671	97	242	2364	48.2587	3.94149
6	201	9669	97	242	2364	48.2587	3.94149
6.5	201	9670	97	242	2362	48.2587	3.94469
7	201	9667	97	242	2359	48.2587	3.94951
7.5	201	9657	97	242	2358	48.2587	3.95112
8	201	9655	97	242	2358	48.2587	3.95112
8.5	201	9655	97	242	2358	48.2587	3.95112
9	201	9654	97	242	2358	48.2587	3.95112
9.5	201	9654	97	242	2358	48.2587	3.95112
10	201	9654	97	242	2358	48.2587	3.95112
10.5	201	9653	97	242	2358	48.2587	3.95112
11	201	9653	97	242	2358	48.2587	3.95112
11.5	201	9653	97	242	2358	48.2587	3.95112
12	201	9653	97	242	2358	48.2587	3.95112
12.5	201	9653	97	242	2358	48.2587	3.95112
13	201	9653	97	242	2358	48.2587	3.95112
13.5	201	9653	97	242	2358	48.2587	3.95112
14	201	9653	97	242	2358	48.2587	3.95112
14.5	201	9653	97	242	2358	48.2587	3.95112
15	201	9653	97	242	2358	48.2587	3.95112
15.5	201	9653	97	242	2358	48.2587	3.95112
16	201	9653	97	242	2358	48.2587	3.95112
16.5	201	9653	97	242	2358	48.2587	3.95112
17	201	9653	97	242	2358	48.2587	3.95112
17.5	201	9653	97	242	2358	48.2587	3.95112
18	201	9653	97	242	2358	48.2587	3.95112
18.5	201	9653	97	242	2358	48.2587	3.95112
19	201	9652	97	242	2358	48.2587	3.95112
19.5	201	9652	97	242	2358	48.2587	3.95112
20	201	9652	97	242	2358	48.2587	3.95112
20.5	201	9651	97	242	2357	48.2587	3.95273
21	201	9651	97	242	2357	48.2587	3.95273
21.5	201	9651	97	242	2357	48.2587	3.95273
22	201	9651	97	242	2357	48.2587	3.95273
22.5	201	9651	97	242	2357	48.2587	3.95273
23	201	9651	97	242	2357	48.2587	3.95273
23.5	201	9650	97	242	2356	48.2587	3.95434

24	201	9650	97	242	2356	48.2587	3.95434
24.5	201	9649	97	242	2357	48.2587	3.95273
25	201	9648	97	242	2357	48.2587	3.95273
25.5	201	9647	97	242	2357	48.2587	3.95273
26	201	9646	97	242	2357	48.2587	3.95273
26.5	201	9648	97	242	2358	48.2587	3.95112
27	201	9644	97	242	2356	48.2587	3.95434
27.5	201	9637	97	242	2354	48.2587	3.95757
28	201	9633	97	242	2353	48.2587	3.95918
28.5	201	9628	97	241	2352	48.2587	3.9608
29	201	9625	97	241	2353	48.2587	3.95918
29.5	201	9620	97	241	2352	48.2587	3.9608
30	201	9605	97	243	2348	48.2587	3.96728
30.5	201	9591	97	242	2351	48.2587	3.96242
31	201	9571	97	241	2351	48.2587	3.96242
31.5	201	9557	97	239	2339	48.2587	3.98194
32	201	9523	96	237	2332	47.7612	3.95387
32.5	201	9494	96	235	2333	47.7612	3.95224
33	201	9460	96	232	2332	47.7612	3.95387
33.5	201	9376	96	230	2307	47.7612	3.99501
34	201	9263	96	232	2289	47.7612	4.02516
34.5	201	9110	96	228	2246	47.7612	4.09906
35	201	8889	95	225	2197	47.2637	4.14485
35.5	201	8642	94	220	2145	46.7662	4.1983
36	201	8337	89	205	2043	44.2786	4.17448
36.5	201	7917	86	188	1930	42.7861	4.26587
37	201	7399	84	181	1788	41.791	4.48718
37.5	201	6810	78	168	1631	38.806	4.56407
38	201	6097	73	154	1459	36.3184	4.76501
38.5	201	5368	66	130	1286	32.8358	4.88166
39	201	4602	56	104	1096	27.8607	4.86111
39.5	201	3868	50	87	912	24.8756	5.19751
40	201	3204	41	73	739	20.398	5.25641
40.5	201	2608	37	64	597	18.408	5.83596
41	201	2106	29	51	468	14.4279	5.83501
41.5	201	1641	24	40	361	11.9403	6.23377
42	201	1285	20	33	280	9.95025	6.66667
42.5	201	994	16	28	220	7.9602	6.77966
43	201	727	13	23	162	6.46766	7.42857
43.5	201	551	11	19	128	5.47264	7.91367
44	201	429	11	18	101	5.47264	9.82143
44.5	201	327	10	17	67	4.97512	12.987
45	201	256	8	15	48	3.9801	14.2857
45.5	201	187	5	11	32	2.48756	13.5135
46	201	144	5	9	27	2.48756	15.625

Table A.3 prediction result of "Coding-gene" for Group 1-22

2. Non-coding gene, at the distance of 600bp

Threshold	reference	j	TP	HitTP	FP	Se	PPV
0	14	8175	8	11	10	57.1429	44.4444
0.5	14	10075	10	17	5	71.4286	66.6667
1	14	9939	8	12	5	57.1429	61.5385
1.5	14	8464	7	9	3	50	70
2	14	7106	6	7	3	42.8571	66.6667
2.5	14	6443	3	4	3	21.4286	50
3	14	6200	3	4	3	21.4286	50

3.5	14	6116	3	4	3	21.4286	50
4	14	6058	3	4	3	21.4286	50
4.5	14	6052	3	4	3	21.4286	50
5	14	6039	3	4	3	21.4286	50
5.5	14	6031	3	4	3	21.4286	50
6	14	6030	3	4	3	21.4286	50
6.5	14	6027	3	4	3	21.4286	50
7	14	6026	3	4	3	21.4286	50
7.5	14	6021	3	4	3	21.4286	50
8	14	6019	3	4	3	21.4286	50
8.5	14	6019	3	4	3	21.4286	50
9	14	6018	3	4	3	21.4286	50
9.5	14	6019	3	4	3	21.4286	50
10	14	6020	3	4	3	21.4286	50
10.5	14	6019	3	4	3	21.4286	50
11	14	6019	3	4	3	21.4286	50
11.5	14	6019	3	4	3	21.4286	50
12	14	6019	3	4	3	21.4286	50
12.5	14	6019	3	4	3	21.4286	50
13	14	6019	3	4	3	21.4286	50
13.5	14	6019	3	4	3	21.4286	50
14	14	6019	3	4	3	21.4286	50
14.5	14	6019	3	4	3	21.4286	50
15	14	6019	3	4	3	21.4286	50
15.5	14	6019	3	4	3	21.4286	50
16	14	6019	3	4	3	21.4286	50
16.5	14	6019	3	4	3	21.4286	50
17	14	6019	3	4	3	21.4286	50
17.5	14	6019	3	4	3	21.4286	50
18	14	6019	3	4	3	21.4286	50
18.5	14	6019	3	4	3	21.4286	50
19	14	6018	3	4	3	21.4286	50
19.5	14	6018	3	4	3	21.4286	50
20	14	6019	3	4	3	21.4286	50
20.5	14	6019	3	4	3	21.4286	50
21	14	6019	3	4	3	21.4286	50
21.5	14	6019	3	4	3	21.4286	50
22	14	6019	3	4	3	21.4286	50
22.5	14	6019	3	4	3	21.4286	50
23	14	6019	3	4	3	21.4286	50
23.5	14	6020	3	4	3	21.4286	50
24	14	6020	3	4	3	21.4286	50
24.5	14	6019	3	4	3	21.4286	50
25	14	6018	3	4	3	21.4286	50
25.5	14	6017	3	4	3	21.4286	50
26	14	6017	3	4	3	21.4286	50
26.5	14	6017	3	4	3	21.4286	50
27	14	6017	3	4	3	21.4286	50
27.5	14	6015	3	4	3	21.4286	50
28	14	6011	3	4	3	21.4286	50
28.5	14	6009	3	4	3	21.4286	50
29	14	6006	3	4	3	21.4286	50
29.5	14	6005	3	4	3	21.4286	50
30	14	5996	3	4	3	21.4286	50
30.5	14	5988	3	4	3	21.4286	50
31	14	5979	3	4	3	21.4286	50
31.5	14	5967	3	4	3	21.4286	50
32	14	5946	3	4	3	21.4286	50
32.5	14	5932	3	4	3	21.4286	50
33	14	5911	2	3	3	14.2857	40
33.5	14	5886	2	3	3	14.2857	40
34	14	5853	2	3	3	14.2857	40

34.5	14	5800	2	3	3	14.2857	40
35	14	5682	2	3	3	14.2857	40
35.5	14	5557	2	3	3	14.2857	40
36	14	5391	2	3	3	14.2857	40
36.5	14	5198	2	4	3	14.2857	40
37	14	4910	2	4	2	14.2857	50
37.5	14	4580	2	4	2	14.2857	50
38	14	4204	2	4	2	14.2857	50
38.5	14	3789	2	4	1	14.2857	66.6667
39	14	3324	2	4	1	14.2857	66.6667
39.5	14	2836	2	4	0	14.2857	100
40	14	2368	2	2	0	14.2857	100
40.5	14	1999	2	2	0	14.2857	100
41	14	1636	2	2	0	14.2857	100
41.5	14	1294	2	2	0	14.2857	100
42	14	1033	0	0	0	0	-1.#IND00
42.5	14	809	0	0	0	0	-1.#IND00
43	14	609	0	0	0	0	-1.#IND00
43.5	14	457	0	0	0	0	-1.#IND00
44	14	364	0	0	0	0	-1.#IND00
44.5	14	276	0	0	0	0	-1.#IND00
45	14	210	0	0	0	0	-1.#IND00
45.5	14	154	0	0	0	0	-1.#IND00
46	14	120	0	0	0	0	-1.#IND00

Table A.4 prediction result of "Non-coding-gene" for Group 1-22

3. Pseudo gene , at the distance of 1000bp

Threshold	reference	j	TP	HitTP	FP	Se	PPV
0	122	4850	52	81	29	42.623	64.1975
0.5	122	6982	70	99	23	57.377	75.2688
1	122	7451	63	82	32	51.6393	66.3158
1.5	122	6679	55	65	39	45.082	58.5106
2	122	5720	42	54	33	34.4262	56
2.5	122	5230	41	53	33	33.6066	55.4054
3	122	5047	35	48	30	28.6885	53.8462
3.5	122	4982	35	47	30	28.6885	53.8462
4	122	4947	35	47	29	28.6885	54.6875
4.5	122	4941	34	46	29	27.8689	53.9683
5	122	4933	34	46	29	27.8689	53.9683
5.5	122	4929	34	46	29	27.8689	53.9683
6	122	4926	34	46	29	27.8689	53.9683
6.5	122	4924	34	46	29	27.8689	53.9683
7	122	4926	34	46	29	27.8689	53.9683
7.5	122	4922	34	46	29	27.8689	53.9683
8	122	4920	34	46	29	27.8689	53.9683
8.5	122	4920	34	46	29	27.8689	53.9683
9	122	4919	34	46	29	27.8689	53.9683
9.5	122	4919	34	46	29	27.8689	53.9683
10	122	4919	34	46	29	27.8689	53.9683
10.5	122	4920	34	46	29	27.8689	53.9683
11	122	4920	34	46	29	27.8689	53.9683
11.5	122	4920	34	46	29	27.8689	53.9683
12	122	4920	34	46	29	27.8689	53.9683
12.5	122	4920	34	46	29	27.8689	53.9683
13	122	4920	34	46	29	27.8689	53.9683
13.5	122	4920	34	46	29	27.8689	53.9683
14	122	4920	34	46	29	27.8689	53.9683
14.5	122	4920	34	46	29	27.8689	53.9683
15	122	4920	34	46	29	27.8689	53.9683
15.5	122	4920	34	46	29	27.8689	53.9683

16	122	4920	34	46	29	27.8689	53.9683
16.5	122	4920	34	46	29	27.8689	53.9683
17	122	4920	34	46	29	27.8689	53.9683
17.5	122	4920	34	46	29	27.8689	53.9683
18	122	4920	34	46	29	27.8689	53.9683
18.5	122	4920	34	46	29	27.8689	53.9683
19	122	4919	34	46	29	27.8689	53.9683
19.5	122	4920	34	46	29	27.8689	53.9683
20	122	4921	34	46	29	27.8689	53.9683
20.5	122	4921	34	46	29	27.8689	53.9683
21	122	4921	34	46	29	27.8689	53.9683
21.5	122	4921	34	46	29	27.8689	53.9683
22	122	4921	34	46	29	27.8689	53.9683
22.5	122	4921	34	46	29	27.8689	53.9683
23	122	4921	34	46	29	27.8689	53.9683
23.5	122	4921	34	46	29	27.8689	53.9683
24	122	4921	34	46	29	27.8689	53.9683
24.5	122	4922	34	46	29	27.8689	53.9683
25	122	4921	34	46	29	27.8689	53.9683
25.5	122	4920	34	46	29	27.8689	53.9683
26	122	4920	34	46	29	27.8689	53.9683
26.5	122	4920	34	46	29	27.8689	53.9683
27	122	4919	34	46	29	27.8689	53.9683
27.5	122	4918	34	46	29	27.8689	53.9683
28	122	4916	34	46	28	27.8689	54.8387
28.5	122	4914	34	46	28	27.8689	54.8387
29	122	4914	34	46	28	27.8689	54.8387
29.5	122	4915	35	47	28	28.6885	55.5556
30	122	4909	36	48	28	29.5082	56.25
30.5	122	4900	35	47	28	28.6885	55.5556
31	122	4897	37	49	28	30.3279	56.9231
31.5	122	4886	35	47	29	28.6885	54.6875
32	122	4873	35	48	29	28.6885	54.6875
32.5	122	4865	35	47	29	28.6885	54.6875
33	122	4846	35	47	27	28.6885	56.4516
33.5	122	4824	35	48	28	28.6885	55.5556
34	122	4819	35	47	26	28.6885	57.377
34.5	122	4781	34	46	27	27.8689	55.7377
35	122	4733	34	47	27	27.8689	55.7377
35.5	122	4658	35	46	26	28.6885	57.377
36	122	4541	34	45	27	27.8689	55.7377
36.5	122	4408	35	46	26	28.6885	57.377
37	122	4213	34	44	24	27.8689	58.6207
37.5	122	4006	32	40	22	26.2295	59.2593
38	122	3708	31	38	22	25.4098	58.4906
38.5	122	3382	30	36	16	24.5902	65.2174
39	122	2976	27	32	15	22.1311	64.2857
39.5	122	2572	23	28	14	18.8525	62.1622
40	122	2169	17	22	11	13.9344	60.7143
40.5	122	1847	14	17	11	11.4754	56
41	122	1520	13	15	9	10.6557	59.0909
41.5	122	1222	10	12	8	8.19672	55.5556
42	122	983	6	7	5	4.91803	54.5455
42.5	122	777	5	6	4	4.09836	55.5556
43	122	593	5	6	3	4.09836	62.5
43.5	122	444	3	3	3	2.45902	50
44	122	353	3	3	3	2.45902	50
44.5	122	268	3	3	3	2.45902	50
45	122	204	3	3	3	2.45902	50
45.5	122	149	2	2	3	1.63934	40
46	122	116	1	1	3	0.81967	25

Table A.5 Prediction result of "Pseudo gene" for Group 1-22

4. Partial gene, at the distance of 1500bp

Threshold	reference	j	TP	HitTP	FP	Se	PPV
0	78	2538	17	24	85	21.7949	16.6667
0.5	78	4440	39	48	102	50	27.6596
1	78	5211	51	58	98	65.3846	34.2282
1.5	78	4956	42	48	98	53.8462	30
2	78	4293	38	41	86	48.7179	30.6452
2.5	78	3882	29	31	77	37.1795	27.3585
3	78	3731	28	30	79	35.8974	26.1682
3.5	78	3673	27	29	71	34.6154	27.551
4	78	3638	26	28	73	33.3333	26.2626
4.5	78	3633	26	28	72	33.3333	26.5306
5	78	3627	26	28	72	33.3333	26.5306
5.5	78	3627	26	28	72	33.3333	26.5306
6	78	3624	26	28	72	33.3333	26.5306
6.5	78	3624	26	28	73	33.3333	26.2626
7	78	3626	26	28	73	33.3333	26.2626
7.5	78	3623	26	28	73	33.3333	26.2626
8	78	3621	26	28	73	33.3333	26.2626
8.5	78	3621	26	28	73	33.3333	26.2626
9	78	3620	26	28	73	33.3333	26.2626
9.5	78	3620	26	28	73	33.3333	26.2626
10	78	3620	26	28	73	33.3333	26.2626
10.5	78	3620	26	28	73	33.3333	26.2626
11	78	3620	26	28	73	33.3333	26.2626
11.5	78	3620	26	28	73	33.3333	26.2626
12	78	3620	26	28	73	33.3333	26.2626
12.5	78	3620	26	28	73	33.3333	26.2626
13	78	3620	26	28	73	33.3333	26.2626
13.5	78	3620	26	28	73	33.3333	26.2626
14	78	3621	26	28	73	33.3333	26.2626
14.5	78	3621	26	28	73	33.3333	26.2626
15	78	3621	26	28	73	33.3333	26.2626
15.5	78	3621	26	28	73	33.3333	26.2626
16	78	3621	26	28	73	33.3333	26.2626
16.5	78	3621	26	28	73	33.3333	26.2626
17	78	3621	26	28	73	33.3333	26.2626
17.5	78	3621	26	28	73	33.3333	26.2626
18	78	3621	26	28	73	33.3333	26.2626
18.5	78	3621	26	28	73	33.3333	26.2626
19	78	3620	26	28	73	33.3333	26.2626
19.5	78	3620	26	28	73	33.3333	26.2626
20	78	3620	26	28	73	33.3333	26.2626
20.5	78	3621	26	28	73	33.3333	26.2626
21	78	3621	26	28	73	33.3333	26.2626
21.5	78	3621	26	28	73	33.3333	26.2626
22	78	3621	26	28	73	33.3333	26.2626
22.5	78	3621	26	28	73	33.3333	26.2626
23	78	3621	26	28	73	33.3333	26.2626
23.5	78	3621	26	28	73	33.3333	26.2626
24	78	3621	26	28	73	33.3333	26.2626
24.5	78	3622	26	28	74	33.3333	26
25	78	3623	26	28	74	33.3333	26
25.5	78	3623	26	28	73	33.3333	26.2626
26	78	3623	26	28	73	33.3333	26.2626
26.5	78	3622	26	28	73	33.3333	26.2626
27	78	3621	26	28	73	33.3333	26.2626
27.5	78	3619	26	28	73	33.3333	26.2626

28	78	3619	26	28	73	33.3333	26.2626
28.5	78	3620	27	29	73	34.6154	27
29	78	3620	27	29	75	34.6154	26.4706
29.5	78	3618	26	28	73	33.3333	26.2626
30	78	3620	26	28	74	33.3333	26
30.5	78	3616	26	28	72	33.3333	26.5306
31	78	3616	26	28	73	33.3333	26.2626
31.5	78	3613	26	28	76	33.3333	25.4902
32	78	3610	26	28	72	33.3333	26.5306
32.5	78	3610	26	29	72	33.3333	26.5306
33	78	3602	27	31	72	34.6154	27.2727
33.5	78	3607	28	32	72	35.8974	28
34	78	3612	26	29	71	33.3333	26.8041
34.5	78	3603	28	32	68	35.8974	29.1667
35	78	3589	27	32	69	34.6154	28.125
35.5	78	3569	26	29	68	33.3333	27.6596
36	78	3539	25	27	71	32.0513	26.0417
36.5	78	3474	24	25	64	30.7692	27.2727
37	78	3385	25	27	60	32.0513	29.4118
37.5	78	3278	26	27	56	33.3333	31.7073
38	78	3099	27	30	50	34.6154	35.0649
38.5	78	2891	26	27	45	33.3333	36.6197
39	78	2607	24	24	40	30.7692	37.5
39.5	78	2296	22	23	35	28.2051	38.5965
40	78	1970	16	16	32	20.5128	33.3333
40.5	78	1709	14	14	30	17.9487	31.8182
41	78	1418	13	13	25	16.6667	34.2105
41.5	78	1157	10	11	20	12.8205	33.3333
42	78	941	9	9	16	11.5385	36
42.5	78	752	7	7	13	8.97436	35
43	78	578	7	7	10	8.97436	41.1765
43.5	78	436	6	6	5	7.69231	54.5455
44	78	347	6	6	5	7.69231	54.5455
44.5	78	265	3	3	4	3.84615	42.8571
45	78	203	1	1	3	1.28205	25
45.5	78	148	1	1	2	1.28205	33.3333
46	78	115	0	0	1	0	0

Table A.6 Prediction result of "partial gene" for Group 1-22

5. IGLV/J, at the distance of 100bp

Threshold	reference	j	TP	HitTP	FP	Se	PPV
0	122	27702	118	469	0	96.7213	100
0.5	122	26151	116	427	0	95.082	100
1	122	21691	103	275	0	84.4262	100
1.5	122	15876	85	206	0	69.6721	100
2	122	12244	76	165	0	62.2951	100
2.5	122	10657	73	151	0	59.8361	100
3	122	10071	73	148	0	59.8361	100
3.5	122	9816	71	144	0	58.1967	100
4	122	9720	70	142	0	57.377	100
4.5	122	9702	70	142	0	57.377	100
5	122	9685	70	142	0	57.377	100
5.5	122	9671	70	142	0	57.377	100
6	122	9669	70	142	0	57.377	100
6.5	122	9670	70	142	0	57.377	100
7	122	9667	70	142	0	57.377	100
7.5	122	9657	70	142	0	57.377	100

8.5	122	9655	70	142	0	57.377	100
9	122	9654	70	142	0	57.377	100
9.5	122	9654	70	142	0	57.377	100
10	122	9654	70	142	0	57.377	100
10.5	122	9653	70	142	0	57.377	100
11	122	9653	70	142	0	57.377	100
11.5	122	9653	70	142	0	57.377	100
12	122	9653	70	142	0	57.377	100
12.5	122	9653	70	142	0	57.377	100
13	122	9653	70	142	0	57.377	100
13.5	122	9653	70	142	0	57.377	100
14	122	9653	70	142	0	57.377	100
14.5	122	9653	70	142	0	57.377	100
15	122	9653	70	142	0	57.377	100
15.5	122	9653	70	142	0	57.377	100
16	122	9653	70	142	0	57.377	100
16.5	122	9653	70	142	0	57.377	100
17	122	9653	70	142	0	57.377	100
17.5	122	9653	70	142	0	57.377	100
18	122	9653	70	142	0	57.377	100
18.5	122	9653	70	142	0	57.377	100
19	122	9652	70	142	0	57.377	100
19.5	122	9652	70	142	0	57.377	100
20	122	9652	70	142	0	57.377	100
20.5	122	9651	70	142	0	57.377	100
21	122	9651	70	142	0	57.377	100
21.5	122	9651	70	142	0	57.377	100
22	122	9651	70	142	0	57.377	100
22.5	122	9651	70	142	0	57.377	100
23	122	9651	70	142	0	57.377	100
23.5	122	9650	70	142	0	57.377	100
24	122	9650	70	142	0	57.377	100
24.5	122	9649	70	142	0	57.377	100
25	122	9648	70	142	0	57.377	100
25.5	122	9647	70	142	0	57.377	100
26	122	9646	70	142	0	57.377	100
26.5	122	9648	70	142	0	57.377	100
27	122	9644	70	142	0	57.377	100
27.5	122	9637	70	142	0	57.377	100
28	122	9633	70	142	0	57.377	100
28.5	122	9628	70	142	0	57.377	100
29	122	9625	70	142	0	57.377	100
29.5	122	9620	70	142	0	57.377	100
30	122	9605	70	142	0	57.377	100
30.5	122	9591	70	142	0	57.377	100
31	122	9571	70	142	0	57.377	100
31.5	122	9557	70	142	0	57.377	100
32	122	9523	70	143	0	57.377	100
32.5	122	9494	70	143	0	57.377	100
33	122	9460	70	142	0	57.377	100
33.5	122	9376	70	137	0	57.377	100
34	122	9263	67	131	0	54.918	100
34.5	122	9110	64	124	0	52.459	100
35	122	8889	64	124	0	52.459	100
35.5	122	8642	64	122	0	52.459	100
36	122	8337	63	117	0	51.6393	100
36.5	122	7917	59	111	0	48.3607	100
37	122	7399	56	109	0	45.9016	100
37.5	122	6810	52	95	0	42.623	100
38	122	6097	52	85	0	42.623	100
38.5	122	5368	48	74	0	39.3443	100
39	122	4602	42	67	0	34.4262	100

39.5	122	3868	40	63	0	32.7869	100
40	122	3204	36	51	0	29.5082	100
40.5	122	2608	29	39	0	23.7705	100
41	122	2106	25	34	0	20.4918	100
41.5	122	1641	21	30	0	17.2131	100
42	122	1285	17	24	0	13.9344	100
42.5	122	994	13	19	0	10.6557	100
43	122	727	12	17	0	9.83607	100
43.5	122	551	8	12	0	6.55738	100
44	122	429	6	10	0	4.91803	100
44.5	122	327	2	4	0	1.63934	100
45	122	256	2	4	0	1.63934	100
45.5	122	187	2	4	0	1.63934	100
46	122	144	2	4	0	1.63934	100

Table A.7 Prediction result of "IGLV/J" for Group 1-22

6. All the 5 kind of genes, at the distance of 600bp

Threshold	reference	j	TP	HitTP	FP	Se	PPV
0	537	8175	346	614	2034	64.432	14.5378
0.5	537	10075	415	676	2397	77.2812	14.7582
1	537	9939	398	593	2398	74.1155	14.2346
1.5	537	8464	334	493	2088	62.1974	13.7903
2	537	7106	281	408	1773	52.3277	13.6806
2.5	537	6443	252	378	1629	46.9274	13.3971
3	537	6200	242	365	1583	45.0652	13.2603
3.5	537	6116	237	360	1558	44.1341	13.2033
4	537	6058	235	358	1552	43.7616	13.1505
4.5	537	6052	234	357	1553	43.5754	13.0946
5	537	6039	234	357	1552	43.5754	13.1019
5.5	537	6031	234	357	1551	43.5754	13.1092
6	537	6030	234	357	1552	43.5754	13.1019
6.5	537	6027	234	357	1550	43.5754	13.1166
7	537	6026	234	357	1549	43.5754	13.1239
7.5	537	6021	234	357	1548	43.5754	13.1313
8	537	6019	234	357	1548	43.5754	13.1313
8.5	537	6019	234	357	1548	43.5754	13.1313
9	537	6018	234	357	1548	43.5754	13.1313
9.5	537	6019	234	357	1548	43.5754	13.1313
10	537	6020	234	357	1548	43.5754	13.1313
10.5	537	6019	234	357	1548	43.5754	13.1313
11	537	6019	234	357	1548	43.5754	13.1313
11.5	537	6019	234	357	1548	43.5754	13.1313
12	537	6019	234	357	1548	43.5754	13.1313
12.5	537	6019	234	357	1548	43.5754	13.1313
13	537	6019	234	357	1548	43.5754	13.1313
13.5	537	6019	234	357	1548	43.5754	13.1313
14	537	6019	234	357	1548	43.5754	13.1313
14.5	537	6019	234	357	1548	43.5754	13.1313
15	537	6019	234	357	1548	43.5754	13.1313
15.5	537	6019	234	357	1548	43.5754	13.1313
16	537	6019	234	357	1548	43.5754	13.1313
16.5	537	6019	234	357	1548	43.5754	13.1313
17	537	6019	234	357	1548	43.5754	13.1313
17.5	537	6019	234	357	1548	43.5754	13.1313
18	537	6019	234	357	1548	43.5754	13.1313
18.5	537	6019	234	357	1548	43.5754	13.1313
19	537	6018	234	357	1548	43.5754	13.1313
19.5	537	6018	234	357	1548	43.5754	13.1313
20	537	6019	234	358	1549	43.5754	13.1239

20.5	537	6019	234	357	1549	43.5754	13.1239
21	537	6019	234	357	1549	43.5754	13.1239
21.5	537	6019	234	357	1549	43.5754	13.1239
22	537	6019	234	357	1549	43.5754	13.1239
22.5	537	6019	234	357	1549	43.5754	13.1239
23	537	6019	234	357	1549	43.5754	13.1239
23.5	537	6020	234	357	1550	43.5754	13.1166
24	537	6020	234	357	1551	43.5754	13.1092
24.5	537	6019	234	357	1550	43.5754	13.1166
25	537	6018	235	358	1550	43.7616	13.1653
25.5	537	6017	234	357	1552	43.5754	13.1019
26	537	6017	234	358	1552	43.5754	13.1019
26.5	537	6017	234	357	1551	43.5754	13.1092
27	537	6017	234	357	1550	43.5754	13.1166
27.5	537	6015	234	356	1548	43.5754	13.1313
28	537	6011	234	356	1548	43.5754	13.1313
28.5	537	6009	234	356	1547	43.5754	13.1387
29	537	6006	234	356	1545	43.5754	13.1535
29.5	537	6005	235	358	1544	43.7616	13.2097
30	537	5996	238	361	1543	44.3203	13.3633
30.5	537	5988	236	360	1544	43.9479	13.2584
31	537	5979	237	361	1543	44.1341	13.3146
31.5	537	5967	236	358	1546	43.9479	13.2435
32	537	5946	237	360	1533	44.1341	13.3898
32.5	537	5932	235	358	1532	43.7616	13.2994
33	537	5911	235	358	1524	43.7616	13.3599
33.5	537	5886	236	361	1522	43.9479	13.4243
34	537	5853	234	354	1511	43.5754	13.4097
34.5	537	5800	230	346	1497	42.8305	13.3179
35	537	5682	224	341	1471	41.7132	13.2153
35.5	537	5557	223	335	1441	41.527	13.4014
36	537	5391	217	325	1397	40.4097	13.4449
36.5	537	5198	207	307	1354	38.5475	13.2607
37	537	4910	203	295	1267	37.8026	13.8095
37.5	537	4580	193	274	1183	35.9404	14.0262
38	537	4204	186	257	1078	34.6369	14.7152
38.5	537	3789	172	227	978	32.0298	14.9565
39	537	3324	152	196	858	28.3054	15.0495
39.5	537	2836	137	174	728	25.5121	15.8382
40	537	2368	112	135	594	20.8566	15.864
40.5	537	1999	95	110	505	17.6909	15.8333
41	537	1636	81	93	399	15.0838	16.875
41.5	537	1294	67	76	313	12.4767	17.6316
42	537	1033	51	58	248	9.49721	17.0569
42.5	537	809	41	47	198	7.63501	17.1548
43	537	609	37	41	153	6.89013	19.4737
43.5	537	457	28	32	118	5.21415	19.1781
44	537	364	26	30	95	4.84171	21.4876
44.5	537	276	18	22	69	3.35196	20.6897
45	537	210	14	17	49	2.60708	22.2222
45.5	537	154	10	13	35	1.8622	22.2222
46	537	120	8	10	27	1.48976	22.8571

Table A.8 Prediction result of "the 5 kind of genes" for Group 1-22

Group 1-16

1. Coding Gene, at the distance of 500bp

Threshold	reference	j	TP	HitTP	FP	Se	PPV
0	201	3428	61	91	808	30.348259	7.019563
0.5	201	4853	81	110	1117	40.298508	6.761269
1	201	5778	99	143	1351	49.253731	6.827586
1.5	201	6148	92	144	1436	45.771145	6.020942
2	201	6287	92	146	1466	45.771145	5.905006
2.5	201	6323	91	150	1462	45.273632	5.859626
3	201	6360	93	151	1470	46.268658	5.950096
3.5	201	6377	93	152	1474	46.268658	5.934907
4	201	6377	93	152	1475	46.268658	5.931122
4.5	201	6378	93	152	1475	46.268658	5.931122
5	201	6380	93	152	1475	46.268658	5.931122
5.5	201	6381	93	152	1475	46.268658	5.931122
6	201	6381	93	152	1475	46.268658	5.931122
6.5	201	6381	93	152	1475	46.268658	5.931122
7	201	6381	93	152	1475	46.268658	5.931122
7.5	201	6381	93	152	1475	46.268658	5.931122
8	201	6381	93	152	1475	46.268658	5.931122
8.5	201	6381	93	152	1475	46.268658	5.931122
9	201	6381	93	152	1475	46.268658	5.931122
9.5	201	6381	93	152	1475	46.268658	5.931122
10	201	6381	93	152	1475	46.268658	5.931122
10.5	201	6381	93	152	1475	46.268658	5.931122
11	201	6381	93	152	1475	46.268658	5.931122
11.5	201	6381	93	152	1475	46.268658	5.931122
12	201	6381	93	152	1475	46.268658	5.931122
12.5	201	6381	93	152	1475	46.268658	5.931122
13	201	6381	93	152	1475	46.268658	5.931122
13.5	201	6381	93	152	1475	46.268658	5.931122
14	201	6381	93	152	1475	46.268658	5.931122
14.5	201	6381	93	152	1475	46.268658	5.931122
15	201	6381	93	152	1475	46.268658	5.931122
15.5	201	6381	93	152	1475	46.268658	5.931122
16	201	6381	93	152	1475	46.268658	5.931122
16.5	201	6381	93	152	1475	46.268658	5.931122
17	201	6381	93	152	1475	46.268658	5.931122
17.5	201	6381	93	152	1475	46.268658	5.931122
18	201	6381	93	152	1475	46.268658	5.931122
18.5	201	6381	93	152	1475	46.268658	5.931122
19	201	6380	93	152	1475	46.268658	5.931122
19.5	201	6380	93	152	1475	46.268658	5.931122
20	201	6380	93	152	1475	46.268658	5.931122
20.5	201	6380	93	152	1475	46.268658	5.931122
21	201	6380	93	152	1475	46.268658	5.931122
21.5	201	6380	93	152	1475	46.268658	5.931122
22	201	6381	93	152	1475	46.268658	5.931122
22.5	201	6381	93	152	1475	46.268658	5.931122
23	201	6381	93	152	1475	46.268658	5.931122
23.5	201	6382	93	152	1476	46.268658	5.927342
24	201	6383	93	152	1476	46.268658	5.927342
24.5	201	6382	93	152	1476	46.268658	5.927342
25	201	6379	93	152	1476	46.268658	5.927342
25.5	201	6378	93	152	1476	46.268658	5.927342
26	201	6378	93	153	1476	46.268658	5.927342
26.5	201	6378	93	152	1477	46.268658	5.923567
27	201	6377	93	152	1475	46.268658	5.931122
27.5	201	6374	93	151	1474	46.268658	5.934907
28	201	6368	93	151	1473	46.268658	5.938697
28.5	201	6367	93	151	1473	46.268658	5.938697
29	201	6363	93	151	1472	46.268658	5.942492
29.5	201	6361	93	151	1471	46.268658	5.946291
30	201	6351	93	151	1471	46.268658	5.946291

30.5	201	6344	93	152	1471	46.268658	5.946291
31	201	6336	93	152	1473	46.268658	5.938697
31.5	201	6322	93	150	1472	46.268658	5.942492
32	201	6303	92	148	1467	45.771145	5.901219
32.5	201	6287	92	148	1463	45.771145	5.916399
33	201	6267	92	149	1458	45.771145	5.935484
33.5	201	6228	92	148	1451	45.771145	5.962411
34	201	6190	92	148	1442	45.771145	5.997393
34.5	201	6125	92	145	1428	45.771145	6.052631
35	201	6001	91	142	1409	45.273632	6.066667
35.5	201	5878	91	143	1381	45.273632	6.182065
36	201	5683	86	138	1338	42.786068	6.039326
36.5	201	5454	83	130	1291	41.293533	6.040757
37	201	5124	83	125	1210	41.293533	6.41918
37.5	201	4774	78	114	1131	38.805969	6.451613
38	201	4366	73	106	1025	36.318409	6.648452
38.5	201	3916	66	93	939	32.835819	6.567164
39	201	3424	56	74	820	27.860697	6.392694
39.5	201	2913	50	63	695	24.875622	6.71141
40	201	2426	41	51	565	20.39801	6.765676
40.5	201	2043	36	43	471	17.910448	7.100592
41	201	1662	29	35	370	14.42786	7.26817
41.5	201	1317	24	29	291	11.940298	7.619048
42	201	1051	20	25	233	9.950249	7.905138
42.5	201	823	16	21	186	7.960199	7.920792
43	201	619	13	17	141	6.467662	8.441559
43.5	201	466	11	15	110	5.472637	9.090909
44	201	371	11	15	88	5.472637	11.111111
44.5	201	281	10	14	63	4.975124	13.69863
45	201	215	8	11	44	3.980099	15.384615
45.5	201	156	5	8	30	2.487562	14.285714
46	201	122	5	7	23	2.487562	17.857143

Table A. 9 Prediction result of "Coding genes" for Group 1-16

2. Non-coding Genes, at the distance of 400bp

Threshold	reference	j	TP	HitTP	FP	Se	PPV
0	14	4254	4	5	1	28.571428	80
0.5	14	5627	3	9	3	21.428572	50
1	14	6503	3	7	3	21.428572	50
1.5	14	6707	3	7	3	21.428572	50
2	14	6776	3	6	3	21.428572	50
2.5	14	6766	3	7	3	21.428572	50
3	14	6786	3	6	3	21.428572	50
3.5	14	6791	3	6	3	21.428572	50
4	14	6786	3	6	3	21.428572	50
4.5	14	6786	3	6	3	21.428572	50
5	14	6789	3	6	3	21.428572	50
5.5	14	6789	3	6	3	21.428572	50
6	14	6789	3	6	3	21.428572	50
6.5	14	6789	3	6	3	21.428572	50
7	14	6789	3	6	3	21.428572	50
7.5	14	6789	3	6	3	21.428572	50
8	14	6789	3	6	3	21.428572	50
8.5	14	6789	3	6	3	21.428572	50
9	14	6789	3	6	3	21.428572	50
9.5	14	6789	3	6	3	21.428572	50
10	14	6789	3	6	3	21.428572	50
10.5	14	6789	3	6	3	21.428572	50

11	14	6789	3	6	3	21.428572	50
11.5	14	6789	3	6	3	21.428572	50
12	14	6789	3	6	3	21.428572	50
12.5	14	6789	3	6	3	21.428572	50
13	14	6789	3	6	3	21.428572	50
13.5	14	6789	3	6	3	21.428572	50
14	14	6789	3	6	3	21.428572	50
14.5	14	6789	3	6	3	21.428572	50
15	14	6789	3	6	3	21.428572	50
15.5	14	6789	3	6	3	21.428572	50
16	14	6789	3	6	3	21.428572	50
16.5	14	6789	3	6	3	21.428572	50
17	14	6789	3	6	3	21.428572	50
17.5	14	6789	3	6	3	21.428572	50
18	14	6789	3	6	3	21.428572	50
18.5	14	6789	3	6	3	21.428572	50
19	14	6788	3	6	3	21.428572	50
19.5	14	6788	3	6	3	21.428572	50
20	14	6788	3	6	3	21.428572	50
20.5	14	6787	3	6	3	21.428572	50
21	14	6787	3	6	3	21.428572	50
21.5	14	6787	3	6	3	21.428572	50
22	14	6788	3	6	3	21.428572	50
22.5	14	6788	3	6	3	21.428572	50
23	14	6787	3	6	3	21.428572	50
23.5	14	6787	3	6	3	21.428572	50
24	14	6788	3	6	3	21.428572	50
24.5	14	6787	3	6	3	21.428572	50
25	14	6784	3	6	3	21.428572	50
25.5	14	6783	3	6	3	21.428572	50
26	14	6784	3	6	3	21.428572	50
26.5	14	6784	3	6	3	21.428572	50
27	14	6784	3	6	3	21.428572	50
27.5	14	6781	3	6	3	21.428572	50
28	14	6775	3	6	3	21.428572	50
28.5	14	6773	3	6	3	21.428572	50
29	14	6769	3	6	3	21.428572	50
29.5	14	6769	3	6	3	21.428572	50
30	14	6762	3	6	3	21.428572	50
30.5	14	6752	3	6	3	21.428572	50
31	14	6745	3	6	3	21.428572	50
31.5	14	6726	3	6	3	21.428572	50
32	14	6703	3	6	3	21.428572	50
32.5	14	6683	3	6	3	21.428572	50
33	14	6665	2	5	3	14.285714	40
33.5	14	6615	2	5	3	14.285714	40
34	14	6570	2	5	3	14.285714	40
34.5	14	6504	2	5	3	14.285714	40
35	14	6367	2	5	3	14.285714	40
35.5	14	6216	2	5	3	14.285714	40
36	14	6002	2	5	3	14.285714	40
36.5	14	5751	2	6	3	14.285714	40
37	14	5392	2	5	3	14.285714	40
37.5	14	5022	2	5	2	14.285714	50
38	14	4553	2	5	2	14.285714	50
38.5	14	4077	2	5	1	14.285714	66.666664
39	14	3569	2	5	1	14.285714	66.666664
39.5	14	3019	2	5	0	14.285714	100
40	14	2510	2	3	0	14.285714	100
40.5	14	2107	2	2	0	14.285714	100
41	14	1715	2	2	0	14.285714	100
41.5	14	1347	2	2	0	14.285714	100

42	14	1083	0	0	0	0	-
42.5	14	841	0	0	0	0	-
43	14	633	0	0	0	0	-
43.5	14	479	0	0	0	0	-
44	14	382	0	0	0	0	-
44.5	14	289	0	0	0	0	-
45	14	222	0	0	0	0	-
45.5	14	162	0	0	0	0	-
46	14	128	0	0	0	0	-

Table A. 10 Prediction result of "Non- coding genes" for Group 1-16

3. Pseudo genes, at the distance of 1000bp

Threshold	reference	j	TP	HitTP	FP	Se	PPV
0	122	3428	31	54	31	25.409836	50
0.5	122	4853	45	68	32	36.885246	58.441559
1	122	5778	40	62	34	32.786884	54.054054
1.5	122	6148	38	59	43	31.147541	46.913582
2	122	6287	36	63	45	29.508196	44.444443
2.5	122	6323	38	63	47	31.147541	44.705883
3	122	6360	38	63	46	31.147541	45.238094
3.5	122	6377	38	63	46	31.147541	45.238094
4	122	6377	38	63	46	31.147541	45.238094
4.5	122	6378	38	63	46	31.147541	45.238094
5	122	6380	38	63	46	31.147541	45.238094
5.5	122	6381	38	63	46	31.147541	45.238094
6	122	6381	38	63	46	31.147541	45.238094
6.5	122	6381	38	63	46	31.147541	45.238094
7	122	6381	38	63	46	31.147541	45.238094
7.5	122	6381	38	63	46	31.147541	45.238094
8	122	6381	38	63	46	31.147541	45.238094
8.5	122	6381	38	63	46	31.147541	45.238094
9	122	6381	38	63	46	31.147541	45.238094
9.5	122	6381	38	63	46	31.147541	45.238094
10	122	6381	38	63	46	31.147541	45.238094
10.5	122	6381	38	63	46	31.147541	45.238094
11	122	6381	38	63	46	31.147541	45.238094
11.5	122	6381	38	63	46	31.147541	45.238094
12	122	6381	38	63	46	31.147541	45.238094
12.5	122	6381	38	63	46	31.147541	45.238094
13	122	6381	38	63	46	31.147541	45.238094
13.5	122	6381	38	63	46	31.147541	45.238094
14	122	6381	38	63	46	31.147541	45.238094
14.5	122	6381	38	63	46	31.147541	45.238094
15	122	6381	38	63	46	31.147541	45.238094
15.5	122	6381	38	63	46	31.147541	45.238094
16	122	6381	38	63	46	31.147541	45.238094
16.5	122	6381	38	63	46	31.147541	45.238094

17	122	6381	38	63	46	31.147541	45.238094
17.5	122	6381	38	63	46	31.147541	45.238094
18	122	6381	38	63	46	31.147541	45.238094
18.5	122	6381	38	63	46	31.147541	45.238094
19	122	6380	38	63	46	31.147541	45.238094
19.5	122	6380	38	63	46	31.147541	45.238094
20	122	6380	38	63	46	31.147541	45.238094
20.5	122	6380	38	63	46	31.147541	45.238094
21	122	6380	38	63	46	31.147541	45.238094
21.5	122	6380	38	63	46	31.147541	45.238094
22	122	6381	38	63	46	31.147541	45.238094
22.5	122	6381	38	63	46	31.147541	45.238094
23	122	6381	38	63	46	31.147541	45.238094
23.5	122	6382	38	63	46	31.147541	45.238094
24	122	6383	38	63	46	31.147541	45.238094
24.5	122	6382	38	63	46	31.147541	45.238094
25	122	6379	38	63	46	31.147541	45.238094
25.5	122	6378	38	63	46	31.147541	45.238094
26	122	6378	38	63	46	31.147541	45.238094
26.5	122	6378	38	63	46	31.147541	45.238094
27	122	6377	38	63	46	31.147541	45.238094
27.5	122	6374	38	63	46	31.147541	45.238094
28	122	6368	38	63	45	31.147541	45.783131
28.5	122	6367	38	63	45	31.147541	45.783131
29	122	6363	38	63	45	31.147541	45.783131
29.5	122	6361	39	64	45	31.967213	46.42857
30	122	6351	39	64	45	31.967213	46.42857
30.5	122	6344	39	64	45	31.967213	46.42857
31	122	6336	39	64	45	31.967213	46.42857
31.5	122	6322	39	64	45	31.967213	46.42857
32	122	6303	39	64	44	31.967213	46.987953
32.5	122	6287	39	64	44	31.967213	46.987953
33	122	6267	39	64	42	31.967213	48.148148
33.5	122	6228	39	65	42	31.967213	48.148148
34	122	6190	39	65	41	31.967213	48.75
34.5	122	6125	37	62	41	30.327869	47.435898
35	122	6001	37	62	41	30.327869	47.435898
35.5	122	5878	38	60	39	31.147541	49.350651
36	122	5683	38	59	41	31.147541	48.101265
36.5	122	5454	38	60	35	31.147541	52.054794
37	122	5124	37	59	31	30.327869	54.411766
37.5	122	4774	36	55	30	29.508196	54.545456
38	122	4366	32	47	29	26.229507	52.459015
38.5	122	3916	31	42	22	25.409836	58.490566
39	122	3424	28	39	21	22.950819	57.142857
39.5	122	2913	24	32	19	19.672131	55.813953
40	122	2426	17	23	16	13.934426	51.515152

40.5	122	2043	14	18	16	11.47541	46.666668
41	122	1662	13	16	13	10.655738	50
41.5	122	1317	10	12	10	8.196721	50
42	122	1051	6	7	6	4.918033	50
42.5	122	823	5	6	5	4.098361	50
43	122	619	5	6	4	4.098361	55.555557
43.5	122	466	3	3	4	2.459016	42.857143
44	122	371	3	3	4	2.459016	42.857143
44.5	122	281	3	3	4	2.459016	42.857143
45	122	215	3	3	4	2.459016	42.857143
45.5	122	156	2	2	4	1.639344	33.333332
46	122	122	1	1	4	0.819672	20

Table A. 11 Prediction result of "Pseudo genes" for Group 1-16

4. partial genes, at the distance of 1500bp

Threshold	reference	i	TP	HitTP	FP	Se	PPV
0	78	656	2	2	24	2.564103	7.692307
0.5	78	1519	10	10	37	12.820513	21.276596
1	78	2247	16	17	50	20.512821	24.242424
1.5	78	2840	23	26	63	29.487179	26.744186
2	78	3159	24	26	68	30.76923	26.086956
2.5	78	3293	24	26	66	30.76923	26.666666
3	78	3362	24	26	67	30.76923	26.373627
3.5	78	3387	26	28	67	33.333332	27.956989
4	78	3404	25	27	69	32.051281	26.595745
4.5	78	3411	25	27	69	32.051281	26.595745
5	78	3415	25	27	69	32.051281	26.595745
5.5	78	3417	25	27	69	32.051281	26.595745
6	78	3417	25	27	69	32.051281	26.595745
6.5	78	3421	25	27	70	32.051281	26.31579
7	78	3422	25	27	70	32.051281	26.31579
7.5	78	3422	25	27	70	32.051281	26.31579
8	78	3422	25	27	70	32.051281	26.31579
8.5	78	3422	25	27	70	32.051281	26.31579
9	78	3422	25	27	70	32.051281	26.31579
9.5	78	3422	25	27	70	32.051281	26.31579
10	78	3422	25	27	70	32.051281	26.31579
10.5	78	3422	25	27	70	32.051281	26.31579
11	78	3422	25	27	70	32.051281	26.31579
11.5	78	3422	25	27	70	32.051281	26.31579
12	78	3422	25	27	70	32.051281	26.31579
12.5	78	3422	25	27	70	32.051281	26.31579
13	78	3422	25	27	70	32.051281	26.31579
13.5	78	3422	25	27	70	32.051281	26.31579
14	78	3423	25	27	70	32.051281	26.31579
14.5	78	3423	25	27	70	32.051281	26.31579
15	78	3423	25	27	70	32.051281	26.31579
15.5	78	3423	25	27	70	32.051281	26.31579
16	78	3423	25	27	70	32.051281	26.31579
16.5	78	3423	25	27	70	32.051281	26.31579
17	78	3423	25	27	70	32.051281	26.31579
17.5	78	3423	25	27	70	32.051281	26.31579
18	78	3423	25	27	70	32.051281	26.31579
18.5	78	3423	25	27	70	32.051281	26.31579
19	78	3422	25	27	70	32.051281	26.31579
19.5	78	3422	25	27	70	32.051281	26.31579

20	78	3422	25	27	70	32.051281	26.31579
20.5	78	3423	25	27	70	32.051281	26.31579
21	78	3423	25	27	70	32.051281	26.31579
21.5	78	3422	25	27	70	32.051281	26.31579
22	78	3423	25	27	70	32.051281	26.31579
22.5	78	3422	25	27	70	32.051281	26.31579
23	78	3424	25	27	70	32.051281	26.31579
23.5	78	3423	25	27	70	32.051281	26.31579
24	78	3423	25	27	70	32.051281	26.31579
24.5	78	3424	25	27	71	32.051281	26.041666
25	78	3427	25	27	71	32.051281	26.041666
25.5	78	3427	25	27	70	32.051281	26.31579
26	78	3430	25	27	70	32.051281	26.31579
26.5	78	3428	25	27	70	32.051281	26.31579
27	78	3426	25	27	70	32.051281	26.31579
27.5	78	3426	25	27	70	32.051281	26.31579
28	78	3425	25	27	70	32.051281	26.31579
28.5	78	3427	25	27	70	32.051281	26.31579
29	78	3423	25	27	71	32.051281	26.041666
29.5	78	3427	25	27	70	32.051281	26.31579
30	78	3424	25	27	70	32.051281	26.31579
30.5	78	3421	25	27	69	32.051281	26.595745
31	78	3425	25	27	69	32.051281	26.595745
31.5	78	3421	25	27	70	32.051281	26.31579
32	78	3425	25	27	69	32.051281	26.595745
32.5	78	3431	25	28	67	32.051281	27.173914
33	78	3420	25	28	69	32.051281	26.595745
33.5	78	3431	27	31	68	34.615383	28.421053
34	78	3450	25	28	68	32.051281	26.88172
34.5	78	3449	25	29	64	32.051281	28.089888
35	78	3453	25	30	66	32.051281	27.472527
35.5	78	3440	23	26	63	29.487179	26.744186
36	78	3424	22	24	66	28.205128	25
36.5	78	3381	22	23	62	28.205128	26.190475
37	78	3299	25	27	58	32.051281	30.120481
37.5	78	3215	26	27	56	33.333332	31.707317
38	78	3045	27	30	50	34.615383	35.064934
38.5	78	2852	26	27	45	33.333332	36.619717
39	78	2580	24	24	39	30.76923	38.095238
39.5	78	2274	22	22	35	28.205128	38.596493
40	78	1954	16	16	31	20.512821	34.042553
40.5	78	1698	14	14	30	17.948717	31.818182
41	78	1414	13	13	25	16.666666	34.210526
41.5	78	1155	10	11	20	12.820513	33.333332
42	78	939	9	9	16	11.538462	36
42.5	78	750	7	7	13	8.974359	35
43	78	577	7	7	10	8.974359	41.176472
43.5	78	435	6	6	5	7.692307	54.545456
44	78	347	6	6	5	7.692307	54.545456
44.5	78	265	3	3	4	3.846154	42.857143
45	78	203	1	1	3	1.282051	25
45.5	78	148	1	1	2	1.282051	33.333332
46	78	115	0	0	1	0	0

Table A. 12 Prediction result of "Partial genes" for Group 1-16

5. IGLV/J, at the distance of 100bp

Threshold	reference	j	TP	HitTP	FP	Se	PPV
0	122	10267	71	197	0	58.19672	100
0.5	122	10916	77	199	0	63.114754	100

1	122	11367	72	160	0	59.016392	100
1.5	122	10631	71	160	0	58.19672	100
2	122	10165	71	149	0	58.19672	100
2.5	122	9878	70	144	0	57.377048	100
3	122	9752	69	142	0	56.557377	100
3.5	122	9673	71	144	0	58.19672	100
4	122	9650	70	142	0	57.377048	100
4.5	122	9650	70	142	0	57.377048	100
5	122	9650	70	142	0	57.377048	100
5.5	122	9649	70	142	0	57.377048	100
6	122	9649	70	142	0	57.377048	100
6.5	122	9649	70	142	0	57.377048	100
7	122	9649	70	142	0	57.377048	100
7.5	122	9649	70	142	0	57.377048	100
8	122	9649	70	142	0	57.377048	100
8.5	122	9649	70	142	0	57.377048	100
9	122	9649	70	142	0	57.377048	100
9.5	122	9649	70	142	0	57.377048	100
10	122	9649	70	142	0	57.377048	100
10.5	122	9649	70	142	0	57.377048	100
11	122	9649	70	142	0	57.377048	100
11.5	122	9649	70	142	0	57.377048	100
12	122	9649	70	142	0	57.377048	100
12.5	122	9649	70	142	0	57.377048	100
13	122	9649	70	142	0	57.377048	100
13.5	122	9649	70	142	0	57.377048	100
14	122	9649	70	142	0	57.377048	100
14.5	122	9649	70	142	0	57.377048	100
15	122	9649	70	142	0	57.377048	100
15.5	122	9649	70	142	0	57.377048	100
16	122	9649	70	142	0	57.377048	100
16.5	122	9649	70	142	0	57.377048	100
17	122	9649	70	142	0	57.377048	100
17.5	122	9649	70	142	0	57.377048	100
18	122	9649	70	142	0	57.377048	100
18.5	122	9649	70	142	0	57.377048	100
19	122	9648	70	142	0	57.377048	100
19.5	122	9648	70	142	0	57.377048	100
20	122	9648	70	142	0	57.377048	100
20.5	122	9647	70	142	0	57.377048	100
21	122	9647	70	142	0	57.377048	100
21.5	122	9647	70	142	0	57.377048	100
22	122	9647	70	142	0	57.377048	100
22.5	122	9647	70	142	0	57.377048	100
23	122	9647	70	142	0	57.377048	100
23.5	122	9646	70	142	0	57.377048	100
24	122	9646	70	142	0	57.377048	100
24.5	122	9646	70	142	0	57.377048	100
25	122	9645	70	142	0	57.377048	100
25.5	122	9644	70	142	0	57.377048	100
26	122	9643	70	142	0	57.377048	100
26.5	122	9645	70	142	0	57.377048	100
27	122	9641	70	142	0	57.377048	100
27.5	122	9634	70	142	0	57.377048	100
28	122	9630	70	142	0	57.377048	100
28.5	122	9625	70	142	0	57.377048	100
29	122	9622	70	142	0	57.377048	100
29.5	122	9617	70	142	0	57.377048	100
30	122	9602	70	142	0	57.377048	100
30.5	122	9588	70	142	0	57.377048	100
31	122	9568	70	142	0	57.377048	100
31.5	122	9554	70	142	0	57.377048	100

32	122	9520	70	143	0	57.377048	100
32.5	122	9491	70	143	0	57.377048	100
33	122	9457	70	142	0	57.377048	100
33.5	122	9373	70	137	0	57.377048	100
34	122	9261	67	131	0	54.918034	100
34.5	122	9108	64	124	0	52.459015	100
35	122	8888	64	124	0	52.459015	100
35.5	122	8640	64	122	0	52.459015	100
36	122	8336	63	117	0	51.639343	100
36.5	122	7916	59	111	0	48.360657	100
37	122	7398	56	109	0	45.901638	100
37.5	122	6809	52	95	0	42.622952	100
38	122	6097	52	85	0	42.622952	100
38.5	122	5367	48	74	0	39.344261	100
39	122	4602	42	67	0	34.426231	100
39.5	122	3868	40	63	0	32.786884	100
40	122	3204	36	51	0	29.508196	100
40.5	122	2608	29	39	0	23.770493	100
41	122	2106	25	34	0	20.491804	100
41.5	122	1641	21	30	0	17.213116	100
42	122	1285	17	24	0	13.934426	100
42.5	122	994	13	19	0	10.655738	100
43	122	727	12	17	0	9.836065	100
43.5	122	551	8	12	0	6.557377	100
44	122	429	6	10	0	4.918033	100
44.5	122	327	2	4	0	1.639344	100
45	122	256	2	4	0	1.639344	100
45.5	122	187	2	4	0	1.639344	100
46	122	144	2	4	0	1.639344	100

Table A. 13 Prediction result of "IGLV/J" for Group 1-16

6. All the 5 kind of genes, at the distance of 300bp

Threshold	reference	j	TP	HitTP	FP	Se	PPV
0	537	5440	203	398	1414	37.802608	12.554112
0.5	537	6660	255	433	1740	47.486034	12.781955
1	537	7459	250	448	1974	46.554935	11.241007
1.5	537	7464	248	459	1967	46.182495	11.196388
2	537	7421	241	448	1960	44.878956	10.949569
2.5	537	7368	240	446	1939	44.692738	11.014227
3	537	7349	241	446	1928	44.878956	11.111111
3.5	537	7338	244	449	1924	45.437618	11.254613
4	537	7328	242	447	1921	45.065178	11.188165
4.5	537	7330	242	447	1921	45.065178	11.188165
5	537	7331	242	447	1921	45.065178	11.188165
5.5	537	7330	242	447	1920	45.065178	11.193339
6	537	7330	242	447	1920	45.065178	11.193339
6.5	537	7330	242	447	1920	45.065178	11.193339
7	537	7330	242	447	1920	45.065178	11.193339
7.5	537	7330	242	447	1920	45.065178	11.193339
8	537	7330	242	447	1920	45.065178	11.193339
8.5	537	7330	242	447	1920	45.065178	11.193339
9	537	7330	242	447	1920	45.065178	11.193339
9.5	537	7330	242	447	1920	45.065178	11.193339
10	537	7330	242	447	1920	45.065178	11.193339
10.5	537	7330	242	447	1920	45.065178	11.193339
11	537	7330	242	447	1920	45.065178	11.193339
11.5	537	7330	242	447	1920	45.065178	11.193339
12	537	7330	242	447	1920	45.065178	11.193339
12.5	537	7330	242	447	1920	45.065178	11.193339
13	537	7330	242	447	1920	45.065178	11.193339

13.5	537	7330	242	447	1920	45.065178	11.193339
14	537	7330	242	447	1920	45.065178	11.193339
14.5	537	7330	242	447	1920	45.065178	11.193339
15	537	7330	242	447	1920	45.065178	11.193339
15.5	537	7330	242	447	1920	45.065178	11.193339
16	537	7330	242	447	1920	45.065178	11.193339
16.5	537	7330	242	447	1920	45.065178	11.193339
17	537	7330	242	447	1920	45.065178	11.193339
17.5	537	7330	242	447	1920	45.065178	11.193339
18	537	7330	242	447	1920	45.065178	11.193339
18.5	537	7330	242	447	1920	45.065178	11.193339
19	537	7329	242	447	1920	45.065178	11.193339
19.5	537	7329	242	447	1920	45.065178	11.193339
20	537	7329	242	447	1920	45.065178	11.193339
20.5	537	7328	242	447	1919	45.065178	11.198519
21	537	7328	242	447	1919	45.065178	11.198519
21.5	537	7328	242	447	1919	45.065178	11.198519
22	537	7328	242	447	1919	45.065178	11.198519
22.5	537	7328	242	447	1919	45.065178	11.198519
23	537	7328	242	447	1919	45.065178	11.198519
23.5	537	7328	242	447	1919	45.065178	11.198519
24	537	7328	242	447	1919	45.065178	11.198519
24.5	537	7327	242	447	1919	45.065178	11.198519
25	537	7325	242	447	1919	45.065178	11.198519
25.5	537	7323	242	447	1919	45.065178	11.198519
26	537	7325	242	448	1919	45.065178	11.198519
26.5	537	7326	242	447	1921	45.065178	11.188165
27	537	7323	242	447	1918	45.065178	11.203704
27.5	537	7322	242	447	1918	45.065178	11.203704
28	537	7316	242	447	1915	45.065178	11.219286
28.5	537	7312	242	447	1915	45.065178	11.219286
29	537	7310	242	447	1914	45.065178	11.22449
29.5	537	7308	242	447	1914	45.065178	11.22449
30	537	7297	242	447	1911	45.065178	11.24013
30.5	537	7285	242	448	1909	45.065178	11.250581
31	537	7276	242	448	1909	45.065178	11.250581
31.5	537	7257	242	446	1910	45.065178	11.245353
32	537	7230	241	445	1897	44.878956	11.272217
32.5	537	7209	241	445	1893	44.878956	11.293345
33	537	7180	241	445	1887	44.878956	11.325188
33.5	537	7116	240	439	1869	44.692738	11.379801
34	537	7069	236	432	1869	43.947857	11.211401
34.5	537	6980	233	424	1846	43.389198	11.207312
35	537	6806	229	416	1804	42.644321	11.264142
35.5	537	6650	228	409	1768	42.458099	11.422846
36	537	6419	223	396	1704	41.527	11.572392
36.5	537	6152	215	380	1632	40.037243	11.640498
37	537	5746	209	352	1516	38.919926	12.115942
37.5	537	5322	197	331	1395	36.685287	12.374372
38	537	4816	189	299	1248	35.19553	13.152401
38.5	537	4300	174	258	1121	32.402233	13.436294
39	537	3752	154	227	983	28.677839	13.544415
39.5	537	3173	137	197	830	25.512104	14.167528
40	537	2636	112	155	679	20.856611	14.159292
40.5	537	2185	96	122	566	17.877094	14.501511
41	537	1768	81	103	435	15.083798	15.697675
41.5	537	1387	67	83	342	12.476723	16.381418
42	537	1107	51	64	269	9.497207	15.9375
42.5	537	863	41	51	214	7.635009	16.078432
43	537	648	37	45	166	6.890131	18.226601
43.5	537	491	28	35	129	5.214153	17.834394
44	537	388	26	33	104	4.841713	20

44.5	537	295	18	23	74	3.351955	19.565218
45	537	227	14	18	54	2.607076	20.588236
45.5	537	165	10	14	39	1.862197	20.408163
46	537	130	8	11	31	1.489758	20.512821

Table A. 14 Prediction result of "IGLV/J" for Group 1-16