國立交通大學

理學院網路學習學程

碩士論文

用計算方法識別哺乳類動物的基因轉錄啟始點

Computationally Identifying Core Promoter Regions of Genes in Mammalian Genomes

研究生：林在營

指導教授：黃憲達　教授

中華民國　九十五　年六月

用計算方法識別哺乳類動物的基因轉錄啟始點

Computationally Identifying Core Promoter Regions of Genes in Mammalian Genomes

研 究 生：林在營　　Student：Tzai-Ying Lin

指導教授：黃憲達　　Advisor：Hsien-Da Huang

國 立 交 通 大 學

理學院網路學習學程

碩 士 論 文

A Thesis

Submitted to Degree Program of E-Learning

College of Science

National Chiao Tung University

in partial Fulfillment of the Requirements

for the Degree of

Master

in

Degree Program of E-Learning

June 2006

Hsinchu, Taiwan, Republic of China

中華民國九十五年六月

用計算方法識別哺乳類動物的基因轉錄啟始點

學生：林在營　　指導教授：黃憲達　教授

國立交通大學理學院碩士在職專班

中文摘要

轉錄是ＲＮＡ從基因染色體的ＤＮＡ片段進行複製的過程，其受到啟動子區域的影響而作用。 核心啟動子（Core promoter）是在轉錄起始點（TSS）鄰近約100個基元的區域，轉錄啟始點是轉錄過程的起始點，而準確的定位出核心啟動子的區域是我們理解基因轉錄規則的第一步。

在這篇論文裡，我們提出一種基於ＤＮＡ穩定性及核苷酸分布及機器學習理論的計算方法用以鑑定哺乳類動物基因組的轉錄啟始點。 已知的轉錄啟始點資料是從DBTSS資料庫取得而哺乳類動物的基因組序列則是由NCBI第35版的資料庫裡取得。我們整合了支持向量機器（Support Vector Machine)來建立預測轉錄啟始點的模型。為了了解我們進行預測的方法的好壞，我們使用了交叉比對(k-fold cross-validation)的方式進行驗證。初步的結果顯示我們的預測方法的準確性達70%以上，而跟其他論文提出的方法進行比較，我們的系統的確較其他方法有較好的效能。

關鍵字： Transcriptional start sites, promoter, Support Vector Machine (SVM), DNA stability

Computationally Identifying Core Promoter Regions of Genes in Mammalian Genomes

student：Tsai-Ying Lin　　Advisors：Dr. Hsien-Da Huang

Institute of Science,
National Chiao Tung University

ABSTRACT

Gene transcription is an extremely important mechanism in the cell, which is regulated by transcription factors (TFs), binding mostly and specifically to the 5' end of genes, the so called promoter region. The core promoter is a region of about 100 base-pairs flanking the transcriptional start site (TSS), which serves as the recognition site for the basal transcription apparatus. To accurately determine the core promoter in gene upstream is the first step to decipher the regulation of gene transcription.

In the study, we incorporated Support Vector Machine (SVM) with three useful regulatory features such as statistically significant 6-mer patterns, nucleotide composition, and DNA stability to identify the transcriptional start sites in mammalian genomes. The experimentally verified transcriptional start sites were obtained from DBTSS, and the genomic sequences of the mammalian genomes were obtained from NCBI build 35. K-fold cross-validation was used to evaluate the prediction performance of the three regulatory features extracted for core promoters, and the preliminary results suggested that the prediction accuracy could be greater than 70%. By comparing to other previously developed approach, our method had better prediction performance than others.

Keywords: Transcriptional start sites, promoter, Support Vector Machine (SVM), DNA stability

誌　謝

時光匆匆，轉眼間我要從交大畢業了，在這二年的日子裡，我要感謝我的指導老師黃憲達博士對我的細心指導，讓我在生物資訊這個領域能從完全不懂進而學得很多有用的知識，也引導我讓我熟悉研究的過程與方法；另外我也要特別感謝博士班的李宗夷學長，不厭其煩的跟我討論及指導我正確的研究方向，讓我的研究能順利的進行而不致有所偏差。

實驗室的其他學長們，也謝謝你們對我的細心指導，實驗室的同學們，謝謝大家在這兩年內的互相幫忙及鼓勵，和大家一起 meeting 的日子，是我成長的動力，實驗室內的點點滴滴更是美好的回憶。

更要感謝專班提供這個進修的機會，讓我能進入交大，能在眾多專班教授的引導下，讓我們收獲良多，尤其是莊祚敏主任、黃大原教授、陳明璋教授…等人。

最後要感謝我的家人的鼓勵，太太美玉的體貼，還有我那配合的小寶貝，在我口試完第三天平安的出生了。

能夠順利完成碩士論文並取得碩士學位，是大家的指導、支持、與鼓勵，誠心的謝謝大家，將這份喜悅及成果與關心我的所有人一同分享。

國立交通大學理學院在職專班

發現生資實驗室 研究生 林在營

謹誌於交通大學　2006 年六月

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction

1.1 Background

The science was stemmed from exploring the universe's secret. The life is a part of the universe, so studying the mysterious biological phenomena clearly had been one of the goals of scientist's efforts. Since James Watson and Francis Crick derived out the structure model of DNA in 1953, the scientist untied the hereditary secret and regulatory mechanism of the gene progressively with this foundation.

1.1.1 Central Dogma

The majority of genes are expressed as the proteins they encode. As shown in the Figure 1.1, the central dogma of molecular biology is based on the principle that the flow of genetic information travels from DNA to RNA and finally to the translation of proteins. In the transcription step, DNA is transcribed to RNA. There are various types of RNA including tRNA (transfer RNA), rRNA (ribosomal RNA) and mRNA (messenger RNA) in the transcription step. The mRNA is the blueprint in the process of protein synthesis. The process of mRNA transform to protein called translation. The direction of transcription and translation is unidirectional, no reverse direction is detected. But one process of RNA transform to DNA called reverse transcription which occur in mRNA

reverse transcript to cDNA (complementary DNA) for RNA amplification.

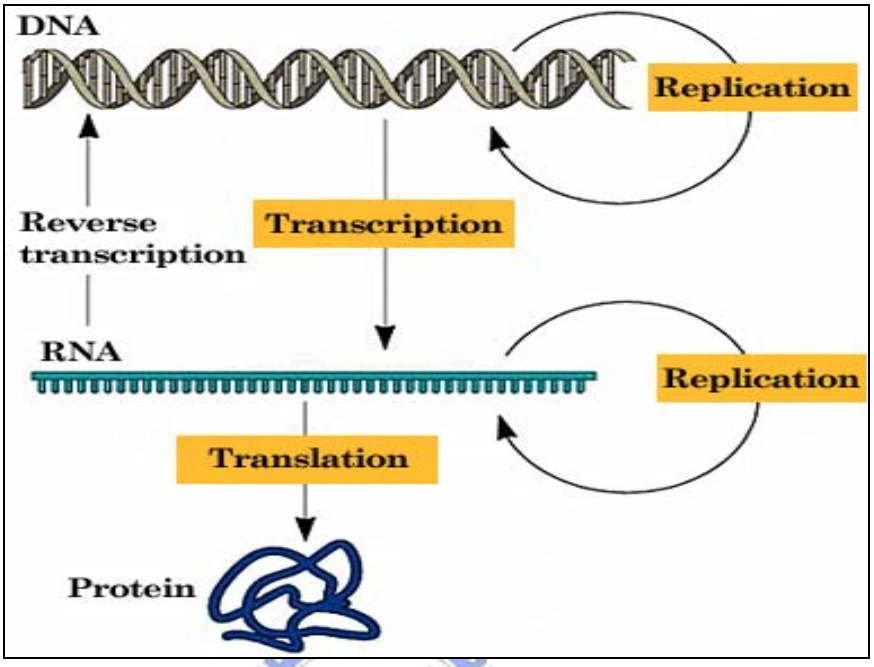


Figure 1.1 Central dogma of molecular biology.

(The figure is obtained from http://cats.med.uvm.edu/.../1_centraldogma_wisc_13.jpg)

1.1.2 Transcriptional Start sites (TSS)

Transcription, the process whereby RNA copies are made from sections of the DNA genome, is directed by promoter regions (Down and Hubbard 2002). The promoter is a DNA sequence which is usually located on the upstream of a gene transcriptional starting site (TSS). The core promoter, a region of about 100 base-pairs flanking the transcriptional start site (TSS), serves as the recognition site for the basal transcription apparatus (Ohler, Liao et al. 2002). Figure 1.2 shows a simplified gene structure and promoter region. When RNA polymerase II and some kinds of transcription factors bind onto the gene promoter region

together, the transcription of a gene will be activated and the messenger RNA (mRNA) will be transcribed from the DNA sequence. We have known that the promoter region is always located on the upstream of a gene and we will try to find out some hallmarks nearby the known transcriptional start sites. Thus, we can identify the transcriptional start sites on the unknown DNA sequences by using these hallmarks.

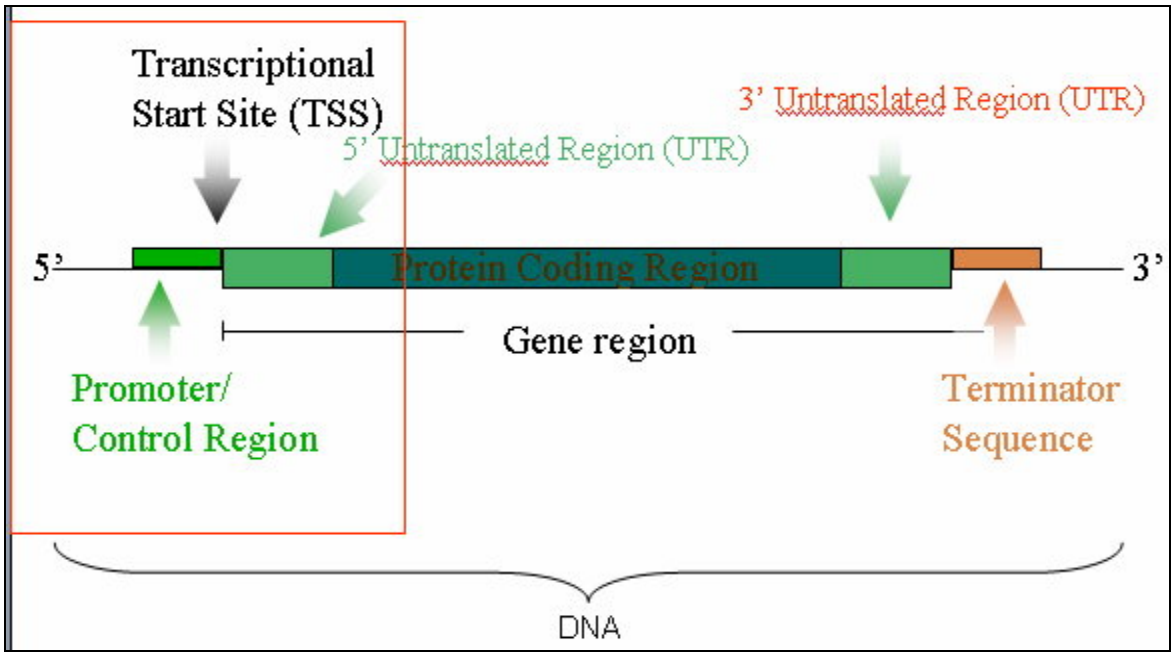


Figure 1.2 A simplified gene structure.

1.1.3 Transcriptional Regulation

Transcriptional regulation is one of the most important means of gene regulation. Uncovering transcriptional regulatory networks helps us to understanding the complex cellular process (Xing and van der Laan 2005). As

shown in the Figure 1.3, gene transcription is regulated by transcription factors (TFs) which are binding mostly and specifically to the 5' end of genes. The RNA polymerase II promoter is the critical region that regulates differential transcription of protein coding genes (Solovyev and Shahmuradov 2003), and is located near the transcription start site (TSS). A typical promoter region is believed to comprise short DNA sequences known as regulatory elements, which includes transcription factor binding sites (TFBSs) (Prakash and Tompa 2005).

About 10~15% of mammalian DNA re-associated very rapidly. This class includes tandem repeats. It includes Satellites (100 kb to over 1Mb), Ministatellites (1kb ~ 20kb), Microsatellites (Short Tandem repeats, 1~ 6 base in a region less than 150 base). Interspersed repeats are repeated DNA sequences located at dispersed regions in a genome. They are also known as mobile elements or transposable elements. In mammals, the most common mobile elements are LINEs (Long Interspersed Nuclear Elements) and SINEs (Short Interspersed nuclear Elements).

The structure of eukaryotic promoters is more complex than prokayotic promoters and they have several sequence motifs, for example TATA box, CCAAT box, GC box, and INR box (Kanhere and Bansal 2005). Therefore, some concepts are also used to analyze the promoter, including the presence of CpG islands close to the transcription start site, the presence of a specific

transcription factor binding site, statistical properties of proximal and core promoters rather than other genomic sequences, the orthologous gene promoters, and restricting the promoter region from using information from mRNA transcripts (Bajic, Tan et al. 2004).

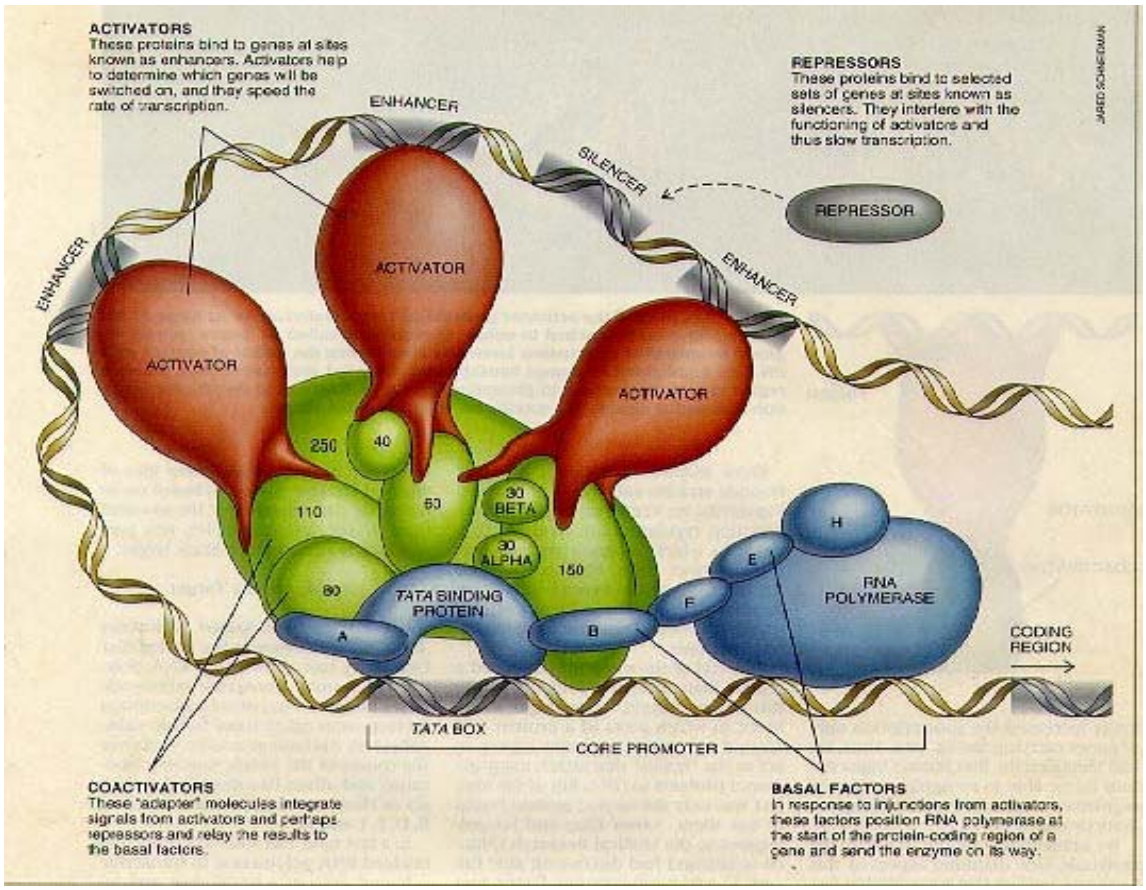


Figure 1.3 Transcriptional regulation.

(The figure is obtained from http://www.wellesley.edu/.../06eukaryotes.jpg)

1.2 Motivation

To accurately determine the core promoter in gene upstream is the first step to decipher the regulation of gene transcription. In recent years, powerful

computational techniques have increasingly been used to analyze annotated DNA sequences to uncover the secret of human genomes. Gene promoter prediction is important for guiding experimental biologists to find novel gene promoter region. However, the existing promoter prediction tools result in low sensitivity or low specificity. The low prediction accuracy stems from the low quality of regulatory features or training data source. Therefore, a computational method integrated with good regulatory features should be proposed.

1.3 Goal

Developing an efficient and effective system to identify gene transcriptional start sites is important in silico tools for guiding experimental biologists. However, the existing promoter prediction programs can only be used for minor reference because of them results in low sensitivity or low specificity. Therefore, the main purpose of this study is to incorporate the powerful computational method with useful regulatory features of core promoters for gene promoter identification that results in high prediction sensitivity and specificity. This study can help researchers identifying gene transcriptional promoter region more efficiently and exactly.

1.4 Challenges

Before extracting the useful regulatory features for core promoters, the

experimentally verified transcriptional start sites need to be obtained. In this thesis, the main source of the known transcriptional start sites were obtained from DBTSS and the genomic sequences of the mammalian genomes were obtained from NCBI build 35. All the promoter sequences obtained from DBTSS were stored into our databank for analyzing the core promoter region. However, the length of sequences obtained from DBTSS was 1201 bps (from -1000 to +200), which was too short to completely analyze the flanking region of TSS. We used the *Blast* program to align the promoter sequence of DBTSS to the genomic sequences of NCBI build 35, then the longer sequences of 6001 (from -3000 to +3000) were obtained.

Besides, the problem of identifying gene transcriptional start sites itself remains some difficult challenges. The most important one challenge is that no reliable dataset of experimentally verified transcriptional start sites can be used to analyze the specifically regulatory elements of promoter. Moreover, the existing promoter prediction tools result in low numbers of true positive predictions and high numbers of false positive predictions. The detailed description of the existing promoter prediction methods will be discussed in chapter 2.

The research scope of this thesis is concentrated on mammalian (human and mouse) genomes. The reason for using the mammalian genomes is that the sequences of gene transcriptional start sites obtained from DBTSS just had human and mouse. In addition, the reason for obtaining the sequences of gene

transcriptional start sites from DBTSS is its amount much more than other databases. Therefore, we constrict our scope to identify transcriptional start sites in mammalian genomes.

Chapter 2 Related Works

2.1 Experimental TSS data source

Two popular experimentally verified transcriptional start site databases, such as Eukaryotic Promoter Database (EPD) and Database of Transcriptional Start Site (DBTSS), were widely used to analyze the promoter region.

2.1.1 Eukaryotic Promoter Database (EPD)

The Eukaryotic Promoter Database (EPD) is an annotated non-redundant collection of eukaryotic POL II promoters, for which the transcription start site has been determined experimentally (Schmid, Praz et al. 2004). EPD is a collection of 4,810 eukaryotic POL II promoters. Tools for analysing sequence motifs around TSSs defined in EPD are provided by the signal search analysis server. EPD can be accessed at http://www.epd. isb-sib.ch.

2.1.2 Database of Transcriptional Start Sites (DBTSS)

Suzuki *et al* developed a novel method namely oligo-capping to collect the full-length cDNA libraries. The different between cloning method and oligo-capping method is shown in Fig. 2.1. The characteristics of oligo-capping method are extensive, high throughput, and high accuracy. DBTSS (Suzuki, Yamashita et al. 2004) was constructed in 2002 based on the full-length cDNA

libraries.

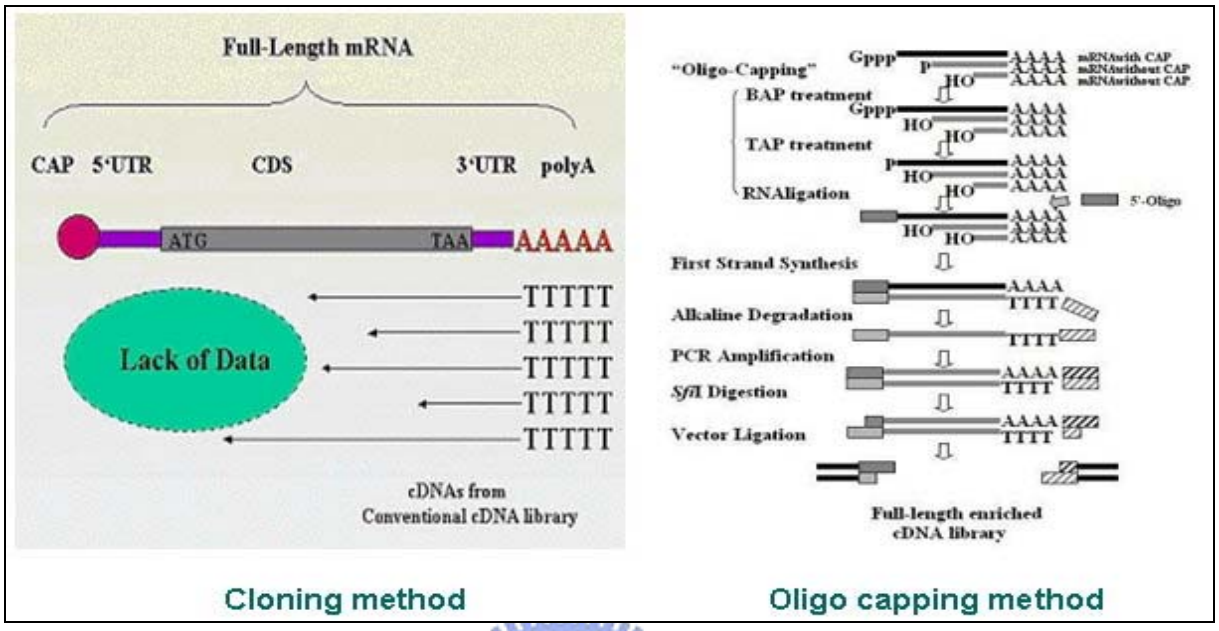


Figure 2.1 The comparison between the cloning method and the oligo-capping method.

(The figure is obtained from http://dbtss_old.hgc.jp/hg17/)

DBTSS is a collection of transcriptional start sites and adjacent promoters, which are experimentally determined by intensive analyses of full-length cDNAs. In order to extract biological insight from the compiled sequence information, search engines for putative transcription factor binding sites are implemented. Also, for molecular evolutionary studies of the transcriptional regulations, detailed sequence alignments of the promoters between human, mouse and other model organisms are provided. DBTSS is available on the web in Japan at http://dbtss.hgc.jp. The positional information of the TSSs, sequences of the promoters and related information can also be downloaded in flat file form from the download site. The current release of DBTSS (5.1) contains TSS

information of 15,262 and 14,162 genes determined by 1.4 and 0.4 million cDNAs in humans and mice respectively.

2.2 Related Works of Gene Promoter Prediction Tools

Gene promoter prediction is important *in silico* analysis for guiding experimental biologists. Although several gene promoter prediction tools had been developed, wet-lab biologists didn't much believe that because of the low prediction accuracy. Until the appearance of PromoterInspector (Scherf, Klingenhoff et al. 2000), gene promoter prediction tools suffered from low accuracy. After the appearance of PromoterInspector, several efficient gene promoter prediction tools have also been developed. In the following, we expand further on their reviews and include more recent research on gene promoter prediction.

2.2.1 NNPP (Version 2.2)

NNPP 2.2 (Reese 2001) is constructed of time-delay neural networks. The network model is a special case of a feed-forward neural network, which has been successfully applied to voice recognition. Time-delay neural networks slightly differ from feed-forward neural networks in the design of the hidden layer. The hidden nodes of a standard feed-forward model are determined by experiment or by lemma. In a time-delay model, the hidden nodes are

determined by the number of the input nodes and the input size of the receptive fields. Therefore, the input layer and the hidden layer are no longer fully connected. The hidden nodes in a hidden layer are only connected to input nodes within a particular receptive field. In a time-delay model, the hidden node is also called a feature node and is known as weight sharing in neural network technology.

In the training process, all the weights in the same receptive filed will be calculated and then copied to each other. However, the weights computed between the hidden nodes and the output nodes are still based on standard feed-forward algorithms. To optimize promoter prediction accuracy, two time-delay neural network models which recognize TATA-box with 30 bp (-40 bp to -10 bp from the TSS) and Inr (-14 bp to +11 bp from the TSS) regions of promoters are used. A combined model with 51 bps (-40 bps upstream to +11 bps downstream of TSS) is used along with the two models mentioned above for promoter prediction. The testing results showed that NNPP demonstrated 75% true positives for a fruit fly genome with a length of 2.9 Mbps.

2.2.2 McPromoter

MCPromoter (Ohler, Stemmer et al. 2000) coordinated three interpolated Markov chains (IMCs) to look for eukaryotic polymerase II TSSs in genomic DNA. It consists of a model for promoter sequences and a mixture model for non-promoter sequences, containing sub-models for coding and non-coding

sequences. To localize TSSs, a window of 300 bases is shifted over the sequence in steps of 10 bases. 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 (see Duda and Hart 1973) to eliminate single false predictions and reduce the high number of neighboring minima that are due to noise. The program then makes a prediction for each local minimum below a pre-specified threshold.

The training dataset are extracted from the EPD which contains a total of 565 vertebrate promoter sequences, and each contained 250 bp upstream and 50 bp downstream from the TSS. After the five-fold cross validation evaluation, the system reached an 84% correlation coefficient for promoters versus coding regions and achieved a 53% correlation coefficient for promoters versus non-coding regions.

2.2.3 Eponine

Eponine (Down and Hubbard 2002) proposed a probabilistic method for detecting transcription start sites (TSS) in mammalian genomic sequence, with good specificity and excellent positional accuracy. Eponine models consist of a set of DNA weight matrices recognizing specific sequence motifs. Each of these is associated with a position distribution relative to the transcription start site.

Eponine has been tested by comparing the output with annotated mRNAs from human chromosome 22. From this work, they estimate that using the default threshold (0.999), it detects >50% of transcription start sites with 70% specificity. However, it does not always predict the direction of transcription correctly. It's an effect which seems to be common among computational TSS finders.

2.2.4 CpGProD

CpGProD (Ponger and Mouchiroud 2002) is a program dedicated to the prediction of promoters associated with CpG Islands in mammalian genomic sequences. In vertebrate genomes, the CpG Islands (CGIs) are involved in DNA methylation of gene transcription. 50-60% of the human genes exhibit a CGI over the transcription start site (TSS) but not all the CGIs are associated with promoter regions (Larsen, Gundersen et al. 1992). CpGProD uses a CGI definition more stringent than that proposed by Gardiner-Garden and Frommer (1987). CpG Island are defined as DNA regions longer than 500 nucleotides (instead 200 bp), with a moving average G + C frequency above 0.5 and a moving average CpG observed/expected (CpG o/e) ratio greater than 0.6. Although it is strictly dedicated to this particular promoter class corresponding to ≈50% of the genes, CpGProD exhibits a higher sensitivity and specificity than other tools used for promoter prediction.

2.2.5 PromoterInspector

PromoterInspector (Scherf, Klingenhoff et al. 2000) is one of the most well-known content-based gene TSSs prediction tools, which gives attention to analyzing genetic context instead of context location. Its main idea is to extract common sequence features from sequences and generate a set of context features called IUPAC (International Union of Pure and Applied Chemistry) word dictionaries. IUPAC word dictionaries are composed of an IUPAC group which is defined by a set of oligonucleotides and a number of undefined base pairs (i.e. the letter "N" could represent A, C, G or T). To optimally distinguish a promoter region from an un-annotated DNA sequence, PromoterInspector introduces not only a promoter region as the training set but also three non-promoter sequences, that is, exon, intron, and 3'UTR. These training data sets then generate three classifiers: promoter and exon, promoter and intron, and promoter and 3'UTR. These three classifiers are the basis of PromoterInspector. In the experiment results, the vertebrate promoter training data were extracted from EPD, and the non-promoter exon and intron were randomly downloaded from NCBI GenBank; the 3'UTR was selected from the UTR database. The valuation data that were collected by Ficktt and Hatzigeorgiou were tested. The greatest advantage of PromoterInspector seems to be that of dramatically reducing the false positives. This advantage can help lab researchers avoid

unnecessary experiments or help other prediction tools further and exactly locate the transcriptional start sites.

2.2.6 Promoter 2.0

Promoter 2.0 (Knudsen 1999) combines several neural networks. Each neural network model uses perceptron-like algorithms. In the training phase, the inputs of the neural networks are a small window of DNA sequence and the output of other networks. For example, giving a network with a given input window size and a set of weights, the network scans along a DNA sequence and records the scores that are generated from every segment. Next, another new network with the same structure but with different weight scans the same DNA sequence. The highest score of each of all the previous networks for the same DNA sequence will be multiplied by a separation function. The result will become the input of this new network. In Promoter 2.0, there are four networks responded to TATA-box, cap site, CCAAT-box and GC box. To optimize these neural networks, genetic algorithms are used to randomly choose and change an individual weight. If the performance of network improved, the new weight was kept. The new weight was ignored if the performance did not improve. Nevertheless, the crossover operation of genetic algorithms is not used in Promoter 2.0. The training and testing data of Promoter 2.0 were 100 vertebrate sequences. For a positive set, 200 bps upstream of the cap site was selected as the promoter sequence. In contrast, 200 bps downstream of the cap site was

assigned as the non-promoter sequence; Promoter 2.0 reported 63% true positives for its testing data.

2.2.7 Dragon Promoter Finder

The Dragon Promoter Finder 1.2 (DPF) (Bajic, Seah et al. 2002) is an gene promoter prediction model for vertebrates. The DPF consists of a nonlinear promoter recognition model, sensors for recognizing specific functional regions of DNA, signal processing, and artificial neural networks. Before using the DPF, a user has to supply a DNA sequence and a selected accuracy range to the DPF system. The DPF then reads the DNA sequence by a sliding window and shifts one base pair each time. The data window's content passes through three sensors. Each sensor responds to a specific functional region, such as a promoter, exon and intron. A non-linear signal processing model further analyzes the sensor's output and feeds it into to a neural network to ascertain the input DNA sequence and determine if a promoter region exists.

The DPF 1.3 extends the capability for recognizing a GC-rich or GC-poor DNA sequence. The DPF used 793 different vertebrate promoter sequences from EPD as positive training set; each sequence was 250 bps long, covering 200 bps upstream and 50 bps downstream from the TSS. For a negative training set, the DPF used 800 coding-exon and 4000 intron sequences. To further tune the system parameters, DPF extended 400 3' UTR human sequences and 200 human exon and 500 human intron sequences to the training set. Each sequence

was also 250 bps long. The final testing data contained 146 human and humanviral sequences, which contained 159 TSSs. The DPF showed 66% true positives, which outperformed other tools significantly.

2.2.8 Dragon Gene Start Finder

Dragon Gene Start Finder (Dragon GSF) (Bajic and Seah 2003) combines three systems. The first one is Dragon Promoter Finder (DPF) (Bajic et al (2003), the second one is the system which estimates the presence of the CpG Islands, and the third one combines information from these two into the final predictions. The third system performs the sensor fusion function utilizing data preprocessing and an artificial neural network (ANN). The combination algorithm will only select the one of possibly many predicted TSS locations in the region [-3700,+3700] relative to the central point of the CpG Island, such that jointly with the other input data the ANN produces the highest score above the selected threshold.

Dragon GSF made every effort not to mislead the user with claims of superior performance. Dragon GSF presented the performance as found on the tests on different genomic sequences including whole human chromosomes 4, 21, and 22. Based on these results Dragon GSF infer that the gene start finding capabilities of Dragon GSF system are among the best currently available. The estimated performance for the human genome of Dragon GSF implies sensitivity of ~65% (relative to all promoters), sensitivity of ~88% relative to the CpG

Island related promoters, and positive predictive value of ~78% relative to all known genes.

2.2.9 First Exon Finder

First Exon Finder (FirstEF) (Davuluri, Grosse et al. 2001) uses different discriminating functions structured as a decision tree to predict the first exons and promoters in a gene. The probabilistic models are designed to find potential first donor sites and CpG-related and non-CpG-related promoter regions based on discriminating analysis. For each potential first donor and upstream promoter region, FirstEF decides whether the intermediate region could be a potential first exon based on a set of quadratic discriminating functions. Test and training sets came from the same database.

FirstEF's accuracy was tested by ten-fold cross-validation analysis and running the program on the complete exon sequences of genes in chromosomes 21 and 22. First exons predictions were considered true positives if the predicted first donor site was identical to the real donor site and the predicted transcriptional start site (TSS) fell within the region between 500 bp before and 200 bp after the real TSS. Recall, precision, and the correlation coefficient were used to judge the accuracy of FirstEF.

2.2.10 Summary of Gene Promoter Prediction Tools

The gene promoter prediction tools introduced in this Session are summarized in Table 2.1. No previously developed tools can achieve the performance with higher sensitivity than 70% and higher specificity than 70%.

Table 2.1 Summary of gene promoter prediction tools.

Tool	Method	Species	Feature Consider	Data Source	Citation	Sn.	Sp.	False positive rate	Acc.	Available
Eponine	Relevance Vector Machine (RVM)	Mammalian	TATA box in a G+C rich domain	EPD	Reese 2001	53.5%	73.5%	-	-	Yes
Promoter 2.0	ANN	Vertebrate	Four TFBSs (TATA box, CCAAT box, GC box, Inr)	EPD	Ohler, Stemmer et al. 2000	68%	-	8%	-	No
NNPP 2.2	ANN	Drosophila	TATA box & Inr	EPD	Down and Hubbard 2002	70%	-	7.2%	-	Yes
CpGProD	Statistics based	mammalian	CpG Island	GenBank	Ponger and Mouchiroud 2002	56%	39%	-	-	Yes
PromoterInspector	Statistics based	Vertebrate	IUPAC	EPD	Scherf, Klingenhoff et al. 2000	48%	85%	-	-	No
Dragon PF	ANN	Human Chr. 22	CpG Island related	EPD	Knudsen 1999	60.17%	-	-	-	No
Dragon GSF	ANN	Human Chr. 4,21,22	G+C rich &G+C poor	DBTSS	Bajic, Seah et al. 2002	65.10%	-	-	77.80%	No
McPromoter	ANN	Human Chr. 22	Interpolated Markov Model	EPD	Bajic and Seah 2003	52.1%	40.3%	-	-	Yes
First Exon Finder	Quadratic discriminating analysis	Human Chr. 21,22	CpG Island related	NCBI	Davuluri, Grosse et al. 2001	79.3%	53.5%	-	-	No

* Sn. = Sensitivity, Sp. = Specificity, Acc. = Accuracy.

2.3 DNA stability

Aditi Kanhere *et al.* (Kanhere and Bansal 2005) devised a novel regulatory feature, DNA stability, for prokaryotic promoter prediction. The DNA stability is the structural property of the fragment of the DNA duplex, which is calculated based on the minimum free energy created by the hydrogen bond of the A-T and C-G pairs. There are several factors that stabilize a DNA double helix. Among the significant factors are:

1. Stacking Interaction

2. Watson-Crick Hydrogen Bonding Interaction

3. Interaction with Water Molecules and Metal Ions

SantaLucia *et al.* (SantaLucia 1998) use the unified standard free energy of ten dinucleotides duplexes, such as AA/TT, AT/TA, TA/AT, CA/GT, GT/CA, CT/GA, GA/CT, CG/GC, GC/CG, and GG/CC (SantaLucia 1998), to calculate the standard free energy change of a DNA oligonucleotide based on dinucleotid e composition. The free energy formula and the calculation example that used by Santalucia were shown in Fig. 2.2 and Fig. 2.3, respectively.

$$\Delta G^0 = -\left(\Delta G^0_{\text{ini}} + \Delta G^0_{\text{sym}}\right) + \sum_{i=1}^{n-1} \Delta G^0_{i,i+1}$$

where,

G^0_{ini} is the initiation free energy for dinucleotide of type *ij*.
ΔG^0_{sym} equals +0.43 kcal/mol and is applicable if the duplex is self-complementary.
$\Delta G^0_{i,j}$ is the standard free energy change for the dinucleotide of type *ij*.

Figure 2.2 The formula of calculating the free energy for the DNA sequence.

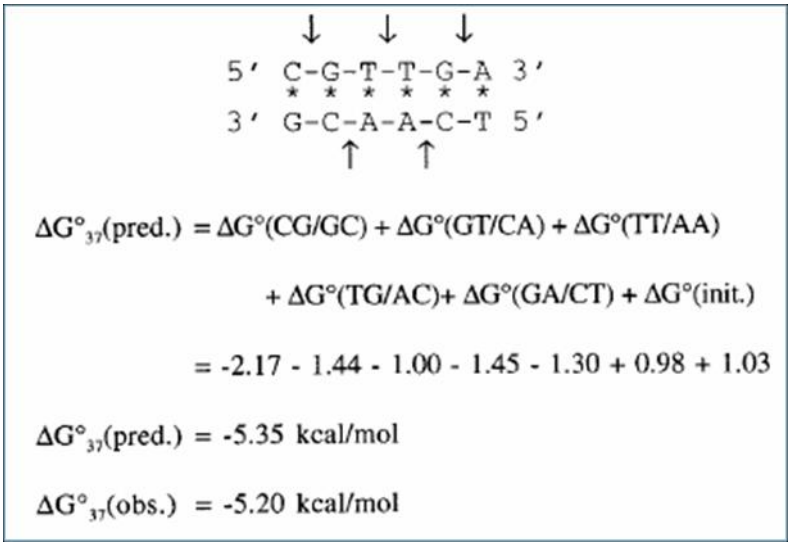


Figure 2.3 The example of free energy computation.

Aditi Kanhere and Manju Bansal presented a novel prokaryotic promoter prediction method based on DNA stability (Kanhere and Bansal 2005). They showed that the promoter region is less stable and hence more prone to melting as compared to other genomic regions. Figure 2.4 shows the distributions of average free energy of DNA duplex formation, and reveals a peak near the TSS, lying

between -10 and -30 region, which corresponds to the TATA box in the eukaryotic promoter sequences. Their analysis also showed that a method of promoter prediction based on the differences in the stability of DNA sequences in the promoter and non-promoter region works much better compared to existing prokaryotic promoter prediction programs, which are based on sequence motif searches. At present the method works optimally for genomes such as that of *Escherichia coli*, which have near 50 % G+C composition and also performs satisfactorily in case of other prokaryotic promoters.

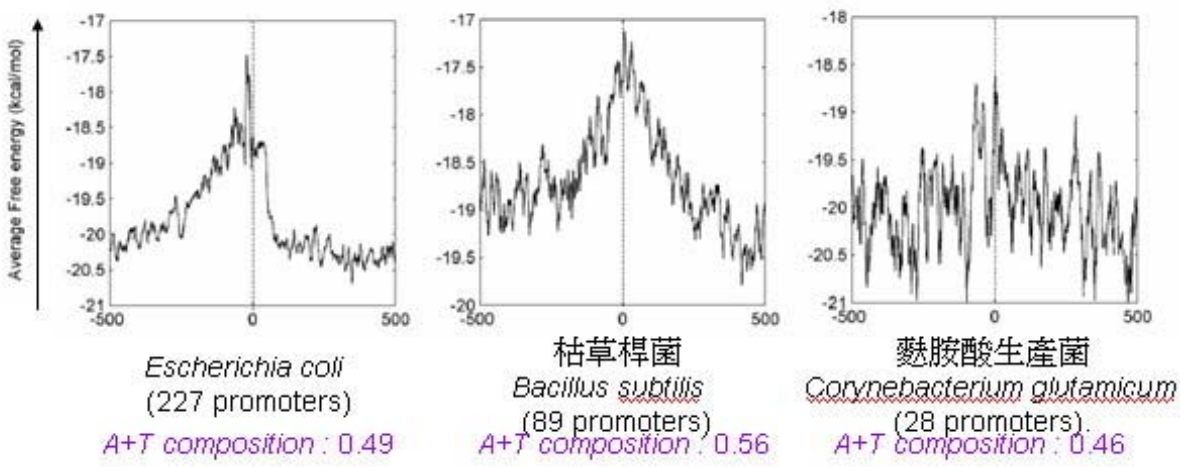


Figure 2.4 Average free energy nearby TSSs of three species.

2.4 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a popular technique for classification (Hsu, Chang et al.). A support vector machine (SVM) is a supervised learning technique from the field of machine learning applicable to both classification and regression.

Rooted in the Statistical Learning Theory developed by Vladimir Vapnik and co-workers at AT&T Bell Laboratories in 1995(Vapnik 1995), SVMs are based on the principle of Structural Risk Minimization. As shown in Fig. 2.5, the goal of SVM is to produce a model which predicts target value of data instances in the testing set which are given only the attributes.

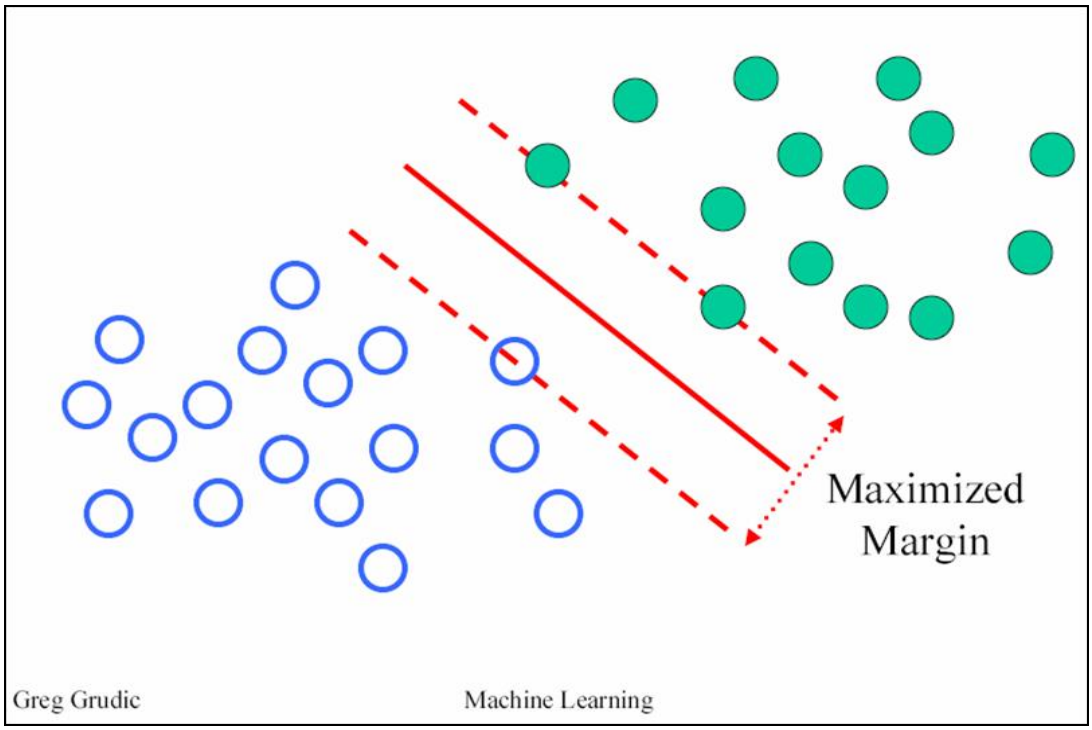


Figure 2.5 The concept of SVM

LIBSVM (Chang and Lin 2001) was developed by Chih-Chung Chang and Chih-Jen Lin. LIBSVM is an integrated software for support vector classification, regression, and distribution estimation. LIBSVM supports the multi-class classification. Since version 2.8, it implements an SMO-type algorithm. Their goal is to help users from other fields to easily use SVM as a tool. The SVM process

overview is shown in Fig. 2.6.

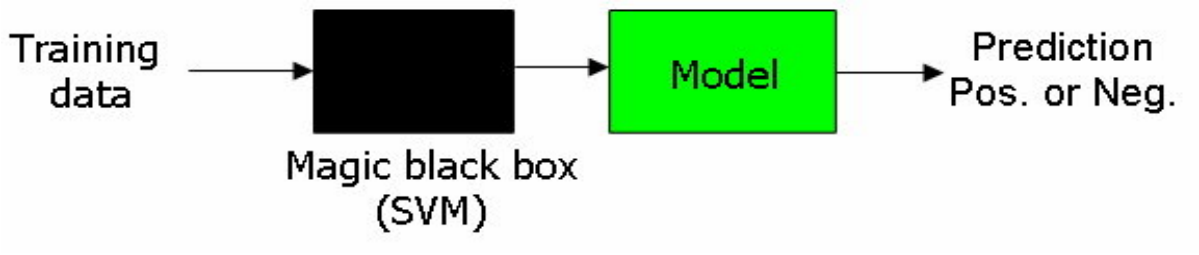


Figure 2.6 SVM process overview

Chapter 3 Material and Methods

3.1 Materials

The experimentally verified transcriptional start sites of DBTSS and human genome sequence of NCBI are used in this research.

3.1.1 Database of Transcriptional Start Site (DBTSS) and EPD

We extract the flanking sequence of experimentally verified transcriptional start sites form DBTSS. The vision of DBTSS we used is release 5.1.0. As given in Table 3.1, it contains 30,964 human promoter sequences which are length 1,201 base pairs (from -1,000 to +200) within 15,262 genes. And 8,308 human genes are found to have putative multiple promoters. As shown in Table 3.2, 1,871 human promoter sequences with length 6,000 base pairs (from -3000 to +3000) were obtained from EPD for the independent test, as well as the 196 mouse promoter sequences.

Table 3.1 The statistics of human and mouse experimentally TSSs in DBTSS.

Species	TSSs	Genes	Region	length
Human	30,964	15,262	-1000 ~ +200	1,201
Mouse	19,924	13,704	-1000 ~ +200	1,201

Table 3.2 The statistics of human and mouse experimentally TSSs in EPD.

Species	TSSs	Region	length
Human	1,871	-3000 ~ +3000	6,000
Mouse	196	-3000 ~ +3000	6,000

3.1.2 NCBI Genome Sequences

We extract the human whole genome sequences from NCBI. The vision we used is NCBI build 35 release 1. The total length of sequences assembled was 3,021,400,000 base pair.

3.2 Methods

We developed a gene promoter prediction method which integrates novel regulatory features with Support Vector Machine (SVM). Figure 3.1 shows the system flow in this works, there are three stages in developing TSSs prediction method. Stage 1 is to construct the training dataset for the TSS prediction. In the Stage 2, three kinds of regulatory features in the flanking region of experimentally verified TSSs are extracted and analyzed. Finally, Stage 3 uses the SVM to construct the classifying model for the three selected features and evaluate the prediction performance. We will discuss every stage in detailed as following.

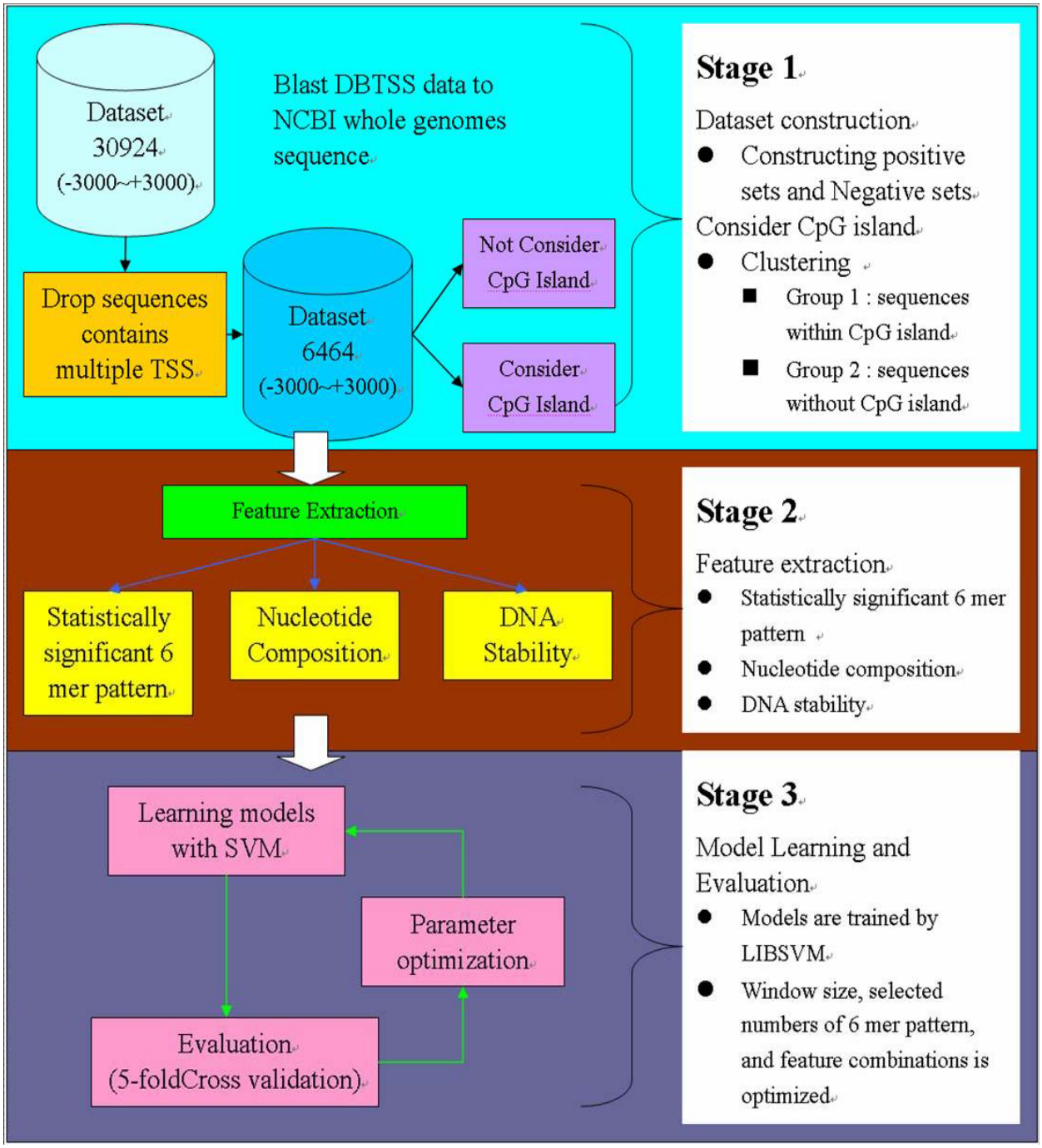


Figure 3.1 System flow of prediction tool developing.

3.2.1 Dataset Construction

The experimentally verified human promoter sequences were extract from DBTSS which contains 30,964 experimentally TSSs. All the promoter sequence

length of DBTSS are 1,201 base pairs (bps), which is extracted from 1,000 bps upstream to 201 bps downstream of the TSS. Because of the sequences we obtained from DBTSS were too short to be analyzed, we use BLAST to align the promoter sequences of DBTSS to the human whole genome sequence. With the alignment preprocess, we got the 6,000 bps sequence length of each promoter sequence (from 3,000 bps upstream to 3000 downstream of the TSS).

After the alignment preprocessing, we should make sure that each promoter sequence only contains just one TSS in the region of 6,000 bps sequence length. Therefore, we dropped the promoter sequences that had multiple experimental TSSs in the region of 6,000 bps sequence length. Finally, it remains 6,464 promoter sequences which have unique TSS in the region of 6,000 bps sequence length.

Furthermore, we also consider if the sequences in our dataset contains CpG Island or not. We used the CpG island prediction tool “CpGproD” to classify all the 6,464 promoter sequences into two groups (has CpG Isalnd or not). Table 3.2 shows the number of sequences of three groups, Group 1: all promoter sequences; Group 2: promoter sequences without CpG Island; Group 3: promoter sequences with CpG Island.

Table 3.3 Numbers of sequence of group 1, 2, and 3.

Group	Numbers of sequence
1 (All)	6,464
2 (Non-CpG)	1,566
3 (CpG)	4,898

3.2.2 Feature Extraction

In the feature extraction stage, we were trying to extract the useful regulatory features of promoter sequences to accurately identify human TSSs. We surveyed many related promoter prediction methods, and finally extracted three classifying features such as statistically significant 6-mer pattern, nucleotide composition, and DNA stability.

3.2.2.1 Statistically Significant 6-mer Pattern

First of all, we must to define the positive set and negative set of training data. Positive set was extracted from flanking regions of TSSs liked the upstream 200 to downstream 100 of TSS. The whole genomic regions other than the regions of positive set were defined as negative set. Following, we computed occurrence probability of 6-mer patterns for positive set and negative set according to the formula:

$$P = \frac{\sum_{i=1}^{n} S}{\sum_{i=1}^{n} Len - 5} \quad (1)$$

, where P denotes the occurrence probability, S denotes the number of 6-mer pattern occurrence in one sequence, and Len denotes the length of a sequence. After that we computed the occurrence ratio of every 6-mer pattern. The formula of Occurrence ratio:

$$O = \frac{Ppos}{Pneg} \quad (2)$$

, where O denotes occurrence ratio, P_{pos} denotes the probability of positive set, and P_{neg} denotes the probability of negative set. Finally, by optimized the top occurrence ratio pattern number, we chose top 60 occurrence rate of these patterns as features.

3.2.2.2 Nucleotide Composition

We try to analyze the nucleotide composition in the promoter region, the average occurrence rate of monomer, dimer, and trimer nucleotide are calculated in window size 20 bps sliding on the promoter sequence. The formula of average nucleotide occurrence rate:

$$AVG = \frac{\sum_{i=1}^{N} P}{N} \quad (3)$$

, where AVG denotes average nucleotide occurrence rate of the sliding window with size 10 bps, P denotes pattern occurrence times of every sequence, and N denotes the total number of sequences in our dataset. Then we specify the monomer nucleotide A and C occurrence rate as standard and computed a Pearson correlation coefficient for the other patterns with these two patterns to determine whether their correlation coefficient values are highly correlated. The cutoff value of the Pearson correlation coefficient is set to 0.8 to decide whether two patterns highly correlated or not.

3.2.2.3 DNA Stability

We computed the average of free energy of the 15 bps nucleotides and shift in the size of 5 bps nucleotides in our dataset sequences. The formula of DNA sequences of free energy is shown in Figure 3.2. The standard free energy change (ΔG_{37}^{0}) corresponding to the melting transition of an 'n' nucleotide (or 'n-1' dinucleotides) long DNA molecule, from double strand to single strand, is calculated as shown in Fig. 3.2 (Kanhere and Bansal 2005). ΔG_{ini}^{0} denotes two types of initiation free energy : "initiation with terminal G·C" and "initiation with terminal A·T"; ΔG_{sym}^{0} is +0.43 kcal/mol and is applicable if the duplex is self-complementary, and $\Delta G_{i,j}^{0}$ represents the standard free energy change for type ij dinucleotide. In the present calculation, each promoter sequence is divided into overlapping windows of 15 bps (or 14 dinucleotide steps), and for each window the free energy is calculated as shown above. We used the free energy of the 15 bps nucleotides and shift in the size of 5 bps nucleotides as features.

$$\Delta G^0 = -\left(\Delta G^0_{\text{ini}} + \Delta G^0_{\text{sym}}\right) + \sum_{i=1}^{n-1} \Delta G^0_{i,i+1}$$

where,

G^0_{ini} is the initiation free energy for dinucleotide of type *ij*.
ΔG^0_{sym} equals +0.43 kcal/mol and is applicable if the duplex is self-complementary.
$\Delta G^0_{i,j}$ is the standard free energy change for the dinucleotide of type *ij*.

Figure 3.2 Formula of DNA sequences free energy.

Aditi Kanhere *et al.* (Kanhere and Bansal 2005) demonstrated that the change in DNA stability appears to provide a much better clue than the usual sequence motifs. Therefore, this work provides the DNA stability of the promoter region to enhance the promoter identification.

3.2.3 Model Learning and Evaluation

After the process of feature extraction, the three kinds of feature are trained by LIBSVM (Chang and Lin 2001) which developed by Chih-Chung Chang and Chih-Jen Lin. LIBSVM is an integrated software for support vector classification, regression, and distribution estimation. The three features of the positive and negative training set are transformed into LIBSVM input format, and the input values are normalized before the SVM model construction. The constructed SVM models of the three kinds of feature are evaluated by K-fold cross-validation.

We used 5-fold cross validation to evaluate the prediction performance of the SVM-trained model. To express the prediction quality of the models, we applied

several criteria. These evaluation criteria included sensitivity, specificity, accuracy, and precision. Figure 3.3 shows the definition of sensitivity, specificity, accuracy, and precision. To optimize our models' performance, we do parameter optimization in the window size, selected numbers of 6-mer pattern, and feature combinations.

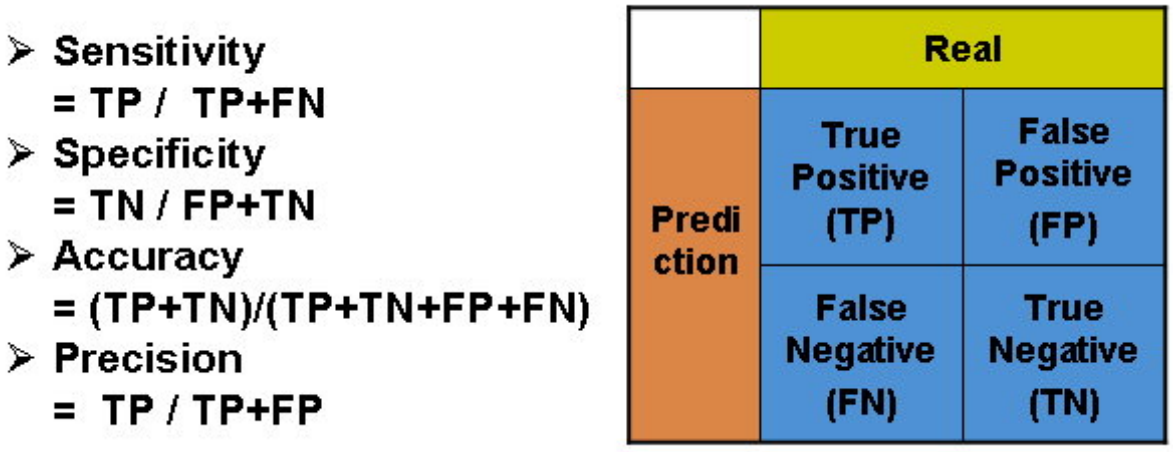


Figure 3.3 The definition of four performance measures.

To fairly evaluate the prediction performance of the gene promoter prediction method, an evaluation benchmark should be constructed. By surveying previous related works, one of the several evaluation benchmarks they used was chosen by us. As shown in Fig. 3.4, the positive sets were extracted from the positive region, and negative sets were randomly extracted from the six kinds of negative regions. There are five kinds of window sizes for positive set including 80, 150, 300, 450, and 600 bps. We show the regions of positive and six kinds of negative in detail in Table 3.3.

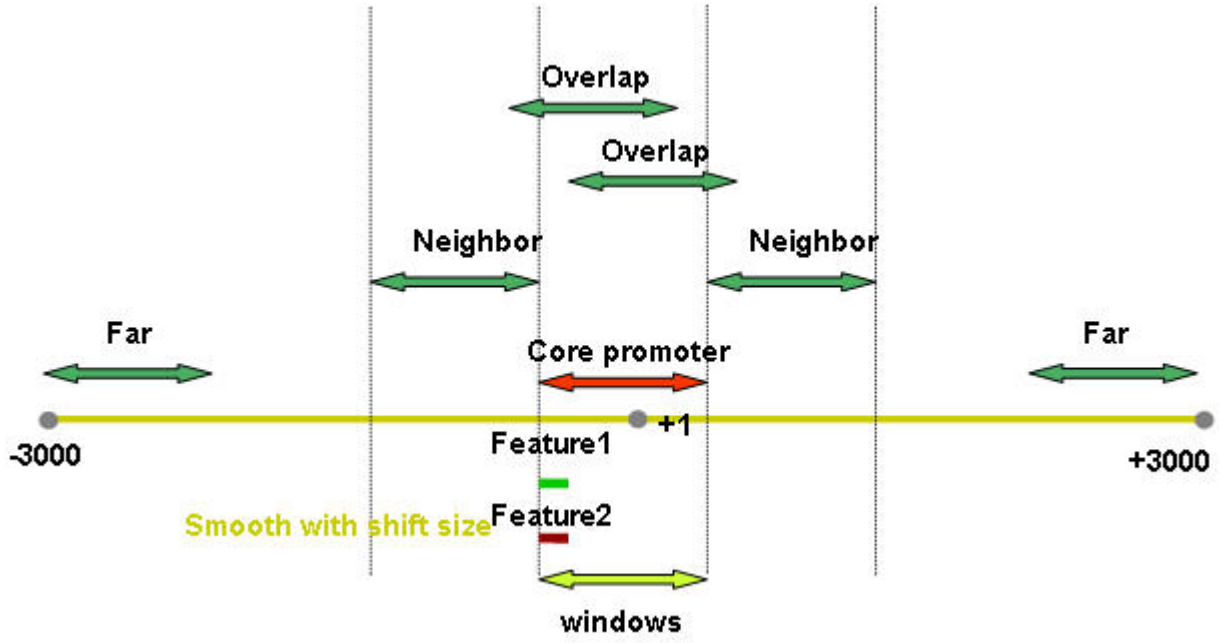


Figure 3.4 Evaluation Benchmark.

Table 3.4 Positive region and six kinds of negative region of window size we optimized.

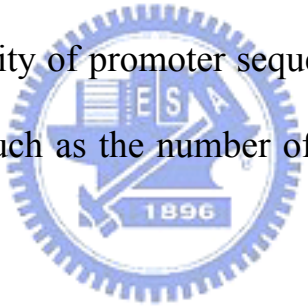
Window size	80	150	300	450	600
Core promoter regions	-60~+20	-100~+50	-200~+100	-300~+150	-400~+200
Overlap regions	-50~+30	-50~+30	-50~+30	-50~+30	-50~+30
Overlap regions	-70~+10	-70~+10	-70~+10	-70~+10	-70~+10
Neighbor regions	+21~+100	+21~+100	+21~+100	+21~+100	+21~+100
Neighbor regions	-140~ -81	-140~ -81	-140~ -81	-140~ -81	-140~ -81
Far regions	+2921~+3000	+2921~+3000	+2921~+3000	+2921~+3000	+2921~+3000
Far regions	-3000~-2921	-3000~-2921	-3000~-2921	-3000~-2921	-3000~-2921

Chapter 4 Results

Here we present the results of our evaluation benchmark to show the prediction performance of the constructed models based on Support Vector Machine (SVM). We also compared our models' accuracy with several existing gene promoter prediction tools that we obtained.

4.1 Feature Observations

Three features such as statistically significant 6-mer patterns, nucleotide composition, and DNA stability of promoter sequences are extracted and observed to decide some parameters such as the number of 6-mer patterns and the window size of positive set.



4.1.1 Statistically Significant 6-mer Patterns

By optimizing the selected numbers of pattern, we selected the top 60 patterns as our prediction feature .The selected top 60 patterns of group 1 are given in Table 4.1. The selected top 60 patterns of group 2 are given in Table 4.2. The top 60 selected patterns of group 3 are given in Table 4.3. We will show top 100 patterns in detail in appendix.

Table 4.1 Top 60 selected pattern of group 1 (all).

CGCGCG	CGCGCA	CGCCGA	CGGTCG
GCGCGC	TGCGCG	TCGGCG	CGGCCG
CGCCGC	CCCGCG	CGGCGA	CGCCCC
GCGGCG	CGCGGG	TCGCCG	GGGGCG
CGGCGC	CGCGAG	CGAGCG	ACGCCG
GCGCCG	CTCGCG	CGCTCG	CGGCGT
CGCGGC	CGCGGA	CCGCGA	GCCGCC
GCCGCG	TCCGCG	TCGCGG	GGCGGC
CCGCCG	TCGCGA	CGCGTC	CCCCGC
CGGCGG	CGCGAC	GACGCG	GCGGGG
CCGGCG	GTCGCG	ACGCGC	CCGCCC
CGCCGG	CGTCGC	GCGCGT	GGGCGG
AGCGCG	GCGACG	CGACGC	GCGGCC
CGCGCT	CCGCGC	GCGTCG	GGCCGC
CCGCGG	GCGCGG	CGACCG	CCGTCG

Table 4.2 Top 60 selected pattern of group 2 (non-CpG island).

CGCGAA	TTCGCG	CGCCCC	GGGGCG
CCGCCC	GGGCGG	CCCCGC	GCGGGG
ATCCGG	CCGGAT	GACGTC	CCGCAG
CTGCGG	CGCGAG	CTCGCG	CGGAAG
CTTCCG	ACGTCA	TGACGT	CGTCGA
TCGACG	CCGGAA	TTCCGG	GCCGGC
CGGCAG	CTGCCG	ACTCGC	GCGAGT
CGCGGA	TCCGCG	CCGCGA	TCGCGG
CGGCGA	TCGCCG	CGACTC	GAGTCG
GCGGCA	TGCCGC	CGGCCC	GGGCCG
TCCGCA	TGCGGA	CGGGGC	GCCCCG
GCCGGA	TCCGGC	GCCCGA	TCGGGC
AGGGCG	CGCCCT	CGTCAC	GTGACG
ACCGGA	TCCGGT	GGGCCC	ACGGCC
GGCCGT	GCCGAC	GTCGGC	CGGCAC

Table 4.3 Top 60 selected pattern of sequences group 3 (CpG island).

CGCGCG	GCGCGC	CGCCGC	GCGGCG
CGGCGC	GCGCCG	CGCGGC	GCCGCG
CCGCCG	CGGCGG	CCGGCG	CGCCGG
AGCGCG	CGCGCT	CCGCGG	CGCGCA
TGCGCG	CCCGCG	CGCGGG	CGCGAG
CTCGCG	CGCGGA	TCCGCG	TCGCGA
CGCGAC	GTCGCG	CGTCGC	GCGACG
CCGCGC	GCGCGG	CGCCGA	TCGGCG
CGGCGA	TCGCCG	CGAGCG	CGCTCG
CGCGTC	GACGCG	ACGCGC	GCGCGT
CCGCGA	TCGCGG	CGACGC	GCGTCG
CGACCG	CGGTCG	CGGCCG	CGCCCC
GGGGCG	ACGCCG	CGGCGT	GCCGCC

GGCGGC GCGGGG	GCGGCC CCGTCG	GGCCGC CGACGG	CCCCGC CCGCCC

4.1.2 Nucleotide Composition

The distributions of monomer, dimer, and trimer nucleotide composition of group 1, 2, and 3 are shown in Fig. 4.1, 4.2, and 4.3, respectively. The cutoff value of the Pearson correlation coefficient is set to 0.8 to decide whether two distribution patterns highly correlated or not. The selected highly correlated patterns of monomer, dimer, and trimer nucleotide composition are given in Table 4.4, 4.5, and 4.6, respectively.

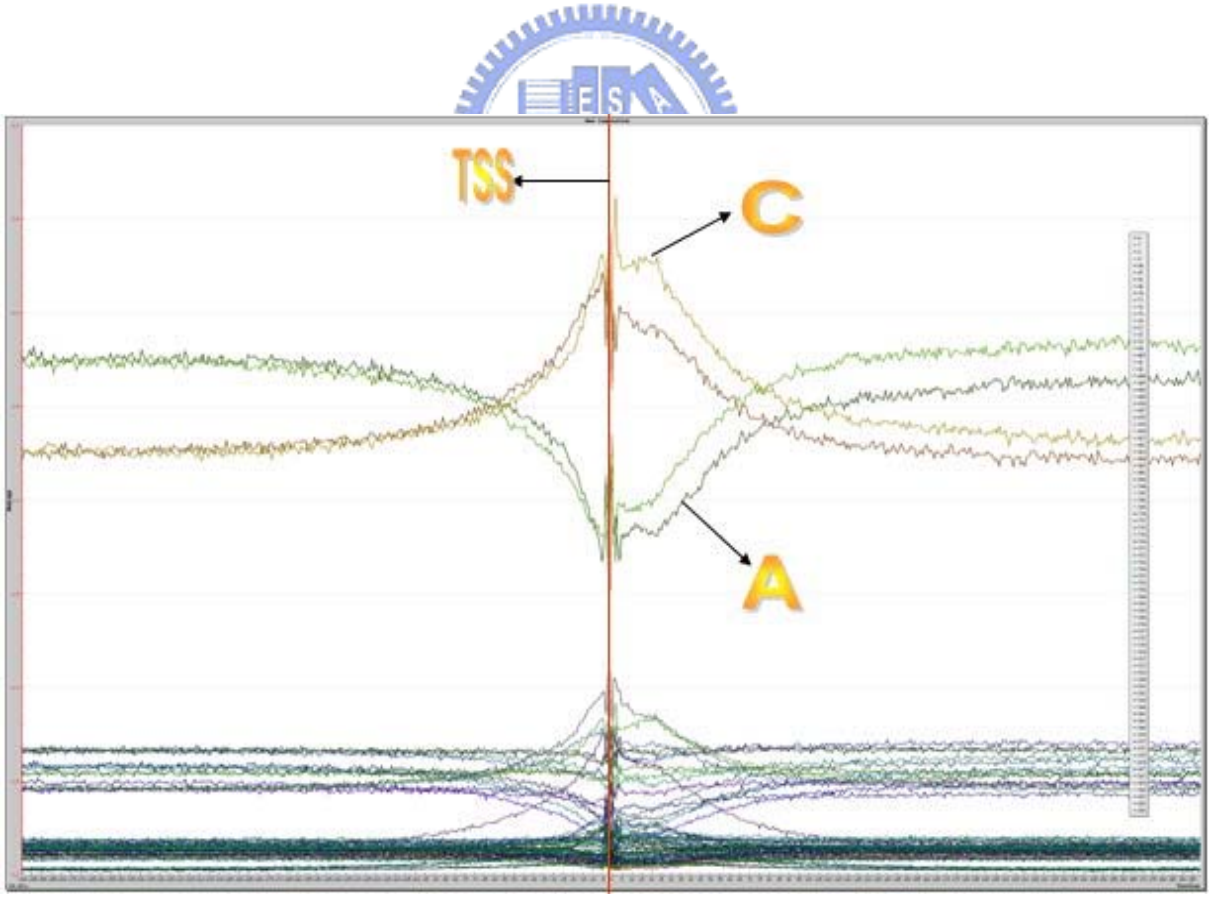


Figure 4.1 The distribution of monomer, dimer, and trimer nucleotide composition of group 1 (all).

Table 4.4 The selected highly correlated patterns of nucleotide composition of group 1 (all).

	C.C. with A >0.8	**C.C. with C >0.8**	**No select**
Monomer and Dimer	T,AA,AT, AC,TA,TT,CA	G,CC,CG, GC,GG	AG,TC,TG, CT,GA,GT
Trimer	AAA,AAT,AAC AAG,ATA,ATC ATT,ATG,ACA ACT,AGATAA TAC,TAT,TAG TTA,TTT,TCA TGA,CAA,CAT CTA	ACG,AGC,TCC,TCG CTC,CCC,CCG,CGA CGT,CGC,CGG,GAC GTC,GCC,GCG,GGA GGC,GGG	ACC,AGT,AGG,TTC TTG,TCT,TGT,TGC TGG,CAC,CAG,CTT CTG,CCA,CCT,GAA GAT,GAC,GAG,GTT GTG,GCA,GCT,GGT

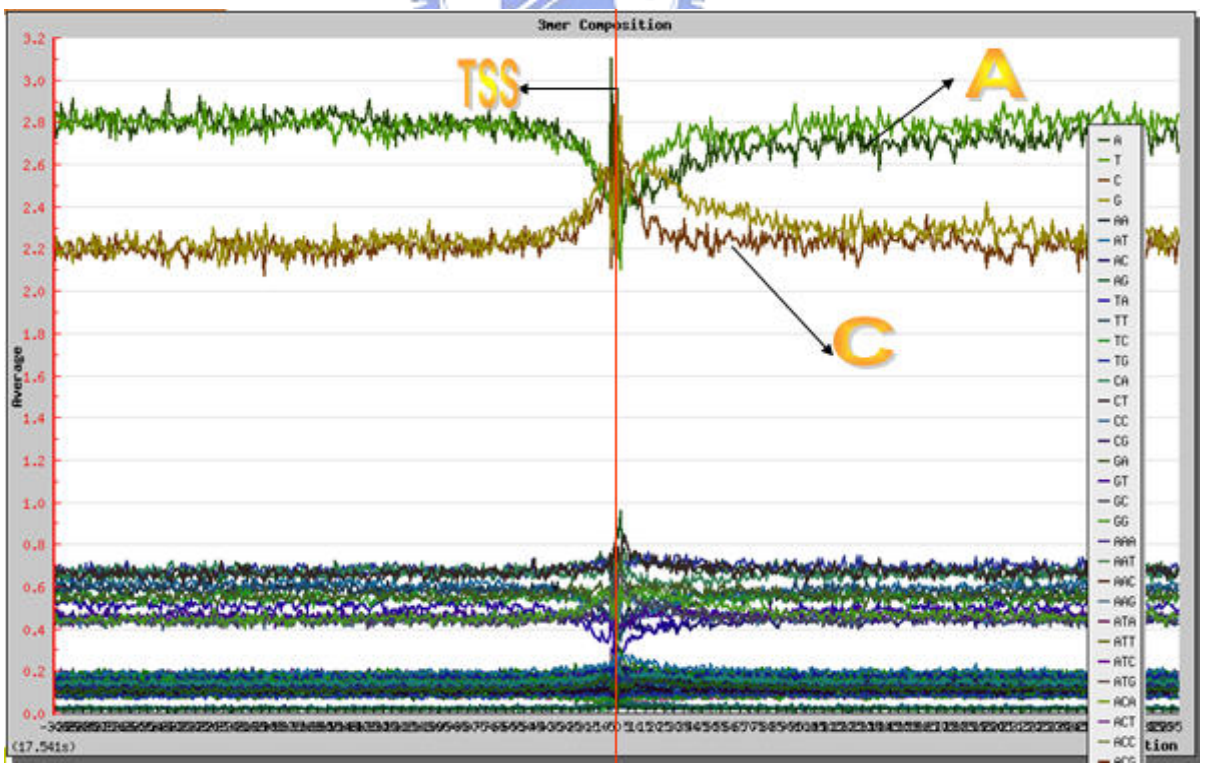


Figure 4.2 The distribution of monomer, dimer, and trimer nucleotide composition of group 2 (non-CpG island).

Table 4.5 The selected highly correlated patterns of nucleotide composition of group 2 (non-CpG island).

	C.C. with A >0.8	**C.C. with C >0.8**	**No select**
Monomer and Dimer	AA,AT,TA	G,AG,CC,CG,CT GC,GG,TC	T,AC,CA,GA,GT,TG,TT
Trimer	AAA,ATA	ACG,AGC,AGG,CAG CCA,CCC,CCG,CCT CGA,CGT,CGC,CGG CTC,CTG,GAC,GAG GCC,GCG,GCT,GGA GGC,GGG,GTC,TCC TGC,TGG	AAC,AAG,AAT,ACA ACC,ACT,AGA,AGT ATC,ATG,ATT,CAA CAC,CAT,CTA,CTT GAA,GAT,GCA,GGT GTA,GTG,GTT,TAA TAC,TAG,TAT,TCA TCT,TGA,TGT,TTA TTC,TTG,TTT

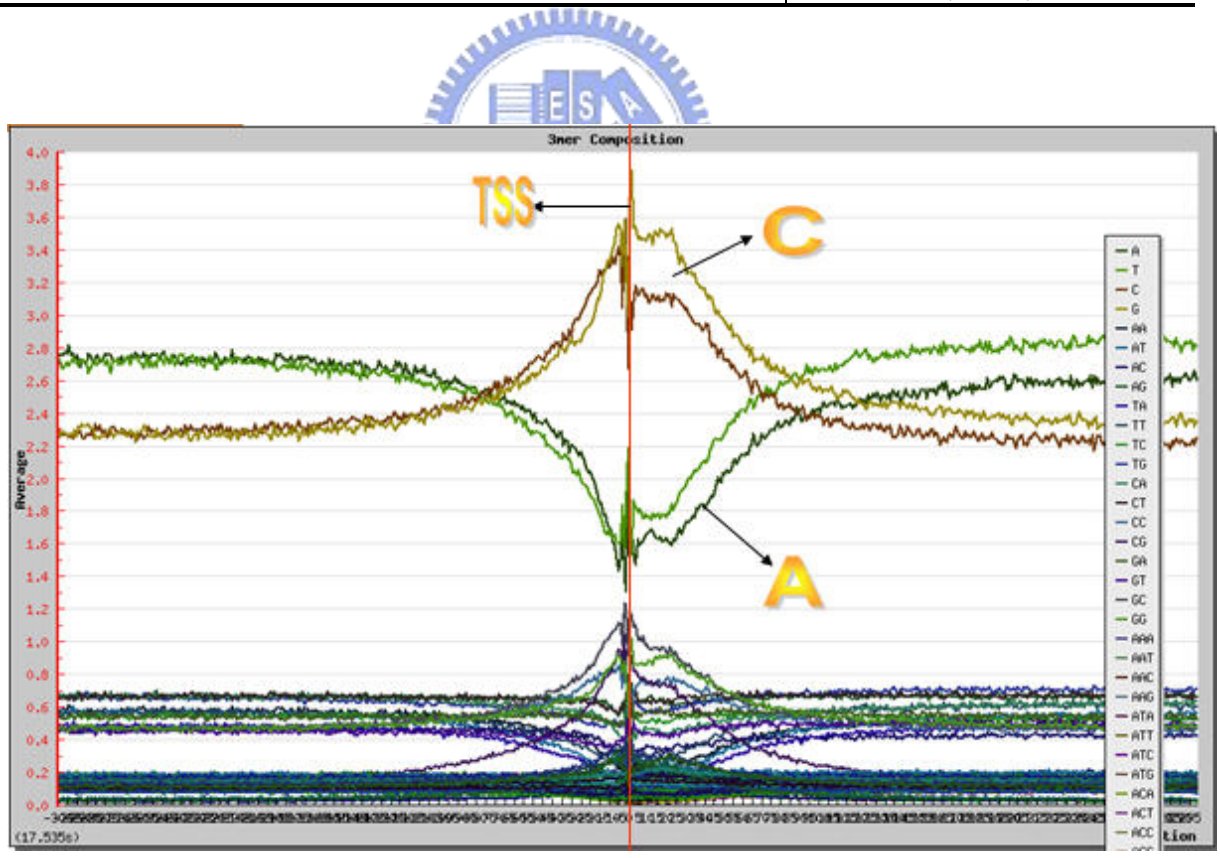


Figure 4.3 The distribution of monomer, dimer, and trimer nucleotide composition of group 3 (CpG island).

Table 4.6 The selected highly correlated patterns of nucleotide composition of group 3 (CpG island).

	C.C. with A >0.8	C.C. with C >0.8	No select
Monomer to Dimer	T,AA,AT, AC,TA,TT,CA	G,CC,CG, GC,GG	AG,TC,TG, CT,GA,GT
Trimer	AAA,AAT,AAC,AAG ATA,ATC,ATT,ATG ACA,ACT,AGA,AGT TAA,TAC,TAT,TAG TTA,TTT,TCA,TCT TGA,CAA,CAT,CTA GAA,GAT,GTA	ACG,AGC,TCC,TCG CTC,CCC,CCG,CGA CGT,CGC,CGG,GAC GTC,GCC,GCG GGA,GGC,GGG	ACC,AGG,TTC TTG,TGT,TGC TGG,CAC,CAG CTT,CTG,CCA CCT,GAG,GTT GTG,GCA,GCT GGT

4.1.3 DNA Stability

The distributions of the average free energy for the 3,000 upstream and 3000 downstream of TSSs with the 15 nucleotides sliding window and shift in 5 nucleotides are shown in Fig. 4.4, 4.5, and 4.6, respectively.

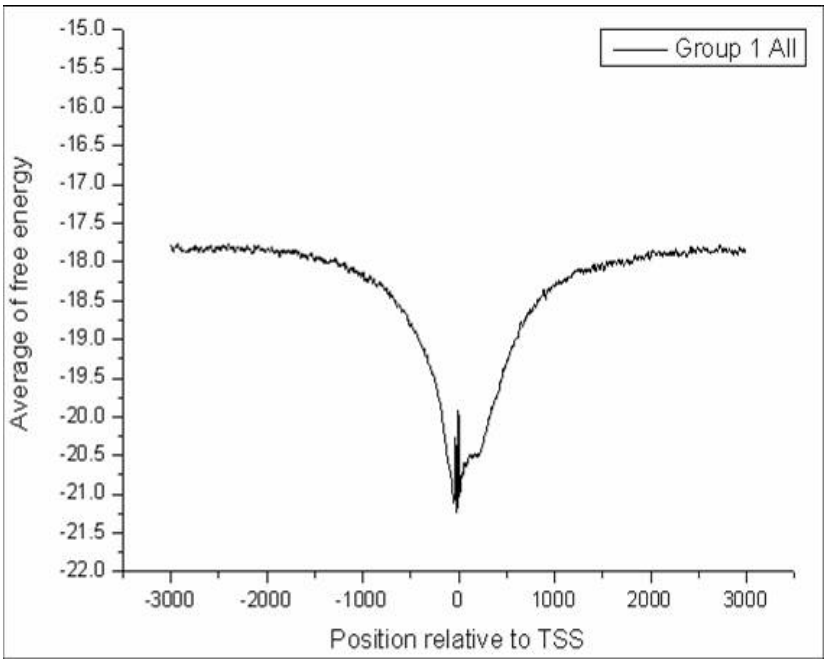


Figure 4.4 The distribution of average free energy relative to TSSs of group 1 (all).

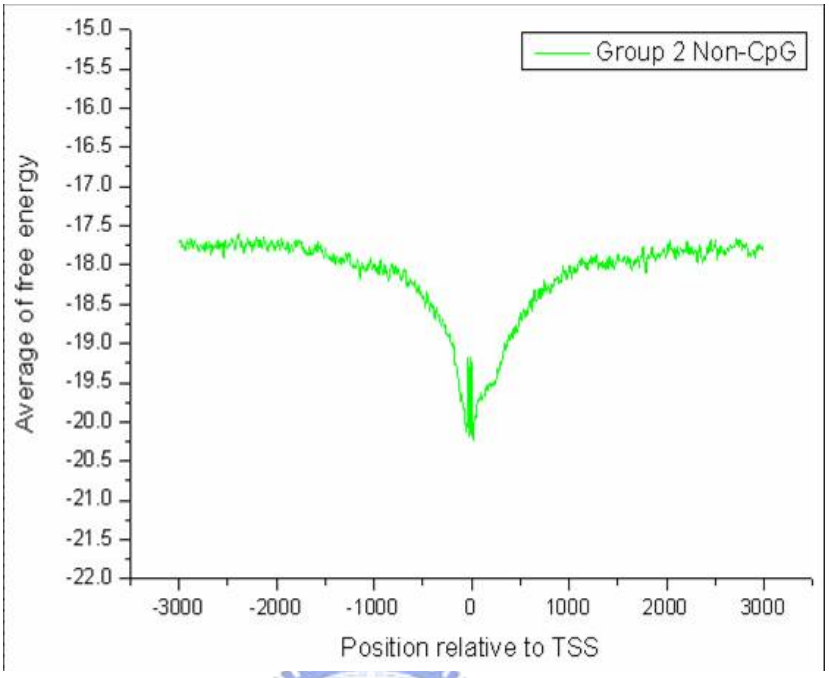


Figure 4.5 The distribution of average free energy relative to TSSs of group 2 (non-CpG island).

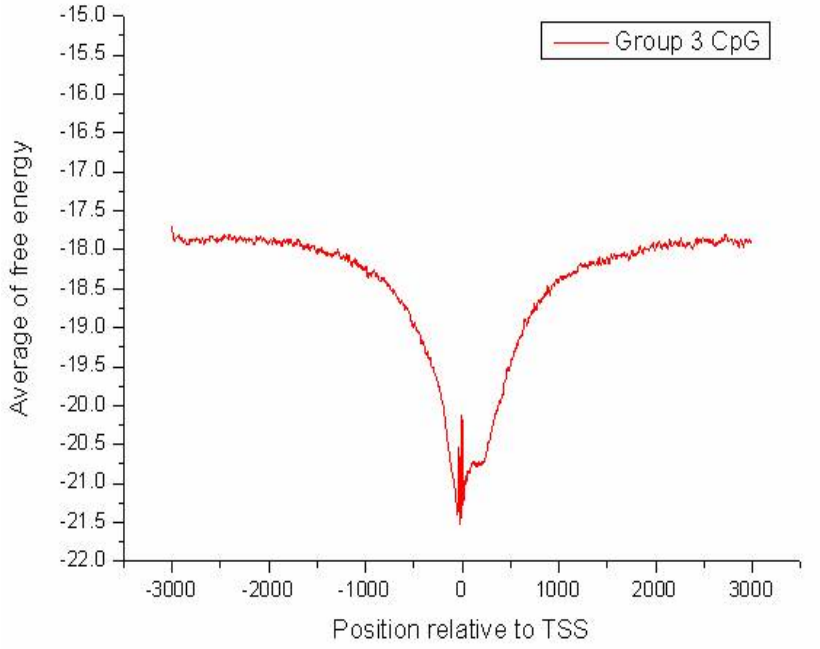


Figure 4.6 The distribution of average free energy relative to TSSs of group 3 (CpG island).

4.2 Prediction Performance

With the evaluation benchmark described previously in chapter 3, the prediction performance of the three SVM models can be evaluated fairly and clearly.

4.2.1 Statistically Significant 6-mer Patterns

The prediction sensitivity (Sen.), specificity (Spe.), accuracy (Acc.), and precision (Pre.) of the constructed SVM model based on statistically significant 6-mer patterns of group 1 (all), 2 (non-CpG), and 3 (CpG) are given in Table 4.7, 4.8, and 4.9, respectively. In addition, the negative set is randomly extracted from six negative regions we define in Table 3.3. As you can see, the larger window size is, the higher prediction performance is. Even so, we still choose the models of window size 300 to be our prediction models because of the prediction accuracy is over 70% and we want to determine the core promoter region accurately. Moreover, the prediction performance of group 3 (CpG) is better than group 1 (all) and 2 (non-CpG).

Table 4.7 The models accuracy of 6mer pattern in group 1(all).

Negative set	Positive set	Window size	Sen.	Spec.	Acc.	Pre.
Random	- 60 ~ + 20	80	51%	77%	64%	69%
Random	-100 ~ + 50	150	58%	76%	67%	71%
Random	-200 ~ +100	300	62%	77%	70%	73%
Random	-300 ~ +150	450	64%	80%	72%	76%
Random	-400 ~ +200	600	65%	83%	74%	79%

Table 4.8 The models accuracy of 6mer pattern in group 2 (non-CpG island).

Negative set	Positive set	Window size	Sen.	Spec.	Acc.	Pre.
Random	- 60 ~ + 20	80	22%	85%	53%	59%
Random	-100 ~ + 50	150	28%	85%	56%	65%
Random	-200 ~ +100	300	32%	83%	58%	66%
Random	-300 ~ +150	450	32%	83%	58%	66%
Random	-400 ~ +200	600	35%	82%	58%	66%

Table 4.9 The models accuracy of 6mer pattern in group 3 (CpG island).

Negative set	Positive set	Window size	Sen.	Spec.	Acc.	Pre.
Random	- 60 ~ + 20	80	56%	75%	66%	70%
Random	-100 ~ + 50	150	67%	74%	71%	72%
Random	-200 ~ +100	300	75%	76%	76%	76%
Random	-300 ~ +150	450	77%	78%	78%	78%
Random	-400 ~ +200	600	79%	82%	80%	81%

Figure 4.7 shows the distribution of predictions by the constructed SVM models (-200 ~ + 100) based on statistically significant 6-mer patterns of group 1, 2, and 3 in the regions of upstream 3,000 bps to downstream 3,000 bps of the TSS. The sliding window size was set to 300 nt which determined by the selective window size (-200 ~ +100) of positive set and shift in the size of 50 bps nt, and the number of predictions were calculated in each window. As it show, the group 1 and 3 have the similar prediction performance, and both of them are better than group 2.

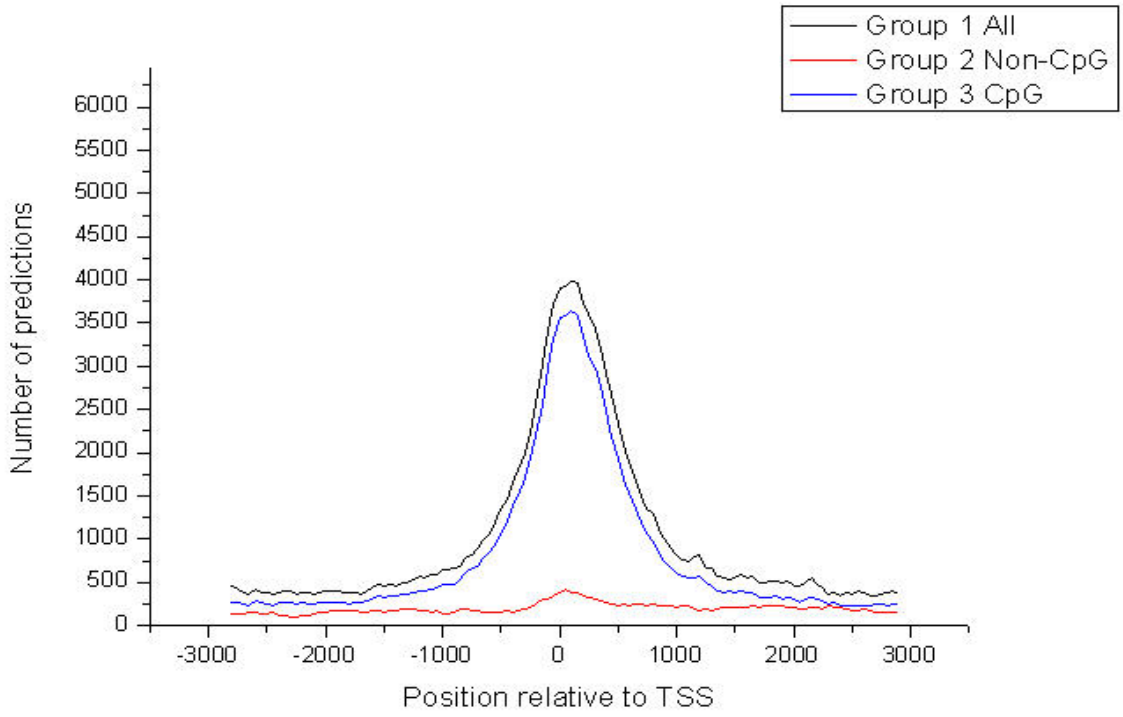


Figure 4.7 Distributions from 6 mer pattern models' predictions of group 1, 2, and 3 in the interval [-3000, +3000] relative to the TSS based on DBTSS.

4.2.2 Nucleotide Composition

The prediction sensitivity (Sn.), specificity (Sp.), accuracy (Acc.), and precision (Pre.) of the constructed SVM model based on monomer to dimer and monomer to trimer of group 1 are given in Table 4.10 and 4.11, respectively. Because the results of monomer to trimer did not have much increases, we chosen monomer to dimer to evaluate group 2, and group 3. Thus The prediction sensitivity (Sn.), specificity (Sp.), accuracy (Acc.), and precision (Pre.) of the constructed SVM model based on monomer to dimer of group 2, and 3 are given in Table 4.12, and 4.13, respectively. In addition, the negative set is randomly

extracted from six negative regions we define in Table 3.3. As you can see, the larger window size is, the higher prediction performance is. Even so, we still choose the models of window size 300 to be our prediction models. Because its' model accuracy is over 70% and we want to determine the core promoter region accurately. Moreover, the prediction performance of group 3 (CpG) is better than group 1 (all) and 2 (non-CpG).

Table 4.10 The models accuracy of monomer to dimer of group 1 (all).

Negative set	Positive set	Window size	Sn	Sp	Acc	Pre
Random	- 60 ~ + 20	80	69%	69%	69%	69%
Random	-100 ~ + 50	150	67%	71%	70%	70%
Random	-200 ~ +100	300	69%	74%	71%	72%
Random	-300 ~ +150	450	70%	75%	72%	74%
Random	-400 ~ +200	600	72%	75%	74%	74%

Table 4.11 The models accuracy of monomer to trimer of group 1 (all).

Negative set	Positive set	Window size	Sn	Sp	Acc	Pre
Random	- 60 ~ + 20	80	66%	71%	69%	70%
Random	-100 ~ + 50	150	67%	73%	70%	71%
Random	-200 ~ +100	300	67%	73%	70%	71%
Random	-300 ~ +150	450	69%	76%	73%	74%
Random	-400 ~ +200	600	71%	75%	73%	74%

Table 4.12 The models accuracy of monomer to dimer of group 2 (non-CpG island).

Negative set	Positive set	Window size	Sn	Sp	Acc	Pre
Random	- 60 ~ + 20	80	61%	71%	66%	68%
Random	-100 ~ + 50	150	62%	68%	65%	66%
Random	-200 ~ +100	300	65%	69%	67%	68%
Random	-300 ~ +150	450	63%	68%	66%	67%
Random	-400 ~ +200	600	63%	68%	66%	66%

Table 4.13 The models accuracy of monomer to dimer of group 3 (CpG island).

Negative set	Positive set	Window size	Sn	Sp	Acc	Pre
Random	- 60 ~ + 20	80	74%	68%	71%	70%
Random	-100 ~ + 50	150	75%	69%	72%	71%
Random	-200 ~ +100	300	76%	71%	74%	73%
Random	-300 ~ +150	450	77%	74%	75%	74%
Random	-400 ~ +200	600	80%	75%	77%	76%

Figure 4.8 shows the distribution of predictions by the constructed SVM models (-200 ~ + 100) based on nucleotide composition of group 1, 2, and 3 in the regions of upstream 3,000 bps to downstream 3,000 bps of the TSS. The sliding window size was set to 300 nt which determined by the selective window size (-200 ~ +100) of positive set and shift in the size of 50 bps nt, and the number of predictions were calculated in each window. As it show, the group 1 and 3 have the similar prediction performance, and both of them are better than group 2.

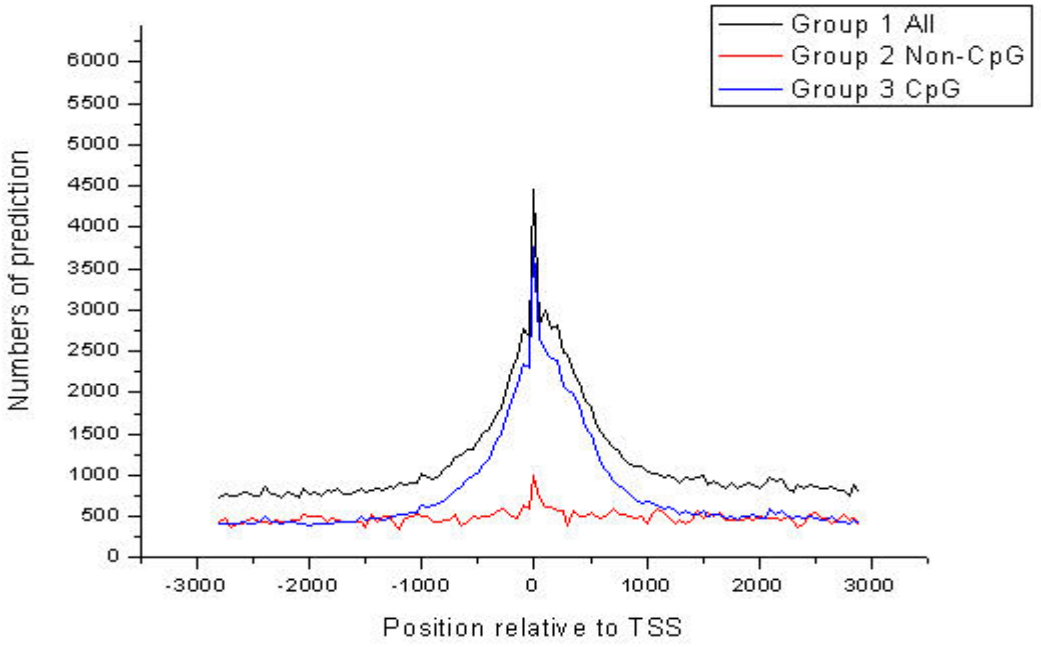


Figure 4.8 Distributions from nucleotide composition models' predictions of group 1, 2, and 3 in the interval [-3000, +3000] relative to the TSS based on DBTSS.

4.2.4 DNA Stability

The prediction sensitivity, specificity, accuracy, and precision of the constructed SVM model based on DNA stability in group 1, 2, and 3 are given in Table 4.14, 4.15, and 4.16, respectively. The negative set is randomly extracted from six negative regions we define in Table 3.3. As you can see, the larger window size is, the higher prediction performance is. Even so, we still choose the models of window size 300 to be our prediction models because of the prediction accuracy of the constructed model is more than 70% and we want to determine the core promoter region accurately. Moreover, the prediction performance of group 3 (CpG) is better than group 1 (all) and 2 (non-CpG).

Table 4.14 The models accuracy of DNA stability in group 1 (all).

Negative set	Positive set	Window size	Sn	Sp	Acc	Pre
Random	- 60 ~ + 20	80	67%	69%	68%	69%
Random	-100 ~ + 50	150	68%	71%	69%	70%
Random	-200 ~ +100	300	70%	74%	71%	73%
Random	-300 ~ +150	450	70%	72%	71%	72%
Random	-400 ~ +200	600	71%	74%	72%	73%

Table 4.15 The models accuracy of DNA stability in group 2 (non-CpG island).

Negative set	Positive set	Window size	Sn	Sp	Acc	Pre
Random	- 60 ~ + 20	80	62%	68%	65%	66%
Random	-100 ~ + 50	150	62%	69%	65%	66%
Random	-200 ~ +100	300	64%	68%	66%	67%
Random	-300 ~ +150	450	65%	69%	66%	67%
Random	-400 ~ +200	600	66%	70%	67%	68%

Table 4.16 The models accuracy of DNA stability in group 3 (CpG island).

Negative set	Positive set	Window size	Sn	Sp	Acc	Pre
Random	- 60 ~ + 20	80	74%	70%	71%	71%
Random	-100 ~ + 50	150	75%	71%	73%	72%
Random	-200 ~ +100	300	76%	72%	74%	73%
Random	-300 ~ +150	450	79%	73%	76%	75%
Random	-400 ~ +200	600	81%	74%	77%	75%

Figure 4.9 shows the distribution of predictions by the constructed SVM models (-200 ~ + 100) based on DNA stability of group 1, 2, and 3 in the regions of upstream 3,000 bps to downstream 3,000 bps of the TSS. The sliding window size was set to 300 nt which determined by the selective window size (-200 ~ +100) of positive set and shift in the size of 50 bps nt, and the number of predictions were calculated in each window. As it show, the group 1 and 3 have the similar

prediction performance, and both of them are better than group 2.

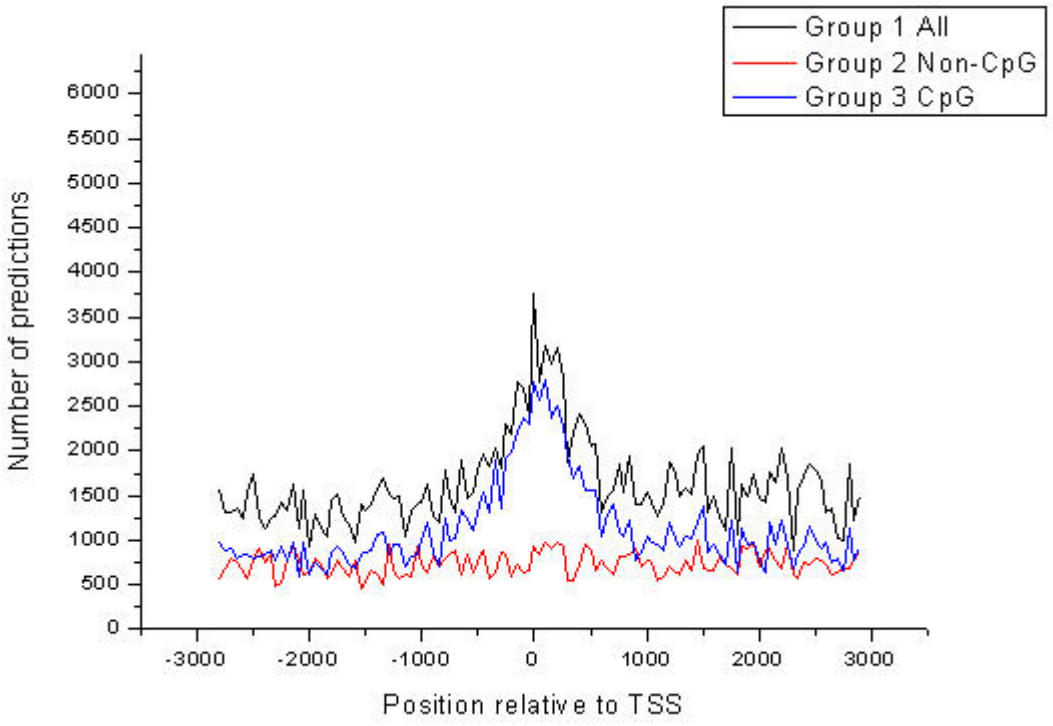


Figure 4.9 Distributions from DNA stability models' predictions of group 1, 2, and 3 in the interval [-3000, +3000] relative to the TSS based on DBTSS.

4.2.5 The Prediction Performance of Combinatorial Features

We try to test all the combinations of the three kinds of regulatory features, and want to find the best combination for increasing the prediction performance. The prediction sensitivity, specificity, accuracy, and precision of the constructed SVM model based on combinatorial models in group 1, 2, and 3 are given in Table 4.17, 4.18, and 4.19, respectively. As you can see, we selected the highest model accuracy models of combination all of three features to be our prediction model. Moreover, the prediction performance of group 3 (CpG island) is better than group

1 (all) and 2 (non-CpG island).

Table 4.17 The model accuracy of Combinational models in group 1 (all).

		Window size 300				
Negative Set	**Positive Set**	**Feature**	**Sn**	**Sp**	**Acc**	**Pre**
Random	-200 ~+100	6m+nc	71%	79%	75%	77%
Random	-200 ~+100	6m+ds	69%	78%	74%	76%
Random	-200 ~+100	nc+ds	74%	76%	75%	75%
Random	-200 ~+100	6m+nc+ds	72%	79%	76%	78%

Table 4.18 The model accuracy of Combinational models in group 2(non-CpG island).

		Window size 300				
Negative Set	**Positive Set**	**feature**	**Sn**	**Sp**	**Acc**	**Pre**
Random	-200 ~+100	6m+nc	64%	70%	67%	68%
Random	-200 ~+100	6m+ds	62%	71%	66%	68%
Random	-200 ~+100	nc+ds	65%	67%	66%	66%
Random	-200 ~+100	6m+nc+ds	65%	71%	68%	69%

Table 4.19 The model accuracy of Combinational models in group 3 (CpG island).

		Window size 300				
Negative Set	**Positive Set**	**feature**	**Sn**	**Sp**	**Acc**	**Pre**
Random	-200 ~+100	6m+nc	81%	79%	80%	79%
Random	-200 ~+100	6m+ds	80%	76%	78%	77%
Random	-200 ~+100	nc+ds	82%	75%	78%	77%
Random	-200 ~+100	6m+nc+ds	81%	78%	80%	79%

Figure 4.10 shows the distribution of predictions by the constructed SVM models (-200 ~ + 100) based on all of those three features of group 1, 2, and 3 in the regions of upstream 3,000 bps to downstream 3,000 bps of the TSS. The sliding window size was set to 300 nt which determined by the selective window

size (-200 ~ +100) of positive set and shift in the size of 50 bps nt, and the number of predictions were calculated in each window. As it show, the group 1 and 3 have the similar prediction performance, and both of them are better than group 2.

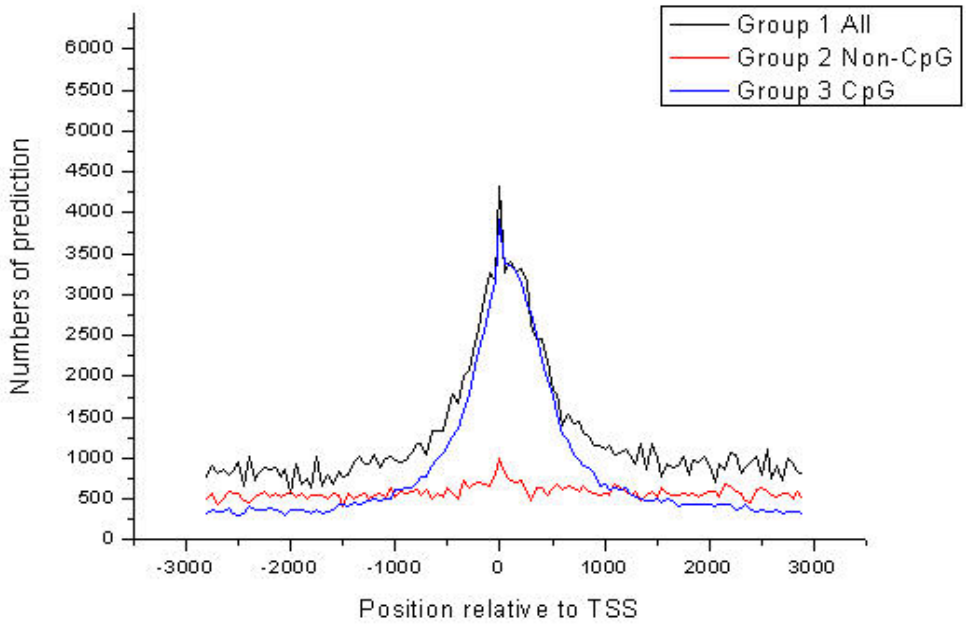


Figure 4.10 Distributions of 6-mer pattern, nucleotide composition, and DNA stability models' predictions of group 1, 2, and 3 in the interval [-3000, +3000] relative to the TSS based on DBTSS.

4.3 Summary of Results

According to the prediction performance of the constructed SVM models described above, it is found that the three models of extracted features have similar accuracy. The model of 6-mer pattern has higher specificity so the numbers of prediction far from real TSS is lower. As shown in Fig. 4.11, the model of nucleotide composition has higher sensitivity so the numbers of prediction of real TSS position is lower.

The prediction accuracy of dataset with CpG island is better than dataset without

CpG island. And by examining all possible combinations of those three models, we demonstrate that such combinations can improve our prediction accuracy.

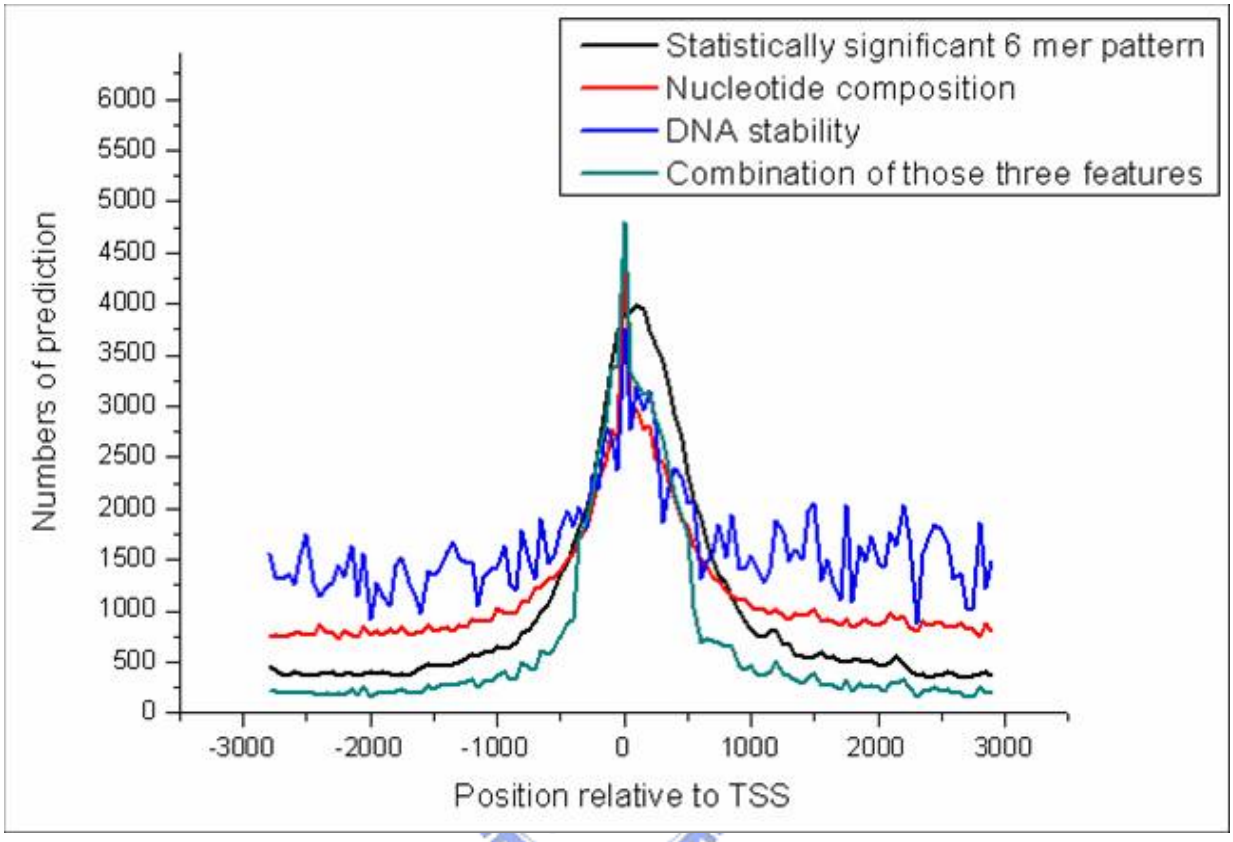


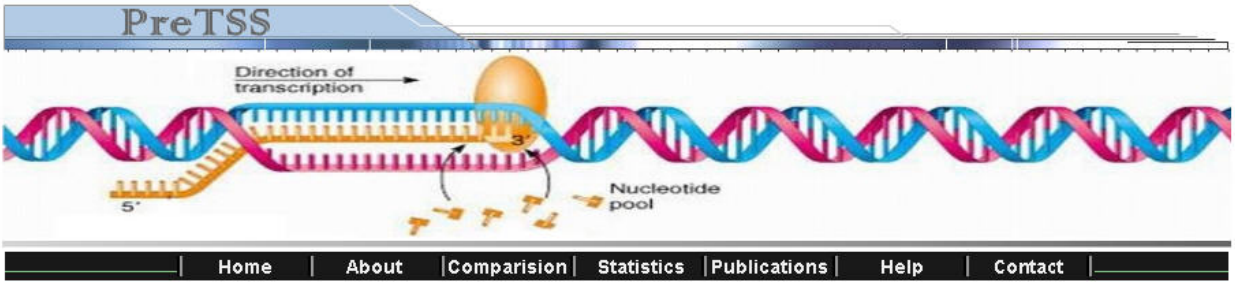
Figure 4.11 The comparison of the prediction performance for the three kinds of feature and Combination of those three features.

4.4 Web Interface

Considering both the prediction performance and core promoter window size, we chose those models with window size 300 bps as the gene promoter identification models. Here we implemented a user friendly interface that can be access from the World Wide Web. This system is implemented on the Linux operation system (Red Hat Enterprise). We use the Apache web server and the PHP4 server side script engine. Those modules are implemented by using PHP program language.

User can input a DNA sequence and specify considering CpG Island or not for the promoter prediction. If users specify two of those features, the combinational models will be use. Here we makes some clustering of the predictions if predictions around 300 nt, and define the midpoint of that as our prediction TSS position. The user web interface is shown in Fig. 4.12 and Fig. 4.13.

Figure 4.12 Web interface [1].



Home | About | Comparision | Statistics | Publications | Help | Contact

Result

Promoter region!

Name	Length	Feature	CpG island	Start pos.	End pos.	TSS pos.
NEW	6000	6m+nc+ds	Yes	2801	3100	3001

Bid Lab, Institute of Bioinformatics, National Chiao Tung University , Taiwan.
Contact us:bryan@mail.nctu.edu.tw with questions or comments

Figure 4.13 Web interface [2].

Chapter 5 Discussion

5.1 Limitations

There are several limitations in the system which we proposed. One of the limitations is that the determined window size of selected model were 300 base pairs, the sequence length inputted from user must be longer than 300 base pairs. The other limitation is that, the proposed method was not implemented as standalone package for user to download. Therefore, users only can use the web-based program to identify the core promoter region in the user input sequence.

5.2 Comparison

There are three promoter prediction tools such as NNPP2.2, McPromoter, and Eponine which could be obtained from the internet, so we compared our method with the three tools. We get 1871 human promoter sequences which length 6000 bps (from -3000 to +3000) form EPD for independence test.

5.2.1 Prediction Accuracy

The comparison of our proposed method with NNPP2.2, McPromoter, and Eponine is shown in Table 5.1. The three promoter prediction tools we obtained from internet were evaluated the prediction performance based on the evaluation

benchmark we proposed. All the combinations of our methods performed better than NNPP2.2, McPromoter, and Eponine. Based on the same evaluation benchmark, NNPP2.2 results in low specificity (45%). Mcpromoter and Eponine result in high specificity 97% and 85%, respectively, but low sensitivity 5% and 25%, respectively. The prediction accuracy of our method is better than the three previous promoter prediction tools.

Maybe someone wonder that does our method perform better than other promoter prediction tools other than NNPP2.2, McPromoter, and Eponine? We could not obtain the other promoter prediction tools, because some of them would not be downloaded for usage in client. Therefore, there is no fair evaluation platform for the other promoter prediction tools such as Promoter 2.0, Dragon GSF, Dragon PF, and so on. However, according to the proposed prediction accuracy on Table 2.1, no promoter prediction tool performs better than our method.

Table 5.1 Comparison of our method with other tools.

		Positive Set	**Sn.**	**Sp.**	**Acc.**	**Pre.**
Our method	6m+nc+ds	92%	64%	78%	72%	92%
Other tools	NNPP2.2	67%	45%	56%	55%	67%
	Mcpromoter 2.0	5%	97%	51%	72%	5%
	Eponine	32%	85%	58%	74%	32%

5.2.2 Characteristics

The study incorporates the powerful computational method with three useful regulatory features of core promoters for gene promoter identification that results in high prediction sensitivity and specificity. Especially the three extracted regulatory features such as statistically significant 6-mer patterns, nucleotide composition, and DNA stability, provided useful discriminating power. Moreover, the incorporated support vector machine also represented the powerful classification ability. Therefore, as shown in Fig. 5.1, the predictions of our method are performed better than other three promoter tools in the regions of upstream 3,000 bps to downstream 3,000 of TSSs.

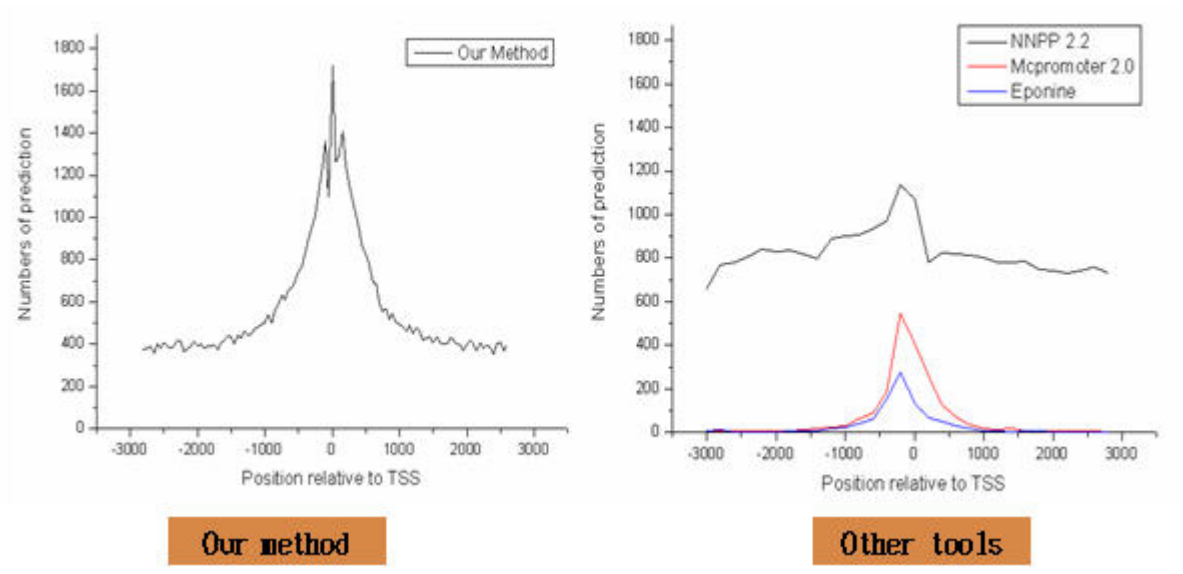


Figure 5.1 Comparison of our method and other tools.

5.3 Future Works

The features we extract in this research are only from one species (human). In

fact, our selected features can be applied to other species. The accuracy of our method is better with smaller DNA sequences. When the evaluated level of DNA sequences region increasing, the accuracy will be down. So it is important for us to improve our method to increase the prediction accuracy in large scale DNA sequences. Now we just consider one promoter sequence containing single TSS. However, we need to consider one promoter sequence containing multiple TSSs in the future.

Chapter 6 Conclusions

This study incorporated the powerful support vector machine with useful regulatory features of core promoters such as statistically significant 6-mer patterns, nucleotide composition, and DNA stability identify the transcriptional start sites in mammalian genomes. By evaluating the prediction performance of the constructed SVM models based on the evaluation benchmark we constructed, our method results in high prediction sensitivity and specificity. The results showed that the accuracy of our method is greater than 70%. Furthermore, the combinatorial SVM model of the statistically significant 6-mer pattern, nucleotide composition, and DNA stability performs better than the individual or pair of them. By comparing our method to other previously proposed gene promoter prediction methods, the performance was also better than others. Therefore, we implement an efficient and effective web interface for guiding biologist to analyze gene promoters and transcriptional start sites.

References

Bajic, V. B. and S. H. Seah (2003). "Dragon gene start finder: an advanced system for finding approximate locations of the start of gene transcriptional units." Genome Res 13(8): 1923-9.

Bajic, V. B., S. H. Seah, et al. (2002). "Dragon Promoter Finder: recognition of vertebrate RNA polymerase II promoters." Bioinformatics 18(1): 198-9.

Bajic, V. B., S. L. Tan, et al. (2004). "Promoter prediction analysis on the whole human genome." Nat Biotechnol 22(11): 1467-73.

Chang, C. C. and C. J. Lin (2001). " LIBSVM: a library for support vector machines."

Davuluri, R. V., I. Grosse, et al. (2001). "Computational identification of promoters and first exons in the human genome." Nat Genet 29(4): 412-7.

Down, T. A. and T. J. Hubbard (2002). "Computational detection and location of transcription start sites in mammalian genomic DNA." Genome Res 12(3): 458-61.

Hsu, C. W., C. C. Chang, et al. "A Practical Guide to Support Vector Classification."

Kanhere, A. and M. Bansal (2005). "A novel method for prokaryotic promoter prediction based on DNA stability." BMC Bioinformatics 6(1): 1.

Kanhere, A. and M. Bansal (2005). "Structural properties of promoters: similarities and differences between prokaryotes and eukaryotes." Nucleic Acids Res 33(10): 3165-75.

Knudsen, S. (1999). "Promoter2.0: for the recognition of PolII promoter sequences." Bioinformatics 15(5): 356-61.

Larsen, F., G. Gundersen, et al. (1992). "CpG islands as gene markers in the human genome." Genomics 13(4): 1095-107.

Ohler, U., G. C. Liao, et al. (2002). "Computational analysis of core promoters in the Drosophila genome." Genome Biol 3(12): RESEARCH0087.

Ohler, U., G. Stemmer, et al. (2000). "Stochastic segment models of eukaryotic promoter regions." Pac Symp Biocomput: 380-91.

Ponger, L. and D. Mouchiroud (2002). "CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences." Bioinformatics 18(4): 631-3.

Prakash, A. and M. Tompa (2005). "Discovery of regulatory elements in vertebrates through comparative genomics." Nat Biotechnol 23(10): 1249-56.

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.

SantaLucia, J., Jr. (1998). "A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics." Proc Natl Acad Sci U S A 95(4): 1460-5.

Scherf, M., A. Klingenhoff, et al. (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.

Schmid, C. D., V. Praz, et al. (2004). "The Eukaryotic Promoter Database EPD: the impact of in silico primer extension." Nucleic Acids Res 32(Database issue): D82-5.

Solovyev, V. V. and I. A. Shahmuradov (2003). "PromH: Promoters identification using orthologous genomic sequences." Nucleic Acids Res 31(13): 3540-5.

Suzuki, Y., R. Yamashita, et al. (2004). "DBTSS, DataBase of Transcriptional Start Sites: progress report 2004." Nucleic Acids Res 32(Database issue): D78-81.

Vapnik, V. N. (1995). The nature of statistical learning theory. New York.

Xing, B. and M. J. van der Laan (2005). "A statistical method for constructing transcriptional regulatory networks using gene expression and sequence data." J Comput Biol 12(2): 229-46.

Appendix A

Table A.1 Top 100 patterns of statistically significant 6-mer pattern in group 1 (all).

SN	pattern	whole	Whole pro.	80num	Pos pro.	OC ratio
1	CGCGCG	93380	0.0000164	1354	0.001396	85.20718
2	GCGCGC	142810	0.0000251	1676	0.001729	68.96482
3	CGCCGC	165495	0.0000290	1916	0.001976	68.03347
4	GCGGCG	165495	0.0000290	1916	0.001976	68.03347
5	CGGCGC	119385	0.0000210	1347	0.001389	66.30266
6	GCGCCG	119385	0.0000210	1347	0.001389	66.30266
7	CGCGGC	123365	0.0000217	1388	0.001432	66.11665
8	GCCGCG	123365	0.0000217	1388	0.001432	66.11665
9	CCGCCG	156119	0.0000274	1558	0.001607	58.64409
10	CGGCGG	156119	0.0000274	1558	0.001607	58.64409
11	CCGGCG	111276	0.0000195	1080	0.001114	57.03421
12	CGCCGG	111276	0.0000195	1080	0.001114	57.03421
13	AGCGCG	72155	0.0000127	683	0.000704	55.62466
14	CGCGCT	72155	0.0000127	683	0.000704	55.62466
15	CCGCGG	129318	0.0000227	1206	0.001244	54.80265
16	CGCGCA	75745	0.0000133	697	0.000719	54.07431
17	TGCGCG	75745	0.0000133	697	0.000719	54.07431
18	CCCGCG	127926	0.0000225	1163	0.001199	53.42372
19	CGCGGG	127926	0.0000225	1163	0.001199	53.42372
20	CGCGAG	63936	0.0000112	562	0.00058	51.65401
21	CTCGCG	63936	0.0000112	562	0.00058	51.65401
22	CGCGGA	61387	0.0000108	525	0.000541	50.25664
23	TCCGCG	61387	0.0000108	525	0.000541	50.25664
24	TCGCGA	29124	0.0000051	238	0.000245	48.0218
25	CGCGAC	35935	0.0000063	289	0.000298	47.25988
26	GTCGCG	35935	0.0000063	289	0.000298	47.25988
27	CGTCGC	44589	0.0000078	329	0.000339	43.35912
28	GCGACG	44589	0.0000078	329	0.000339	43.35912
29	CCGCGC	210901	0.0000370	1521	0.001569	42.38025
30	GCGCGG	210901	0.0000370	1521	0.001569	42.38025
31	CGCCGA	51860	0.0000091	371	0.000383	42.03916

32 TCGGCG	51860	0.0000091	371	0.000383	42.03916
33 CGGCGA	56495	0.0000099	396	0.000408	41.19057
34 TCGCCG	56495	0.0000099	396	0.000408	41.19057
35 CGAGCG	65811	0.0000116	453	0.000467	40.44943
36 CGCTCG	65811	0.0000116	453	0.000467	40.44943
37 CCGCGA	55823	0.0000098	367	0.000379	38.63365
38 TCGCGG	55823	0.0000098	367	0.000379	38.63365
39 CGCGTC	59353	0.0000104	390	0.000402	38.61302
40 GACGCG	59353	0.0000104	390	0.000402	38.61302
41 ACGCGC	66140	0.0000116	428	0.000441	38.02714
42 GCGCGT	66140	0.0000116	428	0.000441	38.02714
43 CGACGC	41872	0.0000074	270	0.000278	37.89237
44 GCGTCG	41872	0.0000074	270	0.000278	37.89237
45 CGACCG	33411	0.0000059	208	0.000215	36.58352
46 CGGTCG	33411	0.0000059	208	0.000215	36.58352
47 CGGCCG	176492	0.0000310	1056	0.001089	35.16022
48 CGCCCC	362006	0.0000635	2130	0.002197	34.57615
49 GGGGCG	362006	0.0000635	2130	0.002197	34.57615
50 ACGCCG	64787	0.0000114	373	0.000385	33.83243
51 CGGCGT	64787	0.0000114	373	0.000385	33.83243
52 GCCGCC	344938	0.0000605	1890	0.001949	32.19831
53 GGCGGC	344938	0.0000605	1890	0.001949	32.19831
54 CCCCGC	426331	0.0000748	2201	0.00227	30.33793
55 GCGGGG	426331	0.0000748	2201	0.00227	30.33793
56 CCGCCC	518352	0.0000910	2647	0.00273	30.00832
57 GGGCGG	518352	0.0000910	2647	0.00273	30.00832
58 GCGGCC	261440	0.0000459	1330	0.001372	29.89456
59 GGCCGC	261440	0.0000459	1330	0.001372	29.89456
60 CCGTCG	44482	0.0000078	226	0.000233	29.85637
61 CGACGG	44482	0.0000078	226	0.000233	29.85637
62 CCGGAA	199323	0.0000350	1000	0.001031	29.48189
63 TTCCGG	199323	0.0000350	1000	0.001031	29.48189
64 CGCACG	77899	0.0000137	381	0.000393	28.74135
65 CGTGCG	77899	0.0000137	381	0.000393	28.74135
66 CGGGGC	371940	0.0000653	1806	0.001863	28.53367
67 GCCCCG	371940	0.0000653	1806	0.001863	28.53367
68 CGCGAA	30243	0.0000053	145	0.00015	28.17444
69 TTCGCG	30243	0.0000053	145	0.00015	28.17444

70 ACGCGG	69430	0.0000122	326	0.000336	27.59189
71 CCGCGT	69430	0.0000122	326	0.000336	27.59189
72 ACGGCG	68155	0.0000120	312	0.000322	26.90103
73 CGCCGT	68155	0.0000120	312	0.000322	26.90103
74 CGCGCC	291748	0.0000512	1310	0.001351	26.38613
75 GGCGCG	291748	0.0000512	1310	0.001351	26.38613
76 CCGACG	45809	0.0000080	200	0.000206	25.6562
77 CGTCGG	45809	0.0000080	200	0.000206	25.6562
78 GCCGGC	246470	0.0000433	1066	0.001099	25.41593
79 CGGAAG	262733	0.0000461	1119	0.001154	25.02812
80 CTTCCG	262733	0.0000461	1119	0.001154	25.02812
81 AACGCG	38488	0.0000068	161	0.000166	24.58179
82 CGCGTT	38488	0.0000068	161	0.000166	24.58179
83 CGGACG	65539	0.0000115	261	0.000269	23.40193
84 CGTCCG	65539	0.0000115	261	0.000269	23.40193
85 GCCGGA	184107	0.0000323	721	0.000744	23.01324
86 TCCGGC	184107	0.0000323	721	0.000744	23.01324
87 CTGCGC	288998	0.0000507	1098	0.001132	22.32648
88 GCGCAG	288998	0.0000507	1098	0.001132	22.32648
89 CACGCG	89405	0.0000157	331	0.000341	21.75602
90 CGCGTG	89405	0.0000157	331	0.000341	21.75602
91 GCGCGA	113177	0.0000199	414	0.000427	21.49582
92 TCGCGC	113177	0.0000199	414	0.000427	21.49582
93 ACGTCG	36991	0.0000065	135	0.000139	21.44626
94 CGACGT	36991	0.0000065	135	0.000139	21.44626
95 GAGCGC	183691	0.0000322	656	0.000677	20.98595
96 GCGCTC	183691	0.0000322	656	0.000677	20.98595
97 ACGCGA	32037	0.0000056	114	0.000118	20.91052
98 TCGCGT	32037	0.0000056	114	0.000118	20.91052
99 TGCGCA	174690	0.0000307	616	0.000635	20.72174
100 CGCGTA	22145	0.0000039	78	8.04E-05	20.69811

Table A.2 Top 100 patterns of statistically significant 6-mer pattern in group 2 (non-CpG island).

SN	pattern	whole	Whole pro.	80num	Pos pro.	OC ratio
1	CGCGAA	30243	0.0000053	6	0.0000255	4.810317
2	TTCGCG	30243	0.0000053	6	0.0000255	4.810317
3	CGCCCC	362006	0.0000635	71	0.0003023	4.759941
4	GGGGCG	362006	0.0000635	71	0.0003023	4.759941
5	CCGCCC	518352	0.0000910	100	0.0004257	4.678166
6	GGGCGG	518352	0.0000910	100	0.0004257	4.678166
7	CCCCGC	426331	0.0000748	82	0.0003491	4.666908
8	GCGGGG	426331	0.0000748	82	0.0003491	4.666908
9	ATCCGG	135765	0.0000238	26	0.0001107	4.650647
10	CCGGAT	135765	0.0000238	26	0.0001107	4.650647
11	GACGTC	138788	0.0000244	26	0.0001107	4.536287
12	CCGCAG	305519	0.0000536	54	0.0002299	4.2889
13	CTGCGG	305519	0.0000536	54	0.0002299	4.2889
14	CGCGAG	63936	0.0000112	11	0.0000468	4.181111
15	CTCGCG	63936	0.0000112	11	0.0000468	4.181111
16	CGGAAG	262733	0.0000461	45	0.0001916	4.155551
17	CTTCCG	262733	0.0000461	45	0.0001916	4.155551
18	ACGTCA	227674	0.0000400	39	0.0001660	4.150702
19	TGACGT	227674	0.0000400	39	0.0001660	4.150702
20	CGTCGA	23424	0.0000041	4	0.0000170	4.143193
21	TCGACG	23424	0.0000041	4	0.0000170	4.143193
22	CCGGAA	199323	0.0000350	34	0.0001447	4.135498
23	TTCCGG	199323	0.0000350	34	0.0001447	4.135498
24	GCCGGC	246470	0.0000433	42	0.0001788	4.129318
25	CGGCAG	310633	0.0000545	52	0.0002214	4.061849
26	CTGCCG	310633	0.0000545	52	0.0002214	4.061849
27	ACTCGC	134642	0.0000236	22	0.0000937	3.968512
28	GCGAGT	134642	0.0000236	22	0.0000937	3.968512
29	CGCGGA	61387	0.0000108	10	0.0000426	3.941788
30	TCCGCG	61387	0.0000108	10	0.0000426	3.941788
31	CCGCGA	55823	0.0000098	9	0.0000383	3.90961
32	TCGCGG	55823	0.0000098	9	0.0000383	3.90961
33	CGGCGA	56495	0.0000099	9	0.0000383	3.862316
34	TCGCCG	56495	0.0000099	9	0.0000383	3.862316
35	CGACTC	145121	0.0000255	23	0.0000979	3.839765

36	GAGTCG	145121	0.0000255	23	0.0000979	3.839765
37	GCGGCA	221533	0.0000389	35	0.0001490	3.830323
38	TGCCGC	221533	0.0000389	35	0.0001490	3.830323
39	CGGCCC	309351	0.0000543	48	0.0002043	3.763209
40	GGGCCG	309351	0.0000543	48	0.0002043	3.763209
41	TCCGCA	176001	0.0000309	27	0.0001149	3.719823
42	TGCGGA	176001	0.0000309	27	0.0001149	3.719823
43	CGGGGC	371940	0.0000653	57	0.0002427	3.716025
44	GCCCCG	371940	0.0000653	57	0.0002427	3.716025
45	GCCGGA	184107	0.0000323	28	0.0001192	3.690392
46	TCCGGC	184107	0.0000323	28	0.0001192	3.690392
47	GCCCGA	172187	0.0000302	26	0.0001107	3.665079
48	TCGGGC	172187	0.0000302	26	0.0001107	3.665079
49	AGGGCG	258763	0.0000454	39	0.0001660	3.657007
50	CGCCCT	258763	0.0000454	39	0.0001660	3.657007
51	CGTCAC	192503	0.0000338	29	0.0001235	3.652568
52	GTGACG	192503	0.0000338	29	0.0001235	3.652568
53	ACCGGA	127794	0.0000224	19	0.0000809	3.610959
54	TCCGGT	127794	0.0000224	19	0.0000809	3.610959
55	GGGCCC	909212	0.0001596	134	0.0005705	3.574887
56	ACGGCC	186522	0.0000327	27	0.0001149	3.515062
57	GGCCGT	186522	0.0000327	27	0.0001149	3.515062
58	GCCGAC	118157	0.0000207	17	0.0000724	3.496194
59	GTCGGC	118157	0.0000207	17	0.0000724	3.496194
60	CGGCAC	202327	0.0000355	29	0.0001235	3.477656
61	GTGCCG	202327	0.0000355	29	0.0001235	3.477656
62	CGCAGA	233353	0.0000410	33	0.0001405	3.426471
63	TCTGCG	233353	0.0000410	33	0.0001405	3.426471
64	TCGGCA	177866	0.0000312	25	0.0001064	3.411162
65	TGCCGA	177866	0.0000312	25	0.0001064	3.411162
66	CCCGGA	256531	0.0000450	36	0.0001533	3.405705
67	TCCGGG	256531	0.0000450	36	0.0001533	3.405705
68	CCGGTC	136422	0.0000239	19	0.0000809	3.38433
69	GACCGG	136422	0.0000239	19	0.0000809	3.38433
70	CCCGCC	746914	0.0001311	104	0.0004427	3.377412
71	GGCGGG	746914	0.0001311	104	0.0004427	3.377412
72	CGGAAC	129310	0.0000227	18	0.0000766	3.375698
73	GTTCCG	129310	0.0000227	18	0.0000766	3.375698

74	CCGGCA	230262	0.0000404	32	0.0001362	3.371985
75	TGCCGG	230262	0.0000404	32	0.0001362	3.371985
76	GCGAAC	79344	0.0000139	11	0.0000468	3.368952
77	GTTCGC	79344	0.0000139	11	0.0000468	3.368952
78	TCGCGA	29124	0.0000051	4	0.0000170	3.332392
79	GGCCCC	1025386	0.0001800	140	0.0005960	3.311782
80	GGGGCC	1025386	0.0001800	140	0.0005960	3.311782
81	GCGGAA	168690	0.0000296	23	0.0000979	3.307906
82	TTCCGC	168690	0.0000296	23	0.0000979	3.307906
83	CGAACG	29492	0.0000052	4	0.0000170	3.28736
84	CGTTCG	29492	0.0000052	4	0.0000170	3.28736
85	CGCGTA	22145	0.0000039	3	0.0000128	3.283134
86	TACGCG	22145	0.0000039	3	0.0000128	3.283134
87	ACGTCG	36991	0.0000065	5	0.0000213	3.279762
88	CGACGT	36991	0.0000065	5	0.0000213	3.279762
89	GCGGAC	110933	0.0000195	15	0.0000639	3.274716
90	GTCCGC	110933	0.0000195	15	0.0000639	3.274716
91	CACCGG	200517	0.0000352	27	0.0001149	3.265413
92	CCGGTG	200517	0.0000352	27	0.0001149	3.265413
93	AGCGTC	171060	0.0000300	23	0.0000979	3.2638
94	GACGCT	171060	0.0000300	23	0.0000979	3.2638
95	AGTCGG	157158	0.0000276	21	0.0000894	3.239121
96	CCGACT	157158	0.0000276	21	0.0000894	3.239121
97	CTCCGA	203551	0.0000357	27	0.0001149	3.219679
98	TCGGAG	203551	0.0000357	27	0.0001149	3.219679
99	ACGCTG	294212	0.0000516	39	0.0001660	3.217599
100	CAGCGT	294212	0.0000516	39	0.0001660	3.217599

Table A.3 Top 100 patterns of statistically significant 6-mer pattern in group 3 (CpG island).

SN	pattern	whole	Whole pro.	80num	Pos pro.	OC ratio
1	CGCGCG	93380	0.0000164	1350	0.001837	112.0417
2	GCGCGC	142810	0.0000251	1674	0.002278	90.77614
3	CGCCGC	165495	0.0000290	1903	0.00259	89.31631
4	GCGGCG	165495	0.0000290	1903	0.00259	89.31631
5	CGGCGC	119385	0.0000210	1338	0.001821	86.7215
6	GCGCCG	119385	0.0000210	1338	0.001821	86.7215
7	CGCGGC	123365	0.0000217	1377	0.001874	86.37025
8	GCCGCG	123365	0.0000217	1377	0.001874	86.37025
9	CCGCCG	156119	0.0000274	1544	0.002102	76.69847
10	CGGCGG	156119	0.0000274	1544	0.002102	76.69847
11	CCGGCG	111276	0.0000195	1067	0.001452	74.47659
12	CGCCGG	111276	0.0000195	1067	0.001452	74.47659
13	AGCGCG	72155	0.0000127	678	0.000923	72.66344
14	CGCGCT	72155	0.0000127	678	0.000923	72.66344
15	CCGCGG	129318	0.0000227	1198	0.001631	71.83249
16	CGCGCA	75745	0.0000133	689	0.000938	70.51111
17	TGCGCG	75745	0.0000133	689	0.000938	70.51111
18	CCCGCG	127926	0.0000225	1155	0.001572	69.86979
19	CGCGGG	127926	0.0000225	1155	0.001572	69.86979
20	CGCGAG	63936	0.0000112	551	0.00075	66.96125
21	CTCGCG	63936	0.0000112	551	0.00075	66.96125
22	CGCGGA	61387	0.0000108	515	0.000701	64.90429
23	TCCGCG	61387	0.0000108	515	0.000701	64.90429
24	TCGCGA	29124	0.0000051	234	0.000318	62.32825
25	CGCGAC	35935	0.0000063	288	0.000392	62.1231
26	GTCGCG	35935	0.0000063	288	0.000392	62.1231
27	CGTCGC	44589	0.0000078	325	0.000442	56.4952
28	GCGACG	44589	0.0000078	325	0.000442	56.4952
29	CCGCGC	210901	0.0000370	1509	0.002054	55.5108
30	GCGCGG	210901	0.0000370	1509	0.002054	55.5108
31	CGCCGA	51860	0.0000091	365	0.000497	54.59356
32	TCGGCG	51860	0.0000091	365	0.000497	54.59356
33	CGGCGA	56495	0.0000099	387	0.000527	53.09936
34	TCGCCG	56495	0.0000099	387	0.000527	53.09936
35	CGAGCG	65811	0.0000116	447	0.000608	52.44928

36	CGCTCG	65811	0.0000116	447	0.000608	52.44928
37	CGCGTC	59353	0.0000104	386	0.000525	50.51774
38	GACGCG	59353	0.0000104	386	0.000525	50.51774
39	ACGCGC	66140	0.0000116	425	0.000578	49.86788
40	GCGCGT	66140	0.0000116	425	0.000578	49.86788
41	CCGCGA	55823	0.0000098	358	0.000487	49.72181
42	TCGCGG	55823	0.0000098	358	0.000487	49.72181
43	CGACGC	41872	0.0000074	266	0.000362	49.25885
44	GCGTCG	41872	0.0000074	266	0.000362	49.25885
45	CGACCG	33411	0.0000059	204	0.000278	47.383
46	CGGTCG	33411	0.0000059	204	0.000278	47.383
47	CGGCCG	176492	0.0000310	1048	0.001426	46.01395
48	CGCCCC	362006	0.0000635	2059	0.002803	44.13393
49	GGGGCG	362006	0.0000635	2059	0.002803	44.13393
50	ACGCCG	64787	0.0000114	368	0.000501	43.93726
51	CGGCGT	64787	0.0000114	368	0.000501	43.93726
52	GCCGCC	344938	0.0000605	1854	0.002523	41.7104
53	GGCGGC	344938	0.0000605	1854	0.002523	41.7104
54	GCGGCC	261440	0.0000459	1306	0.001778	38.72759
55	GGCCGC	261440	0.0000459	1306	0.001778	38.72759
56	CCCCGC	426331	0.0000748	2119	0.002884	38.55843
57	GCGGGG	426331	0.0000748	2119	0.002884	38.55843
58	CCGTCG	44482	0.0000078	221	0.000301	38.51512
59	CGACGG	44482	0.0000078	221	0.000301	38.51512
60	CCGCCC	518352	0.0000910	2547	0.003467	38.09584
61	GGGCGG	518352	0.0000910	2547	0.003467	38.09584
62	CCGGAA	199323	0.0000350	966	0.001315	37.56635
63	TTCCGG	199323	0.0000350	966	0.001315	37.56635
64	CGCACG	77899	0.0000137	376	0.000512	37.35573
65	CGTGCG	77899	0.0000137	376	0.000512	37.35573
66	CGGGGC	371940	0.0000653	1749	0.002381	36.4558
67	GCCCCG	371940	0.0000653	1749	0.002381	36.4558
68	ACGCGG	69430	0.0000122	321	0.000437	35.81254
69	CCGCGT	69430	0.0000122	321	0.000437	35.81254
70	CGCGAA	30243	0.0000053	139	0.000189	35.62954
71	TTCGCG	30243	0.0000053	139	0.000189	35.62954
72	ACGGCG	68155	0.0000120	306	0.000416	34.70804
73	CGCCGT	68155	0.0000120	306	0.000416	34.70804

74 CGCGCC	291748	0.0000512	1303	0.001774	34.63893
75 GGCGCG	291748	0.0000512	1303	0.001774	34.63893
76 CCGACG	45809	0.0000080	194	0.000264	32.84246
77 CGTCGG	45809	0.0000080	194	0.000264	32.84246
78 GCCGGC	246470	0.0000433	1024	0.001394	32.18859
79 AACGCG	38488	0.0000068	159	0.000216	32.06146
80 CGCGTT	38488	0.0000068	159	0.000216	32.06146
81 CGGAAG	262733	0.0000461	1074	0.001462	31.70979
82 CTTCCG	262733	0.0000461	1074	0.001462	31.70979
83 CGGACG	65539	0.0000115	258	0.000351	30.53598
84 CGTCCG	65539	0.0000115	258	0.000351	30.53598
85 GCCGGA	184107	0.0000323	693	0.000943	29.20254
86 TCCGGC	184107	0.0000323	693	0.000943	29.20254
87 CTGCGC	288998	0.0000507	1074	0.001462	28.83276
88 GCGCAG	288998	0.0000507	1074	0.001462	28.83276
89 CACGCG	89405	0.0000157	329	0.000448	28.52241
90 CGCGTG	89405	0.0000157	329	0.000448	28.52241
91 GCGCGA	113177	0.0000199	410	0.000558	28.04276
92 TCGCGC	113177	0.0000199	410	0.000558	28.04276
93 ACGTCG	36991	0.0000065	130	0.000177	27.26394
94 CGACGT	36991	0.0000065	130	0.000177	27.26394
95 GAGCGC	183691	0.0000322	642	0.000874	27.13746
96 GCGCTC	183691	0.0000322	642	0.000874	27.13746
97 ACGCGA	32037	0.0000056	111	0.000151	26.88293
98 TCGCGT	32037	0.0000056	111	0.000151	26.88293
99 TGCGCA	174690	0.0000307	604	0.000822	26.77864
100 CGTACG	22916	0.0000040	78	0.000106	26.4094