CHAPTER V

EXPERIMENTS AND RESULTS

5.1 Sequence Data Sets

5.1.1 Prokaryote Data Set

The *Escherichia Coli* (E.coli) representing prokaryote data set was taken from the UCI Machine Learning Repository. It consisted of 106 sequences, including 53 promoter sequences, which were used as positive data and 53 non-promoter sequences as negative data. All sequences are 57 bp long. Each promoter sequences contains 49 bp upstream and 7 bp downstream of the transcription start site (TSS).

5.1.2 Eukaryote Training Sets

The experimental data sets are taken from two reliable source of promoter sequences, namely, Berkeley Drosophila Genome Project (BDGP) and the Eukaryotic Promoter Database (EPD) available at http://www.epd.isb-sib.ch. Each data set is composed of promoters representing positive training sets and non-promoters sequences (CDS sequences, and non-coding) representing negative training sets. All sequences are 300 bp long. The promoter contains 250 bp upstream and 50 bp downstream of the TSS (Table 5.1).

Table 5.1: **Amount of available Eukaryote Training sets**(No.sequences)

Set	Organism	promoter	CDS	intron
Training Set 1	*human* (BDGP)	565	890	-
Training Set 2	*human* (EPD)	1,871	890	4,345
Training Set 3	*D.Melanogaster* (BDGP)	74,100	389,700	108,900

Table 5.2: **Amount of available test set** (bp) for evaluating performance of predictions.

Set	Length (bp)	No.promoter
Test set A	33,120	24
Test set B		
*AC*002397	227,538	17
*L*44140	219,447	11
*D*87675	301,692	1
*AF*017257	101,569	1
*AF*146793	204,625	4
*AC*002368	324,816	1
Test set C	35,000,000	337
Test set D	853,180	92

5.1.3 Evaluation sets in large genome sequences

For comparing with existing methods, three large genome sequences test sets were proposed (Table 5.2). Test set A was taken from Fickett&Hatzigeorgiou [1], Test set B consisted of six Genbank genomic sequences with a total length of 1.38 Mb and 35 known TSSs on these sequences, Test set C consisted of publicly available sequence for human chromosome 22. The annotation data (Rel.2.3) for human chromosome 22 were produced by the Chromosome 22 Gene Annotation Group at the Sanger Center and were obtained from the http://www.sanger.ac.uk/HGP/Chr22/. Test set D is the sequences of D.melanogaster taken from BDGP.

5.2 Performance Evaluation of Promoter Recognition

Prediction performance is determined by measuring the precision, specificity (sp) and sensitivity (sn) (Figure 5.1). These are defined as Eqs. (5.1)-(5.3), respectively.

$$Precision = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.1}$$

$$Specificity = \frac{TP}{TP + FP} \tag{5.2}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{5.3}$$

where TP, TN, FP, FN are the number of true positives, true negatives, false positives and false negatives, respectively. A *true positive* is the number of promoters correctly predicted. A *false positive* is the number of non-promoters recognized as being promoters. A *false negative* is the number of promoters recognized as being non-promoters. A *true negative* is the number of non-promoters correctly predicted.

Figure 5.1: Performance Evaluation of Promoter Recognition.

5.3 The Results on *E.coli* Data Sets

In this dissertation, our method is compared with Huang and Wang [8]. Leave-one-out method is used for evaluating the performance and comparing the result with several other promoter prediction systems. The leave-one-out method is simply an N-fold cross-validation, where N is the number of samples in the data set (N=106). Each sample was left out in turn, and the learning scheme was trained on the entire remaining samples (105). The procedure was repeated 106 times, so that each sample was excluded once from training set. I used the same data sets and the same evaluation method as [8]. In the *E.coli* data sets, there were 16 input units selected from 64 units (word length 3 (W_3)), 10 hidden units and one output units. Table 5.2 shows the number of errors compared with other methods, including our method. The first number denotes the number of error patterns and the second one is the total number of test patterns. Table 5.3 compares our method with [8] by using precision, specificity and sensitivity criteria. It is clear that our method performs better than the others.

5.4 The results on Eukaryote Training Set

5.4.1 The results on Training Set 1

In data set 1, I used 5-fold cross-validation and the independent test data. For 5-fold cross-validation test, the training data are divided into 5 equal parts. Of these 5 parts, 4 parts are used for training and the fifth is used for testing. This is done repeatly 5

Table 5.3: Number of errors on several methods

Methods	Errors	Methods	Errors
ID3[27]	19/106	BP	8/106
C4.5[28]	18/106	NTSS[30]	7/106
GBI[6]	16/106	HM-layer	6/106
KNN	13/106	KBANN	4/106
O'Neill[29]	12/106	HSVM[8]	3/106
FTSS[30]	9/106	Our method	0/106

Table 5.4: Performances evaluation using precision, specificity and sensitivity

	HSVM[8]	Our method
Precision	97.2%	100%
Specificity	96.2%	100%
Sensitivity	98.1%	100%

times for all 5 parts. Prediction performance is obtained by averaging the results over five tests. In training set 1, there are 87 input units: 20 units from word length 3(W_3), 65 units from word length 4 (W_4), 2 units from CpG island features, 60 hidden units, and one output units. The results of 5-fold cross-validation are shown in Table 5.5. The

Table 5.5: The results on Training Set 1.

Test	TP	TN	FP	FN	Sensitivity	Specificity
1	93	155	23	20	0.823	0.838
2	90	167	11	23	0.796	0.891
3	92	168	10	21	0.814	0.902
4	94	162	16	19	0.832	0.848
5	91	165	13	22	0.805	0.875

results on Training Set 1 in Table 5.6 was compared with Daniel and Karl [20] by using the same data set and evaluation method.

Table 5.6: **Comparing result on Data Set 1.**

Method	Sensitivity(%)	Specificity(%)
Daniel and Karl [20]	79	86
Our method	81	87

5.4.2 The results on Training Set 2

In training set 2, there are 122 input units: 25 units from word length 3(W_3), 50 units from word length 4 (W_4), 45 units from word length 5(W_5), 2 units from CpG island features, 85 hidden units, and one output units. It used as training set of every test sets.

5.4.3 The results on Training Set 3

In training set 3, there are 103 input units: 32 units from word length 3(W_3), 48 units from word length 4 (W_4), 23 units from word length 5(W_5). There are 600 hidden units, and one output units. It used as training set of Test set D.

5.5 The results on Evaluation sets in large genome sequences

The experiment used data set 2 as trining set for evaluating large genome sequences. Results on Fickett&Hatzigeorgiou [1] are presented in Table 5.7. Table 5.8 and 5.9 show the results from Test set B. Table 5.10 shows the result from Test set C. Table 5.11 shows the result from Test set D.

Table 5.7: Comparing results on Test set A.

Method	TP	FP	Sensitivity	Specificity
PromFind [13]	11	24	0.46	0.31
NNPP2.1 [15]	14	59	0.58	0.20
TSSG [14]	10	17	0.42	0.37
TSSW [14]	14	33	0.58	0.30
PromoterInspector [16]	7	3	0.29	0.70
DPF1.4 [19]	10	19	0.42	0.34
PromSearch [31]	8	9	0.33	0.47
Our method	13	15	0.54	0.46

Table 5.8: **The results on Test set B.**

Accession number	No.promoter	TP	FP	Sensitivity	Specificity
*AC*002397	17	5	8	0.30	0.38
*L*44140	11	6	18	0.55	0.25
*D*87675	1	1	2	1	0.33
*AF*017257	1	1	0	1	1
*AF*146793	4	2	3	0.5	0.40
*AC*002368	1	1	1	1	0.50

Table 5.9: **Comparing results on Test set B.**

Method	TP	FP	Sensitivity(%)	Specificity(%)
TSSG [14]	15	449	43	3.3
TSSW [14]	15	501	43	2.9
NNPP2.1 [15]	23	3,533	66	0.6
Promoter2.0	8	1,751	15	0.4
PromoterInspector [16]	15	19	43.0	43.0
PromPredictor [21]	18	43	51.4	29.5
Our method	16	32	45.0	33.3

Table 5.10: **Comparing result on Test set C.**

Method	Sensitivity(%)	Specificity(%)
PromoterInspector	45	33
DPF	64	33
PromPredictor	66	48
Our method	64	55

Table 5.11: **Comparing result on Test set D.**

Method	Sensitivity(%)	Specificity(%)
McPromoter	52.1	40.3
Our method(T2)	59.2	48.7
Our method(T3)	61.4	50.6