

## Genome analysis

## Improved prediction of bacterial transcription start sites

J. J. Gordon<sup>1</sup>, M. W. Towsey<sup>1,\*</sup>, J. M. Hogan<sup>1</sup>, S. A. Mathews<sup>2</sup> and P. Timms<sup>2</sup><sup>1</sup>Faculty of Information Technology and <sup>2</sup>School of Life Sciences, Queensland University of Technology, GPO Box 2434, Brisbane, QLD 4001, Australia

Received on September 9, 2005; revised on November 3, 2005; accepted on November 7, 2005

Advance Access publication November 15, 2005

Associate Editor: T Charlie Hodgman

## ABSTRACT

**Motivation:** Identifying bacterial promoters is an important step towards understanding gene regulation. In this paper, we address the problem of predicting the location of promoters and their transcription start sites (TSSs) in *Escherichia coli*. The accepted method for this problem is to use position weight matrices (PWMs), which define conserved motifs at the sigma-factor binding site. However this method is known to result in large numbers of false positive predictions.

**Results:** Our approaches to TSS prediction are based upon an ensemble of support vector machines (SVMs) employing a variant of the mismatch string kernel. This classifier is subsequently combined with a PWM and a model based on distribution of distances from TSS to gene start. We investigate the effect of different scoring techniques and quantify performance using area under a detection-error tradeoff curve. When tested on a biologically realistic task, our method provides performance comparable with or superior to the best reported for this task. False positives are significantly reduced, an improvement of great significance to biologists.

**Availability:** The trained ensemble-SVM model with instructions on usage can be downloaded from <http://eresearch.fit.qut.edu.au/downloads>

**Contact:** m.towsey@qut.edu.au

## 1 INTRODUCTION

The first step in the initiation of bacterial gene transcription requires an RNA polymerase (RNAP)/sigma factor complex to bind a promoter (Lewin, 1985). Identification of promoters is crucial in the study of gene regulation but they are difficult to find because they lie at a variable distance upstream of their associated genes and because the DNA sequences of known promoters are poorly conserved. Promoters do, however, lie in a well-defined window upstream of the gene transcription start sites (TSSs). Knowing a TSS location, one can predict the promoter location to within a few base pairs (bp) and vice versa. We use the term TSS prediction to refer more generally to this joint identification of TSS and promoter.

This paper describes the use of the support vector machine (SVM) (Vapnik, 1995) to predict TSS locations. We consider TSSs for the major class of *Escherichia coli* promoters bound by sigma-70 ( $\sigma^{70}$ ).  $\sigma^{70}$  binding sites consist of paired hexamers located close to the  $-10$  and  $-35$  positions with respect to the TSS.

The accepted method of finding  $\sigma^{70}$  promoters is to use paired position weight matrices (PWMs) to identify the  $-35$  and  $-10$  motifs, with an additional score or penalty depending on the gap

between them (Stormo, 2000; Huerta and Collado-Vides, 2003). Using information theoretic reasoning, it can be shown that the mapped  $-35$  and  $-10$  hexamers are insufficiently conserved to identify all the expected promoters in the background genome (Schneider *et al.*, 1986).

A  $\sigma^{70}$  promoter can be surrounded by other regulatory sites, including upstream elements (Gourse *et al.*, 2000) and activator and repressor binding sites. The use of machine learning techniques should achieve better TSS prediction by exploiting this expanded set of patterns in the neighbourhood of the promoter.

The SVM is a highly successful supervised learning algorithm that determines the maximum-margin hyperplane between two classes of training examples. When applied to TSS prediction, success depends on an appropriate choice of positive and negative training sequences and on the sequence representation or kernel. Gordon and Towsey (2005) report an SVM method that uses a variant of the mismatch string kernel (Leslie *et al.*, 2004). It significantly outperformed a standard PWM approach on a realistic TSS prediction task when coding sequences were used as training negatives. The work presented here describes two new SVM approaches, an ensemble-SVM and a committee-SVM, both of which yield increased TSS prediction accuracy on the same task. We also describe a segment scoring method, which further reduces the rate of false positive predictions.

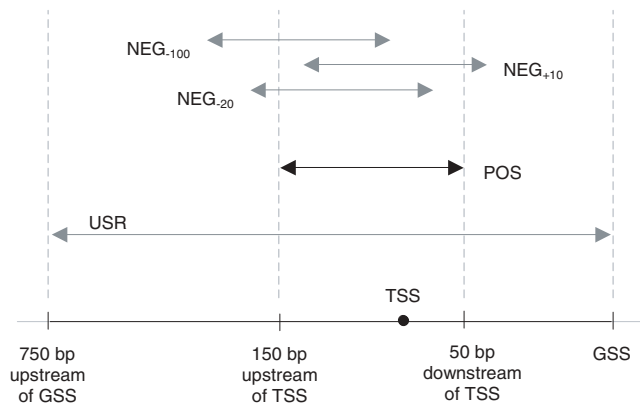
## 2 DATA

We obtained TSS data from the RegulonDB database (Salgado *et al.*, 2001), which contains 676 mapped  $\sigma^{70}$  TSS locations. We extracted sequences from the *E.coli* K12 genome ([www.genome.wisc.edu](http://www.genome.wisc.edu)) and constructed several distinct datasets. The primary dataset consisted of 450 non-overlapping sequences, each extending 750 bp upstream from a gene start codon and each containing exactly one mapped TSS from RegulonDB. These sequences are referred to as gene upstream regions (USRs). Only 450 of the 676 known TSS locations allowed the extraction of a non-overlapped USR containing exactly one known TSS. The TSSs were located at variable positions within these USRs but predominantly near the gene starts (see Section 5). USRs were used to test all methods on a biologically realistic TSS prediction task.

All individual SVMs were trained using 450 positive and 450 negative sequences, each 200 bases long. The positive sequences contained a mapped TSS at position 151. That is, the sequences extended from  $-150$  to  $+50$  bases relative to the TSS.<sup>1</sup> The 450 TSSs

<sup>1</sup>According to biological convention, the TSS position is denoted by  $+1$ . The position immediately upstream is  $-1$ . There is no 0 position.

\*To whom correspondence should be addressed.



**Fig. 1.** A gene upstream region (USR) illustrating the locations of its corresponding positive and negative training sequences. A USR extends 750 bp upstream of the Gene Start Site (GSS). The positive sequence spans the  $[-150, +50]$  neighbourhood around the TSS. Negative sequences are offset from the positive sequence by a designated number of bases. Note that depending on the position of the TSS within the USR, the corresponding positive and negative sequences could extend beyond either end of the USR.

used for the positive sequences were the same as those in the USR sequences.

A total of 40 sets of negative sequences were prepared. These were extracted from the genome at locations offset from the positive sequences by a designated number of base pairs. For example, we use  $NEG_{+25}$  to denote a negative sequence that is shifted with respect to the corresponding positive sequence by +25 bp. Similarly, we use  $\{NEG_{+25}\}$  to denote the set of 450 negative sequences offset from their corresponding positive sequences by +25 bp. As explained in Section 4, our ensemble-SVM approach employed multiple negative sequence sets  $\{NEG_N\}$ , where the offset,  $N$ , took values from  $-150$  to  $+50$  in steps of 5 (excluding zero, which is the set of positive sequences) (Fig. 1).

Position 151 in each training sequence is referred to as the reference position. In the case of positive sequences, this is the TSS position. The 200 bp surrounding the reference position constitute a TSS neighbourhood, within which an SVM searches for information that can be used to classify the reference position either as a TSS or not as one.

### 3 METHODS

#### 3.1 PWM approach

In order to compare our ensemble-SVM method with a standard PWM approach, it was necessary to prepare two PWMs describing the  $-35$  and  $-10$  hexamers, respectively. We assumed that the consensus motifs, TTGACA and TATAAT were known. The first step was to find the best match to the consensus hexamers in a region upstream of each TSS. The 3' end of the best fit TATAAT-like motif was constrained to occur in positions  $[-14, -4]$ . The gap between the hexamers was constrained to the range  $[14, 20]$ .

Each candidate hexamer-pair within these constraints was assigned a score equal to the number of bases matching the consensus plus a weighting to give preference to gaps in the centre of the  $[14, 20]$  range. For each TSS, the hexamer-pair with the highest

score was selected. A PWM derived from the  $-35$  motifs and another derived from the  $-10$  motifs could be constructed for any subset of TSSs, using background nucleotide frequencies sampled from the USRs (Stormo, 2000).

#### 3.2 Ensemble-SVM approach

Our sequence representation was a modification of the mismatch string kernel described by Leslie *et al.* (2004), with strings of length 5 and one mismatch. We incorporated two modifications: (1) Each feature was a 5mer tagged with its location with respect to the TSS. (2) Potential input features having low discriminative value were removed. These two modifications are now described in more detail.

Each 5mer was tagged with the distance of its 5' end from the sequence reference position, rounded to the nearest multiple of 5. Consequently, there were 40 960 potential features ( $4^5$  5mers  $\times$  40 locations). Rounding tag distances accommodates the flexibility of motif locations. The width of the tag window is an important parameter but proved not to be critical in the range 5–10. In the previous work (Gordon and Towsey, 2005) we used a tag window of 10. In this work, a value of 5 was found to yield slightly better results.

The discriminative value of a feature was determined by its symmetric uncertainty (Liu and Wong, 2003), an information theoretic measure derived from counts of the feature in the positive and negative training data. The count of a feature in a set of sequences also included counts of 5mers at the same location differing by a single mismatch.

The features were then ranked in order of decreasing symmetric uncertainty and the list pruned (starting at the top), by eliminating a feature if there was one higher in the list at the same location differing by a single mismatch. The resulting list was truncated to 200 entries. The purpose of this feature pruning step was to select features that were likely to be centers of mismatch neighbourhoods and were likely to have high discriminative value.

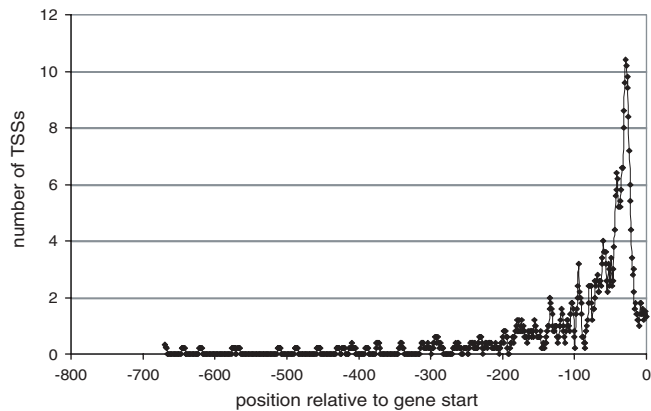
Once the feature list for any particular training set was determined, individual sequences were encoded by weighting the count of the feature by its symmetric uncertainty. Input vectors were normalized to unit length. SVM models were generated using either SVM-Light (Joachims, 1999) or the GPDT package (Serrafini *et al.*, 2004, <http://dm.unife.it/gpdt/>).

The above procedure was used to generate 40 different SVM models, each one trained with the same set of positive sequences but a different set of negatives. The 40 SVMs were treated as an ensemble, with their scores being averaged to give a final score or classification. Ensemble classifiers can achieve greater accuracy by averaging errors in the individual models (Duda *et al.*, 2001).

#### 3.3 DGS approach

For the 450 *E. coli*  $\sigma^{70}$  TSSs employed in this study, Figure 2 shows the distribution of the TSS distance to gene start (DGS). The majority of TSSs occur in the first 100 bp upstream of the gene start, with very few occurring more than 400 bp upstream.

This property was used by Burden *et al.* (2005) to improve promoter recognition using a neural network (NN). The NN score was multiplied by the probability of finding a promoter at the relevant distance from the gene start. In this paper we adopt a similar strategy, using the DGS distribution to modulate PWM and



**Fig. 2.** Distribution of TSS distance to gene start. (Smoothed using a moving average filter, window size = 5.)

SVM scores. However, we also evaluate the DGS distribution as a standalone method of promoter prediction. In this approach, the score at a particular position is simply the DGS probability. Because the DGS distribution does not follow any of the standard functions (Burden *et al.*, 2005), we estimated the distribution empirically from the available data as shown in Figure 2.

### 3.4 Experimental protocol

All methods were evaluated using 10-fold cross-validation. To this end, a unique index from 1 to 450 was assigned to each USR and its corresponding positive and negative sequences. That is, the USR, the positive and the 40 negative sequences associated with the same TSS were assigned the same index. Based on the index, the datasets were divided into 10 equal parts, each of which was successively held in reserve as a test set, while the remaining 90% was used to generate the PWM, SVM and DGS models. The models were then evaluated on the 10% of the data held in reserve. We report performance results for the biologically realistic task of locating the TSS in USR sequences. SVM generalization error (on datasets with equal numbers of positive and negative sequences) is quoted where relevant, although it is not a reliable indicator of performance on the task of TSS prediction in USRs (Gordon and Towsey, 2005).

Performance was measured using detection-error tradeoff (DET) curves. Each model returned a score for each of the 750 positions in all 450 USR sequences. In the case of the PWM method, the score was the highest that could be obtained from two upstream hexamers whose locations were constrained as described in Section 3. In the case of the ensemble-SVM, the score was the perpendicular distance from the decision plane averaged over the 40 component classifiers. In the case of the DGS model, the score was the probability of a TSS occurring at that position as given by the DGS distribution derived from the training data.

Next a threshold,  $T$ , was defined. Mapped or known TSS positions in the USRs were scored as true positives (TPs) if the model score exceeded  $T$  and false negatives (FNs) if below  $T$ . All other positions in the USRs were scored as false positives (FPs) if the score exceeded  $T$  and true negatives (TNs) if below  $T$ .

By varying  $T$  over the range of scores, it was possible to construct DET curves. These are plots of false negative rate (FNR) versus

false positive rate (FPR), where  $FNR = FN/(FN + TP)$  and  $FPR = FP/(FP + TN)$ . The area under a DET curve (Fig. 6) is a measure of the classifier's ability to correctly identify TSS positions over the full range of thresholds: the lower the area under the curve, the lower the overall prediction error. DET area constitutes a single rigorous and objective measure of overall classifier performance, preferable to quoting tables of statistics for various threshold values. Note that a DET curve is similar to an ROC curve except that the latter plots sensitivity versus FPR. By definition, sensitivity or recall =  $1 - FNR$  and specificity =  $1 - FPR$ . Therefore it is easy to calculate sensitivity and specificity for any point (threshold value) on a DET curve in Figure 6.

A variation on the above scoring method was to divide the USRs into non-overlapping segments of length  $N$ , where  $N$  took values 5, 10 and 20. Scoring was then performed on segments rather than individual positions, enabling DET curves to be generated as above. The score assigned to a segment was the maximum position score within the segment. (A scoring segment width of 1 corresponds to scoring individual positions.)

Use of segment scoring was motivated by the fact that in a biologically realistic prediction task, one does not need to identify exact TSS positions in order to motivate a laboratory search. Rather, it may be sufficient to locate TSSs to within a segment width.<sup>2</sup> As shown below, sacrificing a small amount of prediction resolution has the advantage of substantially reducing the number of FP predictions.

## 4 RESULTS

### 4.1 Simple models

For the three methods described above, Table 1 gives DET areas (i.e. areas under DET curves) for scoring segment sizes of 1, 5, 10 and 20. The DET area for the standard PWM method, which assigns a score to every position, is 0.30. This is a useful benchmark result.

The best result (DET area = 0.09) was obtained with the ensemble-SVM and scoring segment size of 5. However for larger scoring segment sizes, performance of the ensemble-SVM declines and all three methods perform similarly.

Perhaps the main surprise in Table 1 is that the simple DGS method performs reasonable well, although it utilizes nothing more than the distribution of DGSs. In fact, it outperforms both methods for a segment size of 20. In order to interpret this result, we repeated the same experiments but with USR sequences truncated to 500, 200 and 100 bp upstream of the GSS.

When confined to shorter USRs, the performance of the DGS method is significantly worse than that of the PWM and SVM methods (right column, Table 2).<sup>3</sup> In summary, the PWM and SVM methods make a higher proportion of FP predictions far from the GSS, while the DGS method makes a higher proportion of FP predictions closer to the GSS.

<sup>2</sup>To verify a TSS in the lab, biologists use primer extension 5' RACE or S1 nuclease mapping to locate the 5' end of mRNA starting from a known downstream position (Sambrook *et al.*, 1989). These techniques locate the TSS to within a few basepairs.

<sup>3</sup>Note that the DET areas in Table 2 ignore positions, including TSSs, outside the USRs. For example, for a USR size of 500, a TSS occurring 520 bp upstream of the gene start is not counted as a FN. The percentage of excluded TSSs can be gauged from Figure 2.

**Table 1.** Area under the DET curves for three TSS prediction models and for four different sizes of scoring segment

Segment size	1	5	10	20
DGS	0.13	0.12	0.12	0.11
PWM	0.30	0.15	0.12	0.13
Ensemble-SVM	0.10	0.09	0.11	0.14

The areas are averages over 10-fold cross-validation. Standard deviations (omitted for clarity) are less than 0.035 in all cases.

**Table 2.** Area under the DET curves for three TSS prediction models and for four lengths of USR sequence

USR size	750	500	200	100
DGS	0.12	0.17	0.28	0.31
PWM	0.15	0.16	0.18	0.18
Ensemble-SVM	0.09	0.10	0.11	0.13

The areas are averages over 10-fold cross-validation. Scoring segment size is 5. Standard deviations (omitted for clarity) are close to 0.03.

Despite its simplicity, the DGS method provides another interesting benchmark against which to compare more sophisticated methods. One might be inclined to question the value of more sophisticated methods if they cannot outperform a simple DGS approach. We note that Burden *et al.* (2005), who used DGS information in conjunction with an NN, did not investigate DGS information by itself. Including DGS information in a TSS prediction task may be useful where one wants to scan an entire genome and the GSSs are known. It will not be useful where one wants to search specific parts of a genome more than 200 bp upstream of the concerned GSS. In these regions the DGS probability curve is essentially flat.

While DET area offers a single value to summarize the overall performance of a prediction method, biologists are more interested in performance at special points on the DET curve. For example, if a laboratory is prepared to investigate 100 *in silico* TSS predictions, it would be useful to know the expected number of TPs in the top 100 predictions. The FP/TP ratio at 90% recall (sensitivity) is an alternative statistic from the other end of the DET curve.

Table 3 gives the FP/TP ratios at a threshold, which yields 90% recall. Table 4 gives the number of TPs in the top 100 predictions (actually a combination of the top 10 predictions from each of the 10 folds). On the basis of these figures, the ensemble-SVM performs better than the PWM method for scoring segment sizes of 1 and 5 but the difference is not significant for sizes 10 and 20. The dominant observation is that increasing the length of the scoring segment greatly reduces the rate of FP predictions. The reduction is most dramatic when the width is increased from 1 to 5. The tradeoff is that TSS predictions are correspondingly less precise. However, as noted above, it may be sufficient to locate TSSs to within a segment, rather than to a precise position.

## 4.2 Ensemble learning

Ensemble learning and segment scoring are the two novel components in our ensemble-SVM prediction algorithm. It is instructive to

**Table 3.** FP/TP ratios at the 10% FNR (sensitivity = 90%) threshold for four scoring segment sizes

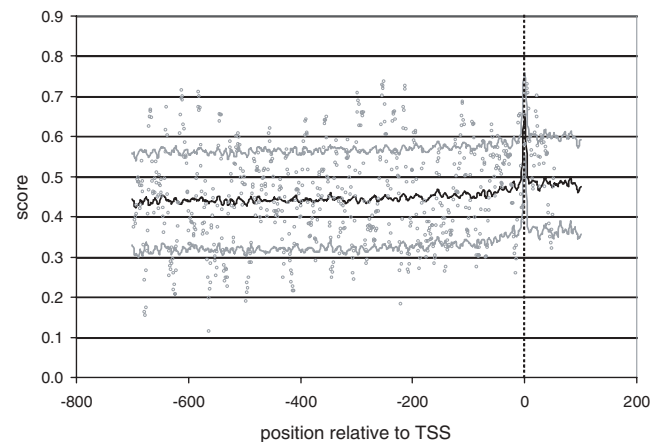
Segment size	1	5	10	20
DGS	300 ± 75	55 ± 15	24 ± 8.3	12 ± 4.8
PWM	560 ± 67	81 ± 16	30 ± 10	15 ± 5.4
Ensemble SVM	270 ± 56	45 ± 10	26 ± 5.7	15 ± 2.0

Ratios are given to two significant figures. Standard deviations obtained from 10-fold cross/validation.

**Table 4.** TP count in the top 100 predictions for four scoring segment sizes

Segment size	1	5	10	20
DGS	1.1 ± 1.5	9.1 ± 3.7	20 ± 4.1	30 ± 4.9
PWM	5.0 ± 5.0	34 ± 9.1	41 ± 9.7	41 ± 9.7
Ensemble SVM	9.8 ± 7.6	35 ± 8.3	49 ± 7.6	48 ± 10

Values are given to two significant figures. Standard deviations obtained from 10-fold cross/validation.

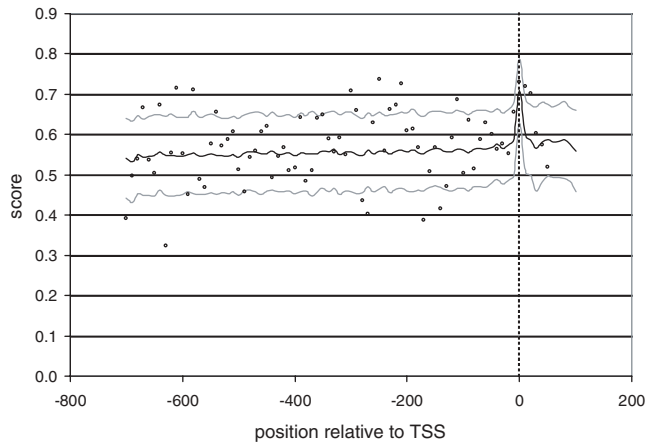
**Fig. 3.** SVM scores for a single USR (circles) for a scoring segment size of 1, superimposed on the average for 450 USRs.

look at the contribution of each. The best result (lowest DET area) reported by Gordon and Towsey (2005) for a single SVM, using the same USR dataset and the same experimental protocol, was  $0.18 \pm 0.02$ . The best result in this work is  $0.10 \pm 0.03$  (for the equivalent scoring segment size of 1, Table 1). We may attribute this improvement in performance to ensemble learning and its ability to smooth the prediction noise of individual models.

This improvement in performance can be used to justify the extra computational expense: a significant number of models must be trained if smoothing is to be achieved.

## 4.3 Segment scoring

With regard to the effect of scoring segments, we observe in Table 1 (row 3) that increasing segment size does not reduce the DET area for the ensemble-SVM method. However it does greatly reduce the



**Fig. 4.** SVM scores for a single USR (circles) for a scoring segment size of 10, superimposed on the average for 450 USRs.

FPR (Table 3), which from the biologists' point of view is a very important result. This is in part elementary—in a sequence of fixed length, the number of scoring segments decreases inversely with segment size. There remains one TSS but the number of potential FPs must decrease with increasing segment size. The simplicity of this result does not diminish its usefulness in the practical task of promoter prediction.

There is an additional reason why segment scoring is effective. When the ensemble-SVM scores are averaged over the 450 USR sequences aligned to their TSSs, there is a well-defined peak at the TSS location (Fig. 3). The grey lines in Figure 3 represent the mean score  $\pm 1$  SD at the position. On average, the SVM model is correctly identifying the TSS position.

However, the standard deviation of the scores is large relative to the peak at the TSS location. This point is illustrated in Figure 3 by the superimposed ensemble-SVM scores for a single USR (grey dots). The plotted USR is for the gene *mali*. This USR extends from  $-697$  to  $+53$  relative to the TSS. The large variation of scores highlights the difficulty of predicting the TSS. The corresponding plot using a scoring segment length of 10 is shown in Figure 4, there being one-tenth the number of scores. But observe in Figure 3 that 72 out of 750 (17.6%) individual position scores exceed the TSS score, whereas in Figure 4 only 1 out of 75 (1.3%) segment scores exceed the TSS score. The use of segment scoring, in addition to controlling the number of FPs, also reduces prediction noise.

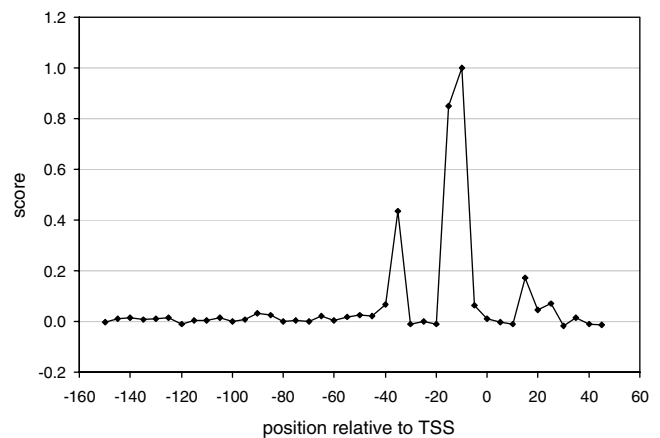
#### 4.4 Important motifs

Of the three TSS prediction methods considered so far, the ensemble-SVM performs best, especially at low scoring segment sizes. As noted in Section 1, one of our objectives was to achieve better TSS (and promoter) prediction by exploiting not just the promoter hexamers themselves, but also UP elements and other regulatory motifs that occur in the neighbourhood of the TSS. In this section we ask if this goal was achieved.

The ensemble-SVM allows us to assign a score to each feature (tagged 5mer) representing its contribution to classification of the training set sequences—the greater the magnitude, the greater the contribution of the motif. Table 5 shows the 40 motifs having the highest scores. As explained in Section 2, the position tag is the motif's position relative to the TSS, rounded to the nearest 5 bp.

**Table 5.** Motifs assigned highest magnitude scores by the motif-based SVM model

Motif	Count	Score	Motif	Count	Score
TTGAC(−35)	384	1.00	AGGAG(+20)	139	0.38
CCCCG(−5)	203	0.88	GGTAC(−15)	117	0.37
GGAGC(+15)	187	0.83	AGTTG(−35)	79	0.37
GTATA(−10)	175	0.64	GGAGG(+15)	87	0.36
GAGCA(+15)	156	0.61	TGCTA(−15)	119	0.36
AAACT(−10)	191	0.59	ACACA(+25)	74	0.36
GTTAG(−15)	140	0.55	TAATG(−10)	319	0.35
GTATA(−15)	216	0.55	CTATG(+25)	66	0.35
TATAG(−15)	162	0.53	TGTGA(−50)	82	0.35
ATAAT(−10)	199	0.52	ATTGC(−35)	92	0.34
ATACT(−10)	186	0.50	GGTAA(−15)	135	0.34
TGACA(−35)	224	0.49	CAATG(+25)	56	0.34
AGAAT(−10)	160	0.47	ACTAC(−10)	77	0.32
CCGTT(+0)	136	0.45	CCTAT(−15)	122	0.32
GCTTG(−40)	130	0.44	AGGGG(+20)	61	0.31
TATGA(+25)	98	0.44	ACAGG(+15)	75	0.30
AGGAC(+15)	109	0.44	TAGAA(−10)	231	0.30
ACTTG(−35)	117	0.44	ACCAT(−10)	50	0.29
TTGCA(−35)	145	0.43	CTTGA(−35)	118	0.29
CAAAC(−65)	87	0.39	ACTAG(−10)	56	0.28



**Fig. 5.** Plot of score versus position relative to the TSS. Scores at each position were obtained by summing the products of motif scores and motif frequencies at that position, across all 450 USRs. Note that high scores can be produced by small numbers of high scoring motifs, or large numbers of low scoring motifs.

To obtain Table 5, motif scores were averaged over the 10 cross-validation models derived from each of the 40 SVM models in the ensemble. The maximum number of models in which a motif could occur was therefore 400. The count columns give the number of models in which that motif occurred. The score columns give the average score of the motif, averaged over the 400 models and normalized so that the largest score is 1.0.

The top scoring motif was the  $\sigma^{70}$  consensus sequence, TTGAC (at  $-35$ ), included in 384 of the 400 models. Table 5 also contains a few TTGACA variants occurring at or around the  $-35$  position, e.g. TGACA and TTGCA. Most of the motifs in Table 5 occur around the  $-10$  and  $-35$  positions and correspond to the consensi,

TATAAT and TTGACA. We note that there is more variability in the  $-10$  motifs.

An alternative view of the distribution of significant motifs is shown in Figure 5. It plots the contribution each position makes (on average) to the score output by the ensemble-SVM. This plot was generated by summing the scores of all motifs with a given tag value across all 450 positive sequences. It reflects not only the scores assigned to individual motifs, but also the frequency of those motifs within the positive sequences.

The dominant peaks (and therefore significant motifs) occur at the expected  $-10$  and  $-35$  positions corresponding to the  $\sigma^{70}$  binding sites. There is also a small peak at  $+25$  which, based on the occurrence of the 5mers TATGA(+25), CTATG(+25) and CAATG(+25) in Table 5, probably represents the ATG start codon. Note that many TSSs occur  $\sim 30$  bp upstream of the gene start, consistent with this observation (refer again to Fig. 2).

Figure 5 also shows a significant spike at  $+15$ . Based on the occurrence of motifs GGAGC(+15), GAGCA(+15), AGGAC(+15), GGAGG(+15) and ACAGG(+15) in Table 5, this peak represents ribosomal binding sites, for which the consensus is AGGAGGU (Schneider *et al.*, 1986).

Finally, there are bumps at  $-50$  and further upstream possibly associated with UP elements (Gourse *et al.*, 2000). Motifs at  $-50$  included TGTGA, AATTA, AAAAA, and AGCAA and at  $-65$  included CAAAC, AACCC, CAAAT, AAACG, AAATC, AAAAC, and AAAAA.

In general the results suggest that when motif-based SVMs are trained on extended  $[-150, +50]$  neighbourhoods around the TSS, they will not only derive the bulk of their information from the  $-10$  and  $-35$  promoter hexamers but also detect and exploit other motifs such as start codons, ribosomal binding sites and UP elements.

#### 4.5 Combined models

It is possible to combine two or more of the primary models—DGS, PWM and ensemble SVM—with the goal of outperforming any single model. As noted in Section 3.3, Burden *et al.* (2005) obtained improved performance when they combined an NN model with the DGS distribution.

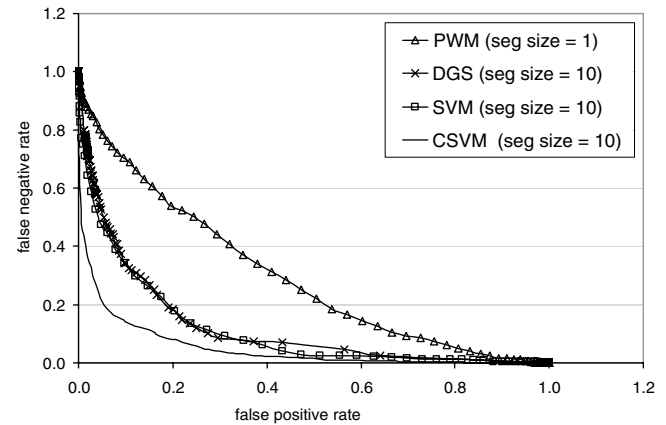
We combined all three primary models using a committee-SVM. The feature space of the committee-SVM is the 3D space of scores produced by the primary models. Committee-SVMs were trained and tested on 450 positive and 450 negative examples, using the same 10-fold cross-validation protocol described in Section 3.4. The positive examples were the primary score triples at the 450 TSS positions inside USRs. The negative examples were score triples for 450 randomly selected non-TSS positions, one for each USR. As with the primary models, the committee-SVM model was then applied to all scoring segments in the 450 USRs, also using 10-fold cross-validation. Best performance was obtained using a polynomial kernel of degree 2.

Table 6 shows the classification errors, DET areas, FP/TP ratios and TP counts in the top 100 predictions obtained by the committee-SVM for four scoring segment sizes. Classification error is the percentage of mis-classified positions in the balanced test set of 450 TSS and 450 non-TSS positions. The classification errors in Table 6 are low for this problem. For a comparable prediction problem involving a balanced dataset, our average error rate of 11.6% (scoring segment = 1) compares favourably with that of

**Table 6.** Classification error, DET area and FP/TP ratios for the committee-SVM model

Segment size	Classification error (%)	DET area	FP/TP FNR 10%	TP count in top 100
1	11.6	0.060	146 $\pm$ 47	14 $\pm$ 8.1
5	9.4	0.047	23 $\pm$ 6.8	57 $\pm$ 11
10	8.4	0.049	11 $\pm$ 3.5	68 $\pm$ 18
20	6.2	0.061	7.3 $\pm$ 2.0	70 $\pm$ 10

DET standard deviations (obtained from 10-fold cross-validation) are close to 0.013.



**Fig. 6.** DET curves for PWM, DGS, ESVM and CSVM models. A PWM with segment size of one represents the standard method of promoter/TSS prediction (see Section 3.1). The DGS, ensemble-SVM and committee-SVM (CSVM) models are described in Sections 3.3, 3.2 and 4.5, respectively.

16.5% reported by Gordon *et al.* (2003) for an SVM using a sequence alignment kernel.

The low DET areas in Table 6 indicate that the committee-SVM model performs substantially better than any of the previous approaches. For segment sizes of 5 and 10, the average DET area is 0.05, which is half of the best result obtained with the ensemble-SVM (Table 1). The DET performance of the methods described in this paper can be shown graphically (Fig. 6). The standard method (PWM with scoring segment size of 1) yields the DET curve having largest area. The committee-SVM yields the DET curve with smallest area, with the other methods lying in-between.

Finally we consider the FP/TP ratios that are of most interest to the biologist. Over the range of scoring segment sizes, the FP/TP ratio of the committee-SVM is about half that of the ensemble-SVM. Correspondingly, the number of true positives in the top 100 predictions is increased by  $\sim 40\%$ , with the best result being 70 correct predictions.

## 5 CONCLUSIONS

This paper has explored three basic methods of bacterial TSS prediction: a standard PWM method, a method based solely on the distribution of TSS DGS and an ensemble-SVM method, which builds on the work of Gordon and Towsey (2005). Performance

of all methods was assessed using 10-fold cross-validation. Comparisons on a biologically realistic task were rigorously measured using the area under a DET curve.

Of the three primary TSS prediction methods considered, the ensemble-SVM performed the best. Our expectation was that this method would achieve better TSS (and promoter) prediction by exploiting the presence of other regulatory motifs in the TSS neighbourhood in addition to the promoter hexamers. Analysis of the motifs recognized by the ensemble-SVM confirmed our expectation, with start codons and ribosomal binding sites making a significant contribution to classification (Table 5 and Fig. 5).

A committee-SVM which combined the three primary models performed significantly better than the three individual methods alone. Importantly, it yielded significantly lower rates of FP predictions, offering a practical *in silico* method to guide laboratory searches for promoter locations. Given the top scoring 100 predictions from the committee-SVM method, laboratory tests could expect to confirm about 70 actual promoters (Table 6), whereas the standard PWM would expect to find only about 5 (Table 4).

An intriguing finding was that the simple DGS method performed better than the standard PWM and almost as well as the ensemble-SVM over gene USRs of length 750 bp. We note however, that the DGS method performed poorly over shorter USRs when compared with the more ‘sophisticated’ methods.

It is difficult to compare our results with other published figures due to differences in data and experimental protocol. However for a comparable TSS prediction task using *E.coli* data, the results shown by Burden *et al.* in their Table 2, indicate that at a recall of 50%, they achieved a precision [defined as  $TP/(TP + FP)$ ] of 16–17%. (They use a scoring segment size of 7 and searched 500 bp upstream of GSSs.) Our comparable figures for the committee-SVM were 26 and 33% for scoring segment sizes of 5 and 10, respectively (data not shown in our tables). Burden *et al.* do not give figures for 90% recall.

Huerta and Collado-Vides (2003) claim ‘the highest predictive capability reported so far’ for the promoter prediction task in *E.coli*. They use a two-stage PWM method, code-named Cover. Figure 8e of their paper indicates that Cover achieved a precision of 33% at 50% recall, comparable with our committee-SVM approach. However, they use a scoring segment size of 11 and their search region is restricted to 250 bp upstream of the GSS, rather than the 750 bp region considered in this work. Their shorter search region lowers the possibility for FP predictions. While direct comparison remains difficult, the comparison suggests that our approach yields superior performance.

In conclusion, our results demonstrate that an ensemble-SVM, using mismatch string kernels, can potentially detect and exploit a range of regulatory motifs for better TSS/promoter detection. This opens up the interesting possibility of ‘focusing’ the SVM approach on special categories of regulatory motifs, e.g. by restricting it to motifs having specified structure or occurring at specified positions.

The models we have constructed from *E.coli* data may in principle be used to find promoters in other bacterial species that have the same promoter consensus and a similar distribution of TSS

locations with respect to gene starts. This intuition has been confirmed by a preliminary investigation of two *Bacillus* and four *Chlamydia* species.

Also, we believe the potential exists to improve our results by including within the committee-SVM approach other motif-based models (e.g. those tuned to inverted repeats) or models based, e.g. on DNA stacking energy. The ultimate goal is to achieve very high levels of TSS prediction accuracy, as has been done with Translation Initiation Sites (Hatzigeorgiou, 2002).

## ACKNOWLEDGEMENTS

We thank administrators of the RegulonDB database for access to their data, and Joachims and Serrafini *et al.* for the SVM Light and GPDT software packages, respectively. This work was supported by an Australian Research Council (ARC) grant and a Queensland University of Technology Strategic Collaborative grant.

*Conflict of Interest:* none declared.

## REFERENCES

- Burden, S. *et al.* (2005) Improving promoter prediction for the NNPP2.2 algorithm: a case study using *Escherichia coli* DNA sequences. *Bioinformatics*, **21**, 601–607.
- Duda, R.O., Hart, P.E. and Stork, D.G. (2001) *Pattern Classification*. John Wiley & Sons, New York.
- Gordon, J.J. and Towsey, M. (2005) SVM based prediction of bacterial transcription start sites. In *Proceedings 6th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL'05)*, Brisbane, Australia, July 2005. Lecture Notes in Computer Science, **3578**, 448, Springer, Berlin.
- Gordon, L. (2003) Sequence alignment kernel for recognition of promoter regions. *Bioinformatics*, **19**, 1964–1971.
- Gourse, R.L. *et al.* (2000) Ups and downs in bacterial transcription initiation: the role of the alpha subunit of RNA polymerase in promoter recognition. *Mol. Biol.*, **37**, 687–695.
- Hatzigeorgiou, A.G. (2002) Translation initiation start prediction in human cDNAs with high accuracy. *Bioinformatics*, **18**, 343–350.
- Huerta, A.M. and Collado-Vides, J. (2003) Sigma-70 promoters in *E. coli*: specific transcription in dense regions of overlapping promoter-like signals. *J. Mol. Biol.*, **333**, 261–278.
- Joachims, T. (1999) Making large scale SVM learning practical. In Scholkopf, B., Burges, C. and Smola, A. (eds), *Advances Kernel Methods—Support Vector Learning*. MIT Press, Cambridge, MA.
- Leslie, C.S. *et al.* (2004) Mismatch string kernels for discriminative protein classification. *Bioinformatics*, **20**, 467–476.
- Lewin, B. (1985) *Genes*. John Wiley & Sons, New York.
- Liu, H. and Wong, L. (2003) Data mining tools for biological sequences. *J. Bioinform. Comp. Biol.*, **1**, 139–167.
- Salgado, H. *et al.* (2001) RegulonDB (version 3.2): transcriptional regulation and operon organization in *E. coli* K-12. *Nucleic Acids Res.*, **29**, 72–74.
- Sambrook, J., Fritsch, E.F. and Maniatis, T. (1989) *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, ISBN: 0-87969-309-6.
- Schneider, T.D. (1997) Information content of individual genetic sequences. *J. Theor. Biol.*, **189**, 427–441.
- Schneider, T.D. *et al.* (1986) Information content of binding sites on nucleotide sequences. *J. Mol. Biol.*, **188**, 415–431.
- Serrafini, T., Zanghirati, G. and Zanni, L. (2004) Parallel GPDT: a parallel gradient projection-based decomposition technique for support vector machines.
- Stormo, G. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.
- Vapnik, V. (1995) *The Nature of Statistical Learning Theory*. Springer-Verlag, NY.