COVER SHEET

SVM Based Prediction of Bacterial Transcription Start Sites

James Gordon and Michael Towsey

Centre for Information Technology and Innovation, Faculty of Information Technology, Queensland University of Technology, GPO Box 2434, Brisbane QLD 4001, Australia
{jj.gordon, m.towsey}@qut.edu.au

Abstract. Identifying bacterial promoters is the key to understanding gene expression. Promoters lie in tightly constrained positions relative to the transcription start site (TSS). Knowing the TSS position, one can predict promoter positions to within a few base pairs, and vice versa. As a route to promoter identification, we formally address the problem of TSS prediction, drawing on the RegulonDB database of known (mapped) *Escherichia coli* TSS locations.

The accepted method of finding promoters (and therefore TSSs) is to use position weight matrices (PWMs). We use an alternative approach based on support vector machines (SVMs). In particular, we quantify performance of several SVM models versus a PWM approach, using area under the detection-error tradeoff (DET) curve as a performance metric. SVM models are shown to outperform the PWM at TSS prediction, and to substantially reduce numbers of false positives, which are the bane of this problem.

1 Introduction

Bacterial promoters are sections of DNA lying upstream of a gene transcription start site (TSS), which regulate transcription via selective binding by an RNA Polymerase (RNAP) / sigma factor complex [1]. They are difficult to find because they lie at an unknown distance upstream of the gene start codon, and their associated DNA is weakly conserved. Importantly, they consist of two binding sites (hexamers) which lie in a well-defined window upstream of the TSS. Knowing the TSS location, one can predict promoter locations to within a few base pairs (bp), and vice versa.

As a route to identifying promoters, this paper uses support vector machines (SVMs) [2] to predict TSS locations. We consider TSSs for the class of *Escherichia coli* sigma-70 promoters. Sigma-70 promoters are bound by the *E. coli* sigma-70 transcription factor, and are located around the –10 and –35 positions with respect to the TSS. The RegulonDB database contains ~ 700 mapped sigma-70 TSS locations [3].

The accepted method of finding promoters is to use a position weight matrix (PWM) to search for matches to known promoter hexamers [4]. This approach utilizes only information contained in the two hexamers and the intervening gap length. Based on information theoretic reasoning, it is known that the mapped hexamers are insufficiently conserved to identify all expected promoters in the background genome [5].

In addition to the promoter hexamers, a TSS is surrounded by a number of other regulatory binding sites. These include binding sites for proteins such as activators and repressors, that enhance or repress the level of transcription initiation. Through the use of machine learning techniques — in our case, SVMs — one might hope to exploit this expanded set of patterns in order to achieve better TSS and promoter prediction. We use the term 'TSS prediction' to refer to this more general approach to TSS and promoter identification.

TSS prediction has an analogue in the problem of translation initiation site (TIS) prediction, the goal of which is to find gene start codons. Recent improvements at TIS prediction have been achieved with a multi-stage approach [6,7]. The problem of sigma-70 TSS / promoter prediction is acknowledged to be a difficult one. It is possible that multi-stage approaches to this problem will likewise improve on results achieved to date. We view the SVM models described in this paper as possible first-stage algorithms, that could be used for identifying likely promoter regions. Future research could involve pairing these models with appropriate second-stage algorithms to achieve higher levels of accuracy.

2 Data

This paper utilized TSS locations derived from the RegulonDB database [3], and sequences extracted from the *E. coli* K12 genome (www.genome.wisc.edu). Two parallel data sets were constructed. The first data set, used to train and test different SVM models, consisted of 450 positive sequences (each containing a single mapped TSS) and 450 negative sequences (not containing known TSSs). The positive sequences extended from –150 to +50 bases relative to each TSS.[1]

Negative sequences were derived from parts of the genome that did not contain a known TSS and were all 200 bases long. They contained a reference position at the 151 position corresponding to the TSS in positive sequences. We derived three sets of negative sequences for the first data set: sequences from (a) coding regions (CDRs), (b) non-coding regions between divergent genes (DNCRs), and (c) non-coding regions between convergent genes (CNCRs). In each case, candidate negative regions were generated from the whole genome and randomly shuffled before selecting 450 examples. Because there were relatively few candidate CNCRs, some CNCR regions were permitted to overlap by 100 bp. Due to the positional nature of the SVM models employed here, this was not expected to influence our results.

Note that because of their location between divergently transcribed genes, DNCRs are inherently likely to contain (unmapped) promoters and TSSs. Like coding regions, CNCRs are inherently less likely to contain promoters and TSSs.

A second data set was used to test the SVM models on a biologically realistic task and to compare their performance with that of a standard PWM. It consisted of 450 sequences extending 750 bp upstream of gene start codons. Hereafter we refer to these as *gene upstream regions* (USRs). Each USR contained a single known TSS (the same

[1] According to biological convention, the TSS position is denoted by +1. The position immediately upstream is -1. There is no 0 position.

TSS as in the corresponding positive SVM training sequence). Of the 450 USRs, only nine consisted entirely of non-coding DNA. The remainder overlapped coding regions to varying degrees. We used only 450 of the 676 known sigma-70 TSSs available in the RegulonDB database [3] to ensure that all USRs were non-overlapping.

3 SVM Approach

In an SVM approach, DNA sequences are represented as vectors in a feature space. The SVM is presented with positive and negative examples. From these it determines an optimal decision plane through the feature space separating positive and negative examples. In real problems one is unlikely to achieve complete separation. SVM performance is therefore measured by 'generalization error' — the percentage of unseen test examples that fall on the wrong side of the decision plane.

The data representation employed for this study was a variant of the string kernel proposed by Leslie et al [8]. Each sequence was represented by a vector of scaled counts of 5-mers occurring within the sequence. A single mismatch was allowed in each 5-mer. After collapsing each mismatch neighbourhood onto a single 5-mer, the number of possible 5-mers used to represent a sequence was $\{A,C,G,T\}^4 = 256$.

Our enhancement of the approach in [8] was to 'tag' each motif with its offset from the sequence's reference (151) position. For example, an occurrence of 5-mer ACCGT in positions [–5,+5] (relative to the reference position) was registered as an occurrence of the position-tagged motif ACCGT(0). The same 5-mer in positions [–15, –6] was registered as ACCGT(–10), and in [+6,+15] as ACCGT(+10). Employing a window of size 10 was intended to allow 'fuzziness' of up to 10 bp in identifying TSSs.

Position-tagging increased the SVM feature space dimension. In sequences extending from –150 to +50, the number of possible position tags is 21 (–150, –140, …, –10, 0, +10, …, +50). The total number of possible position-tagged motifs was therefore 21 x 256 = 5,376. These position-tagged motifs (i.e., features) were ranked according to their symmetric uncertainty as in [6], and the top 200 were retained in SVM vectors. Note that position-tagged motif counts were scaled by their symmetric uncertainty before insertion into SVM vectors. This was intended to increase the margin associated with significant motifs. All SVM vectors were then normalized to unit length. SVMs were generated using either SVM-Light [9] or the GPDT [10].

4 PWM Approach

PWMs were derived from USRs. The first step in PWM construction was to look for the best match to the *E. coli* sigma-70 promoter consensus hexamers TTGACA and TATAAT upstream of each TSS. The 5′ end of the best fit TATAAT-like motif was constrained to occur in the range [–19,–9] relative to the TSS. The gap between the TTGACA and TATAAT-like motifs was constrained to be in the range [14,20].

Each candidate motif pair within these parameter ranges was assigned a score, equal to the number of bases matching the consensus hexamers, plus some gap weight-

ings to give preference to gaps in the centre of the [14,20] range. For each TSS, the motif pair with the highest score was selected as the best fit. Having identified best fit hexamers for every TSS, based on closeness to the consensus, a PWM was then constructed using the background nucleotide frequencies sampled from all USRs [4,5].

5 Method

A unique index from 1 to 450 was assigned to each USR, positive sequence, and negative sequence within each set (CDR, DNCR and CNCR). Note that USRs and positive sequences having the same index were neighbourhoods of the same TSS. Based on index, each of the datasets was then divided into 10 equal parts. Each of these 10 parts was successively held in reserve as a test set, while the remaining 90% of the data was used to train SVM models and generate a PWM. The resulting SVMs and PWM were then evaluated on the 10% of USRs held in reserve (i.e., 10-fold cross-validation). Results reported below represent average performance over the 10 test sets.

Evaluation of the PWM and SVMs involved applying the models to each of the 750 positions within each test USR. The PWMs and SVMs generated a score for each position. In the case of the SVMs, this score was the perpendicular distance from the decision plane of the [–150,+50] neighbourhood of the candidate position. In the case of PWMs, it was the highest PWM score that could be obtained from two upstream hexamers, the first with its 5′ end at –14 bp upstream of the candidate position, and the second lying at a gap of 14 – 20 bp upstream of the first. (Note that these parameters are consistent with those used to generate the SVMs and PWMs.)

Next a threshold T was defined. Within USRs, the TSS position itself and the five positions on either side of it were considered to be positives. These positions were scored as true positives (TP) if the SVM or PWM score exceeded T, and false negatives (FN) if it fell below T. All other positions in the USRs were considered to be negatives and were scored as false positives (FP) if the SVM or PWM score exceeded T, and true negatives (TN) if it fell below T.

By varying T over the range of scores, it is possible to construct detection-error tradeoff (DET) curves. These are plots of false negative rate (FNR) versus false positive rate (FPR), as shown in Figure 1. [FNR = FN / (FN + TP) and FPR = FP / (FP + TN).] Note that the area under a DET curve is a measure of the classifier's ability to correctly identify TSS positions. The smaller the area, the better the performance. DET area constitutes a rigorous and objective measure of performance, similar to the receiver operating characteristic (ROC) curves used in other areas of statistics.

6 Results and Conclusions

Figure 1 compares performance of three SVM models and the PWM at predicting TSSs in the USR sequences of dataset 2. Table 1 gives areas under the DET curves in Figure 1, and corresponding generalization errors on data set 1 (i.e., SVM test sequences). The most notable result is that the CDR and CNCR SVM models perform

substantially better than the PWM on the biologically realistic promoter prediction task. This is true regardless of whether false negatives are scored in a window around the TSS (Table 1), or only at the TSS position itself (results not shown).

The relatively poor performance of the DNCR SVM model is unsurprising. The DNCRs we isolated are likely to contain (unmapped) TSSs and promoters and are therefore a poor source of negative sequences. By contrast, CNCRs are unlikely to contain promoters, but still have characteristics of non-coding regions, and thus are a useful source of negative sequences.

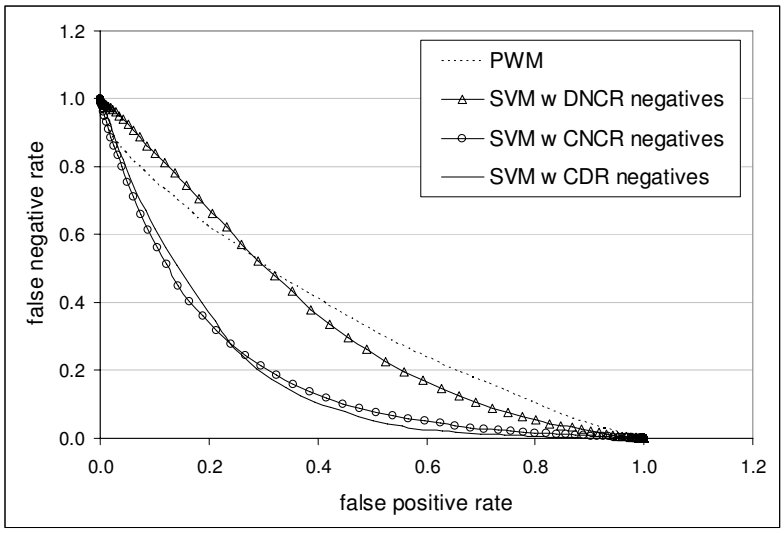

Fig. 1. Detection-Error Tradeoff Curves for PWM and SVMs

An important conclusion from Table 1 is that SVM performance on the TSS classification task is not a reliable indicator of performance on a biologically realistic task. In particular, the CNCR SVM model performed poorly on the classification task but well on the more realistic TSS prediction task where false positives must be reduced. Papers often present promoter identification algorithms tested only on artificial classification tasks with equal numbers of positives and negatives. This is not a reliable indicator of performance in the less constrained biological setting.

The use of coding region negatives is often criticized because they have quite different statistics from non-coding TSS regions resulting in an artificially simple promoter classification task. However in our results, the CDR SVM model trained with CDR negatives performed well on both the classification task and the realistic TSS prediction task. In the compact bacterial genome, some promoters may extend into the upstream gene. In such cases the use of coding region negatives will be helpful.

In the laboratory setting, the biggest problem in using PWMs and other classifiers is the high rate of false positives. When the expected FP/TP ratio is on the order of 1000, in-silico detection of bacterial promoters does not offer the biologist meaningful

Table 1. Generalization errors, DET areas and optimum FP/TP ratios for four methods (standard deviations are derived from 10-fold cross validation)

Method	Gen. Error	DET Area	Optimum FP/TP Ratio
PWM	—	0.36 ± 0.01	155,000 / 300
SVM (DNCR negs)	25.1 ± 2.4%	0.35 ± 0.03	75,000 / 215
SVM (CNCR negs)	27.2 ± 4.3%	0.19 ± 0.03	71,000 / 320
SVM (CDR negs)	15.0 ± 3.8%	0.18 ± 0.02	58,000 / 310

guidance for laboratory testing. Therefore it is particularly important to identify what the expected optimum FP/TP ratio is for any method. Table 1 gives the FP/TP ratio for each of our four models in their optimum configuration (point on DET curve closest to origin). Note that we have both increased the rate of true positives and reduced the FP/TP ratio from 517 to 187. These results suggest that SVM models can outperform PWMs for a realistically constructed promoter prediction task. The goal of our future work will be to further reduce the false positive rate to levels that allow efficient laboratory investigation of in-silico predicted promoters.

Acknowledgements: The authors are grateful to colleagues J. Hogan, P. Timms and S. Mathews for their input to and review of this work. We also thank administrators of the RegulonDB database [3] for access to their data, and Joachims [9] and Serrafini et al [10] for access to the SVM Light and GPDT software packages.

References

1. Lewin,B.: Genes. Wiley (1985).
2. Vapnik,V.: The Nature of Statistical Learning Theory. Springer-Verlag (1995).
3. Salgado,H., Santos-Zavaleta,A., Gama-Castro,S., Millán-Zárate, D., Díaz-Peredo,E., Sánchez-Solano,F., Pérez-Rueda,E., Bonavides-Martínez,C., Collado-Vides,J.: RegulonDB (version 3.2): Transcriptional Regulation and Operon Organization in Escherichia coli K-12. Nucleic Acids. Res. 29: 72-74 (2001).
4. Stormo, G.: DNA Binding Sites: Representation and Discovery. Bioinformatics 16: 16-23 (2000).
5. Schneider,T.D., Stormo,G.D., Gold,L., Ehrenfeucht,A.: Information Content of Binding Sites on Nucleotide Sequences. J. Mol. Biol. vol. 188 pp. 415-431 (1986).
6. Liu, H., Wong,L.: Data Moning Tools for Biological Sequences. J. Bioinformatics and Computational Biology vol. 1 pp. 139 – 167 (2003).
7. Hatzigeorgiou, A.G.: Translation Initiation Start Prediction in Human cDNAs with High Accuracy. Bioinformatics 18 pp. 343 - 350 (2002).
8. Leslie,C.S., Eskin,E., Cohen,A., Weston, J., Noble, W.S.: Mismatch String Kernels for Discriminative Protein Classification. Bioinformatics vol. 20 pp. 467-476 (2004).
9. Joachims,T.: Making Large Scale SVM Learning Practical. Advances in Kernel Methods – Support Vector Learning, B. Scholkopf, C. Burges, A. Smola eds., MIT Press (1999).
10. Serrafini,T., Zanghirati,G., Zanni, L.: Parallel GPDT: A Parallel Gradient Projection-Based Decomposition Technique for Support Vector Machines, http://dm.unife.it/gpdt/ (2004).