THE IN SILICO PREDICTION OF PROMOTERS IN BACTERIAL GENOMES

MICHAEL TOWSEY[1] JAMES M. HOGAN[2]
m.towsey@qut.edu.au j.hogan@qut.edu.au

SARAH MATHEWS[1] PETER TIMMS[1]
s.mathews@qut.edu.au p.timms@qut.edu.au

[1] *Institute for Health and Biomedical Innovation, Queensland University of Technology, Queensland, Australia.*
[2] *Faculty of Information Technology, Queensland University of Technology, Queensland, Australia.*

In silico approaches to the identification of bacterial promoters are hampered by poor conservation of their characteristic binding sites. This suggests that the usual position weight matrix models of bacterial promoters are incomplete. A number of methods have been used to overcome this inadequacy, one of which is to incorporate structural properties of DNA. In this paper we describe an extension of the promoter description to include SIDD (stress induced duplex destabilization), DNA curvature and stacking energy. Although we report the best result to date for a realistic promoter prediction task, surprisingly, DNA structural properties did not contribute significantly to this result. We also demonstrate for the first time, that sigma-54 promoters have a stronger association with SIDD than do other promoter types.

Keywords: bacterial promoters, promoter prediction, SIDD, duplex destabilization, DNA curvature

1. Introduction

The identification of promoters is essential for an understanding of gene regulation. However wet-lab techniques to identify bacterial promoters are costly and time consuming and thus *in silico* methods have strong appeal. Unfortunately, computational approaches to promoter identification are confounded by the poor conservation of their important functional sites. Transcription in bacteria is initiated by a protein complex known as RNA polymerase (RNAP), consisting of five subunits (collectively known as the *core enzyme*) and an additional σ factor. The σ factor is responsible for locating promoters by recognizing two binding sites, typically located at the -10 and -35 positions with respect to the transcription start site (TSS). Once transcription has begun, the σ factor dissociates and transcription continues with core enzyme alone.

In silico identification of promoters has tended to focus on detecting the -10 and -35 binding site motifs which are typically (in the case of the most common housekeeping σ^{70}) separated by a spacer of 14 to 20 base pairs (bp). However, it was established early using information theoretic reasoning, that the known -35 and -10 binding sites are insufficiently conserved to account for all the expected promoters in the background genome [15]. Furthermore, when potential binding sites are scored using position weight matrices (PWMs), it is found that about 50% of the known TSSs are not located at the highest scoring position upstream of a gene start site [7]. Clearly there are other factors involved in the positioning of promoters that are not captured in a simple PWM description. In this regard, it is extremely interesting that experiments have demonstrated that many look-alike promoter sites initiate transcription *in vitro* even though they fail to do so *in vivo* [8].

Several attempts have been made to use more sophisticated machine learning methods to identify promoters, for example neural networks [3] and support vector machines (SVM) [4]. While these methods offer somewhat increased accuracy depending on how the task is constructed, the improvements may not justify the heavy computation required for training the classifiers. Maetschke *et al.* [11] revisited the PWM approach, but this time utilizing information that has recently come to light about the mode of action of RNA polymerase. Incorporating extended -10 motifs [12] and UP elements (AT rich regions upstream of the promoter) [18] into their promoter description slightly improved predictive accuracy to around 50%, but clearly the most important predictive improvement for *E. coli* promoters was obtained by including information about the distance of the putative TSS from the gene start site (hereafter referred to as the TSS-GSS distance). This observation has also been made in [3].

There are at least three explanations advanced to explain the poor predictive performance of existing bacterial promoter models. First, it is possible that potentially strong promoter sites are masked by some mechanism that makes them inaccessible to RNA polymerase. In chlamydial species, for example, DNA is condensed by the binding of histone-like proteins during late development, which plays a role in down-regulating gene expression by removing the accessibility of promoters [6]. Secondly it is well known that some weak promoters can only function in conjunction with activators. Unfortunately, while many transcription factors have been identified, most of their binding sites have not, and it is not clear how transcription factor binding sites can be included in promoter models, except in the case of some well characterized global regulators [16]. Thirdly, it is becoming increasingly clear that structural features of DNA have an important regulatory role in gene expression, for example stacking energy [1], DNA curvature [9, 13] and Stress Induced Duplex Destabilization (SIDD) [19]. Our paper investigates the use of these DNA structural properties to help identify promoters.

Stacking energy refers to the interactions between consecutive base pairs of a ‘stacked’ DNA sequence. It is assumed to be a purely local phenomenon depending only on nearest neighbour interactions and contributes to local duplex stability or meltability. Units are kcal/mole and more negative values correspond to higher duplex stability.

Curved DNA is believed to play an important role in many cell processes such as transcription initiation and termination, DNA replication and nucleosome positioning [9]. DNA curvature influences the binding affinity of regulatory proteins while DNA looping can increase the proximity of separated regulatory sites. Curvature is defined as the inverse of the radius of an arc that approximates a given DNA sequence. A value of one corresponds to the degree of curvature seen in nucleosomal DNA (see Figure 1, right).

SIDD is a thermodynamic quantity whose value for any DNA base pair may be defined as the incremental free energy (kcal/mole) required to force that base pair to remain open. Regions having low SIDD energy are strongly destabilized, that is, they have a high propensity to melt under normal physiological conditions. The SIDD value for any particular base pair depends on the local GC content and on the superhelicity (degree of negative super-coiling) of the DNA molecule. However unlike stacking energy, SIDD is not purely a local property but rather depends on the distribution of SIDD throughout the molecule. Calculating SIDD for an entire bacterial DNA molecule is a computationally demanding exercise. Even in a 4 Mbp genome, every base pair potentially affects every other base pair.

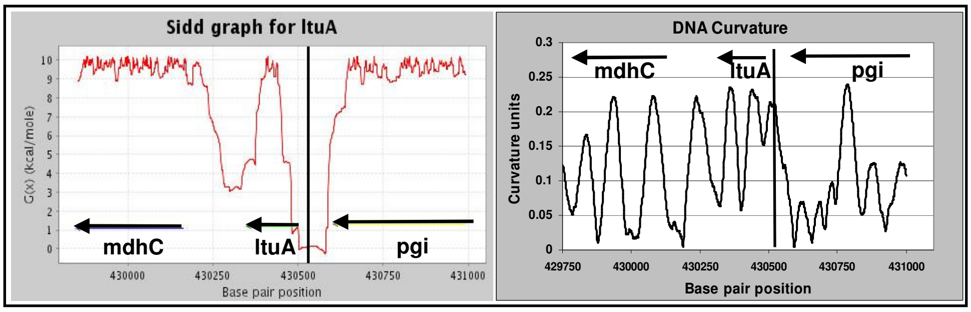

Figure 1. A representation of SIDD (left) and DNA curvature (right) in the vicinity of *ltuA* (*Chlamydia trachomatis*). Vertical lines indicate the TSS location. Horizontal arrows show coding regions. The promoter for *ltuA* lies within a strong SIDD region which occupies the entire upstream non-coding region. It also lies just upstream of a region of high curvature. (Curvature window=100)

Wang and Benham [19] have demonstrated that SIDD energy is a useful predictor of promoter regions. Their reported accuracy of around 80% is, on the face of it, a remarkable result given that the best typical result for promoter prediction in *E. coli* is around 50% [4, 11]. The authors attribute their success to the fact that about 80% of documented promoters contain a strong SIDD site. They define a promoter as extending from positions -80 to +20 with respect to the TSS and they define strong SIDD as any value below 6 kcal/mole. The association of strong SIDD with intergenic regions (see Figure 1, left, for an example) appears to be a general property of all bacterial species [2, 22] although specific association with promoters has been shown only for *E. coli* and *B. subtilis* [19], species for which there is a large number of mapped promoters.

Our group has an interest in the prediction of bacterial promoters using both PWMs and machine learning methods. We have previously shown that the success rates reported for the promoter prediction task are acutely sensitive to the task definition. In particular, the choice of negative instances for the binary prediction task can make the task artificially easy [5] and the degree of focus on regions where TSSs are likely to be found can also bias performance [4].

In this paper we examine the use of DNA structural properties as predictive attributes for finding promoters. Once again we note the sensitivity of the results to the task definition. Our results appear to be the best yet reported for a biologically realistic promoter prediction task. Perhaps surprisingly, structural properties did not contribute appreciably to achieving this result even though they are indeed important for the regulatory activity of many promoters. We also demonstrate for the first time that sigma-54 promoters have a stronger association with SIDD than do promoters associated with sigma-70 and other sigma factors.

2. Methods

2.1. *Data*

All investigations were performed with the genome of Escherichia coli K-12 MG1655 (ACCN:U00096.2) [23]. Experimentally confirmed TSS locations for this genome were obtained from RegulonDB [24]. The data set was filtered for known TSS locations associated with sigma-70 promoters, resulting in 542 records. We extracted 250 bp sequences upstream of those genes closest to the given TSSs. Following Huerta *et al* [7], this approach eliminated all TSSs further than 250 bp from the gene start. We also eliminated seven TSSs that were located within 10 bp of another known TSS because our approach did not discriminate two TSSs closer than 10 bp. The final data set consisted of 439 sequences each 250 bp long, containing a total of 487 annotated TSS locations. Thirty nine of the sequences contained multiple TSS locations.

Stacking energy values were obtained from [1]. SIDD data for the *E. coli* genome were kindly provided by Dr Craig Benham and are available at [22]. DNA curvature was calculated using the CURVA software kindly provided by Dr. Alexander Bolshoy [9].

2.2. *Experimental design*

We approached the promoter prediction task in two steps. First, we constructed a description of a sigma-70 promoter using BioPatML, an XML language for the description of biological patterns [10]. See the next section for more detail of our promoter definition. We scanned all 439 upstream sequences and assigned a score to each position indicating the similarity of that region to our promoter definition as if that position were a TSS. After smoothing the resultant similarity profile with a moving average filter (window = 3), peaks were identified as described in [17] and marked as candidate TSS locations. The rationale is that true TSSs are most likely to be found close to high scoring peaks, that is, locations where the upstream region has high similarity to our promoter definition.

The second step involved using a suitably trained decision tree to classify as *true* or *false* each of the TSS candidates found in step 1. For the decision tree we used the popular WEKA data mining tool [20] and its implementation of C4.5 [14]. This classifier was trained using a selection of promoter features such as the similarity score of the candidate TSS, the -10 and -35 scores of the candidate and a variety of DNA structural features as described later in the paper.

To estimate prediction accuracy, we adopted a 10 fold cross-validation protocol as follows: The 439 sequences were divided into 10 sets. For each fold, nine sets were scanned with our promoter definition to obtain TSS candidates for C4.5 training data. Features were extracted from true TSS locations in each set to obtain positive training instances and from false step 1 TSS candidates to obtain negative training instances.

Testing was performed on candidate predictions obtained from the tenth (holdout) set of sequences. A true positive (TP) was any positive prediction five bp or less from a known TSS. A false positive (FP) was any positive prediction more than five bp from a known TSS. Recall was defined as TP/(TP+FN) and precision as TP/(TP+FP) where FP denotes false positive and FN denotes false negative. Averages were obtained for recall and precision over 10 repeats of 10 fold cross-validation, that is, over 100 folds.

2.3. *Step 1: Use of BioPatML to obtain candidate TSS locations*

Our promoter definition included five elements: an UP element, the -35 element, a spacer, the -10 element and the discriminator (the region between the -10 element and the TSS). The -35 and -10 elements were defined using PWMs prepared from sequence data for known -10 and -35 binding sites available at DPInteract [25]. Scores for the spacer and discriminator widths were calculated using the *accessibility* formula of Shultzberger *et al.* [16, Eq.(2)]. The UP element was defined as a 17 bp sequence, $W_{15}N_2$, directly upstream of the -35 element, where W = A or T and N = any base [11]. Adding an extended -10 element or constraining the TSS to be a Purine did not improve performance.

BioPatML normalises the match score for each pattern element to a value between 0 and 1 - 0 for the minimum possible score and 1 for the maximum. The combined match score is a weighted sum of the normalised element match scores and hence optimisation of the weighting parameters is required. We did this by line search, fixing the weight for the -10 element at 1.0. Interestingly, the optimum weight obtained for the UP element, 0.45, was slightly greater than that for the -35 element, 0.35, indicating the importance of the UP element in *E. coli* promoters.

TSS candidates were obtained by identifying peak locations in the graph of similarity scores. Candidate selection was constrained such that no two candidates could be within 5 bp of each other, i.e. the permitted error tolerance for a correct prediction.

2.4. *Step 2: Classification of candidate TSS locations*

The TSS candidates obtained from step 1 were labeled as *positive* or *negative* depending on their distance from the nearest true TSS. C4.5 training data included the negative candidates from step 1 and positive instances obtained directly from the set of known TSSs. Consequently the available training data consisted of 487 known TSSs (positive class) and 4751 candidate TSSs not biologically confirmed as promoters (negative class)[a].

[a] It is likely that some of these negative instances are indeed as yet unidentified promoters.

This over-representation of negative instances was found to reduce the accuracy of the resultant classifier. Consequently, we trained C4.5 with all the positive instances but only the five top ranking (highest scoring) negative instances from each sequence. Note that this set of negatives includes candidates that are most like positives and hence makes the task difficult, albeit realistic. For training, we used the default C4.5 parameters provided by WEKA except that we set the Laplace parameter *true* since this slightly improved performance.

3. Results and Discussion

3.1 *The TSS Prediction Task*

The first step in our promoter prediction algorithm involved finding candidate TSSs/promoters in each upstream sequence. An average of 11.9 candidates or predictions per sequence was obtained. These were ranked according to their similarity score and each candidate labeled as a TP or FP prediction. Table 1 indicates that of the 217 rank 1 predictions closest to a true TSS, 206 were TP and the remainder FP predictions (>5 bp from the true TSS). The average error of the 217 predictions was 1.98 bp. Recall and precision for the rank 1 predictions were 42% and 47% respectively. This is comparable to the result reported in [11] for the case where TSS-GSS distance was *not* incorporated into the pattern description. Observe that while 90% of true TSSs were within 5 bp of a local maximum, only 42% of them were located within 5 bp of the sequence global maximum. This is consistent with the findings in [7]. As is to be expected, recall increased but precision declined when lower ranked predictions were accepted.

Table 1. Counts of TP TSS predictions obtained from step 1 using a BioPatML description of a promoter. The predictions for each sequence/gene were ranked by BioPatML similarity score.

Rank	# TP predictions	# true TSSs closest to peak	Av error (bp) for predictions
1	206	217	1.98
2	80	92	2.26
3	33	41	2.98
4	33	40	2.88
5	29	32	2.44
6	18	19	2.05
7	12	15	3.47
8	7	9	2.56
9	5	7	2.71
10	7	9	4.44
11	4	4	3.25
12	2	2	1.50
total	436	487	-

The object of step 2 was to design a classifier which could select the true TSS(s) from the candidates identified in step 1. We used the well established C4.5 decision tree because, in our initial investigations, C4.5 outperformed WEKA's implementation of a neural network and an SVM with standard kernels (results not shown).

Success with a classification task depends primarily on identifying appropriate features for the task. Even when a DNA property such as curvature or SIDD is known to play a role in many promoters, finding an appropriate machine learning representation for that feature is not necessarily trivial. We trialed many representations for stacking energy, curvature and SIDD in the vicinity of promoters, the most promising of which are shown in Table 2 along with more obvious features such as TSS-GSS distance.

Table 2. Information Gain merit scores [20] obtained for a range of potential promoter attributes ranked in order of merit. Note that the neighbourhood of a TSS candidate (attributes 4, 10 & 11) is the region -80 to +20 *wrt* the TSS. The promoter upstream region (attributes 8, 9 & 10) refers to -80 to -1 *wrt* the TSS.

Attribute ID	Attribute description	Merit Score
1	Rank of candidate TSS at Step 1.	0.113
2	Distance of candidate TSS from GSS.	0.072
3	Match score of candidate -10 element at Step 1.	0.068
4	Av. similarity score in neighbourhood of candidate TSS at Step 1.	0.066
5	Combined similarity score of candidate TSS at Step 1.	0.061
6	Match score of candidate -35 element at Step 1.	0.028
7	Distance of candidate TSS from position of max. curvature.	0.022
8	GC content of the promoter upstream region.	0.015
9	Stacking energy of the promoter upstream region.	0.014
10	Maximum SIDD gradient in neighbourhood of candidate TSS.	0.008
11	Minimum SIDD value in neighbourhood of candidate TSS.	0.005
12	Maximum curvature in the promoter upstream region.	0.003
13	Is candidate TSS located in low SIDD region? (Boolean)	0.002
14	Is candidate TSS located in the lowest intergenic SIDD region?	0.001

WEKA offers a number of statistical tests to evaluate the efficacy of an attribute when used in isolation for a classification task. Table 2 displays the Information Gain merit scores [20] obtained for a range of potential promoter features ranked in order of merit. The first six features include TSS-GSS distance and various similarity scores obtained from step 1 but do not include DNA structural features, which received low merit scores. The best structural DNA feature was distance of the putative TSS from the position of maximum DNA curvature. Of interest is that maximum SIDD gradient in the vicinity of a promoter was a better feature than minimum SIDD value, so confirming an observation that many TSSs are located near the downstream boundary of a SIDD region.

Based upon the merit scores shown in Table 2, we trained a C4.5 classifier using the first six attributes. This classifier achieved a recall of 50.6% and a precision of 54.0% (Table 3, row 2) on the Step Two task. This was a significant improvement over the results obtained in Step One using the BioPatML promoter description alone (Table 3, row 1). However most of the structural DNA features (such as 7, 8, 9 in Table 2), when added to the basic six features, degraded classification performance. We found two SIDD features and one curvature feature that slightly increased performance when added to the basic six (Table 3, rows 3, 4, 5) but the increase was not significant.

Table 3. Recall and precision for the promoter prediction task obtained after step 1 (selecting promoter candidates) and step 2 (classification of candidates). Feature ID numbers refer to those used in Table 2. Averages are over 10 repeats of 10 fold cross-validation. 95% confidence intervals are shown for the output from step 2.

Step	Feature representation	Include DNA structural features?	Recall	Precision
1	BioPatML	No	42.3%	46.9%
2	1-6	No	50.6±1.6%	54.0±1.4%
2	1-6, 12	Yes	51.1±1.6%	53.9±1.3%
2	1-6, 13, 14	Yes	51.5±1.4%	54.5±1.4%
2	1-6, 12, 13, 14	Yes	51.4±1.4%	54.8±1.4%

A comparison of our best result with previously published results for the TSS prediction task is not straight forward. The difficulty is that there is no standard promoter prediction task and results are sensitive to task definition and the constraints applied. First we must address the issue of task definition. The 80% accuracy achieved by Wang and Benham [19] is readily explained by their task definition. Their positive instances consisted of 500 known promoter sequences which were considered against a set of negative sequences obtained from 500 coding regions and a further 500 convergent non-coding regions, thus yielding a positive-negative ratio of 1:2. The task was then to classify sequences as containing a TSS or not. We have previously argued [5] that this is not the real promoter prediction task because promoters are seldom found in coding sequences or in convergent non-coding regions. The promoter prediction task as addressed in [3, 4, 7 and 11] is to determine the location of promoters/TSSs in regions upstream of gene start sites, since this is where the great majority of promoters are to be found. It is also a much more difficult task because the prediction algorithm must sift through many strong candidates, the majority of which prove to be false instances.

Even where authors agree on the task definition, interpretation is clouded by varying task constraints. Three factors in particular are relevant: (1) the definition of a true positive, (2) the length of the searched upstream region and (3) explicit use (or otherwise) of the TSS-GSS distance. With regard to (1), typically a true positive is a predicted TSS five bp or less from a true TSS. This margin of error is considered acceptable because biological confirmation of an *in silico* prediction does not require greater accuracy. Obviously if the error threshold is tightened, the task becomes more difficult. With regard to (2), 91% of confirmed *E. coli* sigma-70 TSSs are located within 250 bp of the GSS and consequently, most investigations restrict their search to this region. Increasing the distance to 500 bp or more increases the task difficulty as it increases the opportunity to make false positive errors. The search distance also influences the relative performance of algorithms. Using TSS-GSS distance alone as a predictor compares favorably with PWMs where the search distance is 750 bp but not if it is 250 bp [4]. With regard to (3) it has already been noted that prediction accuracy can be increased using TSS-GSS distance because most TSSs are located in a region 30 to 60 bp upstream of the TSS. Whether one considers TSS-GSS distance a valid attribute for this task depends on whether one's goal is to model the behaviour of the RNA polymerase holoenzyme or to find promoters by any means possible.

Table 4. A comparison of the task constraints, recall and precision for several studies of the well defined TSS prediction task.

Authors	Method	Recall	Precision	Error threshold	Explicit use of TSS-GSS distance	Search length
Huerta et al [7]	PWMs	50%	33%	5bp	Yes	250bp
Burden et al [3]	Neural network	50%	17%	3bp	Yes	500bp
Gordon et al [4]	SVM	50%	33%	5bp	No	750bp
Maetschke et al [11]	PWMs+EM	48%	48%	5bp	Yes	250bp
This study	PWMs+C4.5	51%	55%	5bp	Yes	250bp

Our recall and precision values for the TSS prediction task (in Table 3) are compared with four previous sets of published results (see Table 4). Huerta *et al.* [7] in 2003 claimed 'the highest predictive capability reported so far' with a recall of 50% at a precision of 33% (these values are derived from Figure 8e in [7]). We regard this as the benchmark result for a standard set of realistic constraints. Burden *et al.* [3] in 2005 using neural nets, obtained a weaker result probably because they set themselves more difficult task constraints. Gordon *et al.* [4] also obtained a recall of 50% at a precision of 33% but since they searched 750 bp upstream, their task was also notably more difficult. Maetschke *et al.* [11] obtained similar recall but at the significantly higher precision of 48% using an expanded promoter description whose parameters were trained using an Expectation Maximization (EM) approach. Our results offer a modest increase in recall and precision over [11] and therefore represent to our knowledge, the best published result for this task and this set of realistic constraints.

3.2 *SIDD and Promoter Type*

Using the information supplied in RegulonDB [24], we determined the promoter boundaries (-80 to +20 wrt TSS) for all biologically mapped sigma-70, sigma-24, sigma-38, sigma-32 and sigma-54 TSSs. For each TSS, we determined the minimum SIDD value inside its promoter boundaries as defined above. The histograms in Figure 2 show, for each promoter type, the relative frequency of promoters having a given minimum SIDD value. Wang and Benham [19] show similar data but as a probability distribution for all 927 mapped TSS locations in RegulonDB [24]. When we group the promoters according to type, we observe a somewhat uniform distribution of SIDD values for all types except sigma-54, which has 57% of its promoters associated with a SIDD value of less than zero. Only one of the 14 mapped sigma-54 promoters has a minimum SIDD value greater than the strong/weak threshold of 6 kcal/mole set in [19]. It is not surprising that sigma-54 promoters require increased upstream duplex destabilisation. Transcription initiation with sigma-54 requires activation by an enhancer binding protein which binds upstream of the promoter and resulting in interaction with sigma-54 mediated by DNA bending [21].

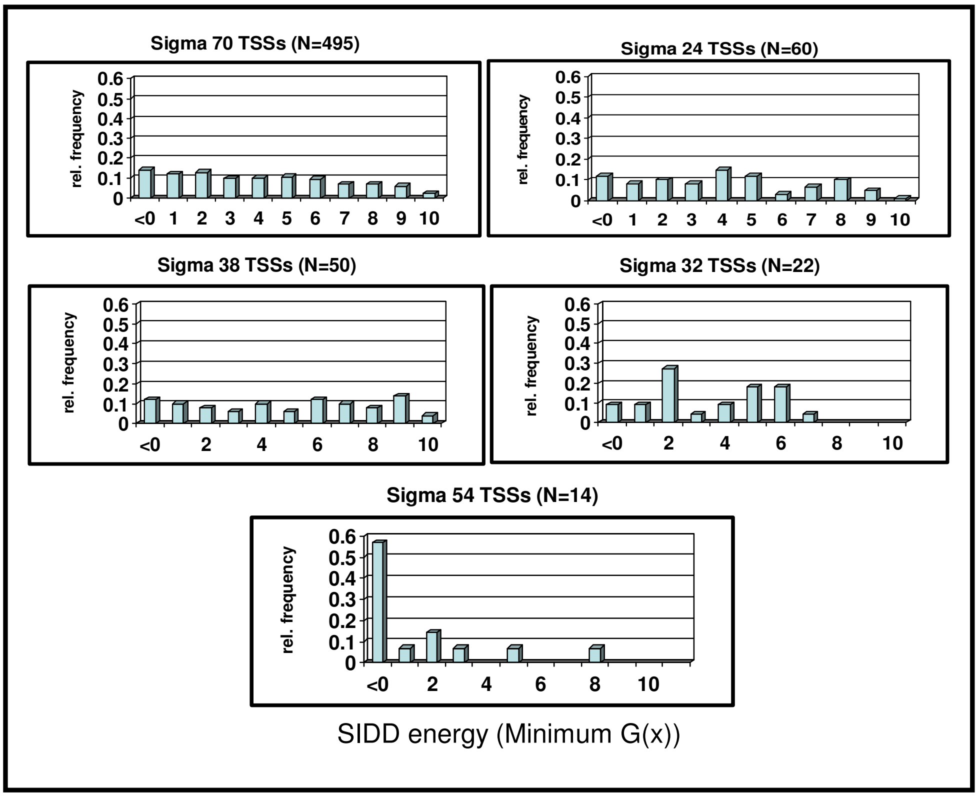

Figure 2. Histograms of the minimum SIDD value associated with five different types of promoter in *E. coli*.

4. Conclusion

We report in this work the best results to date for a well defined and realistic TSS prediction task. We have also demonstrated that sigma-54 promoters have a stronger association with SIDD regions than do other promoter types. Although DNA structural properties are known to be important in the regulation of many sigma-70 promoters, we were not able to find a suitable representation of these features that helped to increase the *in silico* prediction of sigma-70 promoters. This requires some explanation.

It should be noted that the TSS/promoter task that we undertake in this paper is inherently difficult because it involves the selection of a true TSS from of a set of strong candidates. We find that SIDD regions are generally wide enough to contain several strong candidates and therefore a local SIDD value will not be discriminative. Likewise, regions of high curvature in an intergenic region are sufficiently long and numerous that they do not have strong discriminative value. Finally any selection of promoter features implicitly biases the training towards a particular promoter model but it is known that there are many variations on how promoters initiate transcription, so it is unlikely that any one model or set of features can serve as a general purpose predictor of all the known promoters.

Acknowledgments

This work was supported by the Australian Research Council. The authors would like to acknowledge Dr Craig Benham for providing the *E. coli* genome SIDD data and Dr. Alexander Bolshoy for providing the CURVA software.

References

[1] Baldi, P., Chauvin, Y., Brunak, S., Gorodkin, J. and Pedersen, A., Computational Applications of DNA Structural Analysis, *Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology* (ISMB98), 35-42, 1998.

[2] Bi, C. & Benham, C., WebSIDD: Server for Prediction of Stress-induced Duplex Destabilized Sites in Superhelical DNA, *Bioinformatics*, 20, 1477-1479, 2004.

[3] Burden, S., Lin, Y.-X. and Zhang, R. Improving promoter prediction for the NNPP2.2 algorithm: a case study using *Escherichia coli* DNA sequences, *Bioinformatics* 21(5):601-607, 2005

[4] Gordon, J., Towsey, M., Hogan, J., Mathews, S. and Timms, P., Improved prediction of bacterial transcription start sites, *Bioinformatics* 22(2):142-148, 2006.

[5] Gordon, J. and Towsey, M. SVM based prediction of bacterial transcription start sites, *Proceedings 6th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL'05)*, Brisbane, Australia, July 2005. *Lecture Notes in Computer Science*, **3578**:448-453, Springer, Berlin, 2005.

[6] Greishaber, N., Sager, J., Dooley, C., Hayes, S. and Hackstadt, T., Regulation of the *Chlamydia trachomatis* Histone H1-Like Protein Hc2 is IspE Dependent and IhtA Independent, *J. Bact.*, 88(14): 5289–5292, 2006.

[7] Huerta, A. and Collado-Vides, J. Sigma-70 promoters in E. coli: specific transcription in dense regions of overlapping promoter-like signals, *J. Mol. Biol.*, **333:**261-278, 2003.

[8] Kawano, M., Storz, G., Rao, B., Rosner, J. and Robert G. Martin, Detection of low level promoter activity within open reading frame sequences of Escherichia coli, *Nucleic Acids Research*, 33(19):6268-6276, 2005.

[9] Kozobay-Avraham, L., Hosid, S. & Bolshoy, A., Involvement of DNA curvature in intergenic regions of prokaryotes, *Nucleic Acids Research*, 34(8):2316-2327, 2006.

[10] Maetschke, S., Towsey, M. and Hogan, J., BioPatML – an XML description language for patterns in biological sequences, Technical Report, http://eprints.qut.edu.au/archives/00006367, 2007.

[11] Maetschke, S., Towsey, M. and Hogan, J., Bacterial promoter modeling and prediction for *E. coli* and *B. subtilis* with Beagle, *Workshop on Intelligent Systems for Bioinformatics (WISB-2006)*, 9-13, 2006.

[12] Mitchell, J., Zheng, D., Busby, S. and Minchin, S., Identification and analysis of ‘extended -10’ promoters in Escherichia coli, *Nucleic Acids Research* 31(16):4689-4695, 2003.

[13] Perez-Martin, J., Rojo, F. and de Lorenzo, V., Promoters Responsive to DNA Bending: a Common Theme in Prokaryotic Gene Expression, *Microbiological Reviews*, 58(2):268-290, 1994.

[14] Quinlan, J., *C4.5: Programs for machine learning*, San Francisco: Morgan Kaufmann, 1993.

[15] Schneider, T., Stormo, G., Gold, L. and Ehrenfeucht, A., Information content of binding sites on nucleotide sequences, J. Mol. Biol. 188(3):415-431, 1986.

[16] Shultzberger, R., Chen, Z., Lewis, K. and Schneider, T., Anatomy of Escherichia coli σ^{70} promoters, *Nucleic Acids Research* 35(3):771-788, 2007.

[17] Towsey, M., Gordon, J. and Hogan, J. The Prediction of Bacterial Transcription Start Sites using Support Vector Machines, *International Journal of Neural Systems* **16**(5):363-370, 2006.

[18] Typas, A. and Hengge, R. Differential ability of sigma(s) and sigma70 of *Escherichia coli* to utilize promoters containing half or full up-element sites, *Mol. Microbiol*, 55(1):250-260, 2005.

[19] Wang, H. and Benham, C., Promoter prediction and Annotation of Microbial Genomes Based on DNA Sequence and Structural Responses to Superhelical Stress, *BMC Bioinformatics*, 7:248, 2006.

[20] Witten, I. and Frank, E., *Data Mining: Practical machine learning tools and techniques*, 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

[21] Xu, H. and Hoover, T., Transcriptional regulation at a distance in bacteria. *Current Opinion in Microbiology*, 4:138-144, 2001.

[22] http://www.genomecenter.ucdavis.edu/benham/sidd/index.php

[23] ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria/Escherichia_coli_K12/U00096.gbk

[24] http://regulondb.ccg.unam.mx/data/PromoterSet.txt

[25] http://arep.med.harvard.edu/ecoli_matrices/