Computational Biology and Chemistry 32 (2008) 359–366

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Research Article

The cross-species prediction of bacterial promoters using a support vector machine

Michael Towsey[a,b,c], Peter Timms[a,c], James Hogan[b], Sarah A. Mathews[a,c,*]

[a] School of Life Sciences, Faculty of Science, Queensland University of Queensland, GPO Box 2434, Brisbane, Queensland 4001, Australia
[b] School of Software Engineering and Data Communications, Faculty of Information Technology, Queensland University of Technology, GPO Box 2434, Brisbane, Queensland 4001, Australia
[c] Institute of Health and Biomedical Innovation, Cnr Blamey Street & Musk Avenue, Kelvin Grove Urban Village, Kelvin Grove Brisbane, Queensland 4059, Australia

ARTICLE INFO

Article history:
Received 27 June 2007
Received in revised form 1 May 2008
Accepted 6 July 2008

Keywords:
Transcript start site
σ^{70}
Promoter
Support vector machine

ABSTRACT

Due to degeneracy of the observed binding sites, the *in silico* prediction of bacterial σ^{70}-like promoters remains a challenging problem. A large number of σ^{70}-like promoters has been biologically identified in only two species, *Escherichia coli* and *Bacillus subtilis*. In this paper we investigate the issues that arise when searching for promoters in other species using an ensemble of SVM classifiers trained on *E. coli* promoters. DNA sequences are represented using a tagged mismatch string kernel. The major benefit of our approach is that it does not require a prior definition of the typical −35 and −10 hexamers. This gives the SVM classifiers the freedom to discover other features relevant to the prediction of promoters. We use our approach to predict σ^{A} promoters in *B. subtilis* and σ^{66} promoters in *Chlamydia trachomatis*. We extended the analysis to identify specific regulatory features of gene sets in *C. trachomatis* having different expression profiles. We found a strong −35 hexamer and TGN/−10 associated with a set of early expressed genes. Our analysis highlights the advantage of using TSS-PREDICT as a starting point for predicting promoters in species where few are known.

1. Introduction

The identification of bacterial promoters is an essential step in the elucidation of gene regulation and therefore in understanding the life-cycles and the virulence mechanisms of pathogens. As a general rule, the more complex the life-cycle and environmental niche of a bacterium, the greater the number of sigma factors (σ-factors) with corresponding promoter types. Typically however, the most common promoter type is that which regulates the house-keeping genes and the corresponding major σ-factor is shared by all bacteria (σ^{70} in the well studied *Escherichia coli*, and its homologues in other species).

The binding site for the σ^{70}-family of promoters is defined by two consensus hexamers, TTGACA and TATAAT, located at approximately −35 and −10, respectively relative to the transcript start site (TSS) and spaced 15–21 base pairs (bp) apart. RNA polymerase core enzyme associates with the major σ-factor to form the holoenzyme which in turn binds to its cognate promoters to initiate transcription (Browning and Busby, 2004). In addition, transcription factors can bind specific sequences surrounding or overlapping the promoter to either activate or repress transcription (Lloyd et al., 2001).

Molecular techniques for the identification of promoters are both costly and time consuming, hence *in silico* methods are an attractive and well explored alternative. The most common *in silico* method to identify σ^{70} promoters uses position weight matrices (PWMs) and depends on the relative conservation of the −35 and −10 hexamers. Through a judicious combination of PWMs and molecular experimentation, the promoters for σ^{32} and σ^{E} in *E. coli* have been well characterised (Nonaka et al., 2006; Rhodius et al., 2006). But the known σ^{70} binding sites exhibit a greater variation from the consensus than the known σ^{32} or σ^{E} binding sites. This degeneracy of the σ^{70} binding sites means that their *in silico* prediction remains a significant challenge. Successful use of PWMs requires careful choice of a threshold but degeneracy ensures that even the optimum PWM threshold results in unacceptably high numbers of false positive and false negative predictions.

An additional problem is that a satisfactory number of σ^{70} binding sites (that is, sufficient for the purpose of training machine

Abbreviations: TSS, transcript start site; SVM, support vector machine; PWM, position weight matrix; RBS, ribosomal binding site; GSS, gene start site; IC, information content.

* Corresponding author at: Institute of Health and Biomedical Innovation, Cnr Blamey Street & Musk Avenue, Kelvin Grove Urban Village, Kelvin Grove Brisbane, Queensland 4059, Australia. Tel.: +61 7 31386263; fax: +61 7 31386030.
E-mail address: s.mathews@qut.edu.au (S.A. Mathews).

1476-9271/$ – see front matter.
doi:10.1016/j.compbiolchem.2008.07.009

learning algorithms) has been identified in only two species, *E. coli* and *Bacillus subtilis*. There is thus good reason to transfer *in silico* promoter recognition across species but such an attempt requires careful consideration. This is the issue under investigation in this paper.

An alternative method to the discovery of promoters by direct application of promoter models is phylogenetic footprinting. Although a potentially powerful technique for cross-species recognition of promoter binding sites, this approach also presents significant problems. In particular, it requires a careful choice of threshold parameters and of the comparison genomes. If the comparison genomes are too similar to or too dissimilar from the target genome, the method will fail to isolate functionally important conserved sites. Grech et al. (2007) have, however, successfully used a combination of phylogenetic footprinting and PWMs to identify some new σ^{66} promoters in *Chlamydia trachomatis* (σ^{66} is the *C. trachomatis* homologue of σ^{70}).

Our research group has previously published a support vector machine (SVM) algorithm for the *in silico* prediction of TSSs in *E. coli* and their associated promoters (Gordon et al., 2006). The method achieved *state of the art* accuracy but it was only tested on promoters in the same organism (*E. coli*) on which it was trained. In this paper we illustrate the issues that arise when attempting to use the same SVM classifier trained on *E. coli* to predict the location of promoters in other species. We chose *B. subtilis* and *C. trachomatis* as the target species. The former has a large number of identified σ-factor binding sites and therefore offers a useful test of our approach. The latter is of particular interest because the standard genetic manipulation techniques to identify promoters do not work with the *Chlamydia* species.

1.1. The Bacillus Species

The *Bacillus* species are rod shaped gram-positive bacteria. *B. subtilis* is a ubiquitous but harmless bacterium, whose molecular and cell biology is the most well understood of any single-celled organism (Graumann, 2007). Other *Bacillus* species are of great social and economic importance; for example, *B. anthracis* (the plague pathogen), *B. cereus* (a gastro-intestinal pathogen) and *B. thuringiensis* (an important insect pathogen). *B. subtilis* has 18 known or predicted sigma factors (http://dbtbs.hgc.jp/), the homologue of σ^{70} being SigA (σ^{A}). The further *in silico* characterisation of the σ^{A} promoters in *B. subtilis* is important because it provides useful leverage for promoter prediction in the other significant *Bacillus* species.

1.2. The Chlamydia Species

The *Chlamydia* species are obligate intracellular parasites. They are ubiquitous pathogens, causing significant morbidity in humans and other animal hosts where infection can lead to ocular, genital, respiratory and cardiovascular disease (Hogan et al., 2004). The *Chlamydia* are characterised by a unique developmental cycle which involves inter-conversion between the infectious *elementary body* and the non-infectious metabolically active *reticulate body* (extensively reviewed by Abdelrahman and Belland (2005)). In contrast to free-growing bacteria, the *Chlamydia* are difficult to purify in large quantities and are resistant to transformation by standard genetic techniques (Mathews et al., 1999). Hence, relatively little is known about gene regulation in the *Chlamydia* (Mathews and Timms, 2006). The species have only three σ-factors and in the case of *C. trachomatis*, only 26 promoters have been verified (summarised in Tan, 2006; Hefty and Stephens, 2007; Grech et al., 2007). Nevertheless, the modest genome size (~1.2 kbp) and the availability of 10 complete genome sequences (http://www.tigr.org) for *Chlamydia* should make this organism attractive for *in silico* interrogation.

2. Materials and Methods

- *E. coli* K-12 MG1655 (NCBI Accession number NC_000913.1 GI:16127994):
  The list of published TSSs for *E. coli* was obtained from the RegulonDB database (http://regulondb.ccg.unam.mx/data/PromoterSet.txt) version 4.0 corresponding to genome file NC_000913.1. As described in Gordon et al. (2006), we selected 450 of the 676 mapped σ^{70} TSS locations where the upstream gene did not overlap with any other extracted sequence. The database has since been updated to correspond to genome file U00096.2.
- *B. subtilis* (NCBI Accession number NC_000964.2 GI:50812173):
  A list of known *B. subtilis* promoters was obtained from DBTBS, a "database of transcriptional regulation in *B. subtilis*" (Release 4, http://dbtbs.hgc.jp/COG/tfac/SigA.html). We selected 205 of the 275 mapped *B. subtilis* TSSs located within 250 bp upstream of the nearest gene start site.
- *C. trachomatis* (NCBI Accession number NC_000117.1 GI:15604717).
  Lists of the mapped TSS location *C. trachomatis* can be found in Tan (2006), Hefty and Stephens (2007) and Grech et al. (2007).

2.1. Support Vector Machine Prediction of Transcript Start Sites

TSS locations were predicted using an ensemble of 40 SVMs trained and tested on the known *E. coli* σ^{70} promoters as previously described (Gordon et al., 2006). Each of the 40 SVMs was trained on the same set of positive sequences (each 200 nucleotides (nt) long) but with different sets of negative sequences, each set offset upstream or downstream from the positive set in steps of 5 nucleotides. Each positive sequence contained a known *E. coli* TSS located at the fixed position 150 within the sequence. Negative sequences contained a known TSS, but offset either upstream or downstream from position 150, in increments of 5 nt.

When presented with a previously unseen test instance, a single trained SVM classifier returns a value representing the distance of that instance from the optimal hyper-plane dividing positive and negative training examples in feature space. In our case, the ensemble returned the average of 40 SVM scores, where the higher the average score the more likely it was that the test sequence belonged to the positive class, that is, contained a TSS at or near position 150. Using the above method, we obtained a TSS-PREDICT score for all positions from −1 to −250 with respect to the gene start site (see Fig. 1). We then used a peak finding algorithm (Towsey et al., 2006) on the assumption that promoters are most likely to be found at or near the tops of high scoring peaks. The algorithm typically identifies 5–7 peaks upstream of each gene start site, which are accepted as TSS predictions in rank order of peak height.

A false positive (FP) prediction occurs when an accepted prediction (located at the top of a peak as shown in Fig. 1) is more than 5 bp from a known TSS (where this data is available). A true positive (TP) is an accepted prediction located 5 bp or less from a mapped TSS. This degree of error tolerance is acceptable in the context of TSS prediction since the discriminator length between the −10 hexamer and the TSS can be 4–14 bp (Shultzaberger et al., 2007). A false negative (FN) prediction occurs when a known TSS does not appear in the accepted predictions. Recall is defined as the percentage of known TSSs included in the accepted predictions (TP/(TP + FN)), while precision gives the percentage of accepted predictions which are TPs (TP/(TP + FP)).

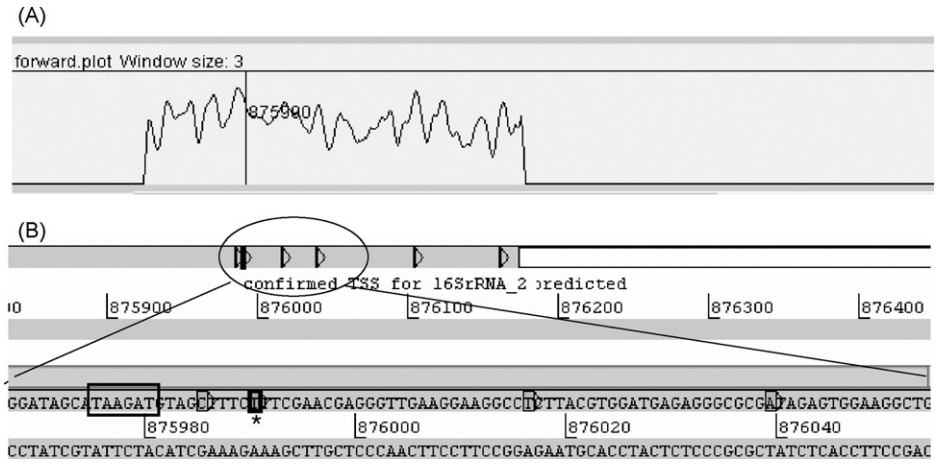

Fig. 1. Portion of the TSS-PREDICT output for the *Chlamydia trachomatis* genome displayed using the Artemis Genome Viewer. (A) Graphical output from TSS-PREDICT which has assigned a score to each of 250 bases upstream of the start site for 16SrRNA_2. Artemis smooths the curve for viewing but identification of the peaks is done on the unsmoothed data. (B) TSS-PREDICT isolated five peaks (indicated by arrow heads) which appear as TSS predictions (standard box, no bold). The true TSS is highlighted by the bold box and * and is 5 nt from the nearest prediction. The predicted promoter −10 hexamer is boxed.

2.2. Promoter Identification

Our search for promoters is confined to the 250 bp upstream of genes which are likely to be the heads of operons, defined as either (i) a gene where the nearest upstream gene is on the opposite strand, or (ii) the nearest upstream gene ends on the same strand with more than a 40 bp gap. The optimal threshold non-coding region (NCR) length for determining operon boundaries varies from one bacterium to the next, but a value of 40 bp is suitable for the three bacterial species involved in this study (Salgado et al., 2000).

Given a known or predicted TSS location, the corresponding predictions for the −10 and −35 hexamers are located using a combination of two PWMs, prepared from 250 known *E. coli* promoter sequences published by RegulonDB. These published sequences extend from the −35 hexamer to the −10 hexamer, enclosing a spacer of length 15–19 bp. The −35 and −10 PWMs were constructed using the first and last six bases of each promoter sequence, respectively. For any known or putative TSS, the −35 and −10 hexamers are located upstream of the TSS by searching for the highest combination of PWM scores, subject to two constraints:

(i) That the *spacer* length (the number of base pairs between the −35 hexamer and the −10 hexamer) should lie in the range {14–20};
(ii) That the *discriminator* length (the number of base pairs between the −10 hexamer and the putative TSS) should lie in the range {4–14}.

This approach to predicting the bound hexamers accommodates the flexibility of RNA polymerase, but leads to anomalies. In particular, shifting the putative TSS position by only one base can sometimes lead to the prediction of very different hexamers. More sophisticated promoter prediction models weight the spacer length (that is to give more weight to a gap of 17 than to a gap of 13), but these models require the fitting of additional parameters, so we chose to use the simpler promoter model. The PWM scores indicated in our results are the sum of the normalised scores for the −10 and −35 hexamers and therefore lie in the interval [0,2], with a score of 2 corresponding to the joint consensus TTGACA (−35) and TATAAT (−10).

2.3. Determining Information Content (IC) and Significant Promoter Features

To determine if a set of promoters of interest had features that differed significantly from an 'average' promoter (our null model), the known or putative promoters were divided into six regions with respect to the TSS, (i) an upstream region from −75 to −36, (ii) the −35 hexamer (−35 to −30), (iii) the first seven positions of the spacer (spacer 1, −29 to −23), (iv) the last seven positions of the spacer (spacer 2, −19 to −13) which allows for variation in spacer length, (v) the −10 hexamer (−12 to −7) and (vi) the TSS region (−6 to +4).

The average values of relevant features were calculated for each region. Relevant features included information content (IC), base frequencies and stacking energy. For the predicted −10 and −35 hexamers the IC at each of the six positions in the motifs was calculated. This analysis was derived only from rank 1 TSS predictions. The *Z*-score of a feature value was calculated with respect to the mean and standard deviation of that feature over 1000 promoter sets, each consisting of 12 randomly selected rank 1 promoter predictions.

3. Results and Discussion

3.1. In silico Whole Genome Mapping of TSSs in B. subtilis

A significant design feature of our ensemble-SVM, TSS-PREDICT, was the incorporation of minimal prior knowledge. The PWM approach depends upon defining the −10 and −35 hexamers to be found. But as has already been noted, this is not particularly helpful in the case of σ^{70} promoters, where we must assume that other features are just as important for promoter recognition. The training of TSS-PREDICT did not depend on any prior sequence definition. Rather sequences were represented using the *tagged mismatch string kernel* (Gordon et al., 2006). We used substrings of length 5 and allowed a single mismatch. Each 5-mer was 'tagged' with its location (rounded to the nearest 10 to accommodate some of the variation in motif locations) with respect to the reference point at 150, the putative TSS. Training then involves the SVM ensemble determining which of the 20,480 potential features (4^5 5-mers × 20 locations) are relevant to a successful prediction.

In order to use TSS-PREDICT (trained for recognition of *E. coli* σ^{70}-like promoters) in the genomes of other species, it is impor-

Table 1
TSS-PREDICT rankings for mapped TSS in *Bacillus subtilis* and *Chlamydia trachomatis* showing % cumulative recall and precision

Rank	*B. subtilis*			*C. trachomatis*		
	TP count	% recall	% precision	TP count	% recall	% precision
1	100	49	49	12	46	46
2	34	65	33	2	54	27
3	14	72	24	1	58	19
4	13	79	20	1	62	15
5+	8	82	<16	1	65	<13
Total	169 (205)	82		17 (26)	65	

TP, true positive.

tant that the feature set used by TSS-PREDICT to make an accurate prediction is relevant to an equivalent class of sigma factor in the new species. As previously described (Gordon et al., 2006), the ensemble-SVM gives greatest prominence to three features: (1) the presence of motifs centred on positions −35 to −40 having up to two mismatches from the consensus TTGACA; (2) the presence of motifs centred on positions −10 to −15 having up to two mismatches from the consensus TATAAT; and (3) the presence of motifs centred on +15 to +25 positions having up to two mismatches from the start codon (ATG) and the RBS consensus (AGGAGGT). These first two features are not surprising but the last is interesting and arises because the distribution of distances from the TSS to nearest downstream gene start site (hereafter referred to as the TSS-GSS distribution) has a dominant peak at 30 bp (Fig. 2A). The importance of the proximity of the TSS (and therefore the promoter) to the gene start site (and RBS) has been observed experimentally since expression from σ^{70} promoter constructs was shown to be greater when an RBS was inserted immediately downstream of a promoter (Kawano et al., 2005).

We infer that TSS-PREDICT can be used to search for σ^{70}-like promoters in other bacterial species if two conditions are satisfied: (1) the two DNA binding domains of the major σ-factor, domains 2.4 and 4.2 corresponding to the −35 and −10 recognition sites, respectively (Campbell et al., 2002) are similar to those of *E. coli* σ^{70}, and (2) the distribution of TSS-GSS separation has a strongly dominant peak near the −30 position.

In the case of *B. subtilis*, the major σ-factor (σ^{A}) has DNA binding domains homologous to those of *E. coli* σ^{70}, and the −10 and −35 PWMs derived from 205 known σ^{A} promoters have high information content at the same positions as the *E. coli* σ^{70} promoters. In addition, the TSS-GSS distribution for the 205 known σ^{A} TSSs in *B. subtilis* (Fig. 2B) is comparable to that of *E. coli* (Fig. 2A) with both having a strong peak near −30 bp. When we applied TSS-PREDICT to the upstream non-coding regions (NCRs) of *B. subtilis*, the rank 1 predictions achieved 49% recall (at 49% precision) of 205 known TSS locations, and the rank 2 predictions recalled an additional 16% (Table 1). Although this prediction accuracy for the rank 1 promoters is slightly less than that achieved by Maetschke et al. (2006) who trained PWMs using a modified *expectation maximisation* algorithm, it should be noted that their method incorporated prior knowledge of the consensus hexamers and was trained and tested on the same organism. The TSS-PREDICT algorithm is more flexible in both respects and we have demonstrated that it can be used for cross-species prediction of σ^{70} promoters.

Of course in the above experiment, determination of the TSS-GSS distribution in *B. subtilis* depended on knowing the location of 205 σ^{A} promoters. What can we do when investigating species where few promoters are known? The two conditions referred to above must still be satisfied. We may easily confirm the homology of the DNA binding domains of the target σ-factor by comparing the relevant amino-acid sequences. We must, however approach

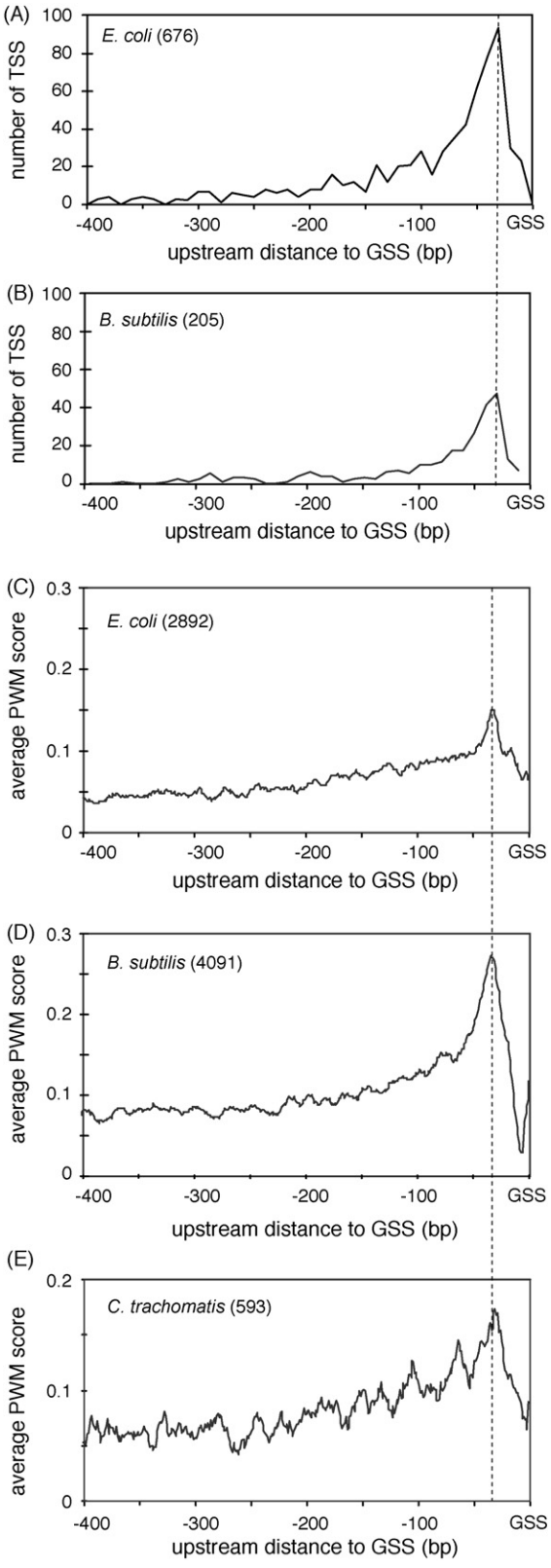

Fig. 2. Using average PWM score as a qualifier for TSS location. TSS location score versus distance to the gene start site for 676 *Escherichia coli* (A) and 205 *Bacillus subtilis* (B) promoters with the maximum TSS locations indicated by the dotted line. Average PWM score versus distance to the gene start site for 2892 *E. coli* (C), 4091 *B. subtilis* (D) and 593 *C. trachomatis* (E) genes (with non-overlapping sequence) with the highest average PWM score position indicated by the dotted line.

Table 2
Comparison of the published[a] to the TSS-PREDICT *C. trachomatis* promoters

Gene	Status	Rank	−35	Spacer (nt)	−10	TSS score	PWM score	Mismatches[b]	TSS location	Error (nt)[c]
infA(*tufA*)	Known		TTGACA	16	TATAAT		2.00		363842	
	*Rank	1	TTGACA	16	TATAAT	2.76	2.00	0	363844	2
dnaK P1(*hrcA*)	Known		TTGACC	17	TATAAT		1.87		449836	
	*Rank	1	TTGACC	17	TATAAT	2.49	1.87	1	449837	1
rpsL	Known		TTGCAA	18	TATATT		1.72		508559	
	*Rank	1	TTGCAA	18	TATATT	2.69	1.72	3	508561	2
tRNAThr2	Known		TTGATA	18	TACTAT		1.71		363424	
	*Rank	1	TTGATA	18	TACTAT	2.06	1.71	3	363424	0
rl14	Known		TTGTTG	15	TATACT		1.65		591695	
	*Rank	1	TTGTTG	15	TATACT	2.57	1.65	4	591693	2
tyrS	Known		TTGCTA	18	TAAGAT		1.65		71883	
	*Rank	1	TTGCTA	16	GATAAG	1.96	1.45	4	71880	3
yscU	Known		TTGAGA	17	TAACTT		1.54		106571	
	*Rank	1	TTGAGA	17	TAACTT	2.12	1.54	4	106572	1
clpC	Known		TTGCAT	19	TATGCT		1.54		317942	
	*Rank	1	TTGCAT	19	TATGCT	2.00	1.54	5	317942	0
ltuA	Known		TGCAGA	19	TATAAT		1.51		430530	
	*Rank	1	TGCAGA	19	TATAAT	2.13	1.51	3	430530	0
CT708	Known		TTGATT	17	TACAAG		1.51		814833	
	*Rank	1	TTGATT	17	TACAAG	1.66	1.51	4	814833	0
16S rRNA P2	Known		TGCATA	17	TAAGAT		1.38		875990	
	*Rank	1	TGCATA	17	TAAGAT	1.76	1.38	5	875985	5
crpA	Known		TAGATG	16	AAGTGT		1.12		511823	
	*Rank[d]	1	TTGAAA	16	TAGACT	3.03	1.74	3	511819	4
omcA P1	Known		TTGATA	19	TAATAT		1.73		514137	
	Rank	1	TTGTTA	19	TAAAAG	2.07	1.55		514080	
	*Rank	2	TTGATA	19	TAATAT	1.55	1.73	1	514137	0
rs1	Known		TTGCCT	17	TACACT		1.62		115708	
	Rank	1	TGGGGA	17	TATAAT	2.05	1.51		115836	
	*Rank	2	TTGCCT	17	TACACT	1.93	1.62	4	115706	2
rpoD	Known		TAGATT	17	TAAACT		1.47		695790	
	Rank	1	CAGTTT	16	TATATT	2.18	1.26		695898	
	*Rank	3	TAGATT	17	TAAACT	1.29	1.47	5	695791	1
CT602	Known		TAGATA	17	TAGGAT		1.53		681191	
	Rank	1	TTGATG	17	TATATT	2.62	1.70		680976	
	*Rank	4	TAGATA	17	TAGGAT	1.70	1.53	4	681189	2
hctA	Known		TTGCAT	16	ACTAAT		1.36		863304	
	Rank	1	TTCAGA	17	TTTTAT	1.61	1.43		863439	
	*Rank[d]	5	ATGAAT	19	AAAAAT	1.23	1.38	5	863301	3

*Rank prediction that matches known promoter.
[a] Published σ^{66} promoters not predicted by TSS-PREDICT are for *yscJ*, exbB, *omcA* P2, 16S rRNA P1, *dnaK* P2, *efp2*, *ltuB*, *ompA* P1 and P2.
[b] Mismatches of promoter hexamers to the eubacterial consensus.
[c] Error (nt) is the difference in position of the mapped and TSS-PREDICT TSS with up to 5 nt difference allowed for a match in prediction.
[d] Hexamers listed for TSS-PREDICT do not match the published promoters, the TSS location mapped correctly but the PWM algorithm found a higher scoring −35 and −10 pairs.

the TSS-GSS distribution indirectly. We prepared a simple model of the canonical σ^{70} promoter consisting of a −35 PWM, a variable spacer and a −10 PWM. We calculated the distribution of the optimum promoter PWM scores (sum of −35 and −10 PWM scores) for each of the 400 positions immediately upstream of all target genes. We may use this distribution as a surrogate for the TSS-GSS distribution because *on average* high scores will emerge where there is a concentration of promoters. Fig. 2C shows the average PWM scores assigned to each position from −1 to −400 with respect to the gene start site for the 2892 *E. coli* genes likely to contain a promoter. This curve shows a distinct peak close to the −40 position, consistent with the highest frequency TSS-GSS position at −30 (Fig. 2A). Indeed, taking the log of the TSS-GSS distribution yields a curve very similar to the distribution of PWM scores. A similar curve was obtained for 4091 upstream sequences analysed for *B. subtilis* (Fig. 2D). Curves of average PWM scores can easily be obtained for new species and any curve peaking at the same place would imply a TSS-GSS distribution similar to that for *E. coli*. This approach relies on the comparatively small variance of the discriminator length compared to the 400 bp upstream region under consideration.

3.2. *In silico* Mapping TSS and Whole Genome Promoter Prediction in *C. trachomatis*

Given the limited TSS and promoter data available for *C. trachomatis* (only 26 known promoters) and the importance of understanding regulatory mechanisms, we used *in silico* prediction to investigate the promoters of this important bacterium. We know the major σ-factor in *C. trachomatis* (σ^{66}) has conserved σ^{70}-like DNA binding domains (Koehler et al., 1990). The distribution

Table 3
Top 20 *C. trachomatis* rank 1 TSS-PREDICT promoters with the highest TSS and PWM scores

Gene	TSS score	PWM score	−35	−10
A. Highest TSS score				
tyrP_2	3.03	1.63	TTGCTA	TAGTAT
crpA[a]	3.03	1.74	TTGAAA	TAGACT
tRNAPhe	2.98	1.74	TTGATT	TAGAAT
CT035[b]	2.97	1.93	TTGATA	TATAAT
tRNAArg1	2.94	1.58	TTGATC	TATCCT
accA	2.94	1.37	TTGGTT	TATTAA
artJ	2.93	1.54	TTTTTT	TATAGT
gcpE	2.86	1.51	TTCCTT	TATACT
CT345	2.83	1.78	TTCTCA	TATAAT
tRNAArg2	2.82	1.76	TTGATT	TAAAAT
CT255	2.76	1.57	TTGAGA	TTTTAT
CT481	2.76	1.40	TTGATT	CATAGC
infA[a,b]	2.76	2.00	TTGACA	TATAAT
oppA	2.76	1.46	GTGTAT	TATACT
tRNAMet1	2.76	1.59	TTGCAT	GATAAT
ytfF	2.75	1.81	TTTACG	TATAAT
hemL	2.74	1.45	TTTTTT	TACACT
rpoC	2.71	1.44	TTGATG	GATGCT
CT683	2.70	1.43	TTGCTC	TATTAC
rpsL	2.69	1.72	TTGCAA	TATATT
B. Highest PWM score				
infA[a,b]	2.76	2.00	TTGACA	TATAAT
CT035[b]	2.97	1.93	TTGATA	TATAAT
CT860	2.16	1.9	TTGACA	TACAAT
CT547	2.00	1.88	TTGACA	TATGAT
hrcA	2.49	1.87	TTGACC	TATAAT
CT646	2.20	1.86	TTGAAA	TAAAAT
CT249	2.60	1.85	TTGATA	TAAAAT
tRNAAla1	1.80	1.85	TTGATA	TAAAAT
dnaA	2.11	1.85	TTGTTA	TATAAT
CT343	2.55	1.85	TTGAAT	TATAAT
CT556	2.62	1.84	TTGATT	TATAAT
pkn5	1.73	1.84	TTGAAA	TACAAT
CT846	2.40	1.83	TTGACA	GATAAT
ypdP	2.67	1.83	TTGCCG	TATAAT
murE	2.10	1.82	TTGACA	TAAACT
fliN	2.19	1.82	TTGAAA	TATGAT
CT016	1.96	1.82	TTGTCA	TACAAT
ptsH	1.68	1.82	TTGAAA	TATTAT
ytfF[b]	2.75	1.81	TTTACG	TATAAT
CT254	1.76	1.81	TTGATA	TATGAT

[a] Promoters have been previously published.
[b] Promoters fall into both highest TSS and highest PWM score list.

of PWM scores for the upstream non-coding regions of the 593 *C. trachomatis* genes (Fig. 2E) reveals a clear maximum at position −40, which aligns well with *E. coli* and *B. subtilis*. The *C. trachomatis* plot displays more variation due to the smaller number of genes available (593 versus 2892 and 4091, respectively for *E. coli* and *B. subtilis*).

TSS-PREDICT was applied to the 250 bp upstream of the 593 *C. trachomatis* genes having an upstream non-coding region of at least 40 bp and promoters predicted for the *in silico* mapped TSS using a pair of PWMs for the −35 and −10 hexamers (with constraints set on the intervening spaces as described in Section 2.2). The results show that the rank 1 predictions included 12 (46%) of the 26 known *C. trachomatis* TSS and the rank 2 predictions included another two known TSS (Table 1). These accuracy levels (46% precision) are comparable to those for *B. subtilis* when considering the small sample size. A summary of the known *C. trachomatis* promoters and their associated TSS-PREDICT output is shown in Table 2 and the rank 1 and rank 2 predictions for the entire *C. trachomatis* genome are provided as Supplementary Data.

Table 3 identifies the top 20 rank 1 promoter predictions for both the highest TSS and PWM scores. Only two of the 26 biologically mapped *C. trachomatis* promoters score highly (*infA* and *crpA*). Furthermore, the correlation between TSS and PWM scores for the predicted promoters is limited, with only *infA*, CT035 and *ytfF* making both lists. This lack of correlation is not surprising, because we know that TSS-PREDICT assigns its TSS score based on features in the entire 200 bp neighbourhood from 50 bp downstream to 150 bp upstream of the predicted TSS location. In fact, this is a significant advantage to TSS-PREDICT as a predictive algorithm.

It is instructive to examine the 14 published *C. trachomatis* promoters which were not included in the list of rank 1 predictions. In seven cases (*omcA* P2, 16S rRNA P1, *dnaK* P2, *ompA* P1 and P2, CT602 and *hctA*) TSS-PREDICT found alternative −35 and −10 hexamers with greater similarity to the consensus (higher PWM score). In the remaining seven incorrect predictions (*yscJ*, *exbB*, *efp2*, *ltuB*, *omcA* P1, rs1 and *rpoD*) TSS-PREDICT selected a lower scoring pair of hexamers because the associated TSS-GSS distance was markedly more probable. In fact, in four of these cases (*yscJ*, *exbB*, *efp*2, *ltuB*) for which the average known TSS-GSS distance is 22 nt, the predicted TSSs were placed at an average distance of 40 nt from their respective gene start sites. This indicates the bias in TSS-PREDICT to locate RBS-like motifs at the +35 position with respect to the predicted TSS.

3.3. Using Whole Genome *C. trachomatis* TSS Predictions to Identify Developmental-stage-specific Regulatory Elements

The regulation of transcription initiation is highly likely to play a key role in coordination of the complex developmental cycle of *C. trachomatis*. We used the TSS-PREDICT predictions (Section 3.1) and the *C. trachomatis* genome micro-array time-course expression profiles (Belland et al., 2003) to investigate any significant motifs associated with *C. trachomatis* development. The first analysis involved mapping SVM score and PWM score against expression levels for the time post-infection when transcripts for each gene were first detected (1, 3, 8, 16 h post-infection as described in Belland et al. (2003)). No significant correlation of either SVM or PWM scores with expression patterns was observed over the entire data set (data not shown).

The next analysis step involved choosing two sets of 12 genes (predicted to be heads of operons) having different expression profiles. The goal was to identify unique promoter features that might correlate with their different expression profiles. Since *C. trachomatis* has a unique life-cycle involving the inter-conversion of elementary body and reticulate body, the first set of genes (CT147, CT474, CT529, *euo*, CT288, CT850, CT473, *map*, CT365, CT734, CT035, *cysQ*) were “early” genes (those expressed at low levels from 1 h post-infection when elementary bodies are attaching and entering the host cell). The second set of genes (CT050, *infA*, *ytgA*, *clpB*, *gcp_1*, *gapA*, CT017, CT102, *zwf*, CT560, *gcpE*, *recA*) belonged to a cluster having consistently high expression from 8 h post-infection when RBs are in an active-stage of replication (termed “mid” genes).

Significant promoter features were determined by comparison with an average set of promoters derived from 1000 sets of twelve randomly selected predicted promoters. Average expression levels and rank 1 TSS-PREDICT promoters are shown in Fig. 3A and Table 4A, respectively. Sequence logos of the predicted −10 and −35 motifs for the two sets of promoters (Fig. 3B) were broadly similar and the information content at individual positions did not differ significantly from the average of 1000 sets of promoters. This is evidence that differential expression of the two sets of genes must be due to other features associated with the two sets of promoters, features that might be detected as conserved motifs.

To look for additional significant motifs, we divided the promoters into six regions with respect to the TSS (upstream, −35 hexamer, spacer 1, spacer 2, −10 hexamer and TSS region as described previously in Section 2.3 and Fig. 3B). Interestingly, there were no

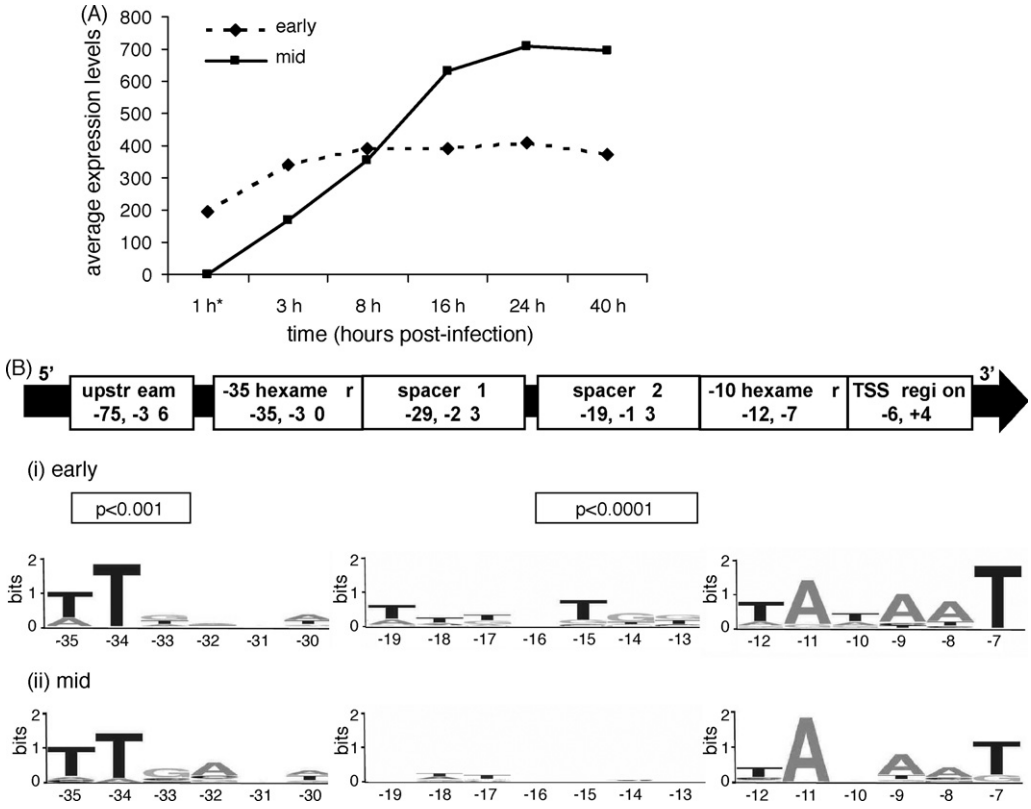

Fig. 3. Analysis of "early" and "late" gene sets for significant features surrounding the promoters. (A) Average expression levels as determined by microarray analysis (Belland et al., 2003) for the "early" (CT147, CT474, CT529, *euo*, CT288, CT850, CT473, *map*, CT365, CT734, CT035, *cysQ* with first expression at 1 h post-infection) and "mid" (CT050, *infA*, *ytgA*, *clpB*, *gcp_1*, *gapA*, CT017, CT102, *zwf*, CT560, *gcpE*, *recA* with expression consistently high from mid-development). (B) Representation of the sequence upstream of a gene by region featuring the "upstream", "−35 hexamer", "spacer 1", "spacer 2", "−10 hexamer" and "TSS region" and indicated by position (with the TSS being +1). The sequence logos for the "−35 hexamer" and "−10 hexamer" are shown for both "early" (i) and "mid" (ii) gene sets with the "spacer 2" sequence logo also shown for the "early" gene set. The significance of the IC is shown for the "early" gene set with probability indicated in the boxes directly under the defined regions.

Table 4
Promoters predicted for the early (A) and highly mid (B) genes

Gene	−35	Spacer 1	Spacer 2	−10	Discriminator	TSS region
A. Early genes						
CT147	TTTATA	tgtgagaa	tagggtat	AATACT	tcctt	caaaaatagt
CT474	ATGAGT	aaatgcagctttttttgg		TAGATT	ttacg	aagtggcgtg
CT529	TTTTCG	gtttaagtaatagaagtgg		TATAAT	atctctc	taaattttgt
euo	TTGATT	aataagtt	ttttgttgg	GAAAAT	gttacc	ttctcttttt
CT288	TTGTAA	aaaaacaa	tatttattc	TAAAAT	aata	accacagtta
CT850	TTAGCT	ttttgttaa	ccaaagcgt	TAGCTT	ctttggag	cagccctaaa
CT473	TTGCTA	tcttaggg	gacgttct	TGAAAT	tcttaat	gaccaactaa
map	TTTCTA	cggctttc	aatataggg	TATTAT	gttacac	cttaaagctc
CT365	ATGAAA	aggatttta	ttttgttgt	TATAAT	taatct	tgttgaaaga
CT734	TTTAGT	taataag	atttgtgc	TATAAT	actacaaa	tttatttttа
CT035	TTGATA	aagcgtt	ttttttgg	TATAAT	gagaaaa	agctttttgt
cysQ	ATATCG	tggcttcatatttttcgtt		AATAAT	ctttaca	gaaatccaag
B. Mid genes						
CT050	ATCAAA	aatgatta	aaaataagc	TAGTAT	agagatttaa	gaagttaaag
infA	TTGACA	ttttctgt	ttagtcga	TATAAT	cgctct	ctcgagtttc
ytgA	TTGAGG	atataaat	tcattctgt	TAAAAG	tatctt	tgcgataagc
clpB	TTGATA	gcctttgtaacaggtgtga		TAAAAG	gtatt	tatggagaaa
gcp_1	TTGCTT	gattaaca	atctcatga	TACGAT	cctctcct	tccaaaatgt
gapA	TTAAAG	ctggagat	ttgtctgct	TATAAT	gttgat	agaagagtca
CT017	TTGACT	ttttcctt	taagtcaat	AATAAT	tccctc	tctagaggat
CT102	TTGGTA	gtgcaggg	caagatata	GAGATT	tcact	aacagttttg
zwf	TAGGCA	catattcg	ttttttgg	TATTGT	gggctc	tctcttaaaa
CT560	CTTACT	cacagctc	tctttcact	CAGAAT	aagccag	ctccaaaaga
gcpE	TTCCTT	taaattga	tctaaagct	TATACT	tcctcag	ccccccaaaa
recA	TTGACT	gttgtctttt	gaaatattg	CAAACT	gacgc	cacgaaatat

significant features for the "mid" genes but there were significant features associated with the "early" genes. The first significant feature ($P < 0.01$) was the high IC in the upstream region. The second, and most significant ($P < 0.001$) region was the high IC in spacer 2 (−19 to −13), which on closer inspection revealed (see sequence logo in Fig. 3B) the presence of an extended −10 promoter which is defined by TGN immediately upstream of the −10 hexamer. The presence of the TGN motif has been shown to compensate for either a weak −35 hexamer (Browning and Busby, 2004) or a weak−10 hexamer (Hook-Barnard et al., 2006). For the early genes investigated in this study, both the −35 and −10 hexamers associated with the TGN motif have similar IC to promoters without the TGN motif. The presence of all three σ^{70} promoter features (−35, TGN, −10) with high IC is likely to be a mechanism that *C. trachomatis* uses to ensure RNA polymerase easily recognises the promoters of the genes required during the very early stages of development, when the elementary body is attaching itself to and entering the host cell. During these early events, the chlamydial DNA is still held in a nucleoid structure by histone-like proteins (Barry et al., 1993) which could effectively reduce promoter recognition by the RNA polymerase. As development progresses, and the chlamydial nucleoid de-condenses, the σ^{66} promoters with less IC would be more easily recognised by RNA polymerase and thus not require additional promoter features.

4. Conclusions

Due to degeneracy of the observed binding sites, the *in silico* prediction of bacterial σ^{70}-like promoters remains a challenging problem. A large number of σ^{70}-like promoters has been biologically identified in only two species, *E. coli* and *B. subtilis*. In this paper we have investigated the issues that arise when searching for promoters in other species using a classifier trained on *E. coli* promoters. Rather than searching for matches to the paired −10 and −35 hexamers, we searched for the most-likely TSS using an ensemble SVM (TSS-PREDICT) and then used these predictions to localise the corresponding promoter. The major benefit of this approach is that it does not require a prior definition of the consensus −35 and −10 hexamers in order to locate a TSS. This gives our SVM classifier the freedom to discover other features. We found that the most significant other feature is the TSS-GSS distribution and we describe a method to estimate this distribution in species where few or no promoters have yet been identified.

We use our approach to predict σ^{A} promoters in *B. subtilis* (which also serves as a check on the accuracy of our method) and σ^{66} promoters in *C. trachomatis*. We extended the analysis to identify specific regulatory features of gene sets in *C. trachomatis* having different expression profiles. We found a strong −35 hexamer and TGN/−10 associated with a set of early expressed genes. Our analysis highlights the advantage of using TSS-PREDICT as a starting point for predicting promoters in species where few are known.

Acknowledgements

The authors would like to thank Dr. Virgil Rhodius for useful discussions. This work was funded by an Australian Research Council Discovery Grant (DP0559750).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.compbiolchem.2008.07.009.

References

Abdelrahman, Y.M., Belland, R.J., 2005. The chlamydial developmental cycle. FEMS Microbiol. Rev. 29, 949–959.
Barry, C.E., Brickman, T.J., Hackstadt, T., 1993. Hc1-mediated effects on DNA structure: a potential regulator of chlamydial development. Mol. Microbiol. 9, 273–283.
Belland, R.J., Zhong, G., Crane, D.D., Hogan, D., Sturdevant, D., Sharma, J., Beatty, W.L., Caldwell, H.D., 2003. Genomic transcriptional profiling of the developmental cycle of *Chlamydia trachomatis*. Proc. Natl. Acad. Sci. U.S.A. 100, 8478–8483.
Browning, D.F., Busby, S.J., 2004. The regulation of bacterial transcription initiation. Nat. Rev. Microbiol. 2, 57–65.
Campbell, E.A., Muzzin, O., Chlenov, M., Sun, J.L., Olson, C.A., Weinman, O., Trester-Zedlitz, M.L., Darst, S.A., 2002. Structure of the bacterial RNA polymerase promoter specificity sigma subunit. Mol. Cell 9, 527–539.
Gordon, J.J., Towsey, M.W., Hogan, J.M., Mathews, S.A., Timms, P., 2006. Improved prediction of bacterial transcription start sites. Bioinformatics 22, 142–148.
Graumann, P. (Ed.), 2007. Bacillus: Cellular and Molecular Biology, 1st ed. Caister Academic Press, ISBN 978-1-904455-12-7.
Grech, B., Maetschke, S., Mathews, S., Timms, P., 2007. Genome-wide analysis of chlamydiae for promoters that phylogenetically footprint. Res. Microbiol. 158, 685–693.
Hefty, P.S., Stephens, R.S., 2007. Chlamydial type III secretion system is encoded on ten operons preceded by sigma 70-like promoter elements. J. Bacteriol. 189, 198–206.
Hogan, R.J., Mathews, S.A., Mukhopadhyay, S., Summersgill, J.T., Timms, P., 2004. Chlamydial persistence: beyond the biphasic paradigm. Infect. Immun. 72, 1843–1855.
Hook-Barnard, I., Johnson, X.B., Hinton, D.M., 2006. *Escherichia coli* RNA polymerase recognition of a sigma70-dependent promoter requiring a −35 DNA element and an extended −10 TGn motif. J. Bacteriol. 188, 8352–8359.
Kawano, M., Storz, G., Rao, B.S., Rosner, J.L., Martin, R.G., 2005. Detection of low-level promoter activity within open reading frame sequences of *Escherichia coli*. Nucleic Acids Res. 33, 6268–6276.
Koehler, J.E., Burgess, R.R., Thompson, N.E., Stephens, R.S., 1990. *Chlamydia trachomatis* RNA polymerase major sigma subunit. Sequence and structural comparison of conserved and unique regions with *Escherichia coli* sigma 70 and *Bacillus subtilis* sigma 43. J. Biol. Chem. 265, 13206–13214.
Lloyd, G., Landini, P., Busby, S., 2001. Activation and repression of transcription initiation in bacteria. Essays Biochem. 37, 17–31.
Maetschke, S., Towsey, M., Hogan, J., 2006. Bacterial promoter modeling and prediction for *E. coli* and *B. subtilis* with Beagle. In: Workshop on Intelligent Systems for Bioinformatics, pp. 9–13.
Mathews, S., Timms, P., 2006. In silico identification of chlamydial promoters and their role in the regulation of development. In: Bavoil, P.M., Wyrick, P.B. (Eds.), Chlamydia: Genomics and Pathogenesis. Horizon Bioscience, UK, pp. 133–156.
Mathews, S.A., Volp, K.M., Timms, P., 1999. Development of a quantitative gene expression assay for *Chlamydia trachomatis* identified temporal expression of sigma factors. FEBS Lett. 458, 354–358.
Nonaka, G., Blankschien, M., Herman, C., Gross, C.A., Rhodius, V.A., 2006. Regulon and promoter analysis of the *E. coli* heat shock factor, sigma 32, reveals a multifaceted cellular response to heat stress. Genes Dev. 20, 1776–1789.
Rhodius, V.A., Suh, W.S., Nonaka, G., West, J., Gross, C.A., 2006. Conservation and variation of the sigma E-mediated envelope stress response in related genomes. PLoS Biol. 4 (1), e2, 43–59.
Salgado, H., Moreno-Hagelsieb, G., Smith, T.F., Collado-Vides, J., 2000. Operons in *Escherichia coli*: genomic analyses and predictions. Proc. Natl. Acad. Sci. U.S.A. 97, 6652–6657.
Shultzaberger, R.K., Chen, Z., Lewis, K.A., Schneider, T.D., 2007. Anatomy of *Escherichia coli* sigma70 promoters. Nucleic Acids Res. 35, 771–788.
Tan, M., 2006. Regulation of gene expression. In: Bavoil, P.M., Wyrick, P.B. (Eds.), Chlamydia: Genomics and Pathogenesis. Horizon Bioscience, UK, pp. 103–132.
Towsey, M., Gordon, J., Hogan, J., 2006. The prediction of bacterial transcription start sites using support vector machines. Int. J. Neural Syst. 16, 363–370.