COVER SHEET

This is the author version of article published as:

Maetschke, Stefan R. and Towsey, Michael W. and Hogan, James M. (2006) Bacterial promoter modelling and prediction for E. coli and B. subtilis with Beagle. In *Proceedings Workshop on Intelligent Systems for Bioinformatics (WISB-2006)* 73, Hobart, Tasmania.

Accessed from http://eprints.qut.edu.au

Bacterial promoter modeling and prediction for *E. coli* and *B. subtilis* with Beagle

Stefan R. Maetschke   Michael W. Towsey   James M. Hogan

Faculty of Information Technology, Queensland University of Technology, GPO Box 2434, Brisbane, QLD 4001, Australia
Email: m.towsey@qut.edu.au

Abstract

We constructed σ^{70}-promoter models of varying complexity to predict promoter locations and to evaluate the importance of specific promoter elements. For this purpose, a novel software, named *Beagle*, was developed that utilizes an easy description language to conveniently specify promoter models. Model specifications are translated into position weight matrices and gap distributions which are refined using data from known promoters.

The method is transparent, fast and allows the rapid exploration of different promoter models. Applied to promoter prediction in *E. coli* and *B. subtilis*, we show that inclusion of UP-elements and extended -10 motifs into the model yields a significant increase in prediction accuracy.

The software, data sets and extended results can be downloaded at http://eresearch.fit.qut.edu.au/Beagle/.

Keywords: Beagle, Promoter, sigma-70, Escherichia coli, Bacillus subtilis

1 Introduction

Promoters are regions of DNA responsible for the initiation of gene transcription. Their identification is crucial for understanding gene regulation but they are difficult to identify *in silico* because their important functional sites are poorly conserved. Identifying promoters using wet-lab techniques is time consuming and given the exponentially growing number of sequenced genomes, there is a definite need for computational methods to detect and study promoters.

Many methods have been devised to identify promoter sites using for example, Regular Expressions (REs), Position Weight Matrices (PWMs), Hidden Markov Models (HMMs), Neural Networks (NNs) and Support Vector Machines (SVMs) (Vanet, Marsan & Sagot 1999). The different model types have strengths and weaknesses which typically involve trade-offs between accuracy, transparency, speed and ease of use. Despite (or perhaps because of) their simplicity, PWMs continue to be a frequently used approach to search for promoters. In addition their use finds theoretical justification in Information Theory (Schneider, Stormo, Gold & Ehrenfeucht 1986).

PWMs have been used in two ways to search for promoters. The direct approach is to search for DNA motifs that bind the RNA Polymerase (RNAP) holoenzyme. In the case of the σ^{70} family of bacterial promoters, with which we are solely concerned in this paper, this means having PWM definitions for two binding sites located at -35 and -10 base pairs (bp) with respect to the Transcription Start Site (TSS). The difficulty with this direct approach is that the known binding sites are highly variable, leading to a high rate of false positive predictions for a satisfactory rate of recall.

The indirect approach to promoter prediction depends on the observation that promoters are accompanied by other binding sites for transcription factors which modulate transcription. Given access to a sufficiently large number of definitions of known transcription factor binding sites (TFBSs), clusters of high scoring hits indicate the presence of a promoter. For example, the well known MatInspector (Cartharius, Frech, Grote, Klocke, Haltmeier, Klingenhoff, Frisch, Bayerlein & Werner 2005) and Cluster-Buster (Frith, Li & Weng 2003) programs both use this strategy which is particularly useful with eukaryotic organisms.

As more becomes known about the structure and function of bacterial RNAP, it is clear that the enzyme interacts with the DNA double helix in more complex ways than just the canonical -10 and -35 interactions (Mitchell, Zheng, Busby & Minchin 2003, Miroslavova & Busby 2006). The purpose of this paper is to revisit the direct approach to identifying bacterial promoters but to build models that incorporate more of what we have recently learned about the DNA-RNAP interaction. To this end, we have developed a software tool, *Beagle*, that utilizes a simple description language to specify bacterial promoter models. Internally, the models are realized as a sequence of PWMs and gap length distributions. The model parameters are refined using experimentally confirmed TSSs. *Beagle* achieves good accuracy compared to more complex machine learning methods but is faster to train and easier to use. In addition, the generated models are transparent and permit direct biological interpretation.

This paper is organized as follows: In Section 2 we discuss related supervised learning algorithms for promoter prediction. The biological background that drives our promoter models is provided in Section 3 and the data utilized to evaluate various models are described in Section 4. Section 5 explains some of the algorithmic detail behind *Beagle*. Prediction results are presented in Section 6 followed by the conclusion in Section 7.

2 Related work

Many methods have been developed for promoter prediction. Vanet *et al.* (Vanet et al. 1999) provides a good overview of the various approaches. We focus our attention on three more recent contributions to the literature that offer interesting comparisons with our work.

Huerta *et al.* (2003) derived PWMs for the -35 and -10 elements of σ^{70} promoters in *E. coli* from multiple alignments of known promoters. The PWMs were optimized using information content and similarity to a known consensus. Typically their derived PWMs ex-

tended two or more bases upstream of the canonical -10 and -35 hexamers and their models also incorporated scores derived from frequency of spacer lengths and distance to the gene start site (GSS). They observed that true promoters tend to occur in regions where there is a cluster of high scoring putative promoters. And in about 50% of cases, the true promoter was not the highest scoring location.

Gordon *et al.* (2006) trained an ensemble of Support Vector Machines (SVMs) for bacterial promoter prediction using a variant of the mismatch string kernel. The SVM approach was more accurate than the PWM approach but highest accuracy was obtained with a model that combined scores from the ensemble-SVM, PWMs and GSS to TSS distance. An obvious drawback with an ensemble of 40 SVMs is the time required to train them – typically several orders of magnitude more than the estimation of parameters for PWM models.

Burden *et al.* (2005) trained a series of Time Delay Neural Networks (TDNNs) to model multiple promoter elements. They demonstrate greatly improved accuracy when distance to GSS is incorporated into the models. However the number and type of model elements was fixed and TDNNs are typically time consuming to train.

The primary motivation for Beagle is the explicit incorporation of additional DNA motifs into promoter models based on our emerging understanding of the action of RNAP. Beagle gives the experimenter control over all elements of the promoter model, enabling a variety of hypotheses to be tested. While the PWM models of Huerta *et al.* (2003) included extended -10 and extended -35 elements, they were not user defined and it was not demonstrated how these contributed to prediction accuracy. In the case of the ensemble-SVM approach, Gordon *et al.* (2006) identified DNA locations important for classification accuracy. Not surprisingly the -10 and -35 locations were most important but also the ribosomal binding site motif figured strongly around the +20 location, indicative of the fact that most promoters lie close to their GSS. Locations upstream of the -35 box and an extended -10 were not identified as important for classification but the method had limited resolution.

3   Biological Background

Bacterial RNAP is a protein complex composed of five subunits, $\alpha_2\beta\beta'\omega$ (Murakami & Darst 2003). To initiate transcription, the core enzyme must first acquire an additional σ subunit whose function is to recognize a promoter (Gross, Chan, Dombroski, Gruber, Sharp, Tupy & Young 1998). DNA binding initiates a series of structural changes that result in DNA strand separation at the -10 site. After several cycles of formation and release of short transcripts, the σ-factor dissociates and gene transcription commences (Murakami & Darst 2003).

It has long been known that domains 2 and 4 of the σ factor bind to the strongly conserved -10 and -35 boxes. More recently, it has been demonstrated that a third domain interacts with a so-called *extended* - 10 element (see Fig. 1) (Miroslavova & Busby 2006). First identified in *B. subtilis*, the extended -10 element is also present in about 20% of *E. coli* promoters. It is located three base pairs upstream of the -10 element with consensus `TG` (Mitchell et al. 2003). Mitchell *et al.* (2003) also identified the importance of a longer extended -16 region (consensus `TRTG`[1]), which is important for some *E. coli* promoters. *In vitro* experiments have demonstrated that domain 3

[1] `N` = any nucleotide, `R` = `A` or `G` and `W` = `A` or `T`, according to the IUPAC DNA alphabet.
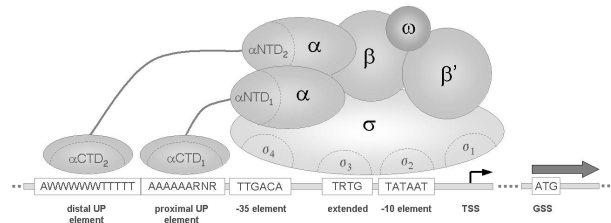

Figure 1: Schematic diagram of the RNA polymerase holoenzyme and its binding elements within the promoter region.

interaction with an extended -10 or -16 consensus site can compensate for weaker -10 or -35 interactions but that a combination of consensus -10, extended -10 and -35 motifs reduces gene expression (Miroslavova & Busby 2006).

The α subunits also play a key role in the initiation of transcription. Each consists of two domains connected by a flexible linker. The amino-terminal domains (αNTD) form part of the main body of the holoenzyme, while the carboxy-terminal domains (αCTD) are free to interact with UP-elements and activators (Estrem, Ross, Gaal, Chen, Niu, Ebright & Gourse 1999).

An UP-element is an `A/T` rich region about 20 bp long located immediately upstream of the -35 element. Each of the two αCTD domains can bind autonomously to the proximal or distal part of an UP-element (Typas & Hengge 2005). It has been shown for some promoters that interactions between one or both α subunits and the UP-elements can increase promoter activity by a factor of 10 or more (Estrem et al. 1999).

The focus of this paper is to determine whether incorporation of these more recently discovered functional sites into promoter models improves the prediction of σ^{70} dependent promoters.

4   Data set

For our experiments we utilized the bacterial genomes of *Escherichia Coli* K-12 MG1655 (ACCN:U00096.2)[2] and *Bacillus subtilis* (ACCN: NC_000964.2)[3].

Experimentally confirmed TSS locations for *E. coli* were obtained from the RegulonDB database[4]. The data set was filtered for unique σ^{70}-promoters with known TSS locations, resulting in 542 records. We then determined the genes in *E. coli* closest to the given TSS locations and extracted the corresponding upstream regions. Following Huerta *et al* (2003), we eliminated all upstream regions (USRs) with a TSS location further than 250 bp from the gene start. The final data set for *E. coli* consisted of 492 sequences, each containing a single annotated TSS location.

A list of TSS locations for *B. subtilis* was was obtained from DBTBS (Release 4)[5], a database of transcriptional regulation in Bacillus subtilis. This list contains 275 TSS predictions from which we selected 205 that were within 250 bp upstream of the nearest gene start site.

5 Beagle

Beagle builds promoter models in two steps. The first step involves initialization of the model using a simple promoter description language and the second step refines the model iteratively. The final model consists of a series of optimized PWMs and gap length distributions.

The initialization phase takes as input a promoter description string which defines a set of consensus motifs and the gaps between them. For instance, the canonical model of a σ^{70}-promoter has a -35 `TTGACA` element, a 15-21 bp spacer, a -10 `TATAAT` element and a 4-13 bp discriminator culminating in the TSS. This canonical promoter can be specified in Beagle by the description string:

`<TTGACA (15,21) TATAAT (4,13) TSS>`

A promoter description can contain an arbitrary number of binding motifs and gap definitions. In particular, models can include the gap between TSS and GSS and incorporate UP elements and extended -10 motifs.

Beagle parses the description string and translates it into a model composed of PWMs and weighted gaps. In the initialization step, the PWM elements are set to represent the required consensus sequences and the gap length frequencies are initialized to a uniform distribution.

The model parameters are optimized during a training phase using an iterative bootstrap approach. At each iteration, the model's TSS position is anchored to the known TSS position of a training sequence and, by exhaustively scoring all valid arrangements of PWM matches taking the current gap distribution into account, the highest scoring combined match is found. Gap weights also contribute to the score[6]. To generate an improved model, maximum likelihood estimates for new PWM and gap weights are calculated from the best match in each of the training sequences. This bootstrapping process continues iteratively until the information content of the PWMs ceases to increase.

For prediction, the model TSS is anchored at each position of the query sequence and the score of the best match is given to that position. The position with the highest overall match score becomes the predicted, putative TSS for that sequence. For more details see the manual which accompanies the software download.

The initial promoter description string may also incorporate a marker for the gene start site (GSS). This permits the definition of models that take the distance to the downstream GSS into account. The GSS marker is always anchored to the nearest gene start site and the weights for the distribution of TSS-GSS gaps are evaluated in exactly the same way as for other gap/spacers in the model. Gaps have a so called *impact factor*, which weights the relative contribution of the gap score to the overall model score. In the following model of a canonical promoter with extended -10 and TSS-GSS gap, gap scores contribute 20% to the overall score:

`TTGACA (12,18,0.2) TGNTATAAT (4,13,0.2) TSS (0,249,0.2) GSS`

The overall match score s_{all} of a sequence to a model consisting of N elements (PWMs or gaps) with element scores s_i and impact factors f_i, is calculated as follows:

$$s_{all} = \frac{\sum_i^N f_i \cdot s_i}{\sum_i^N f_i}, \quad \text{with } s_i, f_i \in \{0, 1\}. \qquad (1)$$

Beagle has some similarity to Meta-MEME (Grundy, Bailey, Elkan & Baker 1997) in that the required patterns are modeled as a set of conserved motifs separated by gaps. But where Meta-MEME uses MEME to obtain an initial PWM description of the conserved motifs, Beagle derives its PWM description from a user supplied consensus. And whereas Meta-MEME then embeds the PWMs into a Hidden Markov Model along with a probabilistic description of the gaps, Beagle preserves the PWMs and gaps as discrete entities.

In the next section, we demonstrate the performance of various promoter models for TSS prediction.

6 Results

We used Beagle to explore extensions to the canonical promoter model by incorporating various combinations of (1) the extended -10 element (consensus `TG`), (2) the -16 element (consensus `TRTG`), (3) UP-elements and (4) distance between TSS and GSS (see Fig. 1). We experimented with three different UP-element sequences that appear to be prominent in several *E. coli* and *B. subtilis* promoters: (1) The most general UP-element is an `A/T`-rich region described in our description language as `NNWWWWWWWWWWWWWWWNN`. (2) For the promoter *rrnB*-P1 in *E. coli*, Estrem *et al.* (Estrem, Gaal, Ross & Gourse 1998) reported an UP-element with the consensus sequence `NNAAAWWTWTTNNAAANNN`. (3) According to Gourse *et al.* (2000), UP-elements can be divided into a more important proximal motif (`AAAAAARNR`) and a distal motif (`NNAAAWWTWTTN`). We incorporated the proximal half of the motif only.

Table 1 shows the prediction accuracies for a variety of promoter models when applied to two sets of known promoters in *E. coli* and *B. subtilis*. The result for the canonical promoter (`TTGACA (15,21,0.2) TATAAT (4,13,0.2)`) is shown in the top left of each table. Prediction accuracy is calculated as the percentage of predicted TSS locations that are at most ± 5 bp from the true TSS[7]. Interpretation of results can be helped by reference to Fig. 2 which illustrates the sequence logos obtained from training data for the most successful model in each genome.

It is immediately apparent that prediction accuracies are up to 50% higher for *B. subtilis* promoters than for *E. coli* promoters. The sequence logos in Fig. 2 illustrate that the *B. subtilis* promoters have higher information content and are more highly conserved. It must also be the case that a larger fraction of *B. subtilis* TSSs are located at the highest scoring location upstream of their genes than is case for *E. coli* promoters. *B. subtilis* has 18 identified sigma factors compared with seven known for *E. coli*. It is thought that this is due to the greater regulatory demands placed on *B. subtilis* given its more variable soil environment. We might expect that having more σ-factors requires *B. subtilis* to conserve the differences between them by keeping binding sites closer to the consensus.

Another interesting difference between the two species is that inclusion of the TSS-GSS distance in the initial promoter definition improves prediction accuracy significantly in *E. coli* but not in *B. subtilis*. Again this can be explained if a larger fraction of *B. subtilis* TSSs are located at the highest scoring location upstream of their genes no matter how far upstream.

The effect of including an UP-element in the promoter definition (in the absence of an extended -10

---

[6]Beagle utilizes the BioPatML pattern matching engine for this purpose. See `http://eresearch.fit.qut.edu.au/BioPatML/` for details.

[7]There is no consistent definition of a true positive TSS prediction in the literature. We follow the definition of Huerta *et al* (Huerta & Collado-Vides 2003).
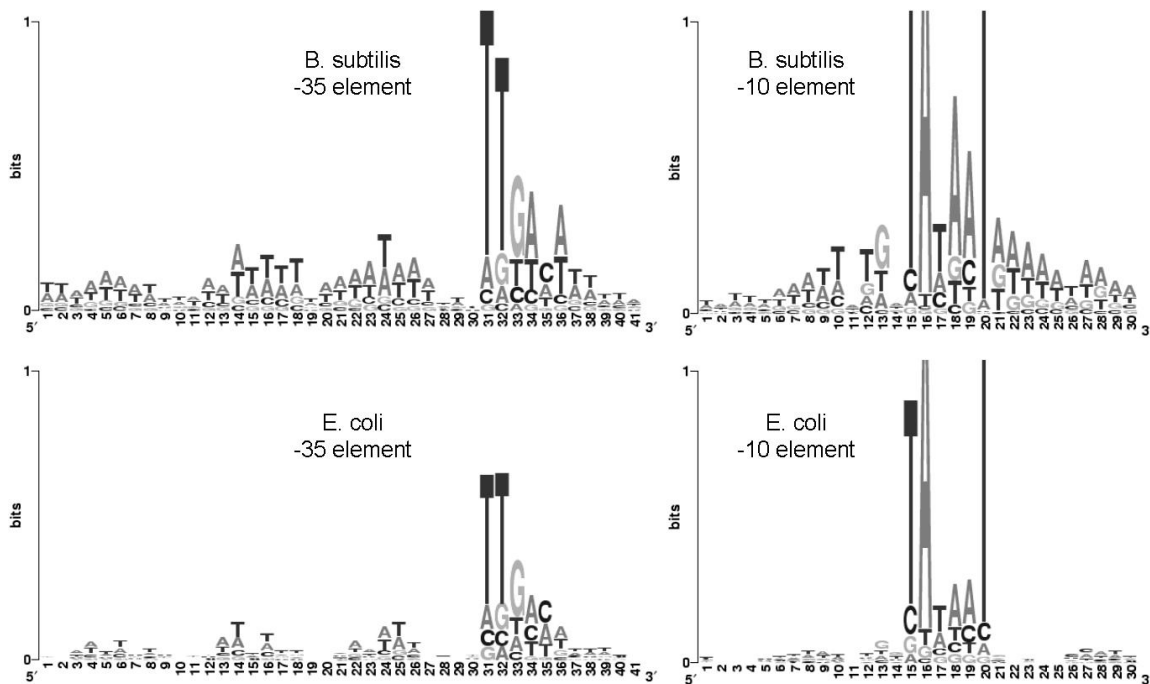
Figure 2: Logos of the vicinity of the -35 and -10 elements of the best performing promoter model in *E. coli* and *B. subtilis*. Note that the y-axis scale has been truncated to 1 bit in order to highlight detail in the upstream region. Logos created with WebLogo at `weblogo.berkeley.edu`.

motif) was variable. The fully defined UP-element NNAAAWWTWTTNNAAANNN had a deleterious effect on prediction performance while the `A/T`-rich UP-element `NNWWWWWWWWWWWWWWWNN` and the proximal UP-element (`AAAAAARNR`) both improved prediction accuracy.

In *E. coli*, use of the extended -10 (`TRTG`) had a deleterious effect on promoter prediction in all cases. Interestingly, use of the `TG` extended -10 also had a deleterious effect on prediction accuracy except when used in conjunction with the `A/T`-rich UP-element. This interaction between the extended -10 and `A/T`-rich UP-elements is one of the novel findings of Beagle that has not, to our knowledge, been reported in the literature previously.

In the case of *B. subtilis*, the `TG` extended -10 motif increases prediction accuracy only when accompanied by an UP-element. And in contrast to *E. coli*, use of the `TRTG` extended -10 increases prediction accuracy more than the `TG` extended -10. These differences between the species become clearer when we compare the sequence logos in Fig 2.

The best performing *E. coli* promoter model achieved 48% recall at 48% precision. In order to compare this result with other publications it is important to ensure that the experimental protocols are similar. In particular the prediction error tolerance and the length of upstream sequence being searched must be the same. We set up our experimental design to be similar to that of Huerta *et al.* (2003). Table 8e of their paper indicates a precision of 33% at a recall of 50%. For different experimental conditions, Burden *et al.* (2005) report 25% precision at 32% recall. When we modify our protocol to match theirs, we achieve 32% precision at 32% recall. The advantage of Beagle lies in the more complex promoter definition and in the iterative refinement of the PWMs. Different experimental conditions do not allow us to compare results with Gordon *et al.* (2006).

7 Conclusion

In this paper we introduced the software, *Beagle*, that enables the convenient description and exploration of PWM based promoter models. Beagle is a technically simple and fast method but nevertheless achieves state-of-the-art accuracy for TSS prediction.

Beagle has several additional attractive features. More complex promoter models can be constructed easily with an arbitrary number of PWMs and spacers. Training and prediction are fast, which allows an interactive study of promoter models and their elements. No negative examples are required for the training process, which can be a serious problem when building discriminative models such as SVMs. The generated models are completely transparent which is helpful for the testing of hypotheses.

We utilized Beagle to investigate a variety of models for σ^{70} promoters prediction in *E. coli* and *B. subtilis*. The results demonstrate an interesting interaction between UP-elements and extended -10 elements that has not been reported previously. The Beagle software, training and test data sets and extended results are publicly available at `http://eresearch.fit.qut.edu.au/Beagle/`.

Further work will examine the properties of wrongly predicted promoters. We also intend to apply Beagle to other transcription factors and genomes.

UP-element	extended -10	E. coli		B. subtilis	
		-	dist. GSS	-	dist. GSS
not used	-	37.5 ±1.4	43.3 ±1.2	61.6 ±1.8	61.2 ±1.7
	TG	36.1 ±1.4	41.6 ±1.3	59.4 ±1.8	62.5 ±1.8
	TRTG	32.5 ±1.3	37.6 ±1.3	59.2 ±1.8	62.6 ±1.8
proximal	-	39.0 ±1.3	44.3 ±1.4	65.2 ±1.9	66.4 ±2.0
	TG	35.4 ±1.3	43.7 ±1.3	66.2 ±2.1	68.5 ±2.1
	TRTG	31.5 ±1.2	38.6 ±1.3	67.3 ±1.9	70.3 ±1.9
full	-	34.8 ±1.3	41.4 ±1.2	58.8 ±1.7	62.0 ±1.7
	TG	31.4 ±1.3	39.0 ±1.4	64.8 ±1.6	66.7 ±1.8
	TRTG	25.9 ±1.0	35.4 ±1.2	65.0 ±1.8	66.6 ±1.9
A/T-rich	-	39.1 ±1.1	47.3 ±1.2	64.5 ±1.7	64.8 ±1.8
	TG	40.8 ±1.2	48.3 ±1.5	66.7 ±1.8	68.8 ±1.6
	TRTG	34.9 ±1.3	40.5 ±1.4	69.6 ±1.7	71.2 ±1.7

Table 1: Accuracies and 95% confidence intervals for TSS prediction on test data for different promoter models. Acceptance tolerance was ±5 bp. Averages are over 10-fold cross-validation, repeated 10 times.

References

Burden, S., Lin, Y.-X. & Zhang, R. (2005), 'Improving promoter prediction for the NNPP2.2 algorithm: a case study using *Escherichia coli* DNA sequences.', *Bioinformatics* **21**(5), 601–607.
*http://dx.doi.org/10.1093/bioinformatics/bti047

Cartharius, K., Frech, K., Grote, K., Klocke, B., Haltmeier, M., Klingenhoff, A., Frisch, M., Bayerlein, M. & Werner, T. (2005), 'Matinspector and beyond: promoter analysis based on transcription factor binding sites.', *Bioinformatics* **21**(13), 2933–2942.
*http://dx.doi.org/10.1093/bioinformatics/bti473

Estrem, S. T., Gaal, T., Ross, W. & Gourse, R. L. (1998), 'Identification of an UP element consensus sequence for bacterial promoters.', *Proc Natl Acad Sci U S A* **95**(17), 9761–9766.

Estrem, S. T., Ross, W., Gaal, T., Chen, Z. W., Niu, W., Ebright, R. H. & Gourse, R. L. (1999), 'Bacterial promoter architecture: subsite structure of UP elements and interactions with the carboxy-terminal domain of the RNA polymerase alpha subunit.', *Genes Dev* **13**(16), 2134–2147.

Frith, M. C., Li, M. C. & Weng, Z. (2003), 'Cluster-buster: Finding dense clusters of motifs in dna sequences.', *Nucleic Acids Res* **31**(13), 3666–3668.

Gordon, J. J., Towsey, M. W., Hogan, J. M., Mathews, S. A. & Timms, P. (2006), 'Improved prediction of bacterial transcription start sites.', *Bioinformatics* **22**(2), 142–148.
*http://dx.doi.org/10.1093/bioinformatics/bti771

Gourse, R. L., Ross, W. & Gaal, T. (2000), 'UPs and downs in bacterial transcription initiation: The role of the alpha subunit of RNA polymerase in promoter recognition', *Mol Micro* **37**(4), 687–695.

Gross, C. A., Chan, C., Dombroski, A., Gruber, T., Sharp, M., Tupy, J. & Young, B. (1998), 'The functional and regulatory roles of sigma factors in transcription.', *Cold Spring Harb Symp Quant Biol* **63**, 141–155.

Grundy, W. N., Bailey, T. L., Elkan, C. P. & Baker, M. E. (1997), 'Meta-meme: motif-based hidden markov models of protein families.', *Comput Appl Biosci* **13**(4), 397–406.

Huerta, A. M. & Collado-Vides, J. (2003), 'Sigma70 promoters in *Escherichia coli*: specific transcription in dense regions of overlapping promoter-like signals.', *J Mol Biol* **333**(2), 261–278.

Miroslavova, N. S. & Busby, S. J. W. (2006), 'Investigation of the modular structure of bacterial promoters', *Biochem. Soc. Symp.* **73**, 1–10.

Mitchell, J. E., Zheng, D., Busby, S. J. W. & Minchin, S. D. (2003), 'Identification and analysis of 'extended -10' promoters in *Escherichia coli*.', *Nuc Acids Res* **31**(16), 4689–4695.

Murakami, K. S. & Darst, S. A. (2003), 'Bacterial RNA polymerases: the wholo story.', *Curr Opin Struct Biol* **13**(1), 31–9.

Schneider, T. D., Stormo, G. D., Gold, L. & Ehrenfeucht, A. (1986), 'Information content of binding sites on nucleotide sequences.', *J Mol Biol* **188**(3), 415–431.

Typas, A. & Hengge, R. (2005), '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.
*http://dx.doi.org/10.1111/j.1365-2958.2004.04382.x

Vanet, A., Marsan, L. & Sagot, M. F. (1999), 'Promoter sequences and algorithmical methods for identifying them.', *Res Microbiol* **150**(9-10), 779–799.