The Journal of Microbiology, October 2005, p.411-416 Vol. 43, No. 5

Computational Detection of Prokaryotic Core Promoters in Genomic Sequences

Ki-Bong Kim[1,*] and Jeong Seop Sim[2]

[1]Department of Bioinformatics Engineering, Sangmyung University, Cheonan 330-180, Republic of Korea
[2]Department of Computer Science and Engineering, Inha University, Incheon 402-751, Republic of Korea

(Received December 22, 2004 / Accepted April 14, 2005)

The high-throughput sequencing of microbial genomes has resulted in the relatively rapid accumulation of an enormous amount of genomic sequence data. In this context, the problem posed by the detection of promoters in genomic DNA sequences via computational methods has attracted considerable research attention in recent years. This paper addresses the development of a predictive model, known as the dependence decomposition weight matrix model (DDWMM), which was designed to detect the core promoter region, including the -10 region and the transcription start sites (TSSs), in prokaryotic genomic DNA sequences. This is an issue of some importance with regard to genome annotation efforts. Our predictive model captures the most significant dependencies between positions (allowing for non-adjacent as well as adjacent dependencies) via the maximal dependence decomposition (MDD) procedure, which iteratively decomposes data sets into subsets, based on the significant dependence between positions in the promoter region to be modeled. Such dependencies may be intimately related to biological and structural concerns, since promoter elements are present in a variety of combinations, which are separated by various distances. In this respect, the DDWMM may prove to be appropriate with regard to the detection of core promoter regions and TSSs in long microbial genomic contigs. In order to demonstrate the effectiveness of our predictive model, we applied 10-fold cross-validation experiments on the 607 experimentally-verified promoter sequences, which evidenced good performance in terms of sensitivity.

Key words: contig, core promoter, DDWMM, MDD procedure, TSS, 10-fold cross-validation.

With the increasing completion of a host of genome sequences, one of the more interesting challenges in molecular biology today is to characterize the mechanisms by which gene expression is regulated (Ko *et al.*, 2002; Jones, 2005). In this context, the problem posed by the identification of promoters in genomic DNA sequences, as well as the determination of the significant patterns they harbor via computational methods, has attracted a considerable amount of research attention in recent years (Pedersen *et al.*, 1999; Sinha *et al.*, 2002). One point of view holds that this problem is intimately related with the fundamental biochemical issues surrounding the specification of the precise sequence determinants associated with transcription and translation (Hernandez *et al.*, 2002). Another point of view holds that the resolution of this problem may contribute to improvements in gene identification, as well as the prediction of gene expression contexts. Moreover, the location and decryption of promoters is an interesting proposition in its own right. Existing algorithms for the prediction of promoters are predicated on the identification of regulatory signals, specifically binding sites for transcription factors. Although the problem posed by the prediction of regulatory sites has been addressed in a number of studies for at least 15 years, it is still far from being solved (Fickett *et al.*, 1997; Ohler *et al.*, 2001). One reason for this is that the learning sample rarely contains more than 20-30 sites. However, even when working with large samples, it has proven extremely difficult to construct a good recognition rule. The physics of protein-DNA interaction remains poorly understood, making it virtually impossible to derive a proper set of features for use in either statistical or pattern recognition algorithms. Furthermore, the latter algorithms cannot take biological context into account. In particular, interaction between different regulatory sites, and structural properties of DNA, remain impossible to integrate into current algorithmic schemes. In many cases, simple profile methods perform reasonably well, in that they can correctly identify true sites if the number of alternatives is not too great (Frech *et al.*, 1997). In this regard, position-specific scoring matrices (PSSMs) carry the

* To whom correspondence should be addressed.
(Tel) 82-41-550-5377; (Fax) 82-41-550-5184
(E-mail) kbkim@smu.ac.kr

important advantage that they are simple, easy to understand, and easy to use (Mount, 2001). In addition, PSSM is probably the most effective model to use in cases in which relatively few (say, a dozen up to a few hundred) signal sequences are available. A PSSM can be constructed from a frequency matrix via the conversion of frequencies to probabilities or scores, in which each element contains the frequency of a given member of the alphabet, observed at a given position within an aligned set of sequences. An important limitation inherent to PSSMs, however, involves the assumption of independence between positions. To compensate for the weak points inherent to PSSMs, we have attempted to develop a more sophisticated model, which is able to address the most significant observed dependencies between positions, and allow for both non-adjacent and adjacent dependencies.

The target organism used in this study was *Escherichia coli*. This is the microorganism about which we have the most knowledge, with regard to the mechanisms underlying gene regulation, metabolism, etc., and appears to be an appropriate choice for use in a long-standing model system for the study of gene regulation. In the prokaryote, *E. coli*, the form of RNA-polymerase responsible for the recognition of promoter sequences possesses the protein subunit composition $\alpha_2\beta\beta'\sigma$. This so-called holo-enzyme can be further divided into two functional components: the core enzyme ($\alpha_2\beta\beta'$, also designated as E), and the sigma factor (σ). The sigma factor performs an important function in the recognition of promoter sequences, and after successful initiation, it is released from the holoenzyme (Gross *et al.*, 1992). Several different sigma factors have been detected, each of which recognizes a specific subset of promoters (Kim *et al.*, 2004). These subsets exhibit different nucleotide sequences. The biological significance of this is that each promoter group exerts control over genes which are needed under physiologically similar conditions, and that, therefore, require simultaneous expression. In fact, this knowledge has already spurred the development of several models or theories dealing with gene regulation (Collado-Vides, 1992), as well as several dedicated databases (Salgado *et al.*, 2004). Also, several computational methods have been developed for the prediction of the occurrence of promoters or regulatory sites in the DNA sequences of *E. coli* (Hertz *et al.*, 1990; Thieffry *et al.*, 1998). Several algorithms and programs for promoter recognition are currently available. They are predicated primarily on machine learning and homology-based string matching approaches. These techniques have not paid enough attention to the structural aspects inherent to transcription initiation processes. We believe that any effective computa-tional method should take into account the structural aspects which reflect biological situations such as those referenced earlier. Therefore, we have explored the pos-sible correlations in the relevant promoter regions. Our predictive model, referred to as the dependence decomposition weight matrix model (DDWMM), was constructed in order to reflect and compensate for the overall structure and biological context of these promoter regions. This model was also designed to predict promoters *ab initio*, i.e., using the DNA sequences of the organisms of interest as the sole information. We have attempted to devise a reliable probabilistic model which reflects the underlying biology and biological context, which is also able to deal with the diverse variations (i.e., variation in positional nucleotides and transcription elements) and arbitrary interactions (i.e., non-adjacent and adjacent dependencies) which are inherent to promoter sequences. The term "non-adjacent dependencies", refers to dependencies between positions with a variety of separations.

Materials and Methods

The primary goal for the construction of the DDWMM was to identify the TSSs and core promoter regions in the genomic sequences. In order to achieve this goal, the DDWMM was constructed, to reflect the overall structure of the core promoter region and the non-adjacent as well as the adjacent dependencies within or between transcription elements in the region. In order to construct such a model, we employed several processes. The construction of the DDWMM began with the collection of the data sets required for this work. The target promoter region of biological significance, judging from positional information content, was defined by the alignment of all collected promoter sequences, and the careful scanning of information content at each position. The target promoter region sequences were then iteratively subclassified into subsets, via maximal dependency decomposition (MDD) (Burge *et al.*, 1997). This process was required in order to incorporate biological context and long-range dependency into our predictive model. This resulted in the construction of a binary tree with many leaf nodes, each of which represented a subset of the target promoter region data. In order to account for the adjacent positions of the dependencies, we applied a first-order Markov model for each subset generated by the MDD procedure.

Data set of promoter sequences

From the 'promoter' table in RegulonDB (Salgado *et al.*, 2004), which is a database containing information regarding transcription regulation and operon organization in *E. coli*, we collected all of the experimentally-verified possible core promoter sequences, a total of 607 of which were ultimately identified. Each promoter sequence was 81 bases long, including 60 bases upstream and 20 bases downstream from the transcription initiation position. According to the results of the classification by each sigma factor, 548 of the 607 promoter sequences belonged

to the σ^{70}-class, and 10, 21, 22, 3, 3, and 83 belonged to the σ^{54}, σ^{38}, σ^{32}, σ^{28}, σ^{24}, and so forth classes, respectively. The ambiguous nucleotide symbol '*N*' within the sequences was randomly converted to either *A*, *T*, *G*, or *C*. A portion of this data set was also employed in the assessment of predictive accuracy. The use of *E. coli* promoter sequences in the construction of a predictive model for prokaryotes is justified, considering the similar transcription machinery employed by *E. coli*, as well as other prokaryotes.

Limiting the promoter region with significant information to be modeled

Using the promoter sequences collected as described above, we initially limited the promoter region in the construction of our predictive model. Most existing algorithms use a promoter region which has been determined arbitrarily, based on the available biological domain knowledge and experimental evidence. However, we used an objective criterion, namely, positional information content, to limit the most reasonable region for modeling. We selected this criterion because the positions with higher information content tend to exhibit a substantial amount of nucleotide conservation, which is probably a fact with a great deal of biological significance. In this context, we have employed the definition provided by Tom Schneider (Schneider *et al.*, 1990) whose work followed the work of Claude Shannon, and who defined the uncertainty measure as the following:

$$H(l) = -\sum_{B=A}^{T} f(B,l)\log_2 f(B,l) \quad \text{(bits per position)} \quad (1)$$

where $H(l)$ is the uncertainty at position l, B is one of the bases (A, C, G, or T), and $f(B,l)$ is the frequency of base B at position l. Total information at any position can be represented by a reduction in uncertainty:

$$R_{sequence}(l) = 2 - [H(l) + e(n)] \quad \text{(bits per position)} \quad (2)$$

where $R_{sequence}$ is the amount of information present at position l, 2 is the maximum uncertainty at any given position (i.e. $\log_2 M$: in this case, the number of symbols or choices, M, is 4), and $e(n)$ is a correction factor required in cases in which the number of sample sequences (n) is small. In this work, we have ignored the correction factor, as we had sufficient sample sequences. The entire set of the $R_{sequence}(l)$ values results in the formation of a curve which represents the importance of each position in the aligned sequences. In this study, all of the promoter sequences were aligned using the transcription initiation point (position 61 in Fig. 1) to determine positional information content, as defined by formula (2). No gaps, i.e. insertions or deletions, were permitted in this alignment. Hereafter we follow the position convention in which the transcription initiation position is zero, then decreasing upstream from initiation position by increments of one, and increasing downstream by increments of one. Based on the information content at each position, we then limited the reasonable region in the original promoter stretch (81 bases), which was relatively rich in information. As a result, our region covers bases -21 to +4 (26 bases), which were used to construct a predictive model in this work.

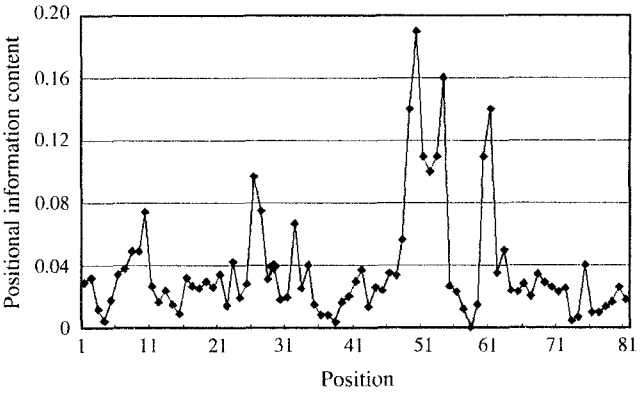

Fig. 1. Positional information content of 607 experimentally confirmed promoter sequences. In order to compute positional information content, the promoter sequences were aligned by TSS (the position 61), permitting no gaps (i.e. insertions/deletions). The height of the curve signifies the information content of the sequences at that position, and the curve itself displays both significant positions and subtle sequence patterns. The transcription start site and the -10 region (around the position 50) show relatively high information content.

Construction of the DDWMM and using it for prediction

In order to construct a model which allows for the possibility of "long-range" interactions between positions which are further than two or three nucleotides apart, we considered the degree of dependence existing between arbitrary positions in the promoter region (-21 to +4). Chi-square statistics were used to determine dependencies between the N_i and N_j variables (which take on the four possible values *A*, *C*, *G*, *T*), indicating the nucleotides at positions i and j of the sequence, i.e., to determine whether an association existed between the occurrence of a particular nucleotide(s) at position i, and the occurrence of other nucleotide(s) at position j within the same sequence. As an example of such a comparison, positions -11 versus -7 in the set of 607 promoter sequences is illustrated, using the standard 4×4 contingency table representation, in Table 1. The observed value of $X^2 = 119.96$, indicates a significant degree of dependence between positions at the $P < 0.001$ level. An examination of the contingency table data revealed that most of the dependence was the result of a positive association between *A* at position -11 and *T* at position -7, with a corresponding increase in the incidence of *T* at -7, in which N_{-11} is *A*. We also determined there to be a somewhat weaker negative association between the occurrence of *A* at position -11 and *A* at -7, and so on. Overall, the most notable feature of this table is that the distribution of nucleotides at posi-

Table 1. Contingency table for the nucleotide variables, N_{-11} versus N_{-7}, in the set of all 607 promoter sequences. For each pair of nucleotides X, Y, the observed count ("O") of the event that $N_{-11} = X$ and $N_{-7} = Y$ is given first, followed by the expected value ("E"). $X^2 = 119.96$ ($p < 0.001$, degree of freedom (df) is 9)

N_{-11}	N_{-7}								
	A		T		G		C		Total
	O	E	O	E	O	E	O	E	O
A	29	63	179	117	25	33	27	46	260
T	66	47	46	87	27	25	54	34	193
G	31	20	23	36	16	10	10	14	80
C	22	18	25	33	10	10	17	13	74
Total	22	18	25	33	10	10	17	13	74

tion -7 seems to depend on whether N_{-11} is or is not A (the consensus at position -11). Then, using the dependency measure as a yardstick, we subclassified all of the experimentally-verified promoter sequences, which constitute a learning data set. The dependence decomposition weight matrix model (DDWMM) is predicated on the maximal dependence decomposition (MDD) procedure proposed by Chris Burge, who initially applied it to human donor splice sites (Burge *et al.*, 1997). The MDD procedure can be applied to generate, using an aligned set of signal sequences, a model which captures the most significant dependencies between positions (allowing for non-adjacent as well as adjacent dependencies), essentially via the substitution of unconditional PSSM probabilities by appropriate conditional probabilities. This, of course, depends on sufficient data being available to reliably do so. Given a data set, D, which consists of n aligned sequences of length k, we first calculate the chi-square statistic $X^2_{i,j}$ for N_i and N_j for each i, j pair with $i \neq j$. If strong dependencies exist between non-adjacent positions, as well as adjacent positions, then we proceed as follows:

(1) Calculate, for each position i, the sum $S_i = \sum_{i \neq j} X^2_{i,j}$, which is a yardstick of the degree of dependency between the variable N_i and the nucleotides at all other positions of the promoter sequences' target regions.

(2) If the data set D exhibits a significant degree of dependency between the positions with variable separation, we choose the value i_1 such that S_{i_1} is maximal, then partition D into two subsets: D_{i_1}, all sequences which agree with the consensus nucleotide(s) with high frequency at position i_1; and $D_{i_1}-$, all sequences which do not.

(3) The first two steps are then repeated on the subsets D_{i_1} and $D_{i_1}-$ and on the subsets thereof, which soon yields a binary "tree" (often called a decision tree), with a theoretical maximum level of k-1 (Fig. 2).

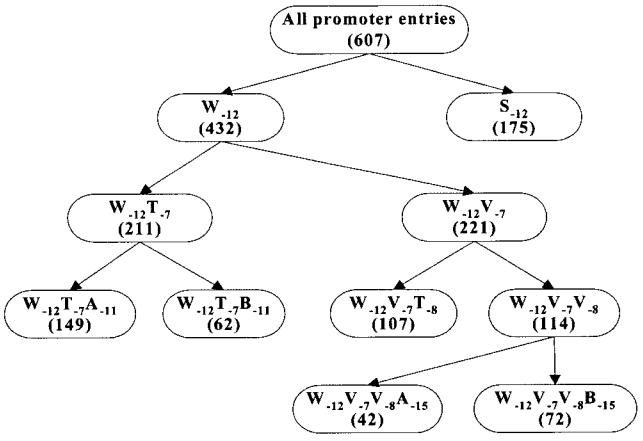

Fig. 2. Recursive subclassification of all promoter sequences (607) using the MDD procedure. Each rounded box represents a subset of promoter sequences, which correspond to a pattern of matches/mismatches to the consensus nucleotide(s) or to the nucleotide with the highest frequency at a set of positions, which exhibit the highest row sum of the X^2 statistics. For example, $W_{-12}T_{-7}$ is the set of promoter sequences with the consensus nucleotides *A/T* and *T* at positions -12 and -7. The number of corresponding promoter sequences in each subset is provided in parentheses beneath the pattern description. IUB single letter symbols are used to represent groups of nucleotides (for example, *W* means *A* or *T*).

This process of subdivision is successively conducted on each leaf node of the tree, until one of the following three conditions is satisfied: (i) the theoretical maximum level of the tree, i.e. (k-1)th level, is achieved (so that no further subdivision is possible); (ii) no significant dependencies between positions in a subset are detected between positions in a subset (so that further subdivision is not indicated); or (iii) the number of sequences in a resulting subset falls below a preset minimum value, M, so that reliable conditional probabilities could not be determined after further subdivision. The results of the application of the MDD procedure to all 607 promoter sequences were illustrated in Fig. 2. The initial subdivision was done, predicated on the consensus W (meaning *A* or *T*) at position -12, resulting in the W_{-12} and S_{-12} subsets (*S* indicating *C* or *G*), which contain 432 and 175 sequences, respectively. No significant dependencies between the positions in subset S_{-12} were detected, and so this was not divided further. However, subset W_{-12} was sufficiently large, and exhibited significant dependence between positions (data not shown). Therefore, it was further subdivided according to consensus *T* at position -7, thereby generating the subsets $W_{-12}T_{-7}$ and $W_{-12}V_{-7}$, and so on (*V* indicating *A*, *C*, or *G*). Thus, the essential concept underlying this method is the iterative analysis of data, initially accounting for the most significant dependence present, and then for the dependencies which remain after the previously chosen dependencies have been accounted for by the subdivision of the data. In order to reflect the dependencies of adjacent positions, we applied a first-

order Markov model for each subset, after the MDD procedure. With a first-order Markov model, the probability of generating the sequence $X = x_1, x_2, \ldots, x_\lambda$ is as follows:

$$P_{WAM}(X) = p^{(1)}(x_1)p^{(2)}(x_2|x_1)p^{(3)}(x_3|x_2)\ldots p^{(\lambda)}(x_\lambda|x_{\lambda-1})$$
$$= p^{(1)}(x_1)\prod_{i=2}^{\lambda} p^{(i)}(x_i|x_{i-1}) \quad (3)$$

where $p^{(i)}(z|y)$ is the conditional probability of generating nucleotide z at position i given nucleotide y at position $i-1$, which can be estimated from the corresponding conditional frequency in this work.

The final predictive model for the generation of promoter sequences is, then, essentially a recapitulation of the subdivision procedure, as described below:

(1) N_{-12} is generated from the first-order Markov model for all of the combined promoter sequences.

(2a) If $N_{-12} \neq W$, then the first-order Markov model for subset S_{-12} is used to generate nucleotides at the remaining positions in the promoter sequence.

(2b) If $N_{-12} = W$, then N_{-7} is generated from the first-order Markov model for the subset W_{-12}

(3a) If ($N_{-12} = W$ and) $N_{-7} \neq T$, then the first-order Markov model for subset $W_{-12}V_{-7}$ is used.

(3b) If ($N_{-12} = W$ and) $N_{-7} = T$, N_{-11} is generated from the first-order Markov model for $W_{-12}T_{-7}$.

(4) ... and so on, until the entire 26 bp sequence has been generated.

Thus, this model actually represents the most significant dependencies between positions, allowing for both non-adjacent and adjacent dependencies, which may adequately reflect biological situations.

Results and Discussion

The main goal of this work was to develop a predictive model which could be employed to computationally recognize core promoters in long, contiguous prokaryotic DNA sequences, such as long contigs. This would obviously be of significant importance in the context of DNA sequence annotation, which requires the collection of as much information as possible. Furthermore, such a computational model could facilitate our determination of which information in a sequence is vital for reliable recognition, and could also allow us to accumulate a substantial amount of knowledge on the biology of gene expression. The approach employed in this work is rightfully referred to as *general* promoter prediction methodology, the primary goal of which is the identification of the transcription start site (TSS) and the core promoter region for all protein-coding genes. We attempted to devise a reliable predictive model, which reflects the underlying biology and biological context, and which is capable of compensating for the diverse variations (i.e., variations in positional nucleotides and transcription elements) and arbitrary interactions (i.e., non-adjacent and adjacent dependencies) inherent to the promoter sequences.

In order to test the performance of the DDWMM, we applied 10-fold cross-validation experiments on the 607 experimentally-verified promoter sequences of fixed lengths of (-21 ~ +4), as mentioned above. In the 10-fold cross-validations, these 607 promoter sequences were divided randomly into 10 subsets of approximately similar size. For each "fold", the DDWMM was trained using all but one of the 10 subsets, then tested on an unseen subset. This procedure was repeated for each of the 10 subsets. The average cross-validation score was assessed according to the average performance across each of the ten training runs. In order to compare the performance of the DDWMM with that of the PSSM, we also applied the same 10-fold cross-validation experiments on the same sequence data, under the same conditions, using the PSSM. Table 2 shows the success percentages of the ten DDWMM and 10 PSSM experiments. The average sensitivity of DDWMM was determined to be 82.5%, whereas the average sensitivity for PSSM was 54%, at a threshold value of 80. Our results indicated that DDWMM performed quite well, and also that DDWMM yielded results superior to those generated by the simpler PSSM. Although the approach followed in this work resulted in good performance, there may be a lot of room for improvement. As an example, the significant disadvantage of N_i and N_j comparisons is that, for positions i and j with strongly-biased compositions, the expected values of the contingency table may become so small (such as < 10) that the X^2 test becomes unreliable. In order to resolve this problem, consensus versus N_j comparison may be an appropriate choice, as lower frequency nucleotides are pooled in a consensus versus N_j comparison. This would then make the problem a less acute one. In addition, our method did not take into account enough of the relevant biological data. This does not, however, indicate that the improvements of our methods necessarily have to include explicit modeling of the biological realities inherent to the situation. Rather, it means that it is important to take biological knowledge into account when

Table 2. Performance comparison of DDWMM with PSSM in 10-fold cross-validation. Each success % was computed from 10 partitions of 607 sequences in training (9/10) and test (1/10) sets, with a threshold value of 80. The last column contains the combined results for all test data.

Testing No.	1	2	3	4	5	6	7	8	9	10	Avg.
DDWMM Success (%)	85	75	85	79	84	82	84	82	89	78	82.3
PSSM Success (%)	59	66	52	46	46	56	49	49	67	50	54.0

determining what to predict and what data should be included in the design of the method.

In the future, this method should be applied to other biological signals, e.g. other transcriptional or translational signals in DNA/RNA or, perhaps, even protein motifs. If larger sets of sequences can be sufficiently accumulated, more complex dependencies can be more reliably measured and modeled. One important future challenge involves the development of more flexible and sensitive approaches to the analysis of available sequence data. This may allow for the detection of even more subtle biological features. In the longer term, it may even be possible to construct realistic models of such complex biological processes as transcription and pre-mRNA splicing, *in silico*.

References

Burge, C. and S. Karlin. 1997. Prediction of complete gene structure in human genomic DNA. *J. Mol. Biol.* 268, 78-94.

Collado-Vides, J. 1992. Grammatical model of the regulation of gene expression. *Proc. Natl. Acad. Sci. USA*. 89, 9405-9409.

Fickett, J. and A. Hatzigeorgiou. 1997. Eukaryotic promoter recognition. *Genome Research.* 7, 861-878.

Frech, K., K. Quandt, and T. Werner. 1997. Software for the analysis of DNA sequence elements of transcription. *Comput. Appl. Biosci.* 13, 89-97.

Gross, C.A. and M. Lonetto. 1992. Bacterial sigma factors. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

Hernandez, E., A. Johnson, V. Notario, A. Chen, and J. Richert. 2002. AUA as a translation initiation site *in vitro* for the human transcription factor Sp3. *J. Biochem. Mol. Biol.* 35, 273-282.

Hertz, G.Z., G.W. Hartzell III, and G.D. Stormo. 1990. Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Comput. Applic. Biosci.* 6, 81-92.

Jones, B.D. 2005. Salmonella invasion gene regulation: a story of environmental awarenss. *J. Microbiol.* 43, 110-117.

Kim, E.Y., M.S. Shin, J.H. Rhee, and H.E. Choy. 2004. Factor influencing preferential utilization of RNA polymerase containing sigma-38 in stationary-phase gene expression in *Escherichia coli. J. Microbiol.* 42, 103-110.

Ko, J., D.S. Na, Y.H. Lee, S.Y. Shin, J.H. Kim, B.G. Hwang, B.I. Min, and D.S. Park. 2002. cDNA microarray analysis of the differential gene expression in the neuropathic pain and electroacupunction treatment models. *J. Biochem. Mol. Biol.* 35, 420-427.

Mount, D.W. 2001. Bioinformatics : sequence and genome analysis. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

Ohler, U. and H. Niemann. 2001. Identification and analysis of eukaryotic promoters: recent computational approaches. *Trends in Genetics.* 17, 56-60.

Pedersen, A., P. Baldi, Y. Chauvin, and S. Brunak. 1999. The biology of eukaryotic promoter prediction - a review. *Comput. Chemistry.* 23, 191-207.

Salgado, H., S. Gama-Castro, A. Martinez-Antonio, E. Diaz-Peredo, F. Sanchez-Solano, M. Peralta-Gil, D. Garcia-Alonso, V. Jimenez-Jacinto, A. Santos-Zavaleta, C. Bonavides-Martinez, and J.H. Collado-Vides. 2004. RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in *Escherichia coli K*-12. *Nucleic. Acids Res.* 29, 72-74.

Schneider, T.D. and R.M. Stephens. 1990. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* 18, 6097-6100.

Sinha, S. and M. Tompa. 2002. Discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res.* 30, 5549-5560.

Thieffry, D., H. Salgado, A.M. Huerta, and J. Collado-Vides. 1998. Prediction of transcriptional regulatory sites in the complete genome sequence of *Escherichia coli K*-12. *Bioinformatics.* 14, 391-400.