Sigma70 Promoters in *Escherichia coli*:
Specific Transcription in Dense Regions of Overlapping Promoter-like Signals

Araceli M. Huerta and Julio Collado-Vides*

Program of Computational Genomics, Nitrogen Fixation Center, UNAM, Cuernavaca A.P. 565-A, Morelos 62100 Mexico

We present here a computational analysis showing that sigma70 housekeeping promoters are located within zones with high densities of promoter-like signals in *Escherichia coli*, and we introduce strategies that allow for the correct computer prediction of sigma70 promoters. Based on 599 experimentally verified promoters of *E. coli* K-12, we generated and evaluated more than 200 weight matrices optimizing different criteria to obtain the best recognition matrices. The alignments generating the best statistical models did not fully correspond with the canonical sigma70 model. However, matrices that correspond to such a canonical model performed better as tools for prediction. We tested the predictive capacity of these matrices on 250 bp long regions upstream of gene starts, where 90% of the known promoters occur. The computational matrix models generated an average of 38 promoter-like signals within each 250 bp region. In more than 50% of the cases, the true promoter does not have the best score within the region. We observed, in fact, that real promoters occur mostly within regions with high densities of overlapping putative promoters. We evaluated several strategies to identify promoters. The best one uses an intrinsic score of the −10 and −35 hexamers that form the promoter as well as an extrinsic score that uses the distribution of promoters from the start of the gene. We were able to identify 86% true promoters correctly, generating an average of 4.7 putative promoters per region as output, of which 3.7, on average, exist in clusters, as a series of overlapping potentially competing RNA polymerase-binding sites. As far as we know, this is the highest predictive capability reported so far. This high signal density is found mainly within regions upstream of genes, contrasting with coding regions and regions located between convergently transcribed genes. These results are consistent with experimental evidence that show the existence of multiple overlapping promoter sites that become functional under particular conditions. This density is probably the consequence of a rich number of vestiges of promoters in evolution. We suggest that transcriptional regulators as well as other functional promoters play an important role in keeping these latent signals suppressed.

© 2003 Elsevier Ltd. All rights reserved.

*Corresponding author

Keywords: sigma70 promoters; *Escherichia coli*; computational predictions; overlapping signals

Introduction

The regulation of transcription initiation is a central piece of the cellular machinery that controls the behavior and dynamic capabilities of the cell. Whole genome expression profiles of completely sequenced genomes give us the possibility to analyze, with unprecedented completeness, the different
states of expression available to the cell. The promoters define the sites for gene transcription initiation. Their identification is crucial in order to anchor the regulatory targets controlling the expression of each gene and of each transcription unit of a genome.

The flexibility of the DNA motif bound by the RNA polymerase (RNAP) holoenzyme has been difficult to capture in an efficient computational predictive algorithm. The understanding and modeling of *Escherichia coli* sigma70 promoters was initiated years ago when Prihnow identified the so-called Prihnow box, or −10 box, based on an alignment of only six known promoter sequences.1 Subsequently, as more promoters were identified, computational methods were able to define what has been called the canonical model of the sigma70 DNA promoter: a −35 hexamer, separated by 15 bp to 21 bp from the −10 hexamer, with consensus sequences TTGACA and TATAAT, respectively.7−10 However, there are other elements that can modify this sigma70 model. Several experiments have described a (TG) sub-region 1 bp immediately upstream of the −10 box, which may render the −35 box dispensable.7−10 This motif, called the extended −10, seems to appear in 20% of known promoters.11,12 Upstream activator sites may substitute the role of a −35 region. A classic example is that of the *pho* operon, where the PhoB binding site is centered at −40, occupying what would have been the −35 box.13,14 Other studies, such as string search analysis,4 revealed a third element upstream from the −35 box, the UP element, originally associated with “bent DNA”.15 and more recently associated with the binding region of the carboxy-terminal domain of the alpha subunit of RNAP.16 In about 3% of the *E. coli* promoters, the UP element is located approximately 4 bp upstream of the −35 region conferring additional strength to the promoter.17

Different particular combinations of the elements mentioned above, some of them mutually substitutable, allow a region to be recognized by the RNAP and transcription to initiate, defining a functional promoter. The dispensability of a −35 element is reflected in the high level of heterogeneity of sequences in that region when comparing several promoters. The −10 box is essential for every promoter. The double-stranded −10 box stabilizes the initial binding of the polymerase and subsequently, as single-stranded DNA, it promotes enzyme isomerization to the functional form. Two highly conserved bases (the first TA) of this box are the most important for interaction with RNAP, while all other positions provide an accessory contribution.18,19 On average, only 7.9 of the 12 canonical bases of the −10 and −35 boxes are conserved among promoters, whereas 10% of the efficient promoters contain either a −10 or a −35 box that resembles the consensus, with five out of six bases of the other box conserved.5,20 This correlates with the enormous diversity among functional *E. coli* promoters, with promoter strengths varying by factors of 100-fold or more.21 It has been shown that mutations in the −10 or −35 that bring the promoter sequence closer to the consensus tend to increase the strength of the promoter, whereas mutations that bring the promoter further from consensus significantly decrease the strength of the promoter.3 However, the binding of the RNAP to a consensus promoter is so strong that it prevents the subsequent steps towards promoter clearance and elongation initiation.3,22,23

The current version of RegulonDB, a relational database of transcriptional regulation and operon organization in *E. coli* K12 contains, among other data, 599 sigma70 promoters for transcription initiation mapped in the genome.24 Using this dataset, we decided to re-evaluate the variability of sigma70 promoters from the canonical model, to implement a predictive computational method for sigma70 *E. coli* promotors, and to lay some features of nucleotide sequences that distinguish promoter regions from regions with no promoter at the genomic level. We conceived this work with two main motivations: to enhance our understanding of what constitutes a promoter sequence, and to implement a predictive computational method. Our computational method and its associated model proved efficient in identifying correctly, from most tested promoters, the +1 position, which had been characterized by experimental analysis. We present our results in the logical order necessary to build the predictive program, together with a description of the biological insight gained at each step.

**Results and Discussion**

**Outline of the problem**

A simple search for the −10 box allowing only two mismatches from the consensus −10 sequence produces putative promoters approximately once every ~30 bp in the complete genomic sequence of *E. coli* K12.25 Searching for entire promoters, with more flexible mismatch restrictions, would produce a huge amount of promoter-like signals, as can be calculated easily from the work published by Robison et al.26 Besides this high probability of false positive signals, only 20% of the known promoters have a score higher than the corresponding to a false positive signal found within a 100,000 bp random sequence.27

What is the adequate length of sequences to search for promoters in fully sequenced microbial genomes? Figure 1 shows that 89.77% of all 791 known mRNA promoters of *E. coli*, or 90.31% of those recognized by sigma70, fall within the first 250 bp upstream from the genes. This fits with the inter-operon average space of 190 bp in *E. coli*,24 together with the fact that 18% of the known promoters are found within the coding region of the immediately preceding gene (see the inset in Figure 1). Based on the distribution of promoter
the corrections upstream The distribution bp with being a promoters co upstream immediately using bp (w deviat run to to selected inte set hexame 22, was bp train a in six program). hexam first alignme criteria of T of a alignme of and we values in in conditio promoters. in promoters. in location, the problem we address here is finding and evaluating a method for promoter recognition within the 250 bp upstream of the first codon of a gene.

Re-evaluating the computational model of a promoter

The much larger collection of promoter sequences currently available justifies a broader analysis without taking a priori any pre-defined consensus sequence or weight matrix as a starting point. We used WCONSENSUS, a program that extracts and aligns motifs within a given input of unordered sequences producing a representative frequency matrix from the alignment,25 to generate multiple alignments and weight matrices using a subset of 116 sigma70 known promoters published by Gralla & Collado-Vides (see Methods).26 The strategy applied to select the best matrices obtained from this training set using a given set of input combinations involves several steps as shown in Figure 2.

In the first step, we selected the best matrices on the basis of four different criteria. Two of the selection criteria were based on values produced by the program: maximal information content and lowest expected frequency. We selected the alignments that generated the narrowest distribution of promoter scores in the training set; that is to say, the distribution with the largest amount of promoters scoring inside one standard deviation of the mean score ($\mu \pm \sigma$), and finally, we chose the alignments that produced matrices where the canonical −10 and −35 hexamers could be identified clearly. The next step of selection involved the evaluation of accuracy (see definitions in Methods) on a testing set of 392 sigma70 promoters. In this way, the complete cycle from generation of the representation of the promoter to the evaluation of its performance was completed.

A library of 288 different weight matrices was thus created. The strategy, illustrated in Figure 2, is as follows: (i) 116 promoters conforming the training set were aligned with respect to the position of the transcription initiation (+1). The first 18 bp upstream from the +1 were used as input for WCONSENSUS to identify the motif corresponding to the −10 conserved region. This size was selected considering that the distance from the −10 hexamer to the +1 varies from 4 bp to 12 bp. (ii) WCONSENSUS was run in four different conditions with this set of 18 bp sequences, corresponding to corrections of standard deviations of 0.5, 1.0, 1.5, and 2.0 (an internal parameter of the program). (iii) In order to identify the −35 box, we performed a re-alignment of the promoter sequences anchoring the −10 boxes identified by the program. New sequences of various lengths (15, 16, 18, 20, 22, and 24 bp long), initiating at 13 bp upstream of the −10, were selected. (iv) WCONSENSUS was run with this new set of sequences using the same four standard deviations as above. Thus, for each one of the −10 alignments, the six alternative sequence sizes were analyzed with the four different standard deviations, making thus 24 possible −35 matrices. The number of possible combinations between the four
matrices previously generated for the −10 box, and the 24 generated for the −35 box gives the total of 96 different pairs of matrices for the whole promoter site.

WCONSENSUS gives the user the option to define the background frequencies of the alphabet elements. We tested three alphabet frequencies: the frequency of bases present in the input sequences; the frequency of strictly non-coding sequences in the genome; and the frequency of bases within coding regions. Thus, we generated a final set of 96 × 3, or 288 alternative promoter models. From all these matrices we selected the best ones on the basis of the four different criteria mentioned above. Table 1 shows the best four classes of weight matrices corresponding to the four criteria used in their selection. Note that in order to compare these matrices with the canonical patterns (TTGACA and TATAAT), the spacers of 13 bp to 19 bp between the two boxes correspond to the 15 bp to 21 bp reported in the literature, as the GT pair is considered as part of the −10 box.

The −10 TATAAT classic pattern is retained in all of the selected matrices. The first two matrices, as well as the matrices with shorter consensus patterns, also partially retained the TG pattern immediately upstream of the −10 box. The consensus shown in Table 1 are composed of the most frequent nucleotides at each position of the alignment. There is no promoter matching the exact consensus sequence; rather there is a great variability among individual promoter sequences. The weight matrices capture this variability by assigning a weight to each possible nucleotide at each position of the promoter sequences aligned. The score assigned to a putative promoter is derived from the added products of the frequencies of nucleotides at each position. All the corresponding matrices may be found on the web. Notice that the matrices with lower expected frequencies are the largest, as they have more positions contributing to the score. As expected, considering the relevance of the −10 and −35 boxes for the conformation of a functional promoter, the information content for the −10 motif was higher than for the −35 motif, and the expected frequency was much smaller for the −10 motif than for the −35 (see the web site).

This consistency of the −10 pattern occurring irrespective of the scoring or selection function contrasts with the much less recoverable TTGACA classic pattern traditionally used to define the −10 box. If we anchor the alignment of this region with the upstream AAA tract found in nearly all consensi, the pattern found can be represented by

† http://www.cifn.unam.mx/Computational_Genomics/PromoterTools/Data
Table 1. Best weight matrices

<table>
<thead>
<tr>
<th>Information content</th>
<th>Spacers (~35 and ~10)</th>
<th>~35 signal</th>
<th>~10 signal</th>
<th>Expected freq.</th>
<th>Inf. Content</th>
<th>%Prom (μ - σ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matrix_18_24_13_0.5_2</td>
<td>AAAA/TCACCTTC/TG</td>
<td>13.25</td>
<td>TGGTATAATG</td>
<td>-141.29</td>
<td>7.11</td>
<td>70</td>
</tr>
<tr>
<td>Expected frequency</td>
<td>Matrix_18_24_13_1_1</td>
<td>13.15</td>
<td>TTTGTATAATG</td>
<td>-206.84</td>
<td>6.33</td>
<td>67</td>
</tr>
<tr>
<td>%Prom in μ - σ</td>
<td>Matrix_18_22_13_1.5_0.5</td>
<td>13.19</td>
<td>GTATATATGA/ TA</td>
<td>-7.31</td>
<td>5.32</td>
<td>74</td>
</tr>
<tr>
<td>Homology to consensus</td>
<td>Matrix_18_15_13_2</td>
<td>13.20</td>
<td>GTATATATGA/ TG</td>
<td>19.65</td>
<td>5.17</td>
<td>74</td>
</tr>
</tbody>
</table>

The 11 best matrices out of 288 generated using the procedure described in the legend to Figure 2. The first column describes the criteria of selection, and the names of the best matrices. The first two numbers in the name of a matrix are the size (in bp) of the sequences used to search the ~10 and ~35 hexamers, respectively. The third number gives the smallest spacer size. The last numbers are the standard deviations used in WCONSENSUS to generate the ~10 and ~35 matrices, respectively. Criteria include the highest information content, the lowest expected frequency, the most compact distribution within μ(±σ), and similarity to classic consensus patterns. The length of the ~10 box varies from 8 bp to 12 bp and that of the ~35 box from 8 bp to 22 bp. It can be seen that the information content and the expected frequency of the different matrices have considerable variation. Large matrices have smaller expected frequency values, since they are less likely to be found by chance. Any pair X/Y means X only in that position.

this regular expression:

\[\text{AAAT}[\text{T/A}]\text{T/A/C}[\text{A}]\text{[A/C/T/G]}\text{[T/A][T/A/C]}\text{[T/T/C][G][A]}\]

This pattern shows that the different alignments do not equally retain the classic ~35 motif. Unlike the ~35 motif, the AAA track appeared in all three classes of alignments. This is remarkable, given that the AAA track is part of an UP proximal element that was previously reported to occur in only 3% of the promoters. Given the compromise of the ~35 region in accommodating additional regulatory elements for activator sites, some of its bases are less conserved (<---ACA) than any A in the consensus A-tract. The AAA motif is present in 17%, TTG in 38%, and ACA in 8% of the complete collection of promoters analyzed here. These frequencies explain why the WCONSENSUS program might produce consensus different from the canonical model of sigma70 DNA promoters. The difference in frequency between the UP-proximal element and the AAA tract lies in the strict search we performed to identify the UP elements published by Estrem et al., requiring five A bases in an UP-proximal element (allowing two mismatches) at either ~5 or ~1 position from the ~35 box. Nonetheless, the strong rmbp1 has only three A bases (AAATTTCCT).

More than 50% of promoter regions have a promoter-like sequence with a higher score than that of the true promoter

The selected matrices were then used as predictors to identify known promoters. We searched for promoter-like signals within the 250 bp upstream regions containing 111 promoters of the training set (the other five promoters in this set are found farther upstream than 250 bp), and selected, for each region, the signal with the highest score. The score corresponding to a particular ~10 or ~35 motif was obtained by the PATSER program, and the global score was defined as the sum of the ~35 and ~10 scores plus the score of the spacer (see Methods). Using this scoring approach, the average predictive capacity of the all pairs of matrices was of 49.6(±3.6)% correctly identified promoters (notice that the pair of matrices denoted Matrix_18_15_13_2_1.5, see the legend to Table 1 for notation, was able to retain only 45.9% of the analyzed promoters, which is slightly lower than the average for all the matrices on Table 1, we will refer to this matrix below). These results imply that a search, in which the highest scored putative promoter is proposed as the real one, would generate 50.4% of false positives. When we tested all other 392 known promoters in the testing set, we obtained 60% false positives, as illustrated in Figure 3(b).

It is surprising that the scores of the ~35 and ~10 hexamers for known functional promoters
are not always the best within the regions upstream of the genes transcribed. Whether these higher scored sites are bound by RNAP is hard to know, unless we accept that similarity to the consensus means only strong binding and has no implication in terms of promoter activity. This is not an unreasonable interpretation, given that the methodology of weight matrices provides a rationale of precisely binding interpretation. \(^{23,24}\) Efficient promoter activity involves binding of the RNAP holoenzyme, as well as several other kinetic steps to open complex, minimal abortive initiation, and subsequent steps to elongation. \(^{32}\) None of these kinetic components are necessarily reflected in the precise promoter sequence we model with weight matrices here. On the other hand, it has been reported that closed and open complexes are formed in sites that are unproductive for transcription. \(^{33}\)

The resulting scores for the 116 promoters in the training set were rather variable, covering a sevenfold range (1.4–10.07), being contained within \(\mu \pm 3\sigma\) (see column two in Table 2). The observation described in Figure 3(b) means that the strategy of searching, based on selecting the highest scoring signal, fails in more than 50\% of the cases. Hence, we define a functional promoter as an opaque promoter when there is another promoter-like signal with a higher score within the same 250 bp upstream region, otherwise we define it as a visible promoter. A visible promoter is not necessarily one with a sequence highly homologous to the consensus.

**Abundance of putative promoters: is it noise or design of promoter regions?**

To be able to detect the opaque promoters, we did a search with the weight matrices for the −10 and −35 boxes using a cutoff of \(\mu \pm 3\sigma\). We obtained an average of 38 signals per 250 bp upstream region, see Figure 3(a) for a picture of
this density of promoter-like signals found. We
decided to use a cutoff of three standard
deviations below the mean score of the
training set for our searches in order to recover
most of the known promoters. This assumption
of a normal distribution of scores appears to be
valid, very few functional promoters have in fact
lower scores for their $-10$ and/or $-35$ boxes than
this cutoff. The $fadBA$ and $lac$ promoter regions,
shown in Figure 4, are examples that show this
density of signals as series of overlapped poten-
tially competing RNAP binding sites.

Table 2 shows how the number of detected true
promoters, the total number of promoter-like sig-
als obtained, and the average number of signals
per region, vary as a function of the threshold
selected. This information redefines the task of
identifying a promoter in $E. coli$ as a matter of
implementing a method able to select one out of 38
promoter-like signals within a given 250 bp region.

At first glance, this excess of potential signals
may be rationalized as noise, due to limitations
of the methods and to the degeneracy of the promoter
signal itself. We decided then to evaluate the fre-
cquency of putative promoters found every 10 bp
within the regions known to contain sigma70
promoters, within non-coding regions separating
divergent genes, and within coding regions. As a
control, we determined the same frequency within
non-coding regions separating convergently tran-
scribed genes, where no functional promoter
should be present, as shown in Figure 5. It can be
seen that, as expected, the frequency of putative
promoters is similar in regions with known prom-
oters and in non-coding regions separating diver-
gent genes. This high frequency of promoter-like
signals is much lower in the two contrasting sets
of coding and convergent non-coding regions,
clearly suggesting that this abundance of signals is
not simply noise due to methodological limi-
tations. The peak observed around the end of the
genes in convergent regions may be explained by
interactions of RNAP with termination sites.

Given that these additional sites are restricted to
non-coding regions where functional promoters
are found in the genome, some of these sites are
expected to play a role either functionally, evolu-
tionarily, or both. The apparent excess of putative
promoters might result from a high evolutionary
rate that changes the functional promoters into

![Diagram](https://example.com/diagram)

**Figure 4.** Putative signals and their location in the 250 bp upstream region of the $fadB$ and $lacZ$ genes. The true promoter is indicated in red, and the promoter-like signal with the highest homology score in blue. Notice that in the $fadBA$ region the higher scored signal is not the real one, which is not the same case for the $lac$ region. A zoom-in of the $-50$ to $-100$ region from of start gene of the $lac$ region is shown, a density of signals can be seen overlapping the real $lac$ promoter.
non-productive ones as mutations occur in the limited region of 250 bp upstream of genes, or facilitate new promoters. Alternatively, these overlapping promoter-like signals might play a regulatory role, negative if those sites had a competitive relationship, or positive if they helped in the channeling of the RNAP into the promoter.

It is possible that some of these apparently false positives will be revealed as functional promoters or RNAP binding sites in new experimental conditions. However, the site of transcription initiation is known to be rather precise. If a large number of putative promoters were indeed functional initiation sites, a smearing in gels of primer extension experiments would be observed, which is not the case in general. There is only one region reported to have seven functional promoters that transcribe the same downstream gene, the *arcA* promoter region. Thus, as mentioned, even in those cases where the computational score shows another stronger promoter-like signal, RNAP identifies with great specificity its precise promoter, and initiation region, in the middle of this forest of plausible attractive sites. Of the 470 regulatory regions reported in RegulonDB 25% contain multiple functional promoters, with three promoters coexisting on average. These promoters are of the same kind of sigma in most (73 or 63.5%) of the cases, whereas in 42, or 36% of the cases they involve promoters with different sigma factors. Nonetheless, it is interesting that the average number of promoter-like signals found is not significantly higher (36.84 signals per region, in average) in these regions with multiple promoters than in regions with a single functional promoter (33.69 average signals per region).

How, then, can the RNAP bind and use, as a functional promoter, a sequence with a computational score lower than those of other signals within the 250 bp long region? A second stronger putative promoter might help bring RNAP closer to the functional promoter, and then be re-positioned to such weaker but effective promoter by means of additional protein interactions, for instance with an activator. Precisely such a case for a re-positioning of an activator from a strong to a weaker site was reported, involving CRP and MalT. Re-positioning might, in principle, occur within regions of multiple promoters by interaction of a second RNAP molecule located nearby. Interference among closely located promoters has been documented for some divergent promoters. It has been shown experimentally that additional RNAP binding sites occur in the regions near functional promoters. Three of these promoter-like signals are embedded within A/T-rich sequences, perhaps to facilitate the localized melting required for transcription initiation process. Promoter-specific sequences are also A/T-rich. Thus, the presence of these signals could be a consequence of the overall high A/T base composition. *E. coli* has a slight tendency toward A/T-richness (58.4% AT within the strictly non-coding regions). However, we have found the same signal density phenomenon in the
GC-rich genome of Pseudomonas aeruginosa, which is 39% AT within the strictly non-coding regions (data not shown).

**Regions with latent promoters**

Knowing that the computational score or information content somehow reflects the energy of binding of the protein to the site, we wonder how to create an algorithm that emulating the RNAP is able to identify its true promoter when there is an apparently stronger signal nearby.

In some cases, the promoter-like signals did not seem to be distributed evenly within each upstream region (see Figure 4). There was a tendency for them to form clusters in regions of 10–30 bp where putative promoters overlap. One could wonder if the true promoter is part of a cluster in the abundant cases when it is not the strongest signal in its region. When multiple sites for attraction to the RNAP are closely located, the probability of binding would increase by the addition of each contributing site. This then provides a possible rationale for the role of what is observed. Assuming these sites are mutually exclusive, the probability that RNAP binds at any of these sites in the clusters would be calculated as:

$$p_{\text{cluster}} = \sum_{\text{s}} e^s$$

where $s$ denotes binding energies or scores for the RNAP binding sites. This formula represents the probability that RNAP binds at any of those sites in the cluster or region. However, each cluster is an aggregate formed by a different number of sites and scores. Measuring the density of signals and scores composing each aggregate would give us an empirical metric to compare them. We will call this pair of measures of density, the Effective score of an aggregate, or region. Since the promoter occupies around 60 bp, it provides room for the overlapping of several signals, where the Effective score will be the pair formed by the sum of the contribution of each signal, and the total number of signals found, that is:

$$\text{Effective score}_{\text{cluster}} = (\text{Total score, Total signals})$$

where:

$$\text{Total score} = \sum_{i=1}^{L} \text{score}(\text{promoter}_i)$$

$$\text{Total signals} = \sum_{i=1}^{L} \text{promoter}_i,$$

where $i$ denotes the $i$th promoter signal within the cluster or region of $L$ bp sized.

In order to identify regions with a rich density of putative promoters, we used a sliding window of 60 bp where we measured the Effective score. We then identified contiguous regions with the highest Total signals and those with the highest Total score. We will call these contiguous regions, the dense regions with the highest effective score or DRHES. When evaluating the predictive capacity of the scoring in equation (2) above, we were able to identify the DRHESs containing 91% of the 111 promoters in the training set, and 80% of the 392 in the testing set. Table 3 shows the results of this screening with each of the different classes of matrices obtained by the different selection criteria.

Using the Matrix.18,22,13,1,1 matrices, the DRHESs have an average size of 93.6 bp. Given that the length of the −10 and −35 boxes of this pair of matrices are 11 bp and 15 bp, respectively, and that the longest distance between them is 20 bp, there is a maximum of ≈48 possible choices to locate the real promoter within those DRHESs (93.6 − (11 + 15 + 20) = 48), which is not of great help for the accurate prediction of the site of transcription initiation.

The Effective score can be calculated base by base instead of using the sliding window of 60 bp. Figure 6 shows the Effective scores per base along four different 250 bp upstream regions, where the Total score at position $i$ is the added score of all promoters overlapping position $i$. Figure 6 also shows the Total signals, which is the frequency of signals overlapping each nucleotide (see equation (2)). As a result, we find the DRHESs corresponding to 92% of the promoters in the training set and 81% in the testing set, with ~40% of the signals being false positives in the testing set. Columns six and eight in Table 3 show the sensitivity of the Effective score method applied per base using the selected matrices of Table 1. These predictions generate DRHESs of 5.27 bases on average, where the precise +1 is not identified. An average of 14 promoter-like signals overlap with these DRHESs. We cannot, however, determine the site of transcription initiation employing this analysis.

Knowledge of the exact transcription initiation site of a gene is of utmost importance for many purposes, such as to delimit the entire protein-coding region, and to provide a powerful alternative for elucidating functional features of genes with no detectable sequence similarity, by linking them to other genes on the basis of their common promoter structure. Finally, the knowledge of the +1 position will help to predict and confirm the function of regulators given the distance of their binding sites to the transcription initiation point. A third alternative for more accurate predictions will be discussed below.

**The Cover function and improved predictions**

The problem of choosing one signal from the conglomerate of putative promoters is addressed in this section by a “divide and conquer” strategy that we have called Cover. It uses a relation of partial order on the set of promoter-like signals found within an upstream region. The selection of
Table 3. Sliding windows to sense high dense regions

<table>
<thead>
<tr>
<th>Matrices</th>
<th>Analysis set (111 promoters)</th>
<th>Testing set (392 promoters)</th>
<th>Analysis set (111 promoters)</th>
<th>Testing set (392 promoters)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sensitivity</td>
<td>Signals</td>
<td>Sensitivity</td>
<td>Signals</td>
</tr>
<tr>
<td>Matrix_18_24_13_0.5_2</td>
<td>87.39</td>
<td>0.97</td>
<td>77.55</td>
<td>0.85</td>
</tr>
<tr>
<td>Matrix_18_24_13_1_1</td>
<td>90.09</td>
<td>0.97</td>
<td>79.08</td>
<td>0.85</td>
</tr>
<tr>
<td>Matrix_18_22_13_1_5_0.5</td>
<td>85.59</td>
<td>0.98</td>
<td>78.83</td>
<td>0.84</td>
</tr>
<tr>
<td>Matrix_18_22_13_1_1</td>
<td>90.99</td>
<td>0.97</td>
<td>80.36</td>
<td>0.84</td>
</tr>
<tr>
<td>Matrix_18_15_13_2_1.5</td>
<td>84.68</td>
<td>0.98</td>
<td>73.47</td>
<td>0.87</td>
</tr>
<tr>
<td>Matrix_18_15_13_5_0.5_1.5</td>
<td>82.88</td>
<td>0.97</td>
<td>78.83</td>
<td>0.83</td>
</tr>
<tr>
<td>Matrix_18_15_13_1.5_1.5</td>
<td>83.78</td>
<td>0.97</td>
<td>79.34</td>
<td>0.85</td>
</tr>
<tr>
<td>Matrix_18_16_13_2_1.5</td>
<td>83.78</td>
<td>0.98</td>
<td>75.51</td>
<td>0.87</td>
</tr>
<tr>
<td>Matrix_18_16_13_2_2</td>
<td>83.78</td>
<td>0.98</td>
<td>75.51</td>
<td>0.87</td>
</tr>
<tr>
<td>Matrix_18_16_13_2_1</td>
<td>83.78</td>
<td>0.97</td>
<td>73.98</td>
<td>0.87</td>
</tr>
<tr>
<td>Matrix_18_16_13_2_0.5</td>
<td>88.29</td>
<td>0.97</td>
<td>73.72</td>
<td>0.87</td>
</tr>
</tbody>
</table>

Results using a sliding window and selecting the regions with the best Effective score calculated as described in equation (2). Sensitivity corresponds to the number of real promoters that have bases overlapping with those of the DRHESs (see Methods). The Signals column is the average number of DRHESs found by region. The first four columns show the results of calculating the higher Effective score using a sliding window of 60 bp. The average size of the DRHESs was 93.7 bp. The last four columns show the results with a sliding window of one base; in this case the average size of the DRHESs was 5.27 bp. The pair of matrices named Matrix_18_22_13_1_1 showed the highest capacity of recognition using either window size. With a window of 1 bp, this procedure yields a sensitivity of 81% with less than two possible DRHESs per region. Their corresponding consensus is: GTAATTAATATGAA[13..20][CTTATAATG]. Promoters are detected more frequently (in 65% of the cases) by the Total signals value, as opposed to only 15% by adding the scores, the Total score value. This again points to the multiple occurrence of overlapping putative promoters.

Figure 6. Distributions of the Total score (in blue) and Total signals (in red) by position in the upstream 250 bp regions of ada, ahpC, araE and argCBH promoter regions. Distances to the gene start are given as positive numbers. (a) An accumulation of signals is obtained at positions 52 to 54, within the 22 to 64 bp region occupied by the functional promoter, but the DRHES is far from the +1 (the ada − 10 box is at position 37). (b) Position 68 of the ahpC promoter region had the highest score for both curves. The promoter stretches from 24 to 68 (its − 10 box is at position 39). Note that the 40–67 bp region is the second highest scored region. (c) The peak in the araE region is at position 29, the functional promoter occupies from base 28 to base 68 having its − 10 box at position 29. (d) The 129 and 130 bases are the highest signals in the argCBH promoter region. The real promoter occupies bases 116 to 154 with the − 10 box located at base 132.
the best promoter is based on a competition among all promoters that have the same spacer between the \(-10\) and \(-35\) hexamers. The complete set of all promoter-like signals found with a cutoff of \(\mu - 2.5r\) were grouped into classes based on the length of the spacers between their \(-10\) and \(-35\) boxes. For example, the Matrix_{18.15.13.2.1.5} (see Table 1) produced seven different classes of spacers. The method we implemented to select the best signals within each class is based on a well-known partial order relation, the inclusion relation (see Methods). In our case, the relation uses both, an “intrinsic score”, the score resulting from adding the \(-10\) and the \(-35\) scores, and an “external score” based on the relative position of the predicted \(-10\) in relation to the beginning of the gene. This external, or positioning score, has been coded as a relative probability that follows the distribution shown in Figure 1. A promoter A is said to be included by another promoter B if, and only if, the scores of promoter B are each better than the corresponding scores of promoter A. More precisely, “B includes A” or, \(A \subseteq B\) when (3):

\[
(A \subseteq B) \text{ if and only if} \\
\text{(Score}_{10}A + \text{Score}_{35}A) = \text{Score}_{10}B + \text{Score}_{35}B
\]

and \(\text{Position}_{10}A \leq \text{Position}_{10}B\)  \hspace{1cm} (3)

This inclusion relationship defines a cover set for each of the different spacer classes, where each candidate resides within any subset of the cover set (see Methods for an example). The subsets with the highest cardinality within each class of spacers were selected, and the predicted “true” promoter was defined to be the upper border of such subsets. In other words, the promoter-like signal with the highest number of included signals is taken to be the predicted promoter.

The complete schema of the strategy of the Cover method is available on the web†. Table 4A and B show the results for both the training and testing sets, respectively. The first three columns correspond to phase 1 in the methodology. In this phase, we searched for all the signals with scores higher than \(\mu - 2.5r\), because this cutoff produced the best overall results. The spacers varied from 13 bp to 19 bp. We tested all four classes of weight matrices discussed in Table 1. Columns four to six show, for comparative purposes, the results of selecting as the predicted promoter the signal with the highest score. The last three columns show the results using the Cover method, which in the testing set reached a sensitivity of 86% with an accuracy of 0.53 using the Matrix_{18.15.13.2.1.5} pairs of matrices. These selected promoters have a well-defined \(-10\) and \(-35\) box, supporting the prediction of the precise +1 box based on the \(-10\) box.

In principle, this method generates seven predicted promoters per region, one for each spacer subclass. However, many of the selected promoters had the same \(-10\) box position with different spacers for different \(-35\) boxes. Using this information, the number of predicted promoters (more accurately the number of predicted +1s) is reduced to an average of 4.7 candidates per region. We tested if those 4.7 candidates were forming some kind of arrangements. An average of 3.7 signals appeared in clusters in which these signals are overlapping; 97% of the promoter regions analyzed here showed these clusters. Cover function was able to correctly detect 337 promoters of the 392 of the training set, 74% of those 337 are in clusters. Another interesting fact is that 70% of the promoter-like signals found by Cover are upstream of real promoters; this might suggest a regulatory role. Figure 7 illustrates the result of Cover in the lac promoter region. Four of the five results given by Cover are clustered (lac PC1, lac PC2, lac PC3, and lac PC4) being lac PC2 the real one. PC stands for promoter by cover.

Even though the canonical picture of a promoter is a 40 bp to 60 bp region defined by the transcription start-point (+1) and the two conserved hexanucleotide sequences centered 10 bp and 35 bp upstream from the +1, it has been shown that promoters exist in clusters, as a series of overlapping potentially competing RNAP interaction sites. In a small sample of four promoter regions (lac, IS50, tet, and gal), it was found that three promoter regions presented an arrangement of at least three overlapping promoters. The lac region presented six promoter-like signals plus the real one, the IS50 promoter region has two promoter-like signals near the functional one, and one promoter-like signal is in the region of the gal promoters. Figure 7 shows the positions of the six promoters-like signals found experimentally for the lac promoter region. If we calculated the performance of the RNAP in these experiments, we found a sensitivity of 1, and precision of 0.30, finding an average of 4.3 signals per region.

The sensitivity, specificity and precision of searching using only weight matrices, selecting the signal with the highest score, obtaining the DRHEs base by base, and applying the Cover method are shown for comparative purposes in Figure 8. These measures were calculated using the equations (4) described in Methods. The specificity using weight matrices at the lowest threshold is initially practically given the large number of accepted signals (Figure 8(a)). Selecting the strongest signal, for instance with the set of matrices denoted Matrix_{18.15.13.2.1.5} gives one candidate per region with a precision of 46%, but the sensitivity never reaches more than 40% in the testing set (Figure 8(b)). Figure 8(d) shows that the Cover method produces the desired balance of sensitivity and specificity (86% and 85%, respectively). This method also improves in sensitivity when compared to the strategies based on overlapping

†http://www.cifn.unam.mx/Computational_Genomics/PromoterTools/Tutorials
of signals (Figure 8(c)), which do not give a precise prediction of a promoter +1. From the graph in Figure 8(c), one can observe that for the same number of false negatives, the number of predictions is much larger in the Cover method than the strongest signal and the DRHEs approaches. We reason that the capability of the Cover strategy is a result of putting separately into competition each subclass of signals based on their spacers. This procedure prevents competition among overlapping signals. The most frequent overlapping occurs for alternative −35 boxes sharing the same −10 box. Also, since in TATAAT the uppercase bases are the most conserved in the −10 region, shifting two base-pairs from a TATAAT box can generate a second good scored −10 box (.TA...). It has been shown that sometimes shifting a −10 box by adding bases to the spacer between the −10 and −35 elements, produces a second −10 functional box recognized by RNAP.

Since the Cover function suggests a sorting of signals, it might be that if we pick up the four strongest promoters, we could obtain similar results to those generated with our Cover method. The sensitivity and precision obtained when selecting the best four signals are 72% and 26%, respectively, with an accuracy of 49%. The poorer performance of this approach when compared to that using Cover is explained by the fact that Cover imposes two requirements on the best candidates, one internal, based on the scores generated by similarity to consensus, and one external, based on their relative distances to the beginning of the gene. What makes a sequence to be a functional promoter may depend on the specific context within the upstream region in which it is

Table 4. Cover function method as compared with weight-matrix search, and strongest promoter naive methods

<table>
<thead>
<tr>
<th>PATSER</th>
<th>Strongest</th>
<th>Cover function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>Sensitivity</td>
<td>Signal</td>
</tr>
<tr>
<td>A. Training set Information content</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Matrix_18_24_13_0.5_2</td>
<td>0.51</td>
<td>0.99</td>
</tr>
<tr>
<td>Expected frequency</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Matrix_18_24_13_1_1</td>
<td>0.51</td>
<td>0.99</td>
</tr>
<tr>
<td>%Prom in μ − σ</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Matrix_18_22_13_1.5_0.5</td>
<td>0.51</td>
<td>0.98</td>
</tr>
<tr>
<td>Matrix_18_15_1_0.5_1.5</td>
<td>0.5</td>
<td>0.98</td>
</tr>
<tr>
<td>Homology to consensus</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Matrix_18_15_1_2_1.5</td>
<td>0.5</td>
<td>0.98</td>
</tr>
<tr>
<td>B. Testing set Information content</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Matrix_18_24_13_0.5_2</td>
<td>0.49</td>
<td>0.94</td>
</tr>
<tr>
<td>Expected frequency</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Matrix_18_24_13_1_1</td>
<td>0.49</td>
<td>0.95</td>
</tr>
<tr>
<td>%Prom in μ − σ</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Matrix_18_22_13_1.5_0.5</td>
<td>0.5</td>
<td>0.96</td>
</tr>
<tr>
<td>Matrix_18_15_1_0.5_1.5</td>
<td>0.5</td>
<td>0.96</td>
</tr>
<tr>
<td>Homology to consensus</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Matrix_18_15_1_2_1.5</td>
<td>0.49</td>
<td>0.92</td>
</tr>
<tr>
<td>Matrix_18_15_1_2_2</td>
<td>0.49</td>
<td>0.93</td>
</tr>
<tr>
<td>Matrix_18_16_1_2.2</td>
<td>0.49</td>
<td>0.93</td>
</tr>
<tr>
<td>Matrix_18_16_1_2.1</td>
<td>0.49</td>
<td>0.93</td>
</tr>
<tr>
<td>Matrix_18_16_1_2.0_5</td>
<td>0.49</td>
<td>0.93</td>
</tr>
</tbody>
</table>

Results in the training and testing sets of alternative methods for predicting promoters. In both sets, we calculated accuracy, sensitivity, and the number of predicted promoter-like signals per 250 bp region, using three methods. In the first approach, we employ only the weight matrices to score a signal with μ − 2.5σ as a cutoff. The PATSER program was used to score the signals. The second column shows the result of selecting the strongest signals in terms of the added scores of −10, −35, and spacers. The third alternative was applying the Cover function with the results shown in column 3. A. Results for the training set. This set was used to build the weight matrices, and then define additional parameters such as the distribution of distances from the −10 boxes to the beginning of the genes and the distribution of spacers lengths between −35 and −10 boxes. The generated sensor matrices together with their additional parameters were employed for the analysis of the testing set. B. Results in the testing set of 392 promoters.
embedded. The use of this relative distance to the beginning of the gene makes our method hard to reconcile with a direct computational modeling of RNAP binding and transcription initiation, since it is unlikely that the polymerase could detect such a distance.

Complexity in upstream regions: multiple true and putative promoters

A more detailed analysis is required to dissect the biology of the observed overlapping of signals. Out of the 392 sigma70 promoters used as testing set, 157 (40%) are at regions where additional promoters in the same direction of transcription, using sigma70 or a different sigma factor, are present (see Table 5). The 28% of the promoters in the training set are in this situation, with 19% of the upstream regions containing two or more active promoters.

On the other hand, there are 238 (61%) true opaque promoters, those that have a lower score with respect to other signals within the same 250 bp region, while the remaining 154 (39%) promoters are visible, they are the strongest signal in the region. We did not see a significant distinction between the opaque and the visible classes of functional promoters depending on the type of regulation (positive, negative or dual), or in the position of the promoter in the transcription unit (upstream or internal), or any other properties. As shown in Table 5, in general, all samples separated by these properties contain around 60% upstream regions where the true promoters are opaque and 40% where the true promoter is visible.

In terms of binding strength, which might correlate with scores, it has been thought that activator proteins should regulate weak promoters, and that strong promoters should be regulated negatively.48 Our data do not show a correlation between positive

---

**Figure 7.** Results of applying Cover in the region of the lac promoter, and the promoter-like signals detected experimentally in this region. P1, P2, P3, P-2A, and P + 1T corresponding to Figure 1 presented by Reznikoff, and P4 to Figure 1 presented by Czarnecki et al.4 Four of five Promoter by Cover (PC) predictions form a cluster similar to that seen experimentally. PC2 corresponds to P1, the functional lac promoter. PC3 and PC4 correspond to the P3 and P2 promoter-like elements identified both in vitro and in vivo to have weak promoter activity. PC1 is equivalent to P + 1T, a mutation that changes the A base, the + 1 position of P1, into T. The + 1 of the lac promoter is located 39 bp upstream of the lacZ start codon. Positions of the first base of the −10 boxes, with respect to the start codon, are marked as positive numbers above each promoter drawing. The scores (−10 box score + −35 box score) of the PC predictions are shown in the legend.

**Table 5.** Structural and functional characteristics of the promoter testing set

<table>
<thead>
<tr>
<th>Regulation type</th>
<th>% Opaque real promoter (238)</th>
<th>% Visible real promoter (154)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unknown (228)</td>
<td>(139) 58.4</td>
<td>(89) 57.8</td>
</tr>
<tr>
<td>Repressor (43)</td>
<td>(23) 9.7</td>
<td>(20) 12.9</td>
</tr>
<tr>
<td>Dual (54)</td>
<td>(38) 15.97</td>
<td>(16) 10.4</td>
</tr>
<tr>
<td>Activator (67)</td>
<td>(38) 15.97</td>
<td>(29) 18.8</td>
</tr>
<tr>
<td>Strand</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Forward (191)</td>
<td>(114) 47.9</td>
<td>(77) 50</td>
</tr>
<tr>
<td>Reverse (201)</td>
<td>(124) 52.10</td>
<td>(77) 50</td>
</tr>
<tr>
<td>Promoter kind</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Single promoter (235)</td>
<td>(127) 53.4</td>
<td>(108) 70.1</td>
</tr>
<tr>
<td>Multi-promoter* (157  in 96 regions)</td>
<td>(111) 46.6</td>
<td>(46) 29.8</td>
</tr>
<tr>
<td>Region kind</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coding (91)</td>
<td>(54) 22.7</td>
<td>(37) 24</td>
</tr>
<tr>
<td>Divergent (87)</td>
<td>(55) 23.1</td>
<td>(32) 20.8</td>
</tr>
<tr>
<td>Non-coding (214)</td>
<td>(129) 54.2</td>
<td>(85) 55.2</td>
</tr>
<tr>
<td>Position in TU</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Start TU (358)</td>
<td>(217) 91.2</td>
<td>(141) 91.5</td>
</tr>
<tr>
<td>Inter TU (34)</td>
<td>(21) 8.8</td>
<td>(13) 8.4</td>
</tr>
<tr>
<td>Binding site overlapped (51)</td>
<td>(35) 14.7</td>
<td>(15) 9.7</td>
</tr>
<tr>
<td>Signals by region (8815)</td>
<td>(5554) 23.3</td>
<td>(3261) 21.2</td>
</tr>
<tr>
<td>Density methods</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Found signals</td>
<td></td>
<td></td>
</tr>
<tr>
<td>window 60</td>
<td>(165) 69.3</td>
<td>(123) 79.9</td>
</tr>
<tr>
<td>Found signals (base X base)</td>
<td>(165) 69.3</td>
<td>(126) 81.8</td>
</tr>
<tr>
<td>Cover function</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Found</td>
<td>(192) 80.7</td>
<td>(145) 94.2</td>
</tr>
</tbody>
</table>

Data of 392 promoters classified as opaque or visible based on scores generated by the Matrix.18,15,13_2.15 matrices. Numbers in parentheses are absolute numbers, followed by percentage values.

* Here, 30% have a strong real promoter.
Concluding Remarks

Although the matrices representing the canonical model of the sigma70 promoter displayed the best predictive ability, they did not result in the best statistical model, either in terms of expected frequency, information content, or in the generation of a more compact distribution of promoter

---

Figure 8. (a) to (d) The different thresholds in the abscissas indicate the number of standard deviations subtracted from the average. The sensitivity in red and specificity in blue are represented in the ordinates. The results correspond to the testing set of 392 promoters. (a) Results with phase 1, using the PATSER program to search for signals. 

(b) Second phase, when the highest scores are selected. (c) Second phase using higher effective score base by base. 

(d) Second phase using the Cover function. 

(e) Number of false negatives versus total number of predictions for each alternative method. 

or negative type of regulation and opaque or visible promoters. There is a slightly greater abundance of promoter-like signals in regions with opaque promoters than in those with visible promoters, suggesting that multiple signals play a role with opaque promoters. Multi-promoter regions also show a slightly higher density of promoter-like signals as compared to single promoter regions.

---

Sigma70 Promoters in E. coli: Dense Overlapping Signals
scores. This indicates that the best statistical model does not reflect the functional nature of the RNAP binding sites due to the high variability present in the promoter sequences.

The predictive capacity of our Cover function is 86%; however, its accuracy is 53%. An interesting explanation for this limitation in promoter prediction would be to consider that most, if not all, promoters are regulated by at least one transcriptional factor. It is certainly hard to know if there really are constitutive promoters, or if constitutive promoters are those whose regulation has not been identified. The transcriptional regulator could contribute when binding, to eliminate alternative promoter-like sites, and to contribute in the precise positioning of the RNAP holoenzyme in the adequate functional promoter site. This would imply that, in addition to their classic activation and repression roles, regulatory sites are playing an additional role of exclusion of other promoter-like signals in the sigma 70 promoters in E. coli. In different contexts, a role of “address” provider has in fact already been suggested for activators of sigma70 promoters, reminiscent of the role of IHF when bending DNA and contributing in this way to the specificity of a regulator and the corresponding sigma54 promoter. The elements allowing the promoter to be identified precisely by sigma70 could be the trans-acting factors, such as those that change the DNA topology and expose regions to RNAP binding, or regulatory proteins that can anchor the RNAP to the promoter site. About 10% of the promoters comprised by RegulonDB have a DNA topology modulator in their region (i.e. Fis), and 47% have at least one regulatory protein reported. Taking into account those trans-acting factors might reduce the number of false positives of our methodology. Another misguiding factor could be the similarity between the sigma70 and sigma38 consensi, which have an almost identical –10 region: 7% of the promoter regions contain sigma70 and sigma38 promoters. Future attempts toward promoter predictions should consider the inclusion of these regulation factors. The performance of the method depends also on the selection of the consensus matrices and their scoring method. Further work around other statistical parameters or ways to score a site given a matrix might improve the results.

As mentioned in Introduction, the scientific challenge of computational prediction of promoters is that of reproducing the performance achieved by the RNAP itself in specifically recognizing its functional sites for transcription initiation. We are still far from a 100% performance, and far from a method that implements computational modules that could be interpreted in terms of the biophysical process of recognition. For instance, there is no known mechanism that would enable the upstream bound RNAP to sense the distance to the beginning of the gene, sometimes 100 or 200 bp downstream. This distance is nonetheless a useful piece of information in our Cover methodology.

The observed high density of overlapping signals suggests that the observed multiple promoters and RNAP binding sites found in the lac promoter is a widespread phenomenon found in practically all promoter regions in the genome of E. coli. As mentioned before, this might not be due to the AT content, since a similar abundance of signals occurs in P. aeruginosa. Our computational method takes advantage of such signals, even though they might be evolutionary vestiges with little direct kinetic effects in promoter activity. We suggest that these signals are latent due to the effective binding of activators, repressors as well as additional functional promoters.

Methods

The computational search for promoters that we performed was limited to regions upstream of genes based on the annotations of the complete E. coli genome. GETools was used to extract all the upstream regions of 250 bp, where 90% of all known promoters are included.

A total of 791 known promoters, out of which 599 are sigma70 promoters, their precise +1 nucleotides of transcription initiation, as well as the information on their regulation, were obtained from RegulonDB. From this collection, 584 sigma70 promoters were selected in this work and divided into two sets, one with the 116 promoters, that were initially analyzed in, was used as training set, and the second and larger set of 392 promoters used as the testing set. The remaining 76 were located further upstream than 250 bp and were not used in this work. The rationale for selecting the training set includes three reasons: first, historical; second, the associated regulation of this dataset is well known; and third, the dataset is a balanced collection of negatively and positively regulated promoters.

WCONSENSUS selects the best matrix based on maximizing the information content of the sequences and minimizing the expected frequency of finding the matrix by chance, given the known sequences. The alphabet or a priori probabilities may be selected by the program or it can be set using the frequencies of the letters in the set of sequences of the input. The correction to the information content of each matrix is provided by a value of expected standard deviation for an alignment generated randomly [0.5, 1.0, 1.5 and 2.0]. Several multiple alignments were generated using WCONSENSUS version v5c. The best alignments and associated weight matrices were always selected using the second cycle, where every input sequence contributes to the final alignment. After the matrices were created, we calculated the spacers between −10 and −35, the distance from the −10 box to the +1, and the position of the −10 box from the gene start (see Figure 2). The spacer scores were calculated as the log of the relative frequencies of the lengths of these elements in the analysis set of 116 promoters. The PATSER program allows one to score the motifs of a given sequence against a

http://www.cifn.unam.mx/Computational_Genomics/GETools
A partial ordered set is a pair (A, R) where R is a partial order relation on the set A. A few examples are:

- \((Z, \leq)\) where \(\leq\) is the common “less than or equal” relation on the integers.
- \((P(X), \subseteq)\) where \(P(X)\) is the power set of \(X\) and the relation \(\subseteq\) is given by the common inclusion of sets.

The logical properties of partial order relation structure the promoter-like signals found on a region in a hierarchical fashion. As example, consider \(P(a,b)\) is a promoter-like signal con score \(a\) and position value \(b\). Let us allow \(P1(1.3,0.3), P2(1.8,0.05), P3(1.5,0.6), P4(2.0,0.8)\) and \(P5(8.5,0.1)\) to be the promoter-like signals found within a given region. Defining the relation \(\subseteq\) as in equation (3) and applying it to the above promoters, we generate the following ordered chains:

\[
\begin{align*}
P1(1.3,0.3) & \subseteq P3(1.5,0.6) \subseteq P4(2.0,0.8) \\
P2(1.8,0.05) & \subseteq P4(2.0,0.8) \\
P2(1.8,0.05) & \subseteq P5(8.5,0.1)
\end{align*}
\]

All the chains generated by an ordered partial on a set A form a set \(A'\) of subsets of A. The set \(A'\) of the promoters of our example is:

\[A' = \{[P1, P3, P4], [P2, P4], [P2, P5]\}\]

The set \(A'\) is called the cover set of A, and we said the subsets of \(A'\) are covering A. The size (cardinality) of the longest chain or subset in \(A'\) is called the order partial length, and the biggest element in that set is called the upper border. In our example, the promoter-like signal predicted to be the “functional” promoter will be P4, since this is the upper border of the longest subset generated by the partial sort.

Acknowledgements

A.M.H. thanks Rob Rohde-Szudy for advice and comments on the text, Ernesto Ortiz-Suri for fruitful discussions and critical review. The authors thank discussions with Gabriel Moreno-Hagelsieb, Jay Gralla, Gary Stormo, M. Carmen Gomez-Eichelmann, Miguel A. Cevallos and Agustino Martinez-Antonio, as well as Victor del Moral and César Bonavides for computational support. A.M.H. has been supported by a PhD grant from CONACyT. This work has been supported by grant 0028 from CONACyT and by grant GM62205 from NIH.

References

36. Richet, E. & Sogaard-Andersen, L. (1994). CRP induces the repositioning of MalT at the *Escherichia*


*Edited by R. Ebright*

(Received 3 April 2003; received in revised form 22 July 2003; accepted 22 July 2003)