# Anatomy of *Escherichia coli* $\sigma^{70}$ promoters

Ryan K. Shultzaberger [a,b] Zehua Chen [a]
Karen A. Lewis [a,c] and Thomas D. Schneider [a,*]

[a] *Center for Cancer Research Nanobiology Program, National Cancer Institute at Frederick, P. O. Box B, Building 469, Room 144, Frederick, MD 21702-1201, USA, 301-846-5581 (-5532 for messages), fax: 301-846-5598.*

[b] *Present address: University of California, Department of Molecular and Cell Biology, 16 Barker Hall, Berkeley, CA 94720-3202.*

[c] *Present address: University of Texas Southwestern Medical Center at Dallas, Department of Physiology, 5323 Harry Hines Blvd., Dallas, TX 75390-9040, (214) 645-6010*

version = 4.23 of flexprom.tex 2007 Feb 6

*Nucleic Acids Research* **35**: 771-788 (2006)

## Abstract

Information theory was used to build a promoter model that accounts for the −10, the −35 and the uncertainty of the gap between them on a common scale. Helical face assignment indicated that base −7, rather than −11, of the −10 may be flipping to initiate transcription. We found that the sequence conservation of $\sigma^{70}$ binding sites is $6.5 \pm 0.1$ bits. Some promoters lack a −35 region, but have a $6.7 \pm 0.2$ bit extended −10, almost the same information as the bipartite promoter. These results and similarities between the contacts in the extended −10 binding and the −35 suggest that the flexible bipartite $\sigma$ factor evolved from a simpler polymerase. Binding predicted by the bipartite model is enriched around 35 bases upstream of the translational start. This distance is the smallest 5′ mRNA leader necessary for ribosome binding, suggesting that selective pressure minimizes transcript length. The promoter model was combined with models of the transcription factors Fur and Lrp to locate new promoters, to quantify promoter strengths, and to predict activation and repression. Finally, the DNA-bending proteins Fis, H-NS and IHF frequently have sites within one DNA persistence length from the −35, so bending allows distal activators to reach the polymerase.

*Key words:* promoter, information theory, extended −10, $\sigma^{70}$, DNA bending

running title: Anatomy of *Escherichia coli* $\sigma^{70}$ promoters

## 1 Introduction

Transcriptional regulation is essential to the viability of the cell [1–3]. In prokaryotes, many molecules can contribute to, or detract from the stability of the initiation complex [4]. The

* Corresponding author.
*Email address:* toms@ncifcrf.gov (Thomas D. Schneider).
*URL:* http://www.ccrnp.ncifcrf.gov/∼toms/ (Thomas D. Schneider).

minimum requirement for RNA polymerase binding is recognition of the promoter by the $\sigma$ factor [5–8]. In general, prokaryotic RNA polymerases can interchange a number of $\sigma$ factors which bind and initiate different groups of genes [9]. $\sigma^{70}$ is the most commonly used $\sigma$ factor in *Escherichia coli* and it is responsible for the initiation of most genes [9]. This paper will only focus on promoters bound by $\sigma^{70}$.

To successfully model initiation, it is necessary to construct a model that unifies multiple components. The conventional model for promoter recognition by $\sigma^{70}$ is the binding of two regions upstream of the transcription start point, named the $-10$ and $-35$ because of their spacing relative to the first transcribed base [10,11]. The initiation complex is also further stabilized by the carboxy-terminal domain of the two $\alpha$ subunits of the core enzyme ($\alpha$CTD), which can either interact directly with upstream DNA, or with regulatory proteins [12]. To add to the complexity of the system, recognition of the $-10$ alone can be sufficient for initiation to occur [13–15]. The initiating polymerase can be thought of as moving ship that needs to be anchored down [16]. The varying affinities of the binding components for the promoter would correlate to varying weights holding the polymerase in place. The sum of these components must have enough energy to stabilize the polymerase against thermal noise. Therefore, in order to model promoter binding, we need to consider the relative affinity of each molecule affecting the stability of the initiation complex.

In addition, the $\sigma$ factor is flexible. That is, the distance between the $-10$ and $-35$ binding sites is not fixed. This flexibility affects the affinity of the polymerase for the sequence [10]. If we treat $\sigma$ factor bound to core as a simple harmonic oscillator, then expansion or contraction of the polymerase when binding to promoters with varying spacings would strain the molecule and reduce the amount of energy available for stabilization. Since the initiation rate is affected by spacing [10,17–21], our model needs to take into account this internal strain.

Traditionally three possible spacings have been proposed at which the $-10$ and the $-35$ bind relative to each other, $17 \pm 1$ bases, but initiation over a larger range, 15 to 20 bases, has been shown [10,22,19]. These spacings correspond to the number of bases between the $3'$ end of the $-35$ hexamer and the $5'$ end of the $-10$ hexamer. The observed optimal spacing of 17 bases [10] places the centers of the two hexamers on the same face of the DNA 23 bases apart, approximately 2 helical twists of B-form DNA [23], suggesting that the polymerase has a DNA-structure dependent contact. There is also a correlation between the extent of negative supercoiling and the amount of transcription from a promoter [22,21], which demonstrates that the $\sigma$ factor is sensitive to genomic structure [24,25].

Although neural networks and hidden markov models have been used to model promoter binding [26–28], constructing these models usually requires the untenable assumption that large stretches of sequence do not contain sites, and the resulting parameters have not been easy to interpret. Other attempts have been made to model promoters using methods not based on HMMs [29–33], but these methods do not uniformly measure the contribution of all components in the initiation complex ($-10$, $-35$, and gap). Hertz and Stormo presented a model in which they subtracted the gap penalty for the optimal spacing from each gap

value [31], so that there was no penalty for having the optimal spacing. Their formula to evaluate gap penalties implies that a set of sites that have several equiprobable gap lengths would have no penalties [31] even though flexibility decreases information [34]. The method used in this paper, which was previously used to investigate ribosome binding sites [34], does account for gap variability with equiprobable gap lengths. In addition, promoter strengths are not determined purely by the binding of the $\sigma$ factor. Transcriptional activators and repressors contribute to and detract from the accessibility of DNA by the RNA polymerase. In order to uniformly model the flexible binding of the $\sigma^{70}$ in conjunction with transcriptional regulators, we used information theory.

Information theory was developed by Claude Shannon to quantify the transfer of information in communications [35,36]. It has proven to be useful when applied to a variety of biological systems [37–41], mainly in quantifying how specific a given DNA-binding protein is, based on the amount of variability within its binding targets. A lower binding site variability corresponds to a higher information content [37]. For convenience, information is generally measured in bits, the choice between two equally likely possibilities. A greater information content for a set of binding sites (more bits of information) generally will have more specific binding and a higher binding affinity.

Prokaryotic ribosomes, like the $\sigma$ factor, have two binding elements separated by a variable distance. In previous work, we used information theory to model 95% of the *E. coli* ribosome binding sites [34]. This flexible model took into account the conservation of both the initiation codon and Shine-Dalgarno regions, and the statistics of the variable spacing between them. Here we applied the same theory to the $\sigma^{70}$ promoter components ($-35$, $-10$ and their spacing) to create a cohesive model of promoter binding.

An important difference between the ribosome model and the promoter model is that the ribosome model is only composed of two binding elements. The promoter model can be made up of a large number of binding elements with parameters governing how each element behaves, and how the elements interact with each other. This paper shows how an internally consistent multi-part model can be constructed directly from experimentally proven sites, and discusses how this model can be used to predict and identify control systems. We are also interested in understanding the fundamental workings of the RNA polymerase. To this end, we examined the variation in the promoter as a function of spacing, global trends in accessory molecule binding relative to the promoters, and the relationship between ribosome binding sites and promoters. We also propose that the flexible bipartite binding site evolved from a rigid extended $-10$ binding mode.

## 2 Materials and Methods

### 2.1 Constructing the promoter model

We built our $\sigma^{70}$ binding model by aligning and refining the sequences upstream of 599 experimentally determined transcription starts reported in the RegulonDb database [42] and 85 starts from the PromEC database [43] that were not included in RegulonDb. To

align binding sites we used the **malign** program to maximize the information of either the −10 or −35 by shuffling the sequences [44]. To refine the model, we iteratively removed all sequences with an information content less than 0 bits of information [45,46] until we converged on a consistent set of sequences. Since there is both a variable spacing between the −10 and the transcription start point, and between the −35 and the −10, this process was not trivial, and we describe below how we converged on our final model.

To align the −10, we embedded the DNA sequences −15 to −3 bases upstream of the transcription start site in random DNA, so that our alignment would not be biased by the −35 or by the preference of adenine at the transcription start point. We realigned the −10 region to maximize the information by using the **malign** program [44] over the range of −12 to −7 bases upstream of the transcription start, allowing for the sequences to shift up to three bases in either direction. Since nearby transcription start points could potentially use the same −10, we identified and removed transcription starts from our dataset that were within 15 bases of another site with a lower genomic coordinate (arbitrarily chosen) and had the same orientation. This prevented the same −10 from appearing multiple times in our model, and decreased the size of our dataset from 684 to 620 starts. We then did a cyclic refinement on these sites to remove sequences from our dataset that were not identified as sites by our model. To do this, all sites that had an information content [46] lower than zero bits were removed, and the model was rebuilt. The zero-bit cutoff was used because it represents a version of the second law of thermodynamics: sites with positive information correspond to negative $\Delta G$ of binding [45,46]. This approach was successfully used for constructing ribosome [34] and splice site models [47]. Removing and rebuilding was continued until no negative sites remained in the set. This reduced our number of sites from 620 to 559. The refined multiple alignment gave us a well conserved −10 region (Fig. 1). At each position, the base conservation of the −10 corresponds to the number of mutants found at that position by genetic studies [11].

We do not adhere to the conventional numbering system used in describing the distance between the −10 and the −35. The conventional numbering of the spacer is the number of bases between the two hexamers [10]. That is, ttgaca**NNN**tataat would have a spacing of three. Since the convention for position numbering in asymmetric sequence logos is to choose a strongly conserved base, we will refer to the second base in each hexamer as zero, and all spacings will be reported as the difference between those coordinates. Therefore, all values for our spacing are six bases greater than the numbering previously used. For example, t**T**gacannnt**A**taat would have a spacing of nine (the difference between the capital T in the −35 hexamer and the capital A in the −10 hexamer), rather than three. So the classical spacing of 17 bases is 23 in our notation. The only way we would be able to adhere to the conventional numbering system would be to assign a base outside one of the hexamers as the zero coordinate. This would be confusing in sequence analysis using sequence walkers [48]. Furthermore, each sequence walker always has the integer zero in its coordinate system so that one can easily and unambiguously locate a binding site and then specifically identify bases within the binding site.

Aligning the −35 was more difficult than aligning the −10. The traditional model of RNA

polymerase binding only allows for three different spacings between the −35 and the −10 [11]. Mutational data have shown that this range could be expanded to six positions [10], but that at these expanded spacings the amount of transcription is reduced substantially. The optimal spacing of 23 bp (McClure's spacing of 17 bp) [11] places both hexamers on the same face of the DNA within their respective major grooves [23], indicating that the spacing may be dependent upon DNA structure. Therefore, a new alignment method is needed that takes into account the structure of the DNA.

The algorithm for aligning the −35 is different from our previous approach of aligning upstream sequences in flexible models (*i.e.* modeling the Shine-Dalgarno relative to the initiation codon in ribosome binding sites [34]). As with ribosomes, our approach was to create a model *de novo* from experimental data, so as to avoid biases in previous models. Using a sequence logo [49], we observed a weak conservation upstream of the aligned −10s in the region expected for the −35, 23 bp upstream. We determined that the conservation of this region was low because a number of sites with a different spacing were overlapping, reducing the total sequence conservation. We performed a cyclic refinement of the region which corresponds to the −35 hexamer at the optimal spacing from the −10 in order to pull out a preliminary −35 model. This cyclic refinement gave a reasonably well conserved −35 sequence logo, which matched the conventional hexamer consensus [11,50], and this alignment was used for an initial model. After refinement, we used malign to allow for the sites to be moved 1 base in either direction, so as to maximize the information in this refined −35.

Using the program **multiscan**, the initial −35 model was scanned over the region upstream of the −10 of every promoter in order to find the −35 which most closely matched this model for each site. Of the 559 promoters scanned, 421 had a −35 site $> 0$ bits in the range of 21-26 bases upstream of the −10 (this corresponds to McClure's spacing of 15-20 bases [10]). The alignment having the strongest site was used, and the total site strength was calculated using the flexible information equation described previously [34]:

$$\text{Flexible Site Information} = R_i(-35) + R_i(-10) - GS(d) \quad \text{(bits/site)}, \tag{1}$$

where $R_i(-35)$ is the individual information [46] of the −35 site, $R_i(-10)$ is the individual information of the −10 site, and $GS(d)$ is the gap surprisal for each spacing $d$, which is based on the major groove accessibility curve of B-form DNA [51,41]. The equation used to generate a distribution which corresponds to the accessibility of the DNA by the RNA polymerase is:

$$\text{Accessibility} = n(d) = 1 + cos(\frac{2\pi}{w}(d - \text{center})), \tag{2}$$

where $w$ is 10.6 bases (one turn of B-form DNA), center is 23 bases (the optimal spacing between the −35 and the −10), and $d$ is the distance of the −35 from the −10, in bases. This equation describes the major groove accessibility [51,41] in which direct contacts are more accessible and contacts on the opposite face of the DNA are not accessible. The surprisal

was then calculated for each position using the gap surprisal equation published previously:

$$GS(d) = -\log_2 \frac{n(d)}{n} + e(n) \quad \text{(bits/spacing)}. \tag{3}$$

$n(d)$ is the accessibility at spacing $d$, and $n$ is the sum of all accessibility values over the allowed values of $d$. $e(n)$ is a small sample correction value [37,34]. By using these equations to model the structure of DNA, we gave preference to $-35$ sequences which were in physically reasonable positions.

Once we had identified the $-35$, the $-10$ and the spacing for each promoter, we dispensed with the "scaffolding" equations described above and built a flexible model directly from the sequence data. We did a further cyclic refinement on this set by removing promoters with a flexible information less than zero bits, reducing the number of promoters from 421 to 401. Our final model, therefore, contains 59% of the sites that are in our original database. Transcriptional regulators can provide informational contacts through the $\alpha$CTD and this could account for some of the information used by polymerases [52]. As a result, many promoters may have poorly conserved, or highly variable, $\sigma$ binding sites. The refinement process made the model self-consistent (containing similar sites) and it can therefore be regarded as a basal promoter model. The excluded sites are not consistent with this basal model, and presumably they initiate by some method other than the sole recognition of the $-10$ and $-35$ (such as an extended $-10$ or activation by another protein).

Promoter binding and transcriptional regulation is more complex than our previous flexible modeling system was able to handle because of the contribution of activation proteins [52]. Therefore, we created an algorithm that not only considers the strength of the two-part $\sigma^{70}$ site, but can also include the information contributed by activating proteins. In order to do this we used the multiscan algorithm.

### 2.2 Multiscan Algorithm

**Multiscan** is an extension of the **biscan** program that is used to model the flexible prokaryotic ribosome [34]. Translational initiation in prokaryotes requires contact at both the P site (or initiation region, IR) and the Shine-Dalgarno (SD). Because of the flexibility of the ribosome, these contacts can occur at different spacings, anywhere between 4 and 18 bases. In order to assess the information present in ribosome binding sites, the contributions of the Shine-Dalgarno, the initiation region and the spacing between them all have to be considered. The equation for calculating the information for a two-part model with variable spacing was given in equation (1).

The RNA polymerase is similar to the ribosome in that it makes two contacts (the $-10$ and the $-35$) with some variable distance between them. Therefore the flexible information analysis used with ribosomes can also be used to describe the binding of the $\sigma$ factor to the promoter. The difference between translational and transcriptional initiation is that auxiliary proteins can also bind to either activate or repress transcription. So, in order to

model the promoter correctly, we need to calculate the information of all of the molecules that contribute or interfere.

For activators, as an initial simple model, we assume that their information contributes additively to the total information of the promoter [53], so the new equation is as follows:

$$\begin{aligned}\text{Multi Site Information} = \quad & R_i(-10) + R_i(-35) - GS(d_{-10/-35}) \\ & + \sum(\rho_{Act} R_i(Act) - GS(d_{-35/Act})) \quad \text{(bits/site)}, \end{aligned} \tag{4}$$

where $R_i$(Act) is the individual information of some activator protein site, and $GS(d_{-35/Act})$ is the gap surprisal value at spacing $d$ between the activating protein site and the $-35$. Since the initiation complex is stabilized by the contacts between the $\alpha$CTD and the regulatory protein [12], the information contribution of an activator is in those contacts, and not in the contacts between the activator and the DNA. Data on these protein-protein contacts are not available through DNA sequences. However, the higher the affinity of the activator for the DNA, the greater the probability of it being bound simultaneously with the polymerase. Therefore, to a first approximation, we can model the indirect contribution of information by the activator as a modulation of the activator information by the protein-protein interaction. We represent this modulation by $\rho_{Act}$, with $0 \leq \rho_{Act} \leq 1$. Since the informatics of interactions with the $\alpha$CTD are unknown, we will use a $\rho_{Act} = 1$.

This algorithm only includes the activator site and corresponding gap surprisal if they contribute positive information, since that corresponds to favorable binding [46]. In addition, the number of potential activators is limited only by the length of the sequence. That is, if multiple activators bind in a range relative to the RNA polymerase that has been observed to be advantageous to transcription initiation, then they are all included into the total information for the site. At present the algorithm does not account for the possibility of repression of one activator by another [54].

Although it seems reasonable to assume that activator protein information can be scaled by $\rho_{Act}$ and added to the total information of the promoter, it is not clear that repressor information should be subtracted. Since repressors block the binding of the polymerase to the DNA, or, in cases such as GalR, cause DNA loops that block binding [55], they do not decrease the strength of the contact but, if present, totally prevent contact from occurring. Therefore, it does not matter what the strength of the repressor is, because a repressor bound to a 1 bit site will prevent initiation as well as a repressor bound to a 10 bit site. The difference between the two is that the 10 bit site will be bound more frequently, so the relative site strengths between the polymerase and the repressor (as well as the concentration of both molecules) can be used to predict the frequency of transcription, but not the ability of the polymerase to bind.

### 2.3 Promoter analysis using the $\sigma^{70}$ model

Sequence logos for promoter components were made using the programs **delila, alist, encode, rseq, dalvec** and **makelogo** as previously described [49,51]. We used the programs

**diffinst, genhis** and **genpic** to generate the spacing distribution between the binding components.

During our refinement process, we identified 138 experimentally verified promoters that did not have an upstream −35. We used this subset to build the extended −10 model (Fig. 3). These sites contained a weakly conserved TG two bases upstream of the −10 hexamer. We cyclicly refined [34] these sites over the range −4 to −3 and automatically isolated a subset of 84 sites that turned out to resemble the extended −10 9mer previously reported [13].

To determine spacings between the −10 and the translational initiation codon (Fig. 4), we scanned our promoter model over the 250 bases upstream of all genes in *E. coli* [56]. The site range we used for both the −10 and the −35 was −1 to +4, the range of the hexamer (Fig. 1). We plotted the number of occurrences at each distance between the zero position of the −10 and the translational start point of the strongest upstream promoter. To eliminate interference between the −10 model and the initiation codon, we only included sites in our plot that were at least 4 bases upstream of the gene start. We also plotted the number of occurrences at each distance between the −10s of the experimentally determined sites in our model and their respective downstream translational start.

In order to analyze individual sequences using our $\sigma^{70}$ model and a transcriptional regulator, we used sequence walker technology [46,48,57]. Flexible sites were located using **multiscan** and displayed as sequence walkers using **lister**. For Fig. 5 and Fig. 6, we merely scanned our models over these regions to see how our analysis compared to biochemical data known about these systems. Our Fur model was built from 24 biochemically characterized binding sites (manuscript in preparation) and our Lrp model comes from [34]. Once we were confident of the predictive capabilities of our models, we searched for uncharacterized Fur controlled genes in *E. coli*, two of which are presented in this paper (Fig. 7).

To identify these novel control elements, we scanned the entire genome for Fur sites that overlapped the $\sigma^{70}$ binding sites within 200 bases of translational start points. These two were chosen because of the strength of both the Fur and $\sigma^{70}$ sites, and their proximity to each other. Gel shifts confirmed that these sites are bound by Fur (data not shown).

Finally, the relative binding plots (Fig. 8, circles) were generated by scanning Fis, H-NS and IHF models over the range −1000 to +1000 bases relative to the transcription start, the −10 and the −35 of all promoters in our model. The frequency of sites was determined at each position relative to a promoter component by dividing the number of predicted sites by the number of promoters. The frequency of sites in the genome was determined by scanning the entire genome with all three regulators, and then dividing by the genome size. We only presented data over the range −400 to +200 because the distributions were flat outside of those ranges and matched the frequency of sites in the entire genome. The Fis and IHF models come from previously published works [39,58,41], while our H-NS model has not been published but resembles another published model [59].

To compute the intergenic density distributions (Fig. 8, red curves), we determined the distance from each promoter part (start, −10, and −35) to the closest upstream and down-

stream coding region for all sites in our model. We used the EcoGene12 genome annotation to determine coding regions [56]. We only used promoters whose zero coordinates are not within coding regions. Out of 401 promoters, 359 transcriptional starts, 356 −10s and 349 −35s were not within a coding region. We then counted the number of sequences that had a non-coding base at each spacing upstream and downstream of the promoter. This curve was then normalized and fit to the relative binding plots in Fig. 8, scaled so that the curves are just above the other data points in the graphs, and the zero occurrences level was set equal to the genomic frequency.

# 3 Results

## 3.1 The $\sigma^{70}$ model

Our dataset for $\sigma^{70}$ promoters consisted of both the RegulonDb and the PromEC databases [42,43]. Unlike eukaryotic start points, which contain about 3 bits of information [60], there did not appear to be much information at this prokaryotic transcription start point, only $0.39 \pm 0.06$ bits (Fig. 1) . There does appear to be a slight preference for an adenine as the first base, but it is weak. Interestingly, this preference is more conserved in early promoters of bacteriophage T4 [61] where, presumably, it contributes to the phage taking over the cell. Realignment of the start region by allowing shifts of 1 base in either direction [44] gives a pattern of 1.33 bits (data not shown). Since a choice of 1 in 3 requires $\log_2 3 = 1.58$ bits, this pattern is not significant. ⇐Fig 1

We were quite easily able to align the −10 relative to the transcription start points (Fig. 1). The distance between the −10 and the transcription start point varied between −14 and −8 bases, with the most common spacing of −11 (this is the distance between the transcription start point and the zero position of the −10 logo). The −10 sites contained $4.78 \pm 0.11$ bits of information over the range of −1 to +4, the range of the hexamer.

The most striking feature of the −10 logo is the strongly conserved T (position +4) where the protein is likely to face the minor groove of the DNA. In other logos for DNA binding proteins, conservation of bases rarely exceeds 1 bit in the minor groove [62], because in B-form DNA the exposed groups in the minor groove can only be used to distinguish A or T from C or G, but not all four bases individually [63]. High conservation in the minor groove suggests that this base is being contacted atypically. Several possibilities are that the helix is distorted when bound, it is interacting with $\sigma^{70}$ in the open complex [64], or that it is being flipped out of the helix to initiate open complex formation, as may occur in DNA replication [41].

We had greater difficulty aligning the −35 perhaps because the −35 is often replaced by activators [52,1]. Traditionally, the placement of the −35 relative to the −10 is ±1 base relative to the most frequent position of 23 bases (McClure's 17 [11], see Materials and Methods). Experimental data have shown initiation at a range of −2 to +3 relative to the most frequent position [10,22,19]. When we allowed for our model to include sites in this expanded spacing, we identified 107 promoters (> 0 bits) that had no possible −35

in the traditional range, suggesting that binding does occur at these peripheral spacings. The amount of conservation in the −35, as in the −10, was low compared with other DNA-binding proteins [37] ($4.02 \pm 0.09$ bits). The conserved region of the −35 appears to fill only one-half of the major groove, as there is an abrupt termination of conservation on the 5′ edge (Fig. 1), which is consistent with the abrupt termination of contacts in the $\sigma^{\mathrm{A}}/-35$ co-crystal and the 5′ most edge of the polymerase [65]. ($\sigma^{\mathrm{A}}$ is the primary sigma factor in *Thermus aquaticus*, corresponding to $\sigma^{70}$ in *E. coli* [8].) This suggests that there is space for other binding components to come in and to help stabilize the complex. This is supported by the observation that positive control mutants are immediately adjacent to the portion of $\sigma$ which binds into the −35 groove [65]. As shown by the sine waves in Fig. 1, when we place the G at +1 of the −35 close to the center of the major groove (as observed in the co-crystal [65]), at the optimal spacing of 23 bases, the T at +4 of the −10 is exactly positioned in the minor groove. This agrees with the proposal that position +4 (−7 in conventional numbering) is contacted atypically, for example by base flipping to initiate transcription [41].

We allowed for the spacing range between the two hexamers to be between 21 and 26 bases. The optimal spacing between the zero coordinates of the −35 and −10 was 23 bases. The spacing distribution appeared to be approximately Gaussian with an uncertainty of $2.32 \pm 0.04$ bits.

The total sequence conservation ($R_{sequence}$) for the $\sigma^{70}$ model (−35, gap, −10) is $6.48 \pm 0.14$ bits (Fig. 1). It contains 401 of the 684 sites in our combined RegulonDb-PromEC database. As discussed in Materials and Methods, the number of sites in our dataset was reduced through a series of refinements in order to identify a consistent subset of promoters that presumably can initiate without accessory molecules.

To see how the $\sigma^{70}$ promoter varies with spacing, logos were made for each of the spacing classes (Fig. 2). The logos for the three most frequently used spacings (22, 23, 24) look fairly similar. In these three cases, the bulk of the −35 logo falls in the major groove on the same DNA face as the −10, two helical turns away. The further spacings do look a little different, but this could be from smaller sample sizes. The logo for spacing 25 is the most unique, having what appears to be a distorted −10 (prominent T at position +1 and poorly conserved Ts at −1 and at +4). Also, at spacing 21 the logo is a little different with perhaps a slightly more conserved T (22 cases) instead of G (18 cases) in the third position of the −35 hexamer (position −20). ⇐Fig 2

The conservation of the −35 at each spacing in Fig. 2 appears to follow the sine wave. That is, the conserved bases do not go above the wave, except at position −26 of spacing 25. This suggests that the polymerase preferentially contacts the two components when they are on the same face of the DNA [23], but that the $\sigma$ 4.2 region (which contacts the −35) can rotate relative to the $\sigma$ 2.4 region (which contacts the −10) or that the DNA twist can change. It has been suggested that models be made for each spacing class [32]; this may be useful for sequences outside of the central three spacings. Differences in logos outside of the three similar central spacings (22-24) could be caused by awkward contacts with the polymerase at the extremes of rotation of the −35 relative to the −10.

We looked at the −10 as a function of spacing relative to the transcription start point (data not shown). There was a slight increase in the conservation of the 5′ T of the −10 hexamer at greater spacings. Besides that, there was little variability in the logos of the different spacing classes. The amount of information at the transcription start point was small ($\sim 0.4$ bits) for all spacings, and the slight variability between them could be accounted for by noise.

To avoid duplicate sites, we had excluded 64 promoters from our dataset which were within 15 bases of another transcript start (see Materials and Methods). Having built the model, we went back and scanned it over these regions to see if we could predict promoters upstream of more complex overlapping transcripts (data not shown). In 25 cases, for every experimentally determined transcriptional start point there were one or more distinct predicted promoters. In 17 of the 64 cases there was only one predicted promoter and both transcripts fell within the known distance distribution of 8 to 14 bases downstream from the −10 (Fig. 1). In 10 cases there was only one predicted promoter and only one of the transcripts fell within 8 to 14 bases downstream. 12 of the starts had no predicted promoter upstream, suggesting that these transcripts are regulated. These results confirm that the basal model functions reasonably well on sequences from which it was not constructed.

To further verify our promoter model, we scanned it over the starts of 36 small RNAs presented by Hershberg *et al.* [66] (data not shown). The model identified promoters upstream of 23 of the 36 starts. That is, approximately 64% of the sites had a promoter with a total information $> 0$ bits from 8 to 14 bases (Fig. 1) upstream of the small RNA transcription start. This percentage is similar to that of empirically determined promoters that formed a coherent 'basal' set in our refinement process (59%), suggesting that initiation by the basal machinery may only occur about 60% of the time both in general and at small RNAs. None of these sites had been used to build our model. Therefore, this result again shows that our model can identify the promoters of transcriptional starts that were not included in the model.

The conservation of bases that we observed in our model resembled previous non-information theory based alignments [11], mutation data [10,11,20], an *in vivo* selection assay [67], and the −10 sequence logo previously published for a smaller dataset [41]. The mutation data of Moyle *et al.* [68] had a 0.6 correlation coefficient to the predicted individual information for our complete promoter model (data not shown). These results are consistent with observations by Mirny and Gelfand [69] who demonstrated a good correlation between sequence conservation and the number of base contacts a protein makes with DNA.

Creating a model for the −35 was difficult, presumably because many promoters are activated and the activator could take over the sequence conservation from the −35, as proposed by Raibaud and Schwartz [52]. To test this hypothesis we scanned the 14 sequences in the *E. coli* genome reported to be positively activated by Raibaud and Schwartz and determined which −35 was strongest in the 10 bp window they allowed. In contrast to the $4.0 \pm 0.1$ bits in our −35 model, these −35 sequences in activated promoters were no more than $1 \pm 4$ bits. The weak conservation of positively activated sites probably does explain why creating a −35 model is difficult. To our knowledge this is the only published dataset of confirmed

positively activated *E. coli* promoters.

### 3.2 The extended minus 10

Initiation has been shown to occur in the absence of a −35 in conjunction with an extension to the −10 region [13,14]. During our refinement process, a subset of 138 promoters did not have a predicted −35 binding site, and these were subsequently removed from the flexible basal model. A sequence logo revealed that the removed subset of promoters contained a weakly conserved TG upstream of the −10 hexamer, in the region identified as the extended −10. We therefore did a cyclic refinement of the two bases containing the weakly conserved TG, and a well-conserved extended −10 emerged (Fig. 3). There was no conservation of bases observed outside of the range −4 to +4. Interestingly, the new bases in the −10 appear to follow the sine wave [41], and this region is protected, suggesting that it is bound in the major groove. The $R_{sequence}$ for the extended −10 over the range of −4 to +4 bases is $6.74 \pm 0.25$ bits, almost the same value as the total $R_{sequence}$ for the flexible $\sigma^{70}$ model, which includes the −35 and the gap surprisal. There was also a slight increase in the strength of the −10 hexamer (−1 to +4) from $4.78 \pm 0.11$ bits in the flexible model to $5.05 \pm 0.23$ bits in the extended −10 model. ⇐Fig 3

### 3.3 The relationship between the promoter and the ribosome binding site

In order to determine if there are any spacing preferences between the zero coordinate of the −10 and the translational initiation codon, we scanned our $\sigma^{70}$ model upstream of all 4122 annotated genes in *E. coli* [56]. We saw one substantial peak in the spacing histogram of predicted promoters, around 30 to 40 bases upstream of the ATG (Fig. 4). We also plotted the distance between the −10s of the experimentally verified transcription starts and their corresponding translational start codons, which gave a similar peak around 35 bases (Fig. 4). In both plots, promoters were predicted as far as 200 bases upstream of the translational start. Similar results were obtained by Huerta and Collado-Vides [70]. ⇐Fig 4

For all 401 promoters in our model, we did not see any correlation between promoter strength and the strength of its downstream ribosome binding site as measured by the individual information contents of each flexible model (data not shown). We did find that the lowest combined individual information of a promoter and an RBS was 3.49 bits.

### 3.4 Transcriptional Regulation

We used the $\sigma^{70}$ model in conjunction with transcriptional regulator models to study promoter structures in non-basal conditions. As an example, we show the experimentally verified Fur-controlled gene *tonB* [71] and the degree to which Fur represses it (Fig. 5, manuscript in preparation). The results are conveniently displayed using sequence walkers, which show the individual contribution of each base to a binding site as the height of a letter, with the scale being in bits [46,48,57]. Multiscan found a strong 11.7 bit RNA polymerase site, which was in our model, 8 bases upstream of the experimentally proven transcription start point [72]. This is displayed as two sequence walkers, one for the −35 and one for the −10. To show that they are part of the same promoter, a dashed line is shown connecting the zero ⇐Fig 5

coordinates of each walker. Over this site there were two Fur binding sites (8.7 and 11.5 bits), one of which shows clear sequence competition with the −35 sequence walker (red box). That is, both sequence walkers show positive contributions of bases in the same major groove, so binding by $\sigma^{70}$ and the dimeric Fur protein cannot occur simultaneously. This is a good example of how these models can show not only the affinity of the promoter and its competing repressor, but also the mechanism of transcriptional regulation. The translational start is 43 bases downstream of the −10, close to the optimal spacing indicated by our data (Fig. 4).

As a second example, we used the transcriptional regulator Lrp [40], which can both activate and repress transcription in *E. coli* (Fig. 6). As with Fur at *tonB* (Fig. 5), the sequence walkers readily show that Lrp repression of the *dad* operon [73–75] is probably caused by an occlusion of the promoter by the repressor in the −10 region (Fig. 6A). ⇐Fig 6

Since there are at least seven identified transcription start points for the *dad* operon [74,75], we used our model to see if we could identify the corresponding promoters. The three most downstream starts marked in Fig. 6A (transcripts 4, 5 and 7) presumably use the same promoter (total information 6.3 bits). These starts are 8, 11 and 13 bases from the −10, respectively, which are all reasonable distances between the −10 and transcription start (Fig. 1). The next upstream start (transcript 3) is clearly the result of the binding of a 6.2 bit promoter 10 bases upstream. Transcript 6 may be initiated by a 3.0 bit promoter only 6 bases upstream, and transcript 2 is probably initiated by a 5.6 bit promoter 8 bases upstream (data not shown). Our model did not identify a potential promoter upstream of the start at transcript 1, suggesting that initiation at this point is stabilized by other accessory molecules; likely candidates are CRP (there is a 15.3 bit site 44 bases upstream of transcript 1) and Lrp [75].

Interestingly, the computed strength of the *dad* promoters increases as they get closer to the gene start point but this effect is not observed in *arcA* (data not shown), which also has 7 verified transcript starts [70]. Two bound Lrp molecules (11.1 and 11.7 bits) could block the binding and initiation of the two downstream *dad* promoters and four subsequent downstream transcripts (3, 4, 5, 7), and possibly prevent the opening of transcript 6. The most downstream promoter (transcript 7) is 38 bases away from the translational start point, which produces a transcript only a few bases longer than needed to contain the conserved Shine-Dalgarno region [34], showing an optimization of cellular resources by minimizing the length of mRNAs. We predicted Lrp binding in the region protected by the upstream footprint, but the site was relatively weak at 1.5 bits (data not shown).

Lrp activation of the *gltBDF* [76] operon can also be predicted using sequence walkers (Fig. 6B). In this instance, a strong Lrp binding site (11.2 bits) is just upstream of a strong promoter ($5.3 + 7.6 - 3.3 = 9.6$ bits). Since bits are additive, we suggest that the upstream Lrp site increases the overall affinity of the initiation complex for the promoter, giving a total information that could be as high as 20.8 bits. As mentioned in Materials and Methods, this value is an upper bound on the contribution to promoter binding by Lrp through the $\alpha$CTD.

Based on the two examples in Fig. 6, we propose that Lrp is stabilizing the $\alpha$CTD [12] and

promoting initiation when bound upstream of the $\sigma^{70}$ binding site, but repressing initiation by occlusion when overlapping the $\sigma^{70}$ binding site.

Besides being able to dissect well-understood genetic control systems, we would like to be able to predict new ones for testing. Using a Fur model and the flexible sigma model, we identified a number of potential Fur repressed genes in the *E. coli* genome. We report two of these cases here (Fig. 7). For *yoeA* (Fig. 7A), two strong Fur binding sites (27.4 and 9.0 bits) overlap an average promoter (6.4 bits). The −10 is located 48 bases upstream of the translational start point. A similar result is seen with the *fhuA* gene (Fig. 7B), where two strong Fur sites (10.7 and 19.5 bits, confirmed by gel shift experiments, data not shown) overlap an above-average promoter (7.3 bits) [77,78]. The −10 is located 29 bases upstream of the *fhuA* coding region. ⇐Fig 7

### 3.5 Where DNA bending proteins bind relative to promoter components.

We searched for the relative placement of transcriptional regulators near the starts of all of the promoters in our model (Fig. 8). We present data for models of Fis [39], H-NS [59] and IHF [58,41]. All three proteins are transcriptional activators that are involved in chromosomal compaction [79]. We plotted the number of sites predicted at each position relative to the transcription start point, the −10, and the −35 in order to determine which component the regulators were primarily interacting with. We observed in these graphs that there tends to be a peak in the range of −300 to +100, maximizing at the −35 and the −10 alignments, but not at the transcription start. In all cases, fewer sites were predicted downstream than upstream of the promoter. Although these curves are not necessarily linear, to quantify this we did a linear regression for the regions −400 to 0 and from 0 to +200. The slope of the −400 to 0 line is always smaller than that for 0 to +200. The intersection of these two lines is consistently to the left of the alignment point for alignment by the start base, but close for the −10 and the −35. This is reasonably consistent with the idea that the transcriptional factors tend to cluster around the −35, as might be expected from the $\alpha$CTD contact [53]. Other proteins (Fnr, Fur, LexA, ArgR, CRP, TrpR and LacI) were also analyzed, but because they have higher information the models predicted fewer binding sites, so their graphs were too noisy to interpret. ⇐Fig 8

We determined the range of non-coding sequences surrounding each promoter component in our dataset of 401 promoters (Fig. 1). We counted the number of intergenic regions at various distances from each promoter component. These curves were then normalized to the graphs in Fig. 8 so that they fit just above the data, and so that zero occurrences of intergenic regions was matched to the genomic baseline frequency for each protein (Fig. 8, red curves). The curves matched fairly well, in that the curves downstream of the components are consistently steeper than the curves upstream of the promoter components.

Previous analysis by Robison *et al.* had also identified a preference for genetic control elements to bind in intergenic regions [59], but that analysis was not done in reference to the alignment of promoter components.

# 4 Discussion

Genetic control systems often consist of multiple binding components with variable distances between them. These variable distances can affect the stability of the binding complex. Our approach is to use experimentally demonstrated binding sites to construct models. Unlike neural networks, this approach avoids the assumption that untested stretches of nucleotide sequence do not contain binding sites, and it sets the model upon a firm foundation. Our model construction uses information theory, which not only allows measurements of the patterns at the binding sites, but can also account for distance preferences on the same quantitative and universal scale of bits [37]. We previously used a flexible modelling method for ribosome binding sites [34]. That successful application suggested that information theory can be applied to any multi-part binding system where binding is affected by the spacing between components. In this paper we show that the same approach works well to quantify prokaryotic promoters, which have two binding components at approximately −10 and −35 bases from the start of transcription [11].

## 4.1 A $\sigma^{70}$ model based on information theory

The amount of sequence conservation in the −10 and in the −35 is fairly low, about 5 and 4 bits respectively. As is found for most DNA binding proteins [62,41], both sequence logos appear to follow a sine wave, which represents the 10.6 base helical twist of B-form DNA (Fig. 1). There are unique characteristics to each logo though.

We used the *Thermus aquaticus* $\sigma^{A}/-35$ co-crystal [65] to determine the location of where the $\sigma$ protein faces the major groove with respect to the −35. Using the average gap distance, this assignment places the major groove of the T-A base pair at position +4 of the −10 on exactly the opposite face of the DNA as the −35 (Fig. 1). With this assignment, the well-conserved T at position +4 in the −10 logo exceeds the sine wave. In instances where conservation exceeds 1 bit in the minor groove, DNA distortion or base flipping was proposed [41]. DNA breathes, opening base pairs on a millisecond time scale [80]. By stabilizing this specific flipped-out base, the polymerase could initiate promoter opening. An enzymatic mechanism to initiate this process was proposed by Dubendorff *et al* [81].

In addition, with the exception of +4, the pattern of sequence conservation of the extended −10 follows the sine wave (Fig. 3). This effect has been observed in numerous other sequence logos [62,51], and it can be used to precisely assign the location of protein contacts [41,82], so the extended −10 pattern further suggests that the T at +4 faces the polymerase through the minor groove. Because position −1 has predominantly T instead of equiprobable A and T (Fig. 1), it is unlikely to be bound by a minor groove contact in B-form DNA [62]. This is consistent with our assignment that the major groove side of base −1 faces the polymerase.

Sclavi *et al.* used hydroxy radical footprinting to look at intermediates in open complex formation [83]. They observed that protection at position 0 (−11 in conventional numbering) occurs after protection in the region of +3 to +5 (−8 to −6 in conventional numbering). This is consistent with the T at +4 (-7 in conventional numbering) initiating DNA melting

through a base flipping mechanism [41], which would explain why this position appears anomalous in the sequence logo. In contrast, it has been proposed that flipping of the A at position −11 in the minus ten (our number 0) initiates DNA melting to form the open complex [84–90]. If this is the case, why is our groove assignment 5 bases (180°) different? One possibility is that the DNA helix could be distorted between the −35 and the −10, which our model does not account for. However, Young *et al.* showed that a small part of the $\sigma$ factor and the first 314 amino acids of the $\beta$ subunit are sufficient to initiate promoter melting, which probably excludes DNA bending or twisting [91]. Furthermore, a co-crystal structure of a fork-junction DNA bound to a holoenzyme [7] shows smoothly bent B-form DNA from the −35 to just before the −10. In this structure the extended −10 is contacted in the major groove, consistent with Fig. 3. We conclude that DNA distortions are not sufficient to explain the discrepancy. If base −11 is flipping first, then it may be that −7 is bound specifically after promoter opening, but this mechanism is apparently inconsistent with the hydroxyl radical footprinting data [83]. We have not found a satisfactory explanation for why our clear groove assignment implies that the flipped base is −7 while previous reports suggest −11.

The amount of conservation in the −35 is fairly weak. The clear absence of sequence conservation in the major groove immediately upstream of the −35 could leave room for activating proteins to bind and to stabilize the polymerase (this is supported by the $\sigma^{\mathrm{A}}$/−35 co-crystal structure [65]). By interacting with the polymerase near the −35 contact, accessory molecules could make $\sigma^{70}$ a much more discriminate binder.

Penotti found that the distance between the human TATA sites and the transcriptional start is variable, with an uncertainty of about 3 bits [60]. He also observed that there are about 3 bits of information at the start point itself. In other words, the information of the start ($R_{sequence}$) is just sufficient for it to be located with respect to the TATA ($R_{frequency}$), which is the smallest known example of this evolutionary principle [37,92]. Unlike eukaryotic transcription starts, there is a low conservation of bases at the transcription start point of *E. coli* ($0.39 \pm 0.06$ bits). Because the average gap surprisal between the −10 and the transcriptional start ($2.56 \pm 0.04$ bits) exceeds the information at the start point, we propose that the determination of which base to begin polymerization is influenced more by the detailed path of the RNA through the open complex [7], than by the actual base at the start.

The conventional spacing allowed between the −10 and the −35 only varies by three bases [11], but to account for experimental data [10,17,18], we allowed six bases. Most promoters do fall into the traditional 3 spacing classes, but binding at further spacings seems experimentally and statistically (Fig. 2) reasonable since 24% of the promoters have their strongest −35 outside of the three central spacings. The sequence logos for the most common spacings of 22, 23 and 24 are fairly similar, while those for the outside spacings are a little different. This suggests that at extreme spacings the polymerase may be contacting the promoter differently. At a spacing of 25, the least frequently bound spacing class, the most peculiar −10 logo is seen. The −10 at spacing 25 is unique, with a prominent T at position +1 and poorly conserved Ts at positions +4 and −1. Although this anomaly may simply be an

artifact of the small sample ($n = 25$), this spacing could be conformationally awkward for the polymerase, and this may in itself account for the rarity of promoters with a spacing of 25 bases. It has been shown that at higher superhelical densities, the rate of transcription significantly increased for promoters with this spacing [22], suggesting that activity from this spacing may require helical distortions.

The overall low information content of the entire $\sigma^{70}$ binding site suggests that the RNA polymerase binds frequently along the genome [70,93]. With a total information of $6.48 \pm 0.14$ bits, $\sigma^{70}$ would bind approximately once in every 90 bases in random equiprobable DNA. This would lead to 10 times more transcripts than genes in *E. coli*. The promiscuous nature of the polymerase may be necessary to allow transcription of many different genes in the genome. The polymerase must bind independently of gene function, so it must be indiscriminate enough to bind to a variety of control regions. This suggests that transcription is frequently influenced not only by the strength of the sigma binding site, but also by regulatory molecules. It is also possible that many small RNAs are generated, as has been discovered recently [94,66,93].

In fragments from *E. coli* with lengths of $163 \pm 24$ bp, Kawano *et al.* found 0.76 promoters in one orientation [93]. From this we compute $R_{frequency} = -\log_2(2 \times 0.76/163 \pm 24) = 6.7 \pm 0.2$ bits per site. This is remarkably close to the value for our model, $R_{sequence} = 6.48 \pm 0.14$ bits per site, and it shows that, as with other genetic systems, the information in the binding sites is sufficient to locate the sites in the genome [37,92]. This quantitatively demonstrates that information theory provides a reasonable basal model of polymerase binding.

### 4.2 Evolutionary implications of the extended minus 10

Surprisingly, we were easily able to isolate 84 promoters that lack a $-35$ and exhibit an extended $-10$ [13,14]. The information content of extended $-10$ promoters is almost identical to the information content of the entire $\sigma^{70}$ model (6.7 and 6.5 bits respectively). That is, the information contribution of the $-35$ hexamer in the flexible promoter (Fig. 1) is approximately compensated by a short extension of the $-10$ (Fig. 3). The additional information of the larger conserved region in the $-35$ is restricted by the information penalty of the gap surprisal. (The gap surprisal accounts for how the variable spacing between the $-10$ and $-35$ affects transcription [10,17,18].)

There is a correlation between the amount of conservation within binding sites (the average information content or $R_{sequence}$) and the amount of information needed to locate binding sites in the genome [37,92]. Also, it appears that the information of a site (or group of sites) relates to the energetics of the system [45]. Therefore, since these two promoter classes have a similar information content, we assume that they are equally able to be identified in the genome and to stabilize the polymerase. A single binding element, such as the extended $-10$, is a much simpler machine to evolve than a two-part flexible binder. The bacteriophage T7 RNA polymerase [95] has only one binding element [96], so having two widely separated parts is not essential for transcription. Therefore, we suggest that in prokaryotes the extended $-10$ may be an evolutionary predecessor to the modern bipartite promoter. Another possibility

is that the bipartite promoter is the evolutionary predecessor of the extended −10, but this does not explain the origin of bipartite promoters.

Although they have a similar amount of information, the single-element promoter (Fig. 3) and the more complicated bipartite promoter (Fig. 1) have differences in their sequence conservation. Not only is the conservation of the two upstream bases of the extended −10 lost in the bipartite promoter, but there is also a slight decrease in strength of the part of the extended −10 which corresponds to the bipartite −10 (5.05 bits to 4.78 bits respectively, over the range −1 to +4). The information in the extended −10 that was lost from the bipartite promoter is important for the polymerase to function correctly. That information was apparently reallocated to the −35, so as to produce a promoter that was functionally equivalent to its predecessor. The energy lost by the internal strain of the flexing polymerase is compensated by the additional information of the −35. This is reminiscent of the apparent evolutionary information flow from the exon to the intron sides of both donor and acceptor splice junctions [97].

An advantage of having two widely separated binding components may be to increase promoter strength disparities through interactions with transcriptional regulators. By having a larger binding region, there are more spatial opportunities for accessory proteins to affect the initiation complex. As shown in Fig. 5 and Fig. 6, both the −10 and −35 are targeted by transcriptional repressors, so having two binding elements provides a larger target region within which to evolve repression.

Why would the cell evolve a flexible bipartite binding mechanism? A possible explanation could be that this mechanism allows a polymerase bound to the promoter to sense genomic structure. Indeed, transcriptional initiation has been observed to vary with the superhelicity of the DNA [22,25,98,99]. These differences in the rate of transcription could be from differences in the meltability of the promoter or the stability of the closed complex [11,22]. Also, the spacing between the −10 and −35 is large, two helical turns of DNA, which increases polymerase sensitivity to the overall structure, since twist or bending effects are amplified over larger distances. Twist and bending strain could affect polymerase contacts at both the −10 and −35, as shown in Fig. 2. Therefore, a large flexible bipartite polymerase may have the advantage of higher sensitivity for superhelical regulation, whereas a rigid single-groove binder could be much less sensitive.

The two extra bases of the extended −10 are similar to positions 0 and +1 of the −35 (Fig. 3 and Fig. 1). In both cases there is a T > G next to a G > T. Both T-A and G-C base pairs have an exposed hydrogen acceptor and donor in the major groove (O4-N6 contact for T-A, O6-N4 contact for G-C) [51], suggesting that the polymerase could contact either base pair at these moieties. Preferences for T > G or *vice versa* should depend on the exact positioning of amino acid contacts between the polymerase and the base, since these contacts are in slightly different positions on the base. Indeed, according to models constructed by Barne *et al.* [14], glutamic acids E458 and E585 contact the extended −10 G at −3 and the −35 G at +1, respectively, suggesting a correspondence between these two regions. Furthermore, there is a gap in sequence conservation at position −2 between the −10 hexamer and the

additionally conserved bases of the extended −10 (Fig. 3). Correspondingly, the amino acids that contact the two bases at −1 and −3, T440 and E458, are separated by 17 amino acids [14]. These observations suggest that there are two separable binding elements contacting the −10 region. The similarities between the −35 and the extension of the −10 region suggest that the protein element recognizing the −10 extension was duplicated and then the duplicate merely drifted away from the −10 to form the −35. This is consistent with the structure of the $\sigma$ factor, which has two parts separated by an extended polypeptide [23].

In comparison to the *E. coli* $\sigma^{70}$ and $\sigma^{32}$, both of which appear to contain two helix-turn-helix DNA binding domains, the $\sigma^{55}$ factor produced by bacteriophage T4 for late transcription, contains only one helix-turn-helix motif [1]. Correspondingly, the T4 $\sigma^{55}$ only recognizes a −10 region which contains about 16.2 bits of information [61]. This is close to the 17.6 bits required to locate the 50 known late promoters in the *E. coli* genome [37]. A pared-down RNA polymerase is able to recognize and open an extended −10 [91]. These observations are consistent with the hypothesis that −10 recognition evolved first, followed by appearance of the −35.

### 4.3 The relationship between the promoter and the ribosome binding site

The information content of the flexible ribosome binding site [34] is greater than the $\sigma^{70}$ model, $9.28 \pm 0.06$ versus $6.48 \pm 0.14$ bits, respectively. This most simply suggests that there is often more than one promoter per coding region in the cell as, for example, shown in Fig. 6 and suggested by Huerta and Collado-Vides [70]. In addition, the information in the promoters is lower than that in ribosome binding sites because many promoters rely on activation. A quantitative estimation of these contributions would require detailed knowledge of the number of transcripts and activator binding sites throughout the genome.

When we scanned our promoter model upstream of all annotated genes in *E. coli* [56], our model frequently identified sites at a spacing of about 35 bases between the zero coordinate of the −10 and the first base of the start codon (Fig. 4). The peak represents transcripts starting 11 bases downstream from the −10 (Fig. 1) to produce an approximately 24 bp mRNA leader. Accounting for the 12 to 14 bases of mRNA inside the polymerase [23,100] this leader leaves about 10 to 12 bases exposed. This is just sufficient space to encode for a Shine-Dalgarno (3 bases on the 5′ side), a typical spacer (most commonly 9 bases) and an initiation codon (2 bases on the 3′ side, for a total of 14 bases) [34], and for the ribosome to dock as soon as 0 to 2 more bases have been synthesized [101]. The 35 base spacing allows the earliest possible loading of ribosomes onto the mRNA. There is a gradual decrease in the number of sites upstream of the peak around −35, which suggests a preference for the polymerase to bind close to the translational start.

### 4.4 Transcriptional regulation

Our original dataset of experimentally verified transcription start points was larger than the number of sites in our final model (684 *vs.* 401). The cyclic refinement that removed sites focused the original group down by selecting a subset that is coherent. The excluded sites were weak (information content is less than zero bits) compared to the retained subset and

are therefore presumably activated or use a different sigma factor. Also, as previously noted, the average information of unregulated promoters is fairly low ($6.48 \pm 0.14$), implying that the polymerase binds frequently throughout the genome. These observations are consistent with the role of regulatory proteins to help stabilize weak promoters.

Intergenic regions have a composition that is different from coding regions, and protein binding domains have evolved to bind the intergenic regions [59]. As shown in Fig. 4, the polymerase has a tendency to bind close to the translational start. Regulatory proteins have the same tendency (Fig. 8, circles), but this appears to be because the density of non-coding regions varies relative to the promoter (Fig. 8, red curves). Interestingly, the steepness of the density curve is greater downstream of the promoter than upstream. This suggests that in each intergenic region, RNA polymerase binding sites tend to be located where they will maximize the amount of upstream regulatory DNA, while minimizing the length of transcribed mRNA. These effects would preserve cellular resources.

The $-35$ is the most upstream component of the promoter, and it is closest to the $\alpha$CTD, to which activator proteins bind [53,12]. This explains why the alignments shown in Fig. 8 are best matched by the $-35$ and $-10$ in contrast to the start point of the flexible promoter.

Without DNA bending, activators more than 20 bases upstream would have difficulty binding to the $\alpha$CTD [53], because at distances shorter than the persistence length (150 to 200 bp) DNA is like a rigid rod [102]. Furthermore, pairs of DNA sites come together most easily when they are about 300 bp apart [103,104]. These physical limits mean that activators would be restricted to be either immediately upstream of the polymerase, as previously noted [105], or at least 300 bases away. DNA bending proteins and curved DNA, which is in the intergenic regions [79,106], loosen these restrictions and allow activators bound within 300 bases upstream of the promoter to function. This explains why Fis, H-NS and IHF are often found within 300 bases of the promoter (Fig. 8), as previously noted by Ussery *et al.* [79]. Given that DNA bending proteins exist, intergenic regions could evolve to be smaller than 300 bases and still allow activation by proteins further than a few bases upstream of the promoter. This leads to another cellular evolutionary conservation principle in which the cell maximizes potential activator access while minimizing total genome length. Thus the intergenic region distribution follows from the distributions of the DNA bending proteins, and the DNA bending protein locations are in turn a consequence of the persistence length of DNA.

The number of Fis dimers in the cell has been shown to drastically increase in response to nutritional upshifts [107]. If a major role of Fis in the cell is to facilitate activation through DNA-bending within the persistence length, then these results show how Fis can act as a powerful global regulator, linking transcription to cellular nutrition [39,108,79].

Our individual information analysis (Fig. 5, Fig. 6, and Fig. 7) confirms that the mechanism of transcriptional control can be predicted by the spatial position of transcription factors [105]. Sequence walkers reveal that proteins whose positions share information with the promoter will probably bind antagonistically, while those which do not interfere with the $-35$ or $-10$ may be activators. Quantification of promoter strengths, however, is a difficult

task because of the number of components that can be involved in the stabilization, or occlusion, of the initiation complex. Our approach is simple, clear, powerful, and useful in the design and analysis of experiments aimed at understanding genetic control mechanisms.

## 5 Acknowledgements

We would like to thank Dmitry Vassylyev for providing atomic coordinates for the closed promoter and Heladia Salgado for proving us with the RegulonDb database. We would also like to thank Brent Jewett, Danielle Needle, Michael Levashov, Aidan Ryan and Pete Rogan for their comments. This research was supported in part by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research.

## References


[1] Horwitz, M. S. and Loeb, L. A. (1990) Structure-function relationships in *Escherichia coli* promoter DNA. *Prog Nucleic Acid Res Mol Biol,* **38**, 137–164.

[2] Gralla, J. D. (1990) Promoter recognition and mRNA initiation by *Escherichia coli* $E\sigma^{70}$. *Methods Enzymol,* **185**, 37–54.

[3] deHaseth, P. L., Zupancic, M. L., and Record Jr, M. T. (1998) RNA polymerase-promoter interactions: the comings and goings of RNA polymerase. *J. Bacteriol.,* **180**, 3019–3025.

[4] Browning, D. F. and Busby, S. J. (2004) The regulation of bacterial transcription initiation. *Nat Rev Microbiol,* **2**, 57–65.

[5] Burgess, R. R., Travers, A. A., Dunn, J. J., and Bautz, E. K. (1969) Factor stimulating transcription by RNA polymerase. *Nature,* **221**, 43–46.

[6] Gross, C. A., Chan, C., Dombroski, A., Gruber, T., Sharp, M., Tupy, J., and Young, B. (1998) The functional and regulatory roles of sigma factors in transcription. *Cold Spring Harb Symp Quant Biol,* **63**, 141–155.

[7] Murakami, K. S., Masuda, S., Campbell, E. A., Muzzin, O., and Darst, S. A. (2002) Structural basis of transcription initiation: an RNA polymerase holoenzyme-DNA complex. *Science,* **296**, 1285–1290.

[8] Young, B. A., Gruber, T. M., and Gross, C. A. (2002) Views of transcription initiation. *Cell,* **109**, 417–420.

[9] Gross, C., Lonetto, M., and Losick, R. (1992) Bacterial sigma factors. In McKnight, S. L. and Yamamoto, K. R., (eds.), *Transcriptional Regulation*, New York: Cold Spring Harbor Laboratory Press Vol. I, pp. 129–176.

[10] Hawley, D. K. and McClure, W. R. (1983) Compilation and analysis of *Escherichia coli* promoter DNA sequences. *Nucleic Acids Res.,* **11**, 2237–2255.

[11] McClure, W. R. (1985) Mechanism and control of transcription initiation in prokaryotes. *Annu. Rev. Biochem.,* **54**, 171–204.

[12] Benoff, B., Yang, H., Lawson, C. L., Parkinson, G., Liu, J., Blatter, E., Ebright, Y. W., Berman, H. M., and Ebright, R. H. (2002) Structural Basis of Transcription Activation: The CAP-$\alpha$CTD-DNA Complex. *Science,* **297**, 1562–1566.

[13] Keilty, S. and Rosenberg, M. (1987) Constitutive function of a positively regulated promoter reveals new sequences essential for activity. *J. Biol. Chem.,* **262**, 6389–6395.

[14] Barne, K. A., Bown, J. A., Busby, S. J., and Minchin, S. D. (1997) Region 2.5 of the *Escherichia coli* RNA polymerase $\sigma^{70}$ subunit is responsible for the recognition of the 'extended -10' motif at promoters. *EMBO J.,* **16**, 4034–4040.

[15] Kumar, A., Malloch, R. A., Fujita, N., Smillie, D. A., Ishihama, A., and Hayward, R. S. (1993) The minus 35-recognition region of *Escherichia coli* sigma 70 is inessential for initiation of transcription at an "extended minus 10" promoter. *J. Mol. Biol.,* **232**, 406–418.

[16] Eichenberger, P., Dethiollaz, S., Buc, H., and Geiselmann, J. (1997) Structural kinetics of transcription activation at the *malT* promoter of *Escherichia coli* by UV laser footprinting. *Proc. Natl. Acad. Sci. USA,* **94**, 9022–9027.

[17] Mandecki, W. and Reznikoff, W. S. (1982) A lac promotor with a changed distance between -10 and -35 regions. *Nucleic Acids Res.,* **10**, 903–912.

[18] Aoyama, T., Takanami, M., Ohtsuka, E., Taniyama, Y., Marumoto, R., Sato, H., and Ikehara, M. (1983) Essential structure of *E. coli* promoter: effect of spacer length between the two consensus sequences on promoter function. *Nucleic Acids Res.,* **11**, 5855–5864.

[19] Dombroski, A. J., Johnson, B. D., Lonetto, M., and Gross, C. A. (1996) The sigma subunit of *Escherichia coli* RNA polymerase senses promoter spacing. *Proc. Natl. Acad. Sci. USA,* **93**, 8858–8862.

[20] Stefano, J. E. and Gralla, J. D. (1982) Mutation-induced changes in RNA polymerase-lac $p^s$ promoter interactions. *J. Biol. Chem.,* **257**, 13924–13929.

[21] Borowiec, J. A. and Gralla, J. D. (1987) All three elements of the *lac* $p^S$ promoter mediate its transcriptional response to DNA supercoiling. *J. Mol. Biol.,* **195**, 89–97.

[22] Aoyama, T. and Takanami, M. (1988) Supercoiling response of *E. coli* promoters with different spacer lengths. *Biochim Biophys Acta,* **949**, 311–317.

[23] Vassylyev, D. G., Sekine, S., Laptenko, O., Lee, J., Vassylyeva, M. N., Borukhov, S., and Yokoyama, S. (2002) Crystal structure of a bacterial RNA polymerase holoenzyme at 2.6 Å resolution. *Nature,* **417**, 712–719.

[24] Lim, H. M., Lewis, D. E., Lee, H. J., Liu, M., and Adhya, S. (2003) Effect of varying the supercoiling of DNA on transcription and its regulation. *Biochemistry,* **42**, 10718–10725.

[25] Peter, B. J., Arsuaga, J., Breier, A. M., Khodursky, A. B., Brown, P. O., and Cozzarelli, N. R. (2004) Genomic transcriptional response to loss of chromosomal supercoiling in *Escherichia coli*. *Genome Biol,* **5**, R87.

[26] Lukashin, A. V., Anshelevich, V. V., Amirikyan, B. R., Gragerov, A. I., and Frank-Kamenetskii, M. D. (1989) Neural network models for promoter recognition. *J. Biomol. Struct. Dyn.,* **6**, 1123–1133.

[27] Weller, K. and Recknagel, R. D. (1994) Promoter strength prediction based on occurrence frequencies of consensus patterns. *J Theor Biol,* **171**, 355–359.

[28] GuhaThakurta, D. and Stormo, G. D. (2001) Identifying target sites for cooperatively binding factors. *Bioinformatics,* **17**, 608–621.

[29] Galas, D. J., Eggert, M., and Waterman, M. S. (1985) Rigorous pattern-recognition methods for DNA sequences. Analysis of promoter sequences from Escherichia coli. *J. Mol. Biol.,* **186**, 117–128.

[30] Mulligan, M. E., Hawley, D. K., Entriken, R., and McClure, W. R. (1984) *Escherichia coli* promoter sequences predict *in vitro* RNA polymerase selectivity. *Nucleic Acids Res.,* **12**, 789–800.

[31] Hertz, G. Z. and Stormo, G. D. (1996) *Escherichia coli* promoter sequences: Analysis and prediction. *Methods Enzymol,* **273**, 30–42.

[32] O'Neill, M. C. (1989) Consensus methods for finding and ranking DNA binding sites: Application to *Escherichia coli* promoters. *J. Mol. Biol.,* **207**, 301–310.

[33] Harley, C. B. and Reynolds, R. P. (1987) Analysis of *E. coli* promoter sequences. *Nucleic Acids Res.,* **15**, 2343–2361.

[34] Shultzaberger, R. K., Bucheimer, R. E., Rudd, K. E., and Schneider, T. D. (2001) Anatomy of *Escherichia coli* Ribosome Binding Sites. *J. Mol. Biol.,* **313**, 215–228 http://www.ccrnp.ncifcrf.gov/˜toms/paper/flexrbs/.

[35] Shannon, C. E. (1948) A Mathematical Theory of Communication. *Bell System Tech. J.,* **27**, 379–423, 623–656 http://cm.bell-labs.com/cm/ms/what/shannonday/paper.html.

[36] Pierce, J. R. (1980) An Introduction to Information Theory: Symbols, Signals and Noise, Dover Publications, Inc., New York second edition.

[37] Schneider, T. D., Stormo, G. D., Gold, L., and Ehrenfeucht, A. (1986) Information content of binding sites on nucleotide sequences. *J. Mol. Biol.,* **188**, 415–431 http://www.ccrnp.ncifcrf.gov/˜toms/paper/schneider1986/.

[38] Rogan, P. K., Faux, B. M., and Schneider, T. D. (1998) Information analysis of human splice site mutations. *Human Mutation,* **12**, 153–171 Erratum in: Hum Mutat 1999;13(1):82. http://www.ccrnp.ncifcrf.gov/˜toms/paper/rfs/.

[39] Hengen, P. N., Bartram, S. L., Stewart, L. E., and Schneider, T. D. (1997) Information analysis of Fis binding sites. *Nucleic Acids Res.,* **25**(24), 4994–5002 http://www.ccrnp.ncifcrf.gov/˜toms/paper/fisinfo/.

[40] Shultzaberger, R. K. and Schneider, T. D. (1999) Using sequence logos and information analysis of Lrp DNA binding sites to investigate discrepancies between natural selection and SELEX. *Nucleic Acids Res.,* **27**(3), 882–887 http://www.ccrnp.ncifcrf.gov/˜toms/paper/lrp/.

[41] Schneider, T. D. (2001) Strong minor groove base conservation in sequence logos implies DNA distortion or base flipping during replication and transcription initiation. *Nucleic Acids Res.,* **29**(23), 4881–4891 http://www.ccrnp.ncifcrf.gov/˜toms/paper/baseflip/.

[42] Salgado, H., Santos-Zavaleta, A., Gama-Castro, S., Millan-Zarate, D., Diaz-Peredo, E., Sanchez-Solano, F., Perez-Rueda, E., Bonavides-Martinez, C., and Collado-Vides, J. (2001) RegulonDB (version 3.2): transcriptional regulation and operon organization in *Escherichia coli* K-12. *Nucleic Acids Res.,* **29**, 72–74.

[43] Hershberg, R., Bejerano, G., Santos-Zavaleta, A., and Margalit, H. (2001) PromEC: An updated database of Escherichia coli mRNA promoters with experimentally identified transcriptional start sites. *Nucleic Acids Res.,* **29**, 277.

[44] Schneider, T. D. and Mastronarde, D. (1996) Fast multiple alignment of ungapped DNA sequences using information theory and a relaxation method. *Discrete Applied Mathematics,* **71**, 259–268 http://www.ccrnp.ncifcrf.gov/˜toms/paper/malign.

[45] Schneider, T. D. (1991) Theory of molecular machines. II. Energy dissipation from molecular machines. *J. Theor. Biol.,* **148**, 125–137 http://www.ccrnp.ncifcrf.gov/~toms/paper/edmm/.

[46] Schneider, T. D. (1997) Information content of individual genetic sequences. *J. Theor. Biol.,* **189**(4), 427–441 http://www.ccrnp.ncifcrf.gov/~toms/paper/ri/.

[47] Rogan, P. K., Svojanovsky, S., and Leeder, J. S. (2003) Information theory-based analysis of CYP2C19, CYP2D6 and CYP3A5 splicing mutations. *Pharmacogenetics,* **13**, 207–218.

[48] Schneider, T. D. (1997) Sequence walkers: a graphical method to display how binding proteins interact with DNA or RNA sequences. *Nucleic Acids Res.,* **25**, 4408–4415 http://www.ccrnp.ncifcrf.gov/~toms/paper/walker/, erratum: NAR 26(4): 1135, 1998.

[49] Schneider, T. D. and Stephens, R. M. (1990) Sequence logos: A new way to display consensus sequences. *Nucleic Acids Res.,* **18**, 6097–6100 http://www.ccrnp.ncifcrf.gov/~toms/paper/logopaper/.

[50] Schneider, T. D. (2002) Consensus Sequence Zen. *Applied Bioinformatics,* **1**(3), 111–119 http://www.ccrnp.ncifcrf.gov/~toms/papers/zen/.

[51] Schneider, T. D. (1996) Reading of DNA sequence logos: Prediction of major groove binding by information theory. *Meth. Enzym.,* **274**, 445–455 http://www.ccrnp.ncifcrf.gov/~toms/paper/oxyr/.

[52] Raibaud, O. and Schwartz, M. (1984) Positive control of transcription initiation in bacteria. *Annu Rev Genet,* **18**, 173–206.

[53] Busby, S. and Ebright, R. H. (1999) Transcription activation by catabolite activator protein (CAP). *J. Mol. Biol.,* **293**, 199–213.

[54] Hengen, P. N., Lyakhov, I. G., Stewart, L. E., and Schneider, T. D. (2003) Molecular flip-flops formed by overlapping Fis sites. *Nucleic Acids Res.,* **31**(22), 6663–6673.

[55] Semsey, S., Virnik, K., and Adhya, S. (2006) Three-stage regulation of the amphibolic *gal* operon: from repressosome to GalR-free DNA. *J. Mol. Biol.,* **358**, 355–363.

[56] Rudd, K. E. (2000) EcoGene: a genome sequence database for *Escherichia coli* K-12. *Nucleic Acids Res.,* **28**, 60–64.

[57] Schneider, T. D. and Rogan, P. K. Computational analysis of nucleic acid information defines binding sites, United States Patent 5867402. (1999).

[58] Goodrich, J. A., Schwartz, M. L., and McClure, W. R. (1990) Searching for and predicting the activity of sites for DNA binding proteins: compilation and analysis of the binding sites for *Escherichia coli* integration host factor (IHF). *Nucleic Acids Res.,* **18**, 4993–5000.

[59] Robison, K., McGuire, A. M., and Church, G. M. (1998) A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete *Escherichia coli* K-12 genome. *J. Mol. Biol.,* **284**, 241–254.

[60] Penotti, F. E. (1990) Human DNA TATA boxes and transcription initiation sites: A statistical study. *J. Mol. Biol.,* **213**, 37–52.

[61] Miller, E. S., Kutter, E., Mosig, G., Arisaka, F., Kunisawa, T., and Ruger, W. (2003) Bacteriophage T4 genome. *Microbiol Mol Biol Rev,* **67**, 86–156.

[62] Papp, P. P., Chattoraj, D. K., and Schneider, T. D. (1993) Information analysis of sequences that bind the replication initiator RepA. *J. Mol. Biol.,* **233**, 219–230.

[63] Seeman, N. C., Rosenberg, J. M., and Rich, A. (1976) Sequence-specific recognition of double helical nucleic acids by proteins. *Proc. Natl. Acad. Sci. USA,* **73**, 804–808.

[64] Brodolin, K., Zenkin, N., and Severinov, K. (2005) Remodeling of the sigma70 subunit non-template DNA strand contacts during the final step of transcription initiation. *J. Mol. Biol.,* **350**, 930–937.

[65] Campbell, E. A., Muzzin, O., Chlenov, M., Sun, J. L., Olson, C. A., Weinman, O., Trester-Zedlitz, M. L., and Darst, S. A. (2002) Structure of the bacterial RNA polymerase promoter specificity $\sigma$ subunit. *Mol Cell,* **9**, 527–539.

[66] Hershberg, R., Altuvia, S., and Margalit, H. (2003) A survey of small RNA-encoding genes in *Escherichia coli*. *Nucleic Acids Res.,* **31**, 1813–1820.

[67] Oliphant, A. R. and Struhl, K. (1988) Defining the consensus sequences of *E.coli* promoter elements by random selection. *Nucleic Acids Res.,* **16**, 7673–7683.

[68] Moyle, H., Waldburger, C., and Susskind, M. M. (1991) Hierarchies of base pair preferences in the P22 *ant* promoter. *J. Bacteriol.,* **173**, 1944–1950.

[69] Mirny, L. A. and Gelfand, M. S. (2002) Structural analysis of conserved base pairs in protein-DNA complexes. *Nucleic Acids Res.,* **30**, 1704–1711.

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

[71] Young, G. M. and Postle, K. (1994) Repression of *tonB* transcription during anaerobic growth requires Fur binding at the promoter and a second factor binding upstream. *Mol. Microbiol.,* **11**, 943–954.

[72] Postle, K. and Good, R. F. (1983) DNA sequence of the *Escherichia coli tonB* gene. *Proc. Natl. Acad. Sci. USA,* **80**, 5235–5239.

[73] Zhi, J., Mathew, E., and Freundlich, M. (1999) Lrp binds to two regions in the *dadAX* promoter region of *Escherichia coli* to repress and activate transcription directly. *Mol. Microbiol.,* **32**, 29–40.

[74] Mathew, E., Zhi, J., and Freundlich, M. (1996) Lrp is a direct repressor of the *dad* operon in *Escherichia coli*. *J. Bacteriol.,* **178**, 7234–7240.

[75] Zhi, J., Mathew, E., and Freundlich, M. (1998) In vitro and in vivo characterization of three major *dadAX* promoters in *Escherichia coli* that are regulated by cyclic AMP-CRP and Lrp. *Mol Gen Genet,* **258**, 442–447.

[76] Wiese II, D. E., Ernsting, B. R., Blumenthal, R. M., and Matthews, R. G. (1997) A nucleoprotein activation complex between the leucine-responsive regulatory protein and DNA upstream of the *gltBDF* operon in *Escherichia coli*. *J. Mol. Biol.,* **270**, 152–168.

[77] Tsolis, R. M., Baumler, A. J., Stojiljkovic, I., and Heffron, F. (1995) Fur regulon of *Salmonella typhimurium*: identification of new iron-regulated genes. *J. Bacteriol.,* **177**, 4628–4637.

[78] Masse, E., Vanderpool, C. K., and Gottesman, S. (2005) Effect of RyhB small RNA on global iron use in *Escherichia coli*. *J. Bacteriol.,* **187**, 6962–6971.

[79] Ussery, D., Larsen, T. S., Wilkes, K. T., Friis, C., Worning, P., Krogh, A., and Brunak, S. (2001) Genome organisation and chromatin structure in *Escherichia coli*. *Biochimie,* **83**, 201–212.

[80] Leroy, J. L., Kochoyan, M., Huynh-Dinh, T., and Guéron, M. (1988) Characterization of base-pair opening in deoxynucleotide duplexes using catalyzed exchange of the imino proton. *J. Mol. Biol.,* **200**, 223–238.

[81] Dubendorff, J. W., deHaseth, P. L., Rosendahl, M. S., and Caruthers, M. H. (1987) DNA functional groups required for formation of open complexes between *Escherichia coli* RNA polymerase and the $\lambda$ $P_R$ promoter. Identification via base analog substitutions. *J. Biol. Chem.,* **262**, 892–898.

[82] Lyakhov, I. G., Hengen, P. N., Rubens, D., and Schneider, T. D. (2001) The P1 Phage Replication Protein RepA Contacts an Otherwise Inaccessible Thymine N3 Proton by DNA Distortion or Base Flipping. *Nucleic Acids Res.,* **29**(23), 4892–4900 http://www.ccrnp.ncifcrf.gov/˜toms/paper/repan3/.

[83] Sclavi, B., Zaychikov, E., Rogozina, A., Walther, F., Buckle, M., and Heumann, H. (2005) Real-time characterization of intermediates in the pathway to open complex formation by Escherichia coli RNA polymerase at the T7A1 promoter. *Proc. Natl. Acad. Sci. USA,* **102**, 4706–4711.

[84] Helmann, J. D. and deHaseth, P. L. (1999) Protein-nucleic acid interactions during open complex formation investigated by systematic alteration of the protein and DNA binding partners. *Biochemistry,* **38**, 5959–5967.

[85] Fenton, M. S., Lee, S. J., and Gralla, J. D. (2000) *Escherichia coli* promoter opening and -10 recognition: mutational analysis of $\sigma$70. *EMBO J,* **19**, 1130–1137.

[86] Lim, H. M., Lee, H. J., Roy, S., and Adhya, S. (2001) A "master" in base unpairing during isomerization of a promoter upon RNA polymerase binding. *Proc. Natl. Acad. Sci. USA,* **98**, 14849–14852.

[87] deHaseth, P. L. and Tsujikawa, L. (2003) Probing the role of region 2 of *Escherichia coli* $\sigma^{70}$ in nucleation and maintenance of the single-stranded DNA bubble in RNA polymerase-promoter open complexes. *Methods Enzymol,* **370**, 553–567.

[88] Roy, S., Lim, H. M., Liu, M., and Adhya, S. (2004) Asynchronous basepair openings in transcription initiation: CRP enhances the rate-limiting step. *EMBO J,* **23**, 869–875.

[89] Lee, H. J., Lim, H. M., and Adhya, S. (2004) An unsubstituted C2 hydrogen of adenine is critical and sufficient at the -11 position of a promoter to signal base pair deformation. *J. Biol. Chem.,* **279**, 16899–16902.

[90] Heyduk, E., Kuznedelov, K., Severinov, K., and Heyduk, T. (2006) A consensus adenine at position −11 of the nontemplate strand of bacterial promoter is important for nucleation of promoter melting. *J. Biol. Chem.,* **281**, 12362–12369.

[91] Young, B. A., Gruber, T. M., and Gross, C. A. (2004) Minimal machinery of RNA polymerase holoenzyme sufficient for promoter melting. *Science,* **303**, 1382–1384.

[92] Schneider, T. D. (2000) Evolution of biological information. *Nucleic Acids Res.,* **28**(14), 2794–2799 http://www.ccrnp.ncifcrf.gov/˜toms/paper/ev/.

[93] Kawano, M., Storz, G., Rao, B. S., Rosner, J. L., and Martin, R. G. (2005) Detection of low-level promoter activity within open reading frame sequences of *Escherichia coli*. *Nucleic Acids Res.,* **33**, 6268–6276.

[94] Wassarman, K. M., Zhang, A., and Storz, G. (1999) Small RNAs in *Escherichia coli*. *Trends Microbiol,* **7**, 37–45.

[95] Yin, Y. W. and Steitz, T. A. (2002) Structural basis for the transition from initiation to elongation transcription in T7 RNA polymerase. *Science,* **298**, 1387–1395.

[96] Schneider, T. D. and Stormo, G. D. (1989) Excess information at bacteriophage T7 genomic promoters detected by a random cloning technique. *Nucleic Acids Res.,* **17**, 659–674.

[97] Stephens, R. M. and Schneider, T. D. (1992) Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites. *J. Mol. Biol.,* **228**, 1124–1136 http://www.ccrnp.ncifcrf.gov/˜toms/paper/splice/.

[98] Travers, A. and Muskhelishvili, G. (2005) DNA supercoiling - a global transcriptional regulator for enterobacterial growth?. *Nat Rev Microbiol,* **3**, 157–169.

[99] Chen, Y. C. and Jeng, S. T. (2000) Binding affinity of T7 RNA polymerase to its promoter in the supercoiled and linearized DNA templates. *Biosci Biotechnol Biochem,* **64**, 1126–1132.

[100] Darst, S. A. (2001) Bacterial RNA polymerase. *Curr Opin Struct Biol,* **11**, 155–162.

[101] Rudd, K. E. and Schneider, T. D. (1992) Compilation of *E. coli* ribosome binding sites. In Miller, J. H., (ed.), *A Short Course in Bacterial Genetics: A Laboratory Manual and Handbook for* Escherichia coli *and Related Bacteria*, Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press pp. 17.19–17.45.

[102] Hagerman, P. J. (1988) Flexibility of DNA. *Annu Rev Biophys Biophys Chem,* **17**, 265–286.

[103] Halford, S. E. and Szczelkun, M. D. (2002) How to get from A to B: strategies for analysing protein motion on DNA. *Eur Biophys J,* **31**, 257–267.

[104] Ringrose, L., Chabanis, S., Angrand, P. O., Woodroofe, C., and Stewart, A. F. (1999) Quantitative comparison of DNA looping *in vitro* and *in vivo*: chromatin increases effective DNA flexibility at short distances. *EMBO J,* **18**, 6630–6641.

[105] Collado-Vides, J., Magasanik, B., and Gralla, J. D. (1991) Control site location and transcriptional regulation in *Escherichia coli*. *Microbiol. Rev.,* **55**, 371–394.

[106] Bolshoy, A. and Nevo, E. (2000) Ecologic genomics of DNA: upstream bending in prokaryotic promoters. *Genome Res,* **10**, 1185–1193.

[107] Ball, C. A., Osuna, R., Ferguson, K. C., and Johnson, R. C. (1992) Dramatic changes in Fis levels upon nutrient upshift in *Escherichia coli*. *J. Bacteriol.,* **174**, 8043–8056.

[108] Travers, A., Schneider, R., and Muskhelishvili, G. (2001) DNA supercoiling and transcription in Escherichia coli: The FIS connection. *Biochimie,* **83**, 213–217.

[109] Siebenlist, U., Simpson, R. B., and Gilbert, W. (1980) E. coli RNA polymerase interacts homologously with two different promoters. *Cell,* **20**, 269–281.

[110] Roberts, C. W. and Roberts, J. W. (1996) Base-specific recognition of the nontemplate strand of promoter DNA by E. coli RNA polymerase. *Cell,* **86**, 495–501.

[111] Althaus, E. W., Outten, C. E., Olson, K. E., Cao, H., and O'Halloran, T. V. (1999) The ferric uptake regulation (Fur) repressor is a zinc metalloprotein. *Biochemistry,* **38**, 6559–6569.

[112] Chen, Z. and Schneider, T. D. (2006) Comparative analysis of tandem T7-like promoter containing regions in enterobacterial genomes reveals a novel group of genetic islands. *Nucleic Acids Res.,* **34**, 1133–1147 http://www.ccrnp.ncifcrf.gov/~toms/papers/t7island/.

[113] Blattner, F. R., Plunkett III, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., Gregor, J., Davis, N. W., Kirkpatrick, H. A., Goeden, M. A., Rose, D. J., Mau, B., and Shao, Y. (1997) The complete genome sequence of *Escherichia coli* K-12. *Science,* **277**, 1453–1474.

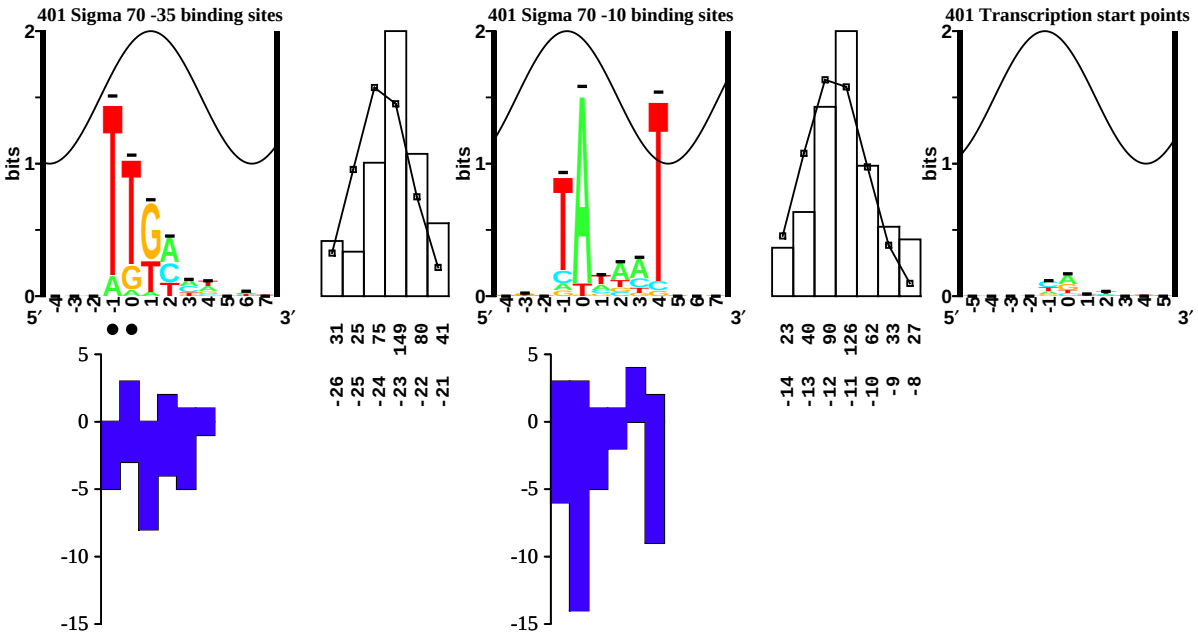


Fig. 1. Sequence logos of $\sigma^{70}$ binding components.
From left to right: sequence logo of −35 binding sites, spacing distribution between −35 and −10 binding sites, sequence logo of −10 binding sites, spacing distribution between −10 binding sites and the transcription start point, sequence logo of transcription start points. In a logo, the height of each letter is proportional to the frequency of that base at each position, and the height of the letter stack is the conservation in bits [49]. Error bars are shown at the top of the stacks. The total information in the −35 and −10, less the gap uncertainty between them, is $(4.02 \pm 0.09) + (4.78 \pm 0.11) - (2.32 \pm 0.04) = 6.48 \pm 0.14$ bits. The sine wave on each logo represents the 10.6 base helical twist of B-form DNA for the optimal spacing of 23 bases, with the major groove centered at +1 of the −35 [51,41]. Black dots indicate the location of important 5-methyl groups on thymine and hence determine the location where the major groove faces the sigma factor [81], along with co-crystal data [65]. The top row of numbers in each gap distribution gives the number of cases and the bottom row is the difference between the zero coordinates. A Gaussian curve was fit to each of the two gap distributions (thin black line). Mutational data presented by Hawley and McClure [10,11] are shown under the logos by blue bars. Bars above the abscissa represent the number of observed mutations at each position that have strengthened a promoter, while bars below the abscissa represent the number of mutations that have weakened a promoter. Promoter locations and the information contents of their parts are given in Supplementary Data.

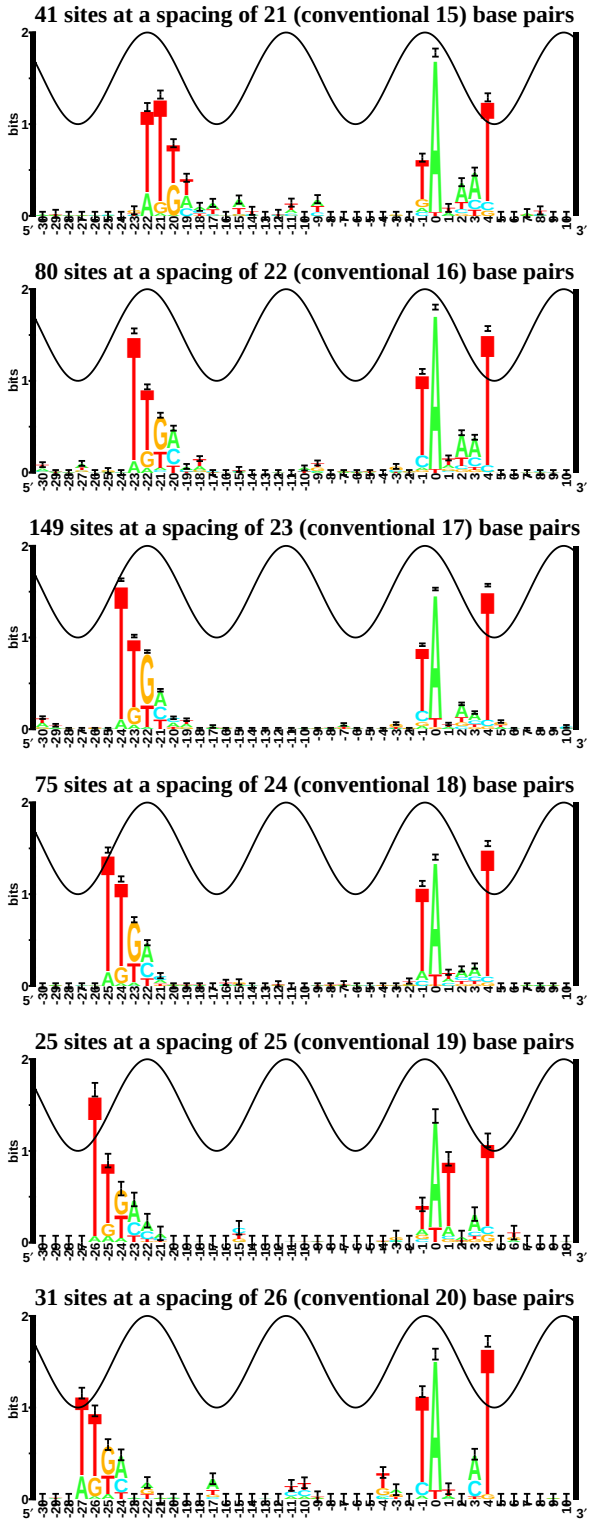


Fig. 2. Sequence logos for $\sigma^{70}$ promoters as a function of spacing.
The spacings correspond to the $-35$ to $-10$ gap distribution in Fig. 1. Conventional spacings are given in parentheses.

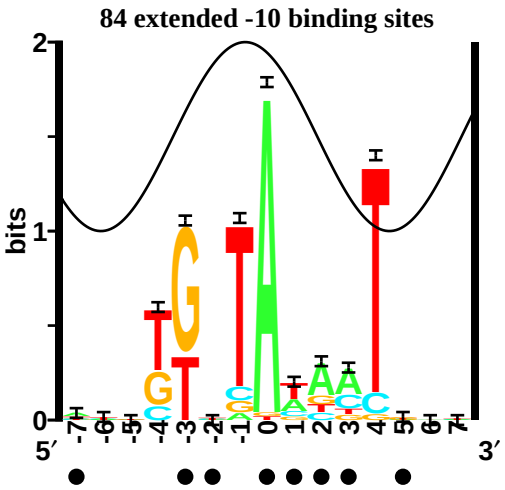


Fig. 3. The extended −10 has two additionally conserved bases.
This is a sequence logo of −10 regions that have no −35 based on our model, but show conservation at positions −4 and −3. Purines protected from DMS methylation and bromouracil substituted thymines protected by the polymerase are indicated by closed circles (•) [109,110].

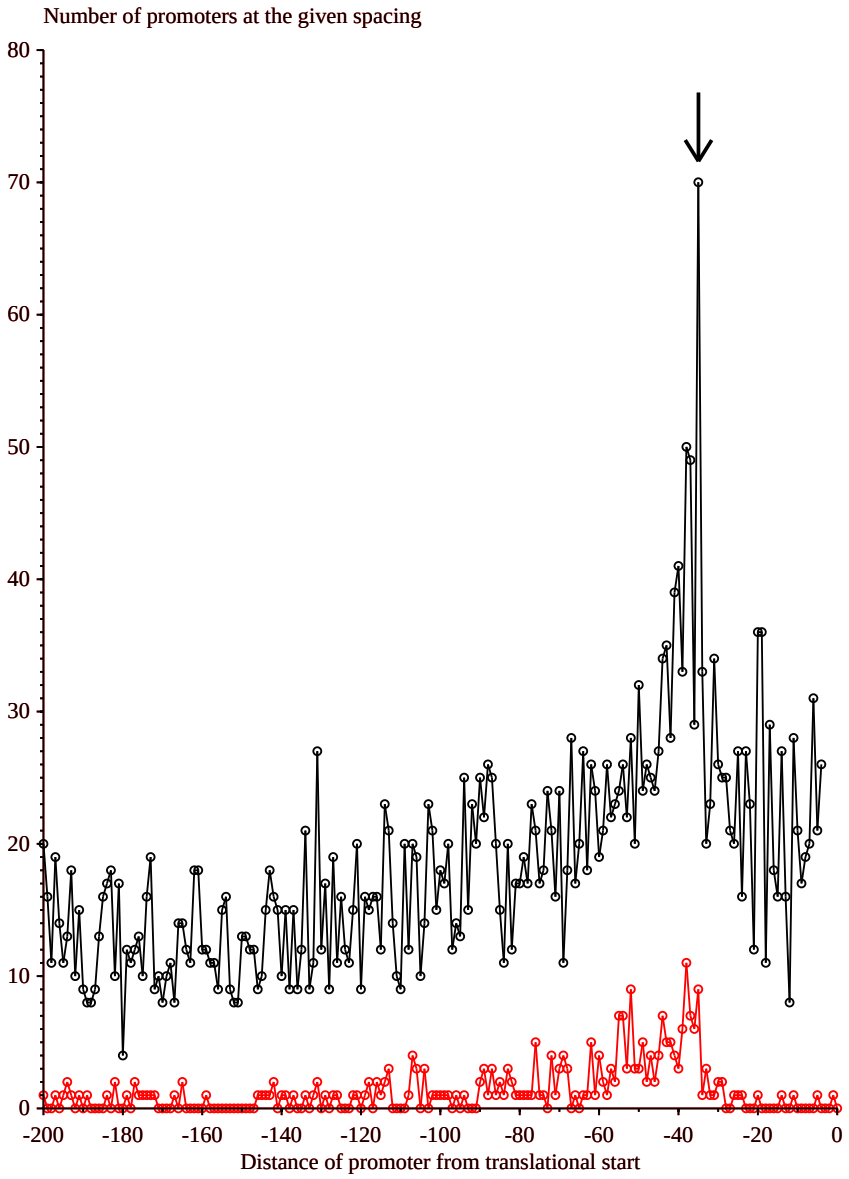


Fig. 4. The optimal spacing of the −10 to the translational initiation codon is around 35 bases. We plotted the distance between the zero coordinate of the −10 and the translational start point on the abscissa, and the number of promoters at that distance on the ordinate. The upper curve (black) represents data from a scan using our promoter model over the upstream regions of all 4122 genes in *E. coli* [56]. The lower curve (red) represents the location of the −10 relative to translational initiation codons for experimentally verified transcription start points. The arrow pointing to the black curve indicates a peak at −35 bases.

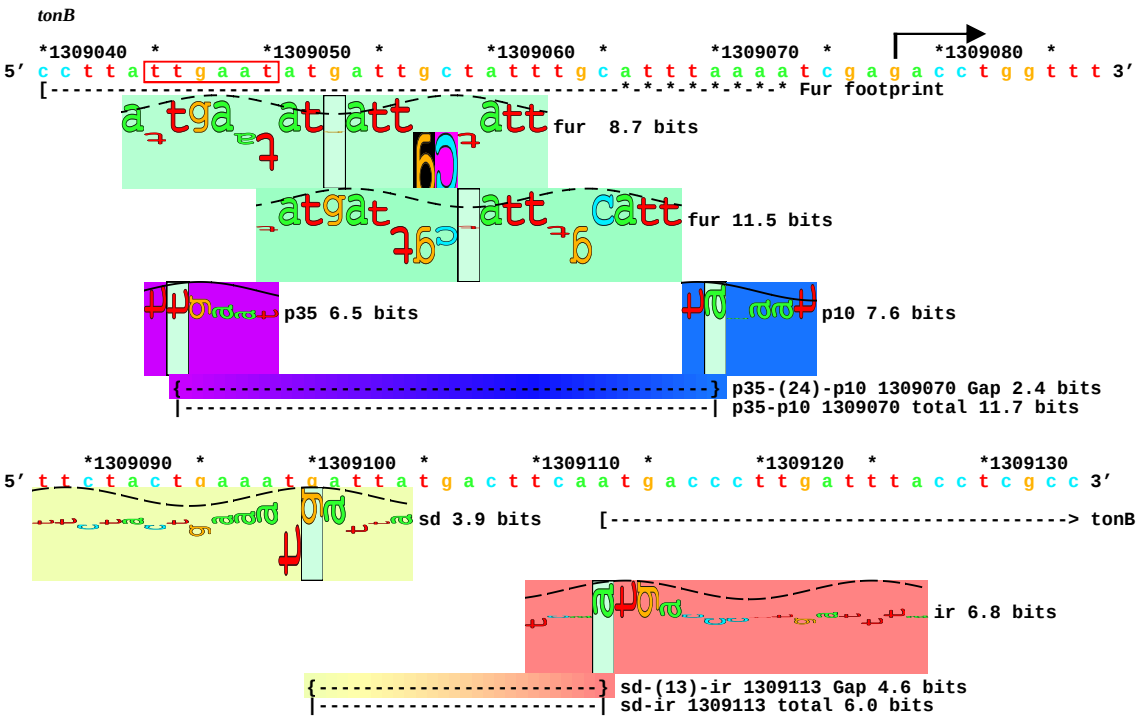


Fig. 5. Fur represses transcriptional initiation of *tonB*.

We present here an individual information analysis [46] of the Fur controlled *tonB* region [71,111] using the sequence walker method [48]. Colored rectangles ('petals') behind the walkers identify the kind of site (by hue) and the strength of the site (by saturation) [112]. The connecting bar between parts of a flexible site transitions linearly between the corresponding colors. The $\sigma^{70}$ binding site and the ribosome binding site were both located using flexible binding models [34], in that there are variable distances between binding components. The horizontal dashed line underneath the −10 walker (p10) and the −35 walker (p35) that is labeled 'Gap', gives the gap surprisal for whatever distance separates the two components, as well as the coordinate of the downstream component. The dashed line that is labeled 'total' gives the total information for the flexible site. Similar lines are underneath the ribosome binding site components (SD, IR). A Fur dimethyl sulfate protection footprint is marked [71,111], and two sequence walkers for Fur fall below it. The downstream side of the Fur protected sequence is somewhere in the region marked by asterisks (*). The sequence that the polymerase and Fur would both bind is marked with a red box. The transcription start point for *tonB* is marked with a black arrow, and the translational initiation start point is marked at position 1309113 with a bracket and an arrow. The sequences and coordinates on the map are from GenBank accession number **U00096** [113].

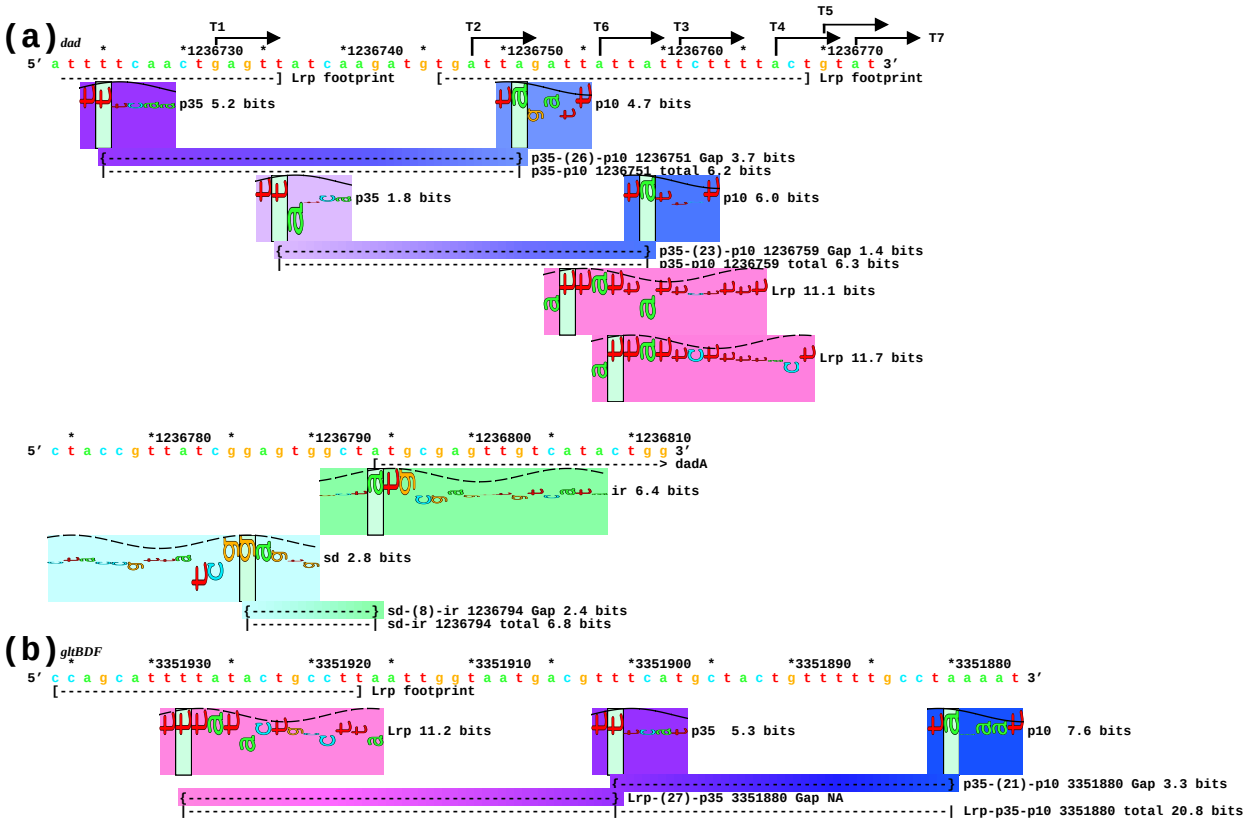


Fig. 6. Transcriptional control by Lrp is based on its spacing relative to the polymerase. (a) An individual information analysis of the Lrp repressed *dad* operon [73,75,74]. (b) An individual information analysis of the Lrp activated *gltBDF* operon [76]. As in Fig. 5, the $\sigma^{70}$ and ribosome binding sites are each internally connected by lines that report the gap surprisal and total information. Experimentally verified transcription start points are identified with black arrows and named according to Zhi *et al.* [75], and the *dadA* gene start is marked with a bracket and arrow at position 1236794. In (b), since Lrp helps to stabilize the initiation complex, its information is added into the total strength of the promoter. Since data on the distance between Lrp sites and the −35 are not available, we did not subtract a gap surprisal and therefore the gap surprisal is marked as NA (not applicable). The sequence and coordinates on the map are from GenBank accession number **U00096** [113].

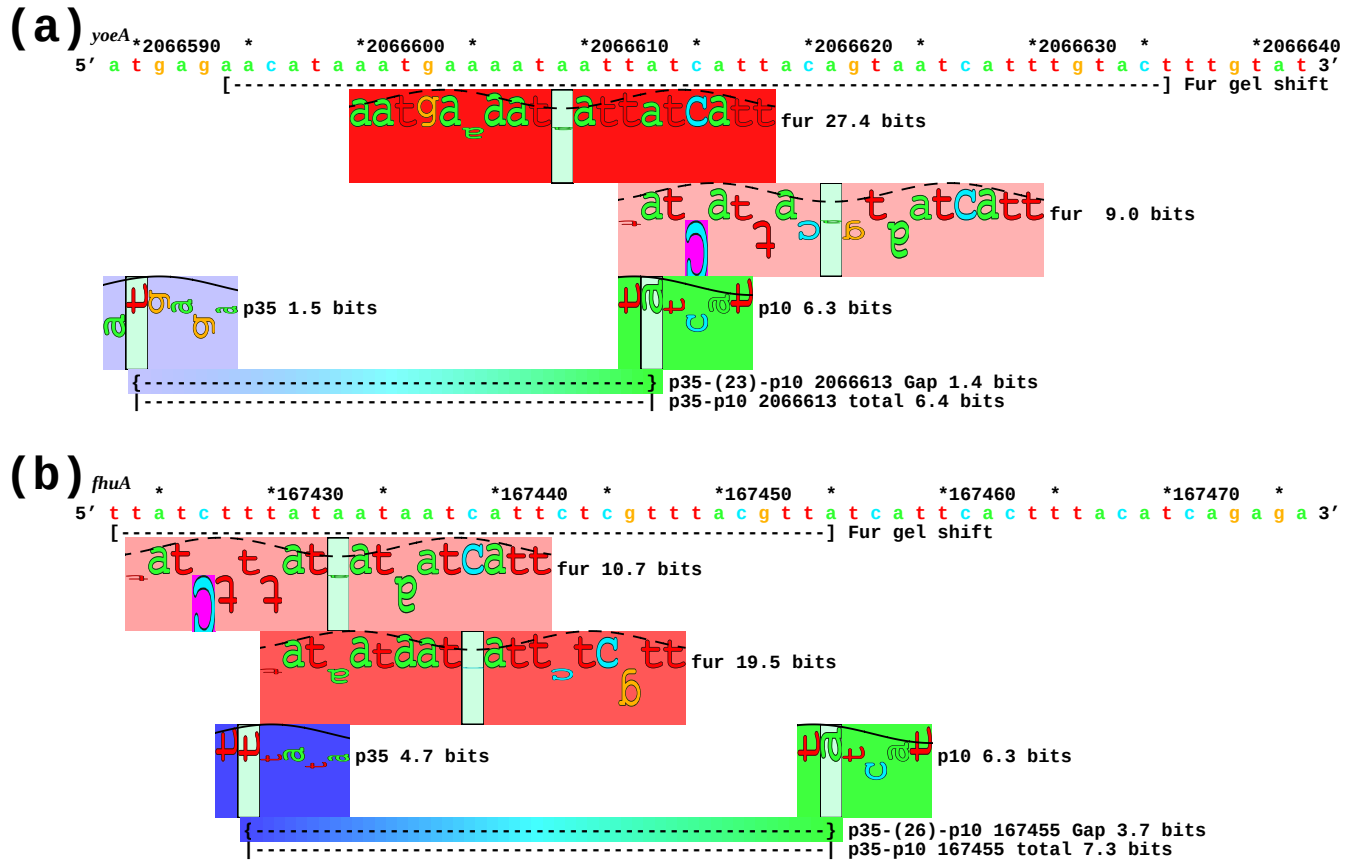


Fig. 7. Sequence walkers for $\sigma^{70}$ and Fur protein upstream of the (a) *yoeA* and (b) *fhuA* genes suggests that these genes are controlled by Fur.
Synthetic oligonucleotides that contain sequences marked by brackets under the DNA showed gel mobility shifts by Fur protein (data not shown). The sequence and coordinates on the map are from GenBank accession number **U00096** [113].

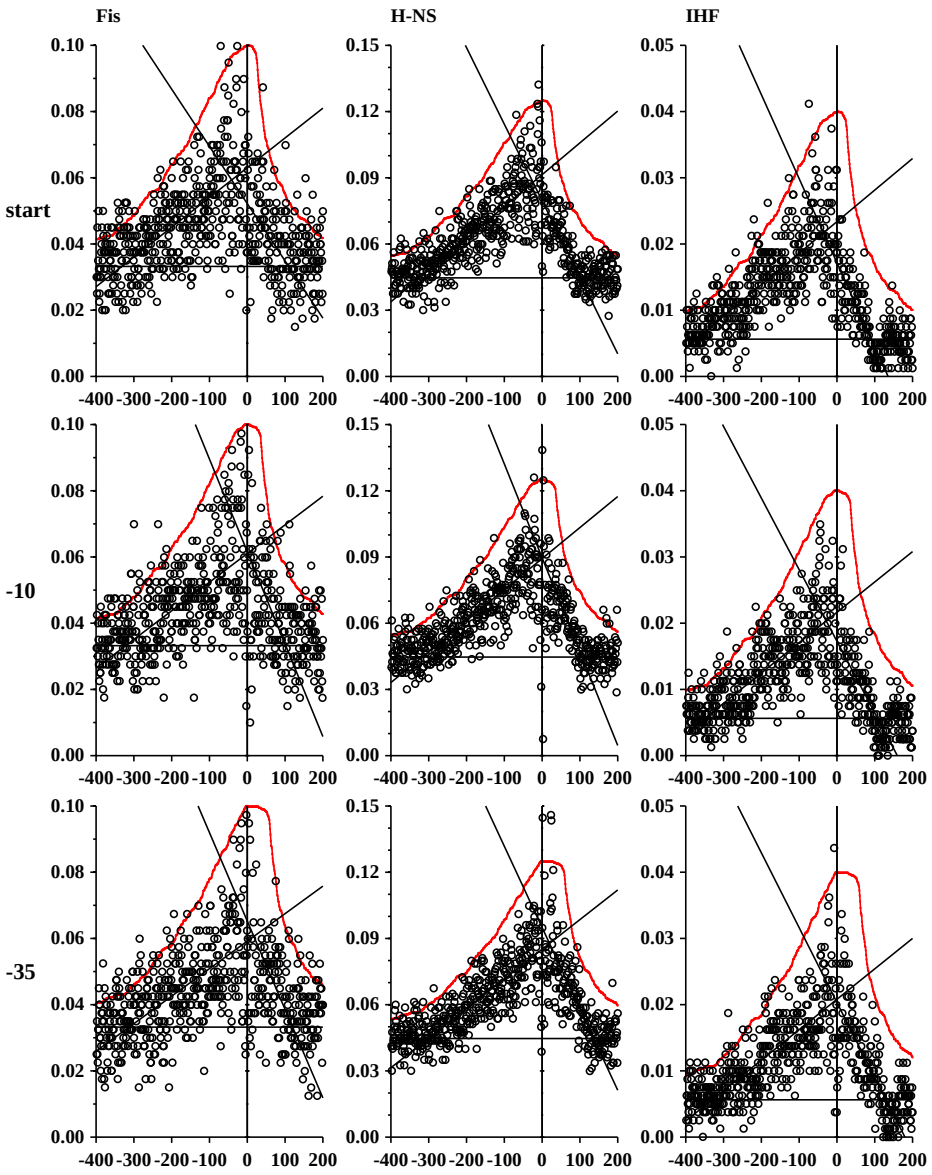


Fig. 8. Intergenic binding of DNA-bending proteins relative to promoter components. These curves allow one to directly compare the density of non-coding regions to the number of DNA binding protein sites at each position relative to experimentally determined promoter components. For all graphs, the abscissa is the position of the regulator binding site (either Fis, H-NS or IHF) relative to either the transcription start, the −10, or the −35 in our promoter model. A vertical line marks the zero coordinate of the promoter component. The ordinate is the frequency of sites at that spacing (sites per base). A solid horizontal line marks the frequency of sites per base predicted for the entire genome. Linear regression lines for −400 to 0 and 0 to 200 are shown. A distribution corresponding to the density of intergenic regions surrounding the experimentally verified promoters was fit to the data in each graph, and is shown as a solid red curve (see Materials and Methods).