Integrative Biology

RSC Publishing

PAPER

View Article Online
View Journal | View Issue

Cite this: *Integr. Biol.*, 2013, **5**, 796

Efficient transcription initiation in bacteria: an interplay of protein–DNA interaction parameters†

Marko Djordjevic*

As the first, and usually rate-limiting, step of transcription initiation, bacterial RNA polymerase (RNAP) binds to double stranded DNA (dsDNA) and subsequently opens the two strands of DNA (the open complex formation). The rate determining step in the open complex formation is opening of a short (6 bp) DNA called the −10 region, which interacts with RNAP in both dsDNA and single stranded (ssDNA) forms. Accordingly, formation of the open complex depends on (physically independent) domains of RNAP that interact with ssDNA and dsDNA, as well as on parameters of DNA melting and sequences of −10 regions. We here aim to understand how these different interactions are mutually related to ensure efficient open complex formation. To achieve this, we use a recently developed biophysical model of transcription initiation, which allows the calculation of the kinetic parameters of transcription initiation on the scale of whole genome. We consequently investigate kinetic properties of sequences derived from all *E. coli* intergenic regions, and from more than 300 experimentally confirmed *E. coli* σ^{70} promoters. We find that interaction specificities of σ^{70} DNA binding domains reduce the number of sequences where RNAP binds strongly, but forms the open complex too slowly to achieve functional transcription (so-called poised promoters). However, we find that, despite this reduction, there is still a significant number of such poised promoters in the intergenic regions, which may provide a major source of false positives in genome-wide searches of transcription start sites. Furthermore, we surprisingly find that sequences of −10 regions of the functional promoters increase the extent of RNAP poising, which we interpret in terms of an extension of a recently proposed model of promoter recognition ('mix-and-match model') to kinetic parameters. Overall, our results allow better understanding of the design of σ^{70} DNA binding domains and promoter sequences, and place a fundamental limit on accuracy of methods for promoter detection that are based on strong RNAP binding (*e.g.* ChIP-chip).

Received 18th September 2012,
Accepted 11th February 2013

DOI: 10.1039/c3ib20221f

www.rsc.org/ibiology

Insight, innovation, integration

In bacteria, transcription is initiated by RNA polymerase (RNAP) binding to double-stranded DNA, and by subsequent opening of the two DNA strands. Kinetic parameters, which characterize these two steps, have to be measured individually for each sequence of interest, through an arduous experimental procedure. To investigate the kinetics of transcription initiation on the whole genome scale, here we use a recently developed biophysical model, which allows calculating the kinetic parameters for any sequence of interest. We find that RNAP DNA-interaction domains are designed to reduce kinetic trapping of RNAP in the genome. On the other hand, we, surprisingly, find that sequences of functional promoters increase RNAP poising, which we interpret in terms of a recently proposed model of promoter recognition.

1 Introduction

Transcription initiation is both the first step and a major control point in gene expression. Transcription cannot be initiated by core RNA polymerase alone, so a complex between core RNA polymerase and a σ factor, which is called RNA polymerase holoenzyme (RNAP), is formed.[1] Different σ factors interact with double-stranded DNA (dsDNA) and single-stranded DNA (ssDNA) in a sequence specific manner, and they are responsible for transcription under different conditions.[2] In this work we concentrate on σ^{70} (the major σ factor in *E. coli*), which is responsible for transcribing housekeeping genes.[3]

Transcription is initiated from the sequences called core promoters. The main elements of core promoters in bacteria

Institute of Physiology and Biochemistry, Faculty of Biology, University of Belgrade, Studentski trg 16, 11000 Belgrade, Serbia. E-mail: dmarko@bio.bg.ac.rs; Fax: +381 11 2639 882; Tel: +381 63 1312 976

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c3ib20221f

are −35 element and −10 element, where −35 and −10 refer to typical distances of these elements from transcription start sites.[4] As the first step of transcription initiation, RNAP reversibly binds to dsDNA of promoter elements, which is called the closed complex formation, and is described by the binding affinity K_B. The binding affinity is, therefore, determined by interactions of σ^{70} with dsDNA, which is exhibited through interactions of σ^{70} domain 4.2 with −35 box, and σ^{70} domain 2.4 with −10 box in the dsDNA form.[2]

This binding of RNAP leads to opening of the two DNA strands (promoter melting), so that a transcription bubble is formed. This transcription bubble extends from the upstream edge of the −10 element to about two bases downstream of the transcription start site, which roughly corresponds to positions −12 to +2 (+1 is the transcription start site).[5] The (inverse) time needed to form the transcription bubble (*i.e.* to open the two DNA strands) is described by the transition rate from the closed to open complex (k_f). The transition rate, therefore, crucially depends on interactions of σ^{70} with −10 element ssDNA, which are exhibited through σ^{70} domain 2.3.[6]

Since almost the entire −10 element is a part of the transcription bubble, this element interacts with σ^{70} in both dsDNA and ssDNA forms. While sequences from the downstream edge of the −10 element to the transcription start site are also part of the transcription bubble, mutating these sequences does not affect the bubble formation,[7] and it is considered that these sequences do not interact with σ^{70} in a sequence specific manner. Furthermore, both theoretical studies[8] and single molecule experiments[9] show that opening of −10 element is the rate limiting step in the transcription bubble formation. Since the −10 box is a part of both the closed and the open complex, there is a complex interplay of biophysical interactions associated with this element: (i) DNA melting energies,[10] since the −10 box dsDNA is opened (melted) in the open complex, (ii) interaction energies of σ^{70} with dsDNA through σ^{70} subdomain 2.4,[11] and (iii) interaction energies of σ^{70} with ssDNA through σ^{70} subdomain 2.3. These three types of interactions are physically independent, since they either correspond to intrinsic DNA properties (for melting energies) or are exhibited through physically distinct σ^{70} binding domains (for σ^{70}–dsDNA and σ^{70}–ssDNA interactions).[6]

Given the complex set of physically independent interactions at the −10 element described above, there is a question of how their mutual relationship leads to efficient transcription. In particular, the RNAP binding affinity (K_B) depends on interactions of −10 box dsDNA with σ^{70} subdomain 2.4,[6] where the stronger interaction leads to larger binding affinity. On the other hand, a stronger interaction of σ2.4 with dsDNA of −10 element leads to a slower transition from the closed to open complex.[8] The transition rate (k_f) also depends on interactions of −10 box ssDNA with σ^{70} subdomain 2.3 and on the −10 element melting energy, both of which are physically independent of σ2.4.[8,12] Due to this, K_B and k_f should *a priori* be negatively correlated, and there may be a large number of sequences in the genome that correspond to high K_B but low k_f. We call such sequences where RNAP is strongly bound to dsDNA (high K_B), but proceeds to the open complex too slowly to achieve functional transcription (due to small k_f), poised promoters; more generally, the term poised promoter is used for all instances where RNAP is bound strongly to DNA, but fails to proceed to functional transcription.[13] Naively, RNAP poising appears particularly detrimental for sequences that should be transcriptionally active (functional promoters), since these sequences should result in efficient transcription.

Given the kinetic issues discussed above, we here aim to understand the following questions: (i) what is the extent of RNAP poising in the genome? (ii) Are binding specificities of σ^{70} interaction domains, and/or sequences of *E. coli* intergenic regions, designed to minimize the number of poised promoters? (iii) Do sequences of functional σ^{70} promoters (additionally) suppress RNAP poising? We here concentrate on the intergenic regions, rather than on the whole genome, since these regions are relevant for transcription regulation, *i.e.* both transcription start sites and regulatory elements are located in the intergenic regions.

The questions posed above are important not only from the point of design of σ^{70}–promoter DNA interactions, but also from the point of searches for functional promoters in the genome. In particular, the most common experimental method to search for core promoters on a genome-wide scale is ChIP-chip[14] and its alternatives (*e.g.* ChIP-seq[15]). However, immunoprecipitation (ChIP) detects DNA sequences that are strongly bound by the protein (RNAP), rather than sequences with a high rate of transcription initiation – which is the parameter that defines a functional promoter. Consequently, the high number of false positives, which is commonly associated with ChIP-chip experiments aimed for promoter detection,[16] may indicate extensive RNAP poising in the genome. The goal of this paper is to investigate a relationship between physical interactions at the −10 element and RNAP poising, which provides a basis for better understanding of the nature of false positives in ChIP-chip experiments.

Along the same lines, DNA footprinting experiments detected sequences that are strongly bound by RNAP, but which result in transcriptionally inactive complexes; these inactive complexes were shown to be due to inefficient formation of the open complex (*i.e.* due to RNAP poising).[17] Such observations seem particularly important from the point of computational searches of transcription start sites (core promoters) in the genome, which typically lead to a very high number of false positives. It was consequently proposed that kinetic effects – an extreme example of which are poised promoters – can significantly contribute to accuracy of the weight matrix (computational) searches of promoters.[18] Furthermore, an understanding of the kinetic effects, which we will achieve in this paper, will motivate their inclusion within more physical methods of TSS recognition. With regard to this, it was frequently observed that coupling biophysical models with sequence statistics provides a significantly better prediction accuracy compared to simple statistical models.[19]

In order to analyze how the interplay of different interaction parameters leads to efficient transcription, one must be able to investigate kinetics of transcription initiation on a genome wide scale. This analysis cannot be done through experiments, since K_B and k_f have to be measured through work-intensive τ-plot measurements,[20] individually for each sequence of interest. We here instead approach the problem computationally, where we use a recently developed biophysical model of the open complex formation,[8] which allows the calculation of the kinetic parameters (K_B and k_f) for each sequence of interest. This model showed a very good agreement with both biochemical and genomics data, with no free parameters used in comparing the model with the experimental data.[8]

We will here show that binding specificities of σ^{70} DNA interaction domains are designed to prevent extensive RNAP poising in the intergenic regions, but that the number of poised promoters is still sufficient to significantly affect accuracy of core promoter searches. Surprisingly, we will find that sequences of functional −10 elements *increase* the extent of RNAP poising; on the other hand, overall, the sequences in the intergenic regions have no tendency to affect RNAP poising. Though seemingly counter-intuitive, we will argue that this result fits well within the recently proposed mix-and-match model of promoter recognition.[21]

2 Results

2.1 Design of *in silico* experiments

Our goal is to investigate how the interplay of physical interactions at the −10 promoter region provides for efficient transcription. We, consequently, systematically investigate relations between the kinetic parameters as the −10 element sequence is varied. To achieve this, we design a number of *in silico* experiments, where we start from a sequence of the lacUV5 promoter. This promoter has a consensus −10 element – which is convenient as a reference for calculating kinetic parameters – but has an imperfect −35 element as is characteristic for most functional promoters.[5] In the analysis/*in silico* experiments presented in the following subsections, we will substitute the consensus −10 element of lacUV5 promoter with different sets of DNA segments.

The biophysical model of transcription initiation[8] allows the calculation of the relevant kinetic parameters for sets of DNA segments at the scale of the entire genome (see Methods and ESI†). In particular, in the analysis below, we will substitute the consensus −10 element of lacUV5 promoter with: (i) all 6 bp long segments from *E. coli* intergenic regions, (ii) all −10 elements that correspond to experimentally detected *E. coli* transcription start sites, (iii) segments that correspond to randomized intergenic regions and randomized −10 elements of experimentally detected promoters; the computational procedure allows randomizing DNA sequences multiple times, so that statistics of the relevant quantities can be calculated.

In the analysis below, we will also address how relevant σ^{70} DNA-interaction domains contribute to the kinetic properties that we investigate. Experimentally, contributions of different protein domains to the properties of interest would be assessed by mutating amino-acid sequences of these domains. We will computationally assess contributions of σ^{70} domains by randomizing interaction specificities of these domains; similarly as with DNA sequences, we can perform multiple randomizations in order to calculate statistics of the relevant quantities. Finally, we will also substitute binding specificities of σ^{70} domains with binding specificities of different *E. coli* transcription factors, in order to ensure that the reported relationships are not a consequence of generic properties of protein–DNA interactions.

2.2 Kinetic properties of *E. coli* intergenic regions

We start from the sequence of the lacUV5 promoter, and substitute its consensus −10 element with all 6 bp long segments from *E. coli* intergenic regions. For all these substitutions we calculate the relative binding affinity (K_B) and the relative transcription initiation rate (φ), by using eqn (1) and (3) (see Methods). The relationship between logarithms of K_B and φ is shown in Fig. 1A, so that the quantities on the two axis correspond to the appropriate interaction energies that determine the relevant kinetic parameters. Specifically, the horizontal axis ($\log(K_B)$) corresponds to the σ^{70}–dsDNA binding energy, while the vertical axis corresponds to a combination of the energy terms that we refer to as the effective energy and which directly determines the transcription initiation rate (see eqn (3) and (4) in Methods).

Both K_B and φ, which are shown in Fig. 1A, are calculated relative to the binding affinity and the transcription initiation rate of the lacUV5 promoter. Note that we substitute (vary) only the −10 element of lacUV5 promoter, and that −10 element of this promoter corresponds to the consensus sequence ('$^{-12}$TATAAT^{-7}'). Consequently, zeros on the horizontal and the vertical axis correspond to the consensus −10 element, and stronger interaction energies correspond to larger (less negative) values on the two axes. The horizontal line in Fig. 1 (transcription rate threshold) indicates the transcription rate below which transcript levels cannot be detected, while the vertical line (binding threshold) indicates the binding affinity above which a sequence is considered to be strongly bound by RNAP. The transcription rate threshold is set based on the estimate that the minimal rate of transcription is 1/400 per second, while the transcription rate of the reference lacUV5 can be estimated at 1/3 per second.[22] The binding threshold is set so that it corresponds to the binding affinity of a weak Plac promoter, with sequences of −35 element and −10 element that correspond, respectively, to '$^{-36}$TTTACA^{-31}' and '$^{-12}$TATGTT^{-7}';[23] this definition is in accordance with an intuitive notion that strongly bound sequences should have a larger binding affinity than a weak promoter.

Fig. 1A shows that there is a high positive correlation (with a Pearson correlation coefficient of R = 0.85) between the transcription activity and the binding affinity for −10 elements derived from *E. coli* intergenic regions. One should note that the determinants of binding affinity and transcription activity are physically independent (see the previous section), so the good correlation has to be due to the design of σ^{70} interaction domains or due to the sequence of DNA intergenic segments, which is further explored in the next subsection. However, despite this high

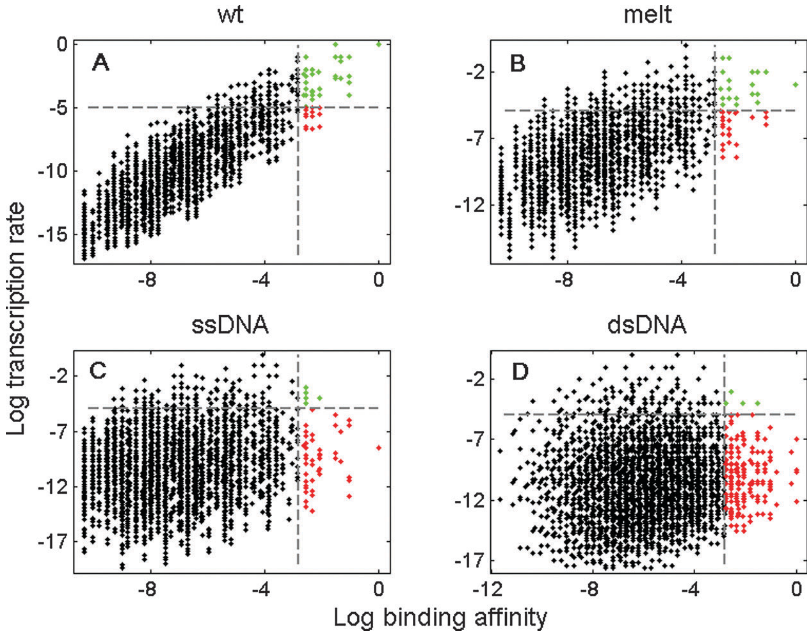

Fig. 1 Log transcription rate ($\log(\varphi)$) *vs.* log binding affinity ($\log(K_B)$) for the intergenic segments. −10 element of the lacUV5 promoter is substituted by all 6 bp long segments from *E. coli* intergenic regions. $\log(K_B)$and $\log(\varphi)$ are calculated for each of these substitutions and shown, respectively, on the horizontal and the vertical axes on each of the panels. Both $\log(K_B)$and $\log(\varphi)$ are calculated relative to the values for the lacUV5 promoter, so that zero of the energies corresponds to the consensus −10 element, and the stronger interactions correspond to larger (less negative) values on the two axes. The horizontal and the vertical dashed lines correspond, respectively, to the transcription rate threshold and the binding affinity threshold. Green and red dots in the figure correspond to the strongly bound DNA sequences that are, respectively, functional promoters and poised promoters. The four panels correspond to the relationship between $\log(\varphi)$ and $\log(K_B)$ for: (A) actual σ^{70} interaction parameters, (B) permutation of the melting energies, (C) randomized interaction specificities of σ^{70} subdomain 2.3 (σ^{70}–ssDNA interactions), and (D) randomized interaction specificities of σ^{70} subdomain 2.4 (σ^{70}–dsDNA interactions). The figure shows a good correlation between the binding affinity and the transcription initiation rate, though a significant number of poised promoters is still present (panel (A)). Panels (B)–(D) indicate that good correlation depends on all three types of the interaction parameters that are relevant for transcription initiation.

correlation, a significant fraction of the strongly bound sequences corresponds to poised promoters: in Fig. 1, the green dots mark strongly bound DNA segments that correspond to the functional promoters (*i.e.* to sequences that are above both the binding and the transcription activity threshold), while the red dots mark the sequences that correspond to the poised promoters (*i.e.* to sequences that are above the binding, but below the transcription activity threshold). One can see that a significant fraction of the strongly bound sequences (∼30%) correspond to poised promoters. Such poised promoters can be falsely identified as targets by computational and experimental searches of core promoters, which we will further discuss in the next section.

2.3 Analyzing the good correlation between the transcription rate and the binding affinity

In this subsection, we concentrate on the properties of σ^{70}–DNA interactions that lead to the good correlation between the transcription activity and the binding affinity, which is observed in Fig. 1A. As discussed above, K_B depends on σ^{70} interactions with −10 element dsDNA, while φ depends on interactions of σ^{70} with −10 box ssDNA and on DNA melting energies.[8] Since K_B and φ are physically independent of each other, there is a question of why there is a good correlation between the transcription rate and binding affinity that is observed in Fig. 1A. The first possibility is that this good correlation is due to the sequence of *E. coli* intergenic regions, *i.e.* the presence of poised promoters is suppressed in these sequences. This possibility might be reasonable, since existence of a large number of poised promoters could be detrimental for efficient transcription initiation (see also Discussion). The second possibility is that the good correlation is due to the design of σ^{70} DNA interaction domains (specifically due to the binding specificities of σ^{70} subunits 2.3 and 2.4). We test these two possibilities below.

In order to generate an appropriate ensemble to test the possibility that the good correlation is due to the DNA sequence, we next randomize the DNA sequence of *E. coli* intergenic regions 50 times. The randomizations are performed so that frequencies of the nucleotides are preserved (see Methods). We next re-calculate the correlation coefficient between the transcription rate and the binding affinity for each of the 50 randomized sequences, and obtain the mean for these 50 randomizations as $\bar{R} = 0.84$ (the relationship between the transcription rate and the binding affinity for one such randomization is shown in ESI,† Fig. S1). This value ($\bar{R} = 0.84$) is only somewhat smaller compared to the correlation coefficient for the actual *E. coli* intergenic regions ($R = 0.85$). Consequently, the design of the DNA sequence of the intergenic regions is not a reason for the high correlation between the transcription rate and the binding affinity.

As the second possibility, we analyze if the high correlation is due to the design of the binding specificities of σ^{70} DNA interaction domains. To test this possibility, we randomize the binding specificities that correspond to σ^{70} subunit 2.3 (σ^{70}–ssDNA interactions) and 2.4 (σ^{70}–dsDNA interactions) and DNA melting energies (see Methods). We first permute the two parameters that – in the single nucleotide approximation – characterize DNA melting (melting energies of A:T and G:C pairs – see Methods); the effect of this permutation is shown in Fig. 1B. In Fig. 1C and D we show the effect of randomization of, respectively, σ^{70} binding domains 2.3 and 2.4.

Fig. 1B–D show that (separately) randomizing each of the interaction energies leads to a large decrease in the correlation coefficient, and to a consequent large increase in the fraction of poised promoters (the red dots in Fig. 1B–D). In particular, note that not only randomizations of the interaction domain specificities (Fig. 1C and D), but also the permutation of the melting energies (Fig. 1B) lead to a significant decrease in the correlation coefficient. This indicates that the reduction of RNAP poising in the genome depends on an interplay of all the relevant parameters (*i.e.* on the mutual relation between ssDNA, dsDNA and melting energy parameters).

To test statistical significance of the results, in Fig. 1C and D, we calculate correlation coefficients for 50 randomizations of ssDNA interaction parameters (σ^{70} subunit 2.3), and for 50 randomizations of dsDNA interaction parameters (σ^{70} subunit 2.4). The mean values and 95% confidence intervals for these randomizations are shown in the histogram (see Fig. 2). For comparison, the correlation coefficient for the actual (wild type) interaction parameters and for the permutation of the melting parameters are also indicated. We see that all the randomizations indeed lead to a statistically significant (and large) decrease in the correlation coefficient. Consequently, the reduction in the number of poised promoters in the intergenic regions depends on the mutual relationship of all physical parameters that are relevant for opening the −10 element.

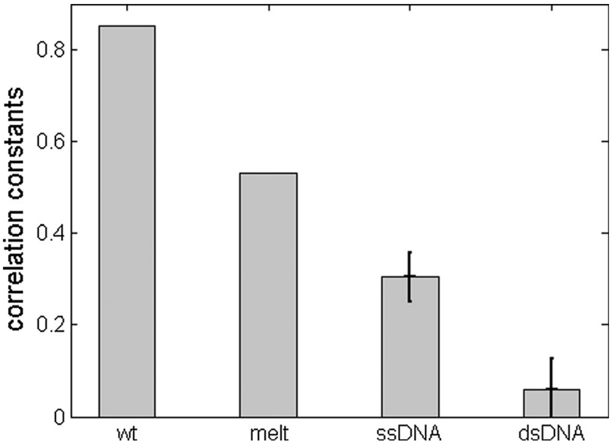

Fig. 2 Statistics of the interaction parameter randomizations. The first two bars in the histogram indicate the correlation coefficients between the transcription rate and the binding affinity for, respectively, the actual (wild type) σ^{70} binding specificities (labeled as ''wt''), and for the permutation of the melting parameters (labeled as ''melt''). For the last two bars, the ssDNA (σ^{70} subunit 2.3) and dsDNA (σ^{70} subunit 2.4) binding specificities are randomized 50 times, and the correlation coefficients between the transcription rate and the binding affinity are calculated for each of these randomizations. The mean value and 95% confidence limits, which correspond to the randomizations for ssDNA interaction parameters (the third bar, labeled as ''ssDNA'') and for dsDNA interaction parameters (the last bar, labeled as ''dsDNA''), are shown. The figure shows that the correlation coefficient is significantly reduced by randomization of any of the relevant sets of parameters, so that the high correlation constant depends on an interplay of all these parameters.

Finally, from Fig. 2 one can also note that randomization of dsDNA interaction parameters (σ^{70} domain 2.4) leads to an almost complete loss of the correlation. The reason for this loss is that the binding affinity depends exclusively on dsDNA interactions, while the transcription rate depends on dsDNA interactions through only one out of six bases of the −10 element (base −12) (see eqn (1), (3) and (4)). Consequently, randomization of dsDNA interactions leads to an almost complete loss of the relation between the binding affinity and the transcription rate.

2.4 Substitutions of σ^{70} DNA interaction domains

In this subsection, we provide further evidence that the binding specificities of σ^{70} interaction domains are designed to prevent extensive RNAP poising. Specifically, while we established that the good correlation is due to the specificities of σ^{70} DNA-binding domains, it remains to be confirmed that the effect is not an artificial consequence of some generic property of protein–DNA interactions. For example, such an artifact would arise if protein–DNA binding domains would have a general tendency to recognize similar AT rich sequences. To test this, we substitute specificities of binding domains 2.3 and 2.4 with specificities of different *E. coli* DNA binding proteins. Parameters of protein–DNA interactions are inferred from binding sequences assembled in DPInteract database,[24] by using the QPMEME algorithm.[19b]

From DPInteract database we can infer, with a high reliability, interaction specificities of 8 *E. coli* transcription factors (see Methods). We then substitute specificities of RNAP binding domains 2.3 and 2.4 with these inferred specificities, which makes a total of 56 substitution pairs; note that we do not allow for the same *E. coli* transcription factor specificity to substitute both σ^{70} domains 2.3 and 2.4. For each of these substitutions we calculate correlation between the rates of transcription and binding affinities, as described in the previous subsection. The distribution of the correlation coefficients for the substitutions is shown in Fig. 3, and the correlation for the actual σ^{70} binding domains is also indicated in the figure for comparison. We see that the correlation in the case of the actual σ^{70} binding domains is significantly larger compared to all the substitutions, with a very high statistical significance (P value of $\sim 10^{-24}$). Therefore, the good correlation is not an artificial consequence of some generic property of protein–DNA interactions, and interaction domains of RNAP are indeed ''hardwired'' so as to reduce RNAP poising in the genome.

2.5 Kinetic properties of experimentally detected σ^{70} promoters

We next investigate kinetic properties of −10 elements associated with 342 experimentally confirmed transcription start sites.

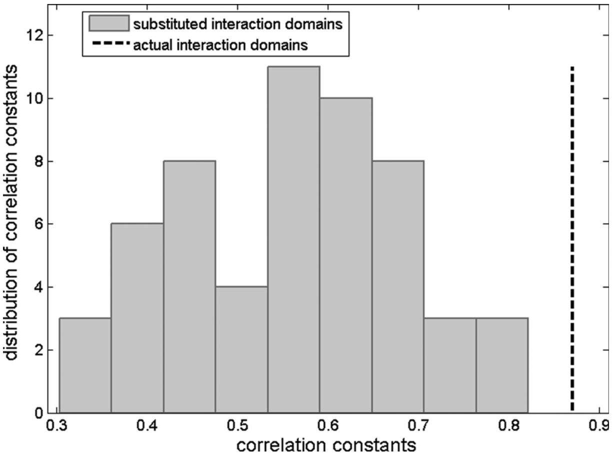

Fig. 3 Substitutions of σ^{70} binding specificities. The binding specificities of σ^{70} domains 2.3 and 2.4 were substituted with different combinations of the interaction specificities for 8 different *E. coli* transcription factors. Gray bars correspond to the histogram of the correlation coefficients between the transcription rate and the binding affinity for these substitutions. Correlation coefficient in the case of the actual RNAP binding domains is indicated by the vertical dashed line. The figure shows that the binding specificities of the actual (wild type) σ^{70} interaction domains lead to a significantly larger correlation – and consequently a larger reduction in RNAP poising – compared to all of the substitutions.

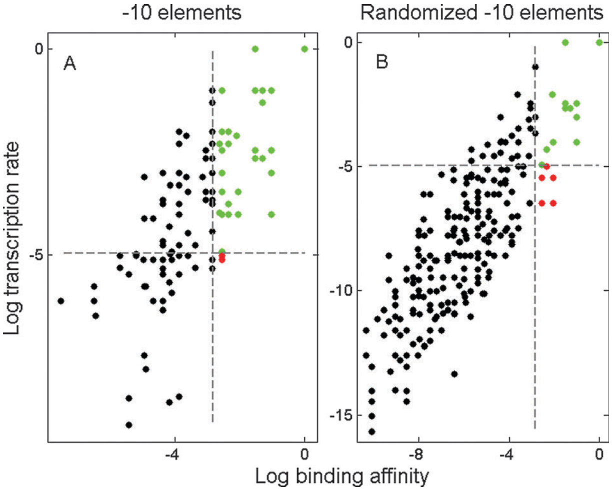

Fig. 4 The relationship between the transcription activity and the binding affinity for actual and randomized −10 elements of experimentally inferred σ^{70} promoters. Transcription start sites with experimentally established transcription activity were selected from RegulonDB database (see Methods). These transcription start sites were next used to align 342 −10 elements, according to the procedure described in Methods; we further refer to these segments as *actual* −10 elements. The actual −10 elements were then randomized such that the nucleotide frequencies are preserved (see Methods). The relationship between the log transcription rate and the log binding affinity is calculated and plotted for (A) the actual −10 elements and (B) the randomized −10 elements. The figure shows that, surprisingly, sequences of actual (experimentally established) −10 elements result in increased RNAP poising.

Selection of the transcription start sites with experimentally confirmed transcription activity from RegulonDB database,[25] and alignment of −10 elements associated with these transcription start sites, is described in Methods. We substitute the consensus −10 element of the lacUV5 promoter with these aligned −10 elements, and for each of these substitutions we calculate the transcription rate and the binding affinity; the obtained relationship between these two quantities is shown in Fig. 4A.

One may expect that RNAP poising at the transcriptionally active sequences should be suppressed to a larger extent compared to the generic segments from the intergenic regions. However, in contrast to this expectation, we find that the correlation in the case of the transcriptionally active −10 elements is notably smaller than the correlation for the intergenic segments (0.75 *vs.* 0.84, compare Fig. 4A with Fig. 1A); to further assess this result, we analyze how the correlation changes when functional −10 elements are randomized. To obtain appropriate statistics, we randomize the set of aligned −10 elements 50 times, and then calculate the correlation coefficient for each randomization. Consistent with the result obtained above, the mean of the correlation coefficients for these randomizations is notably larger compared to the correlation for the actual −10 elements (0.85 *vs.* 0.75), with a very high statistical significance ($P \sim 10^{-39}$). Therefore, the DNA sequences of the transcriptionally active −10 elements indeed significantly decrease the correlation between the transcription rate and the binding affinity, and consequently increase the extent of RNAP poising.

Finally, to visualize the effect of −10 element randomization, we show the relationship between the transcription rate and the binding affinity, for one instance of −10 element randomization (Fig. 4B). Despite the fact that the correlation coefficient in Fig. 4A (actual −10 elements) is notably smaller compared to Fig. 4B (randomized −10 elements), almost all strongly bound sequences for the actual −10 elements correspond to functional promoters. The small number of poised promoters in Fig. 4A is due to the fact that the binding affinity and the transcription rate are (as expected) 'shifted' toward the higher values for the transcriptionally active −10 elements. Consequently, higher values of the kinetic parameters for the transcriptionally active −10 elements are, as expected, the main mechanism for evading RNAP poising. In the next subsection we discuss a more subtle mechanism for evading poised promoters, which we relate to the 'mix-and-match' model of promoter recognition. Finally, note that randomizing −10 elements (Fig. 4B) leads to roughly the same fraction of poised promoters as for the intergenic segments (Fig. 1A), since, upon randomization, magnitudes of both the transcription rate and the binding affinity decrease.

2.6 Extension of the mix-and-match model to kinetic parameters

We here establish a connection between the surprising decrease in the correlation coefficient for functional −10 elements and a recently proposed mix-and-match model of promoter recognition.[21] The mix-and-match model initially proposed that the strengths of the promoter elements, that interact with dsDNA, complement each other so as to achieve a necessary level of overall binding affinity. Subsequently, a more detailed statistical analysis showed that promoter elements match each other to achieve a necessary level of total promoter strength.[26] We here consider an extension of this model to the kinetic parameters, where we propose that the binding affinity and the transition rate match each other to achieve a necessary level of transcription activity.

To test such extension of the mix-and-match model, we start from the intergenic segments (analyzed in Fig. 1A), and from the transcriptionally active −10 elements (analyzed in Fig. 4A). From each of these two sets of sequences, we select the following two subsets: (i) 30% of the sequences with the highest value of the transition rate from the closed to open complex (k_f) and (ii) 30% of the sequences with the lowest value of the transition rate. The transition rates from the closed to open complex (k_f) are calculated according to eqn (2) (see Methods). We next calculate the distribution of the binding affinities for these two subsets – *i.e.* for the sequences with the high and the low values of the transition rate – by using eqn (1) (see Methods). For the intergenic segments, the distributions for the two subsets are shown together in Fig. 5A. Similarly, the two distributions for transcriptionally active −10 elements are shown together in Fig. 5B.

In Fig. 5A, we see that, for the intergenic segments, the mean binding affinity is significantly smaller for the group with small k_f values than for the group with high k_f values ($P < 10^{-100}$). This property decreases the extent of RNAP poising for the intergenic segments, *i.e.* sequences with low values of the transition rates are generally not characterized by high values of the binding affinities. Note that this result is directly related to the high value of the correlation between the binding affinity and the transcription rate for the intergenic segments.

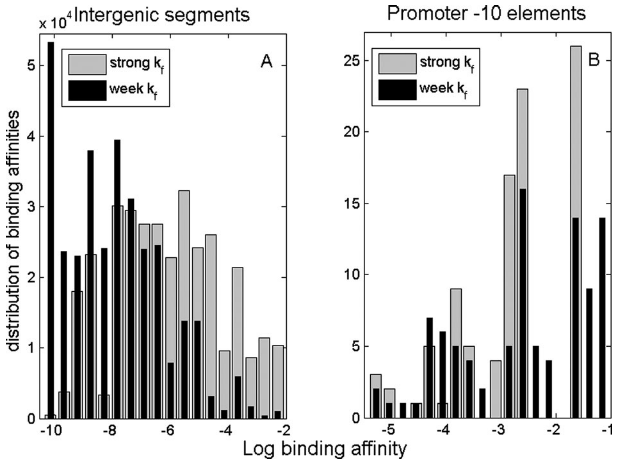

Fig. 5 Complementing kinetic parameters. Substitutions of −10 element of the lacUV5 promoter with (A) intergenic segments and (B) transcriptionally active −10 elements were divided into two groups. The first group corresponds to 30% of the sequences with the highest transition rates from the closed to open complex (k_f), while the second group corresponds to 30% of the sequences with the lowest transition rates. The distribution of the binding affinities (K_B) is calculated and plotted for both the low k_f group (dark gray), and the high k_f group (light gray). The figure shows that the kinetic parameters for functional promoters complement each other, so that −10 elements with low k_f have tendency to have high K_B. This is in contrast to the tendency for the intergenic segments, where sequences with low k_f are generally also accompanied with low K_B.

On the other hand, for the transcriptionally active −10 elements, the distribution of the binding affinities for the group with low k_f is shifted towards the stronger binding affinities, relative to the same distribution for the intergenic segments. As a consequence, for transcriptionally active −10 elements, the group of promoters with high k_f values has smaller mean binding affinities compared to the group with low k_f values (with $P < 0.05$). This result is a consequence of the decrease in the correlation coefficient between the transcription rate and the binding affinity for the transcriptionally active −10 elements relative to the intergenic segments (Fig. 4A *vs.* Fig. 1A), and is analyzed below in terms of the mix-and-match model for promoter recognition.

Though unexpected, the result in Fig. 5B is straightforward to interpret in terms of the extension of the mix-and-match model to kinetic parameters. This figure shows that K_B and k_f complement each other, so that lower k_f is accompanied by higher K_B; this is notably different from the intergenic regions, where sequences with low k_f have tendency to have low K_B. This match of the kinetic parameters for the transcriptionally active −10 elements allows us to achieve a sufficient level of transcription activity (which is proportional to the product of K_B and k_f). This result, and the extension of the mix-and-match model to kinetic parameters, is further discussed in the next section.

3 Discussion

Interactions of σ^{70} with −10 promoter elements are crucial for initiation of transcription. These interactions involve σ^{70} binding domains that interact with dsDNA and ssDNA, as well as DNA melting energies. We here analyzed how the interplay of these interactions affects kinetics of transcription initiation. A prominent example of such kinetic effects are poised promoters, which are sequences where RNAP strongly binds to dsDNA, but has a too slow transition from the closed to open complex to achieve detectable transcription levels. Extensive RNAP poising could be detrimental for efficient transcription, since unproductively bound RNAP can disrupt normal transcription regulation – *e.g.* note that the bound RNAP molecule protects ∼75 bps of DNA, which is often comparable to the size of *E. coli* intergenic regions.[27] Such unproductive binding can also require a significantly larger RNAP production, in order to achieve a sufficiently high RNAP concentration for function of transcriptionally active promoters.

Consequently, it seems plausible that specificities of different interactions and DNA sequences, which are involved in transcription initiation, are somehow tuned to prevent RNAP poising. We here investigated this possibility and showed that σ^{70}–DNA interaction domains, though physically independent, are designed to reduce the extent of RNAP poising in the intergenic regions. This reduction depends on a mutual relationship between all three types of the interaction parameters (ssDNA, dsDNA and melting energies), which strongly suggests that binding specificities of σ^{70}–DNA interaction domains are tuned to evade a large number of poised promoters in the intergenic regions. As another evidence that reduction of RNAP poising is a major 'design' constraint on specificities of σ^{70}–DNA interaction domains, we found that the actual σ^{70} binding specificities lead to a much larger correlation between binding affinity and transcription rate compared with substitutions of these domains with specificities of other *E. coli* transcription factors.

It is interesting that the reduction in the number of poised promoters depends on the binding specificities of σ^{70} interaction domains, rather than on the sequence of the intergenic regions. Such design may allow modularity in reduction of RNAP binding through different bacterial species: while binding specificities of σ^{70} interaction domains are known to be well conserved across different bacteria,[5] DNA sequences of the intergenic regions are widely different. Therefore, imposing the reduction in the number of poised promoters at the level of (conserved) interaction domains, rather than at the level of (variable) DNA sequence, provides a straightforward strategy to impose reduction of RNAP poising in diverse bacterial sequences. Furthermore, there are likely numerous simultaneous constraints on bacterial regulatory (intergenic) regions, since these regions must accommodate a number of functional motifs (*e.g.* core promoters, transcription factor binding sites, terminators). Due to this, tuning the binding specificities of σ^{70} interaction domains may be easier than imposing the absence of poised promoters at the level of DNA sequence.

The fact that σ^{70} interaction domains are designed to reduce the number of poised promoters implies that any DNA sequence will have a tendency for high correlation between the binding affinity and the transcription activity. Such high correlation was also observed for DNA sequences of transcriptionally active promoters. However, we found that DNA sequences of these promoters have a tendency to decrease this correlation, *i.e.* to increase the extent of RNAP poising. This finding is surprising, since one may expect that transcriptionally active sequences should evade RNAP poising.

To better understand this result, it is useful to discuss it from the point of the recently proposed mix-and-match model of promoter recognition. This model proposes that strengths of promoter elements mix with each other, and match each other strengths, so as to achieve the necessary level of promoter strength.[21,26] For example, a weaker −10 element may be complemented by a stronger −35 element, so that a necessary level of transcription activity is achieved.[28] Actually, Fig. 4A shows that many substitutions of the −10 element of the lacUV5 promoter with −10 elements that correspond to the experimentally detected TSS fall below either the binding affinity or the transcription rate threshold. It is likely that, for a substantial number of such −10 elements, the strengths of the other elements within the promoter (−35 element, spacer) are adjusted ('matched') so that the kinetic parameters for the entire promoter are above the thresholds. Furthermore, one should note that some of the known promoters depend on transcription factors in order to achieve sufficient binding affinity and transcription rate, so that their basal values of the kinetic parameters are below the relevant thresholds.

We here proposed to extend the mix-and-match model to the kinetic parameters; consequently, the observed decrease in the correlation between the binding affinity and the transcription activity can be explained by the need to match the lower transition rate from the closed to open complex with higher binding affinity. Our results show that, though statistically significant, this decrease in the correlation is still small enough as not to turn a transcriptionally active promoter into a poised promoter. That is, the observed increase of RNAP poising at functional promoters is such that to allow matching of the kinetic parameters, but not such to cause dysfunctional transcription.

We here predicted that a significant fraction of the strongly bound sequences correspond to poised promoters. This prediction may have a direct consequence on experiments that identify transcription start sites by detecting sequences to which RNAP strongly binds, such as ChIP-chip or ChIP-seq experiments; such measurements provide experimental strategy to detect transcription start sites on a genome-wide scale. Actually, it is interesting that the number of poised promoters, which is estimated here (∼30% of the strongly bound sequences), roughly matches with the reported number of false positives in ChIP-chip experiments.[16] However, care must be taken when literally comparing false positives in ChIP-chip experiments with our *in silico* results, due to possible different choices of the binding thresholds. That is, the binding threshold is to a good degree provisional in ChIP-chip experiments, *i.e.* it depends on the signal intensity above which the sequences are considered to be targets. Therefore, the binding threshold is likely different from one ChIP-chip experiment to the other, and may also be different from the choice of binding threshold in our study. Consequently, it is likely that false positives in ChIP-chip experiments come from both sequences that are poised promoters and from technical issues such as biases in DNA amplification or imperfect immunoprecipitations of DNA fragments cross-linked to protein.

Furthermore, the importance of the kinetic effects strongly suggests that they should be incorporated in bioinformatic methods for TSS detection. In fact, TSS detection in bacteria is a classic bioinformatic problem, where available methods show poor accuracy.[18b,24,29] An alternative to current methods, which are based on information theory, is a biophysics method that would detect promoters based on the calculated transcription rate. A major difficulty in developing such a method is that interactions of σ^{70} with −35 element have (to our knowledge) not been measured until now. Note that in our *in silico* experiments we varied the −10 element, while the sequence of −35 element remained constant. While such a design is evidently useful for studying the interplay of physical interactions at the −10 element, it is not convenient for promoter detection, since promoters sample sequences with variable −35 elements. A solution to this problem may be a mixed bioinformatic and biochemical parameterization, which is our work that is currently in progress.

4 Conclusion

In this work, we investigated kinetic effects of transcription initiation on a genome-wide scale. Such analysis is, to our knowledge, the first of its kind, since there is currently no high-throughput method for measuring kinetic parameters of transcription initiation for sequences of interest. Consequently, the kinetic parameters have to be experimentally measured through classical, but time-consuming, τ-plot measurements, individually for each sequence of interest. To overcome this difficulty, we here used a quantitative model of transcription initiation, which showed a very good agreement with experimental data, and which allows efficient calculation of the kinetic parameters. The computational procedure also allowed repeatedly altering both specificities of σ^{70}–DNA interaction domains, and relevant DNA sequences, which is experimentally not feasible. We consequently designed a set of *in silico* experiments, which use a model of the specific biochemical process (transcription initiation), in order to study kinetics of transcription initiation on a much larger (whole genome) scale.

Through the *in silico* experiments we found that the extent of RNAP poising in the genome is highly suppressed, where this suppression is at the level of σ^{70} interaction domains, rather than the DNA sequence. However, despite this suppression, a significant fraction of the sequences that are strongly bound by RNAP correspond to poised promoters. This significant fraction of poised promoters is directly relevant for interpreting results of experimental and computational searches of transcription start sites. Furthermore, we surprisingly found that sequences of the functional promoters increase the extent of RNAP poising, which we interpreted in terms of the mix-and-match model of promoter recognition. Overall, the analysis presented here strongly suggests that the kinetic effects are important, and that they should be incorporated in methods for core promoter detection. It is likely that this will allow both increasing the accuracy of computational predictions and better understanding the results of the experimental searches.

5 Methods

5.1 Calculation of the kinetic parameters

To calculate the relevant kinetic parameters, we use a biophysical model of transcription initiation.[8] For completeness, in ESI,† we summarize elements of this model that are directly relevant for the analysis presented here. Briefly, the model is used to express the rate by which RNAP opens the two DNA strands, in terms of the interactions of σ^{70} with ssDNA and dsDNA, and DNA melting energies. To parameterize the model, we use a widely used independent nucleotide approximation,[30] according to which the interaction energies are given by the sum of the terms that correspond to different bases at different positions. Also, in this study we vary only the sequence of the −10 element, so that the energy terms that are associated with −35 element interactions and spacer lengths do not enter the relevant equations. Consequently, the binding affinity K_B, the rate of transition from the closed to open complex k_f, and the rate of transcription initiation φ are given below, respectively, by eqn (1), (2) and (3) (see ref. 8 and ESI†):

$$\log(K_\mathrm{B}(S)) \sim c - \sum_{i=1}^{6}\sum_{\alpha=1}^{4}\left(\Delta G_{i,\alpha}^{(\mathrm{ds})}\Big/k_\mathrm{B}T\right)S_{i,\alpha} \tag{1}$$

$$\log\left(k_\mathrm{f}\left(S_{(-10)}\right)\right) = c + \sum_{i=2}^{6}\sum_{\alpha=1}^{4}\left(\Delta G_{\alpha}^{(\mathrm{m})}\Big/k_\mathrm{B}T + \Delta G_{i,\alpha}^{(\mathrm{ds})}\Big/k_\mathrm{B}T - \Delta G_{i,\alpha}^{(\mathrm{ss})}\Big/k_\mathrm{B}T\right)S_{i,\alpha} \tag{2}$$

$$\log(\varphi(S)) = c + \sum_{i=1}^{6}\sum_{\alpha=1}^{4}\Delta G_{i,\alpha}^{(\mathrm{eff})}S_{i,\alpha} \tag{3}$$

where in the last equation we introduced the effective binding energy $\Delta G_{i,\alpha}^{(\mathrm{eff})}$:

$$\Delta G_{i,\alpha}^{(\mathrm{eff})} \equiv \begin{cases} \left(-\Delta G_{i,\alpha}^{(\mathrm{ss})} + \Delta G_{\alpha}^{(\mathrm{m})}\right)\Big/k_\mathrm{B}T \text{ for } i \in (2,6) \\ -\Delta G_{i,\alpha}^{(\mathrm{ds})}\Big/k_\mathrm{B}T \text{ for } i = 1 \end{cases} \tag{4}$$

In the equations above, the index i denotes different positions within the −10 box, so that i = 1 corresponds to the position −12, while i = 6 corresponds to the position −7, relative to the transcription start site. Further, α denotes the four different bases (A, T, C or G), while $S_{i,\alpha}$ is equal to one if base α is present at position i in sequence S, and is equal to zero otherwise. Furthermore, $\Delta G_{\alpha}^{(\mathrm{m})}$ denotes the melting energies of different bases, $\Delta G_{i\alpha}^{(\mathrm{ss})}$ denotes the interaction energies of σ with different bases at different positions of the non-template strand in the open complex, and $\Delta G_{i\alpha}^{(\mathrm{ds})}$ denotes the interaction energies of σ with different bases at different positions of duplex DNA for the −10 box. Note that the base −12 (i = 1) appears asymmetrically in the expression for the effective energy (see eqn (4)), since this is the only base of the −10 element that remains double stranded in the open complex.[6] Also, note that due to the symmetry of the two DNA strands $\Delta G_\mathrm{A}^{(\mathrm{m})} = \Delta G_\mathrm{T}^{(\mathrm{m})}$ and $\Delta G_\mathrm{C}^{(\mathrm{m})} = \Delta G_\mathrm{G}^{(\mathrm{m})}$, so that there are effectively two parameters that determine melting energy in the single nucleotide approximation.

5.2 Alignment of −10 promoter elements

To align −10 elements, we use the assembly of transcription start sites from RegulonDB database.[25] This assembly includes both experimentally verified promoters and computational predictions, and corresponds to both σ^{70} and alternative σ factors. For our alignment, we select only experimentally verified σ^{70} transcription start sites, *i.e.* we disregard all transcription start sites that are either not experimentally validated, or correspond to alternative σ factors. This selection results in the total of 342 σ^{70} transcription start sites, and we use the obtained start sites in order to extract DNA segments that correspond to positions −17 to −2, relative to the transcription start sites. These positions were chosen having in mind that the position of −10 element can deviate for 5 bps relative to its canonical position (−12 to −7).[31]

To identify the 6 bp long −10 elements within the selected DNA segments, we used the Gibbs sampler.[32] The Gibbs sampler implements a version of the Gibbs search algorithm,[33] which is used to perform unsupervised motif alignment. Only the DNA strand defined by the direction of transcription was searched, since both −10 box and −35 box motifs are not palindrome symmetric. The search was done with the initial assumption that one motif element is present in each DNA segment; however, in the end of the Gibbs sampler search, individual motif elements are added in or taken out, in a single pass of the algorithm, depending upon whether or not their inclusion improves the value of the alignment score. The last step allows excluding from the alignment those sequences that do not contain −10 box motifs, *e.g.* due to database miss-assignments. The search resulted in the identification of 322 aligned −10 boxes that correspond to the experimentally confirmed σ^{70} transcription start sites in *E. coli*; these aligned −10 elements were used in the further analysis.

5.3 Randomization of interaction specificities and DNA segments

We aim to randomize the interaction specificities, without changing the overall strength of σ^{70}–DNA interactions. To achieve this, it is useful to visualize the interaction parameters in the form of a matrix, where index i corresponds to different positions within the −10 element, while index α corresponds to four different bases. Overall interaction strength for energy matrix $\varepsilon_{i,\alpha}$ can be defined as $\sum_{i,\alpha} {\varepsilon_{i,\alpha}}^2$.[19b] Consequently, to randomize the interaction specificities, we randomly permute elements of the interaction matrix, which randomizes the interaction specificity but does not change $\sum_{i,\alpha} {\varepsilon_{i,\alpha}}^2$. In order to obtain statistics for quantities of interest, we randomize a given matrix 50 times, according to the procedure described above. In order to randomize the interactions corresponding to DNA melting, we simply permute energies that correspond to AT ($\Delta G_{\rm A}^{\rm (m)} = \Delta G_{\rm T}^{\rm (m)}$) and GC base pairs ($\Delta G_{\rm C}^{\rm (m)} = \Delta G_{\rm G}^{\rm (m)}$). This procedure results in a single randomization, and is a consequence of the fact that in the single nucleotide approximation there are only two parameters that describe DNA melting (see above).

We randomize DNA sequences, *i.e.* intergenic regions and −10 elements that correspond to the experimentally confirmed transcription start sites (see above), by randomly permuting the bases within the sequences. Note that such randomization preserves nucleotide (GC) content of the sequences. Similar to σ^{70}–DNA interaction domains, to obtain appropriate statistics we randomize a given DNA sequence 50 times.

5.4 Interaction parameters for *E. coli* transcription factors

We use protein–DNA interaction parameters that were obtained in ref. 19*b*. These interaction parameters were inferred from *E. coli* transcription factor binding sites which were assembled in DPInteract database.[24] The interaction parameters were inferred from the example binding sites by using the QPMEME (Quadratic Programming Method of Energy Matrix Estimation) algorithm.

To ensure a high accuracy of the inferred protein–DNA interaction parameters, we select those transcription factors (*i.e.* their corresponding interaction parameters), for which the following two conditions are satisfied: (i) the number of the example binding sites assembled in DPInteract database is larger than 10, (ii) over representation for the transcription factor is also larger than 10. The first condition ensures that too few example binding sites do not lead to overfitting of the interaction parameters. The second condition (over representation) is related to a measure of significance/functionality of the inferred interaction parameters.[19b] This procedure results in selection of the interaction parameters for eight *E. coli* transcription factors.

We then use the inferred interaction parameters for the selected *E. coli* transcription factors in order to substitute interaction specificities of σ2.3 (σ^{70}–ssDNA interactions) and σ2.4 (σ^{70}–dsDNA interactions) binding domains. A technical difficulty is that the length of σ2.3 and σ2.4 binding sites (5 bps and 6 bps, respectively) is generally different (shorter) than the length of binding sites of the selected *E. coli* transcription factors. To resolve this difficulty, we select a subset of adjacent positions that correspond to maximal binding specificity within the interaction domain of each transcription factor; the length of the selected adjacent positions corresponds to the length of σ2.3 or σ2.4 binding positions (*i.e.* 5 bps or 6 bps). To select the adjacent positions with maximal specificity, we use a definition of the binding specificity s_i at position i of the energy matrix $\varepsilon_{i,\alpha} : s_i = \sum_{\alpha} {\varepsilon_{i,\alpha}}^2$.

References

1 R. H. Ebright, RNA Polymerase: Structural Similarities Between Bacterial RNA Polymerase and Eukaryotic RNA Polymerase II, *J. Mol. Biol.*, 2000, **304**, 687–698.

2 S. Borukhov and E. Nudler, RNA polymerase holoenzyme: structure, function and biological implications, *Curr. Opin. Microbiol.*, 2003, **6**, 93–100.

3 M. S. B. Paget and J. D. Helmann, The sigma 70 family of sigma factors, *Genome Biol.*, 2003, **4**, 203–208.

4 P. L. DeHaseth, M. L. Zupancic and M. T. Record Jr, RNA polymerase-promoter interactions: the comings and goings of RNA polymerase, *J. Bacteriol.*, 1998, **180**, 3019–3025.

5 S. Borukhov and K. Severinov, Role of the RNA polymerase sigma subunit in transcription initiation, *Res. Microbiol.*, 2002, **153**, 557–562.

6 K. S. Murakami and S. A. Darst, Bacterial RNA polymerases: the wholo story, *Curr. Opin. Struct. Biol.*, 2003, **13**, 31–39.

7 A. Niedziela-Majka and T. Heyduk, *Escherichia coli* RNA Polymerase Contacts outside the -10 Promoter Element Are Not Essential for Promoter Melting, *J. Biol. Chem.*, 2005, **280**, 38219.

8 M. Djordjevic and R. Bundschuh, Formation of the Open Complex by Bacterial RNA Polymerase—A Quantitative Model, *Biophys. J.*, 2008, **94**, 4233–4248.

9 A. Revyakin, R. H. Ebright and T. R. Strick, Promoter unwinding and promoter clearance by RNA polymerase: detection by

single-molecule DNA nanomanipulation, *Proc. Natl. Acad. Sci. U. S. A.*, 2004, **101**, 4776–4780.

10 J. Santa Lucia Jr, A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics, *Proc. Natl. Acad. Sci. U. S. A.*, 1998, **95**, 1460–1465.

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

12 K. S. Murakami, S. Masuda, E. A. Campbell, O. Muzzin and S. A. Darst, Structural Basis of Transcription Initiation: An RNA Polymerase Holoenzyme–DNA Complex, *Science*, 2002, **296**, 1285–1290.

13 S. J. Lee and J. D. Gralla, Osmo-regulation of bacterial transcription *via* poised RNA polymerase, *Mol. Cell*, 2004, **14**, 153–162.

14 T. I. Lee, N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph, G. K. Gerber, N. M. Hannett, C. T. Harbison, C. M. Thompson, I. Simon, J. Zeitlinger, E. G. Jennings, H. L. Murray, D. B. Gordon, B. Ren, J. J. Wyrick, J. B. Tagne, T. L. Volkert, E. Fraenkel, D. K. Gifford and R. A. Young, Transcriptional regulatory networks in *Saccharomyces cerevisiae*, *Science*, 2002, **298**, 799–804.

15 D. S. Johnson, A. Mortazavi, R. M. Myers and B. Wold, Genome-wide mapping of *in vivo* protein–DNA interactions, *Science*, 2007, **316**, 1497–1502.

16 (*a*) D. C. Grainger, D. Hurd, M. Harrison, J. Holdstock and S. J. Busby, Studies of the distribution of *Escherichia coli* cAMP-receptor protein and RNA polymerase along the *E. coli* chromosome, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 17693–17698; (*b*) C. D. Herring, M. Raffaelle, T. E. Allen, E. I. Kanin, R. Landick, A. Z. Ansari and B. O. Palsson, Immobilization of *Escherichia coli* RNA polymerase and location of binding sites by use of chromatin immunoprecipitation and microarrays, *J. Bacteriol.*, 2005, **187**, 6166–6174.

17 D. K. Hawley, A. D. Johnson and W. R. McClure, Functional and physical characterization of transcription initiation complexes in the bacteriophage lambda OR region, *J. Biol. Chem.*, 1985, **260**, 8618–8626.

18 (*a*) G. D. Stormo and D. S. Fields, Specificity, free energy and information content in protein–DNA interactions, *Trends Biochem. Sci.*, 1998, **23**, 109–113; (*b*) A. M. Huerta and J. Collado-Vides, Sigma 70 Promoters in *Escherichia coli*: Specific Transcription in Dense Regions of Overlapping Promoter-like Signals, *J. Mol. Biol.*, 2003, **333**, 261–278.

19 (*a*) J. M. Vilar, Accurate prediction of gene expression by integration of DNA sequence statistics with detailed modeling of transcription regulation, *Biophys. J.*, 2010, **99**, 2408–2413; (*b*) M. Djordjevic, A. M. Sengupta and B. I. Shraiman, A biophysical approach to transcription factor binding site discovery, *Genome Res.*, 2003, **13**, 2381–2390; (*c*) M. Djordjevic and A. M. Sengupta, Quantitative modeling and data analysis of SELEX experiments, *Phys. Biol.*, 2006, **3**, 13–28.

20 W. R. McClure, Rate-Limiting Steps in RNA Chain Initiation, *Proc. Natl. Acad. Sci. U. S. A.*, 1980, **77**, 5634–5638.

21 I. G. Hook-Barnard and D. M. Hinton, Transcription initiation by mix and match elements: flexibility for polymerase binding to bacterial promoters, *Gene Regul. Syst. Biol.*, 2007, **1**, 275.

22 (*a*) J. E. Stefano and J. D. Gralla, Mutation-induced changes in RNA polymerase–lac ps promoter interactions, *J. Biol. Chem.*, 1982, **257**, 13924–13929; (*b*) G. Zocchi and K. Sneppen, *Physics in Molecular Biology*, Cambridge University Press, Cambridge, 2005.

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

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

25 S. Gama-Castro, H. Salgado, M. Peralta-Gil, A. Santos-Zavaleta, L. Muñiz-Rascado, H. Solano-Lira, V. Jimenez-Jacinto, V. Weiss, J. S. García-Sotelo and A. López-Fuentes, RegulonDB version 7.0: transcriptional regulation of *Escherichia coli* K-12 integrated within genetic sensory response units (Gensor Units), *Nucleic Acids Res.*, 2011, **39**, D98.

26 M. Djordjevic, Redefining *Escherichia coli* sigma(70) promoter elements: −15 motif as a complement of the −10 motif, *J. Bacteriol.*, 2011, **193**, 6305–6314.

27 R. Wagner, *Transcription regulation in prokaryotes*, Oxford University Press, Oxford, 2000.

28 (*a*) I. Hook-Barnard, X. B. Johnson and D. M. Hinton, *Escherichia coli* RNA polymerase recognition of a sigma 70-dependent promoter requiring a −35 DNA element and an extended −10 TGn motif, *J. Bacteriol.*, 2006, **188**, 8352–8359; (*b*) B. Thouvenot, B. Charpentier and C. Branlant, The strong efficiency of the *Escherichia coli* gapA P1 promoter depends on a complex combination of functional determinants, *Biochem. J.*, 2004, **383**, 371–382.

29 (*a*) L. Argaman, R. Hershberg, J. Vogel, G. Bejerano, E. Wagner, H. Margalit and S. Altuvia, Novel small RNA-encoding genes in the intergenic regions of *Escherichia coli*, *Curr. Biol.*, 2001, **11**, 941–950; (*b*) S. Burden, Y. Lin and R. Zhang, Improving promoter prediction for the NNPP2. 2 algorithm: a case study using *Escherichia coli* DNA sequences, *Bioinformatics*, 2005, **21**, 601–607; (*c*) M. Towsey, J. Hogan, S. Mathews and P. Timms, The *in silico* Prediction of Promoters in Bacterial Genomes, *Genome Inf.*, 2008, **21**, 178–189.

30 (*a*) P. V. Benos, M. L. Bulyk and G. D. Stormo, Additivity in protein–DNA interactions: how good an approximation is it?, *Nucleic Acids Res.*, 2002, **30**, 4442–4451; (*b*) Y. Zhao and G. D. Stormo, Quantitative analysis demonstrates most transcription factors require only simple models of specificity, *Nat. Biotechnol.*, 2011, **29**, 480–483.

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

32 W. Thompson, E. C. Rouchka and C. E. Lawrence, Gibbs Recursive Sampler: finding transcription factor binding sites, *Nucleic Acids Res.*, 2003, **31**, 3580–3585.

33 C. Lawrence and S. Altschul, Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment, *Science*, 1993, **262**, 208.