Author's Accepted Manuscript

Modeling promoter search by E. Coli RNA polymerase: One-dimensional diffusion in a sequence-dependent energy landscape

Johanna Weindl, Zaher Dawy, Pavol Hanus, Juergen Zech, Jakob C. Mueller

PII: S0022-5193(09)00232-X
DOI: doi:10.1016/j.jtbi.2009.05.006
Reference: YJTBI5562

www.elsevier.com/locate/jtbi

To appear in: *Journal of Theoretical Biology*

Received date: 4 February 2009
Revised date: 12 May 2009
Accepted date: 12 May 2009

Cite this article as: Johanna Weindl, Zaher Dawy, Pavol Hanus, Juergen Zech and Jakob C. Mueller, Modeling promoter search by E. Coli RNA polymerase: One-dimensional diffusion in a sequence-dependent energy landscape, *Journal of Theoretical Biology* (2009), doi:[10.1016/j.jtbi.2009.05.006](https://doi.org/10.1016/j.jtbi.2009.05.006)

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting galley proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Modeling Promoter Search by *E. coli* RNA Polymerase: One-Dimensional Diffusion in a Sequence-Dependent Energy Landscape

Johanna Weindl¹, Zaher Dawy², Pavol Hanus¹,
Juergen Zech³ and Jakob C. Mueller⁴

¹ Institute for Communications Engineering
Technische Universität München

Arcisstraße 21, 80290 München, Germany

Email: johanna.weindl@mytum.de, pavol.hanus@tum.de

Tel: 0049 - 89 28925002, Fax: 0049 - 89 28923490

² Department of Electrical and Computer Engineering
American University of Beirut

P.O.Box 11-0236, Riad El-Solh, Beirut, Lebanon

Email: zaher.dawy@aub.edu.lb

³ MRC Clinical Sciences Centre
Imperial College London

Du Cane Road, London W12 0NN, UK

Email: juergen.zech@csc.mrc.ac.uk

⁴ Department of Behavioral Ecology and Evolutionary Genetics

Max Planck Institute for Ornithology

Seewiesen, 82305 Starnberg, Germany

Email: mueller@orn.mpg.de

Corresponding author: Johanna Weindl

Abstract

We present a biophysical model of promoter search by *E. coli* RNA polymerase. We use an unconventional weight matrix derived from promoter strength data to extract the energy landscape common to a large set of known promoters. This exhibits a continuous strengthening of the binding energy when approaching the transcription start site from either side. During promoter search, the RNA polymerase slides along the DNA double helix (one-dimensional diffusion) after randomly binding to it. We discuss the possibility that the sliding has a sequence-dependent component, which implies that the energy landscape influences the movement with respect to speed, direction and efficiency. Based on this assumption, we relate the obtained energy landscape around the promoters to the one-dimensional diffusion of the RNA polymerase. Our analytical results suggest that the sequence-dependent random walk slows down and gets directed upon entering a region of 500 bp around the transcription start site, which significantly increases the efficiency of promoter search. These results may explain how the RNA polymerase is able to find the promoter in biologically relevant times out of a vast excess of non-target sites. Moreover, they provide evidence for a sequence-dependent component of one-dimensional diffusion.

Key words: sliding model; facilitated target location; binding energy; protein-DNA interaction; random walk.

1 Introduction

The exact details of promoter detection by the prokaryotic RNA polymerase (RNAP) holoenzyme or, more generally, how DNA-binding proteins find their cognate site has been in the focus of research for decades now. More than 30 years ago, Riggs et al. (1970) measured the association rate of the *E. coli lac* repressor to its operator in the DNA to be much higher than the rate achievable by three-dimensional diffusion. In 1981, Berg, Winter, and von Hippel published a seminal series of articles presenting a theory for protein-DNA interactions which provided a first explanation for this faster-than-diffusion search (Berg et al., 1981; Winter et al., 1981). They conjectured that the dimensionality changes while the protein searches for its target site: The protein at first randomly binds to the DNA in a round of three-dimensional diffusion through the cell and subsequently moves along it in a process of one-dimensional diffusion ("facilitated diffusion"). On short ranges, the one-dimensional search round was assumed to be a sliding process along the double helix. However, the assumed random walk nature of sliding with the same probability for stepping forward and backward (1D Brownian motion) appeared to make this process very inefficient (Widom, 2005). Two additional mechanisms of facilitated target location were suggested to supplement sliding, namely hopping and intersegment transfer (see (von Hippel and Berg, 1989) and references therein). Experiments were designed to investigate the mechanisms involved in target search and to observe them in real-time, which led to controversial results. Ricchetti et al. (1988) were the first to present experimental results supporting sliding of the *lac* repressor along

the DNA. In 1993, the process could first be visualized for the RNAP (Kabata et al., 1993), and the possible sliding length was later estimated to lie between 350 and 1000 bp (Shimamoto, 1999). Opposed to that, Gowers and his co-workers presented experimental results for the heterodimeric restriction enzyme BbvGI proposing that translocation over distances of more than 30 base pairs in conditions of physiological salt concentration always includes at least one dissociation/rebinding event (Gowers et al., 2005; Widom, 2005). Recently, Blainey et al. (2006) reported molecules of the human oxoguanine DNA glycosylase 1 (hOgg1) rapidly diffusing along stretches of DNA and computed the mean sliding length at 440 bp for high salt concentrations. As pointed out by these few contributions, a detailed understanding of one-dimensional diffusion and the possible sliding length could not yet be obtained.

In the case of the RNAP holoenzyme, it still remains disputed how it is able to find the promoter in biologically relevant times out of a vast excess of non-cognate sites if considering a random walk nature of the sliding process without an influence of the bound DNA sequence. Von Hippel suggested that DNA-binding proteins might have two distinct conformational states with different modes of binding (von Hippel and Berg, 1989). In the search mode, the protein has a weaker and non-specific binding to the DNA which enables fast sliding over a smooth energy landscape. In the recognition mode, it binds tightly to the DNA sites. Solution structures for both conformations of a dimeric *lac* repressor DNA binding domain were presented in (Kalodimos et al., 2004). Transition between the modes is proposed to occur at low-energy sites where the protein is trapped for a

certain time (Slutsky and Mirny, 2004; Hu et al., 2007). However, in order to be sensitive to these positions, the protein needs to interact sequence-dependently with the DNA also in the search mode, even though weaker than in the recognition mode and possibly not via hydrogen bonds with the bases. Therefore, it was concluded by several researchers that the search mode also has a sequence-dependent component (Barbi et al., 2004a; Bruinsma, 2002; Slutsky et al., 2004). This can for example be based on helix stability/geometry (von Hippel and Berg, 1986) or molecular interactions like van der Waals contacts and specific electrostatic bonds. Bruinsma (2002) even reported for the *lac* repressor that the equilibrium constant for the search mode varies over two orders of magnitude when the underlying DNA sequence is varied. Further evidence for the theory of a sequence-dependent sliding is given by the following contributions:

- 1) Sakata-Sogawa and Shimamoto (2004) showed that the RNAP tracks a DNA groove during its sliding process which facilitates the scanning of DNA sequences;
- 2) Erie et al. (1994) used scanning force microscopy to demonstrate that the *Cro* repressor protein bends the DNA both at its operator as well as at random sites. Since local DNA bendability is known to vary strongly based on the underlying sequence, it is likely to constitute a sequence-dependent component of non-specific protein-DNA interactions;
- 3) Harada et al. (1999) showed for λ -phage DNA that the on-site time of the RNAP for AT-rich sequences is higher than for GC-rich sequences, which thus constitutes a sequence-dependent parameter of the sliding process;
- 4) Jeltsch et al. (1994) showed that the sliding process of the restriction endonuclease EcoRI is slowed down when encountering sites on

the DNA that resemble its recognition site. They conjectured that the protein undergoes a continuous transition between the search mode and the recognition mode depending on the similarity between the bound sequences and the target sequence.

In the following, we present an approach of modeling prokaryotic promoter search performed by the RNAP holoenzyme. Our objective is to investigate the diverging theories about the nature of sliding (1D Brownian motion vs. sequence-dependent random walk) and to rate the influence of the promoter surrounding on the search process. For this purpose, we derive the average energy landscape around 651 experimentally verified promoters of *E. coli* by extrapolating a binding energy estimate obtained from promoter strength data (Section 2). Thereafter, we analytically relate the obtained binding energies to kinetic parameters of the sliding process (assuming a sequence-dependent random walk) such as speed, direction and efficiency (Section 3). The results suggest that the movement of the RNAP is slowed down and directed towards the promoter upon entering a region of 500 bp around the transcription start site. As we subsequently discuss (Section 4), our calculations not only present evidence for a sequence-dependent component of the sliding process of the RNAP during promoter search but also imply that the movement is guided to the promoter.

2 Materials and Methods

In 1999, it was impressively visualized by scanning force microscopy that the RNAP searches its target site (the promoter) by randomly binding the DNA and subsequently sliding along several hundred base pairs before dissociating or moving to another site through hopping or intersegment transfer (Bustamante et al., 1999; Guthold et al., 1999). The sliding process was assumed to be a 1D Brownian motion along the DNA. In contrast to that, the following considerations are based on the assumption that the sliding process has a sequence-dependent component, i.e., that the RNAP holoenzyme encounters a specific energy landscape that depends on the bound DNA sequence (see also discussion in Section 1). Consequently, it does not perform a random walk with equal probabilities of stepping forward or backward, but the sliding is influenced by the binding energy at each position. In this case, the transition rates $w_{i,i+1}$ and $w_{i,i-1}$ from site i to site $i + 1$ and site $i - 1$, respectively, depend on the binding energies E between RNAP and the DNA at these sites through

$$w_{i,i\pm 1} = \nu \times \begin{cases} e^{-\beta[E(i\pm 1)-E(i)]} & \text{if } E(i\pm 1) > E(i) \\ 1 & \text{otherwise} \end{cases}, \quad (1)$$

with $\beta = (k_B T)^{-1}$, where ν denotes the affective attempt frequency, k_B the Boltzmann constant, and T the ambient temperature in Kelvin (Slutsky and Mirny, 2004). According to Eq. 1, known as the regular activated transport form or Arrhenius equation, the transition rate $w_{i,i\pm 1}$ is under constant conditions solely dependent on the difference between $E(i\pm 1)$ and $E(i)$, which corresponds to the gradient g of the function $E(i)$. The

constant rate ν of $w_{i,i\pm 1}$ in the case $E(i \pm 1) \leq E(i)$ is used since any thermal fluctuation will push the RNAP independently of the value $\Delta E = E(i \pm 1) - E(i)$. Hence, it steps uphill (i.e., to a site with higher energy) with a rate smaller than ν (depending on ΔE) and steps downhill (i.e., to a site with lower energy) with constant rate ν .

2.1 Average energy landscape

In order to obtain the binding energy values $E(i)$ between the *E. coli* RNAP and the DNA around the promoters, we use a weight matrix derived from promoter strength data from (Kiryu et al., 2005). The core promoter elements in prokaryotes are two conserved hexamers, the -35 and the -10 region, separated by a variable spacing of 15 to 21 bp. The promoter strength, i.e., the promoter's ability to initiate transcription of the corresponding gene, is mainly determined by the binding energy between the sigma subunit of the RNAP holoenzyme and the promoter, which depends on the DNA sequence and the spacing between the -35 and the -10 promoter region (Kobayashi et al., 1990). This binding energy $E(i_p)$ between the sigma factor and a promoter p at position i_p can be written as

$$E(i_p) = \underbrace{\sum_{k=1}^6 e(n_k, k)}_{\text{-35 region}} + \underbrace{\sum_{k=7}^{12} e(n_{k+s}, k)}_{\text{-10 region}} + \underbrace{\epsilon(s_p)}_{\text{Spacing}}, \quad (2)$$

where $e(n, k)$ denotes the partial binding energy between the nucleotide $n \in \{A, C, G, T\}$ and the binding site of the sigma factor associated with promoter position k . Thus, n_k and n_{k+s} refer to the nucleotides at positions k and $k + s$, respectively, of a given promoter sequence p . The term $\epsilon(s_p)$ is

the contribution of the spacing s between the promoter regions to the binding energy. This represents the energy that the kidney-shaped sigma factor needs to expand or compress in order to bind to promoters with non-ideal spacing (i.e., more or less than 17 bp) (Dombroski et al., 1996; Mulligan et al., 1985).

2.1.1 Weight matrix

The values of $e(n, k)$ and $\epsilon(s)$ are extracted from (Kiryu et al., 2005). The former are used to build up a $[4 \times 12]$ weight matrix $\mathbf{W}$ containing the contribution of the 12 nucleotides in the promoter regions to the binding energy:

$$\mathbf{W} = \begin{bmatrix} e(A, k=1) & e(A, k=2) & \dots & e(A, k=11) & e(A, k=12) \\ e(C, k=1) & e(C, k=2) & \dots & e(C, k=11) & e(C, k=12) \\ e(G, k=1) & e(G, k=2) & \dots & e(G, k=11) & e(G, k=12) \\ e(T, k=1) & e(T, k=2) & \dots & e(T, k=11) & e(T, k=12) \end{bmatrix}.$$

The first six columns of $\mathbf{W}$ reference the positions in the -35 region and the last six columns reference those in the -10 region. The values from (Kiryu et al., 2005) were obtained by extracting the relation between promoter sequences and the binding energy by applying support vector regression on gene expression data. There, Kiryu et al. defined the partial binding energies such that positive values indicate a strengthening effect on the overall binding energy, whereas negative values imply a weakening effect. However, since in chemistry binding energies underlying stable interactions are generally given by negative values and, thus, high negative overall energies should indicate candidate target sites (Sengupta et al., 2002), we

multiply all values by -1 . The resulting values of the weight matrix $\mathbf{W}$ can be found in (Weindl et al., 2007). We apply $\mathbf{W}$ using a sliding window that is shifted in single-nucleotide steps over the DNA. As mentioned before, the sigma factor is capable of expanding or compressing and hereby adapting to different promoter spacings s in order to bind to the energetically most favorable site. Therefore, the binding energy $E(i)$ at position i of an analyzed DNA sequence is obtained by minimizing the energy score calculated according to Eq. 2 over all possible spacings $s \in [15; 19]$.

2.1.2 Application to *E. coli* promoter sequences

Our objective is to extract the energy landscape common to the surroundings of experimentally verified promoters. Therefore, we apply the described algorithm "on average", i.e., to a set of N known promoters of σ^{70} (the main sigma factor, responsible for expression of housekeeping genes), and calculate the arithmetic mean (average) of the resulting values $E_l(i)$ for each position i , where the index l references the l -th promoter ($l \in [1; N]$). Additionally, we apply the algorithm to 10,000 random sequences and normalize the energy measure obtained through averaging over all $E_l(i)$ by the resulting value $\overline{E}_{\text{ran}} = -0.4423$ (see (Weindl et al., 2007) for more details). Due to this normalization by $\overline{E}_{\text{ran}}$, positive values of $\overline{E}(i)$ indicate a below-average binding strength, while negative values refer to an above-average binding strength between sigma factor and DNA sequence. In our previous work (Weindl et al., 2007), we showed that the application of $\mathbf{W}$ to a large set of σ^{70} -promoters exhibits a characteristic energy landscape. Interestingly, the binding energies were demonstrated to

decrease constantly beginning 300 bp before the promoter and to increase again in the 200 bp after the promoter. For a discussion as well as a thorough verification of the characteristic energy landscape see (Weindl et al., 2007). Figure 1 shows the average binding energy $\overline{E}(i)$ around 651 experimentally verified σ^{70} -promoters downloaded from RegulonDB (version 5.7, see (Salgado et al., 2006)) along with its linear approximation by four straight lines with different gradients g . Region 1 and region 4 correspond to random sequences, since $g = 0$ and $\overline{E}(i) = 0$. In region 2 and region 3, however, we observe a negative and positive gradient g , respectively. The properties of the promoter site itself with respect to detection by the RNA polymerase were investigated in (Weindl and Hagenauer, 2007).

3 Results

In Section 2, we derived the average energy landscape around experimentally verified σ^{70} -promoters of *E. coli*. The obtained energy values and the presented linear approximation (see Figure 1) allow us to relate the obtained binding energies to kinetic parameters of the sequence-dependent sliding process of the RNAP holoenzyme along the DNA based on Eq. 1.

3.1 Speed of sliding

In the sliding model, the protein is assumed to slide in single nucleotide steps along the DNA. The escape rate of the protein at site i to one of the neighboring sites and, therefore, the speed of the sliding process is given by

$$r_i = \frac{1}{\tau_i} = w_{i,i+1} + w_{i,i-1} = \nu(e^{-\beta|g|} + 1), \quad (3)$$

where τ_i denotes the time the protein spends bound to site i (Slutsky and Mirny, 2004). Refer to Appendix A for a detailed derivation of Eq. 3.

Applying Eq. 3 to the four regions of our approximation yields the following estimations for r_i :

$$\text{Region 1: } |g| = 0 \rightarrow r_i = 2\nu.$$

$$\text{Region 2: } |g| > 0 \rightarrow r_i < 2\nu.$$

$$\text{Region 3: } |g| > 0 \rightarrow r_i < 2\nu.$$

$$\text{Region 4: } |g| = 0 \rightarrow r_i = 2\nu.$$

It can be seen that the escape rate r_i is lower in the direct surrounding of the promoters than on random sequences ($|g| = 0$). Consequently, the on-site time τ_i of the protein is higher and thus the speed of the sliding

process is lower than on random sequences. This is consistent with results by Jeltsch et al. (1994) who showed that the sliding process of the restriction endonuclease EcoRI is slowed down when encountering sites on the DNA that resemble its recognition site. Since our energy landscape reflects an increasing sequence similarity to the promoter when approaching the transcription start site, the presented calculations of the escape rate are in agreement with the experimental results for EcoRI.

3.2 Direction of sliding

At each site i , the protein eventually escapes to the site $i + 1$ with probability p_i and to the site $i - 1$ with probability $q_i = 1 - p_i$. This probability p_i depends on the transition rates $w_{i,i+1}$ and $w_{i,i-1}$ through (Slutsky and Mirny, 2004)

$$p_i = \frac{w_{i,i+1}}{w_{i,i+1} + w_{i,i-1}} = \frac{w_{i,i+1}}{r_i}. \quad (4)$$

Applying Eq. 4 to our approximation of $\overline{E}(i)$ yields the following estimations of the transition probability p_i :

Region 1: $w_{i,i+1} = \nu$ and $w_{i,i-1} = \nu \rightarrow p_i = 0.5, q_i = 0.5$.

Region 2: $w_{i,i+1} = \nu$ and $w_{i,i-1} < \nu \rightarrow p_i > 0.5, q_i < 0.5$.

Region 3: $w_{i,i+1} < \nu$ and $w_{i,i-1} = \nu \rightarrow p_i < 0.5, q_i > 0.5$.

Region 4: $w_{i,i+1} = \nu$ and $w_{i,i-1} = \nu \rightarrow p_i = 0.5, q_i = 0.5$.

Note that the transition probability p_i from site i to site $i + 1$ increases when approaching the promoter regions from upstream and decreases when leaving the promoter regions. In case the RNAP approaches the promoter

from downstream, the probability q_i increases upon entering region 3. Thus, in both cases the surrounding of the transcription start site seems to direct the sliding towards the target site. Guthold et al. (1999) showed that the RNAP has no defined direction of sliding when binding to promoterless DNA fragments, which is consistent with the calculated values of p_i and q_i for region 1 and region 4.

3.3 Efficiency of promoter search

It was shown in previous studies (Bustamante et al., 1999; Guthold et al., 1999) that the RNAP slides forward and backward several times during its search for the promoters. The efficiency of the sliding process can be regarded as the number of steps the protein needs to reach from one site to another. Therefore, we use the mean first-passage time (MFPT), which is defined as the mean number of steps the protein will make to slide from site $i = 0$ to site $i = L$ assuming a certain set of transition probabilities $\{p_i\}$ (see (Murthy and Kehr, 1989) for more information). The mean first-passage time $\bar{t}_{0,L}$ is given by

$$\bar{t}_{0,L} = L + \sum_{k=0}^{L-1} \alpha_k + \sum_{k=0}^{L-2} \sum_{i=k+1}^{L-1} (1 + \alpha_k) \prod_{j=k+1}^i \alpha_j, \quad (5)$$

where $\alpha_i = q_i/p_i$ (see (Slutsky and Mirny, 2004) and references therein).

For negative values of L (i.e., a sliding backward), the definition of α changes to $\alpha_i = p_i/q_i$ and L in Eq. 5 has to be exchanged for its absolute value $|L|$. If approximating the characteristic energy landscape by four straight lines as depicted in Figure 1, the values of α_i are constant over sufficiently wide ranges. Therefore, we assume $\alpha_0 = \alpha_1 = \dots = \alpha_L := \alpha$ in

each region, which leads after a straightforward simplification of Eq. 5 (see Appendix B) to

$$\bar{t}_{0,L}' = L + L \times \alpha + (1 + \alpha) \sum_{k=1}^{L-1} (L - k) \alpha^k, \quad (6)$$

i.e., the mean first-passage time $\bar{t}_{0,L}'$ for our approximation is given by a polynomial of order L depending on α . Figure 2 shows the MFPT for $L = 50$. It can be seen that even small decreases of α from the random walk case with equal transition probabilities ($\alpha = 1$, i.e., $p_i = q_i = 0.5$) significantly decrease the mean first-passage time. At the same time, small increases of α lead to dramatic increases of $\bar{t}_{0,L}'$. For the four regions of our approximation and for $L = 50$ ($\alpha_i = q_i/p_i$), this can be summarized by:

$$\text{Region 1: } p_i = 0.5, \quad \text{i.e., } \alpha = 1 \rightarrow \bar{t}_{0,L}' = 2550.$$

$$\text{Region 2: } p_i > 0.5, \quad \text{i.e., } \alpha < 1 \rightarrow \bar{t}_{0,L}' \downarrow$$

$$\text{Region 3: } p_i < 0.5, \quad \text{i.e., } \alpha > 1 \rightarrow \bar{t}_{0,L}' \uparrow$$

$$\text{Region 4: } p_i = 0.5, \quad \text{i.e., } \alpha = 1 \rightarrow \bar{t}_{0,L}' = 2550.$$

For $L = -50$ ($\alpha_i = p_i/q_i$), i.e., reaching 50 positions backward of position i , we obtain the following values of $\bar{t}_{0,L}'$:

$$\text{Region 1: } p_i = 0.5, \quad \text{i.e., } \alpha = 1 \rightarrow \bar{t}_{0,L}' = 2550.$$

$$\text{Region 2: } p_i > 0.5, \quad \text{i.e., } \alpha > 1 \rightarrow \bar{t}_{0,L}' \uparrow$$

$$\text{Region 3: } p_i < 0.5, \quad \text{i.e., } \alpha < 1 \rightarrow \bar{t}_{0,L}' \downarrow$$

$$\text{Region 4: } p_i = 0.5, \quad \text{i.e., } \alpha = 1 \rightarrow \bar{t}_{0,L}' = 2550.$$

The symbols $\downarrow$ and $\uparrow$ denote a slight decrease and a strong increase,

respectively. The value $\bar{t}'_{0,L} = 2550$ shows clearly the inefficiency of the sliding process on random sequences: According to this calculation, the RNAP needs on average 2550 steps to bridge a distance of only 50 bases. Apparently, the decrease of binding energies the RNAP faces when approaching the promoters strongly influences the efficiency of promoter search. In case the RNAP approaches the promoter from downstream, region 3 increases the efficiency in direction of the promoter. If approaching the promoter from upstream, the search becomes more efficient upon entering region 2. Hence, the RNAP is directed towards the promoter from either side. Opposed to that, it seems to be nearly impossible for the RNAP to move further downstream if the promoter has been missed (entering region 3 from upstream), i.e., if no transcription initiation has taken place. Since it is known that the RNAP is able to slide backward, this suggests that the RNAP is potentially guided back to the promoter in case it missed the detection. It is important to note that in reality the assumption of constant α -values does not hold true. Instead, it may vary strongly from position to position and the protein can get trapped at positions with low energy. Nonetheless, the results visualize the huge difference between the promoter surrounding and random sequences, where the trapping effect occurs in the same way.

4 Discussion

Based on the results obtained by analyzing the characteristic energy landscape in the surrounding of σ^{70} -promoters (Figure 1), we can draw several conclusions about the kinetics of promoter detection in *E. coli*. In Section 3.1, we showed that the escape rate r and thus the speed of sliding is lower in the surrounding of the promoter regions, which implies that the movement of the RNAP holoenzyme along the DNA slows down when approaching the promoter regions. In Section 3.2, we showed that the probability p of moving from site i to site $i + 1$ increases some hundred base pairs upstream of the promoter regions evincing that the sequence-dependent interaction between sigma factor and promoter surrounding directs the RNAP holoenzyme towards the promoter as soon as it enters the respective region. In Section 3.3, we showed that the mean first-passage time decreases when approaching the promoter regions proposing that promoter search becomes more efficient in this region. Interestingly, the dramatic increase of the mean first-passage time after the promoter regions indicates a mechanism of guiding the RNAP back to the promoter in case it was missed. These results suggest that the surrounding of the promoters contains important information to guide the RNAP and its sigma subunit in a way increasing the probability of transcription initiation by slowing down the sliding process and controlling the direction and efficiency of the movement.

Further evidence for our results can be derived from earlier publications: Since the consensus sequences of the -35 and the -10 region are highly AT-rich, our binding energy score correlates positively with the

AT-content (Weindl et al., 2007). Harada et al. (1999) showed that the dissociation rate of bacterial RNA polymerase from AT-rich sequences is lower than from GC-rich sequences. Lower dissociation from the currently bound position indicates longer on-site times τ of the protein which implies a slow-down of the sliding process along the DNA. Furthermore – as mentioned in Section 1, bending of the DNA by proteins was shown to constitute a sequence-dependent component of non-specific DNA-interactions (see (Erie et al., 1994)). Hosid et al. (2004) later demonstrated that bendability is associated with periodicities of AA/TT dinucleotides in intergenic regions of *E. coli*. Since our energy landscape correlates with the AT-content, it should thus exhibit this reported periodicity. In fact, the observed energy landscape (Figure 1) shows a periodicity of ~ 10 and ~ 20 bps, respectively, in regions 2 and 3, while no clear periodicity is observed outside the promoter flanking regions (regions 1 and 4). Moreover, Bruinsma (2002) speculates about a possible slowing down of the *lac* repressor on DNA sequences that structurally resemble the cognate site. He proposes that this might give the protein more time to test if the currently bound sequence is the operator sequence. Additionally, Jeltsch et al. (1994) showed that the sliding process of the restriction endonuclease EcoRI is slowed down when encountering sites on the DNA that resemble its recognition site. Since the presented energy landscape reflects an increasing sequence homology to the consensus sequence if approaching the promoter from either side, the conclusions of Jeltsch et al. can be considered as experimental evidence for the calculated sliding parameters. Furthermore, Barbi et al. (2004b) showed through simulations

that the sliding of proteins along the DNA is a process of anomalous subdiffusion depending on a set of trapping sites. A similar interpretation was recently derived for three-dimensional diffusion by Saxton (2007). As presented in our previous work (Weindl et al., 2007), the characteristic energy landscape is only observed around strong promoters. Considering our results for the movement of the RNAP along the DNA, this suggests that nature makes use of the surrounding in addition to the promoter sequence itself as another mechanism for determining the expression rate of the respective gene. In the case of weak promoters, the RNAP might only rarely be successful in promoter detection, whereas for strong promoters the surrounding guides the RNAP to the transcription start site.

It is important to note that transcription initiation involves several other factors that were not taken into account in our algorithm. To name but few, we did not consider occupancy of the DNA by ligands (Flyvbjerg et al., 2006) and possible interruptions of the sliding process through other ways of facilitated target location (hopping, intersegment transfer) (Halford and Marko, 2004). However, the obtained results should give a strong indication for future biological experiments. Due to the similarity of main sigma factors in prokaryotes (Gruber and Gross, 2003; Paget and Helmann, 2003), the results are expected to apply in the same way to other prokaryotic organisms.

Appendix A: Escape Rate

The escape rate of a protein at site i is given by (Slutsky and Mirny, 2004)

$$r_i = \frac{1}{\tau_i} = w_{i,i+1} + w_{i,i-1}, \quad (\text{A-1})$$

where $w_{i,i+1}$ and $w_{i,i-1}$ are defined by the Arrhenius equation as

$$w_{i,i\pm 1} = \nu \times \begin{cases} e^{-\beta[E(i\pm 1) - E(i)]} & \text{if } E(i \pm 1) > E(i) \\ 1 & \text{otherwise} \end{cases}. \quad (\text{A-2})$$

We now do a case differentiation for the gradient g in order to derive a closed expression for the escape rate r_i :

Case 1: $E(i+1) > E(i)$, i.e., $g = E(i+1) - E(i) > 0$

$$\begin{aligned} r_i &= w_{i,i+1} + w_{i,i-1} \\ &= \nu \times e^{-\beta[E(i+1) - E(i)]} + \nu \\ &= \nu \times e^{-\beta g} + \nu \\ &= \nu(e^{-\beta g} + 1). \end{aligned} \quad (\text{A-3})$$

Case 2: $E(i+1) < E(i)$, i.e., $g = E(i+1) - E(i) = -[E(i-1) - E(i)] < 0$

$$\begin{aligned} r_i &= w_{i,i+1} + w_{i,i-1} \\ &= \nu + \nu \times e^{-\beta[E(i-1) - E(i)]} \end{aligned}$$

Since g is constant in each of the four regions of the approximation (see Figure 1), we can easily deduce that

$$g = \overline{E}(i+1) - \overline{E}(i) = -[\overline{E}(i-1) - \overline{E}(i)], \quad (\text{A-4})$$

which leads to

$$\begin{aligned} r_i &= \nu \times e^{-\beta(-g)} + \nu \\ &= \nu(e^{\beta g} + 1). \end{aligned} \tag{A-5}$$

Case 3: $E(i+1) = E(i)$, i.e., $g = 0$

$$\begin{aligned} r_i &= w_{i,i+1} + w_{i,i-1} \\ &= \nu + \nu \\ &= 2\nu. \end{aligned} \tag{A-6}$$

Straightforwardly, the three cases (Eq. A-3, Eq. A-5, Eq. A-6) combine to

$$r_i = \nu(e^{-\beta|g|} + 1). \tag{A-7}$$

Appendix B: Mean First-Passage Time

The mean-first passage time defined in (Slutsky and Mirny, 2004) is

$$\bar{t}_{0,L} = L + \sum_{k=0}^{L-1} \alpha_k + \sum_{k=0}^{L-2} \sum_{i=k+1}^{L-1} (1 + \alpha_k) \prod_{j=k+1}^i \alpha_j. \quad (\text{B-1})$$

Since the values of α_i are constant over sufficiently wide ranges, we assume $\alpha_0 = \alpha_1 = \dots = \alpha_L := \alpha$. Plugging this into Eq. (B-1), we obtain

$$\bar{t}_{0,L}' = L + L \times \alpha + (1 + \alpha) \sum_{k=0}^{L-2} \sum_{i=k+1}^{L-1} \alpha^{i-k}, \quad (\text{B-2})$$

which leads with

$$\begin{aligned} \sum_{k=0}^{L-2} \sum_{i=k+1}^{L-1} \alpha^{i-k} &= (\alpha^1 + \dots + \alpha^{L-1}) + (\alpha^1 + \dots + \alpha^{L-2}) + \dots + (\alpha^1) \\ &= (L-1) \times \alpha^1 + \dots + 1 \times \alpha^{L-1} \\ &= \sum_{k=1}^{L-1} (L-k) \alpha^k \end{aligned}$$

to

$$\bar{t}_{0,L}' = L + L \times \alpha + (1 + \alpha) \sum_{k=1}^{L-1} (L-k) \alpha^k. \quad (\text{B-3})$$

Acknowledgments

We would like to thank Leonid Mirny for helpful information on the Arrhenius equation. Moreover, we are indebted to Peter H. von Hippel for comments and discussions on the manuscript. All authors were partly supported by Deutsche Forschungsgemeinschaft (grants MU 1479/1-1, HA 1358/10-2) and Bund der Freunde der Technischen Universität München e. V. (grant 22502).

References

- Barbi, M., Place, C., Popkov, V. and Salerno, M. (2004a).
Base-sequence-dependent sliding of proteins on DNA. *Phys Rev E Stat Nonlin Soft Matter Phys*, **70**, 041901.
- Barbi, M., Place, C., Popkov, V. and Salerno, M. (2004b). A model of sequence-dependent protein diffusion along DNA. *J Biol Phys*, **30**, 203–226.
- Berg, O.G., Winter, R.B. and von Hippel, P.H. (1981). Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. Models and theory. *Biochemistry*, **20**, 6929–6948.
- Blainey, P.C., van Oijen, A.M., Banerjee, A., Verdine, G.L. and Xie, X.S. (2006). A base-excision DNA-repair protein finds intrahelical lesion bases by fast sliding in contact with DNA. *Proc Natl Acad Sci U S A*, **103**, 5752–5757.

- Bruinsma, R.F. (2002). Physics of protein-DNA interaction. *Physica A*, **313**, 211–237.
- Bustamante, C., Guthold, M., Zhu, X. and Yang, G. (1999). Facilitated target location on DNA by individual *Escherichia coli* RNA polymerase molecules observed with the scanning force microscope operating in liquid. *J Biol Chem*, **274**, 16665–16668.
- 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 U S A*, **93**, 8858–8862.
- Erie, D.A., Yang, G., Schultz, H.C. and Bustamante, C. (1994). DNA bending by Cro protein in specific and nonspecific complexes: implications for protein site recognition and specificity. *Science*, **266**, 1562–1566.
- Flyvbjerg, H., Keatch, S.A. and Dryden, D.T.F. (2006). Strong physical constraints on sequence-specific target location by proteins on DNA molecules. *Nucleic Acids Res*, **34**, 2550–2557.
- Gowers, D.M., Wilson, G.G. and Halford, S.E. (2005). Measurement of the contributions of 1D and 3D pathways to the translocation of a protein along DNA. *Proc Natl Acad Sci U S A*, **102**, 15883–15888.
- Gruber, T.M. and Gross, C.A. (2003). Multiple sigma subunits and the partitioning of bacterial transcription space. *Annu Rev Microbiol*, **57**, 441–466.

- Guthold, M., Zhu, X., Rivetti, C., Yang, G., Thomson, N.H., Kasas, S., Hansma, H.G. and Smith, B. (1999). Direct observation of one-dimensional diffusion and transcription by *Escherichia coli* RNA polymerase. *Biophys J*, **77**, 2284–2294.
- Halford, S.E. and Marko, J.F. (2004). How do site-specific DNA-binding proteins find their targets? *Nucleic Acids Res*, **32**, 3040–3052.
- Harada, Y., Funatsu, T., Murakami, K., Nonoyama, Y., Ishihama, A. and Yanagida, T. (1999). Single-molecule imaging of RNA polymerase-DNA interactions in real time. *Biophys J*, **76**, 709–715.
- Hosid, S., Trifonov, E.N. and Bolshoy, A. (2004). Sequence periodicity of *Escherichia coli* is concentrated in intergenic regions. *BMC Mol Biol*, **5**, 14.
- Hu, L., Grosberg, A.Y. and Bruinsma, R. (2007). Thermodynamic restrictions on evolutionary optimization of transcription factor proteins. *arXiv.org e-Print archive*, **arXiv:0709.1495**.
- Jeltsch, A., Alves, J., Wolfes, H., Maass, G. and Pingoud, A. (1994). Pausing of the restriction endonuclease EcoRI during linear diffusion on DNA. *Biochemistry*, **33**, 10215–10219.
- Kabata, H., Kurosawa, O., Arai, I., Washizu, M., Margaron, S.A., Glass, R.E. and Shimamoto, N. (1993). Visualization of single molecules of RNA polymerase sliding along DNA. *Science*, **262**, 1561–1563.
- Kalodimos, C.G., Biris, N., Bonvin, A.M.J.J., Levandovski, M.M., Guennuegues, M., Boelens, R. and Kaptein, R. (2004). Structure and

- flexibility adaptation in nonspecific and specific protein-DNA complexes. *Science*, **305**, 386–389.
- Kiryu, H., Oshima, T. and Asai, K. (2005). Extracting relations between promoter sequences and their strengths from microarray data. *Bioinformatics*, **21**, 1062–1068.
- Kobayashi, M., Nagata, K. and Ishihama, A. (1990). Promoter selectivity of *Escherichia coli* RNA polymerase: effect of base substitutions in the promoter -35 region on promoter strength. *Nucleic Acids Res*, **18**, 7367–7372.
- Mulligan, M.E., Brosius, J. and McClure, W.R. (1985). Characterization *in vitro* of the effect of spacer length on the activity of *Escherichia coli* RNA polymerase at the TAC promoter. *J Biol Chem*, **260**, 3529–3538.
- Murthy, K.P.N. and Kehr, K.W. (1989). Mean first-passage time of random walks on a random lattice. *Phys Rev A*, **40**, 2082–2087.
- Paget, M.S.B. and Helmann, J.D. (2003). The σ^{70} family of sigma factors. *Genome Biol*, **4**, 230.
- Ricchetti, M., Metzger, W. and Heumann, H. (1988). One-dimensional diffusion of *Escherichia coli* DNA-dependent RNA polymerase: A mechanism to facilitate promoter location. *Proc Natl Acad Sci U S A*, **85**, 4610–4614.
- Riggs, A.D., Bourgeois, S. and Cohn, M. (1970). The lac repressor-operator interaction. 3. Kinetic studies. *J Mol Biol*, **53**, 401–417.

- Sakata-Sogawa, K. and Shimamoto, N. (2004). RNA polymerase can track a DNA groove during promoter search. *Proc Natl Acad Sci U S A*, **101**, 14731–14735.
- Salgado, H., Gama-Castro, S., Peralta-Gil, M., Diaz-Peredo, E., Sanchez-Solano, F., Santos-Zavaleta, A., Matrinez-Flores, I., Jimenez-Jacinto, V., Bonavides-Martinez, C., Sequra-Salazar, J., Martinez-Antonio, A. and Collado-Vides, J. (2006). RegulonDB (version 5.0): *Escherichia coli* K-12 transcriptional regulatory network, operon organization, and growth conditions. *Nucleic Acids Res*, **34**, D394–D397.
- Saxton, M.J. (2007). A biological interpretation of transient anomalous subdiffusion. I. Qualitative model. *Biophys J*, **92**, 1178–1191.
- Sengupta, A.M., Djordjevic, M. and Shraiman, B.I. (2002). Specificity and robustness in transcription control networks. *Proc Natl Acad Sci U S A*, **99**, 2072–2077.
- Shimamoto, N. (1999). One-dimensional diffusion of proteins along DNA. Its biological and chemical significance revealed by single-molecule measurements. *J Biol Chem*, **274**, 15293–15296.
- Slutsky, M. and Mirny, L.A. (2004). Kinetics of protein-DNA interaction: Facilitated target location in sequence-dependent potential. *Biophys J*, **87**, 4021–4035.
- Slutsky, M., Kardar, M. and Mirny, L.A. (2004). Diffusion in correlated random potentials, with applications to DNA. *Phys Rev E Stat Nonlin Soft Matter Phys*, **69**, 061903.

- von Hippel, P.H. and Berg, O.G. (1986). On the specificity of DNA-protein interactions. *Proc Natl Acad Sci U S A*, **83**, 1608–1612.
- von Hippel, P.H. and Berg, O.G. (1989). Facilitated target location in biological systems. *J Biol Chem*, **264**, 675–678.
- Weindl, J. and Hagenauer, J. (2007). Applying techniques from frame synchronization for biological sequence analysis. *Proc IEEE Internat Conf Commun (ICC)*, 833–838.
- Weindl, J., Hanus, P., Dawy, Z., Zech, J., Hagenauer, J. and Mueller, J.C. (2007). Modeling DNA-binding by *Escherichia coli* σ^{70} exhibits a characteristic energy landscape around strong promoters. *Nucleic Acids Res*, **35**, 7003–7010.
- Widom, J. (2005). Target site localization by site-specific, DNA-binding proteins. *Proc Natl Acad Sci U S A*, **102**, 16909–16910.
- Winter, R.B., Berg, O.G. and von Hippel, P.H. (1981). Diffusion-driven mechanisms of protein translocation on nucleic acids. 3. The *Escherichia coli lac* repressor-operator interaction: kinetic measurements and conclusions. *Biochemistry*, **20**, 6961–6977.

Figure legends

Figure 1:

Energy landscape: Average energy landscape $\overline{E}(i)$ around 651 experimentally verified promoters of σ^{70} downloaded from RegulonDB (version 5.7, see (Salgado et al., 2006)) and its linear approximation by four straight lines. $\overline{E}(i) = 0$ corresponds to the energy E_{ran} of random sequences.

Figure 2:

Mean first-passage time: Mean first-passage time for $L = 50$ ($\alpha_i = q_i/p_i$) or $L = -50$ ($\alpha_i = p_i/q_i$) if assuming constant values of α_i .

Figure 1

Figure 2