Classes of fast and specific search mechanisms for proteins on DNA

M Sheinman\textsuperscript{1,2}, O Bénichou\textsuperscript{3}, Y Kafri\textsuperscript{2} and R Voituriez\textsuperscript{3}

\textsuperscript{1} Department of Physics and Astronomy, Vrije Universiteit, Amsterdam, The Netherlands
\textsuperscript{2} Department of Physics, Technion, Haifa 32000, Israel
\textsuperscript{3} UMR 7600, Université Pierre et Marie Curie/CNRS, 4 Place Jussieu, 75255 Paris Cedex 05 France

Received 7 April 2011, in final form 20 October 2011
Published 24 January 2012
Online at stacks.iop.org/RoPP/75/026601

Abstract
Problems of search and recognition appear over different scales in biological systems. In this review we focus on the challenges posed by interactions between proteins, in particular transcription factors, and DNA and possible mechanisms which allow for fast and selective target location. Initially we argue that DNA-binding proteins can be classified, broadly, into three distinct classes which we illustrate using experimental data. Each class calls for a different search process and we discuss the possible application of different search mechanisms proposed over the years to each class. The main thrust of this review is a new mechanism which is based on barrier discrimination. We introduce the model and analyze in detail its consequences. It is shown that this mechanism applies to all classes of transcription factors and can lead to a fast and specific search. Moreover, it is shown that the mechanism has interesting transient features which allow for stability at the target despite rapid binding and unbinding of the transcription factor from the target.

(Some figures may appear in colour only in the online journal)

This article was invited by E Frey.
1. Introduction

Many biochemical processes require both an appropriate speed and a high specificity for proper biological functions to occur—a fast desirable process should not be accompanied by a significant acceleration of undesirable ones. With typical energy scales of a few $k_B T$, where $k_B$ is the Boltzmann constant and $T$ is the temperature, evolution has devised many efficient mechanisms which overcome the noisy environment and the speed requirements. These range from mechanisms which rely on the consumption of chemical energy, such as kinetic proofreading [1], to cooperativity, such as in the specific regulation of hemoglobin oxygen concentration [2, 3]. Unraveling these mechanisms is an important step towards understanding how cells function.

Being based on biopolymers, specificity in biological systems implies that two (or more) well-defined subsequences of two given polymers attach to each other, but not to other subsequences of the same polymers or to other polymers. The two polymers can be proteins (for example, in enzymes [3]), RNA molecules (for example, in ribosomal action [4, 5]), a single-stranded and a double-stranded DNA (for example, in the homologous recombination [6]) or a transcription factor (TF) and a DNA molecule. The last example highlights the challenges which a biological system faces.

Consider, for example, a prokaryotic cell (throughout the review we focus on these simpler systems). Its typical DNA length is $N \simeq 10^7$ base pairs. In a particularly simple case a TF has to bind to a specific subsequence (target) of a length of about $10-20$ base pairs on the DNA. The typical binding energy between a protein and the DNA subsequence is of the order of tens of $k_B T$, about one $k_B T$ per base pair. Without using chemical energy (which is true for almost all transcription factors) this gives rise to a classical conflict between entropy and energy which puts a hamper on the stability of the TF at the target. Specifically, the entropy associated with the protein bound to non-target DNA is $k_B \ln N \simeq 16 k_B$ and therefore its contribution to the free energy is of the same order as the binding energy. Unless the TF is designed to have a binding energy at the target that is much lower than to the rest of the sequence the probability of finding it on the target site will be very low. Of course, the copy number of a TF, which in a cell typically ranges from about tens to thousands [8–11], can increase the occupation probability of the target site to a desired level (see, for instance, [12]). This, however, comes at a cost of producing many proteins and possibly activating or repressing unwanted genes and losing specificity, meaning that the TF is likely to occupy non-specific sites (below this argument is presented in a quantitative manner).

Following this line of thought early works [13–15] considered designed targets with a gapped binding energy which is much lower than the rest of the DNA sequence. A sufficiently large energy gap at the target can then yield an arbitrarily large occupation probability of the target site even for one TF. When this is assumed the interesting question becomes that of the speed of the search. To address this question various mechanisms, collectively called facilitated diffusion, were suggested. These combine one-dimensional diffusion along DNA with three-dimensional diffusion or intersegmental transfers. The combination of the various search modes has been observed experimentally [16–33] and shown theoretically to be capable of decreasing the search time significantly [13, 15, 34–60]. More recently the influence of facilitated diffusion on the noise level in gene regulation was analyzed in [61, 62].

However, as realized early [63] the assumption of a designed target is far from obvious. In an alphabet of four letters a target sequence of length 12, quite common in TFs, will occur with essentially probability 1 in a random sequence of length $\simeq 10^3$. Therefore, for target sequences shorter than 12 bases, identical and almost identical sequences will occur on the DNA. These competing sites can easily ruin the stability of the target site. Furthermore, as discussed in detail below, these almost identical sequences act as traps [64] that hinder the search process and lead to an antagonism between the stability of the TF at the target site and the speed of the target location. This problem, raised in [17], is commonly referred to as the speed-stability paradox.

Recently, motivated by new experiments there has been renewed interest in this rather old problem. To date there are now several reviews (some very recent) which cover different aspects of the problem [14, 15, 65–68]. We believe that this review complements these and presents the problem using a somewhat new angle. To this end we give an overview of the current status of the speed-stability paradox and its implications on regulation dynamics. We present the problem using both theoretical considerations and experimental data. As we argue it is preposterous to group all TFs in a single class [69]. Different search mechanisms are likely to apply to different proteins grouping them into different classes. We show that three broad classes can be defined, which we term gapped, marginally gapped and non-gapped TFs. The applicability of previously suggested search mechanisms to each of the groups is analyzed in some detail. Using this we turn to discuss in detail a recently proposed barrier controlled search mechanism [70] which can in principle resolve the speed-stability paradox for all classes of proteins. The possibility of such a mechanism suggests that experiments should also probe activation barriers and not, as commonly done, binding energies (see discussion below). Moreover, this mechanism allows for a rich transient behavior and for transcription factors which are efficient despite binding and unbinding rapidly from the target.

The structure of the review is as follows. In section 2 we discuss in detail the energetics associated with protein–DNA interaction. We argue for the classification of transcription factors into the three classes defined above. The classification is illustrated using experimental data. In section 3 we review the kinetics of simple search mechanisms which have been discussed in the literature. In section 4 we introduce the speed-stability paradox and its possible resolution for each class of TFs. In section 5 we introduce and analyze in detail the barrier controlled search mechanism. In section 6 an effective model for the barrier controlled search mechanism is introduced and
used to study transient behaviors. We summarize the results in section 7.

2. Protein–DNA energetics

Due to the sequences heterogeneity of the non-target DNA the binding energy of a protein to a DNA is location dependent. The structure of this disordered, non-specific, energy landscape is crucial for understanding the stability of a TF at its target site and which search strategies can or cannot be efficient. To this end, in this section we consider the energy landscape both from a theoretical point of view and by looking at experimental data. Throughout what follows we use units where $k_B T = 1$.

Equilibrium measurements [71] reveal that to a good approximation the binding energy, $U(s)$, of a transcription factor which binds to a sequence of $l_p$ bases $s = (s_1, s_2, \ldots, s_{l_p})$ on the DNA is given by [37]

\[ U(s) = \sum_{i=1}^{l_p} E(s_i, i). \]  

(1)

Here $s_i = A, T, C, G$ is the nucleotide type on the $i$th binding location of the protein and $l_p$ is the number of binding sites on the protein (see figure 1). The binding energies $E(s_i, i)$ are usually estimated experimentally by measuring the probability, $Pr(s, i)$, that a nucleotide $s$ is bound to a location $i$ on the protein in equilibrium *in vitro* experiments. Namely, one uses

\[ Pr(s, i) = \frac{e^{-E(s, i)}}{Z_i} \quad \text{where} \quad Z_i = \sum_{s'=(A,T,C,G)} e^{-E(s', i)}. \]  

(2)

We use this simple convention throughout. For a more detailed discussion of other possible choices see [72]. The matrix $Pr(s, i)$ has $4 \times l_p$ elements and is called the weight matrix (also known as position-specific scoring matrix (PSSM) or ‘profile’). It is important to note that these probabilities are measured only for sequences which are close in structure to the target site. The reason for this lies in the existence of other conformations of the protein–DNA complex which we will allude to later [73]. In figure 2 we illustrate a sample binding energy probability distribution for several *E. coli* proteins.

The structure of the binding energy implies that it can be described by three parameters instead of the $4^{l_p}$ entries. Specifically, the energy is a sum of contributions (see equation (1)) which can be assumed independent, if the DNA sequence is uncorrelated, and can therefore be modeled to a good approximation by a Gaussian random variable. (The assumption that the DNA sequence is uncorrelated is believed

5 Since the binding probability is measured only in places close to the target sequence on a finite sample there are cases where one or more of the letters does not appear. To correct for this the probability of a letter to appear at a given site is derived from $(n_s + 1/4)/(1 + \sum_s n_s)$, where $n_s$ is the number of occurrences of the letter $s$. This, standard procedure, ensured that when no measurements are made the probability is 1/4. Note that other choices are possible. These can lead to somewhat different estimates, especially for data points with a small sample size.
Figure 3. A histogram of the binding energy for three different TFs. (a) BaeR, \( l_p = 29, \sigma = 6.86, U_T = -50.16, U_{\text{min}} = -33.5, P^T = 1 \). (b) DcuR, \( l_p = 15, \sigma = 4.93, U_T = -21.76, U_{\text{min}} = -21.75, P^T = 0.074 \). (c) AscG, \( l_p = 7, \sigma = 4.2, U_T = -14.55, U_{\text{min}} = -14.55, P^T = 0.0006 \). The red lines are Gaussian approximations to the distributions using the measured variance calculated from equation (4). (d) A histogram of the estimated gap values. The data were based on 89 weight matrices of *E. coli* DNA-binding proteins and were taken from the RegulonDB database [83].

The target energy is given by

\[
U_T = \sum_{i=1}^{L_p} \min [E(A,i), E(T,i), E(C,i), E(G,i)].
\]

The statistical properties of the binding energy are now encoded by \( \sigma_U \) and \( U_T \) and the mean binding energy which we set to be zero. Note that the Gaussian form is unchanged even if one allows for corrections to the weight matrix which depend, say, on near-neighbor configurations, as suggested in [76–79].

The assumption that the DNA sequence is uncorrelated also implies that the binding energies \( U_i \) and \( U_j \) at different sites \( i \) and \( j \) are independent. Strictly speaking this holds only for \( |i - j| > l_p \). In what follows we neglect these, unimportant, short range correlations.

Another quantity which is important for understanding the binding is the minimal energy, \( U_{\text{min}} \geq U_T \), which occurs randomly on a typical DNA sequence among the non-target sites. Note that typically not all sites of the DNA can be reached by TFs and in eukaryotic cells the number of accessible sites can be significantly smaller than the number of DNA sites. This site competes most strongly with the target site. In a
sequence of $N \gg 1$ accessible uncorrelated base pairs $U_{\text{min}}$ is narrowly distributed (with a variance scaling as $1/\ln N$) and well approximated by [80]

$$\int_{-\infty}^{U_{\text{min}}} P(U) \simeq \frac{1}{N}$$ \hspace{1cm} (6)

or

$$U_{\text{min}} \simeq -\sigma_U \sqrt{2 \ln N}$$ \hspace{1cm} (8)

and

$$U_T < -\sigma_U \sqrt{2 \ln N}.$$ \hspace{1cm} (9)

(b) Marginally gapped transcription factors. Here there is no energetic gap between the target and the rest of the DNA but the number of sites with an energy close to $U_T$ is small (of the order of 1). This happens when

$$U_{\text{min}} \simeq U_T \simeq -\sigma_U \sqrt{2 \ln N}.$$ \hspace{1cm} (10)

Note that within the additive binding energy model, equation (1), the possible existence of a gapped TF is directly related to its length. In that case

$$U_T = l_p E_c$$ \hspace{1cm} (12)

where $E_c < 0$ is the average lowest binding energy per base and

$$\sigma_U^2 = \frac{l_p E_c^2}{3}.$$ \hspace{1cm} (13)

Here we assumed that each base appears with equal probability along the DNA. Then equation (9) implies that to produce an energetic gap between $U_T$ and $U_{\text{min}}$ a TF has to be long enough. Namely, one finds

$$l_p > \frac{2}{3} \ln N.$$ \hspace{1cm} (14)

This has a particularly simple interpretation. It is equivalent to demanding that on a DNA sequence, of length $N$, sites which are identical to the target site do not appear randomly so that $1/4^l < 1/N$. For a typical bacterial DNA length, $N = 10^7$, this gives $l_p > 11$. The argument can be refined using information theoretic arguments (see appendix A and [69] for a similar line of reasoning) to give a stronger bound of $l_p > 22$.

As we discuss below, the structure of the energy landscape, gap existence and the properties of the target have important consequences on the equilibrium probability of finding the protein on the target and the search time. Interestingly, as we show below, experimental data suggest that there are transcription factors which belong to each of the above categories.

2.1. Target occupation probability in equilibrium

Next, we turn to consider the probability of a TF to be at the target, $P^T$, in equilibrium. For TFs which appear in small numbers (as believed to be the case in many examples [8]) this quantity has to be of the order of 1 for proper control over gene expression. Otherwise, assuming equilibration (we discuss other scenarios later), the TF has to be present in a large copy number. Naively $P^T$ will be of the order of 1 as long as the TF is gapped. As we now show this is not guaranteed and we outline the conditions for this to occur. For clarity in the discussion we ignore in the stability discussion the free-energy contribution from configurations where the protein is off the DNA. These can only hamper the stability at the target. To achieve stability the free energy of the off-DNA states must be, of course, higher than the target free energy.

In equilibrium to ensure $P^T$ close to 1 the partition function has to be dominated by the target energy. Namely, for stability we require

$$Z = \sum_{i=1}^{N} e^{-U_i} \simeq e^{-U_T}.$$ \hspace{1cm} (15)

The typical partition function can be approximated, using equation (3), by

$$Z \simeq e^{-U_T} + N \int_{U_{\text{min}}}^{\infty} e^{-\frac{U}{\sigma_U^2}} e^{-U} dU + N \int_{U_{\text{min}}}^{\infty} e^{-\frac{U}{\sigma_U^2}} dU.$$ \hspace{1cm} (16)

Note that as standard in disordered systems, this can be different from the average partition function which is obtained by setting the lower bound of the integrations on the right-hand side to $-\infty$. This gives in the large $N$ limit

$$Z \simeq \begin{cases} e^{-U_T} + e^{\sigma_U^2 \sqrt{2 \ln N}} \sigma_U & \text{for } \sigma_U \gg \sqrt{2 \ln N}, \\ e^{-U_T} + Ne^{E_c^{\gamma}} & \text{for } \sigma_U \ll \sqrt{2 \ln N}. \end{cases}$$ \hspace{1cm} (17)

We therefore identify two regimes: large disorder strength $\sigma_U \gg \sqrt{2 \ln N}$ and small disorder strength $\sigma_U \ll \sqrt{2 \ln N}$.
Note that the physics is very close to that of the random energy model (REM) [82].

For large disorder strength \( \sigma_U \gg \sqrt{2 \ln N} \), which corresponds to the frozen phase of the REM, gapped TFs or marginally gapped TFs are stable on the target. Together with definitions (9) and (10), this condition reads

\[
U_T \leq -\sigma_U \sqrt{2 \ln N}.
\]

(18)

To satisfy the stability requirement in the small disorder case, which corresponds to a system above the freezing point of the REM, it is required that

\[
U_T \leq -\ln N - \frac{\sigma_U^2}{2},
\]

(19)

so that only gapped TFs can be stable on the target. Using the additive binding model, so that \( U_T = l_p E_c \) and \( \sigma_U^2 = \frac{6 \ln N}{E_c} \), implies that the small disorder regime corresponds to \( l_p \ll \frac{6 \ln N}{E_c} \) and the stability condition translates in this case to the constraint

\[
l_p \geq \frac{\ln N}{E_c (1 + E_c / 6)},
\]

(20)

This is possible only for \( -E_c < 6 \). As expected the bound on \( l_p \) grows when \( E_c \) approaches zero.

Note that for both large and small disorder strengths, the larger \( N \), the more stringent the condition on \( U_T \). With \( E_c \) of the order of \(-1\) the above conditions give \( l_p \geq 16 \) for small disorder and \( l_p \geq 32 \) for large disorder. We comment that in principle a simple way to satisfy conditions (18) or (19), is for example to introduce large enough cooperative interactions between different TF’s binding domains. In this case the binding energy is not additive so that equation (1) is not valid. These can single out the target and generate an arbitrarily large gap between the target and the rest of DNA sites.

In summary, TFs with non-gapped targets cannot be stabilized on their targets. Marginally gapped TFs can be stabilized on their targets if the disorder strength is large enough. Below, we show that this requirement gives rise to a conflict with the speed of the target location. A gapped TF is stable on its target when the disorder strength is large or in the small disorder regime if it is large enough (or if cooperative effects are present). Without any cooperative interaction between different TF’s parts, such a gap may be achieved in both small and large disorder regimes for reasonable TF length (for a biochemically reasonable energy scale of about 1). Below we show that combining these requirements with another set of constraints related to the speed of the search gives much more stringent conditions on the length of the protein.

2.2. Experimental data

In recent years much experimental data have been accumulated. Specifically the weight matrix has been measured for many TFs. We now use data from RegulonDB [83] which contains 89 weight matrices to try and single out the different classes of proteins discussed theoretically above. As we proceed to show the three classes can be identified in the data. Three examples are shown in figure 3. These correspond to a gapped (figure 3(a)), marginally gapped (figure 3(b)) and non-gapped (figure 3(c)) proteins.

To analyze the stability of all the proteins in the database we look at several quantities. (i) Their minimal possible binding energy \( U_T = U(s^*) \), where \( s^* \) is defined to be the target of the protein. (ii) The minimal binding energy on a typical disordered sequence of length \( N \), \( U_{min} = U(s^*) \), where \( s^* \) is the strongest binder on the sequence. (iii) The standard deviation \( \sigma_U \) for the different proteins and finally (iv) the occupation probability at the target, \( P_T \). Some of the results presented below are demonstrated in appendix A using the language of information theory (for a related discussion see [69]).

It is useful to present the data by plotting \( U_{min} \) and \( U_T \) as a function of \( \sigma_U \) (see figure 4). Each protein on the graph is represented by two points with the same abscissa. The graph shows several interesting features.

(i) First, as expected, a significant part (about 3/4) of the TFs are gapped with a gap size ranging from a few \( k_b T \) to about \( 20 k_b T \). An histogram of the gap size is shown in figure 3(d). As stated above such gapped proteins are stable only when the gap is large enough, see equations (19) and (18). For an E. coli DNA length this requires \( U_T < -15 \) in the small disorder regime \( \sigma_U \ll \sqrt{2 \ln N} \approx 5.5 \) and \( U_T < -30 \) in the large disorder regime \( \sigma_U \gg 5.5 \). Note that indeed for \( \sigma \gtrsim 5.5 \) a large fraction of the values of \( U_T \) are below \(-30\) and therefore correspond to stable TFs. The stability criterions for both small (equation 19) and large (equation 18) disorder strengths are shown in figure 5 and indicate that most proteins with a large gap are stable. Note also that the theoretical prediction for \( U_{min} \) (shown in figure 4) fits reasonably well with the experimental results.

(ii) Second, for about 1/4 of the TFs \( U_T \approx U_{min} \). This indicates that they are either non-gapped or marginally gapped. Recall that for such proteins a minimal criterion for being stable at the target is that the disorder is large \( \sigma_U \gg \sqrt{2 \ln N} \approx 5.5 \). This does not seem to be satisfied for most of the marginally gapped proteins. Therefore, figure 4 hints that most of the non-gapped and marginally gapped TFs are actually unstable on the target. This is more clearly illustrated in figure 6 which shows that indeed about 1/4 of the proteins have a very small probability (less than \( 10^{-1} \) with about half of them with a probability less than \( 10^{-2} \)) for being on the target. This indicates that non-gapped and marginally gapped TFs seem to break the stability requirement. We return to these proteins later and suggest that either non-equilibrium effects or large copy numbers could stabilize them on the target.

It is interesting to present the same data, but instead of as a function of \( \sigma_U \), as a function of \( l_p \). This is shown in figures 7 and 8. As is clearly seen there is a close relation between the existence of a gap and \( l_p \) being large enough. In fact, in agreement with our simple arguments, a gap begins to form at \( l_p \approx 13 \). The data for \( P_T \) as a function of \( l_p \) are even more
striking. Essentially all proteins with a binding site of size \( l_p \simeq 13 \) or smaller are unstable on the target while those with \( l_p \simeq 16 \) or larger are mostly stable at the target. The close correspondence between \( l_p \) and the gap is a direct result of a similar binding energy per base for all TFs.

The above discussion focused on the stability characteristics. We identified several distinct classes of TFs based on their stability properties. An important question for transcription factors is their speed of operation. The discussion above suggests that different TFs could have different search strategies. Before attempting to map these out in what follows we first review the different possible reactive pathways which have been suggested in the literature.

Before discussing the reactive pathways it is useful to have a simple picture of DNA packing in prokaryotic cells. In typical systems the DNA has a total length of \( L \sim 10^6 \) nm, a persistence length \( L_0 \sim 50 \) nm, a cross section radius \( \rho \sim 1 \) nm, and is contained in a volume of \( \Lambda^3 \sim 10^8 - 10^9 \) nm\(^3\).

The typical distance between segments of DNA of length \( L_0 \) is therefore much smaller than \( L_0: \frac{\Lambda^3}{L_0^2} \ll L_0^3 \). Under these conditions, using \( \Lambda \gg L_0 \), it is easy to check that the radius of gyration of free DNA, which is of the order of \( L_0 \sqrt{\frac{\Lambda}{L_0}} \), is much larger than the cell size \( \Lambda \)—the DNA is densely packed even though its fractional volume in the container \( L_0^2 / \Lambda^3 \), is small (about 1%). By way of comparison, typical protein sizes...
For a cell to properly function the search process has to, typically, be of the order of seconds. In principle, when the target is perfectly reactive this can be achieved by a search driven by pure three-dimensional diffusion. However, driven by experimental results, mostly on the Lac repressor [88, 89], which seem to give search times that are faster than three-dimensional diffusion, various search strategies were suggested. We now give simple arguments that quantify these different search strategies. For a similar discussion see [39, 45].

3.1. Searching with three-dimensional diffusion

Naively, one might expect the protein to search for its target (or, equivalently, its specific binding site on the DNA) using only three-dimensional diffusion. Neglecting interactions of the protein with the environment and the DNA (apart from the target site), one then finds, using results first obtained by Smoluchowski [90] or by dimensional analysis, that the search time, $t_{\text{search}}$, defined as the mean first-passage time at the target, is given by

$$t_{\text{search}} \sim \frac{\Lambda^3}{D_3 r}.$$  \hspace{1cm} (21)

Here $D_3$ is the three-dimensional diffusion constant of the protein, $r$ is the target size and $\Lambda^3$ is the volume that needs to be searched. Assuming a target size of the order of a base pair $r \approx 0.34$ nm, a typical nucleus (or bacterium) size as above and using the measured three-dimensional diffusion coefficient for a protein in a cell ($D_3 \approx 10^5$ nm$^2$ s$^{-1}$ for the GFP protein [91] and $D_3 \approx 10^6$ nm$^2$ s$^{-1}$ for the Lac repressor protein [29]), one finds $t_{\text{search}}$ of the order of hundreds of seconds. We comment that $r$ can be increased significantly by changing the electrostatic interactions between the protein and its target, for example, by changing the salt concentration. Under typical physiological conditions the Debye screening length is of the order of 1 nm [92].

These long time scales can be easily reduced if several proteins are searching for the target. Namely, if $n_p$ proteins are searching for the same target the average search time is given by\textsuperscript{8} $t_{\text{search}} \sim t_{\text{search}}/n_p$. This suggests that about 10–100 proteins could find a target in reasonable time for cells to function properly. As we discuss below this simple relation between the search time of one protein and $n_p$ proteins can fail in some cases.

3.2. Searching with one-dimensional diffusion

In real systems, due to the interactions of proteins with non-specific DNA sequences and the environment [93], the picture is more complex. Indeed, \textit{in vitro} experiments have suggested that mechanisms other than three-dimensional diffusion are used by many proteins to locate their targets. The simplest extension of the pure three-dimensional diffusive search is using three dimensional diffusion to reach the DNA and then

\textsuperscript{8} The relation between the search time $t_{\text{search}}$ for one protein and search time $t_{\text{search}}$ for $n_p$ proteins remains unchanged up to section 5. In the next section it is shown that in the case of wide distributions of the search time the dependence on $n_p$ is more sensitive.

For an exception see, for example, [84].
scan it using one-dimensional diffusion along its contour. This follows closely ideas of Delbruck and Adam [34], introduced in a different context. If the DNA is very long the search time is clearly controlled by the one-dimensional diffusion along the DNA which is given by

\[ t_{\text{search}} \sim \frac{L^2}{D_1} \sim O(\text{hours}). \] (22)

Here \( L \sim 10^6 \text{ nm} \) is the genome length and \( D_1 \) is the one-dimensional diffusion coefficient that was measured indirectly [36] and directly [28, 29] to be much smaller than the three-dimensional diffusion coefficient \( D_3 \sim 10^7 \text{ nm}^2 \text{s}^{-1} \) [91]. Effects of disorder can be incorporated into an effective value of \( D_1 \) [39] (see discussion below). The above result renders this search strategy useless for long DNA. However, if the sequence scanned is short then it is easy to see that the search time is given by

\[ t_{\text{search}} \sim \frac{L^2}{D_1} + \frac{L^3}{D_3 L}. \] (23)

Using the numbers cited above it is easy to check that search times of the order of a 100 s (so that about ten proteins can find the target within seconds) can be obtained as long as \( L \), the length of the sequence scanned is smaller than \( 10^4 \text{ nm} \), about 30 kilobases long. The results are mildly modified if the sequence has a globular shape.

### 3.3. Facilitated diffusion

Motivated by experiments [88, 89] an extension of the Delbruck and Adam model was suggested in [13]. The model combines one-dimensional diffusion (sliding) along the DNA which is interrupted by periods of three-dimensional diffusion (typically called jumping or hopping in this context). This combined strategy, called facilitated diffusion, has been studied and debated extensively both in the context of in vivo and in vitro systems [13, 15, 29, 36, 39, 41, 43–46, 94, 95]. There is now a large body of evidence showing that such a mechanism plays an important role for several TFs. It is illustrated in figure 9 and is believed to speed the search process.

Each of the individual search mechanisms described above, when applied alone, has shortcomings and advantages over the other. When using only three-dimensional diffusion, the number of distinct three-dimensional positions probed grows linearly in time but the protein spends much time probing sites where there is no DNA present. In contrast, during a one-dimensional diffusion the protein is constantly bound to the DNA but suffers from a slow increase in the number of distinct positions probed as a function of time \((\sim t^{1/2})\), where \( t \) denotes time) [96]. It is known that by intertwining one- and three-dimensional search strategies and tuning the properties of both one can in fact decrease the search time significantly [13].

The discussion below follows [15, 39] closely. We imagine a single protein searching for a single target located on the DNA. The search is composed of a series of intervals of one-dimensional diffusion along the DNA (sliding) and three-dimensional diffusion in the solution (jumping). The mean time of each is denoted by \( \tau_1 \) and \( \tau_3 \), respectively. Following a jump, the protein is assumed to associate on a new randomly chosen location along the DNA. Note that one might be worried if the structure of the packed DNA molecule invalidates this approach. Numerics on typical frozen DNA conformations indicate that as long as average search times are considered the structure can be ignored [97]. Nonetheless much more complicated structured may arise in nature (for example, in eukaryotic cells [45, 98–100]) and these are ignored in the discussion below.

Under the above assumptions, during each sliding event the protein covers a typical length \( l \), where \( l \sim \sqrt{D_1 \tau_1} \) (often called the antenna size) [96]. To complete the search process

\[ N_r \sim \frac{L}{l} \] (24)

rounds of sliding and jumping are needed on average. While this can be intuitively understood since the correlations between the locations of the protein before and after the jump are neglected the exact nature of the relation is in fact somewhat more subtle. As shown in [40] the average length scanned before the target is reached is half the total length. Nonetheless for the average search time the expression is exact in the large \( L \) limit. The total time needed to find a specific site is then

\[ t_{\text{search}} = N_r \tau_r, \] (25)

where \( \tau_r = \tau_1 + \tau_3 \) is the typical time of a round. Using equations (24) and (25) one obtains

\[ t_{\text{search}} \sim \frac{L}{l} (\tau_1 + \tau_3) \sim \frac{L}{\sqrt{2D_1}} \left( \sqrt{\tau_1 + \frac{\tau_3}{\sqrt{\tau_1}}} \right). \] (26)

Furthermore, from dimensional analysis it is easy to argue that

\[ \tau_3 \sim \frac{A^3}{D_3 L}. \] (27)
As shown in [86] this result holds up to a logarithmic correction which diverges as the DNA cross section, \( \rho \), vanishes. The analysis leads to three distinct regimes. (i) For \( \tau_1 \ll \tau_3 \) there is no dependence on \( L \) and the search time is given to a good approximation by equation (21). (ii) For \( \frac{L^2}{D_1} \gg \tau_1 \gg \tau_3 \) the dependence on the DNA length is linear. This is the regime typically considered relevant for experiments. (iii) For \( \frac{L^2}{D_1} \ll \tau_1 \) one finds \( t_{\text{search}} \propto L^2 \).

It is natural to ask which \( \tau_1 \) optimizes \( t_{\text{search}} \) when \( \tau_3 \) is held fixed. Using equation (26) it is easy to verify that

\[
\tau_1^{\text{opt}} = \tau_3. \tag{28}
\]

It can be shown that this result is exact in the large \( L \) limit [40]. Alternatively, one can consider an optimal antenna size \( l_{\text{opt}} = \sqrt{2D_1 \tau_3} \). When this condition is met, the total search time scales as

\[
t_{\text{search}}^{\text{opt}} \sim \sqrt{\frac{\tau_3}{D_1}} L \sim \sqrt{\frac{\Lambda^3 L}{D_1 D_3}} \tag{29}
\]

Note that the \( \sqrt{D} \) dependence is obtained by optimizing, say \( \tau_1 \), as \( L \) is varied. This model, at the optimal \( \tau_1 \) and assuming known values for \( D_1 \), \( L \) and \( \tau_3 \), predicts reasonable search times in vitro and is commonly believed to give a possible explanation for the efficiency of the target location process in experiments.

The combined strategy, while better than the pure three-dimensional or one-dimensional search strategies, comes at a cost of being sensitive to changes in the properties of either the three-dimensional or the one-dimensional diffusive processes. Given the many constraints on the protein to function, it is restrictive to demand an optimization of the search process. Specifically, within the model an optimal search process requires fine tuning of the antenna size, \( l \), as a function of the parameters \( D_1 \) and \( \tau_3 \). These parameters depend on various cell and environmental conditions such as the size of the cell, the DNA length and the ionic strength. The dependence can be quite significant: for example, the parameter \( \tau_3/\tau_1 \) has been argued to have an exponential dependence on the square root of the ionic strength [101]. Deviations of this parameter from the optimum value might be crucial to the search time since \( \frac{t_{\text{search}}}{t_{\text{opt}}} = \frac{1}{2} (\sqrt{\frac{\tau_2}{\tau_1}} + \sqrt{\frac{\tau_1}{\tau_2}}) \). Indeed, a strong dependence of the search time on the ionic strength was found in in vitro experiments [89].

Interestingly, in vivo, when the DNA is densely packed, no effect of the ionic strength on the efficiency of the Lac repressor was revealed [102]. Other experiments also suggest that \( \tau_1 \) is not optimized. In particular, equilibrium measurements [103], as well as recent single molecule experiments [28, 29], find a value of \( \tau_1 \) for the Lac repressor that is much larger than the predicted optimum (the value of \( \tau_3 \) in vivo).

Other mechanisms have been suggested to decrease the search time, for example intersegmental transfers (IT) [104, 105]. During an IT the protein moves from one site to another by transiently binding both at the same time. This mechanism is expected to be important for systems with a high DNA density [106]. In principle, the new site can be either close along the one-dimensional DNA sequence (or chemical distance) or distant (see figure 9). An analysis shows that the average search time remains similar to the combined one-dimensional and three-dimensional diffusion described above but with \( \tau_1 \) which obtains a different dependence on the DNA length. This has been discussed in detail in [97].

Finally, we comment that, in principle, all search strategies can be made arbitrarily fast by increasing the number of searchers. This allows, in principle, any of the above discussed mechanisms, one-dimensional diffusion, three-dimensional diffusion, facilitated diffusion with or without intersegmental transfers, to be at work for different proteins [107]. This statement, however, becomes more problematic when the stability requirements discussed above are included. As stated above, it is clear that the TFs also interact with non-target sites so that pure three-dimensional searches are unlikely. This implies that facilitated diffusion is hard to avoid. In addition, it was shown that high copy numbers for all searchers may actually slow down the search process, because the proteins start to become roadblocks for other searchers [108]. This poses a great physical constraint on the global number of DNA-binding proteins. To this end, in what follows we analyze in detail the problems which arise when facilitated diffusion is combined with the stability requirements.

### 4. The speed-stability paradox and possible solutions

It has been recognized early that there is a tight connection and antagonism between the stability of the TF at the target and the search speed. The conflict is commonly termed the speed-stability paradox [17, 39]. As we have seen, the experimental data show that the proteins can be classified into several classes. These classes call for different mechanisms. We begin by introducing the speed-stability paradox and then discuss possible solutions for each class of proteins.

#### 4.1. The speed-stability paradox

Recall the search time of one protein is proportional to \( 1/\sqrt{D_1} \) (we later return to the case of \( n_p \) proteins and discuss it in detail). Then note that the binding energies between the transcription factor and the DNA on each site \( i, U_i \), is an independent random variable with a Gaussian distribution with a variance \( \sigma_U^2 \). This disorder in the binding energies of the protein to different sites implies that this diffusion takes place on a disordered potential. For long times this leads to an effective diffusion constant whose value is given by [39]

\[
D_1 = D_{11} \left( \sigma_U = 0 \right) \sqrt{1 + \frac{\sigma_U^2}{2} e^{-\frac{\sigma_U^2}{2}}} \tag{30}
\]

The important thing to note is the exponential dependence of the diffusion coefficient on \( \sigma_U^2 \). It can be understood up to prefactors by recalling that the diffusion is an activated process so that \( D_1 \propto \int dU e^{-\frac{U^2}{2}} \). This implies that even for \( \sigma_U = 5.5 \) (the borderline between the small and the large disorder regimes for an E. coli genome) the one-dimensional diffusion constant becomes 19 orders of magnitude smaller...
than the diffusion constant on a flat energy landscape. This in turn leads to a very slow search process. Essentially, speed requirements prohibit the large disorder regime discussed above. For the search to be fast $\sigma_U$ has to be kept small, of the order of 0.5, to ensure a diffusion coefficient of the same order as that on a flat energy landscape.

On the other hand, this requirement conflicts with the stability requirements for proteins which demand (see section 2.1) either a large value of $\sigma_U$ (for marginally gapped and gapped TFs) or a large TF length $l_p$ (for gapped TFs in the small disorder regime), to create a significant gap between the energy at the target and the rest of the DNA. From the analysis above, a priori only gapped TFs might satisfy both speed and stability requirements. Below we analyze in detail the speed and stability requirements for gapped and marginally gapped TFs and discuss possible solutions of the paradox.

More puzzling are non-gapped proteins which are unstable at the target. A new possible mechanism which ensures both speed and stability for those is discussed in later sections.

4.2. Possible solutions of the speed-stability paradox for gapped, marginally gapped and non-gapped transcription factors

4.2.1. Gapped transcription factors. In principle, both speed and stability requirements can be easily satisfied using, for example, cooperative interactions between distinct searchers (which may increase the effective size of the binding domain) or through cooperative interactions between distinct TF binding domains on the target sequence and small binding energy, $\sigma_U$. Therefore the probability distribution of the total free energy on site $i$ is then given by [37]

$$ G_i = - \ln \left( e^{-U_{ns}} + e^{-U_i} \right) \simeq \min \left( U_{ns}, U_i \right). $$

Therefore the probability distribution of the total free energy has a cutoff as discussed above and is given by

$$ \Pr(G_i) \simeq \begin{cases} 
\frac{1}{\sqrt{2\pi \sigma_U^2}} \left( e^{-\frac{G_i}{\sigma_U^2}} \right), & G_i < U_{ns} \\
0, & G_i \geq U_{ns}. 
\end{cases} $$

Clearly by tuning the value of $U_{ns}$ the resulting free-energy landscape can be made flat on most of the DNA allowing for a fast one-dimensional diffusion. This happens roughly when the protein is mostly in the non-specific conformation, which yields a first constraint:

$$ e^{-U_{ns}} \gg \left( e^{-U_i} \right) = e^{-\frac{U_i}{2}}. $$

This procedure cannot be carried out in an arbitrary manner as very low values of $U_{ns}$ might destroy the stability of the target.

$\sigma_U^2 \gtrsim 3l_p \left( \sqrt{1 - \frac{2\ln N}{3l_p}} - 1 \right)^2. $ (33)

For $N \sim 10^7$ one may check that both the stability criterion and the speed requirement, equation (31), can only be met for $l_p > 70$. This argument suggests that both speed and stability requirements demand a very large target size. The database studied above does not contain proteins of that size $^9$.

4.2.2. Two-state models. The previous solution relies on having a gapped TF. Another possible resolution of the speed-stability paradox, which applies also to marginally gapped TFs, lies in introducing another conformation of the DNA–TF complex. This conformation, usually attributed to non-specific binding, modifies the properties of the energy landscape experienced by the protein during its one-dimensional diffusion. Specifically, it was suggested [17, 37, 39] that another conformation may introduce an effective cutoff on the TF–DNA-binding energy distribution which will lower its variance and hence lead to a quick one-dimensional diffusion thus resolving the speed-stability paradox.

The two-state model assumes that the protein (or protein–DNA complex) switches rapidly between its two conformations so that the two can be assumed to be equilibrated. We assume that in the first non-specific conformation the protein has a constant binding energy, $U_{ns}$, on all sites and that in the second, specific conformation the binding energy, $U_i$, is, as before, an independent random variable with a Gaussian distribution (3) with a variance $\sigma_U^2$.

The total free energy $E$ has a cutoff as discussed above and is given by

$$ e^{-\frac{U}{2}} \simeq N e^{-\frac{U}{2}}. $$

Thus, to ensure stability we need $E_c < 3(\sqrt{1 - \frac{2\ln N}{3l_p}} - 1)$. Note that a solution for $E_c$ exist only when $l_p \geq \frac{3}{2} \ln N$—as expected the target length has to be large enough to ensure stability. This can be re-expressed in terms of the variance to read

$$ \sigma_U^2 \gtrsim 3l_p \left( \sqrt{1 - \frac{2\ln N}{3l_p}} - 1 \right)^2. $$

During the process of homologous recognition the length of the searcher and the target may be much larger [6].
To avoid this the non-specific energy $U_{ns}$ also has to obey a second constraint

$$N e^{-U_{ns}} < e^{-U_T}, \quad (37)$$

where $U_T$, as before, is the target binding energy.

For marginally gapped TFs these conditions are very restrictive and demand fine tuning:

$$\sigma_U \approx e^{\frac{5}{2} \ln N}. \quad (38)$$

Clearly, most marginally gapped TFs do not satisfy this constraint (see figure 4).

For gapped TFs the constraint is not as severe. It is easy to see that here we need

$$-U_T > \frac{\sigma_U^2}{2} + \ln N. \quad (39)$$

Within the additive binding energy model this implies that $l_p \geq c \ln N$ where $c$ is a constant which depends on $E_c$. It is interesting to check this criterion, which is actually the same as demanding stability of a one state TF in the small disorder regime (see equation (19)) using the protein weight matrices. The results are shown in figure 5. Note that more than half the proteins do not satisfy the criterion. For these the two-state model presented above does not seem to apply.

This gives a clear condition on when this mechanism alone is sufficient to resolve the speed-stability paradox. Recently, such a cutoff was measured in a eukaryotic transcription factor [73]. However, the non-specific binding energy was estimated to be only $5k_B T$ larger than the target energy. The fraction of the time that the protein would spend on the target having such a small energy gap is of the order of $e^5 / N$. In a eukaryotic cell, where one typically has $N \sim 10^{10}$, such a gap looks inexplicably low.

Finally, we comment that all the above considerations ignore the free-energy associated with the protein being off the DNA. At the optimal antenna size (see section 3.3) this has the same contribution as the non-specifically bound state. As mentioned above, it can only reduce the stability on the target.

4.2.3. Multiple transcription factors. In this subsection we assume a single state TF model and comment on the applicability of the results to two-state models at the end. An easy resolution to the slow search speed is increasing the copy number of each TF. When $n_p$ TFs are searching for the target the mean search time is reduced by a factor of $n_p$ (for cases where this does not apply see section 5). For facilitated diffusion the disorder increases the mean search time by a factor of about $e^{\sigma^2 / 2}$. Therefore, $n_p \approx e^{\sigma^2 / 2}$ TFs can compensate for the effects of the disorder.

Multiple TFs may also fulfill the stability requirements since the occupation probability of the target increases with $n_p$. Ignoring interaction between TFs copies the occupation probability of the target, namely the probability to find at least one TF at the target, is given by

$$P^T(n_p) = \frac{n_p P^T(n_p = 1)}{1 + (n_p - 1) P^T(n_p = 1)}. \quad (40)$$

Namely, by taking $n_p$ to be larger than $1 / P^T(n_p = 1)$ the occupation probability of the target becomes of order of 1, so that one satisfies the stability requirement.

However, there is a worry that by increasing the number of TF copies the specificity will be reduced. Namely, the TFs will activate or repress other genes by tightly binding to unwanted sequences on the DNA. Consider the case when in addition to the target there are $N_d$ ‘dangerous’ sites. A site is defined as dangerous if a significant occupation probability of this site affects the transcription of a non-target gene. It seems reasonable to take $N_d$ to be of the same order of magnitude as the total length (in base pairs) of all DNA promoters (of course, evolution could in principle have made this number much smaller). We assume that the binding energy distribution of the dangerous part of the DNA is the same as the rest of the DNA.

The lowest energy among $N_d$ dangerous sites on a typical sequence is given by $-\sigma_U \sqrt{2 \ln N_d}$ so that the occupation probability of the most occupied dangerous site, for a one TF case, is given by

$$P_d(n_p = 1) \approx e^{\sigma_U \sqrt{2 \ln N_d}} \sum_{i=1}^{n_p} e^{-U_i}, \quad (41)$$

while in the case of $n_p$ proteins the occupation probability of the most occupied dangerous site is

$$P_d(n_p) = 1 - \left[ 1 - P_d(n_p = 1) \right]^{n_p}. \quad (42)$$

Thus to ensure that the dangerous part is not significantly occupied $n_p$ has to be much smaller than $1 / P_d(n_p = 1)$.

In summary, three conditions limit the possible value of the TF copy number.

**Rapid search (speed):**

$$n_p \gg e^{\sigma^2}. \quad (43)$$

**Significant occupation probability of the target (stability):**

$$n_p \gg \frac{1}{P^T(n_p = 1)}. \quad (44)$$

**Small occupation probability of a dangerous site (specificity):**

$$n_p \ll \frac{1}{P_d(n_p = 1)}. \quad (45)$$

Of course, these ignore the obvious but hard to quantify cost involved in the production of the TFs.

In appendix C we analyze these conditions in detail. We find that a large copy number can resolve the speed and stability issues without affecting specificity only in the small disorder regime. Finally, we comment that similar considerations hold also for the two-state model discussed above. In that case the specificity condition becomes more stringent for large $N_d$ while for small $N_d$ they are less stringent. Moreover, one can check that adding $n_p$ TFs only eases the criterion given in equation (39) by a $\ln n_p$ term (where we assume that the criterion on the speed in the two-state model is unchanged). This can help the search process only for very large $n_p$ values.
In the previous sections we saw that many proteins have a low occupation probability at the target. Moreover, demanding that the protein reaches the target quickly posed many more constraints on, for example, the length of the protein recognition site and the number of proteins searching for the target. For about 10% of the proteins the problem seems particularly severe. Their occupation probability is so low that they demand thousands of TFs for a high occupation probability. In what follows we suggest a new mechanism which, in principle, may apply to any of the classes above. In particular, in the next section we show how it applies even to TFs with a very low occupation probability through what we call transient stability.

The model assumes that the protein–DNA complex can assume two conformations. Namely, when the protein is bound to the DNA, it can switch between two conformations separated by a free-energy barrier. A closely related model was introduced in previous works in order to solve the speed-stability paradox (see section 4.2.2 and [37, 39, 109]). In these works the barrier between the two states of the protein was assumed to be low enough so that the two conformations were equilibrated with each other. Moreover, the barrier was assumed to be a constant for all DNA sites. A two state structure was demonstrated experimentally in transcription factors [110–113] and type II restriction endonucleases (for a review see [114]). Furthermore, there are simple theoretical arguments for their existence [115].

In contrast to the two-state model discussed above and in previous studies, here an important role is played by a difference in the association rates to different DNA sites. As we show, and intuitively clear, this may supply an additional discriminating factor. Some evidence for the possible importance of association rates is found in the purine repressor. There it was shown that, when activated, it changes its association rate to the target by two orders of magnitude while the dissociation rate changes only by one order of magnitude [116]. Note that the association rate may be very large with a small binding energy and vice versa. Differences between association rates to different DNA sites were also observed in [117, 110]. Interestingly, in the latter work association rates which correspond to very high energetic barriers of tens of $k_B T$ were observed, albeit for eukaryotic cells. In [117] a higher barrier was associated a stronger binding energy.

Assuming two conformations, we call one the search state. In this conformation the protein is loosely bound to the DNA and can slide along it. In the second, recognition state, it is trapped in a deep energetic well (see figure 10). Note that equilibrium measurements of binding energies to the DNA are controlled by the recognition state. To make the discussion clear, below we analyze search processes where the recognition is only based on barrier discrimination. This implies that equilibrium properties of the target site are identical to those of non-target sites.

Based on a quantitative analysis of this model, we argue that due to the occurrence of several time scales in the search process the widely used definition of the reaction rate of a single protein as the inverse of the average search time \( \tau_{\text{ave}} ^{\text{true}} \) [118] is generally irrelevant as a measure of the efficiency of target location on DNA. When \( n_p \) proteins are searching for the target, the relevant quantity is the probability \( R_{n_p} (t) \) for a reaction to occur before time \( t \). We show below that \( R_{n_p} (t) \) can reach values close to 1 on a time scale \( t_{np} ^{\text{ave}} \) which can be orders of magnitude smaller than the average search time, \( \tau_{\text{ave}} ^{\text{true}} \). Both the typical and the average search times can be orders of magnitude smaller than the naive approach based on a one time scale assumption which gives \( \tau_{\text{ave}} ^{\text{true}} / n_p \).

Our analysis has several important merits. First, it reports a fast search time despite a very strong binding of the protein in the recognition state to any site on the DNA. This renders the question of stability in the recognition state irrelevant. We suggest that within this model the measured binding energies of proteins to the DNA are irrelevant to the kinetics of the search process; the relevant quantities are transition rates (specified below). Second, we show that with a proper choice of parameters one may solve the speed-stability paradox without designing the target. We make two comments: (i) while there is
no equilibrium stability within this model it will be shown that the protein is present on the target site for an extended period of time; (ii) within this model the kinetics are independent of the equilibrium properties. Therefore, it is straightforward to add equilibrium stability within it.

The model consists of \( n_p \) proteins which can each be in three states: (i) an unbound state, \( u \), in which it performs three-dimensional diffusion (jumping); (ii) a search state, \( s \), where it is weakly bound to the DNA, performing one-dimensional diffusion (sliding); and (iii) a recognition state, \( r \), where it is tightly bound to the DNA. We assume, for simplicity, that in the recognition state the protein is trapped in a deep energy well (as justified by the experimentally measured strong binding energies) and is unable to move. The transition rates, \( \lambda_i \), between the different states are defined in figure 10. To model sliding, in the \( s \) state the protein can move with transition rate \( \lambda_0/2 \) to neighboring sites on the DNA. Note that the transition rates \( \lambda_i \) and \( \lambda'_i \) expected in general to depend on the location \( i = 1, \ldots, N \) along the DNA. In principle \( \lambda_0 \) and \( \lambda_0 \) also have a dependence on \( i \).

As justified later this will have a weaker effect on our results and we omit it for clarity. After a jump we assume that the protein relocates to a random position on the DNA due to its packed conformation. Finally, the search begins when the protein is produced or activated by a signal, starting from the \( u \) or \( s \) state but not from the \( r \) state. We assume that before the activation (and, obviously, the production) the \( r \) state is inaccessible for the protein.

The presentation of the model gives many details of the derivations of the results. However, we have made an effort to end each subsection with a highlight of the main results. Furthermore, some subsections focus only on results.

5.1. Non-disordered case

To gain an understanding of the difference between the two time scales \( t_{\text{rep}} \), \( t_{\text{ave}} \) and the naive estimation \( t_{\text{ave}}/n_p \) we first consider a single searcher, \( n_p = 1 \), in a simplified model where the transition rates \( \lambda_0 \), \( \lambda'_0 = \lambda_0 \) are independent of \( i \) except at the target site \( T \) (see figure 11(a)). The target site in this section is designed such that the transition rates on the target are different from the transition rates on the rest of DNA. At the target site the transition rate from the \( s \) state to the \( r \) state is denoted by \( \lambda^T \) and the transition rate from the \( r \) state to the \( s \) state is denoted by \( \lambda^s \) (\( \lambda^T \) is irrelevant for the calculation of the first-passage time properties). As stated above, in our considerations we analyze a process of search and recognition based only on barrier discrimination and \( P^T = 1/N \). Therefore, the relation

\[
\lambda^s / \lambda^r = \lambda^T / \lambda^s
\]

holds.

To analyze the model we first consider the probability

\[
R(t) = \int_0^t P(t') \, dt'
\]

that the protein finds its target before time \( t \), where \( P(t) \) is the distribution of the first-passage time (FPT) to the target (we drop the subscript when \( n_p = 1 \)). For simplicity we take a centered target site (labeled 0). The Laplace transform,

\[
\tilde{P}(s) = \int_0^\infty e^{-st} P(t) \, dt,
\]

of \( P(t) \) can be obtained exactly, using. In what follows we first calculate the probability distribution of the number of transitions between the \( s \) and \( r \) conformations during a certain time and the probability to cross the barrier being located on the target site. Combining this with the joint probability density for a protein to find the target at time \( t \) starting from a location \( x_0 \) at \( t = 0 \) before unbinding from the DNA, we calculate the first-passage time probability density. Readers who are not interested in the details of the calculation may skip to the next subsection.

Consider first, the joint probability density for a protein to find the target (in its \( r \) state) at time \( t = t_i + t_r \) starting from a location \( x_0 \) at \( t = 0 \) before unbinding from the DNA. Here \( t_i \) is the total time spent in the \( s \) state and \( t_r \) is the total time spent in the \( r \) state. The probability that exactly \( n \) transitions occurred from the \( s \) state to the \( r \) state is given by \( \mathcal{P}(n, \lambda_s, t_r) \)

\[
\mathcal{P}(n, \mu, t) = (\mu)^n e^{-\mu t} / n!
\]

is a Poisson distribution. The probability to spend a time \( t_i \) in the \( r \) state given that \( n \) transitions occurred from the \( s \) state to the \( r \) state is \( \lambda_r \mathcal{P}(n - 1, \lambda_s, t_r) \) (with the convention \( \mathcal{P}(-1, \mu, t) \equiv \delta(t)/\mu \)). The probability to stay on the DNA up to time \( t = t_i + t_r \), starting at \( t = 0 \) is given by \( e^{-\lambda^s x_0} \).

Finally, the probability to cross the barrier at the target at each visit of its \( s \) state is given by

\[
p_1 = \frac{\lambda^r / \lambda^s}{1 + \lambda^s / \lambda^r + \lambda^s / \lambda^s}.
\]

Therefore, the joint probability density for a protein to find the target at time \( t = t_i + t_r \) starting from a location \( x_0 \) at \( t = 0 \) before unbinding from the DNA is

\[
P_n(t_i, t_r | x_0) = \lambda_r \mathcal{P}(n - 1, \lambda_s, t_r) \mathcal{P}(n, \lambda_r, t_r)
\]

where \( j_{P_1}(t | x_0) \) is the FPT density at the target \( x = 0 \) for a usual random walk starting from \( x_0 \) given that the probability to cross the barrier at the target at each visit in its \( s \) state is \( p_1 \). The FPT density before unbinding starting from \( x_0 \) then reads

\[
J(t | x_0) = \sum_{n=0}^{\infty} \int_0^\infty \int_0^\infty dt_i dt_r \delta(t_i + t_r - t) P_n(t_i, t_r | x_0).
\]

After a Laplace transform and using

\[
\tilde{P}(n, \mu, s) = \frac{\mu^n}{(s + \mu)^{n+1}},
\]

we find

\[
\tilde{J}(s | x_0) = \tilde{j}_{P_1}(u(s) | x_0)
\]

[119]
with \[ u(s) = \frac{s + \lambda_r + \lambda_b + \lambda_{s0}}{s + \lambda_r} \]. (55)

Following [40, 65] we write the probability to find the target, \( P(t) \) as
\[
P(t) = \left\{ \sum_{m=0}^{\infty} \int_0^\infty dt_m J (t_m | x_m) \delta \left( t - \sum_{l=1}^{m-1} (t_l + \tau_l) - t_m \right) \right. \times \prod_{k=1}^{m-1} dt_k J (t_k | x_k) \lambda_k e^{-\lambda_k \tau_k} \}
\] (56)

where \( \langle \rangle_{\{t_l\}} \) denotes an average over the DNA-binding sites and \( J(t|x_0) \) is the probability to unbind before finding the \( r \) state of the target starting from site \( x_0 \). This is given by
\[
\tilde{J} (t|x_0) = \lambda_a e^{-\lambda_a t} \left( 1 - \int_0^t dt' J (t'|x_0) e^{\lambda_a t'} \right).
\] (57)

We assume that each DNA binding event occurs at a random position on the DNA. Thus
\[
P(t) = \sum_{m=0}^{\infty} \int_0^\infty dt_m J (t_m) \delta \left( t - \sum_{l=1}^{m-1} (t_l + \tau_l) - t_m \right) \times \prod_{k=1}^{m-1} dt_k J (t_k) \lambda_k e^{-\lambda_k \tau_k},
\] (58)

where \( J(t) \equiv \langle J(t|x_0) \rangle_{x_0} \) and \( \tilde{J}(t) \equiv \langle \tilde{J}(t|x_0) \rangle_{x_0} \). We then obtain the Laplace transformed FPT distribution as
\[
\tilde{P}(s) = \tilde{J} (s) \left[ 1 - \frac{\lambda_b \lambda_u}{s + \lambda_b} \frac{1 - \tilde{J} (s)}{u(s)} \right]^{-1}.
\] (59)

Using equation (54) and defining \( \tilde{j}_{p_1}(s) \equiv \langle \tilde{j}_{p_1}(s|x_0) \rangle_{x_0} \) one obtains
\[
\tilde{P}(s) = \tilde{j}_{p_1} (u (s)) \left[ 1 - \frac{\lambda_b \lambda_u}{s + \lambda_b} \frac{1 - \tilde{j}_{p_1} (u (s))}{u (s)} \right]^{-1}.
\] (60)

Finally, \( \tilde{j}_{p_1}(s) \) may be calculated using
\[
\tilde{j}_{p_1} (t) = \langle \tilde{j}_{p_1} (t|x_0) \rangle_{x_0} = \langle j (t|x_0) \rangle_{x_0} p_1 + (1 - p_1) \langle j (t|x_0) \rangle_{x_0} \ast j_0 (t) p_1 + (1 - p_1)^2 \langle j (t|x_0) \rangle_{x_0} \ast j_0 (t) \ast j_0 (t) p_1 + \ldots
\] (61)

where \( j(t|x_0) \) is the FPT density at the target \( x = 0 \) for a usual random walk starting from \( x_0 \) and \( j_0(t) \) is the generating function of the first return time to site 0 of a simple random walk. The * symbol denotes a convolution. The Laplace transform of (61) gives
\[
\tilde{j}_{p_1} (s) = \frac{p_1 \tilde{j} (s)}{1 - (1 - p_1) \tilde{j}_0 (s)},
\] (62)

where
\[
\tilde{j} (s) \equiv \langle \tilde{j} (s|x_0) \rangle_{x_0} \simeq \frac{1 + e^{-s/\lambda_0}}{N \sqrt{1 - e^{-s/\lambda_0}}}
\] (63)

and
\[
\tilde{j}_0 (s) \simeq 1 - \sqrt{1 - e^{-2s/\lambda_0}}
\] (64)

for large \( N \) [120] (see appendix B for details).

Applying the Laplace transform to equation (47) and using equation (60) one obtains
\[
\tilde{R} (s) = \frac{\tilde{P}(s)}{s} = \frac{\tilde{j}_{p_1} (u (s))}{s} \left[ 1 - \frac{\lambda_b \lambda_u}{s + \lambda_b} \frac{1 - \tilde{j}_{p_1} (u (s))}{u (s)} \right]^{-1}.
\] (65)

5.1.1. Large barrier regime. By analyzing the pole structure of equation (60) (see appendix E) one can show that in the large barrier regime
\[
\lambda_s \ll \lambda_r \ll \lambda_u, \lambda_b, \lambda_0
\] (66)

(with \( \lambda_u, \lambda_b, \lambda_0 \) of comparable order) the reaction probability simplifies to
\[
R (t) \simeq 1 - q e^{-t/\tau_1} - (1 - q) e^{-t/\tau_2}
\] (67)
\[ R(t) = \frac{1}{\tau} \int_0^t \frac{1}{\lambda_r} \, dt. \tag{67} \]

Rearranging equation (67), we get

\[ R(t) = \frac{1}{\tau} \int_0^t \frac{1}{\lambda_r} \, dt = \frac{1}{\tau} \left[ \frac{1}{\lambda_r} t \right]_0^t = \frac{t}{\lambda_r \tau} = \frac{1}{\tau} \left( 1 - \frac{t}{\lambda_r \tau} \right). \tag{68} \]

From equation (68), we can see that $R(t)$ approaches $1$ as $t$ becomes large, indicating that the protein has reached the target state. The parameter $\lambda_r$ represents the rate at which the target is reached.

Equation (68) shows that the repressor concentration $R(t)$ is inversely proportional to the time $t$ spent in the state and directly proportional to the repressor rate $\lambda_r$. This indicates that as the repressor rate increases, the concentration decreases, unless the time spent is very long.

### Figure 12

A plot of $R(t)$ for $N = 10^6$ (empty circles) and $N = 10^6$ (filled squares) for non-disordered case. Lines correspond to equation (67), with $\tau_1$, $\tau_2$ and $q$ derived analytically (equations (68), (70), (71)). Here $p_1 = 1$, $\lambda_0 = 10^{-5} \lambda_0$, $\lambda_0 = 0.1 \lambda_0$, $\lambda_r = 10^{-3} \lambda_0$ and $\lambda_0 = 10^{-3} \lambda_0$, in agreement with [69]. These correspond to energies, measured relative to the energy of the unbound state, of $E_1 = -4.6 k_B T$, $E_2 = 11.5 k_B T$ and $E_3 = -9.2 k_B T$. The value of $\lambda_0$ was taken to be 1. Experiments suggest $\lambda_0 \approx 10^8 s^{-1}$ for the Lac repressor [28].

This figure illustrates the dynamics of the repressor-lac system, highlighting the impact of repressor rate on the concentration of the repressor over time. The non-disordered case reflects the system's behavior under typical conditions, emphasizing the role of $\lambda_r$ in determining the rate of repressor binding.

### Figure 13

A plot of $R_n(t)$ for $n_p = 1$ (empty circles) and $n_p = 10$ (filled squares) for non-disordered case. Here $N = 10^6$, $p_1 = 1$, $\lambda_0 = 10^{-4} \lambda_0$, $\lambda_0 = 0.1 \lambda_0$, $\lambda_r = 10^{-7} \lambda_0$ and $\lambda_0 = 10^{-9} \lambda_0$ (see [69]). These correspond to energies, measured relative to the unbound state, of $E_1 = -9.2 k_B T$, $E_2 = 6.9 k_B T$ and $E_3 = -13.8 k_B T$. The value of $p_1$ was taken to be 1. Lines correspond to equation (67) with calculated values of $\tau_1$, $\tau_2$ and $q$. Note that here $\lambda_0$ is different from figure 12.

Thus, we see how the repressor concentration evolves over time under different conditions, with $\lambda_0$ and $\lambda_r$ playing key roles in determining the system's dynamics. The figure highlights the impact of these parameters on the repressor's behavior, providing insights into the system's response under various conditions.
experimental data summarized in [69] which are specified in the caption. Since to the best of our knowledge there are no direct measurements of the barrier height for different DNA sequences, we assume this quantity to be of the same order of magnitude as the experimentally measured binding energies [37]. It is found that $\mathcal{R}(t)$ reaches a plateau close to 1 on a typical time scale $t^{\text{typ}}$ which, for $N = 10^6$, is smaller than the average search time $t^{\text{ave}}$ by two orders of magnitude. In next sections we show that results of this simple model may be applied to more realistic models.

5.2. Several searching proteins

The interesting regime $t^{\text{typ}} \ll t^{\text{ave}}$ requires a rather large barrier between the $s$ and $r$ state in the case of long DNA molecules (namely, $\lambda_s \ll \lambda_q k/N$). One might argue that in general this condition may not be met by all proteins. Despite this we now argue that this constraint can be, to a large extent, relaxed. As shown the disorder leaves the mathematical structure of the non-disordered case intact but with effective values of $q$, $\tau_1$ and $\tau_2$. In the next section we discuss the disordered model in detail. As shown the disorder leaves the mathematical structure of the non-disordered case intact but with effective values of $q$, $\tau_1$ and $\tau_2$. Therefore all the results presented below and obtained for the non-disordered case can be easily extended to the disordered one.

5.3. Calculating the average and typical search times

We showed above that the cumulative FPT distribution is given by

$$\mathcal{R}_{n_1}(t) = 1 - \left[ q e^{-t/\tau_1} + (1-q) e^{-t/\tau_2} \right]^{n_0}$$

for the non-disordered model. In this section we calculate the typical and average search times for given values of $q$, $\tau_1$ and $\tau_2$. In the next section we discuss the disordered model in detail. As shown the disorder leaves the mathematical structure of the non-disordered case intact but with effective values of $q$, $\tau_1$ and $\tau_2$. Therefore all the results presented below and obtained for the non-disordered case can be easily extended to the disordered one.

5.3.1. Typical search time. When $\tau_1 \ll \tau_2$, the typical search time $t^{\text{typ}}_{n_1}$, defined through

$$\mathcal{R}_{n_1}(t^{\text{typ}}_{n_1}) = 1 - \frac{1}{e}$$

(77)

can be obtained by assuming $t^{\text{typ}}_{n_1} \gg \tau_1$ or $t^{\text{typ}}_{n_1} \ll \tau_2$ and checking these assumptions self-consistently. Using this method we obtain for $n_1 > 1$

$$t^{\text{typ}}_{n_1} = \begin{cases} \tau_2 \left[ \frac{1}{n_0} + \ln(1-q) \right] & \text{for } n_0 < \frac{1}{\ln(1-q)} \ln\left( \frac{1}{1-q} \right), \\ \tau_1 \ln \frac{q}{e^{-1/n_0} + q - 1} & \text{for } n_0 > \frac{1}{\ln(1-q)} \ln\left( \frac{1}{1-q} \right). \end{cases}$$

(78)

Therefore, for large enough $n_0$ it is widely independent of the binding energy in the $r$ mode.

5.3.2. Average search time. The average search time in the case of $n_0$ proteins is given by

$$t^{\text{ave}}_{n_0} = \int_0^{\infty} (1 - \mathcal{R}(t))^{n_0} dt$$

$$= \sum_{n=0}^{n_0} \binom{n_0}{n} \frac{q^n (1-q)^{n_0-n}}{n!} = \frac{q^{n_0}}{n_0!} (1-q)^{n_0}.$$  

(79)

This sum may be estimated using a saddle point approximation. The saddle point is at $n^* = n_0 q$ as expected (in the limit of a large $\frac{n_0}{n}$ ratio and using the Stirling approximation). Note that the saddle point approximation breaks when $n^* < 1$. In this case the dominant term is $n^* = 0$. When this is not the case we find

$$t^{\text{ave}}_{n_0} = \frac{\tau_2}{n_0} (1-q)^q + \frac{\tau_1}{n_0 q} \frac{1}{\ln\left( \frac{1}{1-q} \right)}.$$  

(80)

In the limit of $\tau_1 \ll \tau_2$ the average time is given by

$$t^{\text{ave}}_{n_0} \simeq \begin{cases} \tau_2 \frac{n_0}{n_0 q} (1-q)^q & \text{for } n_0 < \frac{1}{\ln(1-q)} \ln\left( \frac{1}{1-q} \right), \\ \tau_1 / n_0 q & \text{for } n_0 > \frac{1}{\ln(1-q)} \ln\left( \frac{1}{1-q} \right). \end{cases}$$  

(81)

In figure 14 the average and typical search times are shown and compared with the approximations given by equations (78) and (81). The data shown correspond to a choice of parameters where for $n_0 = 1$ the typical and average search times are roughly the same. Note that there is a large range of $n_0$ values for which $t^{\text{typ}}_{n_0} \ll t^{\text{ave}}_{n_0}$ and that they coincide again at very large values of $n_0$. The range of values of $n_0$ for which the typical and the mean search times differ scales as $\ln(\tau_2/\tau_1)$. Remarkably, for small values of $n_0$ the average search time decreases faster than exponentially with the protein copy number.
Figure 14. In this plot the average and the typical search times and their approximations are shown as a function of the number of proteins, \( n_p \). Circles represent the average search time. Squares represent the typical search time, \( t_{\text{typ}} \) (77). The blue, solid and thick line represents the approximation for the average time (81), the dashed red line represents the approximation for the typical search time (78) and the thin black line represents the naive estimate \( \frac{\tau_{\text{typ}}}{\sqrt{nm}} \).

Parameters chosen for this plot are \( \tau_1 = 1 \), \( \tau_2 = 10^6 \) and \( q = 0.1 \).

5.4. Disordered case

In this section we study a disordered version of the model. Since the barrier plays a key role in the search we focus on effects of disorder in its height. To account for this we consider the case where the barrier height, \( E^t_b \), is drawn from a Gaussian distribution:

\[
p(E^t_b) = \frac{e^{-\frac{(E^t_b - E_b^0)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma}} \tag{82}
\]

such that the transition rate from the site to the state at site \( i \) is given by

\[
\lambda^t_i = \lambda_0 \min\left(1, e^{-E^t_b}\right). \tag{83}
\]

Note that \( \lambda_0 \) is a microscopical time scale related to sliding. In principle a different prefactor could occur for the barrier crossing. This constant, however, can be absorbed by a redefinition of the barrier height. Transitions over negative barrier are controlled by \( \lambda_0 \). However, the contribution of these times to the overall search process are negligible. Introducing an energy difference between the state and the state (taken to be equal for all sites), \( E_r \), the transition rate from the state to the state at site \( i \) is given by

\[
\lambda^r_i = \lambda_0 e^{E_r} \min\left(1, e^{-E^r_b}\right) = \lambda^r_i e^{E_r}. \tag{84}
\]

Similar to the affinity properties of the target site that is typically very close to the highest affinity among the non-target sites (see discussion above), we propose an intrinsic definition of the target as the site with the lowest barrier with no specifically designed properties (see figure 11(b)). Indeed, our previous assumption in section 5.1 that \( \lambda^t_i \) is large at the target site and \( \lambda^r_i \) small everywhere else is a rather strong demand and corresponds to a designed target. Although we show below that the barrier discrimination mechanism may supply an efficient search even for the non-designed target, any special design of the target may significantly increase the search effectiveness.

In the next subsection we analyze the disordered model using a mean-field approach and check the results using a numerical simulation in section 5.4.2.

5.4.1. Mean-field analysis. Within the mean-field approach we replace the different quantities by their disorder average and account for the barrier at the target site. We first compute the disorder averaged probability of crossing the barrier at the target at each visit. The probability density of the barrier height on the target site, \( E^t_b \), (the probability density of the minimal energy among \( N \) normally distributed identical and independent random variables with a mean \( E_0 \) and variance \( \sigma^2 \)) is [80]

\[
\Pr\left(E^t_b\right) = \frac{d}{dE^t_b} \left[ \int_{-\infty}^{E^t_b} e^{-\frac{(E-\bar{E}^t_b)^2}{2\sigma^2}} dE \right]^N \\
= -\frac{d}{dE^t_b} \left[ \frac{1}{2} \text{erfc}\left(\frac{E^t_b - E_0}{\sqrt{2\sigma}}\right)\right]^N. \tag{85}
\]

For a given value of \( E^t_b \), the probability to pass over a barrier to the state is \( e^{-E^t_b} \). Thus the disorder averaged probability of crossing the barrier at the target at each visit is given by

\[
\bar{p}_i = \int_{-\infty}^{\infty} dE^t_b \left[ \frac{1}{2} \text{erfc}\left(\frac{E^t_b - E_0}{\sqrt{2\sigma}}\right)\right]^N. \tag{86}
\]

Here we set the time scale of the activation process across the barrier to be \( \lambda_0 \). We finally assume that the expression for \( u(s) \) of the non-disordered model (67) holds with \( \lambda^t_i \) replaced by its average over the barrier energy. Using equation (83) this is given by

\[
\bar{\lambda}_i = \lambda_0 \left[ \int_{-\infty}^{0} \frac{e^{-\frac{(E^t_b - E^r_b)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma}} dE_b + \int_{0}^{\infty} e^{-E_b} \frac{e^{-\frac{(E^t_b - E^r_b)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma}} dE_b \right]. \tag{87}
\]

Also using equation (84), within the mean-field approximation \( \lambda^r_i \) is replaced by \( \bar{\lambda}_i e^{E_r} \) and \( p_i \) replaced by \( \bar{p}_i \). In the next section we check these results numerically.

5.4.2. Numerical results and comparison with mean-field analysis. We now check the mean-field results using numerics. First, we show that the two scales scenario described above still holds. Indeed, figure 15 shows that \( R(t) \) is well fitted by equation (67) for realistic values of parameters. Note that in figures 15 and 16 we have chosen the worst scenario \( E_r = -\infty \) so that \( \lambda^r_i = 0 \) and the average search time and the value of \( \tau_2 \) are both infinite. This implies that for \( n_p \) large enough the only relevant time scale is \( \tau_1 \) and the typical search time again takes the form \( \bar{t}^{\text{typ}}_n \sim \frac{m}{m_n} \) (the detailed calculation of the typical and average times is presented above in section 5.3).
are obtained for longer DNA sequences. The parameters for this plot (filled squares) for the disordered model. The lines were obtained by fitting the form \(1 - (qe^{-t/\tau_1} + (1 - q))^{p}\) to the numerical simulations with \(q = 0.2817, \lambda_0 t_1 = 1.7 \times 10^5\) and \(\tau_1 = \infty\). These are close to the mean-field prediction \(q = 0.2827, \lambda_0 t_1 = 1.1 \times 10^5\). Here \(\lambda_u = 10^{-2}\lambda_0, E_r = -4.6k_B T, \lambda_b = 0.1\lambda_0, E_0 = 25.4k_B T, \dot{E}_r = -\infty\) and \(\sigma = 5.3k_B T\). Note that here the average height of the barrier at the target site is 6.25k_B T.

![Figure 15](image1.png)

Figure 15. Plot of \(R_c(t)\) for \(n_p = 1\) (empty circles) and \(n_p = 10\) (filled squares) for the disordered model. The lines were obtained by fitting the form \(1 - (qe^{-t/\tau_1} + (1 - q))^{p}\) to the numerical simulations with \(q = 0.2817, \lambda_0 t_1 = 1.7 \times 10^5\) and \(\tau_1 = \infty\). These are close to the mean-field prediction \(q = 0.2827, \lambda_0 t_1 = 1.1 \times 10^5\). Here \(\lambda_u = 10^{-2}\lambda_0, E_r = -4.6k_B T, \lambda_b = 0.1\lambda_0, E_0 = 25.4k_B T, \dot{E}_r = -\infty\) and \(\sigma = 5.3k_B T\). Note that here the average height of the barrier at the target site is 6.25k_B T.

![Figure 16](image2.png)

Figure 16. Results for the disordered model. Here \(N = 10^6\), \(\lambda_u = 10^{-2}\lambda_0, E_r = -4.6k_B T, \lambda_b = 0.1\lambda_0, E_0 = 25.4k_B T\). (a) \(p_{cat}\) as a function of \(\sigma\) for \(n_p = 1\) and \(n_p = 10\). (b) \(t_{typ}\) for \(n_p = 10\) and \(t_1\) are plotted as a function of \(\sigma\). Using \(\lambda_0 = 10^6\text{s}^{-1}\) [28] for \(n_p = 10\) at the minimal \(p_{cat}\) we find \(t_{typ} \approx 10\text{s}\). Note that by moderate changes in \(E_0\) similar results can be obtained for longer DNA sequences. The parameters for this plot are \(t_1 = 1, t_2 = 10^4, q = 0.1\).

This enables a fast search even in the presence of very deep (even infinite) traps.

The regime of a fast search with \(t_{typ}\) independent of the trap depth \(E_o\) also requires, as above, a small catastrophe probability, \(p_{cat} = (1 - q)^{p}\) (see equation (75)). We now show that this condition holds in a wide range of disorder parameters, \(E_0\) and \(\sigma\). To illustrate this, the dependencies (holding all other variables constant) of \(p_{cat}\) and \(t_{typ}\) on \(\sigma\), obtained from numerics and the mean-field treatment, are shown in figure 16 for realistic values of parameters. Notably, the dependence of the catastrophe probability on the disorder strength is not monotonic so that the value of \(p_{cat}\) can be minimized as a function of \(\sigma\). This reflects the fact that for small values of \(\sigma\) the DNA sequence has to be scanned many times before the target enters in the \(r\) mode. Increasing \(\sigma\) lowers the barrier at the target and therefore reduces the number of scans needed, which diminishes \(p_{cat}\). For larger \(\sigma\) the chance of falling into a trap increases due to lower secondary minima of the barrier, which leads to an increase in \(p_{cat}\). As expected, \(p_{cat}\) is dramatically decreased when \(n_p\) is increased, even by a few units, and can remain small for a wide range of values of \(\sigma\). For larger \(\sigma\), \(p_{cat}\) increases and \(t_{typ}\) rises quickly as it starts to depend on \(t_2\).

Summarizing and using the results of section 5.1, the mean-field approach predicts that in the high barrier regime \(\lambda_u \ll \lambda_r \ll \lambda_u, \lambda_b, \lambda_0\) (with \(\lambda_u, \lambda_b, \lambda_0\) of comparable order) the reaction probability simplifies to

\[
R(t) \simeq 1 - qe^{-t/\tau_1} - (1 - q)e^{-t/\tau_2},
\]

with

\[
q = \left( 1 + \frac{\lambda_r}{\lambda_u k_b N} \right)^{-1},
\]

\[
\kappa = \left( \frac{\lambda_r}{\lambda_u k_b N} \right)^{1/2},
\]

\[
\tau_1 = \frac{\lambda_u + \lambda_r}{\lambda_b + \lambda_u + \lambda_r + \frac{\lambda_r^2}{N}},
\]

and

\[
t_2 = \frac{\lambda_r + \frac{\lambda_r^2}{N}}{\lambda_b + \lambda_u + \frac{\lambda_r^2}{N}}.
\]

In the case of a few proteins, searchers that fall into traps tend to occupy sites with low barriers and, therefore, increase the probability of other TFs to reach the target. Thus, equation (74), in which the searchers are assumed to be independent, provides a lower bound on the probability to reach the target. Here and below we assume that the number of proteins, \(n_p\), is small enough (compared with \(N\)) such that this effect does not play a role and equation (74) is applicable.

Most important, as advertised above, these results show that it is possible to obtain relatively small values of \(t_{typ}\) and \(p_{cat}\) with realistic values of the parameters (see figure 16). Reasonable search times (in the range of seconds) are obtained for a rather large range of \(\sigma\) as long as \(n_p\) is of the order of ten or more proteins suggesting another possible resolution of the speed and stability requirements. We stress that this mechanism can apply to any of the classes of TFs discussed above. This is a direct consequence of the decoupling of the stability and speed requirements. We note that by moderate changes in \(E_0\) similar results can be obtained for much longer DNA sequences. In appendix F we show that by increasing the disorder strength and the average barrier height such that

\[
E_0 = \frac{\sqrt{2}erfc^{-1} \left( \frac{2}{N} \right)}{\sigma},
\]
6. Effective model and outcomes

As we showed in the previous section, by only using barrier discrimination between different DNA sites a transcription factor may, in principle, serve as an efficient searcher and its complex with the target can be arbitrarily stable. Experiments show that different DNA sites are discriminated by their binding energy. Therefore, if a barrier mechanism is at work it is likely to be combined with an energetic discrimination between different sites.

Nonetheless, it is interesting to consider a scenario where there is only barrier discrimination. This could apply for TFs which have a very small target occupation probability \( P^T \). As we now show a barrier mechanism may lead to a high transient occupation probability of the target even with no energetic discrimination. When active processes are included the occupation probability can be made large even in the long-time limit. Furthermore, and in a more speculative manner, we show how the barrier mechanism can lead to a dynamical ordering of gene activation.

To show these we construct an effective model which uses the simple resulting mathematical structure of the previous section. Specifically, we use the cumulative probability

\[
\mathcal{R}(t) = 1 - q e^{-\frac{t}{\tau}} - (1-q) e^{-\frac{t}{\tau}}.
\]  
(95)

In our discussion we concentrate on the target occupation probability. This fact and the simplicity of expression (95) allow one to describe our system using a three states model. Within this approach we only consider the \( r \) state on the target (\( T \)), the \( r \) state off the target (\( D \)) and one state for all other configurations (including \( s \) states and the unbound state) (\( U \)). The transition rates between the states are defined as follows: \( \lambda^D \) is the transition rate from \( U \) to \( D \), \( \lambda^{D-1} \) is the transition rate from \( D \) to \( U \), \( \lambda^T \) is the transition rate from \( U \) to \( T \) and \( \lambda^{T+1} \) is the transition rate from \( T \) to \( U \). The model is illustrated schematically in figure 17. As shown below, this simplification allows us to analyze the behavior of the system beyond its FPT properties.

To proceed we first show that the effective model yields the same cumulative FPT distribution \( \mathcal{R}(t) \) as the original system. Specifically, it is straightforward to show that

\[
\mathcal{R}(t) = 1 - \frac{\lambda^+ - \lambda^D}{\lambda^+ - \lambda^D} - \frac{\lambda^- - \lambda^D}{\lambda^- - \lambda^D} e^{-\frac{t}{\tau}},
\]  
(96)

where

\[
\lambda^\pm = \frac{\lambda^{D-1} + \lambda^D + \lambda^T}{2} \pm \sqrt{(\lambda^{D-1} + \lambda^D + \lambda^T)^2 - 4\lambda^T\lambda^{D-1}}.
\]  
(97)

Comparing equation (67) with equation (96) one obtains relations between transition rates of the effective model and the full model:

\[
\frac{\lambda^+ - \lambda^D}{\lambda^+ - \lambda^-} = q,
\]  

\[
\lambda^T = N \lambda^T \frac{\lambda^+}{\lambda^-} \simeq N \frac{q}{2\tau_2},
\]  

\[
\lambda^{D-1} \ll \lambda^D.
\]  

As stated above we consider a case where the binding energies of all sites (including the target) are the same so that in equilibrium the occupation probability of the target site is equal to the occupation probability of all other sites on the DNA. In this case, in equilibrium the occupation probability of the \( D \) state is \( N \) times larger than the occupation probability of the \( T \) state. This implies within the simplified model that

\[
\frac{\lambda^D}{\lambda^{D-1}} = N \frac{\lambda^T}{\lambda^-},
\]  

so that

\[
\lambda^{T+1} = N \frac{\lambda^D}{\lambda^D} \simeq N \frac{q}{2\tau_2}.
\]  

Thus, we showed that a simple three states effective model has the same dynamical and equilibrium properties as
the original system. Below we use the effective model to analyze the search dynamics beyond FPT properties. For example, we consider equilibration dynamics, the possible existence of an active processes and temporal ordering in the activation/repression of multiple targets.

6.1. Transient behavior

Following the above the occupation probability of the target site, $P^T(t)$, evolves as

$$
\frac{\partial P^T}{\partial t} = \lambda^T \left(1 - P^T - P^D\right) - \lambda^{-1}_T P^T,
$$

$$
\frac{\partial P^D}{\partial t} = \lambda^D \left(1 - P^T - P^D\right) - \lambda^{-1}_D P^D,
$$

$$
P^D = 1 - P^T - P^D,
$$

(103)

where $P^D(t)$ is the occupation probability of the $D$ state, $P^U(t)$ is the occupation probability of the $U$ state and the initial conditions are $P^U(t = 0) = 1$ so that $P^T(t = 0) = P^D(t = 0) = 0$. These equations may be solved exactly. However, here we analyze the equations by noting that there are three time regimes. For $t \ll \frac{1}{\lambda^T}$, the occupation probability of the target is close to its initial value, i.e. $P^T \simeq 0$. For $\frac{1}{\lambda^T} \gg t \gg \frac{1}{\lambda^D}$ the protein equilibrated with the target but not with the rest DNA so that

$$
P^T \simeq \frac{1}{1 + \frac{\lambda^T}{\lambda^D}}.
$$

Of course, this regime exists only when $\lambda^T \gg \lambda^D$. For $t \gg \frac{1}{\lambda^D}$ the system reaches thermal equilibrium and the target occupation probability is given by

$$
P^T \simeq \frac{1}{1 + \frac{\lambda^T}{\lambda^D} \left(1 + \frac{\lambda^D}{\lambda^{-1}_T}\right)}.
$$

In figure 18 the occupation probability of the target site, $P^T(t)$, is shown. Since there is no binding energy discrimination between DNA sites the occupation probability of the target in the long time limit is very small, $P^T = \frac{1}{\lambda^T}$. Note, however, that there is a transient regime where the occupation probability is large. In fact, in this regime the TF binds and unbinds many times from the target site before the system reaches thermal equilibrium.

The above discussion may be generalized to the case of a few proteins, $n_p > 1$. In this case a mean-field generalization of (103) is

$$
\frac{\partial n^T}{\partial t} = \lambda^T \left(n_p - n^T - n^D\right) \left(1 - n^T\right) - \lambda^{-1}_T n^T,
$$

$$
\frac{\partial n^D}{\partial t} = \lambda^D \left(n_p - n^T - P^D\right) - \lambda^{-1}_D n^D,
$$

$$
n^U = n_p - n^D - n^T
$$

(104)

with the initial conditions $n^U(t = 0) = n_p$ and $n^T(t = 0) = n^D(t = 0) = 0$. Here $n^T, n^D$ and $n^U$ represent the mean-occupation number at the target in the $T, D$ and $U$, respectively. The numerical solution of this non-linear equation is shown in figure 18. The qualitative behavior is similar to the $n_p = 1$ case: there are three time regimes. Using the same arguments as the $n_p = 1$ case we find that, for $t \ll \frac{1}{\lambda^T n_p}$ we have $n^T \simeq 0$. For $\frac{1}{\lambda^T} \gg t \gg \frac{1}{\lambda^D n_p}$ we have

$$
n^T \simeq \frac{1}{1 + \frac{\lambda^T}{n_p} \frac{\lambda^D}{\lambda^{-1}_T}},
$$

and for $t \gg \frac{1}{\lambda^D}$ the mean-occupation number is given by

$$
n^T \simeq \frac{1}{1 + \frac{\lambda^T}{n_p} \frac{\lambda^D}{\lambda^{-1}_T} \frac{1}{\lambda^D}}.
$$

The intermediate regime, corresponding to a transient high occupation of the target, exist when $\lambda^T n_p \gg \lambda^D$. Its easy to check that for a large enough number of proteins, $n_p \gg \frac{1}{\lambda^D}$, this regime exists even when $\lambda^T < \lambda^D$.

Using this analysis we have shown that when the target site differs from the rest by a low barrier between the transcription factor’s $r$ and $s$ states its occupation has a transient nature. After a change in the environment, which activates the transcription factors, the occupation probability of the target increases exponentially with a fast time constant and after this decreases exponentially with a slow time constant to its final value. When the only discrimination between sites is the barrier height the final occupation probability of the target is very small, so that in the long time limit the system is in the same state as it was before the activation of the protein. By introducing a free-energy binding energy discrimination between sites, the final occupation probability of the target may be significant such that the long time limit of the system may be different from the initial ‘pre-activated’ state.

In general, when subjected to a change in the environmental condition, a cell typically responds by increasing the activity level of certain genes and decreasing the activity level of others. In many cases, the expression level of a certain gene changes temporarily, exhibiting a sharp increase or decrease, and later changing again, reaching a new

![Figure 18](image_url)

Figure 18. The occupation probability of the target site in the simplified model, $P^T$, is shown as a function of time. The parameters are $\lambda^T = 0.02$, $\lambda^{-1}_T = 0.05$, $\lambda^D = 10^{-3}$ and $\lambda^{-1}_D = 5 \times 10^{-5}$. The blue solid (red dashed) line represents the $n_p = 1 (n_p = 20)$ case.
steady state (which often is similar to the original state). This, two-step transient behavior, is widely observed in different transcriptional responses, from yeast [122, 123] to human [124] and may be explained by a negative feedback of an activated protein [125]. In this section we showed how this kind of behavior naturally arises in a regulation system (composed of only one transcription factor) based on barrier discrimination between distinct DNA sites.

6.2. Steady state and the existence of an active process

In equilibrium the occupation probability on a DNA site depends only on its binding energy. In cases where the only difference between the target and non-target sites is the barrier height between the s and r states, after the equilibration the probability of finding the protein on the target site is very small. As we now show, by introducing an active process that returns the searcher to the initial state, u, with a transition rate $\Omega$ (see figure 19) from any state, one may obtain a high occupation probability of the target site even at steady state. This active process may be loosely thought of as cell division or degradation and production of the protein.

In this case, for $n_p = 1$ the equation for the occupation probabilities are given by

$$\frac{\partial P^T}{\partial t} = \lambda^T (1 - P^T - P^D) - (\lambda_{-1} + \Omega) P^T,$$

$$\frac{\partial P^D}{\partial t} = \lambda^D (1 - P^T - P^D) - (\lambda_{-1} + \Omega) P^D.$$ (105)

In the steady state ($\partial P^T / \partial t = 0$) the target site occupation probability is, therefore,

$$P^T = \lambda^T \frac{1 + \frac{\lambda^T}{\Omega}}{\Omega + \left(\frac{\lambda^T}{\Omega} + \lambda^D + \lambda^D + \lambda_{-1}\right)}.$$ (106)

If in the absence of an active process ($\Omega = 0$) the steady-state occupation of the unbound state is small, the ‘on’ rates, $\lambda^D$ and $\lambda^T$ are much larger than the ‘off’ rates, $\lambda_{-1}$ and $\lambda_{-1}$. In this case one obtains three regimes depending on the value of $\Omega$:

$$p^T \simeq \begin{cases} \frac{\lambda^T}{\lambda^D + \lambda^T + \lambda_{-1}} & \Omega \ll \lambda^D_{-1}, \lambda^T_{-1}, \\ \frac{1}{1 + \frac{\lambda^D}{\lambda^T}} & \lambda^D, \lambda^T \gg \Omega \gg \lambda^D_{-1}, \lambda^T_{-1}, \\ \frac{\lambda^T}{\Omega} & \Omega \gg \lambda^D, \lambda^T. \end{cases}$$ (107)

Note that in the second regime the occupation of the target site can be significant.

Similar to above, the approach can be generalized to several proteins (we use the same notation as in the previous subsection). When a few proteins act together the mean-field equations for the occupation probabilities are given by

$$\frac{\partial n^T}{\partial t} = \lambda^T (n_p - n^T - n^D) (1 - n^T) - (\lambda_{-1} + \Omega) n^T,$$

$$\frac{\partial n^D}{\partial t} = \lambda^D (n_p - n^T - n^D) - (\lambda_{-1} + \Omega) n^D.$$ (108)

Here $n^T$ and $n^D$ are the mean-occupation numbers in the $T$ and $D$ state, respectively. Assuming, as before, that without an active process the protein in equilibrium spends most of its time bound to the DNA we obtain in steady state

$$n^T \simeq \begin{cases} \frac{\lambda^T n_p}{\lambda^D + \lambda^T + \lambda_{-1}} & \Omega \ll \lambda^D_{-1}, \lambda^T_{-1}, \\ \frac{1}{1 + \frac{\lambda^D}{\lambda^T}} & n_p \lambda^D, n_p \lambda^T \gg \Omega \gg \lambda^D_{-1}, \lambda^T_{-1}, \\ \frac{\lambda^T}{\Omega} n_p & \Omega \gg n_p \lambda^D, n_p \lambda^T. \end{cases}$$ (109)

The optimal $\Omega$ (that maximizes the steady-state occupation of the target, $n^T$) is independent of $n_p$ and given by

$$\Omega_{\text{opt}} = \sqrt{\lambda^T_{-1} \lambda^D - \lambda^D_{-1} \lambda^D - \lambda^D_{-1}}.$$ (110)

In figure 20 the steady-state probability of the target site as a function of the rate of the active process, $\Omega$, is shown.

Summarizing, non-equilibrium effects of the barrier discrimination between DNA sites may lead to a high target occupation probability even at steady state.

6.3. The possibility of genetic temporal ordering

It is often the case that each TF activates more than one gene [126]. For example, in E. coli there are 68 transcription factors which individually regulate more than 13 operons [9, 127]. In some cases the activation of different genes, regulated by the same TF, are temporally ordered [128–130]. In these systems it seems that the temporal ordering is not caused by the transcriptional network (for example, by a genetic cascade). It was suggested [37, 125] that different genes have different activation thresholds. In this case a temporarily increased concentration of the transcription factor activates
them one-by-one. Different thresholds arise from non-linear effects, such as cooperativity between the transcription factors. Recently [131–133] it was proposed that genetic temporal ordering may be influenced by different distances between the production location of the TF and its target (this mechanism seems plausible only for prokaryotic cells).

Here we show that a search mechanism based on barrier discrimination can also lead to temporal ordering. This does not rely on cooperativity and appears even for a TF with a constant concentration. To show this we generalize the effective three states model to four states by adding an additional target site (see figure 21). The states of the model are the $r$ state of the first target ($T_1$), the $r$ state of the second target ($T_2$), the $r$ state out of both targets ($D$) and the $U$ states (including the $s$ state and the unbound state).

For $n_p = 1$ the evolution equations for the occupation probability of the first target, $P^{T_1}$, the second target, $P^{T_2}$, and the rest of the DNA, $P^D$ are given by

$$\frac{\partial P^{T_1}}{\partial t} = \lambda_1^{T_1} \left(1 - P^{T_1} - P^{T_2} - P^D\right) - \lambda_{21}^{T_1} P^{T_1},$$
$$\frac{\partial P^{T_2}}{\partial t} = \lambda_2^{T_2} \left(1 - P^{T_1} - P^{T_2} - P^D\right) - \lambda_{21}^{T_2} P^{T_2},$$
$$\frac{\partial P^D}{\partial t} = \lambda^D \left(1 - P^{T_1} - P^{T_2} - P^D\right) - \lambda_{11}^D P^D$$

while the occupation probability of the $U$ state is determined by

$$P^U = 1 - P^{T_1} - P^{T_2} - P^D.$$  

In the case of few proteins the mean-field equations for the evolution of the occupation probability are

$$\frac{\partial n^{T_1}}{\partial t} = \lambda_1^{T_1} \left(n_p - n^{T_1} - n^{T_2} - n^D\right) \left(1 - n^{T_1}\right) - \lambda_{21}^{T_1} n^{T_1},$$
$$\frac{\partial n^{T_2}}{\partial t} = \lambda_2^{T_2} \left(n_p - n^{T_1} - n^{T_2} - n^D\right) \left(1 - n^{T_2}\right) - \lambda_{21}^{T_2} n^{T_2},$$
$$\frac{\partial n^D}{\partial t} = \lambda^D \left(n_p - n^{T_1} - n^{T_2} - n^D\right) - \lambda_{11}^D n^D$$

with

$$n^U = n_p - n^{T_1} - n^{T_2} - n^D.$$  

Here $n^{U}, n^{T_1}, n^{T_2}$ and $n^D$ are the mean-occupation numbers in the $U, T_1, T_2$ and $D$ states, respectively. These equations may be solved analytically for the $n_p = 1$ case and for the $n_p \gg 1$ case. In figure 22 one may see that by the tuning transition rates it is possible to obtain a temporal ordering of gene activation or/and repression for a single protein and few ($n_p = 10$) proteins. Genes $T_1$ and $T_2$ are activated or repressed at different times depending on their association rates. The subsequent deactivation or/and repression of genes $T_1$ and $T_2$ take place if $\frac{\lambda_{21}^{T_1}}{\lambda_1^{T_1}} \gg 1$ and occurs also at different times depending on their dissociation rates. These results, in principle, can be generalized to any number of genes.

Figure 20. The occupation probability of the target site in the steady state as a function of $\Omega$. The parameters are $\lambda^{T_1} = 1, \lambda_1^{T_1} = 10^{-3}, \lambda^{D} = 0.5$ and $\lambda_1^{D} = 10^{-5}$. The solid blue (dashed red) line represents the $n_p = 1 (n_p = 15)$ case.

![Figure 20](image1.png)

Figure 21. Schematic representation of the four-states effective system for temporal ordering. Here $T_1$ denotes the $r$ state of the first target, $T_2$ denotes the $r$ state of the second target, $D$ denotes the $r$ state out of both targets and $U$ denotes the remaining states (including the $s$ state and the unbound state).

![Figure 21](image2.png)

Figure 22. In this graph the occupation probabilities of two target sites, $P^{T_1}$ (blue, thick lines) and $P^{T_2}$ (red, thin lines), are shown as a function of time. The parameters are $\lambda^{T_1} = 0.05, \lambda_1^{T_1} = 0.005, \lambda^{T_2} = 5 \times 10^{-1}, \lambda_2^{T_2} = 10^{-3}, \lambda^{D} = 10^{-3}$ and $\lambda^{D} = 10^{-4}$. The solid (dashed) lines represent the $n_p = 10 (n_p = 1)$ case.
7. Summary

Search and recognition problems appear in many contexts in biological systems. Organisms activate/repress many processes by specifically designed proteins. This may be an enzyme that catalyzes some chemical reaction by binding to specific molecules, a transcription factor that changes the transcriptional activity by binding to some specific locations on the DNA, etc. In fact, the flow of information, from a genotype to a phenotype and vice versa is regulated and implemented by searchers for specific DNA/RNA sequences. Needless to say, every particular biological searcher ‘invented’ its own search strategy. However, there is hope that it is possible to divide all search strategies to a few classes similarly to the division of the transcription factors’ binding domains to a few DNA-binding motifs [134]. Different aspects of a transcription factor may dictate its strategy: the number of copies in the cell, its structure and function, its interactions with other proteins, the number of its targets and many more. The beginning of this review suggested a possible classification of TFs. It is certainly of great interest to extend it and test it on large databases. A particular simple classification arose when the length of the binding site of the protein to the DNA was considered.

The bulk part of the review deals with a possible scenario where TFs locate their target using barrier discrimination. This is different from other mechanisms which assume a local equilibration of the TF with its environment before the target is located. A more detailed discussion of those can be found in many other reviews [14, 15, 65–68]. When assuming equilibration, each site on the DNA can be characterized by a single parameter—the binding energy at the site. According to the equilibrium assumption, this is the only parameter that may discriminate between different sites. In contrast in the barrier model the search kinetics can become widely independent of the binding energy, and controlled by the barrier height. This gives a similarly simple resolution to the speed-stability paradox described in detail in this review. Moreover, it suggests that TFs can be weakly bound to the target site for extended periods. We termed this transient activation. The end of the review dealt with more speculative processes that may occur in a barrier mechanism. These include time ordering in the activation of genes.

Under the conditions described above the search process in the barrier model is not characterized by a single time scale but by two: one short and one long. This leads to a distinction between typical search times, which are relevant to experiments, and average search times. The latter are dominated by rare events.

A clear indication that the barrier mechanism is at work in vivo would lie in measurements of the full FPT distribution of a TF at a target gene, and the observation of more than one time scale. Even if the FPT distribution shows only a single time scale, the model predicts the appearance of a new short time scale as \( n_p \), the number of searchers, is increased. Therefore, the appearance of the new time scale would be a clear indication for the presence of the mechanism. In this case the typical search time will generally not scale as \( 1/n_p \). Such experimental data are not yet available but in principle accessible, and will hopefully provide in the near future a better understanding of the kinetics of gene activation.

Acknowledgments

We are grateful for many stimulating and critical discussions with E Braun, A Finkelstein, K Keren and D Levine.

Appendix A. An information theoretic approach to the calculation of disorder parameters and binding energies

In sections 2 we characterized the ability of a protein to recognize its target by the free-energy gap between the target and the rest of the DNA sites and the equilibrium target occupation probability. As an alternative approach one often uses the information content of a protein, denoted by IC [135, 136]. While our personal preference is to the presentation used in the main text the two can be used interchangeably. In this appendix we provide a brief overview of the relations between the information theory quantities (the information content and the sequence score) and the physical (the disorder strength and the binding energy of a sequence) quantities.

A.1. The information content and the disorder strength

Before turning to the information content of a protein of length \( l_p \) we first discuss the information content associated with a single binding site \( i \). Assuming that the frequency of each nucleotide in the genome is close to \( 1/4 \) [137], this quantity, denoted by IC\(_i\), is given by its maximal amount of information minus its Shannon entropy. Note that we consider the IC of the ‘specific’ protein conformation. The non-specific one, presumably, does not contain any information. With this in mind we have

\[
\text{IC}_i = - \log_2(1/4) + \sum_{s \in \{A,T,C,G\}} \text{Pr}(s,i) \log_2 \text{Pr}(s,i),
\]

where Pr\((s,i)\) is defined in equation (2). This quantity is maximal when only one nucleotide type can bind to site \( i \). In this case Pr\((s,i) = \delta_{s,s^*}\) so that IC\(_i\) = 2bits. If two types of nucleotides can bind with equal probability (Pr\((s,i) = \frac{1}{2} \delta_{s,s^*} + \frac{1}{2} \delta_{s,s^t}\)) the information content is reduced to one bit. In a case when each of the four types of nucleotides can bind with an equal probability of \( \frac{1}{4} \) the information content is zero.\(^{12}\)

With these definitions the total information content of the protein is given by a sum of information contents of all protein binding sites:

\[
\text{IC} = \sum_{i=1}^{l_p} \text{IC}_i.
\]

\(^{12}\) Remarkably the average information content of one binding site of a TF is 1.054 bits for TFs from the RegulonDB database [83].
To identify the target this has to be larger than $\log_2 N$. Using equations (1), (2) and (A2) we obtain

$$IC = 2l_p + \sum_s e^{-U(s)} \left( \sum_s e^{-U(s)} \right) \log_2 \sum_s e^{-U(s)}$$

$$= 2l_p - \sum_s \frac{U(s) e^{-U(s)}}{\sum_s e^{-U(s)}} - \log_2 \sum_s e^{-U(s)}$$

$$= 2l_p - \left( 1 - \frac{\partial}{\partial p} \right) \left( \log_2 \sum_s e^{-pU(s)} \right)_{p=1}. \quad (A3)$$

Since, as we mentioned above, $U(s)$ behaves to a good approximation as a Gaussian random variable the information content is given by

$$IC = 2l_p - \left( 1 - \frac{\partial}{\partial p} \right) \left( \log_2 \sum_{i=1}^{4^p} e^{-pU_i} \right)_{p=1}. \quad (A4)$$

where \{U_i\} is a set of Gaussian random variables with a probability density

$$Pr(U_i) = \frac{e^{-\frac{U_i}{\sigma_U^2}}}{\sqrt{2\pi \sigma_U^2}} \quad (A5)$$

and the angular brackets denote an average over realizations of disorder. The expression $\log_2 \sum_{i=1}^{4^p} e^{-pU_i}$ is similar to the free energy of the REM and in the limit of a long protein ($l_p \gg 1$) may be solved using ideas developed in [138]. This gives

$$\left( \log_2 \sum_{i=1}^{4^p} e^{-pU_i} \right) = 2l_p + \left( \log_2 \int_{-\infty}^{U_{\min}} e^{-\frac{U}{2\sigma_U^2}} dU \right) \quad (A6)$$

where $U_{\min}$ is the minimal observed energy which is well approximated by

$$\int_{-\infty}^{U_{\min}} e^{-\frac{U}{2\sigma_U^2}} dU = \frac{1}{4^p} \quad (A7)$$

or

$$U_{\min} = -\sigma_U \sqrt{2} \text{erfc}^{-1} \left( \frac{2}{4^p} \right) \simeq -2\sigma_U \sqrt{l_p \ln 2}. \quad (A8)$$

In the limit of a long protein one obtains

$$\left( \log_2 \sum_{i=1}^{4^p} e^{-pU_i} \right) = \begin{cases} 0 & \sigma_U \geq 2\sqrt{l_p \ln 2}, \\ 2l_p + \frac{\sigma^2_U}{2 \ln 2} & \sigma_U < 2\sqrt{l_p \ln 2}. \end{cases} \quad (A9)$$

Using equation (A4) the information content is finally given by

$$IC = \begin{cases} 2l_p & \sigma_U \geq 2\sqrt{l_p \ln 2}, \\ \frac{2l_p}{\ln 2} & \sigma_U < 2\sqrt{l_p \ln 2}. \end{cases} \quad (A10)$$

Schneider et al [135] suggested, and showed for a few transcription factors, that the information content is just sufficient for the target to be distinguished from the rest of the genome. Specifically, with $N$ potential binders, where $N$ is twice the genome length ($N \simeq 10^7 \text{ bp for E. coli}$), the amount of the information needed to distinguish a single binder is $\log_2 N$. Therefore, we would expect

$$IC \gtrsim \log_2 N \quad (A11)$$

If this is not fulfilled one expects a wrong sequence to be incorrectly recognized. Comparing equations (A10) and (A11) this translated to a condition on the length of the protein

$$l_p \gtrsim \frac{\log_2 N}{2} \quad (A12)$$

and

$$\sigma_U \gtrsim \sqrt{2 \ln N}. \quad (A13)$$

This is identical to results of the simple arguments presented in section 2. This condition is based on the assumption that the maximal information content per TF length is two bits. However, we find that the average information content per TF length is actually close to one bit. This fact increases the minimal protein’s length from 11 to 22 base pairs. A similar conclusion may be obtained from figure 8.

One can see that the condition for a broad line information content (A12) and (A13) is identical to the conditions for a marginally gapped target with a significant occupation probability (see section 2). In the limit of $N \rightarrow \infty$ both conditions represent the freezing point of the REM [138]. In figure 23 a histogram of the information content of several DNA-binding proteins is presented. Note that many proteins have significantly less information about the target than predicted by equation (A11). This conclusion is independent of any assumption on the binding energy distribution. A more detailed study which includes eukaryotic TFs and other databases was done in [69]. In this work authors found that for an eukaryotic TFs the problem of the not sufficient information content is much more severe.
A.2. The sequence score and the binding energy

We showed above that the information content of many proteins is not sufficient for efficient target location. In section 2 we present the same results, in particular, by comparing the binding energy of the target to the rest of the DNA. To this end we had to obtain the binding energy of a given sequence. In an information theory context this may be evaluated using a sequence score [139]. The score of a sequence $s = (s_1, s_2, \ldots, s_l)$ is defined as

$$\text{Sc} (s) = \sum_{i=1}^{l} \ln [4 \Pr (s_i, i)] = l_p \ln 4 + \ln \Pr (s, i),$$  \hspace{1cm} (A14)

where $\Pr (s_i, i)$ is defined in equation (2) and

$$\Pr (s, i) = \prod_{i=1}^{l_p} \Pr (s_i, i).$$  \hspace{1cm} (A15)

The probability that a sequence $s$ is bound is proportional to the Boltzmann factor $e^{-U(s)}$. Therefore, the binding energy, $U(s)$, is equal to the score of sequence $s$. Note that this convention may be calculated by recalling that we defined the binding energy so that its average is zero. Therefore,

$$U (s) = -\text{Sc} (s) + l_p \ln 4 + \frac{1}{4} \ln \left[ \prod_{i=1}^{l_p} \Pr (s_i, i) \right]$$

$$= -\sum_{i=1}^{l_p} \ln \left[ \prod_{i=1}^{l_p} \frac{\Pr (s_i, i)}{\Pr(s', i)} \right] = \sum_{i=1}^{l_p} u_i (s_i),$$  \hspace{1cm} (A16)

where

$$u_i (s_i) = -\ln \left[ \prod_{i=1}^{l_p} \frac{\Pr (s_i, i)}{\Pr(s', i)} \right]$$  \hspace{1cm} (A17)

is the contribution of site $i$ on the protein to the total binding energy of a sequence $s$. Equations (A16) and (A17) are used in section 2.2 to calculate (for each transcription factor) the binding energy of a given sequence.

Appendix B. A derivation of equations (63) and (64)

In this appendix we analyze the FPT properties of a simple random walk and derive equations (63) and (64). A similar calculation can be found in [120] and is presented here for completeness. The Laplace transformed FPT probability density of the discrete space and continuum time random walk,

$$\tilde{j} (x | x_0; s) = \int_0^\infty e^{-st} j (x | x_0; t) \, dt,$$  \hspace{1cm} (B1)

may be expressed in terms of the $z$-transformed FPT probability density of the discrete time random walk,

$$\tilde{j}_d (x | x_0; z) = \sum_{t=0}^{\infty} z^t \tilde{j}_d (x | x_0; t),$$  \hspace{1cm} (B2)

by [96]

$$\tilde{j} (x | x_0; s) = \tilde{j}_d \left( x | x_0; \frac{1}{1 + s/\lambda_0} \right).$$  \hspace{1cm} (B3)

In discrete space and discrete time the occupation probability at origin starting from the site $x > 0$ at $t = 0$ is given by

$$\tilde{P} (x | x_0; z) = \left( \frac{1 - \sqrt{1 - 2z}}{\sqrt{1 - z^2}} \right)^{|x_0|},$$  \hspace{1cm} (B4)

Using relation [140] that connects the FPT to the origin with the occupation probabilities

$$\tilde{j}_d (x | x_0; z) = \frac{\tilde{P} (x | x_0; z) - \delta_{x,x_0}}{\tilde{P} (x | x_0; z)},$$  \hspace{1cm} (B5)

the $z$-transformed probability density of the first return time to site $0$ of a discrete time random walk is given by

$$\tilde{j}_d (0 | 0; z) = 1 - \frac{1}{\tilde{P} (0 | 0; z)} = 1 - \sqrt{1 - z^2}.$$  \hspace{1cm} (B6)

With equation (B3) we obtain the Laplace transform of the first return time of a continuous time random walk (we denote this quantity in the text by $j (s)$):

$$\tilde{j} (x | x_0; s) = 1 - \sqrt{1 - \left( \frac{1}{1 + s/\lambda_0} \right)^2} \simeq 1 - \sqrt{1 - e^{-2s/\lambda_0}},$$  \hspace{1cm} (B7)

The $z$-transformed probability density of the first-passage time to site $0$ from site $x$ of a discrete time random walk is given by

$$\tilde{j}_d (0 | 0; z) = \frac{\tilde{P} (0 | x_0; z) - \delta_{x,x_0}}{\tilde{P} (0 | x_0; z)} = \left( \frac{1 - \sqrt{1 - z^2}}{z} \right)^{|x_0|}.$$  \hspace{1cm} (B8)

Its average over all possible starting sites in the large $N$ limit gives

$$\left( \tilde{j}_d (0 | 0; z) \right)_{x_0} = \frac{1}{N} \sum_{x_0=-N/2}^{N/2} \left( \frac{1 - \sqrt{1 - z^2}}{z} \right)^{|x_0|} \simeq \frac{2}{N} \frac{1}{1 - \frac{1}{1 + s/\lambda_0}}.$$  \hspace{1cm} (B9)

Using equation (B3) we obtain the Laplace transformed first-passage time to site $0$ averaged over the initial sites of a discrete time random walk (we denote this quantity in the text by $\tilde{j} (s)$):\[\tilde{j} (s) = \sum_{x_0} (j (s | x_0))_{x_0} \simeq \frac{1}{N} \frac{1}{1 + \sqrt{1 - (\frac{1}{1 + s/\lambda_0})^2}} \simeq \frac{1}{N} \frac{1 + e^{-s/\lambda_0}}{1 - e^{-s/\lambda_0}}.\]  \hspace{1cm} (B10)

Appendix C. Specificity

The conditions discussed in sections 4.2 and 4.2.3 imply that

$$P_d (n_p = 1) \ll e^{-\sigma_d^0},$$  \hspace{1cm} (C1)

and

$$P^T (n_p = 1) \gg P_d (n_p = 1)$$  \hspace{1cm} (C2)

have to hold. It is useful to separate the discussion into two cases.
(a) Small disorder $\sigma_U \ll \sqrt{2\ln N}$. In this case

$$P_T(n_p = 1) \simeq \frac{1}{1 + N e^{\sigma_U^2/2N_d}} \quad (C3)$$

so that conditions (C2) take the form

$$\sqrt{2\ln N} \gg \sigma_U \sqrt{1 + \frac{2\sqrt{2\ln N_d}}{\sigma_U}} \quad (C5)$$

$$U_T \ll -\sigma_U \sqrt{2\ln N_d} \quad (C6)$$

Condition (C5) can be satisfied for $2\sqrt{2\ln N_d} \ll \sigma_U \ll \sqrt{2\ln N}$ or for $\sigma_U \ll \min\left(\sqrt{\frac{2\ln N}{2\ln N_d}}, \sqrt{\frac{4\ln N}{4\ln N_d}}\right)$. The second condition (C6) is satisfied automatically in the case of a non-designed TF (so that $U_T = -\sigma_U \sqrt{2\ln N}$), in the case of designed TF with an additive binding energy (so that $U_T = l_p E_c = -\sigma_U \sqrt{N}$ where, as stated above, $l_p > \frac{1}{2} \ln N$) and, obviously, in the case of a gapped TF with cooperative binding energy (so that the value of $U_T$ is not bound). Thus, in the small disorder regime the speed-stability paradox can be resolved by increasing the copy number of a TF without destroying the specificity. For a small number of dangerous sites, $N_d \ll N^{1/4}$, one may do so with a relatively large disorder while in the opposite case the upper bound on the disorder strength decreases. The number of TF copies is given by

$$n_p \sim \max\left(e^{\sigma_U^2/2N_d}, N e^{\sigma_U^2/2U_T}\right). \quad (C7)$$

(b) Large disorder $\sigma_U \gg \sqrt{2\ln N}$. In this case

$$P_T(n_p = 1) \simeq 1 \quad (C8)$$

$$P_0(n_p = 1) \simeq \frac{e^{\sigma_U^2/2N_d}}{e^{\sigma_U^2/2 \ln N}} \quad (C9)$$

so that conditions (C2) take the form

$$e^{\sigma_U^2/2N_d} \ll e^{-\sigma_U^2} \quad (C10)$$

or

$$\sqrt{2\ln N} - \sqrt{2\ln N_d} \gg \sigma_U. \quad (C11)$$

This condition cannot be satisfied simultaneously with the regime assumption $\sqrt{2\ln N} \ll \sigma_U$. Thus, in the large disorder regime the speed-stability paradox cannot be resolved by increasing the copy number of a TF without destroying the specificity.

**Appendix D. Details of the simulations**

The model is simulated using a standard continuous time Gillespie algorithm [121]. The protein on a given site, $i$, in the $s$ mode can perform four possible moves: it can move in one of the possible directions along the DNA with probability $\lambda_i e^{\sigma_U^2/2N_d}$, go to the $r$ mode with probability $\sigma_U e^{\sigma_U^2/2N}$, or dissociate from the DNA and reassociate on a randomly chosen site with probability $\frac{1}{e^{\sigma_U^2/2N_d} + \lambda_i e^{\sigma_U^2/2N}}$. In the first two cases time is advanced by an amount drawn from a Poisson distribution with an average of $\frac{1}{e^{\sigma_U^2/2N_d} + \lambda_i e^{\sigma_U^2/2N}}$. When the protein dissociates, the time is advanced by first, drawing a time from a Poissonian distribution with an average time of $\frac{1}{\lambda_i e^{\sigma_U^2/2N}}$, and adding to it a time, drawn from a Poissonian distribution with an average $\frac{1}{\lambda_i e^{\sigma_U^2/2N}}$. This corresponds to the time needed for a relocation to a new site. The $r$ mode can transform only into the $s$ mode. The time for this step is drawn from a Poissonian distribution with an average time $\frac{1}{\sigma_U}$.  

**Appendix E. Pole structure analysis and derivation of equation (67)**

In this appendix we show that equation (67) holds when there is a sufficient time scale separation in the problem. Following standard practice we perform the inverse Laplace transform by studying the poles of $\tilde{R}(s)$. In our case $R(s)$ has no poles in the region $\Re(s) > 0$ of the complex plane so that

$$R(t) = \frac{1}{2\pi i} \int_{-\infty}^{\infty} e^{st} \tilde{R}(s) \, ds$$

$$= \sum_{s \in \text{poles}} e^{s t} \text{Res} \left[ \tilde{R}(s), s \right]. \quad (E1)$$

As we showed above (see equation (65))

$$\tilde{R}(s) = \frac{\tilde{j}_{p_1}(u(s))}{s} \left\{ 1 - \frac{\lambda_b \lambda_u}{s + \lambda_b} - \frac{\lambda_r \lambda_u}{s + \lambda_r} \right\}^{-1}, \quad (E2)$$

where

$$u(s) = \frac{s (s + \lambda_r + \lambda_s + \lambda_u) + \lambda_s \lambda_u}{s + \lambda_r} \quad (E3)$$

and

$$\tilde{j}_{p_1}(s) = \frac{p_1 \tilde{j}(s)}{1 - (1 - p_1) \tilde{j}_0(s)}, \quad (E4)$$

where

$$\tilde{j}(s) = \left\{ \tilde{j}(s|x) \right\}_x \simeq \frac{1}{N} \left[ 1 + e^{-s/\lambda_0} \right] \quad (E5)$$

and

$$\tilde{j}_0(s) = 1 - \sqrt{1 - e^{-2s/\lambda_0}} \quad (E6)$$

for large $N$ [120]. Finally, $p_1$ is probability of crossing the barrier at the target at each visit of its $s$ state,

$$p_1 = \frac{\lambda_r^2/\lambda_0}{1 + \lambda_s/\lambda_0 + \lambda_r^2/\lambda_0}. \quad (E7)$$

The trivial pole of $\tilde{R}(s)$ is easily found (from equation (E2)) to be

$$s_0 = 0 \quad (E8)$$

and its residue is

$$\text{Res}_{s_0} = 1. \quad (E9)$$

Note that a pole in $\tilde{j}_{p_1}(u(s))$ would not lead to a pole in $\tilde{R}(s)$ due to the occurrence of $\tilde{j}_{p_1}(u(s))$ in the numerator and denominator. Thus, the equation for other poles is given by

$$(s + \lambda_b) u(s) = \lambda_b \lambda_u \left[ 1 - \tilde{j}_{p_1}(u(s)) \right]. \quad (E10)$$
Next we assume that $\lambda_\tau \ll \lambda_s, \lambda_u, \lambda_b, \lambda_0$ with $\lambda_u, \lambda_b, \lambda_0$ of comparable order. The order of $\lambda_s$ will be discussed below in more details. We focus on the interesting regime $s/\lambda_0 \ll 1$ in which the target is not found immediately. The analysis is carried out by considering the pole equation in different regimes. Using

$$u(s) = \frac{1}{\lambda_0} \left[ s(s + \lambda_r + \lambda_s + \lambda_u) + \lambda_u \lambda_s \right] = \frac{s + \lambda_r + \lambda_u}{\lambda_0}$$

(E11)

and that $(-s) \ll \lambda_0$ and $\lambda_r \ll \lambda_0$ one may see that for

$$|\lambda_s - (-s)| \gg \frac{\lambda_s}{\lambda_0}$$

(E12)

This equation has one pole of the order of $\lambda_u, \lambda_b$, which corresponds to trajectories finding the target within the first sliding event. This pole can be discarded in the large $N$ limit since its residue scales as $1/N$. Its second pole reads, to leading order in $\lambda_r/\lambda_u$ and $1/N$:

$$\tilde{j}(u(s)) \approx \frac{p_1 \tilde{j}(u(s))}{1 - (1 - p_1) j_0(s)} \approx \frac{N}{1 + e^{-\lambda_u/\lambda_0}} \frac{1}{1 - e^{-\lambda_u/\lambda_0}} \approx \frac{N}{1 + \frac{1}{1 + p_1} \sqrt{1 - e^{-2\lambda_u/\lambda_0}}}$$

$$\lambda_u \equiv \frac{\sqrt{\coth \left( \frac{\lambda_u}{\lambda_0} \right)}}{N}$$

in this regime (this is verified self-consistently below). Equation (E10) then reduces to

$$(s + \lambda_r)(s + \lambda_r + \lambda_u) = \lambda_u \lambda_s \left( 1 - \frac{\kappa}{N} \right).$$

(E13)

This equation has one pole of the order of $\lambda_u, \lambda_b$, which corresponds to trajectories finding the target within the first sliding event. This pole can be discarded in the large $N$ limit since its residue scales as $1/N$. Its second pole reads, to leading order in $\lambda_r/\lambda_u$ and $1/N$:

$$t_{-1}^{-1} \equiv -s_1 \simeq \frac{\lambda_b \left( \lambda_r + \frac{\lambda_u}{N} \right)}{\lambda_b + \lambda_u}.$$  

(E14)

To ensure that $-s \gg \lambda_s$, as was assumed one should satisfy

$$\lambda_s \ll \frac{\lambda_b \left( \lambda_r + \frac{\lambda_u}{N} \right)}{\lambda_b + \lambda_u}.$$  

(E15)

The corresponding residue of $\tilde{R}(s)$ then reads

$$\text{Res}_1 \simeq -q = - \frac{1}{1 + \frac{\lambda_u}{\lambda_s}}.$$  

(E16)

This second pole corresponds to trajectories which find the target before crossing of the barrier. In the limit $\lambda_r \ll \lambda_u$, such events occur with a high probability $q \simeq 1$ and are characterized by a time scale $t_1 \simeq N \left( \frac{1}{\lambda_b} + \frac{1}{\lambda_u} \right)$. In the limit $\lambda_r \gg \frac{\lambda_u}{N}$ this second pole corresponds to processes which find the target before the typical time which characterizes a fall into the trap $\lambda_r^{-1}$ and without scanning the whole length. Such events are unlikely as shown by $q \ll 1$. To check the self-consistency we calculate condition (E12) for $s = s_1$.

$$\frac{\lambda_b \left( \lambda_r + \frac{\lambda_u}{N} \right)}{\lambda_s \left( \lambda_b + \lambda_u \right)} - 1 \gg \frac{\lambda_r}{\lambda_0}$$

(E17)

but

$$\frac{\lambda_b \left( \lambda_r + \frac{\lambda_u}{N} \right)}{\lambda_s \left( \lambda_b + \lambda_u \right)} - 1 \gg \frac{\lambda_r}{\lambda_s}$$

(E18)

so one may see that condition (E12) for the first pole holds easily.

**Regime II.** Here we consider $-s \ll \lambda_b$. To proceed we take

$$\frac{p_1 \tilde{j}(u(s))}{1 - (1 - p_1) j_0(s)} \simeq \frac{N}{1 + e^{-\lambda_u/\lambda_0}} \frac{1}{1 - e^{-\lambda_u/\lambda_0}} \frac{1}{1 + \frac{p_1}{1 - e^{-2\lambda_u/\lambda_0}}}$$

$$= \frac{1}{N} \frac{1 + \frac{1}{1 - e^{-2\lambda_u/\lambda_0}}}{1 + \frac{1}{1 - e^{-2\lambda_u/\lambda_0}}} = \frac{\kappa}{N}$$

(E19)

as before (this is verified self-consistently below). The equation becomes

$$s + \lambda_s + \lambda_r + \lambda_u = \lambda_u (s + \lambda_s) \left( 1 - \frac{\kappa}{N} \right).$$

(E20)

The interesting pole is given by

$$t_2^{-1} \equiv -s_2 \simeq \frac{\lambda_u}{\lambda_r + \frac{\lambda_u}{N}}.$$  

(E21)

and the corresponding residue of $\tilde{R}(s)$ reads

$$\text{Res}_2 \simeq - (1 - q) = - \left( 1 - \frac{1}{1 + \frac{\lambda_u}{\lambda_s/N}} \right).$$  

(E22)

Similar to the case discussed above when $\lambda_r \gg \lambda_u \kappa/N$ the search involves a high chance of many entrances and exists from the $s$ state, as shown by $1 - q \simeq 1$. In the opposite limit trajectories entering the $s$ state are very unlikely ($q \simeq 1$) and the search time is dominated by the trapping time $1/\lambda_s$. To check the self-consistency we calculate condition (E12) for $s = s_2$:

$$1 - \frac{\lambda_u}{\lambda_r + \frac{\lambda_u}{N}} = 1 - \frac{\lambda_u}{\lambda_u + 1} \gg \frac{\lambda_r}{\lambda_0}$$

(E23)

but

$$\frac{\kappa \lambda_u}{N} \ll \lambda_0.$$  

(E24)

This condition is also easily met. Applying the poles (E8), (E14) and (E21) with the corresponding residues (E9), (E16) and (E22) to equation (E1) one obtains equation (67):

$$\tilde{R}(t) \simeq 1 - q e^{-t/t_1} - (1 - q) e^{-t/t_2}.$$  

(E25)
where

\[ \kappa = \frac{\sqrt{\coth \left( \frac{\lambda_b}{2N} \right)}}{1 + \frac{1}{1 + \frac{\lambda_b}{\lambda_a} \sqrt{1 - e^{-2\sigma^2/\lambda_b^2}}} \cdot \frac{1}{q}}, \quad \text{(E26)} \]

\[ q = \frac{1}{1 + \frac{\lambda_b}{\lambda_a} \sqrt{1 - e^{-2\sigma^2/\lambda_b^2}}}. \quad \text{(E27)} \]

\[ \tau_1 = \frac{\lambda_b + \lambda_a}{\lambda_b \left( \frac{\lambda_r + \lambda_s}{N} \right)}, \quad \text{(E28)} \]

\[ \tau_2 = \frac{\lambda_r + \lambda_s}{\lambda_s \frac{\lambda_r + \lambda_s}{N}}. \quad \text{(E29)} \]

### Appendix F. Conditions for a perfect search

In this appendix we show that using the search and recognition strategy based on a disorder in a barrier height one may in principle achieve a ‘perfect’ search without any design of the target and using only one searcher by optimizing the values of \( E_0 \) and \( \sigma \). We define a ‘perfect’ search as one where one searcher goes to the \( r \) state of the target site with probability 1 \((q = 1)\) within a typical search time \( T^{\text{typ}} \). The last demand may be satisfied by ensuring that the barrier height on the target site, \( E_0^T \), is not positive with probability 1. In this case the number of scans of the DNA that the searcher performs before a transition to the \( r \) state is of order 1. Below we show that this perfectly fast search can be achieved with a perfect recognition of the target, \( q = 1 \). Namely we show that the occupation probability of the target in the infinite time limit may be arbitrarily close to 1. To end this calculation we can estimate below the occupation probability of each site on the DNA, \( P_i \), and in particular, the occupation probability of the target site, \( P_T \).

We take \( \lambda'_i = 0 \) \((E_r = \infty)\) to ensure the stability of the protein–DNA complex after the target is located. We assume now (and check this assumptions self consistently) that \( q = 1 \). This implies that the protein scans the whole DNA in the s state (and, therefore, equilibrates in the s states, but not in the \( r \) state, on all DNA sites) before it passes over the barrier on the target. Thus, the transition rate to the \( r \) state of site \( i \) is given by \( \frac{\lambda_b}{\lambda_b + \lambda_a} \lambda'_i \). The time evolution of the occupation probability of the \( r \) state at site \( i \) is then given by

\[ \frac{dP_i}{dt} = \frac{\lambda_b}{\lambda_b + \lambda_a} \lambda'_i \left( 1 - \sum_{j=1}^{N} P_j \right). \quad \text{(F1)} \]

At steady state the occupation probability of the \( r \) state at site \( i \) is

\[ P_i = \frac{\lambda'_i/\lambda_0}{\sum_{j=1}^{N} \lambda'_j/\lambda_0}. \quad \text{(F2)} \]

The transition rate \( \lambda'_i \) is given by equation (83). As stated above, the barrier height, \( E_b^T \), is drawn from a Gaussian distribution:

\[ \Pr(E_b^T) = \frac{e^{-\frac{(x_b^T - \mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}. \quad \text{(F3)} \]

13 This is the time to reach a designed target on a flat energy landscape.
\[
\phi = \left[ 1 + N \frac{e^{-E_0 \sigma^2/2} \text{erfc} \left( \frac{-E_0 \sigma^2/2}{\sqrt{2} \sigma} \right)}{1 + \text{erf} \left( \frac{E_0 \sigma^2/2}{\sqrt{2} \sigma} \right)} \right]^{-1}. 
\]

Using equations (F5), (F8) and taking the leading order in \( \frac{1}{N} \) we obtain
\[
q \gtrsim \left[ 1 + \frac{2\pi}{N} \left( \frac{N^2}{2\pi} \right) \exp \left( J - 1 \right) \text{erf}^{-1} \left( \frac{2}{N} \right) \right] \left[ \frac{J - 1}{2} \text{erf} \left( J - 1 \right) \text{erf}^{-1} \left( \frac{2}{N} \right) \right]^{-1} \approx \left\{ \begin{array}{ll}
1 + \frac{2}{J} & J > 1, \\
1 + \frac{2}{J} & J \leq 1.
\end{array} \right.
\]

In the limit \( N \to \infty \) one obtains a behavior similar to the usual second order phase transition of the REM:
\[
q = \left\{ \begin{array}{ll}
\frac{J - 1}{J + 1} & J \geq 1, \\
0 & J < 1.
\end{array} \right.
\]

Therefore, for large enough \( J \) the searcher finds its target with a probability close to 1 so that all our assumptions in this section based on \( q = 1 \) are self consistent. Also, since the typical barrier between the \( s \) and \( r \) states on the target is zero, the searcher finds its target within the facilitated diffusion limit. Note that, although the search and the recognition are perfect, the average time is infinite since there is a finite, \( 1 - q \), probability to be trapped on a non-target site.

Summarizing, to ensure a perfect search one should set
\[
\sigma = J \sqrt{2} \text{erf}^{-1} \left( \frac{2}{N} \right) 
\]
with some constant \( J \gg 1 \) and
\[
E_0 = \sigma \sqrt{2} \text{erf}^{-1} \left( \frac{2}{N} \right). 
\]

In this case the probability to find the target is
\[
q = \frac{J - 1}{J + 1} \approx 1 - \frac{2}{J} 
\]
and the typical search time is comparable to the facilitated diffusion limit. Therefore, \( J \) should be as large as possible. For the case of \( n_p \) searchers (assuming that they are independent) the condition on \( J \) becomes
\[
p_{\text{cat}} = (1 - q)^{n_p} = \left( \frac{2}{J + 1} \right)^{n_p} \ll 1. 
\]

In fact, a perfect search may be impossible due to practical reasons. For example, for \( N = 10^6 \) our results suggest that to ensure \( q = 0.5 \) one should take \( E_0 \approx 67.7 \) and \( \sigma \approx 14.3 \). Such large energies may be difficult to achieve. Nevertheless, as we showed in section 5.4 the proposed mechanism of barrier discrimination is very efficient even when parameters are far from perfect search conditions and the mean and the variance of the barrier energy are comparable to the mean and the variance of the experimentally found binding energy distribution.

**References**

[1] Hopfield J J 1974 Kinetic proofreading: a new mechanism for reducing errors in biosynthetic processes requiring high specificity Proc. Natl Acad. Sci. USA 71 4135

[2] Hill A V 1913 The combinations of haemoglobin with oxygen and with carbon monoxide: J Biochem. J. 7 471

[3] Fersht A 1985 Enzyme Structure and Mechanism (New York: Freeman)

[4] Schimmel P and Alexander R 1998 All you need is RNA Science 281 658

[5] Cech T R 2000 The ribosome is a ribozyme Science 289 878

[6] Barzel A and Kupiec M 2008 Finding a match: how do homologous sequences get together for recombination? Nature Rev. Genetics 9 27

[7] Loverdo C, Benichou O, Moreau M and Voituriez R 2007 Enhanced reaction kinetics in biological cells Nature Phys. 4 134

[8] Guttasarma P 1995 Does replication-induced transcription regulate synthesis of the myriad low copy number proteins of Escherichia coli? BioEssays 17 987

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

[10] Ishihama Y, Schmidt T, Rappisler I, Mann M, Harli F U, Kerner M and Frishman D 2008 Protein abundance profiling of the Escherichia coli cytosol BMC Genomics 9 102

[11] Taniguchi Y, Choi P J, Li G-W, Chen H, Babu M, J Hearn, Emilii A and Xie X S 2010 Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells Science 329 533

[12] Yamanaka K and Inouye M 2001 Induction of CspA, an E. coli major cold-shock protein, upon nutritional upshift at 37 °C Genes Cells 6 279

[13] Berg O G, Winter R B and von Hippel P H 1981 Diffusion-driven mechanisms of protein translocation on nucleic acids: I. Models and theory Biochemistry 20 6929

[14] von Hippel P H and Berg O G 1989 Facilitated target location in biological systems J. Biol. Chem. 264 675

[15] Halford S E and Marko J F 2004 How do site-specific DNA-binding proteins find their targets? Nucl. Acids Res. 32 3040

[16] Winter R B and Von Hippel P H 1981 Diffusion-driven mechanisms of protein translocation on nucleic acids: II. The Escherichia coli repressor–operator interaction: equilibrium measurements Biochemistry 20 6948

[17] Winter R B, Berg O G and von Hippel P H 1981 Diffusion-driven mechanisms of protein translocation on nucleic acids: III. The Escherichia coli Lac repressor–operator interaction: kinetic measurements and conclusions Biochemistry 20 6961

[18] Lesser D R, Kurpiewski M R, Waters T, Connolly B and Jen-Jacobson L 1993 Facilitated distortion of the DNA site energetics by sequence variation at 37 °C Genes Cells 6 279

[19] Jeltsch A, Wenz C, Stahl F and Pingoud A 1996 Linear diffusion of the restriction endonuclease EcoRV on DNA
is essential for the in vivo function of the enzyme EMBO 15 1014

[20] Jeltsch A and Pingoud A 1998 Kinetic characterization of linear diffusion of the restriction endonuclease EcoRV on DNA Biochemistry 37 2160

[21] 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 ASMBB 274 16665

[22] Shimamoto N 1999 One-dimensional diffusion of proteins along DNA J. Biol. Chem. 274 15293

[23] Stanford N L P, Szczelkun M D, Marko J F andHalford S E 2000 One- and three-dimensional pathways for proteins to reach specific DNA sites EMBO 19 6546

[24] 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. USA 102 15883

[25] Kampmann M 2005 Facilitated diffusion in chromatin lattices: mechanistic diversity and regulatory potential Mol. Microbiol. 57 889

[26] Blainey P C, Biophys. Chem. Oijen, 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. USA 103 5752

[27] Granelli A, Yekhal C C, Robertson R B and Greene E C 2006 Long-distance lateral diffusion of human Rad51 on double-stranded DNA Proc. Natl Acad. Sci. USA 103 1221

[28] Wang Y M, Austin R H and Cox E C 2006 Single molecule measurement of repressor protein 1D diffusion on DNA Phys. Rev. Lett. 97 048302

[29] Elf J, Li G-W and Xie X S 2007 Probing transcription factor dynamics at the simple single-molecule level in a living cell Science 316 1191

[30] Bonnet I, Biebricher A, Porte P-L, Lovoerdo C, Benichou O, Voituriez R, Escudé C, Wende W, Pingoud A and Desbiolles P 2008 Sliding and jumping of single EcoRV restriction enzymes on non-cognate DNA Nuclei. Acids Res. 36 4118

[31] Fok P-W, Guo C-L and Chou T 2008 Charge-transport-mediated recruitment of DNA repair enzymes J. Chem. Phys. 129 255101

[32] Fok P-W and Chou T 2009 Accelerated search kinetics mediated by redox reactions of DNA repair enzymes Biophys. J. 96 3949

[33] Komazin-Meredith G, Mirchev R, Golan D E, van Oijen A M and Coen D M 2008 Hopping of a processivity factor on DNA revealed by single-molecule assays of diffusion Proc. Natl Acad. Sci. USA 105 10721

[34] Adam G and Delbruck M 1968 Reduction of dimensionality in biological diffusion processes Structural Chemistry and Molecular Biology ed A Rich and N Davidson (San Francisco, CA: Freeman) pp 198–215

[35] Richter P H and Eigen M 1974 Diffusion controlled reaction rates in spheroidal geometry: application to repressor–operator association and membrane bound enzymes Biophys. Chem. 2 255

[36] Berg O G and Blomberg C 1976 Association kinetics with coupled diffusional flows: special application to the lac repressor-operator system Biophys. Chem. 4 367

[37] Gerland U, Moroz J D and Hwa T 2002 Physical constraints and functional characteristics of transcription factor–DNA interaction Proc. Natl Acad. Sci. USA 99 12015

[38] Bruinsma R F 2002 Physics of protein–DNA interaction Physica 313 211

[39] Slatsky M and Mirmir A 2004 Kinetics of protein–DNA interaction: facilitated target location in sequence-dependent potential Biophys. J. 87 4021

[40] Coppey M, Benichou O, Voituriez R and Moreau M 2004 Kinetics of target site localization of a protein on DNA: a stochastic approach Biophys. J. 87 1640

[41] Belotserkovskii B P and Zarling D A 2004 Analysis of a one-dimensional random walk with irreversible losses at each step: applications for protein movement on DNA J. Theor. Biol. 226 195

[42] Widom J 2005 Target site localization by site-specific, DNA-binding proteins Proc. Natl Acad. Sci. USA 102 16909

[43] Zhou H X 2005 A model for the mediation of processivity of DNA-targeting proteins by nonspecific binding: dependence on DNA length and presence of obstacles Biophys. J. 88 1608

[44] Lomholt M A, Ambjørnsson T and Metzler R 2005 Optimal target search on a fast-folding polymer chain with volume exchange Phys. Rev. Lett. 95 260603

[45] Hu T, Grosberg A Y, and Shklovskii B I 2006 How proteins search for their specific sites on DNA: the role of DNA conformation Biophys. J. 90 2731

[46] Hu T and Shklovskii B I 2007 How proteins search for their specific sites on DNA: the role of intersegment transfer Phys. Rev. E 76 051909

[47] Oshanin G, Wio H S, Lindenberg K and Burlatsk S F 2007 Intermittent random walks for an optimal search strategy: one-dimensional case J. Phys.: Condens. Matter 19 065142

[48] Eliazar I, Koren T and Klafter J 2007 Searching circular DNA strands J. Phys.: Condens. Matter 19 065140

[49] Cherstvy A G, Kolomeisky A B and Kornyshev A A 2008 Protein–DNA interactions: reaching and recognizing the targets J. Phys. Chem. B 112 4741

[50] Loverdo C, Benichou O, Voituriez R, Biebricher A, Bonnet I and Desbiolles P 2009 Quantifying hopping and jumping in facilitated diffusion of DNA-binding proteins Phys. Rev. Lett. 102 188101

[51] Lomholt M A, van den Broek B, Kalisch S M J, Wuiteand G L and Metzler R 2009 Facilitated diffusion with DNA coiling Proc. Natl Acad. Sci. USA 106 8204

[52] Florescu A-M and Joyceus M 2009 Description of nonspecific DNA–protein interaction and facilitated diffusion with a dynamical model J. Chem. Phys. 130 015103

[53] Florescu A-M and Joyceus M 2009 Dynamical model of DNA–protein interaction: effect of protein charge distribution and mechanical properties J. Chem. Phys. 131 105102

[54] Givaty O and Levy Y 2009 Protein sliding along DNA: dynamics and structural characterization J. Mol. Biol. 385 1087

[55] Vuzman D and Levy Y 2010 DNA search efficiency is modulated by charge composition and distribution in the intrinsically disordered tail Proc. Natl Acad. Sci. USA 107 21004

[56] Vuzman D, Polonsky M and Levy Y 2010 Facilitated DNA search by multi-domain transcription factors: a cross-talk via a flexible linker Biophys. J. 99 1202

[57] Vuzman D, Azia A and Levy Y 2010 Searching DNA via a ‘monkey bar’ mechanism: the significance of disordered tails J. Mol. Biol. 396 674

[58] Díaz de la Rosa M A, Koslover E F, Mulligan P J and Spakowitz A J 2010 Dynamic strategies for target-site search by DNA-binding proteins Biophys. J. 98 2943

[59] Benichou O, Chevalier C, Meyer B and Voituriez R 2011 Facilitated diffusion of proteins on chromatin Phys. Rev. Lett. 106 038102

[60] Rudniansky A B 2011 Physics of protein–DNA interactions: mechanisms of facilitated target search Phys. Chem. Chem. Phys. 13 2088
[61] Tkáčik G and Bialek W 2009 Diffusion, dimensionality, and noise in transcriptional regulation Phys. Rev. E 79 051901
[62] Tamiya Z, Barkai N and Fouxon I 2011 Physical aspects of precision in genetic regulation J. Biol. Phys. 37 213
[63] von Hippel P H and Berg O G 1986 On the specificity of DNA–protein interactions Proc. Natl. Acad. Sci. USA 83 1608
[64] Jeltsch A, Alves J, Wolfs H, Maass G and Pingoud A 1994 Pausing of the restriction endonuclease EcoRI during linear diffusion on DNA Biochemistry 33 10215
[65] Benichou O, Loverdo C, Moreau M and Voituriez R 2008 Optimizing intermittent reaction paths Phys. Chem. Chem. Phys. 10 6948
[66] Mirny L, Slutsky M, Wunderlich Z, Tafvizi A, Leith J and Kosmrlj A 2009 How a protein searches for its site on DNA: the mechanism of facilitated diffusion J. Phys. A: Math. Gen. 42 434013
[67] Das R K and Kolomeisky A B 2010 Facilitated search of proteins on DNA: correlations are important Phys. Chem. Chem. Phys. 12 2999
[68] Benichou O, Loverdo C, Moreau M and Voituriez R 2007 Intermittent search strategies Rev. Mod. Phys. 83 81
[69] Wunderlich Z and Mirny L A 2009 Different gene regulation strategies revealed by analysis of binding motifs Trends Genetics 25 434
[70] Benichou O, Kafri Y, Sheinman M and Voituriez R 2009 Searching fast for a target on DNA without falling to traps Phys. Rev. Lett. 103 138102
[71] Stormo G D and Field D S 1998 Specificity, free energy and information content in protein–DNA interactions Trends Biochem. Sci. 23 109
[72] Berg O G and von Hippel P H 1987 Selection of DNA binding sites by regulatory proteins statistical-mechanical theory and application to operators and promoters J. Mol. Biol. 193 723
[73] Maerkl S J and Quake S R 2007 A systems approach to measuring the binding energy landscapes of transcription factors Science 315 233
[74] Herzel H, Trifonov E N, Weiss O and Grosse I 1998 Interpreting correlations in biosequences Physica A 249 449
[75] Peng C-K, Buldrev S V, Goldberger A L, Havlin S, Sciortino F, Simons M and Stanley H E 1992 Long-range correlations in DNA sequences Nature 356 168
[76] Stormo G D, Schneider T D and Gold L 1986 Quantitative analysis of the relationship between nucleotide sequence and functional activity Nucl. Acids Res. 14 6661
[77] Zhang M O and Marr T G 1993 A weight array method for splicing signal analysis Comput. Appl. Biosci. 9 499
[78] Ponomarenko M P, Ponomarenko J V, Frolov A S, Podkolodnaya O A, Vorobyev D G, Kolchanov N A and Overton G C 1999 Oligonucleotide frequency matrices addressed to recognizing functional DNA sites Bioinformatics 15 631
[79] Bulyk M L, Johnson P L, F and Church G M 2002 Nucleotide of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors Nucl. Acids Res. 30 1255
[80] de Haan L and Ferreira A 2006 Extreme Value Theory: An Introduction (Berlin: Springer)
[81] Sheinman M and Kafri Y 2011 arXiv:1110.5997v4
[82] Derrida B 1980 Random-energy model: limit of a family of disordered models Phys. Rev. Lett. 45 79
[83] Gama-Castro S et al 2008 RegulonDB (version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and textpresso navigation Nucl. Acid. Res. 36 D120
[84] Berg O G 1978 On diffusion-controlled dissociation Chem. Phys. 31 47
[85] Redner S 2001 A Guide to First-Passage Process (Cambridge: Cambridge University Press)
[86] Condamin S, Benichou O, Tejedor V, Voituriez R and Klafter J 2007 First-passage times in complex scale-invariant media Nature 450 77
[87] Benichou O, Chevalier C, Klafter J, Meyer B and Voituriez R 2010 Geometry-controlled kinetics Nature Chem. 2 472
[88] Riggs A D, Suzuki H and Bourgeois S 1970 Lac repressor–operator interaction: I. Equilibrium studies J. Mol. Biol. 48 67
[89] Riggs A D, Bourgeois S and Cohn M 1970 The Lac repressor–operator interaction: III. Kinetic studies J. Mol. Biol. 53 401
[90] von Smoluchowski M 1917 Mathematical theory of the kinetics of the coagulation of colloidal solutions Z. Phys. Chem. 92 129
[91] Elowitz M B, Surette M G, Wolf P E, Stock J B and Leibler S 1999 Protein mobility in the cytoplasm of Escherichia coli J. Bacteriol. 181 197
[92] Phillips R, Kondaj J and Theriot J 2009 Physical Biology of the Cell (New York: Taylor and Francis)
[93] Lin S Y and Riggs A D 1972 Lac repressor binding to non-operator DNA: detailed studies and a comparison of equilibrium and rate competition methods J. Mol. Biol. 72 671
[94] Terry B J, Jack W E and Modrich P 1985 Facilitated diffusion during catalysis by EcoRI endonuclease nonspecific interactions in EcoRI catalysis J. Biol. Chem. 260 13130
[95] Blaine Y P, van Ojken A M, Banerjee A, Verdine G L and Xie X S 2006 A base-excision DNA-repair protein finds intrahelial lesion bases by fast sliding in contact with DNA Proc. Natl Acad. Sci. USA 103 5752
[96] Hughes B D 1995 Random Walks and Random Environments (Random Walks vol 1) (Oxford: Clarendon Press)
[97] Sheinman M and Kafri Y 2009 The effects of intersegmental transfers on target location by proteins Phys. Biol. 6 016003
[98] Grosberg A, Rabin Y, Havlin S and Neer A 1993 Crumpled globule model of the three-dimensional structure of DNA Europhys. Lett. 23 373
[99] Bancaud A, Huet S, Daigle N, Mozziconacci J, Beaudouin J and Ellenberg J 2009 Molecular crowding affects diffusion and binding of nuclear proteins in heterochromatin and reveals the fractal organization of chromatin EMBO 28 3785
[100] Lieberman-Aiden E et al 2009 Comprehensive mapping of long-range interactions reveals folding principles of the human genome Science 326 289
[101] Lin S Y and Riggs A D 1975 The general affinity of Lac repressor for human genome long-range interactions reveals folding principles of the genome Nucl. Acid. Res. 10 7157
[102] Richey B, Cayley D S, Mossing M C, Kolka C, Anderson C F, Farrar T C and Record M T 1987 Variability of the intracellular ionic environment of Escherichia coli and gene expression J. Mol. Biol. 197 397
[103] Anderson C F, Farrar T C and Record M T 1987 Variability of the intracellular ionic environment of Escherichia coli and gene expression J. Mol. Biol. 197 397
[104] Riggs A D, Bourgeois S and Cohn M 1970 The Lac repressor–operator interaction: I. Equilibrium studies J. Mol. Biol. 48 67
[105] Bresloff J L and Crothers D M 1975 DNA-ethidium reaction kinetics: demonstration of direct ligand transfer between DNA binding sites J. Mol. Biol. 172 263
[106] van den Broek B, Lomholt M A, Kalisch S-M J, Metzler R and Wuite G J J 2008 How DNA coiling enhances target localization by proteins Proc. Natl Acad. Sci. USA 105 15738
[107] Halford S E 2009 An end to 40 years of mistakes in DNA–protein association kinetics Biochem. Soc. Trans. 37 343–8
[108] Li G-W, Berg O G and Elf J 2009 Effects of macromolecular crowding and DNA looping on gene regulation kinetics Nature Phys. 5 294
[109] Hu L, Grosberg A Y and Bruinsma R 2008 Are DNA transcription factor proteins Maxwellian demons? Biophys. J. 95 1151
[110] Ferreiro D U and de Prat-Gay G 2003 A protein–DNA binding mechanism proceeds through multi-state or two-state parallel pathways J. Mol. Biol. 331 89
[111] Kalodimos C G, Biris N, Bonvin A M J, Levandoski M M, Guennuegues M, Boelens R and Kaptein R 2004 Structure and flexibility adaptation in nonspecific and specific protein–DNA complexes Science 305 386
[112] Pingoud A and Wende W 2007 A sliding restriction enzyme pauses Structure 15 391
[113] Townsend S A, Samuelson J C, Bao Y, Xu S-Y and Aggarwal A K 2007 BstYI bound to noncognate DNA reveals a ’hemispherical’ complex: implications for DNA scanning Structure 15 449
[114] Pingoud A and Jeltsch A T 2001 Structure and function of type II restriction endonucleases Nucl. Acids Res. 29 3705
[115] Dahirel V, Paillusson F, Jardat M, Barbi M and Victor J-M 2009 Non specific–DNA interaction: why proteins can diffuse along DNA Phys. Rev. Lett. 102 228101
[116] Xu H, Moraitis M, Reedstrom R J and Matthews K S 1998 Kinetic and thermodynamic studies of purine repressor binding to corepressor and operator DNA J. Biol. Chem. 273 8958
[117] Shultzaberger R K, Roberts L R, Lyakhov I G, Sidorov I A, Stephen A G, Fisher R J and Schneider T D 2007 Correlation between binding rate constants and individual information of E. coli Fis binding sites Nucl. Acids Res. 35 5275
[118] Hanggi P, Talkner P and Borkovec M 1990 Reaction-rate theory: fifty years after Kramers Rev. Mod. Phys. 62 251
[119] Koshland D E 1958 Application of a theory of enzyme specificity to protein synthesis Proc. Natl Acad. Sci. USA 44 98
[120] Montroll E W 1969 Random walks on lattices containing traps J. Math. Phys. 26 6
[121] Gillespie D T 1976 A general method for numerically simulating the stochastic time evolution of coupled chemical reactions J. Comput. Phys. 22 403
[122] Gasch A P, Spellman P T, Kao C M, Carmel-Harel O, Eisen M B, Storz G, Botstein D and Brown P O 2000 Genomic expression programs in the response of yeast cells to environmental changes Mol. Biol. Cell 11 4241
[123] Braun E and Brenner N 2004 Transient responses and adaptation to steady state in a eukaryotic gene regulation system Phys. Biol. 1 67
[124] Ramoni M F, Sebastiani P and Kohane I S 2002 Cluster analysis of gene expression dynamics Proc. Natl Acad. Sci. USA 99 9121
[125] Alon U 2007 An Introduction to Systems Biology: Design Principles of Biological Circuits (Mathematical and Computational Biology Series vol 10) (Boca Raton, FL: Chapman and Hall)
[126] Alon U 2007 Network motifs: theory and experimental approaches Nature Rev. Genetics 8 450
[127] Shen-Orr S S, Milo R, Mangan S and Alon U 2002 Network motifs in the transcriptional regulation network of Escherichia coli Nature Genetics 31 64
[128] Kalir S, McClure J, Pabbaraju K, Southward C, Ronen M, Leibler S, Surette M G and Alon U 2001 Ordering genes in a flagella pathway by analysis of expression kinetics from living bacteria Science 292 2080
[129] Ronen M, Rosenberg K, Shraiman B I and Alon U 2002 Assigning numbers to the arrows: parameterizing a gene regulation network by using accurate expression kinetics Proc. Natl Acad. Sci. USA 99 10555
[130] Zaslaver A, Mayo A E, Rosenberg R, Bashkin P, Sberro H, Tsaluyk S, Surette M G and Alon U 2004 Just-in-time transcription program in metabolic pathways Nature Genetics 36 486
[131] Kolesov G, Wunderlich Z, Laikova O N, Gelfand M S and Mirny L A 2007 How gene order is influenced by the biophysics of transcription regulation Proc. Natl Acad. Sci. USA 104 13948
[132] Wunderlich Z and Mirny L A 2008 Spatial effects on the speed and reliability of protein DNA search Nucl. Acids Res. 36 3570
[133] Benichou O, Loverdo C and Voituriez R 2008 How gene colocalization can be optimized by tuning the diffusion constant of transcription factors Europhys. Lett. 84 38003
[134] Luscombe N M, Austin S E, Berman H M and Thornton J M 2000 An overview of the structures of protein–DNA complexes Gen. Biol. 1 11
[135] Schneider T D, Stormo G D, Gold L and Ehrenfeucht A 1986 Information content of binding sites on nucleotide sequences J. Mol. Biol. 188 415
[136] Hertz G Z and Stormo G D 1999 Identifying DNA and protein patterns with statistically significant alignments of multiple sequences Bioinformatics 15 563
[137] Blattner F R et al 1997 The complete genome sequence of Escherichia coli K-12 Science 277 1453
[138] Derrida B 1981 Random-energy model: An exactly solvable model of disordered systems Phys. Rev. B 24 2613
[139] Staden R 1984 Computer methods to locate signals in nucleic acid sequences Nucl. Acids Res. 12 505
[140] Montroll E W and Weiss G H 1965 Random walks on lattices: II J. Math. Phys. 6 167
[141] Abramowitz M and Stegun I A 1970 Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables (New York: Dover)