Searching fast for a target on a DNA without falling to traps

O. Bénichou, Y. Kafri, M. Sheinman and R. Voituriez
1 UMR 7600, Université Pierre et Marie Curie/CNRS, 4 Place Jussieu, 75255 Paris Cedex 05 France. and
2 Department of Physics, Technion, Haifa 32000, Israel.
(Dated: January 28, 2009)

Genomic expression depends critically both on the ability of regulatory proteins to locate specific target sites on a DNA within seconds and on the formation of long lived (many minutes) complexes between these proteins and the DNA. Equilibrium experiments show that indeed regulatory proteins bind tightly to their target site. However, they also find strong binding to other non-specific sites which act as traps that can dramatically increase the time needed to locate the target. This gives rise to a conflict between the speed and stability requirements. Here we suggest a simple mechanism which can resolve this long-standing paradox by allowing the target sites to be located by proteins within short time scales even in the presence of traps. Our theoretical analysis shows that the mechanism is robust in the presence of generic disorder in the DNA sequence and does not require a specially designed target site.

It is commonly believed that three-dimensional diffusion is too slow for proteins to locate their specific target on a DNA molecule for cells to function properly. To resolve this issue Berg and von Hippel suggested, in series of seminal papers [1, 2], that combining periods of one-dimensional diffusion along the DNA (sliding) with periods of three-dimensional diffusion off the DNA (jumping) can speed up the search time by several orders of magnitude. Since then, sliding (or equivalently binding of proteins to non-specific DNA sequences) has been observed in many experiments [3, 4, 5] and is now believed to be a common mechanism [6, 7, 8, 9, 10, 11]. On the other hand, as pointed out already in [12], experimental and theoretical works have shown that the binding energies of a protein to different DNA sequences are very large - a direct consequence of the required stability of the protein with its target site. The binding energies can be well fitted by a Gaussian with the strongest binding energies of the order of $\sim 30 k_B T$ and a standard deviation of the order of $5 k_B T$ [13]. This casts a cloud on the simple facilitated diffusion picture of Berg and von Hippel - the binding energy distribution suggests an unacceptably slow search with very slow sliding and deep traps [14]. This unresolved conflict is called the speed-stability paradox [2].

Here, motivated by direct experimental observations [14, 15, 16] and the theoretical work by Slutsky and Mirny [10], we consider a model in which the protein, when bound to the DNA, can switch between two conformations separated by a free energy barrier. In one, termed the search state the protein is loosely bound to the DNA and can slide along it. In the second, recognition mode, it is trapped in a deep energy well. Note that equilibrium measurements of binding energies to the DNA are controlled by the recognition state.

In this paper, 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 $\tau_{ave}$ [17], is generally irrelev-

![FIG. 1: An illustration of the model.](image)

vant 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 $R_{np}(t)$ for a reaction to occur before time $t$. We show below that $R_{np}(t)$ can reach values close to one in a time scale $t_{np}(t)$ which can be orders of magnitude smaller than the value $t_{ave}/n_p$ expected from the usual approach.

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. We suggest that 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, it shows that in the realistic case of generic disorder in the barrier height the search can be very effective even if the target site is not designed. If experimentally verified the proposed mechanism will resolve the speed-stability paradox.

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 [10]. The transition rates, \( \lambda_i^s \) and \( \lambda_i^r \), \( \lambda_i^u \) and \( \lambda_i^u \), between the different states are defined in Fig. 1. To model sliding, in the \( s \)-state the protein can move with rate \( \lambda_u / 2 \) to neighboring sites on the DNA. Note that the rates \( \lambda_i^s \) and \( \lambda_i^r \) may depend on the location \( i = 1 \ldots N \) along the DNA. In principle \( \lambda_u \) and \( \lambda_u \) also have a dependence on \( i \). As justified later this will have a weaker effect on our results and we omit it for clarity. Finally, after a jump we assume the protein relocates to a random position on the DNA due to its packed conformation [18].

To gain an understanding of the difference between the two time scales \( t_{\text{typ}}^{\text{bp}} \) and \( t_{\text{avc}}^{\text{bp}} / n_p \) we first consider \( n_p = 1 \) in a simplified model where \( \lambda_i^s \) and \( \lambda_i^r \) are independent of \( i \) except at the target site \( T \) where \( \lambda_f^s = \infty \) and \( \lambda_f^r = 0 \). The disorder of the DNA sequence is neglected and the target is designed such that a reaction takes place at the first visit of the target site. As stated above, we are interested in the probability \( R(t) = \int_0^t P(t') dt' \) that a reaction occurs before time \( t \), where \( P(t) \) is the distribution of the first-passage time (FPT) [19, 20, 21] to the target (we drop the subscript when \( n_p = 1 \)).

The Laplace transform, \( \tilde{P}(s) = \int_0^\infty e^{-st} P(t) dt \), of \( P(t) \) can be obtained exactly. To do this we consider a DNA molecule of \( N \) sites. For simplicity we take a centered target site (labeled 0). Consider, first, the joint probability density for a protein to find the target at time \( t = t_s + t_r \) starting from a location \( x_0 \) at \( t = 0 \) before unbinding from DNA. Here \( t_s \) is the total time spent in the \( s \) state and \( t_r \) is the total time spent in the \( r \) state. If exactly \( n \) transitions occurred from the \( s \)-state to the \( r \)-state this is given by

\[
P_n(t_s, t_r | x_0) = \lambda_s P(n-1, \lambda_s, t_r) \mathcal{P}(n, \lambda_r, t_s) j(t_s | x_0) e^{-\lambda_u t_r},
\]

where \( \mathcal{P}(n, \mu, t) = (\mu t)^n e^{-\mu t} / n! \) is the Poisson distribution and we use the convention \( \mathcal{P}(0, \mu, t) = \delta(t) / \mu \).

\[
j(t_s | x_0) \quad \text{is the FPT density at the target } x = 0 \text{ for a usual random walk starting from } x_0 \text{ whose functional form was derived in [22].}
\]

The FPT density before unbinding starting from \( x_0 \) then reads:

\[
j(t | x_0) = \sum_{n=0}^\infty \int_0^\infty \int_0^\infty ds dt r \delta(t_s + t_r - t) P_n(t_s, t_r | x_0).
\]

After Laplace transform and using \( \tilde{P}(n, \mu, s) = \mu^n / (s + \mu)^{n+1} \), we find \( \tilde{j}(s | x_0) = \tilde{j}(u(s) | x_0) \) with \( u(s) = e^{s(\lambda_s + \lambda_u + \lambda_r) + \lambda_s \lambda_u} \) \( \lambda_s \lambda_u \lambda_r \). Averaging over \( x_0 \) and following [6, 22] we finally obtain

\[
\tilde{P}(s) = \tilde{j}(u(s)) \left\{ 1 - \frac{\lambda_u \lambda_u 1 - \tilde{j}(u(s))}{s + \lambda_u u(s)} \right\}^{-1},
\]

where \( \tilde{j}(s) \overset{\approx}{=} \langle \tilde{j}(s | x) \rangle_x \sim 1 / \sqrt{1 + s^2 / \lambda_u} \) for large \( N \) [22].

The results along with numerics, performed using a standard continuous time Gillespie algorithm, are shown in Fig. 2. As is clearly evident, for a realistic range of parameters (we take barrier heights to be of the same order of magnitude as the experimentally measured binding energies) \( R(t) \) reaches a plateau close to one on a typical time scale \( t_{\text{avc}}^{\text{bp}} \), which, for \( N = 10^6 \), is much shorter than the average search time \( t_{\text{avc}}^{\text{avc}} = -d \tilde{P} / ds (s = 0) \). Quantitatively the typical search time \( t_{\text{avc}}^{\text{avc}} \) can be defined, for
For analytical purposes, we find it useful to define it through
\[ \int_0^\infty e^{-t/t^{up}} P(t) dt = \tilde{P}(1/t^{up}) = 1/2. \] (3)

Experimentally, the relevant time, where almost all search processes end, is \( t^{up} \) and not \( t^{ave} \).

Importantly, the distribution \( R(t) \) has two intrinsic time scales, one short and one long, and can in practice be well approximated by
\[ R(t) \approx 1 - qe^{-t/\tau_1} - (1-q)e^{-t/\tau_2} \] (4)
where \( q, \tau_1 \) and \( \tau_2 \) can be calculated analytically. This form allows an explicit determination of \( t^{up} \) (through Eq. 4) and enables the following interpretation. The short time scale \( \tau_1 = -\frac{1}{q} \frac{dP}{ds}(\lambda_s = 0, s = 0) \) characterizes events where the protein never enters the r state and is therefore independent of the binding energy \( E_b \) (and hence of \( \lambda_s \)) ; \( q = P(\lambda_s = 0, s = 0) \) is the probability of such an event. The time scale \( \tau_2 = (t^{ave} - q\tau_1)/(1-q) \) characterizes events where the protein enters the r state, and is therefore much larger than \( \tau_1 \) in the case of strong binding (\( \lambda_s \) small). As illustrated in Fig. 2, the competition between the two time scales can lead, for DNA lengths which are experimentally relevant, to a significant difference between the typical and average times. More precisely, we find that for DNA lengths \( N < \sqrt{2\lambda_0\lambda_u}/\lambda_s \), \( q \) is of the order of one and \( t^{up} \approx \tau_1 \approx N\sqrt{2\lambda_0/(\lambda_s^{-1} + \lambda_u^{-1})} \) is independent of \( \lambda_u \) - the only rate which depends on the binding energy in \( r \) mode. The relevant time scale of the search process \( t^{up} \) can therefore be much shorter than \( t^{ave} \approx N\lambda_s/\sqrt{2\lambda_0\lambda_u} \) even in the presence of deep traps (\( \lambda_s \) small).

This interesting regime where \( t^{up} \ll t^{ave} \) requires a rather large barrier between the \( s \) and \( r \) state in the case of long DNA molecules (namely, \( \lambda_s < \sqrt{2\lambda_0\lambda_u}/N \)). We now argue that this constraint can be, to a large extent, relaxed when \( n_p \) proteins are searching for the target simultaneously. In this case even when for a single protein \( t^{ave} \approx t^{up} \) the typical search time \( t^{up}_{n_p} \) of \( n_p \) proteins can be significantly shorter than \( t^{ave}/n_p \) even for relatively small values \( n_p \approx 10-15 \). Here, again, \( t^{ave} \) is the average search time of a single protein and \( t^{up}_{n_p} \) is defined as in Eq. 5 where for \( n_p \) proteins the first-passage distribution \( P_{n_p}(t) \) is deduced from the cumulative distribution
\[ R_{n_p}(t) = 1 - (1 - R(t))^{n_p}. \] (5)

In Fig. 4 we show the results of \( R_{n_p}(t) \) for \( n_p = 10 \). Note that as claimed above \( t^{up}_{n_p} \ll t^{ave}/n_p \), whereas \( t^{up} \) is close to \( t^{ave} \) for one protein. This can be understood in the following manner. Using the approximate form, Eq. 5 in Eq. 6 it is obvious that when \( \tau_2 \gg \tau_1 \), the decay of \( R_{n_p}(t) \) is dominated by \( \tau_1 \) as long as \( (1-q)^n \ll 1 \). In essence since only one protein needs to find the target, the probability of a catastrophic event where the search time is of the order of \( \tau_2 \) is \( p_{cat} = (1-q)/n \) which decays exponentially fast with \( n_p \). For large enough values of \( n_p \) the short time scale \( \tau_1 \) controls the behavior of \( R_{n_p}(t) \), even if it is insignificant for the one protein search time. This implies that searches involving several proteins strongly suppress the long time-scales induced by the traps which control \( t^{ave} \). The typical search time is then given by \( t^{up}_{n_p} = \tau_1/m \), where \( m \) is of the order of \( n_p \), and is therefore again widely independent of the binding energy of the \( r \) mode. This makes fast searches possible even in the presence of deep traps - enabling both speed and stability.

We now argue that this mechanism of fast search can still be at play when the binding energy of the protein to the DNA is strongly disordered, as observed in experiments. To account for this we consider the case where the barrier height is drawn from a Gaussian distribution: \( p(E_b) = e^{-(E_b−E_0)/2\sigma^2}/\sqrt{2\pi\sigma^2} \). Importantly, in the presence of disorder we can propose an intrinsic definition of the target as the site with the lowest barrier with no specifically designed properties. Indeed, our previous assumption \( \lambda_r^2 = \infty \) at the target site and \( \lambda^2 \) small anywhere else is a rather strong demand. Since the target sequence is of the order of 10 base-pairs, many sequences with similar properties are very likely to exist, unless the DNA sequence is carefully tailored. To analyze this model we combine numerics with a mean-field analysis. For simplicity, we consider the extreme case where all recognition sites are infinitely long lived \( \lambda_s = 0 \) (or equivalently \( \tau_s = \infty \)), which obviously fulfills the stability requirement. Note that the average search time is then infinite.

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. Knowing the distribution of the minimum of the barrier [24], this is given
by \( p_1 = \int_{-\infty}^{\infty} dE \frac{e^{-E/k_B T}}{1 + \rho \cosh E/k_B T} \frac{1}{2} \text{erfc} \left( \frac{E - E_0}{\sqrt{2} \sigma} \right) \). 

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 holds with \( \lambda_r \) replaced by \( \tilde{\lambda}_r = \lambda_0 \int_{-\infty}^{\infty} e^{-E/k_B T} \frac{(E - E_0)^2}{x \rho} dE \) and \( j \) replaced by

\[
\tilde{j}_p = \frac{p_1 \tilde{j}(z)}{1 - (1 - p_1) \tilde{j}_0(z)}
\]

where \( \tilde{j}_0(z) \) is the generating function of the first return time to site 0 (22).

First, we show that the two scales scenario described above still holds. Indeed, Fig. 4 shows that \( R(t) \) is well fitted by Eq. 4 for realistic values of parameters. 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 \( t^{\text{typ}} \approx \tau_1/m \) with \( m \) of the order of \( n_p \). This enables a fast search even in the presence of infinitely deep traps.

The regime of a fast search with \( t^{\text{typ}} \) independent of the trap depth \( E_r \) also requires, as above, a small \( p_{\text{cat}} \). We now show that this condition holds in a wide range of disorder parameters. To illustrate this, the dependence (holding all other variables constant) of \( p_{\text{cat}} \) and \( t^{\text{typ}} \) on \( \sigma \) is obtained from numerics and the mean-field treatment, is shown in Fig. 5 for realistic values of parameters. Notably, the value of \( p_{\text{cat}} \) 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 is entered in the \( r \)-mode. Increasing \( \sigma \) lowers the barrier at the target and therefore reduces the number of scans needed, which diminishes \( p_{\text{cat}} \). For larger \( \sigma \) the chance of falling into a trap increases due to lower secondary minima of the barrier, which leads to an increase of \( p_{\text{cat}} \). As expected, \( p_{\text{cat}} \) 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 \( \sigma \), \( p_{\text{cat}} \) increases and \( t^{\text{typ}} \) rises quickly as it starts to depend on \( \tau_2 \).

Most important, as advertised above, these results show that it is possible to obtain relatively small values of \( t^{\text{typ}} \) and \( p_{\text{cat}} \) with realistic values of the parameters (see Fig. 5). 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, even in the extreme case of infinitely deep traps suggesting a possible resolution of the speed and stability requirements. We note that by moderate changes in \( E_0 \) similar results can be obtained for much longer DNA sequences.

We thank E. Braun and D. Levine for comments and the support of the High Council for Scientific and Technological Cooperation between France-Israel. Y. K. and M. S. were also supported by the Israeli Science Foundation, and O.B. and R.V. by ANR grant “Dyoptri”.

\[\text{References}\]

[1] O. G. Berg and P. H. von Hippel. J. Biol. Chem., 264, 675, 1989.
[2] R. B. Winter, O. G. Berg, and P. H. von Hippel. Biochem., 20, 6961, (1981).
[3] I. Bonnet et al. Nucl. Acids Res., 36, 4118, (2008).
[4] J. Elf, G.-W. Li, and X. S. Xie. Science, 316, 1191, (2007).
[5] Y. M. Wang et al. Phys. Rev. Lett., 97, 048302, (2006).
[6] M. Coppey et al. Biophys. J., 87, 1640, (2004).
[7] I. Eliazar, T. Koren, and J. Klafter. Journal Of Physics-Condensed Matter, 19, 065140, (2007).
[8] T. Hu, A.Y. Grosberg, and B. I. Shklovskii. Biophys. J., 90, 2731, (2006).
[9] M. A. Lomholt, T. Ambjörnsson, and R. Metzler. Phys. Rev. Lett., 95, 260603, (2005).
[10] M. Slutsky and L. A. Mirny. Biophys. J., 87, 4021, (2004).
[11] B. van den Broek et al. PNAS, 105, 15738, (2008).
[12] O. G. Berg and P. H. von Hippel. J. Mol. Biol., 193, 723, (1987).
[13] U. Gerland, J. D. Moroz, and T. Hwa. PNAS, 99, 12015, (2002).
[14] C. G. Kalodimos et al. Science, 305, 866, (2004).
[15] A. Pingoud and W. Wende. Structure, 15, 391, (2007).
[16] S. A. Townsend et al. Structure, 15, 449, (2007).
[17] P. Hanggi, P. Talkner, and M. Borkovec. Rev. Mod. Phys., 62, 251, (1990).
[18] M. Sheinman and Y. Kafri. Phys. Biol., 6, 016003, (2009).
[19] S. Condamin et al. Nature, 450, 77, (2007).
[20] C. Loverdo et al. Nature Physics, 4, 134, (2008).
[21] S. Redner. A guide to first passage time processes. (Cambridge University Press, Cambridge, England, 2001).
[22] E. W. Montroll. J. Math. Phys, 10, 753, (1969).
[23] O. Benichou et al. Phys. Chem., 10, 7059, (2008).
[24] L. de Haan and A. Ferreira. Extreme Value Theory: An Introduction. (Springer, 2006).
[25] L. Hu, A. Y. Grosberg, and R. Bruinsma. Biophys. J., 95, 1151, (2008).