# Kinetics of Protein-DNA Interaction: Facilitated Target Location in Sequence-Dependent Potential

Michael Slutsky* and Leonid A. Mirny*†
*Department of Physics and †Harvard-MIT Division of Health Sciences and Technology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

ABSTRACT Recognition and binding of specific sites on DNA by proteins is central for many cellular functions such as transcription, replication, and recombination. In the process of recognition, a protein rapidly searches for its specific site on a long DNA molecule and then strongly binds this site. Here we aim to find a mechanism that can provide both a fast search (1–10 s) and high stability of the specific protein-DNA complex ($K_d = 10^{-15}$–$10^{-8}$ M). Earlier studies have suggested that rapid search involves sliding of the protein along the DNA. Here we consider sliding as a one-dimensional diffusion in a sequence-dependent rough energy landscape. We demonstrate that, despite the landscape's roughness, rapid search can be achieved if one-dimensional sliding is accompanied by three-dimensional diffusion. We estimate the range of the specific and nonspecific DNA-binding energy required for rapid search and suggest experiments that can test our mechanism. We show that optimal search requires a protein to spend half of its time sliding along the DNA and the other half diffusing in three dimensions. We also establish that, paradoxically, realistic energy functions cannot provide both rapid search and strong binding of a rigid protein. To reconcile these two fundamental requirements we propose a search-and-fold mechanism that involves the coupling of protein binding and partial protein folding. The proposed mechanism has several important biological implications for search in the presence of other proteins and nucleosomes, simultaneous search by several proteins, etc. The proposed mechanism also provides a new framework for interpretation of experimental and structural data on protein-DNA interactions.

## INTRODUCTION

The complex transcription machinery of cells is primarily regulated by a set of proteins, transcription factors (TFs), that bind DNA at specific sites. Every TF can have from one to several dozens of such specific sites on the DNA. Upon binding to the site, TF forms a stable protein-DNA complex that can either activate or repress transcription of nearby genes, depending on the actual control mechanism. Fast and reliable regulation of gene expression requires 1), fast (~1–10 s) search and recognition of the specific site (referred to as the *target* or *cognate* site below) out of $10^6$–$10^9$ possible sites on the DNA; and 2), stability of the protein-DNA complex ($K_d = 10^{-15}$–$10^{-8}$ M). Despite its apparent simplicity, such a mechanism is not understood in depth, either qualitatively or quantitatively. Here we focus on a simpler case of bacterial TFs recognizing their cognate (target) sites on the naked DNA. Needless to say, eukaryotic protein-DNA recognition is significantly complicated by chromatin packing of the DNA and the multisubunit structure of the TFs. Interestingly, similar problems of specific binding and binding rates arise in the context of oligonucleotides-DNA binding (Lomakin and Frank-Kamenetskii, 1998).

Vast amounts of experimental data available these days provide the structures of protein-DNA complexes at atomic resolution in crystals and in solution (Luscombe et al., 2000; Bell and Lewis, 2001, 2000; Lewis et al., 1996; Schumacher et al., 1994), binding constants for dozens of native and hundreds of mutated proteins (Takeda et al., 1989; Grillo et al., 1999), calorimetry measurements (Spolar and Record, 1994), and novel single-molecule experiments (Shimamoto, 1999). These experimental data contributed most significantly to our present understanding of protein-DNA interaction since the early work of von Hippel and co-workers. In a series of pioneering articles (Berg et al., 1981; Winter et al., 1989; von Hippel and Berg, 1989; Berg and von Hippel, 1987), they created a conceptual basis for describing both the kinetics and thermodynamics of protein-DNA interaction, which has since become a starting point for practically every subsequent theoretical work on the subject.

We start by reviewing the history of the problem and describing the paradox of the faster-than-diffusion association rate. Next, we present the classical model of protein-DNA sliding and explain how this model can resolve the paradox. We outline the problem that the sliding mechanism faces if the energetics of protein-DNA interactions are taken into account. Next we introduce our novel quantitative formalism and undertake an in-depth exploration of possible mechanisms of protein-DNA interaction.

### Faster-than-diffusion search

The problem of how a protein finds its target site on DNA has a long history. In 1970, Riggs et al. (1970a,b) measured the association rate of *Lac*I repressor and its operator on DNA as $\sim 10^{10}\ M^{-1}\ s^{-1}$. This astonishingly high rate (as compared to other biological binding rates) was shown to be much higher than the maximal rate achievable by three-dimensional

*Submitted July 29, 2004, and accepted for publication September 15, 2004.*
Address reprint requests to Leonid A. Mirny, Tel.: 617-452-4862; E-mail: leonid@mit.edu.

0006-3495/04/12/4021/15 $2.00
doi: 10.1529/biophysj.104.050765

diffusion. In fact, if a protein binds its site by three-dimensional diffusion, it has to hit the right site on the DNA within $b = 0.34$ nm. A shift by 0.34 nm would result in binding a site that is different from the native site by 1 bp. Such a site can be very different; e.g., GCGCAATT versus CGCAATTC. Using the Debye-Smoluchowski equation for the maximal rate of a bimolecular reaction (see e.g., Richter and Eigen, 1974; Flyvbjerg et al., 2002; Bruinsma, 2002), with a protein diffusion coefficient of $D_{3d} \sim 10^{-7}$ cm$^2$ s$^{-1}$ (Elowitz et al., 1999) we get

$$k_{DS} = 4\pi D_{3D} b \sim 10^8\ \mathrm{M}^{-1}\,\mathrm{s}^{-1}. \quad (1)$$

This value for the association rate, relevant for in vitro measurements, corresponds to target location in vivo on a timescale of a few seconds, when each cell contains up to several tens of TF molecules.

To resolve the discrepancy between the experimentally measured rate of $10^{10}$ M$^{-1}$ s$^{-1}$ and the maximal rate of $10^8$ M$^{-1}$ s$^{-1}$ allowed by diffusion, Richter and Eigen (1974), and later Berg et al. (1981) and von Hippel and Berg (1989), suggested that the dimensionality of the problem changes during the search process. They concluded that, while searching for its target site, the protein periodically scans the DNA by sliding along it.

## Sliding along the DNA

If a protein performs both three-dimensional and one-dimensional diffusion, then the total search process can be considered as a three-dimensional search followed by binding DNA and a round of one-dimensional diffusion. Upon dissociation from the DNA, the protein continues three-dimensional diffusion until it binds DNA in a different place, and so on. Some experimental evidence supports this search mechanism. These include affinity of the DNA-binding proteins for any fragment of DNA (nonspecific binding), single molecule experiments where one-dimensional diffusion has been observed and visualized, and numerous other experiments where the rate of specific binding to the target site has been significantly increased by lengthening nonspecific DNA surrounding the site (Kim et al., 1987). What are the benefits and the mechanism of one-dimensional diffusion and what limits the search rate?

Here we address this question and consider possible search mechanisms that involve both one-dimensional and three-dimensional diffusion, where one-dimensional diffusion along the DNA proceeds along the rough energy landscape. Quantitative analysis of the search process brought us to the following four main results:

1. When the roughness of the binding energy landscape is $\gtrsim 2\ k_BT$, the diffusion along the DNA becomes extremely slow, with the protein unable to diffuse more than a few basepairs. The total search process is prohibitively slow.
2. If the search proceeds by a combination of one-dimensional and three-dimensional diffusion, nonspecific binding to the DNA plays a very important role in controlling the balance between these two processes. The optimal energy of nonspecific binding can provide the maximal search rate. Although faster than either three-dimensional or one-dimensional search alone, optimal combination of three-dimensional and one-dimensional diffusion cannot expedite the search if the roughness of the landscape is $\gtrsim 2\ k_BT$.
3. Experimentally observed and biologically relevant rates of search can be reached only when one-dimensional sliding proceeds through a fairly smooth landscape with a roughness of the order of $k_BT$.
4. Paradoxically, the stability of the protein-DNA complex at the target site requires a roughness of the binding energy landscape considerably larger than $k_BT$. Rapid search, however, by one-dimensional/three-dimensional diffusion is impossible at such a roughness.

Finally, we formulate this search-speed/stability paradox and suggest a search-and-fold mechanism that can resolve it. The paradox can be resolved if the DNA-binding protein has two distinct (conformational) states in which it exhibits two modes of binding. In the first, which is the mode that has weaker binding and a smoother landscape, it searches for its site. In the second (recognition) mode, which has larger roughness of the binding landscape, the protein tightly binds DNA sites. Correlation between the energy landscapes in the two modes and the energy difference and the barrier between the two protein conformations controls the frequency of transition between the two modes and provides effective preselection of low-energy sites.

We suggest that these modes correspond to two distinct conformational states of the protein-DNA complex (a relatively open complex in the search mode, and a tighter complex in the recognition mode). Transition between the two states can include partial folding of the protein, water extrusion, change in the DNA conformation, etc. Focusing on the conformation of the protein, and without loss of generality, we consider a partially unfolded (disordered) conformation and the folded conformation bound to the cognate site as the two conformations required by our model. In fact, a protein in the partially unfolded conformation may have fewer and/or weaker interactions with DNA allowing rapid sliding. Folded conformation, in turn, provides stronger and more specific interactions required for tight binding.

We also quantify the requirements of this two-mode mechanism to provide both rapid search and stability. Structures of known DNA-binding proteins are known to be flexible and have been reported to exhibit two or more distinct binding modes. This two-state mechanism also agrees well with the results of calorimetric experiments.

The proposed search-and-fold mechanism is not limited to the protein-DNA interaction; it also provides a general

framework for protein-ligand binding and demonstrates the advantages of induced folding, a common theme in molecular recognition.

## THE MODEL

### Search time

In our model, the search process consists of $N$ rounds of one-dimensional search (each takes time of $\tau_{1d,i}$, $i = 1...N$) separated by rounds of three-dimensional diffusion ($\tau_{3d,\,i}$). The total search time $t_s$ is the sum of the times of individual search rounds,

$$t_s = \sum_{i=1}^{N} (\tau_{1d,i} + \tau_{3d,i}). \quad (2)$$

The total number $N$ of such rounds occurring before the target site is eventually found is very large, so it is natural to introduce probability distributions for the essentially random entities in the problem. The first obvious simplification that can be made without any loss of rigor is to replace $\tau_{3d,i}$ by its average $\tau_{3d}$. Each round of one-dimensional diffusion scans a region of $n$ sites (where $n$ is drawn from some distribution $p(n)$). The time, $\tau_{1d}(n)$, that it takes to scan $n$ sites can be obtained from the exact form of the one-dimensional diffusion law (see Appendix A). If, on average, $\bar{n}$ sites are scanned in each round, then the average number of such rounds required to find the site of length $M$ on DNA is $N = M/\bar{n}$. Using average values, we get a total search time of

$$t_s(\bar{n}, M) = \frac{M}{\bar{n}}[\tau_{1d}(\bar{n}) + \bar{\tau}_{3d}]. \quad (3)$$

From Eq. 3 it is clear that, in general, $t_s(\bar{n}, M)$is large for both very small and very large values of $\bar{n}$. In fact, if $\bar{n}$ is small, so few sites are scanned in each round of the one-dimensional search that a large number of such rounds (alternating with rounds of three-dimensional diffusion) are required to find the site. On the other hand, if $\bar{n}$ is large, lots of time is spent scanning a single stretch of DNA, making the search very redundant and inefficient. An optimal value, $\bar{n}_{opt}$, should exist, which provides little redundancy of one-dimensional diffusion and a sufficiently small number of such rounds. For a given diffusion law $\tau_{1d}(n)$, function $t_s(\bar{n}, M)$ can be minimized producing $\bar{n}_{opt}$, the optimal length of DNA to be scanned between the association and the dissociation events. (Naturally, we assume here that $\tau_{1d}(\bar{n})$ grows with $\bar{n}$ at least as $O(\bar{n}^{1+\alpha})$, with $\alpha > 0$.)

### Protein-DNA energetics

While diffusing along DNA, a TF experiences the binding potential $U(\vec{s})$ of every site $\vec{s}$ it encounters. The energy of protein-DNA interactions is usually divided into two parts—*specific* and *nonspecific* (Berg and von Hippel, 1987; Gerland et al., 2002),

$$U_i = U(\vec{s} = s_i, \ldots s_{i+l-1}) + E_{ns}, \quad (4)$$

—where $\vec{s}$ describes a binding DNA sequence of length $l$. As its name suggests, the nonspecific binding energy $E_{ns}$ arises from interactions that do not depend on the DNA sequence that the TF is bound to, e.g., interactions with the phosphate backbone. The specific part of the interaction energy exhibits a very strong dependence on the actual nucleotide sequence. Here and below we use the term *energy* to refer to the change in the free energy related to binding $\Delta G_b$. This free energy includes the entropic loss of translational and rotational degrees of freedom of the protein and amino acids' side chains, the entropic cost of water and ion extrusion from the DNA interface, the hydrophobic effect, etc.

The energy of specific protein-DNA interactions can be approximated by a weight matrix (also known as *PSSM*, or *profile*) where each nucleotide contributes independently to the binding energy (Berg and von Hippel, 1987),

$$U(\vec{s} = s_i, \ldots s_{i+l-1}) = \sum_{j=1}^{l} \epsilon(j, s_j), \quad (5)$$

where $s_j$ is a basepair in position $j$ of the site and $\epsilon(j, x)$ is the contribution of basepair $x$ in position $j$. Most of the known weight matrices of TFs $\epsilon(j, s_j)$ give rise to uncorrelated energies of overlapping neighboring sites, obtained by one basepair shift (Gerland et al., 2002). Fig. 1 presents distributions of the sequence specific binding energy $f(U)$ obtained for different bacterial transcription factors and all possible sites in the corresponding genome. The weight matrices for these transcription factors has been derived using a set of known binding sites and standard approximation (Berg and von Hippel, 1987; Stormo and Fields, 1998). Notice that for a sufficiently long site the distribution of the binding energy of random sites (or genomic DNA) can be closely approximated (see Fig. 1) by a Gaussian distribution with a certain mean $\langle U \rangle$ and variance $\sigma^2$,

$$f(U_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(U_i - \langle U \rangle)^2}{2\sigma^2}\right]. \quad (6)$$

We also assume independence of the energy of neighboring (although overlapping) sites. Binding energies calculated for bacterial TFs support this assumption. Other physical factors such as local DNA flexibility (Erie et al., 1994) can create a correlated energy landscape, providing a different mode of diffusion, as we have described in Slutsky et al. (2004).

### Diffusion in a sequence-dependent energy landscape

The whole DNA molecule can thus be mapped onto a one-dimensional array of sites, $\{\vec{s}_i\}$—each corresponding to a certain binding sequence comprising bases from the $i^{th}$ to the $(i + l–1)^{th}$, $l$ being the length of the motif (see Fig. 2). At each site, there is a probability $p_i$ of hopping to site $i + 1$ and a probability $q_i$ of hopping to site $i–1$. These probabilities depend on the specific binding energies, $U_i$ and $U_{i\pm1}$, at the $i^{th}$ site and at the adjacent sites, respectively, and are proportional to the corresponding transition rates, $\omega_{i,i+1}$ and $\omega_{i,i-1}$. For the latter, it is most natural to assume the regular activated transport form

$$\omega_{i,i\pm1} = \nu \times \begin{cases} e^{-\beta(U_{i\pm1} - U_i)} & \text{if} \quad U_{i\pm1} > U_i \\ 1.0 & \text{otherwise} \end{cases}, \quad (7)$$

where $\nu$ is the effective attempt frequency, $\beta \equiv (k_B T)^{-1}$; $k_B$ is the Boltzmann constant; and $T$ is the ambient temperature. Having defined that, we have a one-dimensional random walk with position-dependent hopping probabilities.

As has been shown in numerous articles throughout the last two decades, the properties of one-dimensional random walks can vary dramatically depending on the actual choice of probabilities, $\{p_i\}$ (for review, see Bouchaud and Georges, 1990). Here we employ the mean first-passage time formalism (Murthy and Kehr, 1989) to derive the diffusion law $\tau_{1d}(\bar{n})$ for protein sliding along the DNA given the sequence-dependent binding energy (Eq. 7).

## RESULTS

Using the model described above, we studied the following problems:

1. How fast is the one-dimensional search on DNA as a function of the roughness, $\sigma$, of the binding energy landscape?

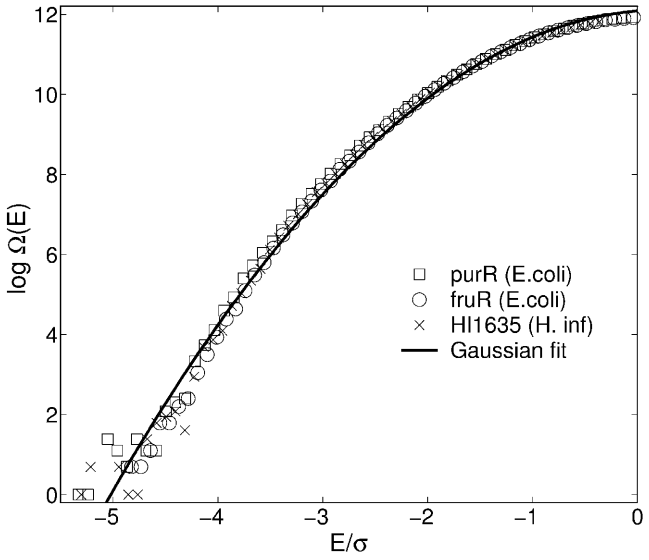


FIGURE 1 Spectrum of binding energy for three different transcription factors and the Gaussian approximation (*solid line*).

2. How significant is the role of nonspecific binding energy, $E_{ns}$, in determining the search time?
3. How fast is the search for the native site under conditions that provide stability to the protein-DNA complex at the target site?

## Diffusion along the DNA

We state here the main results without a derivation (which can be found in Appendix A). For a given set of probabilities $\{p_i\}$, the mean first-passage time (MFPT) from $i = 0$ to $i = L$ (in terms of number of steps) is (Murthy and Kehr, 1989)

$$\bar{t}_{0,L} = L + \sum_{k=0}^{L-1} \alpha_k + \sum_{k=0}^{L-2} \sum_{i=k+1}^{L-1} (1 + \alpha_k) \prod_{j=k+1}^{i} \alpha_j, \qquad (8)$$

where $\alpha_i \equiv q_i/p_i$. The relation in Eq. 8 gives the MFPT for one given realization of probabilities. Assuming that the specific binding energies $\{U_i\}$ have a normal distribution with variance $\sigma^2$ (see above), we plug the probabilities in Eq. 7 into Eq. 8 and after a somewhat lengthy but straightforward calculation, we obtain an expression for the MFPT averaged over genomic sequences for $L \gg 1$,

$$\langle \bar{t}_{FP}(L) \rangle \simeq \tau_0 L^2 e^{7\beta^2\sigma^2/4} (1 + \beta^2\sigma^2/2)^{-1/2}, \qquad (9)$$

where $\tau_0$ is the reciprocal of the effective attempt frequency for hopping to a neighboring site.

The main result is that the one-dimensional search by hopping to neighboring sites proceeds by normal diffusion with $t \sim L^2/2D_{1d}$, where the diffusion coefficient

$$D_{1d}(\sigma) \simeq \frac{1}{2\tau_0} \left(1 + \frac{\beta^2\sigma^2}{2}\right)^{1/2} e^{-7\beta^2\sigma^2/4} \qquad (10)$$

exhibits an exponential dependence on the roughness of the binding energy landscape $\sigma$, dropping rapidly as $\sigma$ becomes greater than a few $k_BT$ (Slutsky et al., 2004). Hence, rapid diffusion of a protein along the DNA is possible only if the roughness of the binding energy landscape is small compared to $k_BT$ ($\beta\sigma < 1.5$). This requirement imposes strong constraints on the allowed energy of specific binding interactions.

## Optimal time of three-dimensional/one-dimensional search

When one-dimensional scanning is combined with three-dimensional diffusion, what is the optimal time a protein has to spend in each of the two regimes? To answer this question we compute the optimal number of sites the protein has to scan by one-dimensional diffusion to get the fastest overall search. Results of this section are rather general and are not limited to the particular scenario of slow one-dimensional diffusion on a rough landscape discussed above.

Each time the protein binds DNA it performs a round of one-dimensional diffusion. If the round lasts $\tau_{1d}$, then, on average, the protein scans $\bar{n} = \sqrt{16 D_{1d}\tau_{1d}/\pi}$ bps (Hughes, 1995). By plugging this relation into Eq. 3 for search time $t_s$, and minimizing $t_s$ with respect to $\bar{n}$, we get the optimal total search time and the optimal number of sites to be scanned in each round,

$$t_s^{opt} = t_s(\bar{n}_{opt}) = \frac{M}{2}\sqrt{\frac{\pi\bar{\tau}_{3d}}{D_{1d}}} \quad \bar{n}_{opt} = \sqrt{\frac{16}{\pi} D_{1d}\bar{\tau}_{3d}}, \qquad (11)$$

which brings us to the following conclusions.

First, and most importantly, we obtain that, in the optimal regime of search,

$$\tau_{1d}(\bar{n}_{opt}) = \tau_{3d}, \qquad (12)$$

i.e., the protein spends equal amounts of time diffusing along nonspecific DNA and diffusing in the solution. This striking result is very general, and is true irrespective of the values of diffusion coefficients $D_{1d}$ or $D_{3d}$, or size of the genome $M$. In fact it follows directly from the diffusion law $\bar{n} \sim \sqrt{\tau_{1d}}$. More importantly this central result can be verified experimentally by either single-molecule techniques or by traditional methods.

Also note that the optimal region of the DNA scanned in a single round of one-dimensional diffusion $\bar{n}_{opt}$ does not depend on $M$—i.e., is the same irrespective of the size of the genomes to be searched for a specific site.

Second, the optimal one-dimensional/three-dimensional combination reached at $\tau_{1d} = \tau_{3d}$ leads to a significant speedup of the search process. In fact, an optimal one-dimensional/three-dimensional search is $\bar{n}_{opt}$ times faster than a search by three-dimensional diffusion alone, and $M/\bar{n}_{opt}$ times faster than a search by one-dimensional diffusion alone. For example, if the protein operates in the

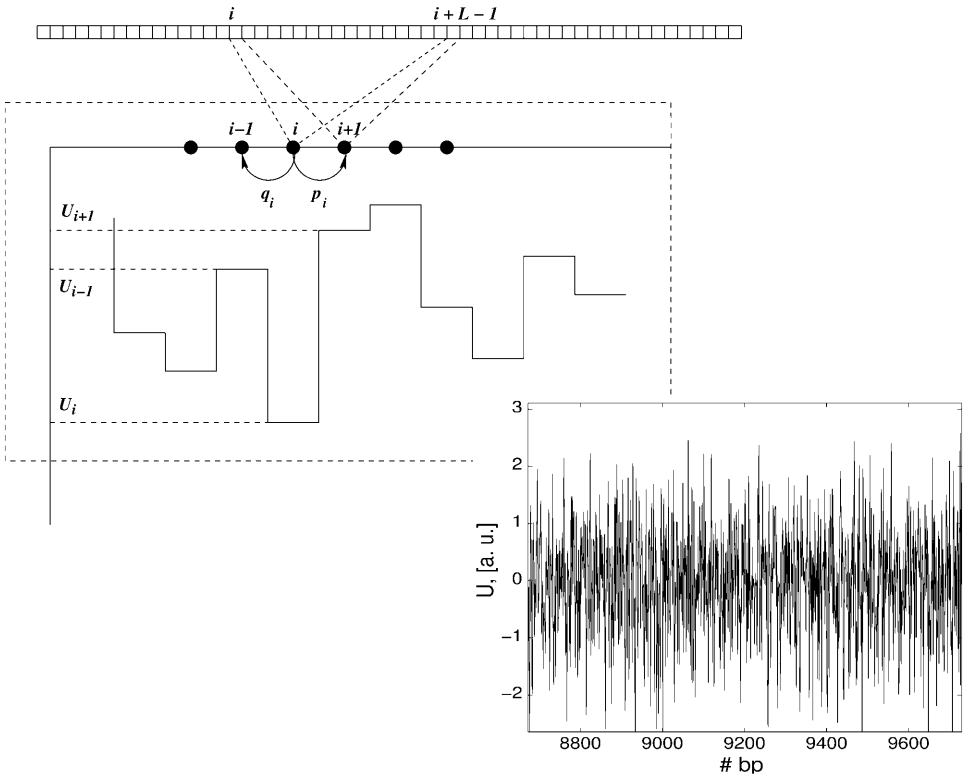


FIGURE 2 The model potential.

optimal one-dimensional/three-dimensional regime and scans $\bar{n}_{\rm opt} = 100$ bp during each round of DNA binding, then the experimentally measured rate of binding to the specific site can be 100 times greater than the rate achievable by three-dimensional diffusion alone.

Third, we can estimate $\bar{n}_{\rm opt}$, the maximal number of sites a protein can scan in each round of one-dimensional search. If we set $D_{1\rm d}$ to its maximum, i.e., $D_{1\rm d} \sim D_{3\rm d}$ and $\bar{\tau}_{3\rm d} \sim l_{\rm d}^2/D_{3\rm d}$, with $l_{\rm m} \sim 0.1$ $\mu$m, we get

$$\bar{n}_{\rm opt}^{\rm max} \sim 500\,{\rm bp}. \tag{13}$$

For a smaller one-dimensional diffusion coefficient, e.g., $D_{1\rm d} \sim D_{3\rm d}/100$, we get $\bar{n}_{\rm opt}^{\rm max} \sim 50$ bp. Again, single molecule experiments can provide estimates of these quantities for different conditions of diffusion.

Finally, we obtain estimates of the shortest possible total search time. If $M \approx 10^6$ bp and one-dimensional diffusion is at its fastest rate, i.e., $D_{1\rm d} \sim D_{3\rm d} = 10^{-7}{\rm cm}^2/{\rm s}$, then using Eq. 11 we get

$$t_{\rm s}^{\rm opt} \sim \frac{M}{2}\sqrt{2\pi\bar{\tau}_{3\rm d}\tau_0} \sim 5\,{\rm s}, \tag{14}$$

where we estimate $\tau_0 \sim a_0^2/D_{3\rm d} \sim 10^{-8}$ s.

One can also estimate the search time using in vitro experimentally measured binding rates in water, $k_{\rm on}^{\rm water} \approx 10^{10}\,{\rm M}^{-1}\,{\rm s}^{-1}$ (Riggs et al., 1970a,b). The diffusion coefficient of a protein in the cytoplasm is 10–100 times lower than that in water, leading to the estimated binding rate of $k_{\rm on}^{\rm cytoplasm} \approx 10^8 - 10^9\,{\rm M}^{-1}\,{\rm s}^{-1}$ (see Appendix D). From this we obtain the time it takes for one protein to bind one site in a cell of 1 $\mu{\rm m}^3$ volume (i.e., [TF] $\approx 10^{-9}$ M) as

$$t_{\rm s}^{\rm exp} = (k_{\rm on}^{\rm cytoplasm}[{\rm TF}])^{-1} \sim 1 - 10\,{\rm s}. \tag{15}$$

One can see perfect agreement between our theoretical estimates and experimentally measured binding rates.

As we mentioned above, there are usually several TF molecules searching in parallel for the target site. Naturally, in this case, the search is sped up proportionally to the number of molecules.

## Diffusion of *PurR* on the *Escherichia coli* genome

To check the applicability of the above considerations, we simulated one-dimensional diffusion of *PurR* transcription factor on the *E. coli* chromosome.

The specific energy profile was built using a weight matrix derived from 35 *PurR* binding sites following a standard procedure described elsewhere (Berg and von Hippel, 1987; Stormo and Fields, 1998). The resulting energy profile is random and uncorrelated and has a standard deviation $\sigma \simeq 6.5\ k_{\rm B}T$. This profile was used as an input for calculating mean first passage time at different temperatures. (Since the magnitude of the interaction is fixed, in these calculations we vary temperature rather than binding strength.) The result of these calculations is presented in Fig. 3. It is clear that when the roughness of the landscape becomes significant at

$\sigma > 2\ k_BT$, the diffusion proceeds extremely slowly. Only ~10–100 bp can be scanned by a TF when $\sigma = 2\ k_BT$. A natural requirement for sufficiently fast diffusion is, as before, $\sigma \sim k_BT$.

## Nonspecific binding

Whereas the diffusion of the TF molecules along DNA is controlled by the specific binding energy, the dissociation of the TF from the DNA depends on the total binding energy, i.e., on the nonspecific binding as well as on the specific one. Moreover, since the dissociation events are much less frequent than the hopping between neighboring basepairs (roughly by a factor of $\bar{\tau}_{3d}/\langle\tau\rangle$), the nonspecific energy $E_{ns}$ makes a sensibly larger contribution to the total binding energy.

For a TF at rest bound to some DNA site $i$, the dissociation rate, $r_i$, would be given by the Arrhenius-type relation,

$$r_i = \frac{1}{\tau_0} e^{-\beta(E_{ns}-U_i)}. \tag{16}$$

Given the specific ($U_i$) and the nonspecific ($E_{ns}$) energy, one can calculate the average time, $\tau_{1d}$, a protein spends before dissociating from the DNA (see Appendix B). We obtain

$$E_{ns} = k_BT\left[\ln\left(\frac{\tau_{1d}}{\tau_0}\right) - \frac{1}{2}\left(\frac{\sigma}{k_BT}\right)^2\right], \tag{17}$$

and in the optimal regime where $\tau_{1d} = \bar{\tau}_{3d}$,

$$E_{ns}^{opt} = k_BT\left[\ln\left(\frac{\tau_{3d}}{\tau_0}\right) - \frac{1}{2}\left(\frac{\sigma}{k_BT}\right)^2\right]. \tag{18}$$

## The parameter space

Since for a given value of $\sigma$, the nonspecific binding controls the dissociation rate, the search time will deviate from the optimum if $E_{ns}$ moves from this predetermined value. In Fig. 4 *a* we plot the search time as a function of the nonspecific binding energy for different values of $\sigma$.

We now define the tolerance factor, $\zeta$, as the ratio between the acceptable value of the search time, $t_s$, and the optimal search time, $t_s^{opt}$. Experimental data suggest $\zeta \leq 5$, but for the moment we allow for much larger values of $\zeta \sim$ 10–100 (this can be done when, for instance, there are many protein molecules searching in parallel). As we can see from Fig. 4 *a*, for each value of $\sigma$, there is a range of possible values of $E_{ns}$ such that the resulting search time is within the region of tolerance (see Appendix B). Note the dramatic increase in the search time as $E_{ns}$ deviates from its optimal value.

Specifying $\zeta$, we can define our parameter space, i.e., the values of specific and nonspecific energy producing a total search time within the region of tolerance. In Fig. 4 *b*, we consider three values of $\zeta$. The most relaxed requirement $\zeta =$ 100 provides a search time of $t_s \leq 500$ s. If 100 proteins are searching for a single site, then the first one will find it after

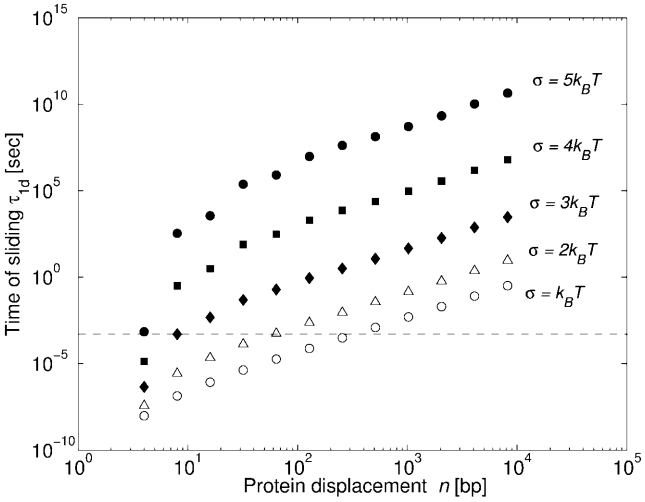


FIGURE 3 The mean first passage time versus traveling distance for *purR* transcription factor on the binding landscapes of different roughness (or at different temperatures). The horizontal line indicates the optimal regime, $\tau_{1d} \sim \bar{\tau}_{3d}$.

~5 s—leading, however, to a fairly low binding rate of $k_{on} \approx 1/500\ \text{s} \cdot 10^9\ \text{M}^{-1} = 2 \cdot 10^6\ \text{M}^{-1}\ \text{s}^{-1}$ (compared to experimentally measured $10^{10}\ \text{M}^{-1}\ \text{s}^{-1}$ in water). Importantly, to comply with even this most relaxed search time requirement, the characteristic strength of specific interaction must be $\lesssim 2.3\ k_BT$.

These results bring us to a very important conclusion that a protein cannot find its site in biologically relevant time if the roughness of the specific binding landscape is $\gtrsim 2\ k_BT$. Although an optimal one-dimensional/three-dimensional combination can speed up the search, it cannot overcome the slowdown of one-dimensional diffusion. Only fairly smooth landscapes ($\sigma \sim 1\ k_BT$) can be effectively navigated by proteins.

## Speed versus stability

Whereas rapid search requires fairly smooth landscapes ($\sigma \sim 1\ k_BT$), stability of the protein-DNA complex, in turn, requires a low energy of the target site ($U_{min} < 15\ k_BT$ for a genome of $10^6$ bp).

In Fig. 5 *a*, we present the equilibrium probability $P_b$ of binding the strongest target site with energy $U_{min} = U_0$ (Gerland et al., 2002) as a function of $\sigma/k_BT$. In equilibrium, $P_b$ equals the fraction of time the protein spends at the target site,

$$P_b = \frac{\exp[-\beta U_0]}{\sum_{i=0}^{M} \exp[-\beta U_i]}. \tag{19}$$

Since the target site is not separated from the rest of the distribution by a significant energy gap, $P_b$ is comparable to 1 (which is the natural requirement for a good regulatory site) only at $\sigma \gg k_BT$.

Fig. 5 *b* shows the optimal search time at the corresponding values of $\sigma/k_BT$. High roughness of $\sigma \gg k_BT$ required for

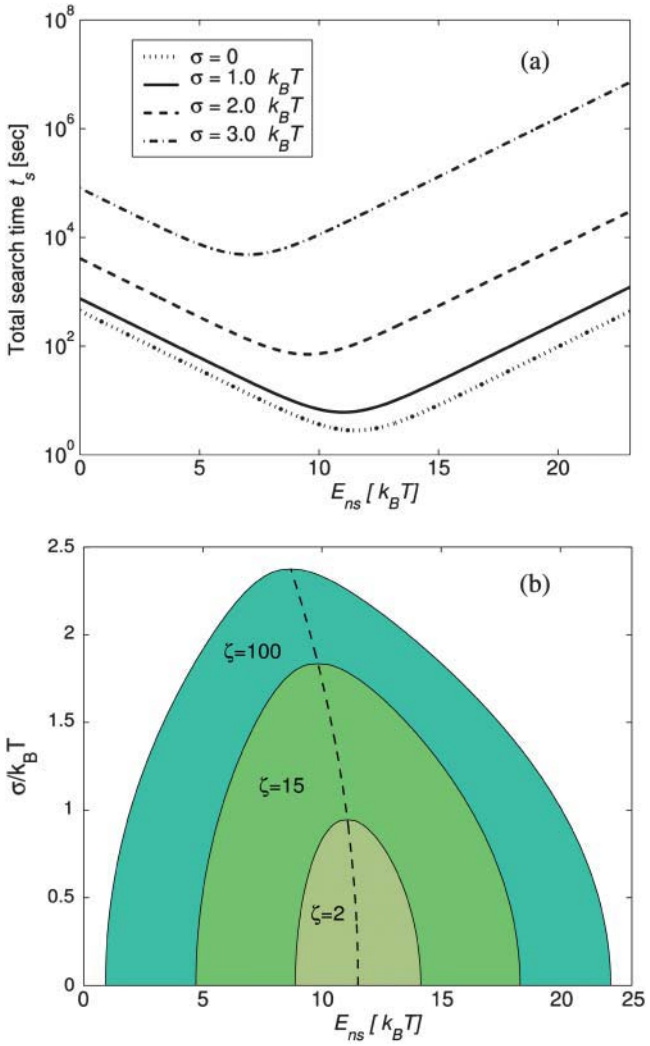


FIGURE 4 (*a*) Dependence of the search time on the nonspecific binding energy. (*b*) The parameter space. The dashed line corresponds to optimal parameters $\sigma$ and $E_{ns}$ connected by Eq. 18.

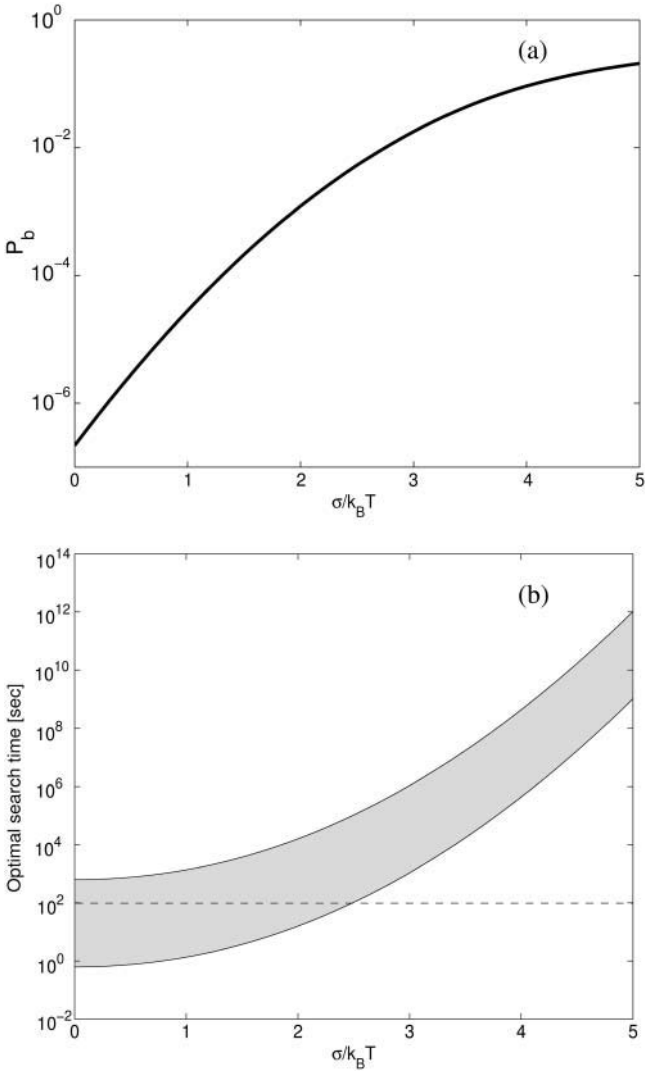


FIGURE 5 (*a*) Stability on the protein-DNA complex on the cognate site measured as the fraction of time in the bound state at equilibrium. (*b*) Optimal search time as a function of the binding profile roughness, for the range of parameters $10^{-4}$ s $\leq \tau_{3d} \leq 10^{-2}$ s, $10^{-10}$ s $\leq \tau_0 \leq 10^{-6}$ s.

stability of the protein-DNA complex leads to astronomically large search times. In contrast, a protein can effectively search the target site at $\sigma < 1$–$2\ k_BT$.

This brings us to the central result that the ability to translocate rapidly along the DNA clearly cannot comply with the stability requirement.

Requirement of high stability at the target site, $P_b \sim 1$ (or $P_b \sim 1/N_p$, if $N_p$ copies of the protein are present), yields an estimate for the minimal $\sigma$ of

$$\sigma \sim k_BT\sqrt{2\ln M} \sim 5\,k_BT, \tag{20}$$

given a genome size $M = 10^6$.

From the above analysis, an obvious conflict arises: the same energy landscape cannot allow for both rapid translocation and high stability of states formed at sites with the lowest energy. This conflict is similar to the speed-stability paradox of protein folding formulated by Gutin et al. (1998): rapid search in conformation space requires a smooth energy landscape, but then the native state is unstable. In protein folding, this conflict is resolved by the presence of a large energy gap between the native state and the rest of the conformations (Finkelstein and Ptitsyn, 2002; Pande et al., 2000).

As evident from Fig. 1, no such energy gap separates cognate sites from the bulk of other (random) sites. In fact, the energy function in the form of Eq. 5 cannot, in principle, provide a significant energy gap. Increasing the number of TFs cannot resolve the paradox either (see Appendices D and E). An alternative solution must be sought.

## The two-mode model

The search-speed stability paradox has already been qualitatively anticipated by Winter et al. (1989), who therefore concluded that a conformational change of some sort

should exist that would allow fast switching between the *specific* and the *nonspecific* modes of binding. In the nonspecific mode, the protein is sliding over an essentially equipotential surface (in our terms, $\sigma_{\text{non-spec}} = 0$), whereas site-binding takes place in the specific mode ($\sigma_{\text{spec}} \gg k_BT$). A protein in the nonspecific binding mode is ''unaware'' of the DNA sequence it is bound to. Thus, it should permanently alternate between the binding modes, probing the underlying sites for specificity.

This model naturally raises a question about the nature of the conformation change. Originally, it was described as a microscopic binding of the protein to the DNA accompanied by water and ion extrusion. However, numerous calorimetry measurements and calculations (Spolar and Record, 1994) show that such a transition is usually accompanied by a large heat capacity change $\Delta C$. This $\Delta C$ cannot be accounted for, unless additional degrees of freedom, namely, protein folding, are taken into account. On-site folding of the transcription factor may involve significant structural change (Flyvbjerg et al., 2002; Bruinsma, 2002; Kalodimos et al., 2004) and take a time of $\sim 10^{-4}$–$10^{-6}$ s (Akke, 2002) (compared to a characteristic on-site time of $\tau_0 \sim 10^{-7}$–$10^{-8}$ s). We conclude that conformational transition between the two modes involves (but is not limited to) partial folding of the TF.

If the TF is to probe every site for specificity in this fashion, it would take hours to locate the native site. We note, however, that if there was a way to probe only a very limited set of sites, i.e., only those having high potential for specificity, the search time would be dramatically reduced. From the previous section it is clear that a relatively weak site-specific interaction (i.e., smooth landscape, $\sigma \sim k_BT$) does not significantly affect the diffusive properties of the DNA and the total search time. If this landscape, however, is correlated with the actual specific binding energy landscape (with $\sigma \sim 5$–$6\ k_BT$), the specific sites will be the strongest sites in both modes. The protein conformational changes should occur therefore mainly at these sites, which constitute *traps* in the smooth landscape. Since such sites constitute a very small fraction of the total number of sites, the transitions between the modes are very rare.

We therefore suggest that there are two modes of protein-DNA binding: the *search* mode and the *recognition* mode (Fig. 6). In the search mode, the protein conformation is such that it allows only a relatively weak site-specific interaction ($\sigma_s \sim 1.0$–$2.0\ k_BT$) (Fig. 6, *top*). In the recognition mode, the protein is in its final conformation and interacts very strongly ($\sigma_r \geq 5\ k_BT$) with the DNA (Fig. 6, *bottom*). If two energy profiles are strongly correlated, then the lowest-lying energy levels (i.e., *traps*) in the search mode ($\leq -5\ k_BT$) are likely to correspond to the strongest sites in the recognition mode (putatively, the cognate sites). The transitions between the two modes happen mainly when the protein is trapped at a low-energy site of the search landscape. In this fashion, the one-dimensional diffusion coefficient $D_{1d}$ is $\sim$10–100 times smaller than the ideal limit, but the search time in the optimal regime is reduced only by a factor of $\sim$3–10 (Eq. 11).

The coupling between the conformational change and association at a site with a low-energy trap is likely to take place through time conditioning. Namely, the folding (or a similar conformation transition) occurs only if the protein spends some minimal amount of time bound to a certain site. This statement is basically equivalent to saying that the free energy barrier that the protein must overcome to transform to the final state must be comparable to the characteristic energy difference that controls hopping to the neighboring sites.

The protein conformation in the recognition mode should be stabilized by additional protein-DNA interactions. If these interactions are unfavorable, the folded structure is destabilized; the search conformation is then rapidly restored and the diffusion proceeds as before. If the new interactions are favorable, however, the folded structure is stable and the protein is trapped at the site for a very long time.

For this mechanism to work, transition between the two modes of search has to be associated with a significant change in the free energy ($\sim$5..10 $k_BT$) of the protein-DNA complex (see Fig. 6 *c*). Such an energy difference between the two states is required to make the majority of the high-energy sites in the recognition mode less favorable than in the search mode. A protein would rather (partially) unfold than bind an unfavorable site. As a result, sites that lay higher in energy than a certain cutoff exhibit a similar nonspecific binding energy (i.e., there is a switch into the search mode of binding). The folding of partially disordered protein loops or helices can provide the required free energy difference between the two modes.

Efficiency of the proposed search-and-fold mechanism depends on the energy difference between the two modes, correlation between the energy profiles, and the barrier between the two states. The barrier determines the rate of partial folding-unfolding transition. If the barrier is too low, then the protein equilibrates while on a single site, having no effect on search kinetics. On the contrary, too high a barrier can lead to rear folding events and the cognate site can be missed. It can be shown that having a barrier of proper size provides for an efficient search and stable protein-DNA complexes. Alternatively, the cognate site can lower the barrier by stabilizing the transition state (i.e., the folding nucleus; see Abkevich et al., 1994; Mirny and Shakhnovich, 2001), whereby it acts as a catalyst of partial folding. (Quantitative analysis of these factors is beyond the scope of this study, and will be published elsewhere.)

## DISCUSSION

### Specificity for free: kinetics versus thermodynamics

The proposed mechanism of specific site location is akin to kinetic proofreading (Hopfield, 1974), which is a very general

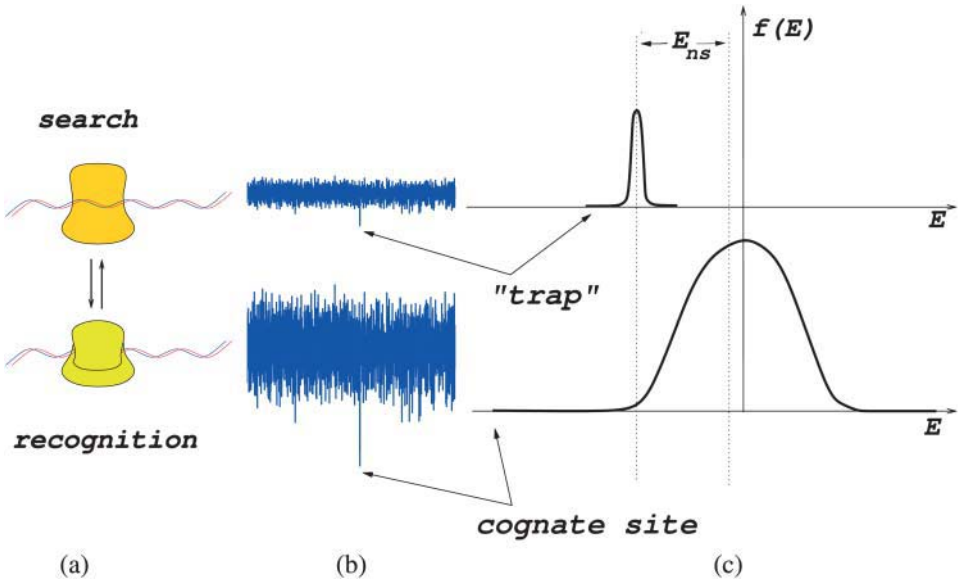


FIGURE 6 Cartoon demonstrating the two-mode search-and-fold mechanism. (*Top*) Search mode; (*bottom*) recognition mode. (*a*) Two conformations of the protein bound to DNA: partially unfolded (*top*) and fully folded (*bottom*). (*b*) The binding energy landscape experienced by the protein in the corresponding conformations. (*c*) The spectrum of the binding energy determining stability of the protein in the corresponding conformations.

concept for a broad class of high-specificity biochemical reactions. The required specificity is achieved in kinetic proofreading through formation of an intermediate metastable complex that paves the way for irreversible enzymatic reaction. If the reaction is much slower than the lifetime of the complex, then substrates that spend enough time in the complex are subject to the enzymatic reaction, whereas substrates that form short-lived complexes are released back to the solvent before the reaction takes place. In other words, the substrates are selected by kinetic partitioning.

In contrast to kinetic proofreading that increases equilibrium specificity for the price of energy consumption, the search-and-fold model does not require any additional source of energy. The two-mode search-and-fold model provides a faster on-rate of binding while keeping the equilibrium binding constant unchanged. Naturally, the off-rate is increased as well. This makes our two-mode model thermodynamically neutral.

## Coupling of folding and binding in molecular recognition

Several DNA- and ligand-binding proteins are known to have partially unfolded (disordered) structures in the unbound state. The unstructured regions fold upon binding to the target. Does binding-induced folding provide any biological advantage?

The idea of coupling between local folding and site binding has been around for some time and was recently reassessed in the much broader context of intrinsically unstructured proteins (Wright and Dyson, 1999; Dyson and Wright, 2002; Uversky, 2002). Induced folding of these proteins can have several biological advantages. First, flexible unstructured domains have an intrinsic plasticity that allows them to accommodate targets/ligands of various sizes and shapes; and second, free energy of binding is required for compenstation for the entropic cost of ordering of the unstructured region. A poor ligand that does not provide enough binding free energy cannot induce folding and, hence, cannot form a stable complex. Williams et al. (2001) have suggested that unstructured domains can be the result of evolutionary selection that acts on the bound (structured) conformation, while ignoring the unbound (unstructured) conformation. Partial unfolding can also increase protein's radius of gyration and, hence, increase the binding rate (Shoemaker et al., 2000; Levy et al., 2004).

Here we propose a mechanism that suggests the role of induced folding in providing rapid and specific binding. Induced folding (or any sort of two-state conformational transition) allows a protein to search and recognize DNA in two different conformations providing rapid binding to the target site. Importantly, this mechanism reconciles rapid search for the target site with a stable bound complex (see above). The rate of induced folding can also play a role in determining the specificity of recognition (M. Slutsky and L.A. Mirny, unpublished).

Structural and thermodynamic data argue in favor of distinct protein conformations for search along noncognate DNA and for recognition of the target site. Proteins such as λcI, *Eco*RV, and GCN4 apparently do not fold their unstructured regions while bound to noncognate DNA (Winkler et al., 1993; Clarke et al., 1991; O'Neil et al., 1990); this supports our hypothesis.

Heat capacity measurements on a vast variety of protein-DNA complexes report a large negative heat capacity change in site-specific recognition, which is a clear indication of a phase transition. These measurements supplemented by x-ray crystallography and NMR structural data were interpreted by Spolar and Record (1994), mainly in terms of hydrophobic and conformational contributions to entropy.

Thus, folding-binding coupling is now considered a well-established effect for a large set of transcription factors.

However, real-time kinetic measurements were not performed until recently, so that the question of the actual mechanism was left open. Serious advances in this direction were made by Kalodimos et al. (2001, 2002, 2004), who observed a two-step site recognition by dimeric *Lac* repressor. The H/D-exchange NMR data unambiguously demonstrates site preselection by $\alpha$-helices bound in the major groove followed by folding of hinge helices that bind to the minor groove elements and complete the specific site recognition. Although the experiments in this field were performed with a single model system, their implications are likely to have a general character.

It should be mentioned that no transition of this kind is observed when the protein is unbound from DNA. A possible reason for this can be a significant reduction of the free energy barrier for folding, entropic in essence, that accompanies protein-DNA association. Entropy barrier reduction is a natural consequence of relative anchoring of the various parts of the protein on the DNA scaffold. Thermal fluctuations that the associated protein is subject to are generally of the order of $\sim k_BT$, and their main effect is protein translocation along the DNA. From the above analysis, it follows that the translocation actually takes place only if the protein encounters barriers of $\sigma_s \sim k_BT$ on its way. In a large enough collection of sites ($M \gg 1$), however, potential wells of depth $\sim\sigma_s\sqrt{2\ln M}$ will be present. If the well depth is larger than the folding barrier height, the probability of on-site (in-well) folding increases, leading eventually to a stable complex formation. (More detailed computational analysis of coupling between folding and binding will be published elsewhere.)

## Biological implications

The mechanism of three-dimensional/one-dimensional search described above has several biological implications. The studied model, as with any quantitative model, is, of course, a gross simplification of protein-DNA recognition in vivo. Despite this simplification, proposed mechanism can be generalized to describe the in vivo binding. Here we briefly discuss some of the biological implications of our model.

## Simultaneous search by several proteins

If several TFs are searching for its site on the DNA, the total search time is given by Eq. 15 and is obviously shorter than the time for a single TF. For example, if 100 copies of a TF are searching in parallel for the cognate site, then assuming $k_{on}^{cytoplasm} \approx 10^8\,M^{-1}\,s^{-1}$ and a cell of 1 $\mu m^3$ volume, we obtain the search time of $t_s \approx 0.1$ s. Increasing the number of TF molecules can further decrease the search time, but can have harmful effects due to molecular crowding in the cell. Note, however, that increasing the number of TF molecules to 100–1000 per cell cannot resolve the speed-stability paradox (see Fig. 5).

## Search inside a cell: molecular crowding on DNA and chromatin

Above we assumed that a TF is free to slide along the DNA. The in vivo picture is complicated by other proteins and protein complexes (nucleosomes, polymerases, etc.) that are bound to DNA, preventing a TF from sliding freely along the DNA. What are the effects of such molecular crowding on the search time?

Our model suggests that molecular crowding on DNA will have little effect on the search time if certain conditions are satisfied. Obviously, the cognate shall not be screened by other DNA-bound molecules/nucleosomes. DNA-bound molecules can interfere with the search process by shortening regions of DNA scanned on each round of one-dimensional diffusion. If, however, the distance between DNA-bound molecules/nucleosomes in the vicinity of the cognate site is greater than $\bar{n}_{opt} \sim 300 - 500$ bp (see Eq. 13 and Kim et al., 1987), then obstacles on the DNA do not shorten the rounds of one-dimensional diffusion and, hence, do not slow down the search process. Our analysis also suggests that sequestration of part of genomic DNA by nucleosomes can even speed up the search process.

If DNA-bound proteins are separated by >300–500 bp, *E. coli* genomic DNA can accommodate $4.6 \times 10^6$ bp/300 bp $\approx$ $1.5 \times 10^4$ proteins. In other words, all 150 known and predicted *E. coli* TFs can be simultaneously present in 100 copies each, and search for their cognate sites without affecting one other (in fact, they can be present in 200 copies each, since optimal search requires 50% of proteins to be in solution at any one time). On the other hand, a short ~50-bp linker between nucleosomes in eukaryotic chromatin can increase the search time ~10-fold. Details of this analysis will be published elsewhere.

## Funnels, local organization of sites

Several known bacterial and eukaryotic sites tend to cluster together. One may suggest that such clustering or other local arrangement of the sites can create a *funnel* in the binding energy landscape, which leads to a more rapid binding of cognate sites. Our model suggests that even if such funnels do exist, they would not significantly speed up the search process. The proposed search mechanism involves $\sim M/\bar{n}_{opt} \sim 10^4$ rounds of one-dimensional/three-dimensional diffusions. So a TF spends all the search time far from the cognate site. Only the last round (out of $10^4$) will be sped up by the funnel, leading to no significant decrease of the search time.

Local organization of sites and other sequence-dependent properties of the DNA structure (flexibility of AT-rich regions, DNA curvature on poly-A tracks, etc.) may influence preferred localization of TFs and lead to faster on-/off-binding

rates and fast equilibration on neighboring sites (see Slutsky et al., 2004, for details).

### Protein hopping: intersegment transfer

Our model assumed that rounds of one-dimensional diffusion are separated by periods of three-dimensional diffusion. Intersegment transfer is another mechanism that can be involved. If two segments of DNA come close to each other, a TF sliding along one segment can hop to another. The benefit of this mechanism is that it significantly shortens the transfer time, $\tau_{3D}$. Several examples of experimental evidence suggest that tetrameric *Lac*I, which has two DNA-binding sites, travels along DNA through one-dimensional diffusion and intersegment transfer.

We did not consider this mechanism because of the two following considerations. First, it is unclear whether TFs that have only one binding site can perform intersegment transfer; and second, for this mechanism to work, distant segments of DNA need to come close to each other. Although DNA packed into a cell/nuclear volume crosses itself every ~500 bp, DNA in solution, at in vitro concentrations, is unlikely to have any such self-crossings. Hence intersegment transfer cannot explain the faster-than-diffusion binding rates observed in vitro. This mechanism, however, may play a role in vivo, especially for proteins that have multiple DNA-binding sites.

### Proposed experiments

Our results propose several experimentally testable predictions.

First, we predict that the maximal rate of binding is achieved when the protein spends half of the time in solution and half sliding along the DNA. This result can be readily verified experimentally by measuring the concentration of free protein in solution that contains DNA but no cognate site. We also show how the search time depends on the energy of nonspecific binding, which, in turn, can be controlled by the ionic strength of solution or by engineering proteins with stronger or weaker nonspecific binding. In vivo observation of the 50/50 rule would suggest that proteins are optimized by evolution for rapid search.

Second, we show how the binding rate depends on the average travel time between two random segments of DNA, $\tau_{3d}$. This time measurement ($\tau_{3d}$) depends on the DNA concentration and the domain organization of DNA. By changing DNA concentration and/or DNA stretching in a single molecule experiment, one can alter $\tau_{3d}$ and thus study the role of DNA packing on the rate of binding. This effect has implications for DNA recognition in vivo, where DNA is organized into domains. Similarly, one can experimentally measure and compare the binding rate, in the presence of other DNA-binding proteins or nucleosomes, with analytical predictions.

Single molecule experiments and AFM/SFM imaging allow direct observation of protein trajectory and measurement of the one-dimensional diffusion coefficient, $D_{1d}$, on noncognate DNA. Our formalism, in turn, allows us to calculate the spectrum of specific binding energy, given $D_{1d}$. Such measurements can be direct tests of our conjecture that one-dimensional search along noncognate DNA proceeds along a smoother energy profile.

Third, using protein engineering one can stabilize unstructured regions of DNA-binding proteins (e.g., $\lambda$cI, *Eco*RV, and GCN4), and study the binding rates of these engineered, rigid proteins. Such experiments can test the proposed search-and-fold mechanism and shed light on the role of unstructured regions in determining stability, specificity, and binding rates.

We also suggest that proteins bound to noncognate DNA are not fully ordered. Unfortunately very few studies (Kalodimos et al., 2001, 2002, 2004) have addressed the mechanisms of binding to noncognate DNA. More studies of structures, thermodynamics, and dynamics of proteins bound to noncognate DNA will deepen our understanding of specific protein-DNA recognition.

## CONCLUSIONS

We have developed a quantitative model of protein-DNA interaction that provides an insight into the mechanism of fast target site location. We found the range of parameters (specific and nonspecific binding energies) that are crucial for fast search and, hence, the robust functioning of gene transcription. Paradoxically, realistic energy cannot provide both rapid searches and strong binding of a rigid protein. This allowed us to formulate the speed-stability paradox of protein-DNA recognition (which is similar to the famous Levinthal paradox of protein folding). To resolve the paradox, we proposed a search-and-fold mechanism that involves the coupling of protein binding and protein folding.

The proposed mechanism has several important biological implications in explaining how a protein can find its site on DNA, in vivo, in the presence of other proteins and nucleosomes and by a simultaneous search of several proteins. Our model provides, for the first time, a quantitative framework for analysis of the kinetics of transcription factor binding and, hence, gene expression. Importantly, our model links molecular properties of transcription factors to the timing of transcription activation. Proper understanding of the entire mechanism will hardly be possible without further experimental effort in these directions.

## APPENDIX A: DIFFUSIVE PROPERTIES OF THE DNA

The derivation consists of two steps. First, we describe the random walk along the DNA in terms of number of steps. Next, we calculate the mean time between successive steps in the random energetic landscape that provides the timescale for the problem. Such a decoupling, strictly speaking,

does not hold when the number of steps is small, i.e., when the number of visited sites is small and the random quantities are not averaged properly. However, since we are dealing with a large number of steps ($\sim 10^5$–$10^6$), this approach is legitimate—as is also confirmed by numerical simulations.

## The MFPT

To derive the diffusion law, we calculate the mean first passage time (MFPT) from site *0* to site *L*, defined as the mean number of steps the particle has to make to reach the site *L* for the first time. The derivation here follows that in Murthy and Kehr (1989).

Let $P_{\mathrm{i,j}}(n)$ denote the probability to start at site $i$ and reach the site $j$ in exactly $n$ steps. Then, for example,

$$P_{\mathrm{i,i+1}}(n) = p_{\mathrm{i}} T_{\mathrm{i}}(n-1), \tag{21}$$

where $T_{\mathrm{i}}(n)$ is defined as the probability of returning to the $i^{\mathrm{th}}$ site after $n$ steps without stepping to the right of it. Now, all the paths contributing to $T_{\mathrm{i}}(n–1)$ should start with the step to the left and then reach the site $i$ in $n–2$ steps, not necessarily for the first time. Thus, the probability $T_{\mathrm{i}}(n–1)$ can be written as

$$T_{\mathrm{i}}(n-1) = q_{\mathrm{i}} \sum_{\mathrm{m,l}} P_{\mathrm{i-1,i}}(m) T_{\mathrm{i}}(l) \delta_{\mathrm{m}+l,\mathrm{n}-2}. \tag{22}$$

We now introduce generating functions

$$\tilde{P}_{\mathrm{i,j}}(z) = \sum_{\mathrm{n}=0}^{\infty} z^{\mathrm{n}} P_{\mathrm{i,j}}(n), \quad \tilde{T}_{\mathrm{i}}(z) = \sum_{\mathrm{n}=0}^{\infty} z^{\mathrm{n}} T_{\mathrm{i}}(n). \tag{23}$$

One can easily show (see e.g., Goldhirsh and Gefen, 1986) that

$$\tilde{P}_{\mathrm{0,L}}(z) = \prod_{\mathrm{i}=0}^{\mathrm{L}-1} \tilde{P}_{\mathrm{i,i+1}}(z). \tag{24}$$

Knowing $\tilde{P}_{\mathrm{i,i+1}}(z)$, one calculates the MFPT straightforwardly as

$$\bar{t}_{\mathrm{0,L}} = \frac{\sum_{\mathrm{n}} n P_{\mathrm{0,L}}(n)}{\sum_{\mathrm{n}} P_{\mathrm{0,L}}(n)} = \left[\frac{d}{dz} \ln \tilde{P}_{\mathrm{0,L}}(z)\right]_{z=1} = \sum_{\mathrm{i}=0}^{\mathrm{L}-1} \left[\frac{d}{dz} \ln \tilde{P}_{\mathrm{i,i+1}}(z)\right]_{z=1}. \tag{25}$$

Using Eqs. 21 and 22, we obtain the recursion relation for $\tilde{P}_{i,i+1}(z)$,

$$\tilde{P}_{\mathrm{i,i+1}}(z) = \frac{z p_{\mathrm{i}}}{1 - z q_{\mathrm{i}} \tilde{P}_{\mathrm{i-1,i}}(z)}. \tag{26}$$

To solve for $\bar{t}_{\mathrm{0,L}}$, we must introduce boundary conditions. Let $p_0 = 1, q_0 = 0$, which is equivalent to introducing a reflecting wall at $i = 0$. This boundary condition clearly influences the solution for short times and distances. However, as numerical simulations and general considerations suggest, its influence relaxes quite fast, so that for longer times, the result is clearly independent of the boundary. The benefit of setting $p_0 = 1$ becomes clear when we observe that

$$\tilde{P}_{\mathrm{0,1}}(1) = 1 \quad \Rightarrow \quad \forall i \quad \tilde{P}_{\mathrm{i,i+1}}(1) = 1. \tag{27}$$

Hence,

$$\bar{t}_{\mathrm{0,L}} = \sum_{\mathrm{i}=0}^{\mathrm{L}-1} \tilde{P}'_{\mathrm{i,i+1}}(1). \tag{28}$$

The recursion relation for $P'_{\mathrm{i,i+1}}(1)$ is readily obtained from Eq. 26,

$$\tilde{P}'_{\mathrm{i,i+1}}(1) = \frac{1}{p_{\mathrm{i}}} + \frac{q_{\mathrm{i}}}{p_{\mathrm{i}}} \tilde{P}'_{\mathrm{i-1,i}}(1) = 1 + \alpha_{\mathrm{i}}[1 + \tilde{P}'_{\mathrm{i-1,i}}(1)], \tag{29}$$

with $\alpha_{\mathrm{i}} \equiv p_{\mathrm{i}}/q_{\mathrm{i}}$. Thus, the expression for $\bar{t}_{\mathrm{0,L}}$ is obtained in closed form as

$$\bar{t}_{\mathrm{0,L}} = L + \sum_{\mathrm{k}=0}^{\mathrm{L}-1} \alpha_{\mathrm{k}} + \sum_{\mathrm{k}=0}^{\mathrm{L}-2} \sum_{\mathrm{i}=\mathrm{k}+1}^{\mathrm{L}-1} (1 + \alpha_{\mathrm{k}}) \prod_{\mathrm{j}=\mathrm{k}+1}^{\mathrm{i}} \alpha_{\mathrm{j}}. \tag{30}$$

This solution expression gives MFPT in terms of a given realization of disorder producing a certain set of probabilities $\{p_{\mathrm{i}}\}$, wherein we are interested in the behavior averaged over all realizations of disorder. The cumulative products in Eq. 30 reduce to the two forms of $e^{\beta(\mathrm{U_i}-\mathrm{U_j})}$, which, after being averaged over uncorrelated Gaussian disorder, produces a factor of $e^{\beta^2\sigma^2}$. After the summations are carried out, the expression for MFPT becomes for $L \gg 1$,

$$\langle \bar{t}_{\mathrm{0,L}} \rangle \simeq L^2 e^{\beta^2\sigma^2}. \tag{31}$$

Thus, the diffusion law appears to be the classical one, with a renormalized diffusion coefficient.

## The time constant

Consider a particle at site $i$. The particle will eventually escape to one of the neighboring sites ($i \pm 1$), the escape rate being

$$r_{\mathrm{i}} = \omega_{\mathrm{i,i+1}} + \omega_{\mathrm{i,i-1}}. \tag{32}$$

To calculate the characteristic diffusion time constant $\langle \tau \rangle$, this rate should be averaged over all configurations of disorder $\{U_{\mathrm{i}}\}$. To obtain an analytic expression for the $\langle \tau \rangle$, we assume the form

$$\omega_{\mathrm{i,i\pm1}} = \nu e^{-\beta(\mathrm{U_{i\pm1}}-\mathrm{U_i})} \tag{33}$$

for both $U_{\mathrm{i}\pm1} > U_{\mathrm{i}}$ and $U_{\mathrm{i}\pm1} < U_{\mathrm{i}}$, as opposed to Eq. 7. Numerics show that this approximation introduces an up to $\sim$15% error for small values of $\beta\sigma$ and is practically exact for $\beta\sigma > 2$. Thus,

$$r_{\mathrm{i}} = \frac{1}{2\tau_0}(e^{-\beta(\mathrm{U_{i+1}}-\mathrm{U_i})} + e^{-\beta(\mathrm{U_{i-1}}-\mathrm{U_i})}), \tag{34}$$

where $\tau_0 = 1/(2\nu)$. The mean time between the successive steps can be calculated therefore as the average over all possible configurations of $U_{\mathrm{i}}$, $U_{\mathrm{i}\pm1}$ of the reciprocal of the escape rate, i.e.,

$$\langle \tau \rangle = \left\langle \frac{1}{r_{\mathrm{i}}} \right\rangle = 2\tau_0 \int_{-\infty}^{\infty} dU_{\mathrm{i}} dU_{\mathrm{i+1}} dU_{\mathrm{i-1}} \frac{f(U_{\mathrm{i}}) f(U_{\mathrm{i+1}}) f(U_{\mathrm{i-1}})}{e^{-\beta(\mathrm{U_{i+1}}-\mathrm{U_i})} + e^{-\beta(\mathrm{U_{i-1}}-\mathrm{U_i})}}. \tag{35}$$

Assuming Gaussian energy statistics as above, this integral is evaluated as

$$\langle \tau \rangle = \frac{\tau_0 e^{\beta^2\sigma^2/2}}{\pi} \int_{-\infty}^{\infty} dx dy \frac{e^{-(\mathrm{x}^2+\mathrm{y}^2)/2}}{e^{-\beta\sigma\mathrm{x}} + e^{-\beta\sigma\mathrm{y}}}. \tag{36}$$

After the change of variables

$$s = \frac{1}{\sqrt{2}}(x + y), \quad t = \frac{1}{\sqrt{2}}(x - y), \tag{37}$$

the integral factorizes leading to

$$\begin{aligned} \langle \tau \rangle &= \frac{\tau_0 e^{\beta^2\sigma^2/2}}{2\pi} \int_{-\infty}^{\infty} ds\, e^{-\mathrm{s}^2/2 + \beta\sigma\mathrm{s}/\sqrt{2}} \int_{-\infty}^{\infty} dt \frac{e^{-\mathrm{t}^2/2}}{\cosh(\beta\sigma t/\sqrt{2})} \\ &= \frac{\tau_0 e^{3\beta^2\sigma^2/4}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} dt\, e^{-\mathrm{t}^2/2 - \ln[\cosh(\beta\sigma\mathrm{t}/\sqrt{2})]} \\ &\simeq \frac{\tau_0 e^{3\beta^2\sigma^2/4}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} dt\, e^{-\mathrm{t}^2(1+\beta^2\sigma^2/2)/2} \\ &= \tau_0 e^{3\beta^2\sigma^2/4} [1 + \beta^2\sigma^2/2]^{-1/2}. \end{aligned} \tag{38}$$

Now, multiplying Eq. 31 by $\langle\tau\rangle$, we obtain the diffusion coefficient as

$$D_{1d}(\sigma) \simeq \frac{1}{2\tau_0}\left(1+\frac{\beta^2\sigma^2}{2}\right)^{1/2} e^{-7\beta^2\sigma^2/4}. \quad (39)$$

## APPENDIX B: NONSPECIFIC ENERGY

To find how the nonspecific energy $E_{ns}$ is related to the average time, $\tau_{1d}$, that a protein spends scanning a single region of the DNA, we use the simple observation that

$$\left\langle \sum_i \tau_i r_i \right\rangle = 1; \left\langle \sum_i \tau_i \right\rangle = \tau_{1d}, \quad (40)$$

which states that, eventually, protein dissociates from the region it scans with probability 1.

Since some massive hopping from site to site takes place before the particle eventually dissociates, the dissociation rates and, consequently, the nonspecific binding energy, should satisfy the equation

$$\begin{aligned}\left\langle \sum_i \tau_i r_i \right\rangle &= \frac{1}{\tau_0}\left\langle \sum_i \tau_i e^{-\beta(E_{ns}-U_i)} \right\rangle \\ &= \frac{1}{\tau_0}\int_{-\infty}^{\infty} e^{-\beta(E_{ns}-U)}\tau(U)f(U)dU = 1,\end{aligned} \quad (41)$$

and this is subject to a condition

$$\left\langle \sum_i \tau_i \right\rangle = \int_{-\infty}^{\infty} \tau(U)f(U)dU = \tau_{1d}, \quad (42)$$

where $\tau_i$ is the time that the TF spends at the $i^{th}$ site and $\tau_{1d}$ is the average time of a one-dimensional search to dissociation. The average lifetime $\tau_i = t(U_i)$ at that site is proportional to $\exp(-\beta U_i)$. In this specific case, the particle usually escapes to one of the neighboring sites, and we should average over their energies. Hence, the explicit form $\tau(U)$ as calculated from Eq. 42 is

$$\tau(U) = \tau_{1d} e^{-\beta^2\sigma^2/2} e^{-\beta U}. \quad (43)$$

Substituting this into Eq. 41, we have

$$\frac{\tau_{1d}}{\tau_0} e^{-\frac{1}{2}\beta^2\sigma^2-\beta E_{ns}} = 1, \quad (44)$$

or

$$E_{ns} = k_B T\left[\ln\left(\frac{\tau_{1d}}{\tau_0}\right) - \frac{1}{2}\left(\frac{\sigma}{k_B T}\right)^2\right]. \quad (45)$$

Next we recall that, in the optimal regime, $\tau_{1d} = \bar{\tau}_{3d}$. Thus, to ensure optimal performance, $E_{ns}$ should be equal to the expression in Eq. 45 with $\tau_{1d}$ replaced by $\bar{\tau}_{3d}$,

$$E_{ns} = k_B T\left[\ln\left(\frac{\tau_{3d}}{\tau_0}\right) - \frac{1}{2}\left(\frac{\sigma}{k_B T}\right)^2\right]. \quad (46)$$

The meaning of this relation is quite transparent. The logarithm gives $E_{ns}$ in a system with zero or constant specific binding energy. The second term introduces suppression of $E_{ns}$ due to disorder, so that the dissociation events in a system with disorder are more frequent, to compensate partially for the one-dimensional diffusion slowdown. This relation obviously holds as long as $E_{ns} > 0$. Negative values of $E_{ns}$ mean simply that the nonspecific interaction became overshadowed by the specific one, and no longer has any direct physical sense.

Since for a given value of $\sigma$, the nonspecific binding controls the dissociation rate, the search time will deviate from the optimum if $E_{ns}$ moves from this predetermined value. In Fig. 3 *a* we plot the search time as a function of the nonspecific binding energy for different values of $\sigma$.

We now define the tolerance factor, $\zeta$, as the ratio between the maximal acceptable value of the search time $t_s$ and the minimal time $t_{s0}$. Experimental data suggest $\zeta \leq 5$, but we for the moment allow for much larger values of $\zeta \sim$ 10–100 (this can be done when, for instance, there are many protein molecules searching in parallel). As we can see from Fig. 3 *a*, for each value of $\sigma$, there is a range of possible values of $E_{ns}$ such that the resulting search time is within the region of tolerance. This range is easily calculated producing the values of nonspecific energy between

$$E_{ns}^{\pm}(\sigma,\zeta) = \frac{2}{\beta}\ln\left[\sqrt{\frac{D_{1d}(\sigma)\bar{\tau}_{3d}}{D_{1d}(0)\tau_0}}\left(\zeta \pm \sqrt{\zeta^2 - \frac{D_{1d}(0)}{D_{1d}(\sigma)}}\right)\right] - \frac{\sigma^2\beta}{2}. \quad (47)$$

## APPENDIX C: ROLE OF DNA CONFORMATION

The central parameter here is $\tau_{3d}$, the interval of time between a dissociation of the protein from DNA until the next binding to DNA. Exact calculation of $\tau_{3d}$ is a very difficult task, considering the nontrivial packaging of the DNA molecule inside a bacterial cell, electrostatic effects, and the inhomogeneity of the cytoplasm.

Considering the microscopic picture one can easily obtain a reasonable estimate of $\tau_{3d}$ as a characteristic time of three-dimensional diffusion across the nucleoid (the region of a bacterial cell to which the DNA is confined). The corresponding diffusion length depends on the conformation of the DNA molecule. Indeed, if the DNA molecule was a single homogeneous globule, there would be a single relevant length scale, which is the molecule characteristic size $l_m$ (the gyration radius). On the other hand, as Fig. 7 shows, diffusion of a protein molecule inside a more realistic non-homogeneous multidomain molecule involves at least one additional length scale $l_d$, which is a characteristic size of a domain. These two lengths may differ by a factor of $\sim$10 (Neidhardt et al., 1996), making the ratio of the resulting diffusion times $\tau_{3d}^m/\tau_{3d}^d \sim 10^2$. In the original problem (a single protein molecule searching for a single site on the DNA), the search process is dominated by the larger timescale, since at least a few domains must be explored before the target site is located. However, there are usually $\sim 10^2$ TF molecules present in a cell, so it is reasonable to assume that the domains are scanned in parallel, making the interdomain transfer processes irrelevant.

## APPENDIX D: STABILITY REQUIREMENT

In fact, it is not hard to estimate analytically the $(\sigma/k_B T)$ ratio for a genome of length $M$ such that the probability of binding to the lowest site is comparable to the probability of binding to the rest of the genome; i.e., their contributions to the partition function are of the same order of magnitude. The partition sum for the Gaussian energy level statistics is

$$\begin{aligned}\Omega &= \frac{M}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty} e^{-\beta U - U^2/(2\sigma^2)}dU \\ &= Me^{\beta^2\sigma^2/2} \sim \exp[-\beta U_{min}] \sim \exp(\beta\sigma\sqrt{2\ln M}),\end{aligned} \quad (48)$$

so that for $M = 10^6$

$$\sigma \sim k_B T\sqrt{2\ln M} \sim 5\, k_B T. \quad (49)$$

Strictly speaking, for a large, albeit finite set of energy levels, the integration limits are cut off at $\pm\sigma\sqrt{2\ln M}$ so that for $\beta\sigma \gg \sqrt{\ln M}$ the partition function is dominated by the lower edge of the distribution. The estimate for

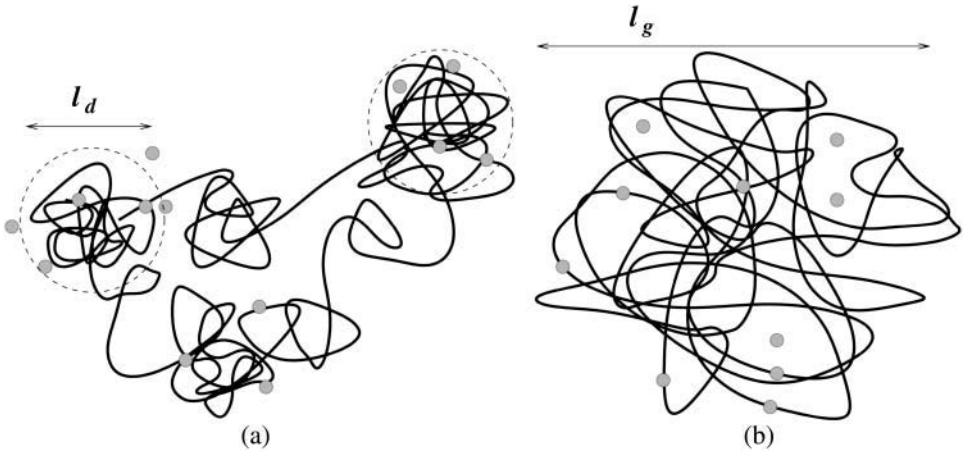


FIGURE 7 Effect of DNA conformation on the effective diffusion distance: (*a*) single globule; (*b*) multidomain conformation.

$\beta\sigma$ gives, therefore, the crossover value between the regime of multiple-site contribution to $\Omega$ and the regime with single-site domination (the analog of this effect would be *thermodynamical freezing*; see Derrida, 1981).

If $N_p$ proteins are searching and binding a single target site, then the probability of being occupied is given by

$$P(N_p) = 1 - (1 - P_b)^{N_p} \approx N_p P_b, \tag{50}$$

where $P_b$ is the probability of the site being occupied by a single protein (see Eq. 19) and the approximation is for $P_b \ll 1/N_p$. As evident from Fig. 4 *b*, requirement of the rapid search is satisfied if $P_b(\sigma/T \approx 1) \sim 10^{-5}$. An unfeasible amount of $10^4$ copies of a single TF is required to saturate such a weak binding site.

## APPENDIX E: ENERGY GAP

Large energy gap between the cognate site $\vec{s}_c$ and the bulk of genomic sites would solve the paradox of rapid search and stability. One may seek parameters, $\epsilon(j, s)$, of the energy function

$$U(\vec{s} = s_i, ..s_{i+l-1}) = \sum_{j=1}^{l} \epsilon(j, s_j) \tag{51}$$

to maximize the energy gap by minimizing the *Z*-score

$$Z(\vec{s}_c) = \frac{U(\vec{s}_c) - \langle U \rangle}{\sigma}, \tag{52}$$

where averaging and variance is taken over all possible sequences of length $l$ (or over genomic words of length $l$). It is easy to see that $Z(\vec{s}_c)$ is minimal if

$$\epsilon^{opt}(j, s) = -\delta(s, s_{cj}), \tag{53}$$

where $\delta(x,y)$ is the Kronecker delta. For $K$ types of nucleotides assuming their equal frequency in genome we obtain the maximal reachable energy gap of

$$Z^{min} = -\sqrt{lK}. \tag{54}$$

For $K = 4$ and $l \approx 8$ we get $Z^{min} \approx -5$. For the genome of $10^6$–$10^7$ bp, the energy spectrum of the genomic DNA ends at $Z \approx -5$. Although sufficient to provide stability of the bound complex (see main text), such an energy gap is unable to resolve the search-stability paradox.

## APPENDIX F: DIFFUSION IN WATER AND IN CYTOPLASM

The diffusion coefficient of a protein molecule in water can be estimated as in Landau and Lifshitz (1987),

$$D \simeq \frac{k_B T}{3\pi\eta d}, \tag{55}$$

where $d$ is the diameter of the molecule and $\eta$ is the water viscosity. Setting $\eta \sim 10^{-2}$ g/(s × cm) and $d \sim 10$ nm, we obtain at room temperature

$$D \sim 10^2 \mu m^2/s. \tag{56}$$

Diffusion coefficient measurements for GFP in *E. coli* (Elowitz et al., 1999) produce values of ~1–10 $\mu m^2$/s. This difference in diffusion coefficients may account for more than an order-of-magnitude difference in the theoretically calculated and measured target location times.

We thank A. Finkelstein, M. Kardar, W. Bialek, and A. van Oijen for useful discussions.

L.M. is an Alfred P. Sloan Research Fellow.

## REFERENCES

Abkevich, V., A. Gutin, and E. Shakhnovich. 1994. Specific nucleus as the transition state for protein folding: evidence from the lattice model. *Biochemistry*. 33:10026–10036.

Akke, M. 2002. NMR methods for characterizing microsecond to millisecond dynamics in recognition and catalysis. *Curr. Opin. Struct. Biol.* 12:642–647.

Bell, C. E., and M. Lewis. 2000. A closer view of the conformation of the *Lac* repressor bound to operator. *Nat. Struct. Biol.* 7:209–214.

Bell, C. E., and M. Lewis. 2001. The *Lac* repressor: a second generation of structural and functional studies. *Curr. Opin. Struct. Biol.* 11:19–25.

Berg, O. G., and P. H. von Hippel. 1987. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.* 193:723–750.

Berg, O. G., R. B. Winter, and P. H. von Hippel. 1981. Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. Models and theory. *Biochemistry*. 20:6929–6948.

Bouchaud, J. P., and A. Georges. 1990. Anomalous diffusion in disordered media: statistical mechanisms, models and physical applications. *Phys. Rep.* 195:172–293.

Bruinsma, R. F. 2002. Physics of protein-DNA interaction. *Phys. A*. 313:211–237.

Clarke, N., L. Beamer, H. Goldberg, C. Berkower, and C. Pabo. 1991. The DNA binding arm of λ-repressor: critical contacts from a flexible region. *Science*. 254:267–270.

Derrida, B. 1981. Random-energy model: an exactly solvable model of disordered systems. *Phys. Rev. B.* 24:2613–2626.

Dyson, H. J., and P. E. Wright. 2002. Coupling of folding and binding for unstructured proteins. *Curr. Opin. Struct. Biol.* 12:54–60.

Elowitz, M. B., M. G. Surette, P. E. Wolf, J. B. Stock, and S. Leibler. 1999. Protein mobility in the cytoplasm of *Escherichia coli*. *J. Bacteriol.* 181: 197–203.

Erie, D., G. Yang, H. Schultz, and C. Bustamante. 1994. DNA bending by Cro protein in specific and nonspecific complexes: implications for protein site recognition and specificity. *Science.* 266:1562–1566.

Finkelstein, A., and O. Ptitsyn. 2002. Protein Physics. Academic Press, New York.

Flyvbjerg, H., F. Jülicher, P. Ormos, and F. David. (Editors.). 2002. Physics of biomolecules and cells. *In* Les Houches, Vol. 75, Chap. 1. Springer-Verlag, Heidelberg, Germany.

Gerland, U., J. D. Moroz, and T. Hwa. 2002. Physical constraints and functional characteristics of transcription factor-DNA interaction. *Proc. Natl. Acad. Sci. USA.* 99:12015–12020.

Goldhirsh, I., and Y. Gefen. 1986. Analytic method for calculating properties of random walks on networks. *Phys. Rev. A.* 33:2583–2594.

Grillo, A. O., M. P. Brown, and C. A. Royer. 1999. Probing the physical basis for Trp repressor-operator recognition. *J. Mol. Biol.* 287:539–554.

Gutin, A., A. Sali, V. Abkevich, M. Karplus, and E. Shakhnovich. 1998. Temperature dependence of the folding rate in a simple protein model: search for a glass transition. *J. Chem. Phys.* 108:6466–6483.

Hopfield, J. J. 1974. Kinetic proofreading: a new mechanism for reducing errors in biosynthetic processes requiring high specificity. *Proc. Natl. Acad. Sci. USA.* 71:4135–4139.

Hughes, B. D. 1995. Random Walks and Random Environments. Clarendon Press, New York.

Kalodimos, C., N. Biris, A. Bonvin, M. Levandoski, M. Guennuegues, R. Boelens, and R. Kaptein. 2004. Structure and flexibility adaptation in nonspecific and specific protein-DNA complexes. *Science.* 305:386–389.

Kalodimos, C. G., R. Boelens, and R. Kaptein. 2002. A residue-specific view of the association and dissociation pathway in protein-DNA recognition. *Nat. Struct. Biol.* 9:193–197.

Kalodimos, C. G., G. E. Folkers, R. Boelens, and R. Kaptein. 2001. Strong DNA binding by covalently linked dimeric *Lac* headpiece: evidence for the crucial role of the hinge helices. *Proc. Natl. Acad. Sci. USA.* 98: 6039–6044.

Kim, J. G., Y. Takeda, B. W. Matthews, and W. F. Anderson. 1987. Kinetic studies on Cro repressor-operator DNA interaction. *J. Mol. Biol.* 196: 149–158.

Landau, L., and E. Lifshitz. 1987. Fluid Mechanics. Butterworth-Heinemann, Oxford, Boston.

Levy, Y., P. Wolynes, and J. Onuchic. 2004. Protein topology determines binding mechanism. *Proc. Natl. Acad. Sci. USA.* 101:511–516.

Lewis, M., G. Chang, N. C. Horton, M. A. Kercher, H. C. Pace, M. A. Schumacher, R. G. Brennan, and P. Lu. 1996. Crystal structure of the lactose operon repressor and its complexes with DNA and inducer. *Science.* 271:1247–1254.

Lomakin, A., and M. Frank-Kamenetskii. 1998. A theoretical analysis of specificity of nucleic acid interactions with oligonucleotides and peptide nucleic acids (PNAs). *J. Mol. Biol.* 276:57–70.

Luscombe, N. M., S. E. Austin, H. M. Berman, and J. M. Thornton. 2000. An overview of the structures of protein-DNA complexes. *Genome Biol.* 1:1–37.

Mirny, L., and E. Shakhnovich. 2001. Protein folding theory: from lattice to all-atom models. *Annu. Rev. Biophys. Biomol. Struct.* 30:361–396.

Murthy, K. P. N., and K. W. Kehr. 1989. Mean first-passage time of random walks on a random lattice. *Phys. Rev. A.* 40:2082–2087.

Neidhardt, F. C., R. Curtiss, and E. C. Lin. (Editors.). 1996. *Escherichia coli* and *Salmonella*, Chap. 12. ASM Press, Washington, DC.

O'Neil, K., R. Hoess, and W. DeGrado. 1990. Design of DNA-binding peptides based on the leucine zipper motif. *Science.* 249:774–778.

Pande, V., A. Grosberg, and T. Tanaka. 2000. Heteropolymer freezing and design: towards physical models of protein folding. *Rev. Mod. Phys.* 72:259–314.

Richter, P. H., and M. Eigen. 1974. Diffusion controlled reaction rates in spheroidal geometry. Application to repressor-operator association and membrane bound enzymes. *Biophys. Chem.* 2:255–263.

Riggs, A. D., S. Bourgeois, and M. Cohn. 1970a. The *Lac* repressor-operator interaction. 3. Kinetic studies. *J. Mol. Biol.* 53:401–417.

Riggs, A. D., H. Suzuki, and S. Bourgeois. 1970b. *Lac* repressor-operator interaction. 1. Equilibrium studies. *J. Mol. Biol.* 48:67–83.

Schumacher, M. A., K. Y. Choi, H. Zalkin, and R. G. Brennan. 1994. Crystal structure of *Lac*I member, *PurR*, bound to DNA: minor groove binding by $\alpha$-helices. *Science.* 266:763–770.

Shimamoto, N. 1999. One-dimensional diffusion of proteins along DNA. Its biological and chemical significance revealed by single-molecule measurements. *J. Biol. Chem.* 274:15293–15296.

Shoemaker, B., J. Portman, and P. Wolynes. 2000. Speeding molecular recognition by using the folding funnel: the fly-casting mechanism. *Proc. Natl. Acad. Sci. USA.* 97:8868–8873.

Slutsky, M., M. Kardar, and L. Mirny. 2004. Diffusion in correlated random potentials, with applications to DNA. *Phys. Rev. E.* 69:061903–061915.

Spolar, R. S., and M. T. Record. 1994. Coupling of local folding to site-specific binding of proteins to DNA. *Science.* 263:777–784.

Stormo, G. D., and D. S. Fields. 1998. Specificity, free energy and information content in protein-DNA interactions. *Trends Biochem. Sci.* 23:109–113.

Takeda, Y., A. Sarai, and V. M. Rivera. 1989. Analysis of the sequence-specific interactions between Cro repressor and operator DNA by systematic base substitution experiments. *Proc. Natl. Acad. Sci. USA.* 86:439–443.

Uversky, V. N. 2002. Natively unfolded proteins: a point where biology waits for physics. *Protein Sci.* 11:739–756.

von Hippel, P. H., and O. G. Berg. 1989. Facilitated target location in biological systems. *J. Biol. Chem.* 264:675–678.

Williams, P., D. Pollock, and R. Goldstein. 2001. Evolution of functionality in lattice proteins. *J. Mol. Graph. Model.* 19:150–156.

Winkler, F., D. Banner, C. Oefner, D. Tsernoglou, R. Brown, S. Heathman, R. Bryan, P. Martin, K. Petratos, and K. Wilson. 1993. The crystal structure of *Eco*RV endonuclease and of its complexes with cognate and non-cognate DNA fragments. *EMBO J.* 12:1781–1795.

Winter, R. B., O. G. Berg, and P. H. von Hippel. 1989. Diffusion-driven mechanisms of protein translocation on nucleic acids. 3. The *Escherichia coli Lac* repressor-operator interaction: kinetic measurements and conclusions. *Biochemistry.* 20:6961–6977.

Wright, P. E., and H. J. Dyson. 1999. Intrinsically unstructured proteins: reassessing the protein structure-function paradigm. *J. Mol. Biol.* 293: 321–331.