Protein−DNA Interactions: Reaching and Recognizing the Targets

A. G. Cherstvy,*,†,|| A. B. Kolomeisky,*,‡ and A. A. Kornyshev*,§

Max-Planck-Institut für Physik komplexer Systeme, Nöthnitzer Strasse 38, D-01187 Dresden, Germany, Department of Chemistry, Rice University, Houston, Texas 77005, Department of Chemistry, Faculty of Natural Sciences, Imperial College London, SW7 2AZ, London, U.K., and Institut für Festkörperforschung, Theorie-II, Forschungszentrum Jülich, D-52425 Jülich, Germany

Received: August 10, 2007; In Final Form: December 21, 2007

Protein searching and recognizing the targets on DNA was the subject of many experimental and theoretical studies. It is often argued that some proteins are capable of finding their targets 10−100 times faster than predicted by the three-dimensional diffusion rate. However, recent single-molecule experiments showed that the diffusion constants of the *protein motion along DNA* are usually small. This controversy pushed us to revisit this problem. We present a theoretical approach that describes some physical−chemical aspects of the target search and recognition. We consider the search process as a sequence of cycles, with each cycle consisting of three-dimensional and one-dimensional tracks. It is argued that the search time contains three terms: for the motion on three-dimensional and one-dimensional segments, and the correlation term. Our analysis shows that the acceleration in the search time is achieved at some intermediate strength of the protein−DNA binding energy and it is partially "apparent" because it is in fact reached by parallel scanning for the target by many proteins. We also show how the complementarity of the charge patterns on a target DNA sequence and on the protein may result in electrostatic recognition of a specific track on DNA and subsequent protein pinning. Within the scope of a model, we obtain an analytical expression for the capturing well. We estimate the depth and width of such a potential well and the typical time that a protein spends in it.

I. Introduction: Facilitated Diffusion

There are many proteins that regulate the activity of DNA, for example, repressor proteins, DNA polymerases, DNA helicases, and endonucleases/restrictases.[1] They have different functions, but most of them have to reach and recognize their targets—distinct short sequences or defects on DNA molecules—a moment before starting their "job". When recognition refers to a precise match with two or three DNA base pairs, it is often called site-specific. If it extends to a somewhat longer DNA sequences, then it is called sequence-specific. Despite multiple experimental and theoretical efforts, how exactly proteins recognize the target places on DNA still remains, in many cases, a puzzle.[2,3]

Reaching the target and recognizing it are two sides of the process. The key questions here are: (i) How fast can a protein reach a given target on DNA? (ii) What exactly causes it to stop at the target? (iii) Once captured, will the protein residence time be long enough for the protein to perform its function? In this paper we are not going to describe the protein performance after reaching the target, but we will try to answer these three questions.

It has been realized that some DNA-binding proteins, for example, *lac* repressor, can find the corresponding targets on DNA much faster than allowed by ordinary three-dimensional diffusion. This phenomenon is called *facilitated diffusion*, and it has attracted the attention of many investigators.[4−12] Typically, these proteins possess high sequence specificity in their interactions with DNA. A current understanding of this phenomenon, supported by some experimental observations, is the following. The search process is a combination of three-dimensional excursions of proteins in solution and their one-dimensional sliding on DNA.[13]

A number of theoretical studies on the subject of facilitated protein diffusion on DNA have been performed in recent years. Several concepts that can allow decrease of the target search time have been implemented. In particular, some models of combined 3D diffusion in solution and 1D sliding along the DNA have been developed;[3,7,49,50] the effects of intersegmental protein transfer on a coiled DNA molecule have been studied and can make the process of DNA sampling by proteins more efficient;[51] the model of attractive "antenna" around the DNA target site has been utilized,[5] and the protein−DNA interaction energy landscape on 1D protein sliding has been studied in refs 7, 8, 31, and 52. Some effects of electrostatic interactions on protein−DNA binding affinity have been considered in particular in ref 53. Computer simulations of facilitated protein diffusion on DNA have been performed in refs 10, 54, and 55. Still, some aspects of this phenomenon require a more detailed theoretical consideration. In particular, (i) how the protein search process takes place out of equilibrium, (ii) what "facilitated diffusion" means in the view of recent experimental findings that 1D diffusion on DNA tracks appeared to be orders (!) of magnitude slower than in the solution bulk, and (iii) how strongly the complementarity of DNA and protein charge lattices can contribute to their electrostatic recognition. It is often claimed that the acceleration of the search process is achieved by reducing the dimensionality for some parts of the searching pathways.[4,3] This picture implicitly assumes that the diffusion constants for 3D (denoted as D_3) and 1D (D_1) motion are of the same order of magnitude, or at least not too different.

However, recent single-molecule experiments,[14,15] as well as old bulk biochemical studies,[16] suggest that one-dimensional protein transport along the DNA is, in fact, much slower (more

* Corresponding author.
† Max-Planck-Institut für Physik komplexer Systeme.
‡ Rice University. E-mail: tolya@rice.edu.
§ Imperial College London. E-mail: a.kornyshev@imperial.ac.uk.
|| Institut für Festkörperforschung. E-mail: a.cherstvy@fz-juelich.de.

10.1021/jp076432e CCC: $40.75
Published on Web 03/22/2008

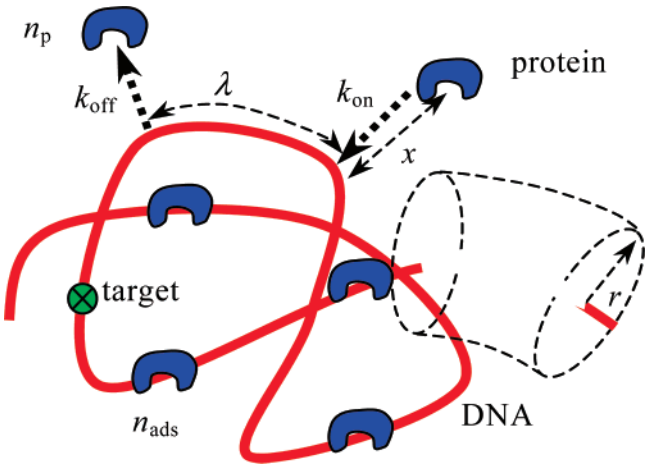

Figure 1. Proteins adsorption and desorption onto and from DNA and protein transport on DNA: a sketch defining the model parameters. n_p is the number of proteins in the bulk; n_{ads} is the number of proteins on DNA; k_{off} and k_{on} are the protein desorption and adsorption rate constants; x is the average distance of a protein in solution from DNA; and λ is the average length that a protein passes in one run on DNA, called the "sliding length". Distance r is an effective radius of DNA "tube"; it is chosen so that the tube then fills the whole space occupied by the DNA coil.

than 1000 times!) than 3D diffusion in the bulk of the solution. Thus, the motion along the DNA *per se* is not what accelerates the optimized search. In addition, in the currently reported theoretical models the rate of association of proteins to the target sequences was found to increase with *decreasing* concentration of free proteins or targets.[3] It is thus obvious that at least in this limit some of the previous approaches break down.

This forces us to revisit the physical and chemical aspects of the main stages of facilitated diffusion. The goal of the first part of the paper is to develop a qualitative picture of the search and detection of targets on DNA. We treat nonequilibrium large-scale properties of the search process, calculate the mean time of reaching the target, and consider the effects of the strength of protein adsorption onto the DNA. In the second part, we present a model of protein−DNA primary recognition based on the complementarity of their charge patterns. We introduce some degree of nonuniformity along the DNA and calculate the forces that bind or model proteins near it. We calculate the shape of the potential well near the DNA−protein complementary region and the mean time of protein escape from this well. We also present the distribution of electrostatic potential for some DNA protein complexes supporting the idea of DNA−protein charge recognition.

Nothing in the derived formulas should be taken literally because we used very simplified models for very complex problems. But this was the only way to reach a certain level of generality in our consideration. So, the qualitative conclusions are what the readers are invited to address their attention to in the first round.

II. Time for Reaching the Target: Diffusion Model

A. Model. Similar to the previous works on protein diffusion on DNA, we consider the process of reaching the target on DNA as a sequence of searching events. On average, in our model each protein binds and unbinds to DNA several times before finding the target. Binding to nontarget segments of DNA is called nonspecific; the average adsorption energy here is smaller than that on the target segments because mostly noncovalent interactions are responsible for the nonspecific binding. For many proteins, electrostatic interactions provide a large contribution to their nonspecific binding affinity to DNA. Each cycle consists of 3D and 1D tracks, explored by protein with different velocities.

Consider a DNA molecule with an average distance between the targets L ($\sim$1 μm), or one target per molecule. The mean first-passage time for any protein molecule to reach a target of size a (typically 3−6 DNA base pares or 1−2 nm) can be calculated as follows. The protein molecule is assumed to move through 3D space some average distance x (the length of a free path of a protein to DNA in solution). It binds to DNA at a random position and then moves along it some average distance λ, the sliding length. The protein scans on average a section of the length λ on DNA during this searching event, see Figure 1.

B. Basic Equations. The mean first-passage time for one searching cycle of a particular protein can be calculated, assuming that the segment of 3D diffusion is considered as effective 1D diffusion with a properly rescaled diffusion constant. The result reads[17]

$$\tau_c = \int_0^{x+\lambda} \frac{\exp[\beta G(z)]}{D(z)}\,dz \int_0^z \exp[-\beta G(z')]dz' \qquad (1)$$

Here $\beta = 1/k_BT$, $G(z)$ is the free energy of the protein at the position z of its path, and $D(z)$ is a position-dependent diffusion constant

$$D(z) = \begin{cases} D_3, 0 < z < x \\ D_1, x < z < x + \lambda \end{cases} \qquad (2)$$

where D_3 and D_1 are 3D and 1D protein diffusion constants, respectively.

We assume that the energy of nonspecific binding to DNA is E_{ads}, and if k_{on} and k_{off} are defined as the rate constants for protein binding and unbinding, respectively, then

$$y \equiv \frac{k_{on}}{k_{off}} = \exp\left(\frac{E_{ads}}{k_BT}\right) \qquad (3)$$

The parameter y plays the role of the adsorption equilibrium constant. At the same time, if the concentration of free proteins in solution is c_p and the effective volume concentration of proteins adsorbed on DNA is c_{ads}, then the difference between free energies of the protein in solution, and in the adsorbed state, E_{eff}, is given by

$$y_{eff} = \frac{k_{on}c_p}{k_{off}\,c_{ads}} = \exp\left(\frac{E_{eff}}{k_BT}\right) \qquad (4)$$

It is important to note that hereafter we will *not* assume equilibrium between association and dissociation processes because the majority of biological processes are generally out of equilibrium.

The free energy profile along the searching trajectory can be written in the following form

$$G(z) = \begin{cases} 0, 0 < z < x \\ -E_{eff}, x < z < x + \lambda \end{cases} \qquad (5)$$

Substituting this expression into eq 1, we obtain

$$\tau_c = \frac{x^2}{2D_3} + \frac{\lambda^2}{2D_1} + \frac{x\lambda}{D_1 y_{eff}} \qquad (6)$$

The terms in this formula can be understood in the following way. The first two terms correspond to the time spent by the protein on 3D and on 1D segments, respectively. The last term

is the correlation term responsible for contributions of trajectories when the protein went from 3D to 1D but unbinds from DNA before it travels the whole length λ. This contribution partially accounts for fluctuations in the length of the 3D and 1D segments (recall that x and λ are parameters averaged over many searching trajectories). Note that this last term was not present in the previous theoretical treatments; however, as will be shown below, it plays an important role in the dynamics of reaching the target.

In order to find the target the protein, on average, should scan the length L/n_{ads}. The number of adsorbed proteins in the denominator, n_{ads}, appears here because the protein is not alone on the contour length L: the average distance between the proteins is L/n_{ads}, and this is the length that each protein should scan. This is so because if not "this" protein, then another one will find the target. We use here an approximation of a low concentration of proteins on DNA, $n_{ads} \ll (L/\lambda)$, which implies a negligible probability of overlap of trajectories of individual proteins sliding on DNA.

Generally, the total mean time to find the target is given by the following expression

$$\tau = \left(\frac{L}{\lambda\, n_{ads}}\right)^{1/\alpha} \tau_c \tag{7}$$

The exponent $\alpha > 0$ reflects the nature of the scanning mechanism. If we assume that after desorption from DNA a protein can rebind with equal probability to *any* point on the length L (which is a reasonable assumption for realistic situations of $D_3 \gg D_1$), as shown in ref 4, then, simply, $\alpha = 1$. Exponents $\alpha > 1$ correspond to superdiffusion, while $\alpha < 1$ correspond to the subdiffusive regime. Because various modes of diffusion are possible in these complicated substrates, for generality it would be better to keep α as a parameter of the model. To get a more practical expression for τ_c in eq 6 and thereby τ, we will need to express x, y_{eff}, and λ through observable quantities.

Similar to refs 3 and 5, we view the DNA molecule as a coil with a contour length L per target, see Figure 1. The volume of such coil per one target is given by

$$V \sim Lr^2 \tag{8}$$

where r is $^1/_2$ of the average distance between the neighboring branches of DNA (an effective DNA radius), see Figure 1. This parameter is responsible in the model for the effect of DNA conformations and 3D structure on the protein diffusion. The concentrations of free and adsorbed proteins can be written as

$$c_p = \frac{n_p}{V}, \qquad c_{ads} = \frac{n_{ads}}{V} \tag{9}$$

Here, n_p is the number of free proteins in the volume V, and n_{ads} is the number of adsorbed proteins on the length L. Both n_p and n_{ads} may be non-integer, and can be even smaller than 1, but typically $n_{ads} \gg 1$.

The scaling relationship coming from the fact that the volume per one free protein molecule in solution can be written as

$$n_p L x^2 = Lr^2 \tag{10}$$

and it gives us

$$x = \frac{r}{\sqrt{n_p}} \tag{11}$$

Next, we recall that by definitions, eqs 3 and 4

$$y_{eff} = y\frac{n_p}{n_{ads}} \tag{12}$$

Finally, let us get an expression for λ. The flux of protein molecules binding to DNA and unbinding from DNA is proportional to, respectively

$$k_{on}c_p = \frac{1}{\tau_{free}} = \frac{2D_3}{x^2}$$

$$k_{off}c_{ads} = \frac{1}{\tau_{ads}} = \frac{2D_1}{\lambda^2} \tag{13}$$

Here, τ_{free} and τ_{ads} are the mean times for a protein molecule to be found in the solution and in the adsorbed state, correspondingly. Combining eqs 4 and 10−13, one obtains

$$y\frac{n_p}{n_{ads}} = \frac{n_p\lambda^2}{r^2 d} \tag{14}$$

where we have introduced a dimensionless ratio of the diffusion constants, $d = D_1/D_3$. Hence the optimal value of the sliding length is

$$\lambda = \frac{r\sqrt{yd}}{\sqrt{n_{ads}}} \tag{15}$$

Substituting expressions 11 and 12 into eqs 6 and 7, it can be shown that

$$\tau = \frac{Lr}{2D_3 n_p}\left(\frac{r}{\lambda}\frac{1}{n_{ads}} + \frac{\lambda}{r}\frac{n_p}{n_{ads}}\frac{1}{d} + \frac{2}{\sqrt{n_p}yd}\right)\left(\frac{L}{\lambda\, n_{ads}}\right)^{(1/\alpha)-1} \tag{16}$$

This time should be compared with the time for the purely 3D search for the target of size a, given by the Smoluchowski theory.[18] To be consistent, we again consider 3D diffusion as an effective 1D process with a corresponding diffusion constant

$$\tau_S = \frac{1}{2D_3 a c_p} = \frac{Lr^2}{2D_3 a n_p} \tag{17}$$

Hence, the *relative* search time is given by

$$\frac{\tau}{\tau_S} = \left(\frac{a}{\lambda}\frac{1}{n_{ads}} + \frac{a\lambda}{r^2}\frac{n_p}{n_{ads}d} + \frac{a}{r}\frac{2}{yd\sqrt{n_p}}\right)\left(\frac{L}{\lambda\, n_{ads}}\right)^{(1/\alpha)-1} \tag{18}$$

If we use, finally, the explicit expression for λ, eq 15, then we get

$$\frac{\tau}{\tau_S} = \frac{a}{r}\left(\frac{1}{\sqrt{n_{ads}yd}} + \frac{n_p\sqrt{y}}{n_{ads}^{3/2}\sqrt{d}} + \frac{2}{\sqrt{n_p}yd}\right)\left[\frac{L}{r\sqrt{n_{ads}yd}}\right]^{(1/\alpha)-1} \tag{19}$$

Let us point out again that, according to recent single-molecule experiments,[14] the value of d, in contrast to the earlier conjectures,[3,4] can be very small, for example, $\sim 10^{-3}$. Can we still expect any acceleration of the search time? In fact, in spite of the smallness of d, we can get $\tau/\tau_S < 1$ for several reasons. First, a/r can be as small as 10^{-2} for relevant DNA lengths and densities[3]. Next, a large number of n_{ads} can also help. The role of adsorption equilibrium constant y is more complex, as discussed below.

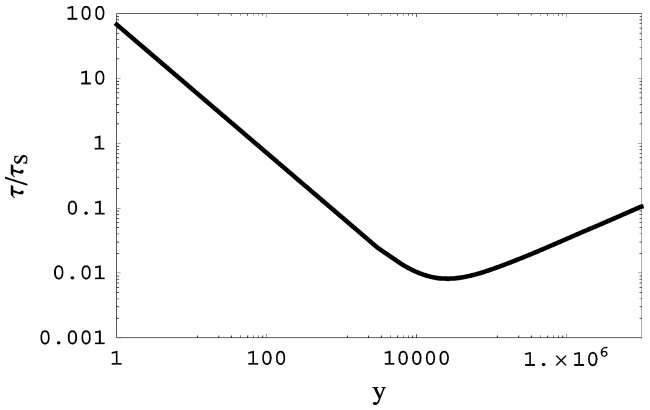

Figure 2. Relative search time (the ratio of the calculated search time as compared to Smoluchowski time, see text) as a function of the adsorption strength for $a = 1$ nm, $r = 30$ nm, $\alpha = 1$, $n_{ads} = 1000$, $n_p = 1$, and $d = 0.001$.

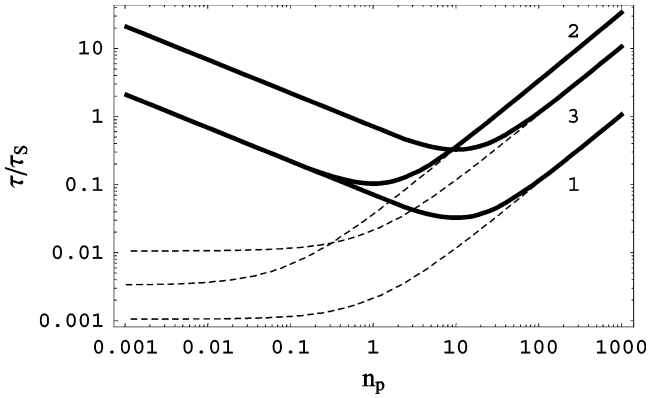

Figure 3. Relative search time as a function of protein concentration for (1) $n_{ads} = 1000$, $y = 1000$; (2) $n_{ads} = 100$, $y = 1000$; and (3) $n_{ads} = 100$, $y = 100$. The dotted curve is the result without the correlation term.

C. Discussion of the Results. As shown in Figure 2, for low values of y, the search time is large because of weak attraction or even repulsion between the protein and DNA, which prevents scanning for the target. The increase of the adsorption energy makes the search time shorter, reaching the optimal value, after which it starts to grow again. The latter is due to the fact that for strong adsorption (large y) protein molecules spend most of the time diffusing along DNA with rare unbinding events. This makes the sliding length λ long. Because the 1D diffusion is slow, this effect increases the overall search time. For very large values of the adsorption energy, at which $y \geq y^* \approx L^2 n_{ads}/(r^2 d)$, the sliding length becomes equal to length L, and the relative time reaches a plateau (not in the range displayed in Figure 2). For $\alpha = 1$, the value of the plateau reads

$$\frac{\tau}{\tau_s} = \frac{a}{n_{ads} L}\left(1 + \frac{n_p}{d}\frac{L^2}{r^2}\right) \qquad (20)$$

Such a complex dependence of the relative time on adsorption energy was first observed in ref 6 and analyzed in ref 5.

The relative time has, as well, a non-monotonic dependence on the concentration of free proteins in solution, see Figure 3. When the concentration is very low (remember that n_p can be substantially smaller than 1), the protein molecule squanders most of its time on binding and unbinding events and it does not scan much of the DNA length. This is because after the protein binds to DNA the thermodynamic drive to *unbind* becomes enormous: at low concentrations, any binding or unbinding event significantly shifts the chemical equilibrium in one direction or another. As a result, the searching time becomes very long in comparison with the ordinary Smoluchowski diffusion mechanism. In this case, the correlation time dominates, and the relative search time increases with increasing n_p.[43] In the opposite limit ($n_p \gg 1$), the density of free proteins is so large that there is always a protein close to the target. Then, there is no need for scanning along the DNA molecule because proteins can reach the target much faster via 3D diffusion. In general, as the value of y grows, the ratio τ/τ_S decreases and the optimal n_p value goes down as well, Figure 3. As n_{ads} grows, the position of the minimum of calculated times shifts to the right and the minimum value goes down.

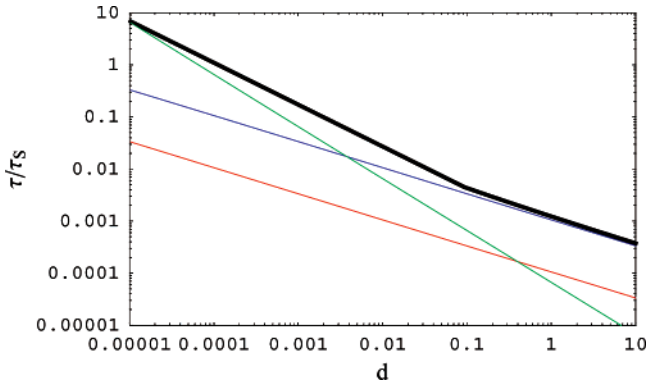

Figure 4. Relative search time as a function of the ratio of the diffusion coefficients d. Notations for the curves: total time (black), time spent in 3D (red), time in 1D (blue), and the correlation term (green). Parameters: $n_{ads} = 100$, $n_p = 1$, $y = 1000$.

Any $\alpha < 1$ will impede the searching process. The simplest model considered here suggests $\alpha = 1$, which corresponds to an uncorrelated, random diffusion. However, one can think of some sophisticated interplay between the bulk and surface diffusion that effectively leads to $\alpha > 1$ ("superdiffusion"), and the latter would accelerate the search. But without a consideration of a particular biophysical model behind such superdiffusion, it would not make sense to speculate about it any further.

It should be stressed once more that in our derivation *we did not assume equilibrium* between adsorption and desorption of proteins from nontarget sequences. The equilibrium, however, is a particular case of our analysis. Here, $y_{eff} = 1$, and, for $\alpha = 1$,[44] our calculation yields

$$\frac{\tau}{\tau_s} = \frac{a}{r}\frac{2}{y\sqrt{n_p}}\frac{\sqrt{d}+1}{d} \qquad (21)$$

In this case, our resulting formula due the account of the correlation term is different from ref 4 (reflected by the $+1$ term in eq 21). This term will significantly increase the search time and make it very difficult to explain the facilitated diffusion for realistic values of $d \ll 1$. Thus, the accelerated search is to a high degree facilitated by the nonequilibrium character of the environment inside the cell. We also obtain that at small d the search time is dominated by the correlation term in eq 19 while at large d the diffusion in 1D gives the dominant contribution to the total search time, see Figure 4. Note also that one can optimize the search time via minimizing the total time in eq 19 over y; naturally, the optimal y value appears to be a decreasing function of d.

III. DNA−Protein Binding: Electrostatic Mechanism of Primary Sequence Recognition

A. Experimental Observations. Similar to the first part, we concentrate here mainly on proteins with a pronounced sequence-specificity of interactions with the DNA. Several mechanisms of DNA−protein sequence-specific recognition have been discussed in the literature.[19,20] Some of them are based upon the formation of hydrogen bonds between protein amino acids and DNA bases approached through the DNA grooves. The others invoke electrostatic, hydrophobic, steric, hydration, or van der Waals interactions. There is, however, *no* unambiguous code for DNA−protein recognition. It is rather a probabilistic than a deterministic process: the same protein can bind to a number of DNA sequences with different affinities and thus can tolerate some degree of mismatch. It is the sequence-dependent DNA structure that determines the positions and strengths for interactions of all types formed by a DNA fragment with a DNA-binding protein.

Such proteins typically possess two binding modes. In the nonspecific mode, the protein remains flexible to allow easier scanning (for the *lac* repressor, the lysine and arginine residues are quite mobile). In the specific binding mode, the protein forms stronger interactions with the DNA that can induce substantial deformations both in the protein (binding-induced protein "folding") and in the DNA structure.[21] The proteins are typically more rigid in this binding mode, see ref 22 for the *lac* repressor.

It has been experimentally observed that the rates of association of many DNA-binding proteins are strongly *salt-dependent*, indicating the importance of electrostatic DNA−protein interactions. In particular, for the *lac* repressor the observed binding constant to DNA drops down dramatically as the concentration of simple salt in solution grows.[23,24] Strong sensitivity to the presence of divalent cations in solution has also been detected. The nonspecific binding mode of the *lac* repressor is entirely electrostatic with about 11 charge−charge interactions of positively charged protein amino acids (Lys, Arg, and probably His) interacting with the negatively charged DNA phosphates. Specific repressor−DNA complexes contain about six to eight electrostatic contacts and in addition seven hydrogen bond interactions.[23,24] The numbers of charge−charge interactions for both complexes are extracted from the slope of the *lac*-repressor binding constant on the log of the ionic strength of the solution.

A number of other gene regulatory proteins, for example, the RNA polymerase,[25] also have a positively charged patch in the DNA-binding domain. The electrostatic interactions of DNA phosphates with positively charged protein amino-acids are believed, however, to be largely sequence-nonspecific. They are thought to provide a general, nonspecific affinity of proteins to DNA that allows proteins to stay close to the DNA surface and not dissolve into the solution. The subsequent formation of hydrogen bonds in the grooves of the double helix does depend strongly on the DNA sequence; thus, the track of DNA to which the protein will bind through building hydrogen bonds must be recognized first.[26] In the following, we focus precisely on this electrostatic contribution to the sequence specific recognition, which can drive the protein to its binding site.

B. Complementarity, Adaptation, and Randomness. What is the fingerprint for the DNA−protein recognition? As we will show below, the complementarity of DNA and protein charge patterns in the recognition region can provide a primary recognition mechanism. For DNA−protein electrostatic interactions, such an option has not been studied before. Note that in each binding mode the protein will tend to maximize the number of corresponding interactions with the DNA. That might involve some *adaptation* of protein and DNA aimed to improve the complementarity of their interaction lattices.[45] Such interaction-induced adaptation of the protein and DNA structure has been visualized recently for nonspecific and specific complexes of the *lac* repressor with DNA.[22,26]

We will show that the electrostatic recognition between the protein and its binding track on DNA results in a potential well that traps the protein. On the contrary, when the degree of complementarity is small (mismatch is large), such a well will be shallow and practically unnoticeable, allowing proteins to *slide* along DNA easily without trapping. Such a model of DNA−protein recognition is conceptually similar to the theory of electrostatic recognition of homologous genes on two juxtaposed DNA molecules, considered earlier by some of us for torsionally rigid[27] and elastic DNA duplexes.[28,29] Also, the suggested model is reminiscent of the model of electrostatic complementarity developed for describing protein−protein electrostatic interactions in their complexes.[30]

Some effects of *randomness* of the energy profile for protein diffusion on DNA originating from the sequence specificity of DNA−protein interactions has been considered recently within several theoretical models.[7,31,32] In particular, for a random sequence nonspecific Gaussian-correlated energy profile the protein diffusion was shown to be strongly *impeded* when the roughness of the potential surface exceeds the thermal energy.[7] In ref 32, the base-pair-specific formation of hydrogen bonds between protein chemical groups and DNA bases has been taken into account. It has allowed the authors to predict the preferred positions of recognition sequences on DNA for binding of RNA polymerase as well as to study the effects of this randomness onto the properties of protein diffusion.

C. Model and Approximations. Describing the "coarse grain recognition", we will make substantial simplifications. For instance, it is known that a repressor protein binding to DNA involves a release of counterions condensed on the double helix because positively charged protein residues replace them in interactions with the negatively charged DNA phosphates. Upon protein sliding along the DNA, a fast equilibrium is established between cations removed from the DNA surface in front of the protein and rebinding to the DNA behind it.[6] This process of "evaporation" of adsorbed cations induced by "ironing" the DNA by the protein involves cation−DNA binding energies larger than the thermal one. We thus neglect the contribution of rearrangement of cations in the hope that this will rather contribute a constant *independent* of the patterns of fixed charge distributions considered in the model. Although in this case the linearized Poisson−Boltzmann model theory may fail easily, some *qualitative* features of the result are expected to be similar to those obtained from the solution of the full nonlinear Poisson−Boltzmann equation (for similar situations, see, e.g., refs 33 and 34).

The DNA and a protein will be modeled hereafter as linear *quasi-periodic* 1D charge lattices with the average separation h between the elementary charges e_0 on both lattices. Only electrostatic interactions are taken into account in the model. The axis-to-axis DNA−protein separation is R, the protein has $M = 2N + 1$ charges, and the DNA has an infinite number of charges, see Figure 5. DNA charges are all negative; protein charges are all positive.

More complicated charge distributions as well as the DNA helicity can be, in principle, incorporated in a more sophisticated model, but a "linear model" is a good starting point, at least because some DNA-binding proteins are known to move in a spiral-like fashion following the DNA helical motif. Indeed, the

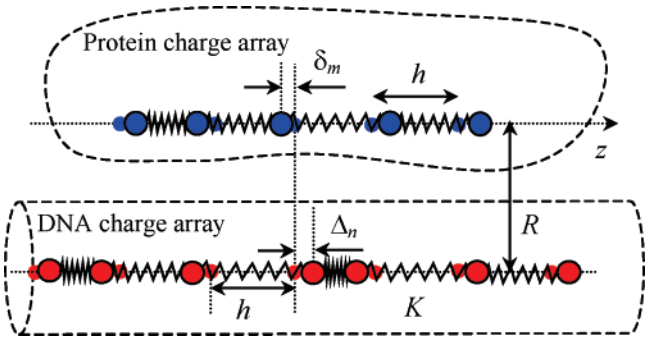

Figure 5. Scheme of protein−DNA electrostatic recognition. The protein and DNA are modeled as linear arrays of point-like charges: protein charges are all positive (blue), and DNA charges are all negative (red). In the model of the long-range order shown in the picture, DNA and protein charges keep the average periodicity *h* (indicated by small semitransparent circles). The charges are displaced randomly from these positions on Δ_n and δ_m on the corresponding DNA and protein cites. The DNA−protein separation is *R*, and the elastic constant of DNA and protein backbone is *K*. This constant is taken infinitely large in the model considered in the text.

quasi-1D sliding of a protein along DNA considered here can be visualized as sliding either along a straight array of charges or along a helical path in the proximity of the DNA phosphate strand. Note that the observed spiraling of some proteins around the helix upon sliding on DNA (e.g., RNA polymerase[35]) is consistent with the picture of electrostatic interactions of proteins with *helical* DNA charge patterns. Namely, upon tracking the negatively charged DNA strands a protein should *not* cross the electrostatic barriers between DNA strands and grooves (see ref 56), contrary to the situation when it just slides along the DNA axis without any spiraling.

Distance *h* characterizes the average separation between the phosphate charges on DNA (∼7 Å along a *B*-DNA helical phosphate strand, ∼3.4 Å along the axis of single-stranded DNA, and ∼1.7 Å along the axis of the double-stranded *B*-DNA) and a typical periodicity of charges on the protein.[46] The actual positions of DNA phosphates and of protein charges *fluctuate* about these regular positions with some dispersion: Δ_m are the variations on the *m*th site on DNA and δ_n are the variations on the *n*th site of the protein. There can be two models to describe the positions of charges on DNA and protein interacting arrays.

In the model of *long range order*, $z_n = nh + \Delta_n$ with $<\Delta_n^2> = \Delta^2$ and for the protein $z_m = mh + \delta_m$ with $<\delta_m^2> = \delta$,[2] see Figure 5. In this model, the periodicity of the charged lattices persists at all distances along the molecules and variations in charge positions are described by a Debye−Waller smearing of the lattice. Physically *this is not a good model for DNA*, see ref 36 for a detailed explanation. We consider it here only for tutorial purposes because it is easier to handle. Furthermore, the notion of a long-range order must be apprehended with a pinch of salt because of course there could be no real long range in one dimension.

Much more realistic for DNA is the *short-range order* model in which the mismatches in positions of charges *accumulate* along the lattices

$$z_k = kh + \sum_{s=0}^{k} \Delta_s \qquad (22)$$

where the actual values of Δ_s are sequence-specific. This model mimics the sequence-specificity of the DNA structure.

The recognition between DNA and protein takes place in the region where the patterns of charges are the same, that is, where the equality $\Delta_n = \delta_m$ holds. For such "recognizable" sequences, the DNA−protein electrostatic interaction energy is expected to be *lower* than that for sequences with uncorrelated patterns of charges. We position the center of the complementary region on DNA at $z = z_*$; the center of the protein is at $z = z_0$.

As mentioned, in what follows, we neglect the elasticity of the protein and of DNA, assuming that the positions of charges *cannot* be affected by mutual electrostatic interactions of the lattices. In a simplest model that does take into account the elastic response, the charges can be connected by elastic springs and their actual positions will be found from the minimization of elastic and electrostatic energy. As a result, the charges adjust their positions to some extent and the depth of the potential well near the recognition region will decrease.[47]

When calculating DNA−protein electrostatic interactions, we set the dielectric constant of the medium between them to be small, that is, $\epsilon_c = 2-5$. This assumption can hold for interaction of charges near the contact, where the water molecules are likely to be removed and interaction of charges takes place predominantly through a low-polarizability environment. The charges far from the close contact, however, interact mainly through the electrolyte solution and the approximation of small dielectric constant is likely to fail there (see ref 37 for some effects of low-dielectric DNA interior on the electric field around the molecule). Larger dielectric constants used for this region would diminish the interaction energy. The main prediction of our simple model is the shape and depth of the recognition well in a relatively tight DNA−protein contact, for which the small ϵ_c value is likely to hold.

D. Expression for the Recognition Energy. For a given set of values of δ and Δ, simple Fourier analysis shows that a general expression for the energy of electrostatic interaction of two linear charge arrays in electrolyte solution with the reciprocal Debye screening length κ can be written as

$$W_{el} = -\frac{e_0^2}{\epsilon_c \pi} \int_{-\infty}^{\infty} dq\, K_0\left(\sqrt{q^2+\kappa^2}R\right) e^{iqz_0} \times \sum_{m=-N}^{N} \sum_{n=-\infty}^{\infty} e^{iqh(m-n)} e^{iq(\delta_m - \Delta_n)} \qquad (23)$$

where $K_0(x)$ is the modified Bessel function of the second order. We now average the energy eq 23 over the realizations of Gaussian uncorrelated fluctuations in charge positions on protein and on DNA.

Long-Range Order. In this case, the Gaussian average of eq 23 over the position of charges is trivial and the total interaction energy of a protein with DNA is given by

$$\langle W_{el}\rangle = -\frac{2e_0^2 M}{\epsilon_c h}\left\{K_0(\kappa R) + 2\sum_{n=1}^{\infty} K_0\left(\sqrt{n^2g^2+\kappa^2}\,R\right) e^{-n^2g^2\Omega^2/2} \times \cos[ngz_0]\right\} - \frac{2e_0^2}{\pi\epsilon_c} M \int_0^{\infty} dq\, K_0\left(\sqrt{q^2+\kappa^2}\,R\right) \times \cos[q(z_* - z_0)](1 - e^{-q^2\Omega^2/2}) \qquad (24)$$

Here $g = 2\pi/h$ determines the reciprocal screening length connected with the charge periodicity and $\Omega^2 = \delta^2 + \Delta^2$. The

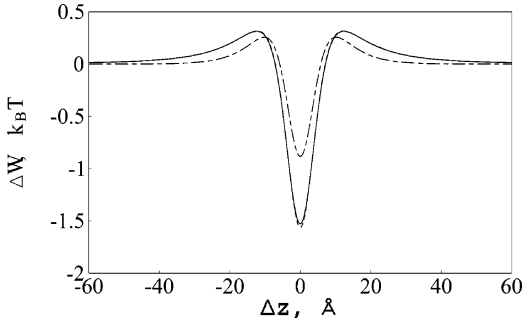

Figure 6. Electrostatic recognition energy for the fictional case of *long-range* order. Numerical integration of eq 24 at $\kappa = 0$ and at $\kappa = 1/(7$ Å) are, correspondingly, the solid and dashed−dotted curve; the simplified result eq 25 is the dotted curve. Parameters: $M = 11$, $R = 10$Å, $\epsilon_c = 2$, $\epsilon = 80$, $\delta^2 = \Delta^2 = 0.5$ Å^2.

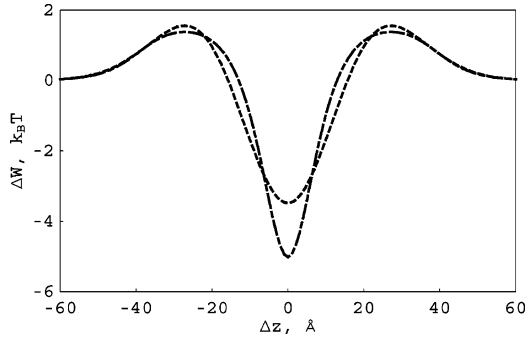

Figure 7. Recognition energy profile for the realistic *short-range* order in the charge positions. Dotted−dashed curve is the exact result eq A3, thick dotted curve is the expansion eq 26. Parameters: $h = 3.4$ Å, $\kappa = 1/(7$ Å) and other parameters are the same as in Figure 6.

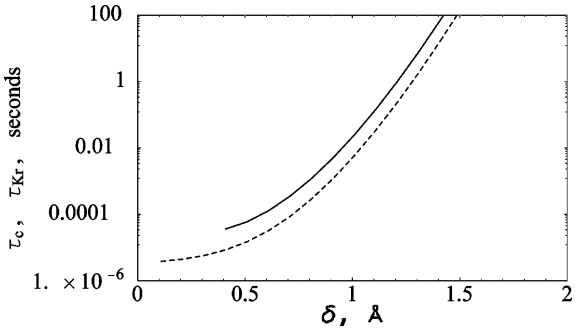

Figure 8. Residence time in the well as calculated from the Kramers equation, τ_{Kr} (solid curve), and the mean first passage time from the bottom of the well to the top of the potential barrier, τ_c, eq 1 (dashed curve). Parameters are the same as in Figure 7, $D_1 = 10^8$ Å^2/s.

first term in eq 24 is the attraction energy of M charges to a homogeneously charged DNA "line". The second term accounts for the energy barriers due to the discreteness of DNA charges. The third term describes the difference in the interaction energy of the protein with a complementary region on DNA as compared to that with a noncomplementary region. This is the electrostatic DNA−protein *recognition energy* that will be denoted hereafter as ΔW. It is this quantity that represents a well for protein trapping on its complementary track on DNA. It is proportional to the number of charges in the complementary region M and not to M^2 because each protein charge is in register with only *one* DNA charge in the recognition domain.

For small fluctuations and in the absence of added salt the recognition energy, being h-independent in this limit, scales like the mean-squared fluctuation amplitude, Ω^2, and returns a particular simple form

$$\frac{\langle \Delta W(\Delta z)\rangle_{\text{long-range}}}{k_B T} \approx -\frac{l_B M \Omega^2 \epsilon}{2\epsilon_c} \frac{R^2 - 2\Delta z^2}{(R^2 + \Delta z^2)^{5/2}} \quad (25)$$

where $l_B = e_0^2/(\epsilon k_B T)$ is the Bjerrum length in water and $\Delta z = z_0 - z_*$.

Short-Range Order. In this case, the derivations are more cumbersome and are presented in the Appendix. The approximate expression for the recognition energy, obtained under similar simplifying assumptions as in derivation of eq 25 reads as

$$\frac{\langle \Delta W(\Delta z)\rangle_{short-range}}{k_B T} \approx -\frac{2\, l_B \epsilon}{\epsilon_c \pi} \gamma K_0(\kappa R) \frac{\sqrt{\pi}\,(2(\alpha+\beta) - \Delta z^2) e^{-(\Delta z^2/4(\alpha+\beta))}}{8(\alpha+\beta)^{5/2}} \quad (26)$$

where M-dependent factors α, β, and γ are defined in the Appendix. We thus obtained two handy but nontrivial expressions for the recognition well, eqs 25 and 26. As we have explained earlier, eq 25 is not structurally justified for DNA, and so a slightly more complicated eq 26 is recommended, that typically results in deeper and wider wells.

E. Shape of the Well and Protein Residence Time. These approximate shapes of the recognition energy for long- and short-range order reveal good agreement with the corresponding exact numerical calculations, see Figures 6 and 7. Interestingly, the well is, of course, symmetric in Δz, being confined on both sides by two potential barriers.[48] The width of the well in the case of short-range order grows with the length of the recognition domain. At physiological salt concentrations the energy well becomes less deep but does not disappear. In realistic situations, there will hardly be any mobile ions between the protein and DNA anyway. By varying the model parameters (R, Ω, M, κ), one can vary the depth of the well in a wide range. The depth diminishes nearly exponentially with the separation R between the protein and DNA axes in the case of short-range order in the presence of salt, and as R^{-3} for the long-range order without salt.

It appears that the width of the recognition energy well in the short-range order is rather small. This well is unlikely to work as a *funnel* directing a protein from far away on DNA toward its actual binding site, and this is not what is physically expected. The protein diffusion will, however, be slowed down in the vicinity of the well. The calculations of the mean first-passage time for the energy wells (eq 25) has revealed, however, the unimportance of the actual well shape:[38] the well works like a Smoluchowski drain and only its width matters.

We have estimated a typical time the protein spends in the well. In Figure 8, we show the results of the Kramers-like approach for the inverse escape rate (τ_{Kr}) and the mean first-passage time from the bottom of the well to the top of the barriers (τ_c), as calculated from eq 1 for the energy well given by eq 26. As expected, for stronger amplitudes of distortions both the depth of the well and the height of the barriers increase, resulting in longer times the protein spends in the well. One can expect that if the residence time is larger than a typical time of protein conformational rearrangements then an interaction-induced protein folding (or unfolding) can occur.

These times lie typically in the microsecond to millisecond range, depending on the size of the protein domains that have to rearrange. Being confined in the well, proteins can adjust to DNA even better via forming stronger interactions (e.g., hydrogen bonds). The latter can be modeled as a delta-like

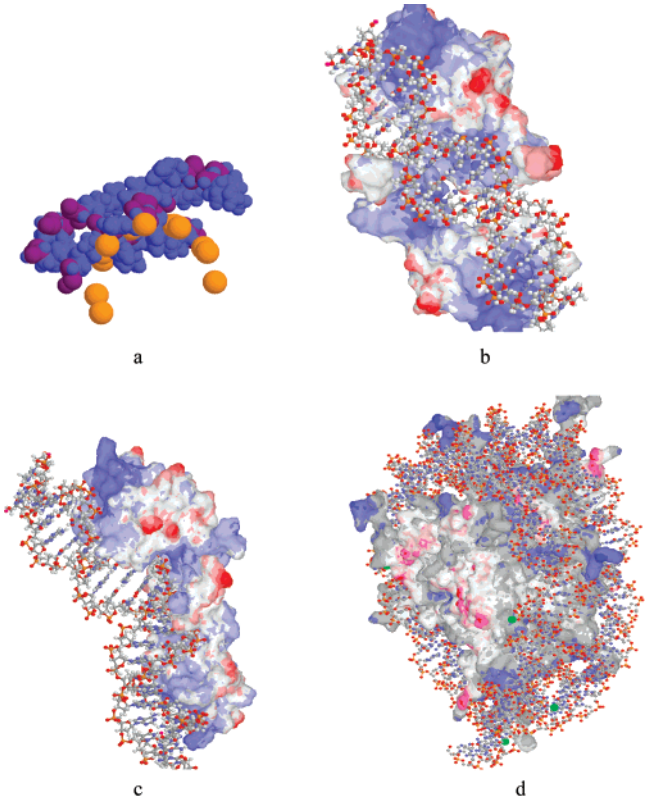

Figure 9. (a) Charge complementarity in DNA−protein interactions is realized via close contacts of positively charged protein residues following the path of negatively charged DNA phosphates. (a) Arrangement of positively charged residues Lys, Arg, and His (shown in blue) in the nonspecifically bound *lac* repressor complex near the contact with the DNA phosphates (shown in yellow). The view along the DNA axis. (b−d) Visualization of the electrostatic potential distribution for nonspecifically (b) and specifically (c) bound *lac* repressor−DNA complexes as well as on the histone proteins in the nucleosome core particle (d). Blue color corresponds to positive values of the electrostatic potential, and the potential is negative in red regions. The DNA potential is not shown. The images are obtained with the help of "MDL Chime" program for the protein visualization and with an integrated electrostatic potential solver in "Protein Explorer 2.80". We have used the Protein Data Bank files for the atomic coordinates in these protein−DNA complexes, 1osl.pdb, 1l1m.pdb, and 1aoi.pdb, correspondingly.

potential, which is switched on in the energy minimum after some time of residence. Such binding-induced changes in protein conformations appear to be necessary to allow proteins to perform a fast diffusion on a nonspecific DNA fragment and at the same time to bind strongly to a specific target site on DNA.

F. Potential Distribution. As an example of importance of the close contacts of opposite charges in DNA−protein interactions, let us consider now the distribution of electrostatic potential around some DNA-binding proteins: the *lac* repressor and the histone octamer. The structure of the nonspecifically bound *lac* repressor has been obtained in ref 22, 1osl.pdb file of the Protein Data Bank (PDB, www.rcsb.org). First, the DNA is almost straight in the nonspecific complex and the hydrophilic positively charged lysine, arginine, and histidine residues of the protein are in close proximity to DNA phosphates. Figure 9a, for example, shows the contact region of these cationic protein residues (represented in blue) and the DNA phosphates (shown in yellow). On the contrary, in the *lac* repressor bound specifically to DNA, PDB entry 1l1m.pdb and ref 39, the DNA is bent by about 36° upon stronger overall protein binding, Figure 9c.

The electrostatic potential distribution in these complexes was obtained using a solver in Protein Explorer and MDL Chime protein visualization programs, Figure 9b and c. Blue and red regions in these figures correspond to positive and negative potential values, correspondingly. The color intensity correlates with the potential absolute value. As one can see, *the distribution of the positive potential on the nonspecifically bound lac protein follows the helical pattern of negative charges in the DNA binding region*. Positive charges are in close contact with the DNA phosphates, while negatively charged residues are further away from the DNA. Positively charged protein residues accumulate near DNA phosphates also for DNA-protein complexes involving zinc finger and leucine zipper recognition motifs. So, we expect the consequences of DNA-protein charge matching to be quite general for DNA-protein recognition. This supports, or at least does not contradict the model of DNA−protein charge−charge recognition suggested above. A step further in the development of DNA−protein charge recognition models would be, using the PDB files for various DNA−protein complexes, to analyze whether the helix of DNA phosphates indeed generates in its proximity an array of *periodically positioned* positively charged protein residues. And the basic question then would be how sequence-specific are these charge-charge DNA-protein interactions.

For the nucleosome core particle, PDB entry 1aoi and ref 40, the 1.75 turns of DNA superhelix are shown wrapping around the histone octamer in Figure 9d. The Mn^{2+} cations bound to DNA (in green) appear to be necessary for obtaining good crystals, affecting the interactions between neighboring core particles in the assembly.[41] One can recognize a ring with a positive charge on the outer histone core surface bound to the DNA. This ring ensures quite uniform electrostatic binding affinity along the wrapped DNA, in addition to specific binding contacts in the nucleosome in places where the DNA minor groove faces the histone core.

IV. Summary

Facilitated protein diffusion on DNA is a complicated process, which exploits a relatively fast diffusion through 3D sections of protein transport toward the DNA in solution and presumably much slower 1D diffusion along the DNA chain. If not for the impeded motion on 1D tracks, then one could have suggested that reaching the target by one protein is accelerated via narrowing down the search space from 3D to 1D. Our analysis shows, however, that *there is no facilitated diffusion for one protein under realistic conditions*. Acceleration of the overall search process is ensured by parallel, simultaneous scanning for the target by many proteins adsorbed on the DNA due to a nonspecific binding.

The patterns of phosphate charges on DNA correlate with its sequence. As we have shown, the complementarity of charges on the protein and on the DNA target sequence can provide a sufficiently deep well, which may slow down a protein diffusing along DNA. The estimated residence time is enough to allow the protein to start performing its specific function, but not passing by the target.

Although some details of the search and sequence recognition thus seem to get clearer, many questions still remain to be answered. All of the conclusions made are "averaged" over many degrees of freedom. Thus, correlated motion of proteins should be investigated, as well as the effects of DNA conformational dynamics. It is also important to take into account the nonequilibrium nature of the cell environment, which was incorporated in our analysis in a simplistic form, in order to

clarify such fundamental issues as passive versus active biological transport.

Acknowledgment. We thank Aaron Wynveen for providing the results of Monte Carlo simulations for the recognition energy. The authors are also thankful to Alexander M. Berezhkovskii, Michael E. Fisher, Jack Mao, Gleb Oshanin, and Michael Urbakh for useful discussions and comments. AGC acknowledges the support from the Deutsche Forschungsgemeinschaft through the DFG grant CH 707/2-1. ABK acknowledges the support from the Welch Foundation through the grant C-1559 and from the US National Science Foundation through the grant CHE-0237105. AAK thanks 2001 Royal Society Wolfson Merit Research Award and EPSRC, grant GR/S31068/01. Part of this work was performed while two of us, ABK and AAK were participating in the program on "Biomachines", 2006, in Kavli ITP at the University of California at Santa Barbara.

Appendix

Recognition Energy in the Short-Range Order Model. The expression for the interaction energy of DNA with the protein charge array is

$$W_{\mathrm{el}} = -\frac{e_0^2}{\epsilon_c \pi}\int_{-\infty}^{\infty} \mathrm{d}q\, K_0\left(\sqrt{q^2+\kappa^2}\,R\right)\sum_{n=-\infty}^{\infty}\sum_{m=1}^{M} \mathrm{e}^{-iqz_n}\,\mathrm{e}^{iqz_m} = -\frac{e_0^2}{\epsilon_c \pi}\int_{-\infty}^{\infty} \mathrm{d}q\, K_0\left(\sqrt{q^2+\kappa^2}\,R\right)\epsilon(q) \quad \text{(A1)}$$

where $z_m = z_0 + mh + \sum_{s'=0}^{m}\delta_{s'}$ and $z_n = z_* + nh + \sum_{s=0}^{n}\Delta_s$ are the positions of charges on the protein and on the DNA, and $M = 2N + 1$ is the number of charges in the recognition domain. The complementary regions on the protein and on DNA start on their left end, at positions z_0 and z_*, correspondingly. The summation over the DNA charges is separated into three parts, $\sum_{n=-\infty}^{\infty} \to \sum_{n=-\infty}^{0} + \sum_{n=1}^{M} + \sum_{n=M+1}^{\infty}$, in order to extract the recognition energy. After performing the averaging over realizations of the random variables, Δ_s and $\delta_{s'}$, the Fourier component of the recognition energy reduces to

$$\Delta\epsilon(q) = \epsilon_{\mathrm{hom}}(q) - \epsilon_{\mathrm{nonhom}}(q) = \sum_{n=1}^{M} e^{-iqnh}\left(\sum_{m=1}^{n}\mathrm{e}^{iqmh}\mathrm{e}^{-q^2\Delta^2(n-m)/2} + \sum_{m=n+1}^{M}\mathrm{e}^{iqmh}\mathrm{e}^{-q^2\delta^2(m-n)/2} - \sum_{m=1}^{M}\mathrm{e}^{iqmh}\mathrm{e}^{-q^2\Delta^2 n/2}e^{-q^2\delta^2 m/2}\right) \quad \text{(A2)}$$

Calculating the sums and putting, for simplicity, $\Delta^2 = \delta^2$ we get a rather cumbersome expression for the average recognition energy

$$\langle \Delta W\rangle = -\frac{e_0^2}{\epsilon_c \pi}\, 2\int_0^{\infty} \mathrm{d}q\, K_0\left(\sqrt{q^2+\kappa^2}\,R\right)\Delta\epsilon(q)\cos[q\Delta z]$$

$$\Delta\epsilon(q) = \frac{M(\mathrm{e}^{\delta^2q^2/2}-1) - 2\mathrm{e}^{-M\delta^2q^2/2}(\cosh[M\delta^2q^2/2] - \cos[Mqh])}{2\mathrm{e}^{\delta^2q^2/2}(\cosh[\delta^2q^2/2] - \cos[qh])} + \frac{2\mathrm{e}^{-M\delta^2q^2/4}(\cos[Mqh/2](AC - BD) + \sin[Mqh/2](BC + AD))}{(\cosh[\delta^2q^2/2] - \cos[qh])^2}$$

$$A = \sinh[M\delta^2q^2/4]\cos[Mqh/2], \quad B = \cosh[M\delta^2q^2/4]\sin[Mqh/2]$$

$$C = 1 - \cosh[\delta^2q^2/2]\cos[qh], \quad D = \sinh[\delta^2q^2/2]\sin[qh] \quad \text{(A3)}$$

In the limiting case of $M = 1$, $\Delta\epsilon(q) \to 1 - \mathrm{e}^{-q^2\delta^2}$ and the result recovers the one for the long-range order. For larger h values, the $\Delta W(\Delta z)$ reveals "oscillations" about the basic shape, in agreement with the computer simulations of A. Wynveen.[42]

To get a simpler expression for the recognition energy, we take the q integral approximately using the Laplace method. Namely, for small q values we write $f(q) \equiv \mathrm{e}^{\ln[K_0(\sqrt{q^2+\kappa^2}R)\Delta\epsilon(q)]} \approx K_0(\kappa R)\mathrm{e}^{-\beta q^2}\gamma q^2\mathrm{e}^{-\alpha q^2}$, where the coefficients are defined by our model parameters as

$$\alpha = \frac{15\delta^2M(M+1) - h^2(2 + 3M - 3M^2 - 3M^3)}{20(1+2M)}, \quad \beta = \frac{RK_1(\kappa R)}{2\kappa K_0(\kappa R)}, \quad \text{and } \gamma = \frac{\delta^2M(1+3M+2M^2)}{6} \quad \text{(A4)}$$

Hence we obtain eq 26 of the main text. Alternatively, one can expand the function $f(q)$ near its maximum at $q = q_0$ using the method of steepest descent. This will give slightly more accurate result but with no simple expression for q_0.

Typically, the approximations considered above work better for smaller M values (when the shape of $\epsilon(q)$ is less complicated), for larger R values (when the decaying Bessel function under the integral suppresses the contributions from large q values to the integral more effectively), and at smaller h values (when the oscillations of $\Delta W(\Delta z)$ are less pronounced or disappear at all). The approximations typically work less satisfactorily near the bottom of the well. For instance, the expansion described above predicts that the depth of the well saturates as a function of M, while the exact calculations result in the increase of well depth with the number of charges in the recognition domain, as one would expect. For the typical values used in this study of $M = 3-10$, the difference between the approximate and exact results is not substantial, taking into account the level of simplification from the initial expression for the recognition energy (eq A3) to the final expression (eq 26).

References and Notes

(1) Alberts, B.; Bray, D.; Lewis, J. *Molecular Biology of the Cell*, 4th ed.; Taylor & Francis, 2002.
(2) Widom, J. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 16909−16910.
(3) Halford, S. I.; Marko, J. F. *Nucleic Acids Res.* **2004**, *32*, 3040−352.
(4) Berg, O. G.; Winter, R. B.; von Hippel, P. H. *Biochemistry* **1981**, *20*, 6929−6948.
(5) Hu, T.; Grosberg, A. Y.; Shklovskii, B. I. *Biophys. J.* **2006**, *90*, 2731−2744.
(6) Winter, R. B.; Berg, O. G.; von Hippel, P. H. *Biochemistry* **1981**, *20*, 6961−6977.
(7) Slutsky, M.; Mirny, L. A. *Biophys. J.* **2004**, *87*, 4021−4035.
(8) Slutsky, M.; Kardar, M.; Mirny, L. A. *Phys. Rev. E* **2004**, *69*, 061903.
(9) Eliazar, I.; Koren, T.; Klafter, J. *J. Phys.: Condens. Matter* **2007**, *19*, 065140.
(10) Klenin, K. V.; Merlitz, H.; Langowski, J.; Wu, C.-X. *Phys. Rev. Lett.* **2006**, *96*, 018104.
(11) Berg, O. G.; Blomberg, C. *Biophys. Chem.* **1976**, *4*, 367−381.
(12) Bruinsma, R. F. *Physica A* **2002**, *313*, 211−237.
(13) Gowers, D. M.; Wilson, G. G.; Halford, S. E. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 15883−15888.
(14) Wang, Y. M.; Austin, R. H.; Cox, E. C. *Phys. Rev. Lett.* **2006**, *97*, 048302.

(15) Elf, J.; Li, G.-W.; Xie, X. S. *Science* **2007**, *316*, 1191−1194.
(16) Schurr, J. M. *Biophys. Chem.* **1979**, *9*, 413−414.
(17) van Kampen, N. G. *Stochastic Processes in Chemistry and Physics*; North Holland: Amsterdam, 1992.
(18) von Smoluchowsky, M. V. *Z. Phys. Chem.* **1917**, *92*, 129−198.
(19) von Hippel, P. H. *Science* **1994**, *263*, 769−770.
(20) von Hippel, P. H.; Berg, O. G. *Proc. Natl. Acad. Sci. U.S.A.* **1986**, *83*, 1608−1612.
(21) Spolar, R. S.; Record, M. Th., Jr. *Science* **1994**, *263*, 777−784.
(22) Kalodimos, C. G.; Biris, N.; Bonvin, A. M. J. J.; Levandoski, M. M.; Guennuegues, M.; Boelens, R.; Kaptein, R. *Science* **2004**, *305*, 386−389.
(23) Record, M. T., Jr.; deHaseth, P. L.; Lohman, T. M. *Biochemistry* **1981**, *16*, 4791−4796.
(24) Winter, R. B.; von Hippel, P. H. *Biochemistry* **1981**, *20*, 6948−6960.
(25) Cramer, P.; Bushnell, D. A.; Kornberg, R. G. *Science* **2001**, *292*, 1863−1876.
(26) von Hippel, P. H. *Science* **2004**, *305*, 350−352.
(27) Kornyshev, A. A.; Leikin, S. *Phys. Rev. Lett.* **2001**, *86*, 3666−3669.
(28) Cherstvy, A. G.; Kornyshev, A. A.; Leikin, S. *J. Phys. Chem. B* **2004**, *108*, 6508−6518.
(29) Kornyshev, A. A.; Wynveen, A. *Phys. Rev. E* **2004**, *69*, 041905.
(30) McCoy, A. J.; Epa, V. C.; Colman, P. M. *J. Mol. Biol.* **1997**, *268*, 570−584.
(31) Hu, T.; Shklovskii, B. I. *Phys. Rev. E* **2006**, *74*, 021903.
(32) Barbi, M.; Place, C.; Popkov, V.; Salerno, M. *J. Biol. Phys.* **2004**, *30*, 203−226.
(33) Cherstvy, A. G. *J. Phys. Chem. B* **2007**, *111*, 7914−7927.
(34) Kornyshev, A. A.; Kuznetsov, A. M. *Electrochem. Commun.* **2006**, *8*, 679−682.
(35) Sakata-Sogawa, K.; Shimamoto, N. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 14731−14735.
(36) Kornyshev, A. A.; Lee, D. J.; Leikin, S.; Wynveen, A. *Rev. Mod. Phys.* **2007**, 943−996.
(37) Cherstvy, A. G. *J. Phys. Chem. B* **2007**, *111*, 12933−12937.
(38) Kornyshev, A. A. unpublished.
(39) Kalodimos C. G.; Bonvin, A. M. J. J.; Salinas, R. K.; Wechselberger, R.; Boelens, E.; Kaptein, R. *EMBO J.* **2002**, *21*, 2866−2876.
(40) Luger, K.; Mäder, A. W.; Richmond, R. K.; Sargent, D. F.; Richmond, T. J. *Nature* **1997**, *389*, 251−260.
(41) Cherstvy, A. G.; Everaers, R. *J. Phys.: Condens. Matter* **2006**, *18*, 11429−11442, and references cited therein.
(42) Wynveen, A. personal communication.
(43) The correlation term was not considered before, but neglecting it may lead to incorrect limiting behavior. For instance, in contrast with earlier predictions that the relative time becomes constant in the absence of proteins, ref 3, our model predicts that in the limit $n_p \rightarrow 0$, τ/τ_S grows to infinity as $\propto n_p^{-1/2}$ due to the correlation term.
(44) To validate the usage of eq 7, we have performed some computer simulations of diffusion of a single protein on DNA. We have calculated the time required for a protein to reach a target positioned randomly on DNA for the first time. The position of every protein attachment to DNA was chosen as random; the direction of scanning for each cycle, to the left or to the right, was also set random. Doing simulations for different λ, we indeed have obtained a linear dependence of the number of search cycles needed to find the target upon the ratio L/λ. This is consistent with eq 7 at $\alpha = 1$. We have also calculated the number of 1D cycles required to scan the whole DNA molecule. The scanning times obtained in this case were considerably longer; the scaling exponent α was, however, still close to 1.
(45) The adaptation may be a rather complicated process, as quoted below, similar to the adaptation effects between interacting DNA molecules that have been studied in detail. We will not take it into account here, limiting our consideration by a simplified, illustrative model. Adaptation itself will decrease the energy of the protein−DNA complex but it can only weaken their recognition because the flexibility of DNA and protein patterns will make them less sensitive to a complementary mismatch.
(46) This is a strong assumption given the diversity of the protein structures known. More detailed models will actually require us to validate this assumption via searching the protein data bases and probably recognizing some common features in their charge patterns in DNA binding domains valid for many proteins.
(47) In a more advanced model of protein elasticity, one can also allow for the possibility of a two-state behavior. In the first state—protein sliding on DNA—the protein is elastically soft and it scans DNA, probing various conformations whether it fits the underlying DNA charge pattern. The interaction energy is not strong enough to enable large protein deformations. Close to the recognition domain, the interaction strength grows and the protein can be driven into another, substantially deeper potential well, now along the "internal reaction coordinates" of a protein.
(48) The fact that two barriers accompany the well is counterintuitive. Its possible interpretation is as follows. Consider two charges, one negative on the DNA and one positive on the protein. Their positions vary randomly in the plane of charges along the same axis (the "fluctuation axis"). When the charges are on top of each other, i.e., when the fluctuation axis is perpendicular to the line connecting the charges, the effective charge−charge separations increase and the attraction between the charges is always weakened due to the fluctuations. However, when the fluctuation axis forms an acute angle with the line connecting the charges, the distortions in charge positions can give rise to a stronger attraction. Indeed, the energy profit from those charge configurations, for which the opposite charges become closer due to fluctuations, can be larger than the energy loss from those for which charge−charge separations eventually increase. Therefore, for complementary lattices of charges on DNA and on the protein, which are shifted longitudinally along their axis for distances larger than $\Delta z > R/\sqrt{2}$, the fluctuations enhance the electrostatic attraction. Therefore, the energy barriers emerge in their recognition energy when we start moving the protein from far away on DNA toward the minimum of the recognition energy well. These barriers appear only for fluctuations of charges along the DNA−protein separating plane (parallel fluctuations). For fluctuations perpendicular to this plane, the case not presented in the paper, there are no barriers. In this case, the separations between DNA and protein charges always increase due to fluctuations, independent of mutual charge positions. The recognition energy well can be modeled in this case by eq 25 without a Δz-dependent term in the nominator.
(49) Halford, S. E. et al. *Eur. Biophys. J.* **2002**, *31*, 257.
(50) Mirny, L. *Nature Physics* **2008**, *4*, 93.
(51) Hu, T.; Shklovskii, B. I. *Phys. Rev. E* **2007**, *76*, 051909.
(52) Barbi, M. et al. *Phys. Rev. E* **2004**, *70*, 041901.
(53) Misra, V. K. et al. *Biophys. J.* **1998**, *75*, 2262.
(54) Sokolov, I. M. et al. *Phys. Rev. E* **2005**, *72*, 041102.
(55) Merlitz, H. et al. *J. Chem. Phys.* **2006**, *124*, 134908.
(56) Cherstvy, A. G.; Winkler, R. G. *J. Chem. Phys.* **2004**, *120*, 9394.