Journal of Biological Physics **30:** 203–226, 2004.

A Model of Sequence-Dependent Protein Diffusion Along DNA

MARIA BARBI[1,2,*], CHRISTOPHE PLACE[3], VLADISLAV POPKOV[4,5]
and MARIO SALERNO[1]
[1]*Dipartimento di Fisica "E.R. Caianiello" and INFM, Università di Salerno, Baronissi (SA), Italy;* [2]*Laboratoire de Physique Théorique des Liquides, Université Pierre et Marie Curie, case courrier 121, 4 Place Jussieu, 75252 Paris cedex 05, France;* [3]*Laboratoire de Physique, CNRS-UMR 5672, École Normale Supérieure de Lyon, Lyon, France;* [4]*Institut für Festkörperforschung, Forschungszentrum Jülich GmbH, Jülich, Germany;* [5]*Institute for Low Temperature Physics, Kharkov, Ukraine*
*(*Author for correspondence, e-mail: barbi@lptl.jussieu.fr)*

Abstract. We introduce a probabilistic model for protein sliding motion along DNA during the search of a target sequence. The model accounts for possible effects due to sequence-dependent interaction between the nonspecific DNA and the protein. Hydrogen bonds formed at the target site are used as the main sequence-dependent interaction between protein and DNA. The resulting dynamical properties and the possibility of an experimental verification are discussed in details. We show that, while at large times the process reaches a linear diffusion regime, it initially displays a sub-diffusive behavior. The sub-diffusive regime can last sufficiently long to be of biological interest.

Key words: anomalous diffusion, DNA-protein interaction, dynamical models, sliding, target site

1. Introduction

The way by which proteins can find their target sites along a DNA chain represents a puzzling problem. In many cases, the reaction rate has been demonstrated to be faster than diffusion controlled [1–4]. Nonspecific sliding along the DNA has been proposed to be the main mechanism for faster search of the specific site on DNA ([5] and references herein). Nevertheless, a precise experimental determination of the statistical law characterizing the diffusive motion of protein along DNA during the specific site search is presently lacking. It is believed that during the sliding motion, the activation barrier for the translocation of the protein to continuous nonspecific positions is high enough to randomize the protein motion through collisions with the solvent molecules, but appropriately small compared to the thermal energy, in order to allow the protein to move [6]. This has induced some authors to propose a model where protein freely slides along DNA under the effect of the thermal fluctuations without any sequence dependent interaction, i.e., the DNA is seen as an homogeneous cylinder on which the protein can diffuse until the specific site is reached [6–8]. During sliding, however, the protein must be able to distinguish the specific region from nonspecific DNA so that a recognition mechanism must be involved. To this regard, the possibility that sliding could imply sequence dependent protein-DNA interaction is rather reasonable.

The aim of the present paper is to investigate this possibility in the context of a simple probabilistic model for a protein sliding along DNA, which accounts for sequence-dependent interaction between the nonspecific DNA and the protein. The model is based on the idea that the protein needs to "read" the underlying sequence during sliding in order to test whether special "signals" associated with the recognition site are present, i.e., a sequence-dependent interaction should be at work during the search. This means that the DNA sequence can influence the dynamics of the protein also far from the target region. In this sense the protein stop at the recognition site should be the extreme effect of a complex dynamics, i.e., the protein should follow a noise-influenced, sequence-dependent motion that includes the possibility of slowing down, pauses and stops. From this point of view the usual assumption of a standard random walk along DNA [2, 9–12] appears inadequate.

To investigate the possibility of a sequence-dependent diffusion motion of the protein along the DNA, we define a base sequence energy landscape from which hopping rates of the protein on the DNA (viewed as a discrete inhomogeneous lattice) can be deduced. The energy landscape is constructed by assuming a sequence dependent protein-DNA interaction inside the target region and extrapolating it to nonspecific regions. The diffusive motion of the protein is then studied by Monte-Carlo simulations of the probabilistic process on the landscape energy both in absence and in presence of thresholds which define different rules for the hopping motion. As a result we show that, while at large times the process reaches a linear diffusion regime, at the initial stage it displays a sub-diffusive behavior. It is remarkable that the anomalous diffusion regime can last for time large enough to be observable in single molecule experiments similar to those that have permitted to visualize sliding for different proteins [5, 12]. Base sequence induced dynamics along DNA was also considered in [13, 14] in connection with a nonlinear model of DNA, and in [15] in connection with the RNA-polymerase motion during the transcription process.

The paper is organized as follows: in Section 1 we introduce a sequence dependent model for protein-DNA nonspecific interaction. An energy landscape with minima corresponding to the recognition sequence is constructed. We then introduce four possible models for the protein translocation on DNA by using the sequence induced energy landscape and its modification as the inclusion of energy thresholds, which allow to describe different possible reading mechanisms. The rate of translocation to the neighboring sites is constructed from the energy landscapes (for the different models) by means of the Arrhenius law. In Section 2 we use Monte-Carlo simulations to study in detail the different dynamical regimes of our models. Finally, in Section 3 we discuss the limits of our analysis and the possibility to check the results with experiments, so as to verify if the inferred mechanism actually corresponds to the real one. In Section 5 we draw our conclusions.

2. The Model

2.1. Protein-DNA interaction and specific site recognition

For a suitable description of the search dynamics it is crucial to determine which sequence-dependent interaction is responsible for the specific recognition. This obviously depends on the protein-DNA specific complex and can be obtained, in general, from biochemical and crystallographic analysis. Nevertheless, some general features have been shown to be common to many specific recognition mechanisms. The probably most important recognition interaction is in fact mediated by a direct hydrogen bonding between protein amino-acids and special DNA chemical groups, that are present on the minor and the major groove sides of any base-pair (bp). Due to the DNA geometry and chemical composition, each different base-pair exposes on the major groove a different pattern of four chemically active sites (Figure 1) which can be either hydrogen bond acceptors and donors, or sites where a hydrogen atom or a methyl group are present [16]. Hydrogen bonding to these groups provides proteins to a highly specific "lock-and-key" mechanism, able to distinguish up to the single base-pair. The use of the chemical composition of base-pairs, which is evidently common to any DNA region, makes this recognition mechanism very general; the specificity of different proteins arising only from the different bps composition of the target DNA sequence. Additional specificity arises from the fact that usually the protein only made a subset of bonds between all the possible hydrogen bonds that can be made to its target sequence: the actual number, position and type of these bonds represent therefore the "footprint pattern" of each specific protein.

Other molecular interactions as van der Waals contacts and specific electrostatic bonds may also occur. From a qualitative point of view, the effect of these different specific bonds on the resulting diffusive dynamics should be, nevertheless, approximately the same: they only add further contacts with more or less specific DNA chemical groups. The quantification of the interaction energies involved in these contacts is instead more difficult to obtain: we will therefore focus, in this work, only on recognition mechanisms based on specific sets of hydrogen bonds. We will neglect other interaction which characterize the final specific complex like flexibility of the DNA [17], major conformational changes [18] or hydrophobic interaction by amino-acid intercalation in the DNA double helix. We suppose that these interactions are posterior to the hydrogen network recognition and are not effective during the protein sliding.

We also assume that, in each position along DNA, the protein "tries" to make the same set of hydrogen bonds as at the specific site, testing in this way the underlying sequence. In other words, we assume that the protein makes use, during the search, of the pattern of active chemicals groups that allows for the best binding to the specific

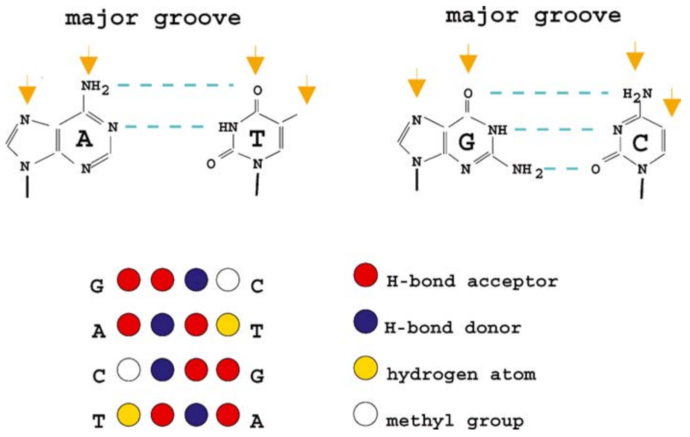

Figure 1. The positions of all the possible major groove interacting sites where base-pairs can make hydrogen bonds (top) and the corresponding base-pair patterns (bottom). Blue and red disks indicates the hydrogen donor and acceptor DNA groups respectively. White positions correspond to hydrogen atoms and yellow ones to methyl groups. Each base-pair is associated with a different 1 × 4 pattern.

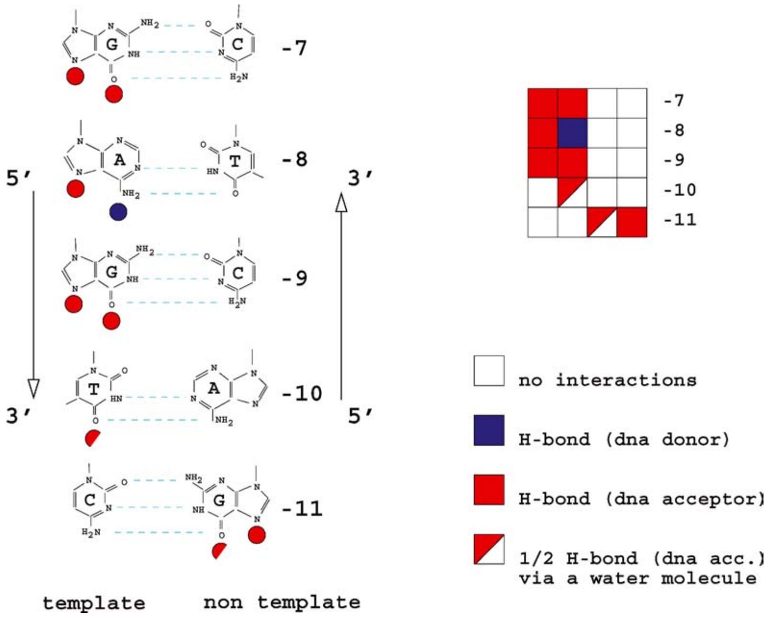

Figure 2. A sketch of the DNA interaction sites at the T7 promoter, where hydrogen bonds with corresponding RNA-polymerase chemical groups are made. Blue and red disks indicate the hydrogen donor and acceptor DNA groups respectively. On the right, the corresponding 5 × 4 pattern that T7 RNA-polymerase recognizes. The two half disks in the left part of this figure and their corresponding positions on the right pattern correspond to a couple of sites that share a water mediated hydrogen bond [26]. The recognition matrix R is constructed directly from the sketch. The entries $+1$, -1, $\frac{1}{2}$, 0 of the matrix correspond, respectively, to donor, acceptor, shared, and neutral hydrogen bonds.

site. This hypothesis can be justified on the basis of some known protein-DNA complex features. The stability of the protein-DNA nonspecific complex is mainly due to electrostatic interaction with the backbone phosphate of DNA [7] and to the entropic release of cations [19–21]. Besides these stabilizing factors, sequence-dependent interaction allows the protein to test the DNA during the target search [17]. Experimental data on endonuclease EcoRI show that pausing of the protein during sliding occurs at sites which resemble the specific sequence [22]. Thus, the *nonspecific* “reading” should be of the same nature as the *specific* recognition[1]. This suggests that a continuous variation between specific and nonspecific binding exists [22]. Once the target is reached, the transition from nonspecific to specific complex can be, eventually, induced by conformational changes of the proteins [18], these being however not relevant for the search phase. One can thus deduce that some interaction observed at the specific complex could be already present during sliding, and might be used in the recognition mechanism.

2.2. SEQUENCE DEPENDENT DIFFUSIVE MODEL

A suitable example of target recognition through a specific set of hydrogen bonds is given by the case of the bacteriophage T7 RNA polymerase. For this enzyme it is known [24–27] that the relevant set of sequence-specific recognition bonds between protein side chains and bases arises in the major groove in correspondence of the 5 bps sequence GAGTC extending from -11 to -7 relatively to the initiation site, via the formation of hydrogen bonds with the appropriate acceptor or donor chemical groups in the base pairs sides. In modeling the sequence dependent diffusion of a generic protein along DNA we will assume, for concreteness, the same 5 bp recognition sequence and the same pattern of hydrogen bonds. In Figure 2 we have depicted the hydrogen bonds made between the T7 RNA-polymerase and DNA at the promoter region, as revealed by crystallographic analysis [26]. We remind that, as already pointed out, the actual interaction can be more complex: it should include a more detailed pattern of molecular bonds, and eventually other kinds of interaction depending, e.g., on the protein and DNA three-dimensional structure and flexibility [17]. Anyway, the qualitative (statistical) features of the resulting sliding motion do not depend strongly on the detailed nature of the protein-DNA interaction, as it will emerge from our numerical study.

The concrete way to represent our simplified protein is by the introduction of a recognition matrix. This matrix should contain the information needed to match the target sequence, i.e., the pattern of active chemical groups that allows for the best binding at the 5 bps long promoter. The comparison of this perfect-matching set of bonds to the actual chemical features of any 5 bps sequence along DNA will give therefore a certain number of made (*matches*) and unmade (*mismatches*) hydrogen bonds between the DNA and the protein. For convenience, we represent the recognition pattern directly in terms of its corresponding binding sites on DNA.

Formally, the interaction pattern will be defined by denoting by $+1$, -1, 0 respectively the acceptor, donor, and noninteracting DNA sites. The DNA sequence is then represented as a list of vectors, $\ldots b_{n-1}, b_n, b_{n+1}, \ldots$, where

$$b_n = \begin{cases} (1, & -1, & 1, & 0)^{\mathrm{T}} & \text{for base A} \\ (0, & 1, & -1, & 1)^{\mathrm{T}} & \text{for base T} \\ (1, & 1, & -1, & 0)^{\mathrm{T}} & \text{for base G} \\ (0, & -1, & 1, & 1)^{\mathrm{T}} & \text{for base C} \end{cases}$$

The sequence of vectors b_n thus represents exactly the DNA sequence in terms of its possible chemical bonds on the major groove. The protein interacts, at position n, with the corresponding sequence of 5 bases, that is therefore represented by a 4×5 matrix $D_n = (b_n, b_{n+1}, b_{n+2}, b_{n+3}, b_{n+4})$. The consensus sequence GAGTC of the considered case corresponds to the matrix

$$D_n = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 \\ 1 & -1 & 1 & 1 & -1 \\ -1 & 1 & -1 & -1 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix}.$$

By considering the subset of hydrogen bonds actually made by the T7 RNA-polymerase with this target sequence (see Figure 2 [26]), we can define the following 5×4 recognition matrix $R(i; j)$, corresponding to Figure 2:

$$R = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 1/2 & 0 & 0 \\ 0 & 0 & 1/2 & 0 \end{pmatrix}.$$

The factors $1/2$ have been introduced in order to reproduce the shared hydrogen bond evaluated, in a first approximation, as half hydrogen bonds everywhere along the chain (see Figure 2). Note that the same kind of model can be defined for any other sliding protein using the same recognition mechanism to fit its target sequence, provided that the specific set of bonds is experimentally determined. DNA will be represented exactly in the same way, while the recognition matrix should change (in composition and length) according to the set of bonds characteristic to the specific protein-target interaction.

Once represented in such a way the recognition pattern and the DNA sequence, we need to introduce a suitable definition of the interaction energy. To this respect, we remark that each match will stabilize the complex, while mismatches will act as to destabilize the protein, which will tend therefore to move away from the "wrong" positions [6, 7]. For each position n along the chain we can therefore define an energy $E(n)$, simply by counting the number of matches and mismatches, and adding a corresponding negative or positive amount of energy, respectively (empty sites in the recognition pattern do not contribute to the energy).

With the notation previously introduced, the interaction energy can be written simply as

$$E(n) = -\epsilon \, \mathrm{tr}(R \cdot D_n) \tag{1}$$

where the dot $\cdot$ denotes the usual matrix multiplication and tr is the trace. Absolute minima correspond to the complete matching and thus to the recognition sequence GAGTC. Each positive or negative contribution to the energy, ϵ, corresponds to a hydrogen bond energy.

Note that the mobility of proteins dramatically depends on $\epsilon/k_B T$. Since there are no direct measurements of the interaction energies during sliding and it is difficult to make an estimate of the involved hydrogen bond strength, we shall use $\epsilon/k_B T$ as a free parameter. The resulting energy E(n) defines an irregular landscape on which the protein can move as discussed in the next subsection.

2.3. Different translocation mechanisms

The actual mechanism allowing for the shift of the set of hydrogen bonds from one position along the DNA to the next one is unfortunately unknown. In order to account for the possible effect of different translocation mechanisms, we introduce in the following four versions of the model, corresponding to different possible physical features of the fundamental step in the protein motion. The various possibilities are related to the presence and nature of an activation barrier separating one position from the next one. The length of hydrogen bonds (up to 3.5 Å in DNA-protein interaction [28]) can roughly reach the same order of magnitude as the distance between base pairs (3.4 Å). Therefore, the protein may eventually shift directly from one position to the next one without activation energy for the one step process. Nevertheless, it is also possible that the protein has to disrupt partially or completely the hydrogen bonds on one site before moving to the next: in this case one has to take into account the additional activation barrier that has to be overcome by the protein in order to move. We will therefore analyze the effect of energy barriers of different heights (models I to III, see below).

Furthermore, it is possible to imagine more complex scenarios. For instance, the protein could have internal flexibility allowing for conformational changes, eventually depending on the local degree of stability. This could allow for an additional modulation of the protein affinity for different DNA sequences. As a first attempt to investigate the possible effect of similar mechanisms, we will include in our model an effective modification in the protein-DNA interaction energy for regions with a low degree of homology. More precisely, following the suggestion of von

Hippel *et al.* [7], we will assume that, in positions where too many mismatches are found, the protein should undergo a conformational change from a "reading" mode to a "sliding" mode in which no hydrogen bonds are made and the interaction energy becomes independent of the sequence. We remark that in the sliding mode the protein is weakly bound to DNA and its dynamics is driven by the hydrodynamic forces of the viscous media surrounding the DNA (see below, model IV). This leads therefore to a "two-states model," for which, if the total energy $E(n)$ is over a threshold E_t, the system passes to a different state of constant energy $E_{\rm sl}$ where the protein can freely slide. We have to stress that we are modeling here the possible interaction modulation, induced by the internal protein flexibility, in a very schematic way: we just modify the energy landscape in correspondence to the unfavorable positions, without any detail on the real kinetic and activation barrier of the protein conformational change itself. The investigation of this simplified model is, anyway, just aimed to get a first insight on the possible relevance of local energy modulation effects on the resulting protein motion.

In order to investigate the differences in the protein dynamics induced by these different scenarios, we define and analyze four different models, sketched in Figure 3 and listed hereafter. Each different model results in a redefinition of the effective energy barrier $\Delta \mathrm{E}_{n\rightarrow n'}$, with $n' = n \pm 1$, initially defined simply on the basis of the local number of mismatches.

I) *no-threshold model* (Figure 3, I): hydrogen bonds can directly translate from one position to another without being destroyed. In this case the energy

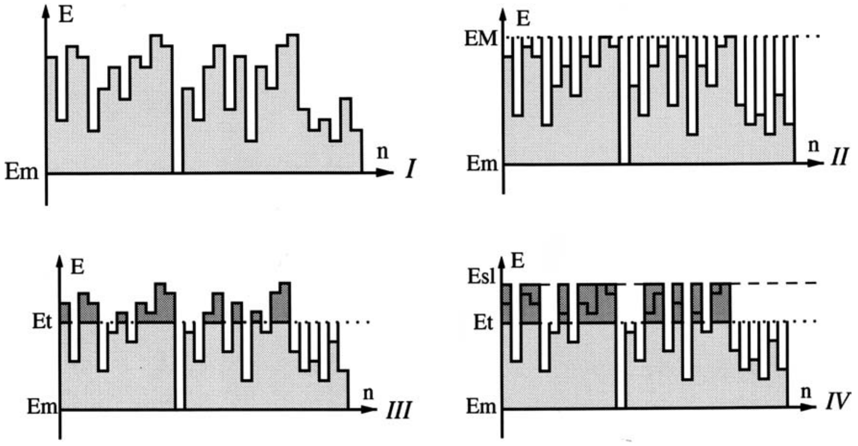

Figure 3. A schematic picture of the four considered variants of the model. On the horizontal axis, we represent a few (30) positions along DNA. Correspondingly we sketch the interaction energy E varying between its minimum (E_m) and its maximum (E_M) values. The interaction energy evaluated on the T7 DNA present similar rapid oscillations between different levels. The dotted lines indicate the threshold level E_t, set to E_M for model II, to an intermediate value for model III and IV. In the case of model IV, all energy levels above the threshold are redefined to a common value $E_{\rm sl}$ (dashed line).

difference $\Delta E_{n \to n'}$ from n to $n' = n \pm 1$ is simply the difference of the two energy levels as defined in Equation (1). Furthermore, $\Delta E_{n \to n'}$ will be set to zero if $E(n') - E(n)$ is negative, in the usual way, in order to allow a coherent definition of the transition rates. We define therefore:

$$\Delta E_{n \to n'} = \max[E(n') - E(n), 0]. \tag{2}$$

II) *maximal-threshold model* (Figure 3, II): in order to reach an adjacent site, the protein must destroy all bonds and pass through a state of "total mismatch." In this case the energy barrier only depends on the energy $E(n)$ and on the threshold level $E_M = \max[E(n)]$, and we get:

$$\Delta E_{n \to n'} = E_M - E(n). \tag{3}$$

III) *intermediate-threshold model* (Figure 3, III): in order to reach a next site, the protein must destroy all bonds passing through an intermediate "zero" state defined by a threshold energy E_t. The energy barrier will therefore depend on the neighboring energy $E(n')$ only if this is larger than the threshold energy E_t. Formally this reads

$$\Delta E_{n \to n'} = \max[E_t - E(n), E(n') - E(n), 0]. \tag{4}$$

Models I and II are actually the two limiting cases of model III when the threshold is set to the minimum and maximum values of the potential energy, respectively. These three models could therefore be considered as three cases of a *unique model*, just dependent on the choice of the energy threshold. We will anyway refer to these three cases as to models I , II and III in the following, for convenience. Note that in the general case of an intermediate threshold, the previous model gives already two different regimes for the protein, because the energy profile is qualitatively different in regions where $E(n)$ is greater or lower than E_t.

Finally, to account for the possibility of two regimes of the protein-DNA interaction associated with a protein conformational change, we propose a fourth model as follows:

IV) *two-regimes model* (Figure 3, IV): a threshold energy E_t separates "reading" *regions*, where the energy is $E(n) < E_t$, from "sliding" *regions*, where no hydrogen bonds are made and the protein can freely diffuse on a flat energy landscape, $E(n) = E_{\mathrm{sl}}$. Below the threshold, the barrier E_t still affects the translocation as in case of model III. For simplicity, we will fix the value of E_{sl} to $E_M = \max[E(n)]$. In this case, one can redefine the energy as

$$E(n) = \begin{cases} E(n) & \text{if } E(n) < E_t \\ E_{\mathrm{sl}} & \text{if } E(n) \geq E_t \end{cases} \tag{5}$$

and $\Delta E_{n \to n'}$ results to be defined as in case III, Equation (4).

Note that our model IV interpolates between straight sequence-dependent walk (model I) and the biological model of the specific site search by protein proposed by von Hippel in [6, 7]. The scenario suggested by von Hippel relies indeed on the idea that the specific interaction is "switched off" by a conformational change if too many mismatches are present. In that picture, the protein is more often in a "sliding" mode, where the specific hydrogen bond interaction is inactive. A quantitative description of this mechanism can be obtained by the introduction of our model IV, where the varying threshold level E_t accounts for the degree of homology which leads to the supposed protein conformational change.

To introduce dynamics we describe the motion of a protein along DNA as a Markov process on a discrete chain with sites representing consecutive DNA base pairs. A protein is represented by a particle on the chain which can hop to its nearest neighboring sites with rates

$$r_{n\to n'} = \frac{1}{2\tau} \times \exp\left(-\Delta E_{n\to n'}/k_B T\right); \quad n' = n \pm 1. \tag{6}$$

The Markovian process assumption (Eq. (6)) implies loss of memory about previous evolution during a single translocation step. This can be justified if the typical energy dissipation during translocation is much higher than the thermal energy $k_B T$. This is indeed the case as shown in the appendix. Notice also that the meaning of the rates defined in Eq. (6) is two-fold. From one side, when the energy barriers between sites are small ($\Delta E \ll k_B T$), the protein does no feel the underlying potential and diffuses "freely" along the DNA helix subjected to purely hydrodynamic forces due to the viscous medium surrounding the DNA. In the case of the lac repressor, Schurr [29] has estimated an upper limit for the diffusion constant as $D_{lac} = 4.5\ 10^{-9}\ \mathrm{cm}^2/\mathrm{s}$, by modeling the protein as a hard ball of radius a_{lac} spiraling along the double helix. To estimate the diffusion constant for a generic protein, one can rescale the result of Schurr accounting for the difference in sizes between the protein and the lac repressor, as $D = D_{lac}(a_{lac}/a)^3$. Thus, to cover the distance $\ell = 3.4$ Å corresponding to one base pair step, a time $\tau = \ell^2/(2D)$ is needed (for the T7 RNAP this time is equal to $\approx 10^{-7}$s). In the Markov process (6) this time corresponds to the typical time t needed for the translocation to the nearest site in absence of the potential, $t = 1/(2r_{n\to n'}) = \tau$. Thus τ in (6) can be viewed as the time the protein needs to cover a distance of one bp along DNA due to hydrodynamic forces. On the other hand, when the protein is strongly bound to DNA, i.e. for deep potential wells ($\Delta E_{n\to n'} \gg k_B T$), it must overcome the barrier to move, and the other process, driven by hydrodynamic forces, becomes negligible. Notice that the contributions of both processes are included in the definition of the rates in Eq. (6).

3. Results: Recognition Efficiency and Anomalous Diffusion

In this section we will study, through Monte-Carlo simulations, the dynamical behavior of the model sliding protein previously defined. We will use in the following

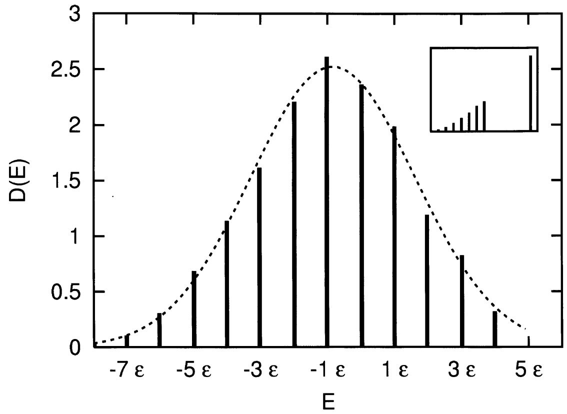

Figure 4. Energy level distribution (models I to III), obtained by averaging on the whole T7 DNA. A Gaussian fit of the resulting histogram (dashed line) is superimposed for comparison. Inset: The corresponding distribution for model $IV\,(E_t = 0)$.

the T7 DNA as testing sequence. An energy landscape can be defined on the base of this sequence according to the rules (1) or (5). The distribution of the energy levels obtained for models I to III and for model IV is shown in Figure 4.

The first important check of the four models is related to their affinity to the target region. Theoretically, one can easily estimate the stationary distribution of a population of proteins on the four different model landscapes as

$$\rho_\infty(n) \propto e^{-E(n)/k_BT}. \tag{7}$$

As usual, the stationary distribution only depends on the site energy, and not on differences and thresholds. Consequently, models I to III have the same distribution, whereas the redefinition of energy in model IV leads to a substantially different result. Eq. (7) straightforwardly implies that the recognition sites, which have the lower energy, will be in average the most populated.

In order to verify that this is indeed obtained in a dynamical context, we simulated numerically the time evolution of models I to IV taking a uniform distribution of independent proteins on a DNA region of 1000 bps as initial condition. Note that the assumption of an uniform initial distribution is statistically equivalent to considering the probability evolution of a single protein binding to DNA at random site. The simulation is performed on the first 3000 base-pairs of the T7 genome, which contains two recognition sequences GAGTC, at positions 1126 and 1435.

After a sufficiently long time, the protein distribution $\rho(n)$ spreads out, as shown in the inset of Figure 5, and shows a series of peaks corresponding to the sites with larger occupancy. Where the border effects can be neglected, this distribution tends to its equilibrium limit; this is shown in Figure 5, where we plot a portion of

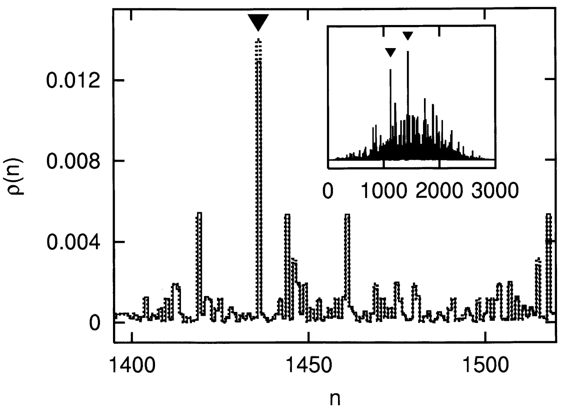

Figure 5. A central portion of the protein distribution $\rho(n)$ for model I after an integration time of 10^6 integration steps, obtained by averaging over 3×10^4 particles initially uniformly distributed in the interval [1000; 2000] (solid line). The analytical equilibrium distribution $\rho_\infty(n)$, (dotted line) is shown for comparison. Here $\epsilon/k_BT = 0.5$. Inset: the whole distribution at the same time. In both plots, the arrows indicate the location of the recognition sequences GAGTC (sites 1126 and 1435).

the distribution obtained after 10^6 time steps for model I , together with ρ_∞. As expected, the larger peaks correspond to energy minima, i.e., to the location of the two recognition sequences GAGTC present in this DNA region. For all the models I to III the final distribution is similar, with the two highest peaks exactly in correspondence to the two recognition sequences, this confirming that the energy landscape defined on the basis of the pattern matching actually guides the protein to the target recognition sequences.

Note that, in case of model IV , the distribution of levels is different, this obviously implying a different shape for $\rho_\infty(n)$. In particular, the case of a sufficiently low threshold energy is reflected on an asymptotic distribution with rarer, larger peaks on a very low constant background (data not shown).

We now investigate the dynamical behavior of the four models, and check if there are some relevant deviations from random walk, induced by the sequence sensitivity. For large enough values of ϵ/k_BT, some positions along DNA could trap proteins for long time, this implying that, at small and intermediate time, diffusion could be substantially different than for a pure random walk. In order to estimate this effect, we calculate the mean square displacement for the protein:

$$\langle \Delta n^2 \rangle = \langle \Delta n^2(t) \rangle = \sum_{i=1}^{n} (n_i(t) - n_i(0))^2. \tag{8}$$

We average over $N = 9\,10^3$ independent particles, initially distributed uniformly in the DNA region [1000, 2000]. This procedure therefore includes both average on

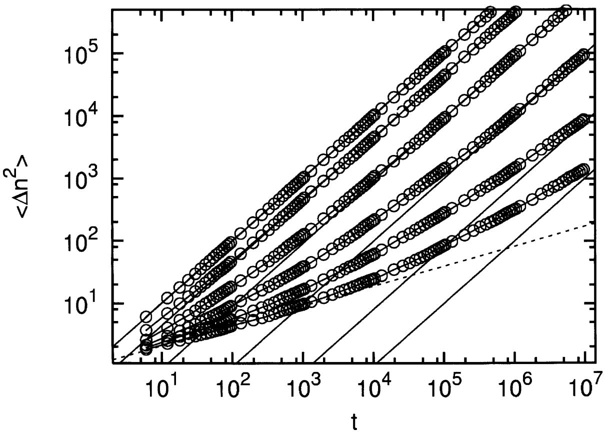

Figure 6. Diffusion behavior of model I for different values of ϵ/k_BT. From the upper curve to the bottom: $\epsilon/k_BT = 0$, 0.3, 0.6, 0.9, 1.2, 1.5. Note the log-log scale: a linear diffusion $\langle\Delta n^2\rangle \propto t$ corresponds in this graph to the straight lines of unit slope (*solid lines*), while slopes lower than 1 correspond to $\langle\Delta n^2\rangle = A\ t^b$, with $b < 1$. The time is measured in units of $\tau \approx 10^{-7}$ seconds, see the discussion after Eq.(6), and the distance is measured in base pair steps (1bps $\approx$ 3.4 Å). So, the unit on the horizontal axis correspond to 10^{-7} seconds, and on the vertical axis to $\approx 11.6 * 10^{-16}$ cm^2. A (dashed) line of slope 0.3 is reported for comparison.

a large number of particles and on a large set of initial conditions. For simplicity, in the following we shall set $\tau = 1$ in (6), i.e., the elapsed time will be measured in the units of τ.

Starting from model I, we investigate the dependence of the diffusive behavior on $\epsilon/\mathrm{k_B T}$. Results are shown in Figure 6. In the limit of $\epsilon/k_BT = 0$, i.e., in the case of a flat potential (or $T = \infty$), the diffusion is of course normal, with $D = 1/2$ and $\langle\Delta n^2(t)\rangle = t$, so that the corresponding curve is a straight line of slope 1 in the log-log plot (upper curve on Figure 6). For larger values of $\epsilon/\mathrm{k_B T}$ (smaller temperatures compared with the energy fluctuations), the dynamics of the model shows at short time large deviations from the normal diffusion: in these finite temperature cases, the motion is initially sub-diffusive, with, locally,

$$\langle\Delta n^2\rangle = A\,t^b, \quad b < 1. \tag{9}$$

The exponent b increases monotonically with time toward its asymptotic value 1. The initial deviation $(1 - b)$ and the crossover to $b = 1$ both increase with $\epsilon/k_BT = 0$. This behavior does not depend on the choice of the initial condition and it is not a transient induced by some $t = 0$ properties: we have verified indeed that qualitatively the same time dependence is reproduced after an initial transient time of 10^4, 10^5 or 10^6 time steps. As expected, once the normal diffusion regime is reached, different temperatures correspond to different diffusion constants D

(in the log-log representation, $2D$ corresponds to the vertical offset of the lines of slope 1, according to the relation $\log \langle \Delta n^2 \rangle = \log 2D + \log t$).

It is evident from Figure 6 that the overall behavior observed for model I can be fitted neither with a linear law $f(t) = 2Dt$ (a line of slope one on the log-log graph), nor with a single $f(t) = At^b$ law (a slope b line). The same is true, as we will see, for the other models as long as the parameter $\epsilon / k_B T$ is large enough. To be more quantitative, one can try to fit with these two laws the whole set of data and check the standard deviation of the fit $\sigma = \frac{1}{N}\sqrt{\sum(\langle \Delta n^2 \rangle - f(t))}$, N being the number of degrees of freedom. For model I e.g., with $\epsilon / k_B T = 1$, we obtain for both laws a value of σ larger than 100, and the fit hypotheses should be rejected, as expected.

Plots of Figure 6 also give a measure of the slowing down in the target search induced by the sequence-dependent interaction. Indeed, in the log-log plot the horizontal offset between different curves, at a given Δn^2, corresponds to the logarithm of the ratio between the time needed to cross the corresponding displacement Δn for different choices of $\epsilon/\mathrm{k_B T}$. Therefore, if Δn is a typical distance to target, the horizontal offset just gives the slowing factor induced by sub-diffusion with respect to normal diffusion. Referring to Figure 6, we can conclude that, if the distance to target is larger than 100 bps (so that $\Delta n^2 > 10^4$), then the time to reach the target should be reduced with respect to standard diffusion roughly by a factor 10 for the case $\epsilon / k_B T = 0.6$, or by a factor 100 for $\epsilon / k_B T = 0.9$. Furthermore, this slowing factor does not depend on Δn, provided that it is large enough to consider the asymptotic regime. In this hypothesis, it is possible to obtain an analytical estimation of the slowing factor [30].

We will now extend the diffusion analysis to the other versions of the model, introduced in Section 1. Resulting curves for models I to IV and for $\epsilon / k_B T = 1$ are presented on Figure 7. As for model I, in all cases we observe a sub-diffusive regime at short times due to the trapping effect of the rough energy landscape.

The initial values of b, fitted in the time range (0, 100) through the function At^b, are the following for the first three models:

$$\begin{aligned} \mathrm{I} &: b = 0.49 \pm 1\% \\ \mathrm{II} &: b = 0.61 \pm 1\% \\ \mathrm{III} &: b = 0.56 \pm 1\%. \end{aligned}$$

A check of the standard deviation for the three fits always gives $\alpha < 0.06$, this implying a very good agreement with the assumed power law in the considered time interval.

Let us remark that, in principle, the obtained anomalous diffusion could be due to some particular spatial correlation properties of the underlying potential. Nevertheless, we have checked that it is only due to the roughness of the landscape, doing the same experiment on an artificial base sequence, completely random. In the conditions described by model I, for instance, and in the same fit range, we obtained $b = 0.52 \pm 1\%$, and a curve similar to the real DNA case (data not shown).

We then studied the behavior of the short time sub-diffusive exponent b as a function of E_t for model III with varying threshold (i.e., including model I and II). The results are shown on Figure 8. For threshold lower that a critical value of about -3ϵ the system displays almost no sensitivity to the threshold level. Indeed, this is

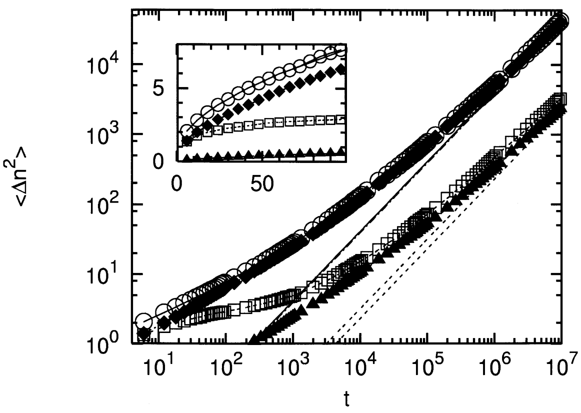

Figure 7. Mean square deviation $\langle \Delta n^2 \rangle$ for the four different models, with $\epsilon / k_B T = 1$ and $E_t = 0$, in the log-log representation. Symbols refer respectively to: open circles, model I; triangles, model II; diamonds, model III; squares, model IV ($E_t = 0$). The straight lines correspond to the fit in the last part of the graphs ($t \in [6\,10^6, 10^7]$). For the definition of the units, refer to Figure 6. Inset: the same curves in a linear representation in the short time regime (symbols have the same meaning).

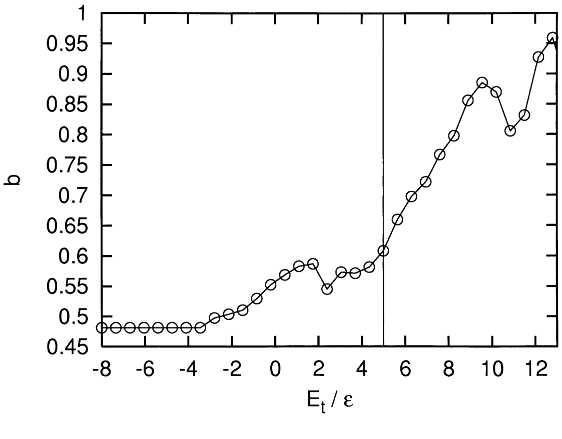

Figure 8. Behavior of the exponent b as fitted in the short time regime $t \in (0, 100)$ as a function of the threshold energy E_t for model III. The vertical line corresponds to $\max[E(n)] = 5\epsilon$.

due to the fact that it is necessary to have not only a site n with $E(n) < E_t$, but at least two neighboring sites should be below the threshold in order to feel its effect (see Eq. (4)). The probability of finding two adjacent sites below the threshold is too low if $E_t \leq -3\epsilon$, thus explaining the observed insensitivity. Interestingly, the exponent b becomes a non-monotonic and very sensitive function of E_t for larger values of E_t. The effect of the threshold in this intermediate regime is in fact twofold: from one side, it induces an additional damping on many low energy sites; from the other, it makes (a fraction of) these same sites "blind" to the energies of their neighbors (the translocation barriers only will depend on $E(n)$ and E_t). The complex balance between the two contributions induces the high instability of the fit results displayed in Figure 8. As the threshold increases above the maximum level ($E_t = 5\epsilon$), the disorder of the underlying energy landscape becomes less and less important, and the system tends to recover a standard diffusive behavior strongly damped, i.e., with $b \to 1$ and $A \to 0$.

Now let us consider the large time limit. The asymptotic diffusion constant depends on the model choice. A linear fit of the large time regime of $\langle \Delta n^2 \rangle$ of Figure 7 has been done in order to estimate the average diffusion constant D, in the random walk approximation where $\langle \Delta n^2 \rangle = 2Dt$. Besides, we checked that an effective linear behavior is reached in the corresponding time range by fitting again with a function $\langle \Delta n^2 \rangle = A\ t^b$ and verifying that b is close to unity. The resulting diffusion constants D and the exponents b for the four models at large time ($t \in [6, 10^6, 10^7]$) are given, for $E_t = 0$, respectively by:

$$\begin{aligned} \text{I} : 2D &= 4.1\ 10^{-3} \pm 1\% & b &= 0.93 \pm 1\% \\ \text{II} : 2D &= 0.23\ 10^{-3} \pm 2\% & b &= 0.86 \pm 1\% \\ \text{III} : 2D &= 4.0\ 10^{-3} \pm 1\% & b &= 0.91 \pm 1\% \\ \text{IV} : 2D &= 0.32\ 10^{-3} \pm 1\% & b &= 0.85 \pm 1\%. \end{aligned} \tag{10}$$

The corresponding fits are the straight lines in Figure 7. We notice that the values of b obtained in this time interval are still smaller than their asymptotic limit of 1. The linear fits are therefore, in these last cases, less good, also due to the much larger time interval used: the $\langle \Delta n^2 \rangle$ behavior is indeed still changing during this time toward its asymptotic regime.

The differences in the equilibrium diffusion constant between different models are explicitly related to the activation barrier in the four cases: the higher is the threshold to overcome in order to move one step, the lower is the diffusion constant. Note that in the case of model IV the boundaries between flat and rough regions act as energy barriers of amplitude $\approx E_{\text{sl}}$: these barriers appear to affect the motion at large times more strongly than the threshold E_t, this resulting in a diffusion constant closer to that of model II than to that of model III.

We have analyzed the dependence on E_t also for the asymptotic diffusion constant D. Figure 9 shows[2] the dependence of D on E_t in model III. Again, almost no

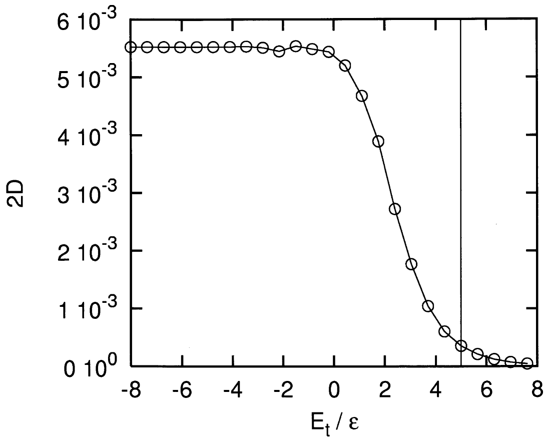

Figure 9. Behavior of the coefficient $2D$ as fitted in the large time regime $t \in (8\,10^5, 10^6)$ as a function of the threshold energy E_t for model III. The level $\max[E(n)] = 5\epsilon$ is represented by a vertical line. D is measured in the units of $(\text{bps})^2/\tau \approx 11.6 * 10^{-9}$ cm^2/seconds.

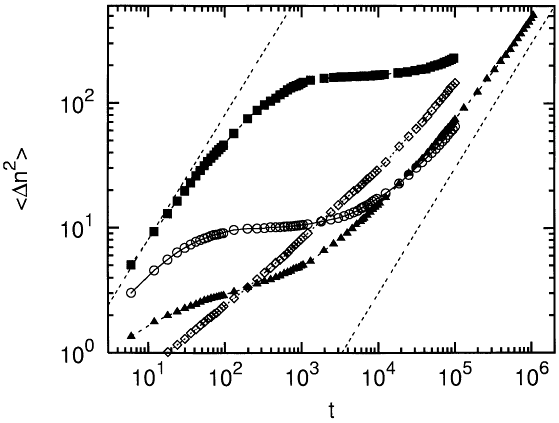

Figure 10. Time behavior of $\langle \Delta n^2 \rangle$ for model IV, in the cases $E_t = -4$ (fullsquares), $E_t = -2$ (circles), $E_t = 0$ (full triangles), $E_t = 2$ (diamonds), and with $\epsilon/k_BT = 1$. For the definition of the units, refer to Figure 6. Two straight lines of slope 1 are shown for comparison.

sensitivity to the threshold level is observed below a critical value, approximatively $E_t = -3\epsilon$. Roughly, between this value and $E_t = 0$, we observe a transition to a regime of strong sensitivity ($E_t > 0$), where the damping effect induced by the threshold is much more enhanced. The diffusion constant decreases rapidly above the maximal energy ($E_M = 5\epsilon$, vertical line), as intuitively expected.

We shall now discuss in detail model IV, since it displays, with respect to the others, a more complicated behavior. Note that, in principle, model IV can be put exactly in the same scheme as the other models, once the underlying potential $E(n)$ is redefined according to Eq. (5). Nevertheless, this redefinition of the energy landscape leads to substantially different features. As can be observed in Figure 10, during an initial time interval the protein diffuses more rapidly, even if still subdiffusively, with initially a larger effective diffusion constant. The initial speeding up of the dynamics becomes more pronounced as the value of the threshold decreases, i.e., as the energy redefinition involves an increasing number of sites. This effect can be explained by considering how the potential landscape is changed for model IV. Among the particles, uniformly distributed at time zero over a large region of the sequence, all those that are initially on flat regions of energy $E_{\rm sl}$ will start diffusing freely with diffusion constant equal to 1, until they fall down in one $E < E_t$ region. These particles contribute initially to the diffusion with a large term, thus making it increase. After an initial transient, however, most of the particles will be almost trapped in the potential wells, and the effective diffusion coefficient will decrease accordingly.

More precisely, the trapping effect will depend on the value of $E_{\rm sl}$, set to max $[E(n)]$ in our calculations. If $E_{\rm sl}$ is big enough, most of the particles will be trapped in $E(n) < E_t$ regions, with activation barriers and only a small probability to escape again toward the flat plateaus. Therefore, in the long time regime, the system will be essentially in the same state as model III, but mostly localized in some finite regions. In other words, the particular equilibrium conditions introduced in model IV are indeed such that one particle needs to spend a large amount of energy (and, therefore, of time) before reaching a high level plateau, but once reached, it can move much faster to the next favorable site. An analytical derivation of the main dynamical quantities as functions of the model parameters discussed in this section will be presented elsewhere [30].

4. Discussion

All the results presented in this work can be checked by a comparison with detailed experimental data. Biochemical experiments have demonstrated diffusion along DNA for several types of enzymes: lac repressor [31], restriction endonuclease (EcoRI [32, 33, 22], EcoRV [34, 35]), methyl transferase (EcoRI [4]). Experiments leading to a rather precise determination of the *E. Coli* RNA-polymerase position along DNA at different times during the promoter search have also appeared [8–11]. No evidence have been presented for T7 RNA-polymerase sliding, but single molecule experiments on this protein are underway in several laboratories [36–38]. These series of experiments will give rapidly, and for the first time, the possibility to estimate the detailed features of protein diffusive motion. As we have shown, a dynamical model which includes both the affinity for the specific site together with the possibility of sliding, leads to a nontrivial sequence dependent dynamics. It is thus important to verify if these effects can actually be observed experimentally.

The sliding distance for RNA-polymerases have been kinetically evaluated in different experiments around 350–1000 bps [5]. Other enzymes also seem to slide along the DNA covering a short distance of about 300 bps before being released in solution [12]. In this space scale, the anomalous diffusion behavior is predominant for our model. Obviously, the model is too simplified to account for the whole possible interaction occurring i.e. in the case of the multimeric *E. Coli* RNA-polymerase, for which subunit β, β' and σ contact the DNA during promoter search and recognition [39], or in the case of lac repressor, which can bind simultaneously to two DNA regions [40].

Nevertheless, it is interesting, in particular, to compare our results with the recent scanning force microscope (SFM) experiment, performed by Guthold *et al.* [11]. The experiment allows for a direct observation of one *E. Coli* RNA-polymerase sliding back and forth on a single DNA chain partially adsorbed on a mica surface, although with some technical limitations (the average lifetime of the nonspecific complex is more than hundred times larger than what measured in solution, probably due to the two-dimensional constraint). The statistical properties of the observed diffusive motion have been fitted by the law $\langle \Delta x^2 \rangle = 2Dt$, in order to confirm the general assumption that RNA-polymerase moves randomly along DNA ([11], Figure 2). Quantitatively, however, in the observed displacement ranges (less than 200 base-pairs), the corresponding data seem to deviate from a pure diffusive motion. This may be due to the experimental constraints and to the limited number of RNA-polymerase sliding trajectories (about 30, with 9 values each). On the other hand, a rough fit of numerical data from Figure 2 of [11] with a power law of the type At^b, gives $b \sim 0.5 \pm 15\%$. We obtain a standard deviation σ for this fit of about 4, to be compared with $\sigma \approx 8$ obtained by fitting with a standard $\langle \Delta x^2 \rangle = 2D\,t$ law. Interestingly, the value of σ for the power law fit can be reduced further (to about 2) if the last point in Figure 2 of [11] is neglected, while it stays practically constant for the normal law fit. We observe that this last point accounts for the detachment of the polymerase from the DNA by setting its position at the end of the DNA chain, which seems a somehow improper statement. If therefore this last data can be reasonably neglected, the anomalous diffusion fit results particularly better than the standard one. It is very interesting to note that these data seem much more compatible with a sub-diffusive behavior than with normal diffusion, as it is usually assumed. This first experiment allowing for a direct visualization of a RNA-polymerase sliding motion gives therefore, from our point of view, intriguing and encouraging results.

We remark that the dynamical features described here depend crucially on the choice of the model parameters: the ratio $\epsilon/k_B T$, the value of the energy threshold E_t, and, in the case of model IV , the energy of the plateaus E_{sl}. As a first check, we can try to compare our rough estimation of the power exponent we extrapolate from the results in [11] with the behavior of the model as a function of $\epsilon/k_B T$. The value of about 0.5 very roughly corresponds to $\epsilon/k_B T \approx 1$ for all values of E_t, this confirming that the parameter choice made in the most part of our simulations could be indeed of the right order of magnitude.

Further experimental investigations, devoted to the detailed determination of the nonspecific interaction, are necessary to improve the model. The version of the model which is compatible with the sliding dynamics of single molecule experiments should emerge from the comparison with the experimental data, using the model parameters as fitting parameters. In practice, the complicated diffusive behavior of the model will allow us to compare theory and experiments by means of more than one dynamical observable. For the case of model IV, where the additional model parameter E_{sl} is needed, the presence of a new short time specific feature could also be used in the fit of the experimental results.

From a biological point of view, the four models offer a framework for defining the pertinent parameters to optimize the target search. For all models, the specific interaction energy ϵ between protein and DNA is crucial and should be close to $k_B T$ in order to allow the protein to move. This adjustment of the interaction energy can be achieved by varying the distance and angle of the hydrogen bonds during sliding or including the effect of the solvent. Perhaps, the more interesting model from a biological point of view is model *IV*, since it allows for a better control of the diffusion pattern, and consequently for the corresponding biological function. An exact balance has to be found in biological system between the reading and sliding mode. E_t, E_{sl}, and $\epsilon/\mathrm{k_B T}$ have to be optimized for the biological purpose which will be physically reflected by the protein-DNA interaction and by the DNA sequence.

Finally, it is important to keep in mind that the recognition mechanism through hydrogen bonds considered here does not allow for a complete identification of the target sites. For the considered case of the T 7 genome, e.g., the recognition sequence GAGTC (or its complementary sequence) appears more than 90 times; however, only 10 of them actually belongs to real promoters. T7 RNA-polymerase recognizes, in fact, a longer sequence, that extends from -17 to $+6$ relatively to the initiation site and consists of two functional domains [41]. Not all these 23 bps have the same importance in the recognition mechanism [42, 27]. Anyway, it is evident that other “signals” must cooperate with the direct core recognition mechanism in order to allow the polymerase to find its target. The weak sequence TAATA (positions -13 to -17), for instance, also interacts with RNA-polymerase through the minor groove [26]. A sensitivity to this minor groove region should probably be included. In this sense, our model can represent a first attempt toward a detailed description of the T7 RNA-polymerase dynamics during the promoter search. The model can also be reformulated for the case of other enzymes by a detailed introduction of their sequence-dependent interaction with nonspecific DNA. We believe that the main idea of the model, which is the link between base sequence and protein dynamics, will be valid in general. Indeed, as far as a sequence dependence is considered, the protein will always interact with DNA through an effective potential with a fluctuating profile. This potential should be induced for different proteins by different kinds of interaction. Its roughness by itself, however, will always generate anomalous diffusion features as those described in this paper. The generality of the subdiffusive behavior is indeed confirmed by the fact that similar results are obtained starting with an artificial random DNA sequence (see

Section 2); furthermore, even completely different diffusion models give analogous qualitative results, provided that a local trapping mechanism is introduced [30].

5. Conclusions

In this paper we have proposed a simple model for the protein sliding motion along DNA, which includes a sequence dependent interaction. Four possible variations of the model were included by considering slightly different translocation probabilities, i.e., by the presence of a varying activation barrier E_t (leading to models I to III), and eventually by distinguishing "reading" regions from "sliding" regions, where no hydrogen bonds are made so that the protein can freely diffuse on an effective constant potential (model IV).

A numerical study of the diffusion properties of the four versions of the model shows that a normal diffusion regime is only achieved after some time. We have shown that all the four models are characterized at shorter times by a sub-diffusive behavior. A rough estimation of the slowing factor induced by the sequence dependence for different values of the energy parameter can be easily obtained. This result is of particular interest because, as we have discussed, the anomalous diffusion is observed in a range that corresponds approximately to the experimentally observed characteristic distance covered by proteins during sliding [5]. The physical reasons underlying the different diffusion behavior have been discussed.

Nowadays the existing nano-technologies and single molecule techniques allow for constraining and manipulating single biological objects. The present paper represents a first step toward theoretical picture where some of the resulting experimental results could be analyzed and connected with the known functional properties of the corresponding biological systems. It is important to keep in mind, anyway, that the in vivo dynamics of the corresponding biological processes occurs in a high density environment, in presence of very complex spatial structures and of water molecules mainly bound and structured [43]. What we usually call the diffusive motion of proteins inside the cell is likely to be instead a motion strongly depending on a complex set of environmental trapping sites, as in the case considered here. Also in this respect, the approach proposed in this paper may have a larger range of application.

Acknowledgements

We are grateful to A. Lesne, M. Peyrard and S. Ruffo for helpful discussions. M.B. wishes to thank the EU and the Physics Department of the University of Salerno, Italy, and the European Community, for a two years post-doctoral research grant during which this work was done. V.P. acknowledges the Physics Department of the University of Salerno for financial support of two short-term visits during which part of this work was done. M.S. acknowledges support from the European grant LOCNET n.o HPRN-CT-1999-00163.

Notes

1. Also note that, for the case of CRP protein, nonspecific binding have been proposed to mimic the specific, c-Amp-dependent binding [23], this confirming the hypothesis of a continuity between nonspecific and specific recognition interaction.
2. For technical reasons, we display data resulting from the fit in the range (8 10^5, 10^6), i.e., in a region where the parameter *b* has not yet reached unity. The curve of Figure 9 represents therefore only a qualitative analysis and shows some small discrepancy with data given in Eq. (10).

Appendix. Markovian Assumption and the Hopping Rates

By following the approach of Schurr [29], it is possible to obtain an estimate of the energy dissipated per unit step just by the effect of friction hydrodynamical forces. Schurr includes in its description the effect of the helical trajectory actually followed by a protein sliding along the DNA major groove; using the Stokes–Einstein relation, he then obtains for the one dimensional diffusion constant the estimation

$$D_1 = \frac{kT}{f_{\rm ef}}\,, \tag{11}$$

depending on an effective friction parameter, $f_{\rm ef} = F/v$. The diffusion constant for sliding particles is known to be approximately of the order of $D_1 = 10^{-13}\,\mathrm{m}^2/\mathrm{s}$. Note then that $f_{\rm ef}$ is related to the energy dissipation rate *over unit distance*, by the following rule:

$$\frac{\Delta E}{\Delta x} = F = \frac{kT}{D_1} \times v \approx kT\,10^{13}\,\mathrm{s/m}^2 \times v. \tag{12}$$

The speed v (within the single step translocation) is $v = \ell/\tau$, with $\tau \approx 10^{-7}$s and $\ell = 3.4\,10^{-10}$ m the inter base-pair distance: one can thus easily derive that

$$\frac{\Delta E}{\Delta x} \approx 10^{13} \times 10^{7}\; kT\;\ell \approx 10\frac{kT}{\ell}. \tag{13}$$

This means that the protein looses all its kinetic energy (of the order of *kT*) before having reached the first neighboring position. This implies that the memory is lost too, or in other words, that the particle velocity decorrelates in less than one step.

For this reason we choose to describe the protein diffusion along DNA as a discrete Markov process. Once having established this first assumption, we have to address the question of how to define the forward and backward hopping rates. At this stage, the dislocation process can be described as a chemical reaction of the type

$$\mathrm{Prot}\cdot \mathrm{DNA}_n + \mathrm{DNA}_{n'} \rightarrow \mathrm{DNA}_n + \mathrm{Prot}\,\cdot\,\mathrm{DNA}_{n'};$$

for what one can reasonably use a speed constant k as usually given in biochemistry handbooks,

$$k = A\,\exp(-\Delta G/RT)\quad(\mathrm{s}^{-1}), \tag{14}$$

or, equivalently, the hopping rates as defined in our work.

References

1. Riggs, A.D., Bourgeois, S. and Cohn, M.: The Lac Repressor-Operator Interaction. 3 Kinetic Studies. *J. Mol. Biol.*, **53** (1970), 401.
2. Berg, O.G., Winter, R.B. and von Hippel, P.H.: Diffusion-Driven Mechanisms of Protein Translocation on Nucleic Acids. 1. Models and Theory. *Biochemistry*, **20** (1981), 6929.
3. Reich, N.O. and Mashhoon, N.: Kinetic Mechanism of the EcoRI DNA Methyltransferase. *Biochemistry*, **30** (1991), 2933.
4. Surby, M.A. and Reich, N.O.: Contribution of Facilitated Diffusion and Processive Catalysis to Enzyme Efficiency: Implications for the EcoRI restriction-modification system. *Biochemistry*, **35** (1996), 2201.
5. Shimamoto, N.: One Dimensional Diffusion of Proteins Along DNA: Its biological and chemical significance revealed by single-molecule measurements. *J. Biol. Chem.*, **274** (1999), 15293.
6. von Hippel P.H. and Berg, O.G.: Facilitated Target Location in Biological Systems. *J. Biol. Chem.*, **264** (1989), 675.
7. von Hippel, P.H., Rees, W.A., Rippe, K. and Wilson, K.S.: Specificity Mechanisms in the Control of Transcription. *Biophys. Chem.*, **59** (1996), 231.
8. Park, C.S., Wu, F.Y. and Wu, C.W.: Molecular Mechanism of Promoter Selection in Gene Transcription. II. *J. Biol. Chem.*, **257** (1982), 6950.
9. Kabata, H., Kurosawa, O., Arai, I., Washizu, M., Margarson, S.A., Glass, R.E. and Shimamoto, N.: Visualisation of Single Molecules of RNA Polymerase Sliding along DNA. *Science*, **262** (1993), 1561.
10. Harada, Y., Funatsu, T., Murakami, K., Nonoyama, Y., Ishihama, A. and Yanagida, T.: Single-Molecule Imaging of RNA Polymerase-DNA Interactions in real time. *Biophys. J.*, **76** (1999), 709.
11. Guthold, M., Zhu, X., Rivetti, C., Yang, G., Thomson, N.H., Kasas, S., Hansma, H.G., Smith, B., Hansma, N.K. and Bustamante, C.: Direct Observation of One-Dimensional Diffusion and Transcription by Escherichia Coli RNA polymerase. *Biophys. J.*, **77** (1999), 2284.
12. Stanford, N.P., Szczelkun, M.D., Marko, J.F. and Halford, S.E.: One-and Three-Dimensional Pathways for Proteins to Reach Specific DNA Sites. *EMBO J.*, **19** (2000), 6546.
13. Salerno, M.: Discrete Model for DNA Promoters Dynamics. *Phys. Rev. A*, **44** (1991), 5292.
14. Salerno, M.: Nonlinear Dynamics of Plasmid pbr322 Promoter, in M. Peyrard (ed.), *Nonlinear Excitations in Biomolecules*, Edition de Physique, Springer, New York (1995) p. 147.
15. Jlicher F. and Bruinsma, R.: Motion of RNA Polymerase Along DNA: A stochastic model. *Biophys. J.*, **74** (1997), 1169.
16. Seeman, N.C., Rosenberg, J.M. and Rich, A.: Sequence-Specific Recognition of Double Helical Nucleic Acids by Proteins. *Proc. Natl. Acad. Sci. USA*, **73** (1976), 804.
17. Travers, A.: *DNA-Protein Interactions, ch.* 3 and 4. Chapman and Hall, London, 1993.
18. Spolar R.S. and Jr. Record, M.T.: Coupling of Local Folding to Site-Specific Binding of Proteins to DNA. *Science*, **263** (1994), 777.
19. deHaseth, P.L., Lohman, T.M. and Jr. Record, M.T.: Nonspecific Interaction of Lac Repressor with DNA: An Association Reaction Driven by Counterion Release. *Biochemistry*, **16** (1977), 4783.
20. Sidorova N.Y. and Rau, D.C.: Linkage of EcoRI Dissociation from its Specific DNA Recognition Site to Water Activity, salt concentration, and pH: Separating their roles in specific and non-specific binding. *J. Mol. Biol.*, **310** (2001), 801.
21. Singer P.T. and Wu, C.W.: Kinetics of Promoter Search by Escherichia Coli RNA Polymerase. Effects of Monovalent and Divalent Cations and Temperature. *J. Biol. Chem.*, **263** (1988), 4208.

22. Jeltsch, A., Alves, J., Wolfes, H., Maass, G. and Pingoud, A.: Pausing of the Restriction Endonuclease EcoRI During Linear Diffusion on DNA. *Biochemistry*, **33** (1994), 10215.
23. Katouzian-Safadi, M., Blazy, B., Cremet, J.Y., Le Caer, J.P., Rossier, J. and Charlier, M.: Photo-Cross-Linking of CRP to Nonspecific DNA in the Absence of cAMP. DNA Interacts with both the N- and c-Terminal Parts of the Protein. *Biochemistry*, **32** (1993), 1770.
24. Schick C. and Martin, C.T.: Tests of a Model of Specific Contacts in T7 RNA Polymerase-Promoter Interactions. *Biochemistry*, **34** (1995), 666.
25. Li, T., Hung Ho, H., Maslak, M., Schick, C. and Martin, C.T.: Major Groove Recognition Elements in the Middle of the T7 RNA Polymerase Promoter. *Biochemistry*, **35** (1996), 3722.
26. Cheetam, G.T., Jeruzalemi, D. and Steitz, T.A.: Structural Basis for Initiation of Transcription from an RNA Polymerase-Promoter Complex. *Nature*, **399** (1999), 80.
27. Imburgio, D., Rong, M., Ma, K. and McAllister, W.T.: Studies of Promoter Recognition and Start Site Selection by T7 RNA Polymerase using a Comprehensive Collection of Promoter Variants. *Biochemistry*, **39** (2000), 10419.
28. Nadassy, K., Wodak, S.J. and Janin, J.: Structural Features of Protein-Nucleic Acid Recognition Sites. *Biochemistry*, **38** (1999), 1999.
29. Schurr, J.M.: The One-Dimensional Diffusion Coefficient of Proteins Absorbed on DNA; Hydrodynamic Considerations. *Biophys. Chem.*, **9** (1979), 413.
30. Barbi, M., Popkov, V. and Salerno, M.: In preparation, 2004.
31. Winter, R.B., Berg, O.G. and von Hippel, P.H.: Diffusion Driven Mechanisms of Protein Translocation on Nucleic Acids. 3. the Escherichia Coli Lac-Operator Interaction : Kinetic Measurements and Conclusions. *Biochemistry*, **20** (1981), 6961.
32. Jack, W.E., Terry, B.J. and Modrich, P.: Involvement of Outside DNA Sequences in the Major Kinetic Path by Which EcoRI Endonuclease Locates and Leaves its Recognition Sequence. *Proc. Natl. Acad. Sci. USA*, **79** (1982), 4010.
33. Ehbrecht, H.J. Pingoud, A., Urbanke, C., Maass, G. and Gualerzi, C.: Linear Diffusion of Restriction Endonucleases on DNA. *J. Biol. Chem.*, **260** (1985), 6160.
34. Dowd D.R. and Lloyd, R.S.: Biological Significance of Facilitated Diffusion in Protein-DNA Interactions. Applications to t4 Endonuclease V-Initiated DNA Repair. *J. Biol. Chem.*, **265** (1990), 3424.
35. Stanford, N.P., Szczelkun, M.D., Marko, J.F. and Halford, S.E.: Contribution offacilitated Diffusion and Processive Catalysis to Enzyme Efficiency: Implications for the EcoRI Restriction-Modification System. *EMBO J.*, **19** (2000), 6546.
36. Heslot, F.: Oral communication, Meeting “DNA in chromatin,” Arcachon, France. 2002.
37. Baumann, C.G.: Oral Communication, Meeting “DNA in Chromatin,” Arcachon, France. 2002.
38. Place, C.: Personal Communication. See also [44], 2002.
39. Park, C.S., Hillel, Z. and Wu, C.W.: Molecular Mechanism of Promoter Selection in Gene Transcription. I Development of a Rapid Mixing-Photocross Linking Technique to Study the Kinetics of Escherichia Coli RNA Polymerase Binding to T7 DNA. *J. Biol. Chem.*, **257** (1982), 6944.
40. Fried, M.G. and Crothers, D.M.: Kinetics and Mechanism in the Reaction of Gene Regulatory Proteins with DNA. *J. Mol. Biol.*, **172** (1984), 263.
41. McAllister, W.T.: Transcription by T7 RNA Polymerase, in F. Eckstein and D. M. J. Lilley, (eds.), *Mechanisms of Transcription*, Springer, Berlin and Heidelberg (1997) p. 15.
42. Ūjvāri A. and Martin, C.T.: Identification of a Minimal Binding Element within the T7 RNA Polymerase Promoter. *J. Mol. Biol.*, **273** (1997), 775. And References Herein.
43. Goodsell, D.S.: A Look Inside the Living Cell. *Amer. Scientist*, **80** (1992), 457.
44. Gueroui, Z., Place, C., Freyssingeas, E. and Berge, B.: Observation by Fluorescence Microscopy of Transcription on Single Combed DNA. *Proc. Natl. Acad. Sci. USA*, **99** (2002), 6005.