Base-sequence-dependent sliding of proteins on DNA

M. Barbi

Laboratoire de Physique Théorique des Liquides, Université Pierre et Marie Curie, case courrier 121,
4 Place Jussieu, 75252 Paris Cedex 05, France

C. Place

Laboratoire de Physique, CNRS-UMR 5672, École Normale Supérieure de Lyon, Lyon, France

V. Popkov

Institut für Festkörperforschung, Forschungszentrum Jülich, 52425 Jülich, Germany

M. Salerno

Dipartimento di Fisica "E.R. Caianiello" and Istituto Nazionale di Fisica della Materia (INFM), Universitá di Salerno,
I-84081 Baronissi (SA), Italy

The possibility that the sliding motion of proteins on DNA is influenced by the base sequence through a base pair reading interaction, is considered. Referring to the case of the T7 RNA-polymerase, we show that the protein should follow a noise-influenced sequence-dependent motion which deviate from the standard random walk usually assumed. The general validity and the implications of the results are discussed.

I. INTRODUCTION

How site-specific DNA binding proteins locate their targets on DNA is an issue of primary importance for understanding the functioning of DNA. With the development of new experimental techniques, this problem is getting much of attention, see, e.g., [1–8]. Sliding, hopping and uncorrelated three-dimensional diffusion are generally taken into account as possible searching mechanisms, and their relative role in target location is being discussed and experimentally investigated. In the seminal work of Berg, Winter and von Hippel (BWH), one-dimensional diffusion (sliding) along DNA was proposed as a necessary ingredient of the target search [9]. More recent papers [4–6] confirm the importance of sliding in the search process, along with three-dimensional paths (disattachment of a protein from DNA and reattachment to a different segment of DNA) [7].

A completely coherent description of the search process is nevertheless still lacking. In a recent paper [4], Bruinsma remarks, e.g., that the time spent by a lac-repressor on each DNA site in the frame of the BWH theory is too short to allow the structural changes necessary for the protein to recognize its target. He thus indicates the need for a slowing down effect and suggests that "indirect read-out" mechanisms, associated to the DNA flexibility, can account for it. Note that the DNA sequence, responsible for the DNA flexibility and shape, is crucial also for this kind of slowing down effect.

On the other hand, all existing models of target search dynamics describe the sliding motion as a *standard random walk*. In theoretical analysis of experiments it is indeed taken for granted that the protein motion is governed by a linear diffusion, $\langle x^2 \rangle = 2Dt$. While the linear diffusion assumption is natural for 3-dimensional paths (when protein is not bound to DNA and diffuses in solution), for the sliding phase of mo-

tion, it implies that the DNA is essentially "seen" by the protein as a homogeneous chain. This homogeneity of DNA, however, seems incompatible with the recognition function, which always involves a form of *reading*, so that it is natural to assume an influence of the DNA sequence on the sliding dynamics [10]. This influence could result in slowing down, pauses and stops which, in turn, could invalidate the random walk assumption. These slowing effects can have have a different origin from that suggested by Bruinsma [4]; note, nevertheless, that different mechanisms can coexist, and that in any case the dynamic effects of (direct or indirect) sequence sensitivity are considered.

Our aim in the present paper is to show that sequence dependence of the DNA-protein interaction can induce strong deviations from standard diffusion for a generic protein sliding on DNA. To this regard, we use a probabilistic model for the sliding motion of a protein on DNA in which the influence of the base sequence is accounted through the DNA-protein reading interaction [11]. As a result we show that the protein follows a noise-influenced sequence-dependent motion which deviates from standard diffusion, reaching normal diffusion only at asymptotically large times. The presence of an anomalous diffusion (AD) regime speeds up the mobility of a protein thus greatly facilitating the target search. The speed-up effect of the anomalous diffusion is shown schematically later in Fig. 4. The cross-over from anomalous to normal diffusion occurs at times typically needed for a protein to cover the distance at which the potential averages out (of order 100 bp in our model). On the other hand, indirect measurements hint on the typical mean path length traversed by the protein during a single DNA binding event, of the same order of magnitude (e.g., around 170 bp in [7]). Thus, the anomalous diffusion (AD) should actually dominate the binding phase, and cannot be neglected. This is the main result of our work. As we will
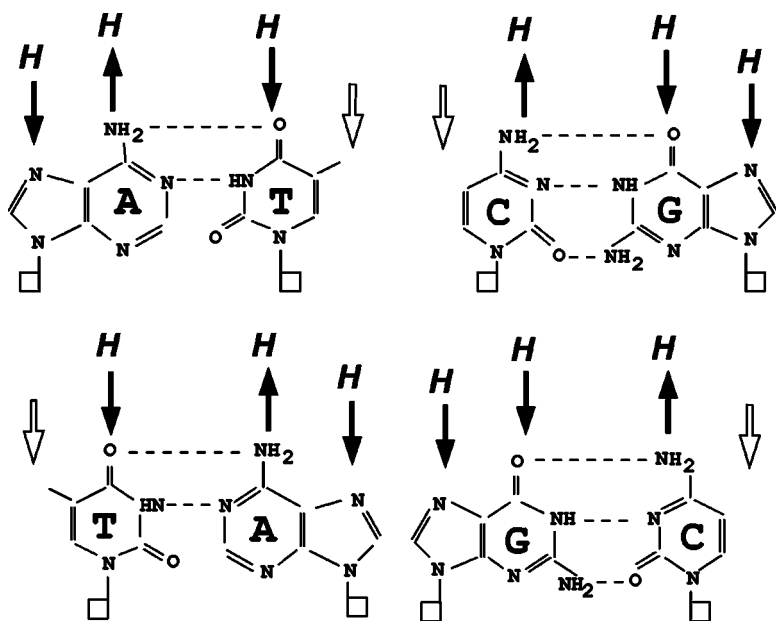
FIG. 1. Four base pairs. Arrows indicate the location of the hydrogen accepting

discuss, the anomalous behavior is very general and only depends, qualitatively, on the presence of a rough sequence-dependent interaction, so that it results to be very robust and not limited to the specific model defined here.

The paper is organized as follows. In Sec. II we introduce the model using T7 RNA-polymerase as a specific example of a sliding protein. In Sec. III we investigate the main properties of the sliding dynamics including the sub-diffusive regime and the crossover to normal diffusion. In Sec. IV we provide some arguments supporting the generality of our results in connection to applications to other enzymes. Finally, in Sec. V, results and conclusions of the paper are summarized.

II. THE MODEL

A target sequence usually consists of few (say, r) consecutive base pairs (bps). Specific sequence recognition is often mediated by hydrogen bonds (H-bonds) to a set of four specific, spatially ordered chemical groups of the bps [12,13]. Each base pair (AT, TA, GC or CG) in DNA exposes indeed, on its major groove side, four specific chemical sites. Among them, some are able to form a hydrogen bond as acceptors, others are hydrogen bonds donors. Other sites are occupied by groups that are not of interest here. The set of chemical sites on each bp provide a specific chemical pattern, allowing at the same time the protein binding and a specific recognition mechanism at the single bp level (see Fig. 1). Besides this mechanism, other features of DNA such as shape and flexibility, as well as electrostatic interactions between protein and DNA [14,15] may also be involved in the recognition process. In this paper, we will focus mainly on the first mechanism, i.e., we assume that proteins check the sequence at each position on DNA by exploiting the same set of hydrogen bonds they form with the DNA at the target site. We thus represent the DNA binding sites at position n as a sequence of r vectors b_n (one for each bp), of the form $B_n = (b_n, b_{n+1}, \ldots, b_{n+r-1})$, according to the rule

$$b_n = \begin{cases} (1,-1,1,0)^T \text{ for AT}, & (0,1,-1,1)^T \text{ for TA}, \\ (1,1,-1,0)^T \text{ for GC}, & (0,-1,1,1)^T \text{ for CG} \end{cases} \quad (1)$$

where $+1$, -1, 0 denote, respectively, an acceptor, a donor, and a missing bond, that each of the four base pairs can form with an external ligand at position n on the DNA [12]. We also assume that the H-bonds formed in the DNA-protein complex at the recognition site are known (this information can be obtained from a crystallographic analysis of the DNA-protein complex). The protein can then be represented by an $(r \times 4)$ *recognition matrix R* describing the pattern of H-bonds formed by the protein and the DNA at the recognition site. The protein-DNA interaction energy is then defined by counting the matching and unmatching bonds between the recognition matrix and the DNA sequence at site n,

$$E(n) = \epsilon \, \mathrm{tr}(R \cdot B_n), \quad (2)$$

where ϵ denotes each H-bond energy, tr the trace, and the dot refers to usual matrix multiplication. The typical H-bond energy is of order of a few kcal/mol, but in fact the actual ϵ could be much less due to screening introduced by the water layer around DNA. In the lack of realistic estimates of this parameter, we will use in the following the dimensionless quantity $\beta\epsilon$ as an adjustable model parameter. The DNA is thus viewed as a one-dimensional vector lattice characterized by a rough on-site potential $E(n)$, on which a random walker (a protein) moves, with rates (probability per unit time)

$$r_{n \to n'} = \min(1/2, 1/2 \exp(-\beta \Delta E_{n \to n'})),$$

$$r_{n \to n} = 1 - r_{n \to n+1} - r_{n \to n-1}; \quad (3)$$

where $n' = n \pm 1$ and $\beta = 1/k_B T$. Time is measured in one-step time units (t.u.). An estimation for the lower bound of the time unit can be obtained through simple hydrodynamic considerations [11,16], yielding 1 t.u. $\approx 10^{-7}$ s.

The presence of an activation barrier for the translocation on neighboring sites can be accounted for by introducing a uniform threshold energy level E_t, so that

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

Note that the effective translocation barrier also depends on the position, through the on-site energy. As a specific example, we consider the case of the T7 RNA-polymerase sliding on the bacteriophage T7 DNA. For this case it is known that the recognition site is the five bps sequence *GAGTC* extending from position −11 to −7 in the *T7* promoter. The interaction matrix R can be inferred from the crystallographic studies of Cheetam *et al.* [17], as

$$R = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 1/2 & 0 & 0 \\ 0 & 0 & 1/2 & 1 \end{pmatrix}, \tag{5}$$

where the presence of $1/2$ is due to one shared DNA-protein H-bond mediated by a water molecule and therefore considered as two half bonds.

It is important to stress that a certain number of assumptions have been made in order to set up our model. Not all of them can be experimentally confirmed, at present. In particular, the additivity of the different H-bond contributions, implicit in Eq. (2) for the binding energy, has to be verified. In the next section we shall present the results for diffusion dynamics, which, qualitatively, are largely independent on the interaction details.

III. THE PROPERTIES: SUBDIFFUSIVITY AND CROSSOVER TO NORMAL DIFFUSION

Theoretically, one can easily calculate the stationary distribution of a population of proteins on the energy landscape as $\rho_\infty(n) \propto \exp(-\beta E(n))$, only dependent on the site energy and on temperature. This implies that the energy minima, that correspond to the recognition sites, will be, on average, the most populated. We then calculate the mean square deviation from the average of the spatial displacement, $\langle \Delta n^2 \rangle = \Sigma_{i=1}^N (n_i(t) - n_i(0))^2$, where the average over initial positions and different histories (Monte-Carlo runs) is made. The three cases $E_t = \min[E(n)] \equiv E_{min}$, $E_t = 0$ and $E_t = \max[E(n)] \equiv E_{max}$ have been examined. In the limit $\beta\epsilon = 0$ the linear diffusion is recovered, as one expects (the limiting value $2D = 1$ is ob-
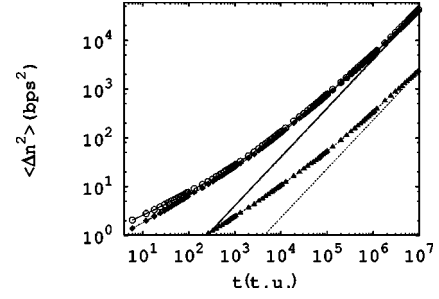

FIG. 2. Diffusion dynamics for the T7 RNA-polymerase-DNA interaction (symbols), for energy parameters: $E_t = E_{min}$, $\beta\epsilon = 0.5$ (squares); $E_t = E_{min}$, $\beta\epsilon = 1$ (open circles); $E_t = E_{max}$, $\beta\epsilon = 1$ (triangles). Time is measured in time units. Solid lines show the dynamics obtained on an artificial Gaussian energy landscape with $E_{min} = -N\epsilon$, $E_{max} \approx N\epsilon/2$ (see the discussion in Sec. IV).

tained in the case $E_t = E_{min}$, i.e., for a flat potential without thresholds). Nevertheless, in the finite temperature case, we obtain large initial deviations from the normal diffusion behavior. More precisely, for all thresholds we find that at the initial stage the diffusion displays anomalous *sub-diffusive* features, with

$$\langle n^2 \rangle = 2At^b, \ b < 1, \tag{6}$$

where A and b depend on the fixed threshold level; see Fig. 2. The appearance of the initial subdiffusive regime is not surprising, and has been observed both for random trap and random barrier potentials; see, e.g., [18]. Our case in Eq. (3), however, represents a mixture of these two cases, for which to our knowledge, there are no studies for the initial time behavior. On the other hand, note that in Eq. (3) the hopping rates $r_{n \to n+1}$, $r_{n \to n-1}$ are not random variables but depend on the gradient of the energy landscape, $\log(r_{n \to n+1}/r_{n+1 \to n}) = (E_{n+1} - E_n)/(k_B T)$. This has the important consequence that in the continuous (Langevin) approximation of the process (see, e.g., [19]), the effective potential U stays Gaussian localized with the typical difference $U(n) - U(n-1) \approx \sqrt{2}\sigma_E$ independent of n, σ_E being the energy variance. This is different from Sinai model where typical $U(n)$ grows with n as $\sqrt{n}$, this leading to anomalous $\langle x^2 \rangle \sim (\ln(t))^4$ behavior. Since the Sinai model is not applicable to our case, we will be using in the following a rather crude approximation (6) to describe the crossover from initial subdiffusion to a linear diffusion regime. A quantitative characterization of the initial transient regime is given in Table I, for the three values of E_t. The diffusion constant D_∞ for the three threshold levels is esti-

TABLE I. The short time sub-diffusive parameters A and b fitted in the initial time interval [0, 100], and those characterizing the asymptotic regime, D_∞ b_∞, fitted in $t \in [8 \times 10^6, 10^7]$. The equilibrium diffusion constant D^* is estimated from the *mfpt* analysis. All values are obtained for $\beta\epsilon = 1$.

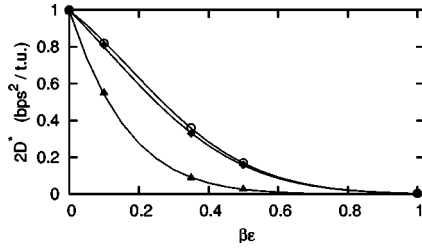| E_t	$2A$	b	$2D_\infty$	b_∞	$2D^*$
E_{min}	$0.82 \pm 2\%$	$0.49 \pm 1\%$	$4.4 \ 10^{-3} \pm 1\%$	$0.94 \pm 1\%$	$4.4 \ 10^{-3}$
0	$0.48 \pm 2\%$	$0.56 \pm 1\%$	$4.3 \ 10^{-3} \pm 1\%$	$0.93 \pm 1\%$	$4.3 \ 10^{-3}$
E_{max}	$0.04 \pm 3\%$	$0.61 \pm 1\%$	$0.25 \ 10^{-3} \pm 2\%$	$0.83 \pm 1\%$	$0.2 \ 10^{-3}$

FIG. 3. $2D^* = \langle(n-n_0)^2\rangle/T_{n_0}^n$ as a function of the adimensional parameter $\beta\epsilon$ (full lines), and the corresponding $2D$ directly evaluated by fitting the large time diffusion (symbols), for corresponding different values of the threshold energy: $E_t = E_{min}$ (open circles), $E_t = 0$ (triangles) and $E_t = E_{max}$ (diamonds). Time is measured in time units (t.u.) and mean square displacement in squared base pairs (bps^2).

mated from the linear fit $\langle\Delta n^2\rangle = 2Dt$ at large times $t \in [8 \times 10^6, 10^7]$. We checked that an effective linear behavior is roughly reached by evaluating the parameter b_∞ in the same range (see Table I). Asymptotically, a standard diffusion is recovered (on the large scale the potential roughness averages to zero). The asymptotic diffusion constant D decreases for increasing $\beta\epsilon$. The initial deviation from a random walk $(1-b)$ and the time needed to reach the asymptotic limit both increase with $\beta\epsilon$; the typical one-step time (or time unit t.u.) should be roughly, for real proteins, of the order of a microsecond [4,16], thus giving crossover times up to seconds corresponding to mean displacements up to hundreds of bps (data not shown; more details will be given elsewhere). A theoretical estimate of the large time effective diffusion constant can be obtained from mean first passage time (mfpt) analysis. According to Ref. [20], for a discrete one step process, such as the one considered here, the mfpt $T_{n_0}^n$ to go from a referring position n_0 to position $n > n_0$ can be evaluated, once a reflecting barrier is fixed in a position $a < n_0$, as

$$T_{n_0}^n = \sum_{i=n0}^n \frac{1}{r_{i\to i-1}\rho_\infty(i)} \sum_{j=a}^i \rho_\infty(j). \qquad (7)$$

Note that $T_{n_0}^n$ depends on the threshold level E_t through the rate $r_{n\to n-1}$, according to Eq. (3). For large enough $T_{n_0}^n$,

$$\langle\Delta n^2\rangle \approx 2DT_{n_0}^n. \qquad (8)$$

Making the choice $a = 0$, the theoretical diffusion constant D^* as a function of $\beta\epsilon$ can be evaluated using Eq. (8). The result is shown in Fig. 3 together with the corresponding numerically evaluated diffusion constants. We observe a good agreement showing that Eqs. (7) and (8) provides a credible analytical estimate of the asymptotic diffusion constant D. Note that the diffusion constant decreases exponentially for $\beta\epsilon \to \infty$ (in practice, it is already $\ll 1$ for $\beta\epsilon \approx 1$) and the corresponding mfpt exponentially increases in the same limit. This behavior reflects the divergence of the typical extent of the sub-diffusive transient, which becomes more and more important as $\beta\epsilon$ approaches 1. At the same time, an increase of $\beta\epsilon$ makes the linear diffusion approximation less and less appropriate. Figure 4 shows qualita-
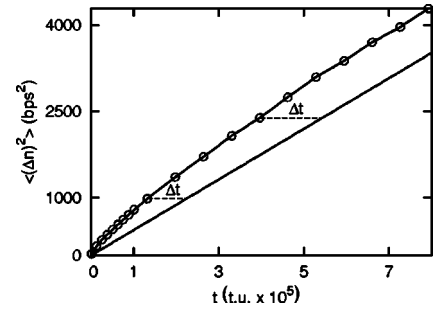

FIG. 4. Dispersion (in bps^2) versus time (in time units t.u.) for our model (upper curve), and for linear diffusion (straight line), for $\beta\epsilon = 1$. The linear diffusion approximation systematically overestimates the time needed for an enzyme to diffuse over a distance $l = \sqrt{(\Delta n)^2}$ base pairs along DNA. Segments between the two curves show differences between the two descriptions. The relative mistake $\Delta t/t$ is $\approx 70\%$ for $l = 30$ and $\approx 37\%$ for $l = 50$.

tively the mistake one makes estimating the time $\Delta t(\beta\epsilon, l)$ needed for an enzyme to cover the distance of l base pairs, induced by the linear diffusion approximation.

Some comments are in order. As one can see from Fig. 4, the relative mistake, $\Delta t(\beta\epsilon, l)/t(\beta\epsilon, l)$ is a monotonically decreasing function of l, diverging for $l \to 0$ and decreasing as $\Delta t_\infty(\beta\epsilon)/l^2$ for times $t = l^2/2D_\infty$ much larger than transition time period, where $\Delta t_\infty(\beta\epsilon) = \lim_{l\to\infty} \Delta t(\beta\epsilon, l)$. Thus, applying a linear diffusion approximation with the diffusion coefficient D_∞ would lead to a systematic overestimation of the time needed for a protein to diffuse along a DNA. Note that protein is likely to attach and disattach many times from the DNA during the target search, each time diffusing over some distance l_i along the DNA so one has to know the distribution of the distances $p(l)$ (or the rate of disattachment) in order to estimate the net effect of initial subdiffusion on a protein efficiency to find its target. Further studies and inputs are needed to clarify this question.

The model allows also us to consider the possibility that very unfavorable positions (with a large number of mismatches) could induce protein conformational changes to a different conformation that does not allow the formation of any H-bond, inducing a regime of "free sliding" [21]. A threshold energy level should in this case separate *reading* regions from *free sliding* regions, where the DNA is seen as homogeneous. The energy landscape should then be redefined above this threshold to a constant value: we will put $E(n) = E_{sl}$ if $E(n) \geqslant E_t$, and refer to this second possibility as "*two-state model*." In this case, the redefinition of the energy landscape leads to a faster diffusion, even if still subdiffusive, in the very initial time. This effect is more evident for low threshold values, i.e., as the energy redefinition involves an increasing number of sites. Indeed, among the many particles initially distributed uniformly over a large region, all those that are on a flat part of the potential will start to diffuse freely with a diffusion constant equal to unity. These particles therefore contribute initially to the diffusion with a large term. After having slid freely for a certain time, however, they will fall in $E < E_t$ regions, and be partially trapped in the potential wells. After a transient time, a subdiffusive behavior similar to the previous case is indeed

reached, that converges, on larger times, to the stationary normal diffusion. A detailed analysis of the "two-state model" is presented elsewhere [11].

Thus, one sees a substantial deviation from the random walk during the sliding phase of a target search. In the next section, we address the question about the generality of the presented results, in application to larger and more complex proteins such as, e.g., E. Coli RNA-polymerase, lac repressor, EcoRI and EcoRV, i.e., for other H-bond reading enzymes.

IV. GENERALIZATION TO OTHER ENZYMES AND BINDING MECHANISMS

First of all, note that the dynamics of the proposed model depends only on the obtained energy profile, and that the most important parameter is the single energy contribution ϵ, that fixes the energy scale. This quantity, though experimentally difficult to access, should nevertheless depend only on the nature of the H-bond: one can thus reasonably expect it to be roughly the same for all proteins. The actual threshold mechanism is also unknown, but again we could reasonably expect that it depends on general properties of the protein-DNA interaction, and does not vary in nature from one protein to another.

What should represent the main difference between different proteins is therefore the length of the recognition sequence [22], or, more precisely, the number of bonds involved in the reading. This parameter should be adapted in order to mimic the sliding of different enzymes.

An examination of the whole set of possible hydrogen bonds that DNA bps can form with external ligands [12,13] shows that, among the 12 possible H-bond sites exposed on the 4 different bps, those that are in central binding sites of the bases ($b_n[2]$ and $b_n[3]$) can induce both matches or mismatches, while the external ones ($b_n[1]$ and $b_n[4]$) are either matches or give zero contribution to the interaction energy. It is thus possible to calculate explicitly the energy level distribution for a generic enzyme looking for a total of N matches with N' of them in the two central binding sites of the bases. The only assumption made is that the matches are uncorrelated, which turns out to be a reasonable approximation for quasi-random DNA sequences. The resulting energy level distribution is a convolution of two binomials that rapidly converges to a Gaussian as N and N' increase [23]. It is then easy to calculate the average and standard deviation of the energy that result to be $\langle E \rangle = (N-N')\epsilon/2$ and $\sigma_E = (N+3N')\epsilon/4$, respectively. The minimum and maximum energies of the resulting distributions are given by $E_{min} = -N\epsilon$, $E_{max} = N'\epsilon$.

This leads us to conclude that, for not too small values of N (and N'), the energy level distribution is approximatively a Gaussian, and its width just depends on N and N' (or, alternatively, E_{min} and E_{max}). Note, furthermore, that if bonds on different positions are equiprobable, N' should be roughly equal to $N/2$, so that one ends with only one parameter. We can expect therefore that the energy landscape for a generic sliding protein, and therefore the sliding motion depends crucially on the number of H-bonds made at the recognition site.
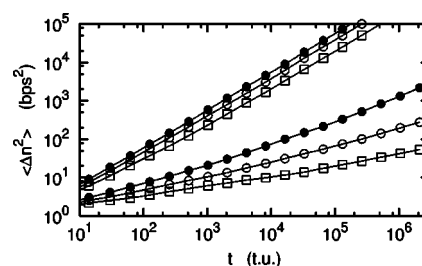

FIG. 5. Dynamic behavior, obtained on the artificial Gaussian energy landscape, for $N=10$ (full circles), $N=14$ (open circles), $N=20$ (squares), with $\beta\epsilon=1$ (upper curves) or 0.2 (lower curves). Time is measured in time units (t.u.) and mean squared distance in squared base pairs (bps^2); see the text for details.

We have tested the previous arguments by building an artificial energy profile, with random levels distributed so as to reproduce the original distribution width and thus the original Gaussian shape. In Fig. 2, simulations of the protein sliding motion on the basis of this artificial energy landscape are compared with previous results for different choices of the model parameters. Despite the certain arbitrariness in the definition of the artificial energy landscape, we obtain essentially the same diffusive behavior as for the true DNA case. In Fig. 5 we depict the diffusive behavior for three different values of N, with $N'=N/2$: as easily predicted, the asymptotic normal diffusion slows down when the number of bonds increases. This parameter thus affects the asymptotic diffusion regime as well as the initial subdiffusion and the transition time.

V. CONCLUSIONS

In this paper we have considered the sliding motion of a protein on DNA by means of a probabilistic model which includes the information about the base sequence through the base pair reading interaction. In the case of the T7 RNA-polymerase we found that the protein executes a random motion which deviates from the standard random walk dynamics usually assumed. We argued that the same qualitative behavior should be valid also for other types of enzymes. The presence of an anomalous diffusion regime at the early stages of the process speeds up the mobility of the protein facilitating the target search. The overall diffusive behavior of the sliding protein can be characterized in terms of few parameters: the typical interaction energy ϵ associated with each DNA-protein bond, and the number N of such bonds formed at the recognition site. We conclude that only few parameters determine the overall diffusive behavior of a sliding protein on DNA: the typical interaction energy ϵ associated with each DNA-protein bond, and the number N of such bonds formed at the recognition site. One can therefore expect the same qualitative behavior described here on the example of T7 RNA-polymerase to be valid also for other types of enzymes (if other kinds of specific chemical bonds intervene in the recognition mechanism, as e.g. water-bridges, minor groove H-bonds or hydrophobic contacts [13,24], the corresponding energies should be evaluated and included in the model; nevertheless, the number of specific bonds is

strictly a characteristic of each different enzyme-DNA interaction, and the diffusing behavior must still depend on this number).

We finally remark that the presence of additional sequence-dependent interaction in the recognition process, such as the one involving geometrical and elastic characteristics of the DNA, can also be included in our model. This additional interaction, being sequence specific, would lead to a redefinition of the energy landscape without affecting much the qualitative results of the paper (they however are much more difficult to model due to the scarcity of experimental data). In particular, the previously discussed anomalous diffusion regime is robust with respect to changes of the energy landscape. Therefore, the influence of the DNA sequence on the sliding motion of a protein on DNA makes the standard random walk assumption for a sliding phase of the target search incorrect for a large set of parameters. Accounting for this anomalous diffusive motion should be included in realistic description of the sliding component of the target search in order to discriminate the relative role of 1D sliding and 3D diffusion in the search process.

[1] M. Guthold, X. Zhu, C. Rivetti, G. Yang, N. H. Thomson, S. Kasas, H. G. Hansma, B. Smith, N. K. Hansma, and C. Bustamante, Biophys. J. **77**, 2284 (1999).

[2] N. Shimamoto, J. Biol. Chem. **274**, 15293 (1999).

[3] U. Gerland, J. Moroz, and T. Hwa, Proc. Natl. Acad. Sci. U.S.A. **99**, 12015 (2002).

[4] R. F. Bruinsma, Physica A **313**, 211 (2002).

[5] N. Stanford, M. Szczelkun, J. Marko, and S. Halford, EMBO J. **19**, 6546 (2000).

[6] S. E. Halford and M. D. Szczelkun, Eur. Biophys. J. **31**, 257 (2002).

[7] D. M. Gowers and S. E. Halford, EMBO J. **22**, 1410 (2003).

[8] M. Slusky and L. A. Mirny, q-bioBM/0402005.

[9] O. G. Berg, R. B. Winter, and P. von Hippel, Biochemistry **20**, 6929 (1981).

[10] In the context of directed translocation of proteins, the importance of sequence heterogeneity is recognized, see, e.g., Y. Kafri, D. K. Lubensky, D. R. Nelson, cond-mat/0310455.

[11] M. Barbi, C. Place, V. Popkov, and M. Salerno, J. Biol. Phys. (in press).

[12] N. C. Seeman, J. M. Rosenberg, and A. Rich, Proc. Natl. Acad. Sci. U.S.A. **73**, 804 (1976).

[13] K. Nadassy, S. J. Wodak, and J. Janin, Biochemistry **38**, 1999 (1999).

[14] S. G. Kamzolova, V. S. Sivozhelezov, A. A. Sorokin, T. R. Dzhelyadin, N. N. Ivanova, and R. V. Polozov, J. Biol. Struct. Dyn. **18**, 325 (2000).

[15] A. Travers, *DNA-Protein Interactions* (Chapman and Hall, London, 1993), Chaps. 3 and 4. ISBN 0-412-25990-7.

[16] J. M. Schurr, Biophys. Chem. **9**, 413 (1979).

[17] G. T. Cheetam, D. Jeruzalemi, and T. A. Steitz, Nature (London) **399**, 80 (1999).

[18] J. Haus and K. Kehr, Phys. Rep. **150**, 263 (1987).

[19] J.-P. Bouchaud and A. Georges, Phys. Rep. **195**, 127 (1990).

[20] N. G. Van Kampen, *Stochastic Processes in Physics and Chemistry* (North-Holland, Amsterdam, 1981), Chap. 11, Eq. (2.14).

[21] P. von Hippel, W. A. Rees, K. Rippe, and K. S. Wilson Biophys. Chem. **59**, 231 (1996).

[22] An interesting discussion on the relevance of the number of base-pairs involved in the recognition mechanism is also presented in Ref. 3.

[23] Detailed calculations will be presented in a future work.

[24] C. G. Kalodimos, A. M. J. J. Bonvin, A. K. Salinas, R. Wechselberger, R. Boelens, and R. Kaptein, EMBO J. **21**, 2866 (2002).