Model of RecA-Mediated Homologous Recognition

Kevin D. Dorfman,[1] Renaud Fulconis,[1] Marie Dutreix,[2] and Jean-Louis Viovy[1,*]

[1]*Laboratoire Physicochimie-Curie, CNRS/UMR 168, Institut Curie, 26 Rue d'Ulm, F-75248 Paris Cedex 5, France*
[2]*Laboratoire Génotoxicologie et Cycle Cellulaire, UMR 2027, Institut Curie, Bâtiment 110 Centre Universitaire, F-91405 Orsay Cedex, France*

(Received 4 March 2004; published 20 December 2004)

We consider theoretically the homology search between a long double-stranded DNA and a RecA-single-stranded DNA nucleofilament, emphasizing the polymeric nature of the search and the ability of double-stranded DNA to overcome the difference in pitch between itself and the nucleofilament by thermally activated stretching from the canonical *B* state to the metastable, stretched *S* state. Our analytical first-passage-time analysis agrees well with experimental data, predicts new dependencies on the intracellular fluid viscosity and ionic strength, and strongly suggests that initial homologous recognition involves a three base-pair seed.

DOI: 10.1103/PhysRevLett.93.268102 PACS numbers: 87.10.+e, 05.40.−a, 87.17.−d

RecA protein is an important part of the homologous recombination machinery in *E. Coli* bacteria. It promotes strand exchange between homologous DNA molecules, i.e., between molecules bearing identical or very similar sequences, allowing the bacteria to repair damaged DNA [1]. A key step in the overall process of homologous recognition is the homology search, where a RecA-single-stranded DNA nucleofilament (RecA-ssDNA) finds its homologous double-stranded DNA (dsDNA) counterpart and then aligns and pairs with it. Successful pairing requires overcoming a structural incompatibility, since native dsDNA has a rise of 3.4 Å/base pair (bp) and RecA-ssDNA has a rise of 5.1 Å/bp [2]. Recognition is achieved without adenosine triphosphate hydrolysis [3], and it is therefore bound to happen by a mere combination of diffusive processes and molecular recognition.

In this Letter, we examine the homology search using the minimal model proposed by Dutreix *et al.* [4]. We suggest that two simple concepts underlie the basic physics. First, dsDNA and RecA-ssDNA are polymers, which strongly affects the diffusive process that brings them into contact. Second, when in contact, the incompatibility in pitch is overcome by the ability of dsDNA to exist in two different states [5–7]: the *B* state, with the standard rise, and the stretched *S* state, with a rise 1.7 times greater than the *B* state. Our intent here is to determine the scaling that results from these assumptions, compare it with experimental data, and suggest other experiments that could test our model.

Consider first what happens on the scale of the entire bacterium (or test tube), the "global search" process depicted in Fig. 1. We begin with the classical problem of a single dsDNA molecule diffusing in a volume R^3. We divide the dsDNA into Kuhn steps of length $l_k = 100$ nm ≈ 300 bp that move through the volume via Rouse diffusion [8,9]. We model the diffusion by jumps of the Kuhn steps on a lattice of cubes of volume l_k^3 at the rate of their self-diffusion time, $\tau = 6\pi\eta l_k^3/k_BT$, where k_BT is the Boltzmann factor and η is the viscosity of the intracellular fluid. The diffusion of a Kuhn step is qualitatively different from the motion of a globular particle of the same size, such as a protein or short oligonucleotide [10], due to the motion of the other parts of the chain. The Kuhn step moves subdiffusively with $\langle x^2\rangle^{1/2} \sim t^{1/4}$, compared to $t^{1/2}$ for proteins. As a consequence, the motion is compact [12]—the Kuhn step makes many visits to the same location in the time required to diffuse through the bacterium's volume.

We now introduce N_{RecA} RecA-ssDNA filaments into the volume. We assume that the dsDNA sequence is exactly homologous to some contiguous part of the ssDNA sequence on each filament, so that there is only one possible alignment of the sequences that leads to homologous recognition along the entire dsDNA strand. Since the persistence length of dsDNA (50 nm) is much less than that of

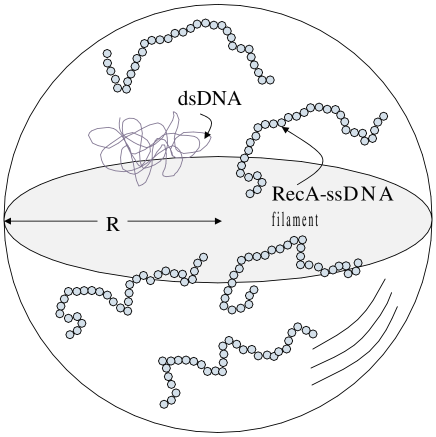

FIG. 1 (color online). In the global search, dsDNA comes into contact with the RecA-ssDNA nucleofilament by Rouse diffusion in the bacterium volume R^3. In this figure, we show $N_{\mathrm{RecA}} = 5$ nucleofilaments.

RecA-ssDNA (920 nm [7]), we can treat the latter as essentially stationary on the time scale of interest.

We assume that there is a sequence-independent attractive energy between dsDNA and RecA-ssDNA that overcomes the entropic penalty for locally aligning the strands, permitting a finer-scale "local" search that we examine shortly. However, experimental evidence [4,13–15] and simulations [16] suggest that this energy also affects the global chain diffusion. In particular, researchers [4,13,14] have suggested that RecA nucleoprotein networks are an intermediate stage in homologous recombination, and there are at least two DNA binding sites on RecA [15], which could allow for weak transient binding. We model this enthalpic attraction by a small energy V that results in rejected jumps when the Kuhn step tries to escape from a RecA-occupied site [17]. We assume that V is at most k_BT, which produces macroscopic condensationlike behavior [13,14] while preventing long-range sliding along the filament [18,19], in accord with experiments. In order to accurately reflect the latter, we choose jumps to adjacent lattice sites with equal probability.

We can now estimate the average time between the contact of any Kuhn step and its homologous counterpart. If the N_{RecA} RecA-ssDNA filaments occupy a fraction α of the total volume, then the time for one Kuhn step to visit all of the sites in the volume is proportional to $[\tau(1-\alpha) + \tau e^{V/k_BT}\alpha](R/l_k)^4$, where the bracketed terms are the average jump times from non-RecA and RecA sites, respectively. On average, each Kuhn step is in contact with each of its homologous lattice sites $(R/l_k)e^{V/k_BT}$ times during each pass through the volume. With N_{RecA} homologous lattice sites for each of the n_k Kuhn steps, the average time between homologous contacts for *any* Kuhn step is $(\tau/n_k)(1/N_{\mathrm{RecA}})(R/l_k)^3[e^{-V/k_BT} + \alpha(1 - e^{-V/k_BT})]$.

We now need to determine the probability that a given homologous contact results in recognition. We thus consider the local search depicted in Fig. 2, where we track a representative "seed" of n base pairs. We assume, when properly aligned and stretched, the seed stabilizes a three-strand complex, halts the global chain motion, and results in the rest of the strand being rapidly "zipped." We assume the $B \rightarrow S$ transition, rather than any steric hindrance, is the rate limiting step. Stretching to the S state unwinds the dsDNA [20] and binding of RecA to ssDNA exposes the ssDNA base pairs [21]. Moreover, dynamics limit steric hindrances. For the short sequences used *in vitro,* thermal fluctuations allow rapid sampling of all conformations, while the presence of topoisomerases *in vivo* allows the DNA to overcome topological constraints. Local fluctuations of RecA-ssDNA may further facilitate base pairing [22].

Since we previously assumed that the attractive energy aligns the strands, the seed moves via 1D diffusion (local sliding). We use a diffusivity $D_0 = k_BT/6\pi\eta l_0$, where l_0 is the size of the seed. This may somewhat overestimate the diffusivity of the seed, but for short times we can be certain that the seed diffuses as least as fast as the Kuhn segment. During this motion, the seed can convert between the native (B) state and the stretched (S) state, with $k_{B\rightarrow S}$ the rate of stretching to the S state and $k_{S\rightarrow B}$ the rate of relaxation back to the B state. Because of the coupled motion between different parts of the chain, it is unlikely that the seed will be able to reach all points along the length l_k, so we consider it to be restricted to a search region $l_s < l_k$ centered around the homologous sequence. The numerical value of l_s ultimately proves inconsequential at the scaling level provided that $l_s/l_k \sim \mathcal{O}(1)$.

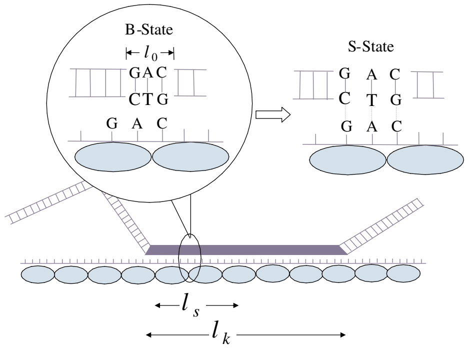

FIG. 2 (color online). After the global search brings a Kuhn step of dsDNA, shaded in the figure, in contact with its homologous counterpart on the RecA-ssDNA, a finer-scale local search occurs. For the contact to result in homologous recognition, a seed element (shown in the inset as 3 bp) of dsDNA must align itself with the homologous ssDNA sequence and stretch from the B state to the S state. The lengths l_0, l_k, and l_s are defined in the text.

In dimensionless form, scaling time with l_s^2/D_0, length with l_s, and the probability densities with $1/l_s$, our two-state reaction-diffusion system is described by

$$\begin{aligned}\partial_t P_B &= \partial_{xx} P_B - k(\epsilon P_B - P_S),\\ \partial_t P_S &= \partial_{xx} P_S + k(\epsilon P_B - P_S),\end{aligned} \tag{1}$$

over the search interval $x \in (-l_s/2, l_s/2)$. In the latter, $\epsilon \equiv k_{B\rightarrow S}/k_{S\rightarrow B} \ll 1$ is the B/S equilibrium constant and $k \equiv k_{S\rightarrow B} l_s^2/D_0$ is the Damköhler number for the relaxation process. We assume the initial position is unbiased and the seed is initially in B/S equilibrium: $P_B(x, 0) = 1/(1+\epsilon)$, $P_S(x, 0) = \epsilon/(1+\epsilon)$. Our previous restriction to a search region of size l_s requires that $\partial P_i/\partial x = 0$, $i = B, S$ at $x = \pm l_s/2$. If the seed reaches the center of the search region and is stretched, we assume that this corresponds to a successful search. Consequently, $P_S(0, t) = 0$ and the probability of a successful search with an n-bp seed, S_n, is computed by integrating the flux of stretched DNA at the properly aligned position,

$$S_n = \int_0^{l_k^3/l_0 l_s^2} \left[\left(\frac{\partial P_S}{\partial x} \right)_{x=0} \right] dt, \qquad (2)$$

where $l_k^3/l_0 l_s^2 \gg 1$ is the dimensionless contact time.

We obtain the leading order flux by substituting the regular perturbation expansion $P_i(x, t; \epsilon) \approx P_i^0(x, t) + \epsilon P_i^1(x, t) + \ldots$ into Eq. (1) and solving at successive orders with Laplace transforms, ultimately arriving at

$$\left(\frac{\partial P_S}{\partial x} \right)_{x=0} = 2\epsilon(1 + k) \left[\frac{\tanh(\sqrt{s + k}/2)}{\sqrt{s + k}} \right] + \mathcal{O}(\epsilon^2), \qquad (3)$$

where s is the Laplace variable. The flux decays rapidly for dimensionless times greater than unity, indicating that the primary contributions to the success rate are those searches which are initially in the neighborhood of $x = 0$. Since $l_k^3/l_0 l_s^2 \gg 1$, we can move the upper bound of the integral in Eq. (2) to infinity, whereupon $S_n \approx 2\epsilon\sqrt{k}$, valid for $k \gg 1$.

We can now estimate the average search time. With n_k Kuhn steps, each Kuhn step contains $1/n_k$ of the n_{bp} base pairs in the entire sequence, with $[(n_{\mathrm{bp}}/n_k) - n]$ possible n-bp seeds. Since $n_{\mathrm{bp}}/n_k \approx 300 \gg n$, the probability of success, S, during the contact of a single Kuhn step is $\approx (n_{bp}/n_k)S_n$ [23]. A search can be successful only once, so the Nth search has a success probability $S(N) = S(1 - S)^N \approx S \exp(-SN)$, and the mean number of contacts required for a successful search is the first moment of $S(N)$, namely $n_k/(2n_{bp}\epsilon\sqrt{k})$. Making use of the latter and our result for the average time between contacts, and further assuming that rejected jumps renew the local search process, we arrive at the average search time

$$\langle t \rangle \approx \left(\frac{\tau}{N_{\mathrm{RecA}}} \right) \left(\frac{R}{l_k} \right)^3 \left[\frac{e^{-V/k_B T} + \alpha(1 - e^{-V/k_B T})}{2\epsilon n_{bp}\sqrt{k}} \right]. \qquad (4)$$

We first check that the search time predicted by Eq. (4) agrees with observations if $n = 3$ bp, then rationalize the seed size in the following paragraph. The typical linear dimension of *E. coli* is $R \sim 1$ μm. The fluid viscosity is approximately 10 times that of water at biological temperatures, so $\tau \sim 10^{-3}$ s [4]. A $B \rightarrow S$ conversion costs $3.75 k_B T$/bp and forming a B/S interface costs $3.6 k_B T$/interface [25], so the equilibrium constant for a 3 bp seed is $\epsilon \approx 10^{-8}$. Although we have a reliable estimate for ϵ, no such data exist for the reaction between states, so we make a simple approximation using Kramers' escape from a potential well [26]. Assuming the inflection of the well is $\mathcal{O}(1)$, then $k_{S \rightarrow B} \approx D_0/l_0^2$, which furnishes $k \approx (l_s/l_0)^2 \approx (l_k/l_0)^2 \approx 10^4$. With a RecA concentration of 1–10 μM ($\alpha \sim 10^{-2}$) and a small V, the effect of attraction is small. Using these values in (4), a 1 kbp ($n_{\mathrm{bp}} = 1000$) dsDNA strand recognizes one RecA-ssDNA ($N_{\mathrm{RecA}} = 1$) in $\langle t \rangle \approx 500$ s, which is on the order of magnitude of experimental data [4,27].

To justify our use of a 3 bp seed, we adopt the rationale Craig *et al.* [28] used to determine the critical number of base pairs for helix formation. We already know that a 3 bp seed with no attractive energy captures the experimentally observed homologous recognition time scale. If the seed size was 2 bp, then $\langle t \rangle \sim 5$ s. Conversely, if the seed size was 4 bp, $\langle t \rangle \sim 18.5$ h. Thus, for small attraction energies, only $n = 3$ bp gives biologically reasonable search times, assuming that the model and the other parameters' values are valid. Our seed size strongly depends upon the $B \rightarrow S$ energy barrier, so it is probably valid *in vitro.* It is possible that the *in vivo* seed size may be larger due to force-acting proteins or structural transitions in the RecA-ssDNA-dsDNA triplex [22]. The role of the energy barrier, and thus the dependence of $\langle t \rangle$ on ϵ, could be investigated experimentally by using magnetic tweezers to control both the extension and the torsion of the dsDNA.

We achieve excellent agreement with the experiments of Julin *et al.* [29], who observed that the rate of recombination increases linearly with dsDNA sequence length and RecA concentration, independent of the filament size (i.e., the ssDNA sequence length). Since the contribution of attraction is typically small, we also recover an essentially linear dependence on RecA concentration and negligible dependence on the size of the nucleofilament.

The effect of diffusivity can be examined experimentally by modifying the viscosity η of the fluid. Our rather unusual and nontrivial scaling $\langle t \rangle \sim \tau/\sqrt{k} \sim \eta^{1/2}$ reflects the opposing influences of diffusivity at the different scales; a high diffusivity aids in quickly locating the homologous portion of the RecA filament in the global search, whereas a low diffusivity allows the seed to stretch while it is properly aligned in the local search. To the best of our knowledge, there has been no systematic study of homologous recombination kinetics as a function of the solvent viscosity [30].

Our elementary treatment of attraction between dsDNA and RecA-ssDNA shows that any attractive energy, however small, decreases $\langle t \rangle$. The time to diffuse through the volume increases with V, but this is more than offset by the increased amount of time a Kuhn step spends on its homologous lattice site. The dependence on V could be verified experimentally, at least qualitatively, by adding either polyvalent cations or anions [13].

We finish by considering the effect of the dsDNA recognizing other short homologous sequences on the RecA-ssDNA and binding to them. We use $n = 3$, and set $V = 0$ for simplicity. On a random sequence of 300 bp, we would expect 1 in 64 of the 3 bp stretches to have the same sequence as our seed, so that the seed could make $q \approx 5$ different "wrong" matches during each contact. We assume that recognizing the wrong sequence (i) terminates the local search and (ii) forces the Kuhn step to unbind from the wrong match before making the next jump on the lattice. Condition (i) does not affect $\langle t \rangle$, since the probability of recognizing two different 3-bp sequences during the same local search is $\mathcal{O}(\epsilon^2)$. The effect of condition (ii) is more subtle. If we assume that each matched base gains an energy $\beta k_B T$, with β a numerical prefactor, then the average jump time, $\langle\tau\rangle$, on the lattice is

$$\frac{\langle\tau\rangle}{\tau} \approx (1-\alpha) + \alpha(1-qS) + \alpha qS \sum_{3}^{12} \frac{e^{\beta n}}{4^n}. \qquad (5)$$

The terms in the latter correspond to jumps from non-RecA sites, jumps from RecA sites without any binding, and jumps from RecA sites where the dsDNA was bound to the wrong sequence. The upper bound on the sum accounts for the fact that local searches of genomic DNA would not typically result in long wrong matches. We chose the numerical value based upon sequences used for hybridization experiments with DNA chips, where it is generally accepted that 12 bp stretches represent unique sequences on the genome (in the absence of repeat sequences). Since $\alpha qS \sim 10^{-6}$, we would need $\beta > 2.5$ to see an effect on the global search. However, such a large value of β is unlikely, since it would result in frequent trapping of the chain due to single-bp matches that do not require a $B \rightarrow S$ transition. Nevertheless, this discussion brings up many interesting issues concerning sequence dependence, which we are addressing by numerical simulations.

K. D. D. acknowledges financial support from the Human Frontier Science Program and Institut Curie. R. F. acknowledges financial support from DGA-CNRS. This work was supported by CNRS Nano-objects/Nanoscience Programme.

*Electronic address: Jean-Louis.Viovy@curie.fr

[1] A. Kuzminov, Microbiol. Mol. Biol. Rev. **63**, 751 (1999); S. Lusetti and M. Cox, Annu. Rev. Biochem. **71**, 71 (2002).

[2] A. Stasiak, E. Di Capua, and Th. Koller, J. Mol. Biol. **151**, 557 (1981).

[3] S. Kowalczykowski and R. Krupp, Proc. Natl. Acad. Sci. U.S.A. **92**, 3478 (1995).

[4] M. Dutreix, R. Fulconis, and J.-L. Viovy, Complexus **1**, 89 (2003).

[5] P. Cluzel *et al.*, Science **271**, 792 (1996).

[6] S. Smith, Y. Cui, and C. Bustamante, Science **271**, 795 (1996).

[7] M. Hegner, C. Smith, and C. Bustamante, Proc. Natl. Acad. Sci. U.S.A. **96**, 10 109 (1999).

[8] M. Doi, *Introduction to Polymer Physics* (Clarendon Press, Oxford, 1996).

[9] Although more complex polymer diffusion models could be considered, Rouse diffusion still represents a significant improvement over a point-sized diffusion model [10], and recent experiments [11] have suggested that dsDNA, in fact, obeys Rouse dynamics.

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

[11] R. Shusterman *et al.*, Phys. Rev. Lett. **92**, 048303 (2004).

[12] P. DeGennes, J. Chem. Phys. **76**, 3316 (1982).

[13] S. Chow and C. Radding, Proc. Natl. Acad. Sci. U.S.A. **82**, 5646 (1985).

[14] S. Tsang, S. Chow, and C. Radding, Biochemistry **24**, 3226 (1985).

[15] M. Kubista, M. Takahashi, and B. Norden, J. Biol. Chem. **265**, 18 891 (1990); B. Muller, T. Koller, and A. Stasiak, J. Mol. Biol. **212**, 97 (1990); A. Mazin and S. Kowalczykowski, EMBO J. **17**, 1161 (1998).

[16] S. Patel and J. Edwards, DNA Repair **3**, 61 (2004).

[17] Coupling between rejected jumps and the motion of the rest of the chain most strongly affects the high wavelength modes of the diffusion, rather than the more relevant lower modes.

[18] D. Gonda and C. Radding, Cell **34**, 647 (1983); J. Biol. Chem. **261**, 13 087 (1986).

[19] K. Adzuma, J. Biol. Chem. **273**, 31 565 (1998).

[20] J. F. Léger *et al.*, Phys. Rev. Lett. **83**, 1066 (1999).

[21] T. Nishinaka *et al.*, Proc. Natl. Acad. Sci. U.S.A. **94**, 6623 (1997).

[22] T. Nishinaka *et al.*, Proc. Natl. Acad. Sci. U.S.A. **95**, 11 071 (1998).

[23] By assuming that each seed searches independently, we qualitatively account for the difference in pitch between *B*-DNA and RecA-ssDNA. As pointed out in a recent article [24], the difference in pitch does not require that *all* base pairs be in register in order to search the RecA-ssDNA sequence, thereby enhancing the local search.

[24] K. Klapstein, T. Chou, and R. Bruinsma, Biophys. J. **87**, 1466 (2004).

[25] P. Cizeau and J. Viovy, Biopolymers **42**, 383 (1997).

[26] H. Risken, *The Fokker-Planck Equation: Methods of Solution and Applications* (Springer, Berlin, 1984).

[27] Estimating D_0 by the diffusivity of the Kuhn step reduces $\langle t\rangle$ by a factor of 10, which is still within the experimental range.

[28] M. Craig, D. Crothers, and P. Doty, J. Mol. Biol. **62**, 383 (1971).

[29] D. Julin, P. Riddles, and I. Lehman, J. Biol. Chem. **261**, 1025 (1986).

[30] The closest analogue is the work of Lavery and Kowalczykowski [31], who added polyethylene glycol or poly(vinyl alcohol) to simulate *in vivo* occupied volume effects caused by the other parts of the cell. While the addition of macromolecules certainly altered the viscosity of their solutions, their main conclusion was that such polymers promote the formation of the RecA-ssDNA complex at Mg^{2+} concentrations which are otherwise prohibitive. This effect is absent from our model, since we assumed *a priori* that the RecA-ssDNA complex has been formed and is stable throughout the homology search.

[31] P. Lavery and S. Kowalczykowski, J. Biol. Chem. **267**, 9307 (1992).