Which way up? Recognition of homologous DNA segments in parallel and antiparallel alignments

Dominic J. (O') Lee, Aaron Wynveen, Tim Albrecht, and Alexei A. Kornyshev

Citation: *The Journal of Chemical Physics* **142**, 045101 (2015); doi: 10.1063/1.4905291

View online: <http://dx.doi.org/10.1063/1.4905291>

View Table of Contents: <http://scitation.aip.org/content/aip/journal/jcp/142/4?ver=pdfcov>

Published by the [AIP Publishing](#)

Articles you may be interested in

[Free-energy calculations for semi-flexible macromolecules: Applications to DNA knotting and looping](#)
J. Chem. Phys. **141**, 174902 (2014); 10.1063/1.4900657

[Equilibration of complexes of DNA and H-NS proteins on charged surfaces: A coarse-grained model point of view](#)
J. Chem. Phys. **141**, 115102 (2014); 10.1063/1.4895819

[Controlling the conformations and transport of DNA by free energy landscaping](#)
Appl. Phys. Lett. **99**, 263112 (2011); 10.1063/1.3673277

[DNA methylation profiling in nanochannels](#)
Biomechanics **5**, 034106 (2011); 10.1063/1.3613671

[Effect of the atmospheric pressure nonequilibrium plasmas on the conformational changes of plasmid DNA](#)
Appl. Phys. Lett. **95**, 083702 (2009); 10.1063/1.3212739

A banner for AIP The Journal of Chemical Physics. It features the AIP logo and the text 'The Journal of Chemical Physics' in white on a dark blue background. Below this, it says 'Meet The New Deputy Editors' in white. There are three circular portraits of the new deputy editors: Peter Hamm, David E. Manolopoulos, and James L. Skinner. The background of the banner is decorated with colorful, abstract, star-like patterns in green, yellow, and purple.

Which way up? Recognition of homologous DNA segments in parallel and antiparallel alignments

Dominic J. (O') Lee,^{1,a)} Aaron Wynveen,² Tim Albrecht,¹ and Alexei A. Kornyshev¹

¹*Department of Chemistry, Imperial College London, SW7 2AZ London, United Kingdom*

²*School of Physics and Astronomy, University of Minnesota, Minneapolis, Minnesota 55455, USA*

(Received 3 December 2014; accepted 19 December 2014; published online 22 January 2015)

Homologous gene shuffling between DNA molecules promotes genetic diversity and is an important pathway for DNA repair. For this to occur, homologous genes need to find and recognize each other. However, despite its central role in homologous recombination, the mechanism of homology recognition has remained an unsolved puzzle of molecular biology. While specific proteins are known to play a role at later stages of recombination, an initial coarse grained recognition step has, however, been proposed. This relies on the sequence dependence of the DNA structural parameters, such as twist and rise, mediated by intermolecular interactions, in particular, electrostatic ones. In this proposed mechanism, sequences that have the same base pair text, or are homologous, have lower interaction energy than those sequences with uncorrelated base pair texts. The difference between the two energies is termed the “recognition energy.” Here, we probe how the recognition energy changes when one DNA fragment slides past another, and consider, for the first time, homologous sequences in antiparallel alignment. This dependence on sliding is termed the “recognition well.” We find there is a recognition well for anti-parallel, homologous DNA tracts, but only a very shallow one, so that their interaction will differ little from the interaction between two nonhomologous tracts. This fact may be utilized in single molecule experiments specially targeted to test the theory. As well as this, we test previous theoretical approximations in calculating the recognition well for parallel molecules against MC simulations and consider more rigorously the optimization of the orientations of the fragments about their long axes upon calculating these recognition energies. The more rigorous treatment affects the recognition energy a little, when the molecules are considered rigid. When torsional flexibility of the DNA molecules is introduced, we find excellent agreement between the analytical approximation and simulations. © 2015 AIP Publishing LLC. [<http://dx.doi.org/10.1063/1.4905291>]

I. INTRODUCTION

The shuffling of homologous genes between DNA molecules is a key process in meiosis and DNA repair. It expedites evolution and facilitates genetic diversity, as well as underpinning the transfer of genetic material between different strains and species of bacteria and viruses. The recombination process requires base pairing between the recombining strands, in which various proteins, for instance, RecA in E-coli, play a crucial role. However, before recombination, there needs to be a search in which homologous genes can find each other. The slow process of random diffusion is incompatible with the rate of recombination,^{1,2} suggesting that another mechanism is at play. What this might be is an important remaining question in molecular biology.

The process of homologous recombination is known to proceed via (i) the breakage of double stranded (ds-)DNA and formation of single strands mediated by specialized proteins,^{3,4} followed by (ii) the single strand recognizing and invading a homologous double helix through base pairing (single strand invasion). One possible stage where the recognition of the homologous genes takes place is at the stage of broken strand exchange, utilizing the sequence complementarity between

single strands.^{4–7} However, if complementarity was the only homology recognition mechanism at work, recognition would occur between homologous fragments of genes with as few as 8–10 base pairs,⁸ suggesting that frequent recognition and recombination errors would be inevitable. It might also take too long for the single strand to search for the homologous duplex. Indeed, there is evidence to suggest that before strand breakage occurs, identical intact double-stranded DNA segments may pair.^{9–12} It is also conceivable that homology recognition between intact DNA tracts may occur in other biological processes. One instance might be the silencing of multiple copies of the same gene.¹⁰ One possible mechanism for homology recognition before double strand breakage is the stem loop kissing model,^{13–16} which again relies on the sequence complementarity of single stranded DNA, at loop ends. A second alternative for homology pairing may arise from the base-pair dependent conformation of DNA.^{17,18}

It has been well established that DNA wraps on histones at defined sequence tracts,¹⁹ the base sequence simultaneously encoding the nucleosome positioning.²⁰ This may arise from base-pair sequence dependence in the bending of DNA,^{21,22} but also from the base pair dependence of DNA structure.^{23,24} Base pair dependent DNA structure may also influence the interactions between two DNA fragments in the modes of interaction that depend on the helix shape of the molecule.^{25–28}

^{a)}Electronic mail: domolee@hotmail.com

There is some compelling evidence to suggest that the intermolecular interactions are indeed affected by the helix structure of DNA.^{29,30} If the helix-dependent forces are important, then a base pair specific pattern of helix distortions³¹ may well provide a mechanism for homology recognition.³² This relies on the fact that two paired DNA fragments with the same pattern of helix distortions have lower interaction energy than those with different patterns^{32–35} due to commensurability between positive counter-ion and negative phosphate charges along the DNA pair, known as an electrostatic zipper.³⁶

Several experimental studies have reported signatures of DNA-DNA recognition, at the molecular level, in protein free solutions. These suggest a possible role for direct interactions between double stranded DNA in the homologous pairing process. The experiments utilized gel electrophoresis,³⁷ cholesteric liquid crystals,³⁸ AFM imaging,³⁹ and surface-confined DNA using magnetic beads.⁴⁰ In addition, Seeman's group reported the formation of PX-DNA complexes presumably resulting from homologue pairing,⁴¹ and homology recognition has also been observed between nucleosomes.³⁹

In two recent publications,^{42,43} in which the forces that depend on helix structure were considered, the effect of sliding one molecule with respect to another on interaction energies between homologous sequences, aligned in the same direction, was investigated. In Ref. 42, the DNA was considered to be rigid, whereas in Ref. 43, torsional elasticity was included. These studies demonstrated that the interaction energy, as a function of the sliding distance, forms a potential well centred at the point where the two homologous base pair sequences are perfectly aligned in parallel register. The width of the well has been found to be much larger than the decay length of DNA-DNA interactions (on the scale of a Debye screening length of the order of angstroms), contrary to naive expectations. Its origin lies in the character of the helix distortions.^{42,43} This finding advocates a possible role for helix structure dependent forces in aligning homologous sequences before single strand invasion. Indeed, no energy well was predicted for non-homologous molecules. The absence of a recognition well between homologous sequences aligned in opposite directions (antiparallel homology) seems intuitively obvious, and was suggested in Ref. 42 without rigorous analysis.

In this study, we test the theoretical results of the previous work^{42,43} against Monte Carlo simulations. We also rigorously investigate, for the first time, the interaction of DNA pairs containing antiparallel homology. One finding is that there is, in fact, a recognition well, but a very shallow one. As the length of the antiparallel homologous sequences in juxtaposition increases, the difference in the pairing energy per base pair between such pairs and those containing totally uncorrelated sequences becomes negligible. Another issue that this study raised, for rigid molecules, is how the optimization of the interaction energy between two molecules with respect to the angles of rotation of the DNA fragments about their long axes should be performed. In previous studies, the optimization was performed after averaging over all realizations of base-pair sequence-dependent distortions of the two DNA fragments. For rigid molecules, the sequence dependent base pair disorder is quenched, which is to say that it is frozen and cannot change, unlike thermal disorder. Thus, strictly speaking, one

should optimise the interaction energy for each possible set of base pair sequences first, before averaging the energies of an ensemble of different sequences. In the present study, we have investigated the impact of the order of averaging on the evaluated recognition energy (the pairing energy difference between DNA fragments containing homologous sequences and those that do not). We find that the recognition energy changes only slightly; moreover, when we introduce torsional flexibility for long homologous sequences, this effect vanishes.

The paper is structured in the following way: In Sec. II, we discuss the nature of paired fragments investigated, as well as the model of helix distortions. Then, we consider the case of rigid molecules, where we present the two ways of optimizing the azimuthal orientations of the two fragments about their long-axes. Finally, we describe the situation of torsionally flexible molecules. Section III presenting the results is divided into two parts: In the first, we consider rigid molecules and in the second, torsionally flexible molecules. In Sec. IV, we discuss our findings and point to future theoretical and experimental work. Finally, we present a summary of the key findings of this study.

II. THEORY

A. General considerations

To investigate DNA homology recognition through helix dependent interactions, we construct a model in which two DNA fragments, each with total length L , lie parallel to each other, along their long axis (z -axis). Three different types of paired DNA fragments are considered. First, we suppose that the two fragments have completely different uncorrelated base pair texts, referring to this as a random pair. The other two types of pairing consist of tracts of DNA which have the same text (or are homologous to each other⁴⁴), and are of length L_H , which, in turn, may be embedded into fragments with random pairing. These final two are distinguished by the homologous tracts being either aligned so that both texts read in the same direction (*parallel homology*) or in opposite directions (*antiparallel homology*). The homologous sequences are positioned with their centres coinciding with the centres of both fragments. Therefore, on either side, we can have tracts of DNA of length $L_{NH} = (L - L_H)/2$, which consist of sequences which are completely uncorrelated with respect to each other. We label the two fragments within a pair $\mu = 1, 2$. This arrangement is illustrated in Fig. 1. The case where $L_{NH} \neq 0$ is not necessarily relevant to the recombination of genes. Nevertheless, we consider it for two reasons. First, this is the situation in homology recognition experiments currently underway. Secondly, this situation could be realized in other biological processes that may require the pairing of identical sequences, for instance, repeat induced point mutation.^{45,46}

To probe how the homologous segments recognize each other, we shift the centre of fragment 2 a distance x along the z -axis. Therefore, the centre of fragment 1 lies at $z = L/2$, whereas that of fragment 2 lies at $z = L/2 + x$. For each pair, we first calculate the pairing energy, either at $x = 0$ or as a function of x . For rigid fragments, the pairing energy is simply the interaction energy, but for the pairing energy of torsionally

FIG. 1. Diagrams illustrating the orientations of the paired fragments. On the left hand side, we show a small section of any of the three types of paired DNA fragments. Here, the trajectory of the point where each minor groove is bisected is shown by blue helical lines. The z -axis runs parallel to the molecular centre lines, illustrated by vertical grey lines. The equation describing these trajectories is given by Eq. (1). The two DNA fragments are separated by interaxial distance R and the respective azimuthal orientations of each trajectory are described by the phases $\Phi_1(z) + \phi_0 = \delta\Phi_1(z) + gz$ and $\Phi_2(z) + \phi_0 = \delta\Phi_2(z) + gz$. Here, ϕ_0 is a constant chosen here for the ease of presentation. It is set so that the points of reference from which $\Phi_1(z)$ and $\Phi_2(z)$ are measured do not lie on the line connecting the two molecular centre lines. The phase difference, which interaction energy depends on, is given by $\Delta\Phi(z) = \delta\Phi_1(z) - \delta\Phi_2(z) = \Phi_1(z) - \Phi_2(z)$, and does not depend on this choice. On the right hand side, we illustrate what we mean by antiparallel and parallel aligned homologies. Here, we present long DNA fragments of arbitrary sequences, which have homologous (identical) segments; parts of the arbitrary base pair text are shown. L_H is the length of homology (the unshown length L corresponds to the whole fragment); the short dashed lines refer to the ends of the homologous segments. Outside these lines, the sequences are uncorrelated between the two fragments. The end of one homology fragment may be shifted a distance x away from the end of the second fragment. Positions of both the 3' and 5' ends of the homologous segments define the direction of DNA polarity. Had we swapped the 3' and 5' ends, as well as rotated the segments 180° , this would be chemically distinct from the previous case in which we just rotated one segment from the parallel orientation; because of this, the pattern of distortions in the “swapped” case could be different.

flexible molecules, elastic energy will also contribute (see below). Upon determining these energies, we calculate the difference, or “recognition energy,” between the interaction energies of fragments with embedded homologous sequences and the energies of pairs containing completely uncorrelated, nonhomologous sequences.

In the model that we consider for homology recognition, differences in pairing energy arise from DNA not being an ideal helix structure. Previous studies³¹ have shown that a particular base pair text leads to its own pattern of helix distortions. To understand what is meant by helix distortions, let us consider the trajectories that the minor grooves trace out along the two fragments. First, we suppose that the molecular axes of the two fragments are separated a distance R apart. Then, the location of the minor groove along the helices, with radius a , can be described through the position vectors

$$\mathbf{r}_\mu(z) = \left(\frac{(-1)^\mu R}{2} + a \cos(\delta\Phi_\mu(z) + gz) \right) \hat{\mathbf{i}} + a \sin(\delta\Phi_\mu(z) + gz) \hat{\mathbf{j}} + z \hat{\mathbf{k}}, \quad (1)$$

in the region $0 < z < L$, for fragment 1, and $x < z < L + x$ for fragment 2. The trajectories described by Eq. (1) are illustrated on the left hand side of Fig. 1. Here, $g = 2\pi/H$, where $H \approx 33.8 \text{ \AA}$ is the average helical pitch of DNA. If the phase $\delta\Phi_\mu(z)$ is constant (or linear) with respect to z , Eq. (1) defines an ideal helix. It is the variations in $\delta\Phi_\mu(z)$ along the helices that describe the pattern of distortions for each fragment. These variations arise from both thermal fluctuations and intrinsic base-pair-specific distortions.^{25,31,47}

When there is no torsional strain on the two molecules (the torsionally relaxed state)

$$\frac{d\delta\Phi_\mu(z)}{dz} = \frac{\delta\Omega_\mu(z)}{h}. \quad (2)$$

The function $\delta\Omega_\mu(z)$ represents the field that characterizes the pattern of base pair distortions of an isolated molecule without thermal fluctuations. A particular function $\delta\Omega_\mu(z)$ representing a particular base pair sequence can be constructed from available DNA structural information.³¹ We assume that (for justification see Ref. 31)

$$\langle \delta\Omega_\mu(z) \delta\Omega_\mu(z') \rangle_{\delta\Omega} = \frac{h^2}{\lambda_c^{(0)}} \delta(z - z'), \quad (3)$$

where the bracket denotes an ensemble average over all possible realizations of $\delta\Omega_\mu(z)$. Here, $h \approx 3.4 \text{ \AA}$ is the average distance (or rise) between base pairs and $\lambda_c^{(0)}$ is the structural contribution to what is termed the helical coherence length,^{25,31,32} which we take to be $\lambda_c^{(0)} = 150 \text{ \AA}$ based on previous estimates.³¹ The parameter $\lambda_c^{(0)}$ is a measure of the degree of helix non-ideality caused by base-pair dependent distortions. It is the length scale over which two distorted helices with different base pair texts, described by Eqs. (1)–(3), lose azimuthal alignment with each other.^{25,32,47}

For random pairs, whose distortions are uncorrelated, $\langle \delta\Omega_1(z) \delta\Omega_2(z') \rangle = 0$. For the other two pairing types, the random sequences on either side of the homologous segments are also uncorrelated with each other. Thus, as before, $\langle \delta\Omega_1(z) \delta\Omega_2(z') \rangle = 0$ when either $0 < z, z' - x < L_{NH}$ or

$L_{NH} + L_H < z, z' - x < 2L_{NH} + L_H$. When $L_{NH} < z < L_{NH} + L_H$, for parallel homologous pairs $\delta\Omega_1(z) = \delta\Omega_2(z - x)$, whereas for the antiparallel pairs $\delta\Omega_1(z) = \delta\Omega_2(L_H + 2L_{NH} - z + x)$. In the main text, we will consider only the case where $L_{NH} = 0$.⁴⁸ Any effect due to non-homologous ends becomes negligible when we introduce torsional flexibility for sufficiently long homologous segments.⁴⁸

If the interaction energy E_{int} depends on the helix structure of the DNA, we may write^{25,32}

$$E_{\text{int}} = La_0(R) + \int_x^L dz (-a_1(R)\cos\Delta\Phi(z) + a_2(R)\cos 2\Delta\Phi(z)), \quad (4)$$

where $\Delta\Phi(z) = \delta\Phi_1(z) - \delta\Phi_2(z)$ is the difference between the sequence dependent helix distortions of paired molecules. In what follows, we will consider two separate situations. First, as in Ref. 42, the molecules are considered as torsionally rigid such that no torsional strain can accumulate. Hence, the relative angle between base-pairs on opposing molecules is solely due to sequence-dependent variations, i.e., Eq. (2). The second case is when we allow for torsional flexibility. Here, changes in $\Delta\Phi(z)$ from the form specified by Eq. (2) may occur in order to reduce (optimize) the interaction energy (described by Eq. (4)) at the cost of torsional elastic energy. In the main text, we shall consider only the ground state, leaving the effect of thermal fluctuations to the supplementary material.⁴⁸

B. Rigid molecules

For the case of rigid molecules, we consider only the $x = 0$ case, where, for homologous segments, the identical patterns of sequence dependent distortions of paired molecules lie directly across from one another. One can easily integrate Eq. (2), which yields

$$\Delta\Phi(z) = \frac{1}{h} \int_0^z [\delta\Omega_1(z') - \delta\Omega_2(z')] dz' + \Delta\Phi_0, \quad (5)$$

where $\Delta\Phi_0$ is a constant of integration to be determined by minimizing the interaction energy either after or before averaging over the realizations of $\delta\Omega_\mu(z')$, the sequence-dependent distortions describing a particular DNA sequence. $\Delta\Phi_0$, in geometric terms, is the azimuthal angle between the lines bisecting each minor groove of the two molecules at the position $z = 0$. In previous work,⁴² $\Delta\Phi_0$ was determined after averaging. In this case, $\Delta\Phi_0$ takes the form

$$\Delta\Phi_0 = \Delta\bar{\Phi}_0 - \frac{1}{h} \int_0^d [\delta\Omega_1(z') - \delta\Omega_2(z')] dz'. \quad (6)$$

Equation (6) ensures that at the location d , $\Delta\Phi(d) = \Delta\bar{\Phi}_0 = \langle \Delta\Phi(z) \rangle_{\delta\Omega}$. In other words, at $z = d$, $\Delta\Phi(z)$ is fixed at its ensemble-averaged value. The ensemble-averaged value $\Delta\bar{\Phi}_0$ is determined by both the values $a_1(R)$ and $a_2(R)$ in the interaction energy (Eq. (4)). By substituting Eq. (6) into Eq. (5), we may obtain general analytical forms for the average pairing energies for our three types of pairing $\langle E_{\text{int}} \rangle_{\delta\Omega}$, using the appropriate forms of $\delta\Omega_\mu(z')$ for each (see Appendix A of

the supplementary material⁴⁸). In this study, we have chosen that $a_1(R) > 4a_2(R)$, so that $\Delta\bar{\Phi}_0 = 0$.²⁵ Here, we find that the minimum average energy with respect to d is given at the location $d = L/2$ for both fragments containing parallel homologous segments and random pairs. For antiparallel fragments, there are two optimal values of d at $d = L_H/2 \pm L_H/4$ (for $L_{NH} = 0$). For these cases, we find the following expressions for the two recognition energies (the energy difference between the homologous pairs and uncorrelated ones), for details see Appendix A of the supplementary material.⁴⁸ For parallel homologous pairs, we have ($L_{NH} = 0$)

$$E_{\text{rec}} = -L_H(a_1(R) - a_2(R)) + 2\lambda_c^{(0)} \left[a_1(R) \left(1 - \exp\left(-\frac{L_H}{2\lambda_c^{(0)}}\right) \right) - \frac{a_2(R)}{4} \left(1 - \exp\left(-\frac{2L_H}{\lambda_c^{(0)}}\right) \right) \right], \quad (7)$$

and for antiparallel homologous pairs ($L_{NH} = 0$)

$$E_{\text{rec}} = \left\{ -a_1(R)\lambda_c^{(0)} \left[2 - 4\exp\left(-\frac{L_H}{4\lambda_c^{(0)}}\right) + 2\exp\left(-\frac{L_H}{2\lambda_c^{(0)}}\right) \right] + a_2(R)\lambda_c^{(0)} \left[\frac{1}{2} - \exp\left(-\frac{L_H}{\lambda_c^{(0)}}\right) + \frac{1}{2}\exp\left(-\frac{2L_H}{\lambda_c^{(0)}}\right) \right] \right\}. \quad (8)$$

We also performed MC simulations over an ensemble of fragments with different sequence-dependent distortions. For some of these simulations, $\Delta\Phi_0$ is found after averaging over all the realizations of $\delta\Omega_1(z)$ and $\delta\Omega_2(z)$. These simulations have been compared to the analytical results given by Eqs. (7) and (8), yielding excellent agreement (cf. Fig. 2).

Considering that we are dealing with disorder that is fixed and cannot change (quenched disorder), unlike thermal fluctuations, a better way of choosing $\Delta\Phi_0$ is to minimize the interaction energy of a particular configuration described by $\delta\Omega_\mu(z)$ and taking the average of the optimized energies for these configurations, rather than ensemble averaging over these configurations and then optimizing over $\Delta\Phi_0$ to find the minimum energy of this ensemble average. For the former case, on inspection, it is an extremely laborious task to obtain the exact analytical solution for $\Delta\Phi_0$ in terms of the arbitrary function $\delta\Omega_1(z') - \delta\Omega_2(z')$, and it does not seem possible to obtain a formula for the average energy if one had such a solution. However, we can perform MC simulations that directly minimize $\Delta\Phi_0$ for each configuration of $\delta\Omega_1(z') - \delta\Omega_2(z')$ in order to determine the ensemble average of the interaction energy.

In addition, we have obtained the analytical result of the leading order term in a perturbation expansion in $\delta\Omega_1(z') - \delta\Omega_2(z')$. Such an expansion assumes that L is not too large compared to λ_c . Details on how such a perturbation expansion is performed are given in Appendix B of the supplementary material.⁴⁸ To leading order, we find that

$$\Delta\Phi_0 \approx \Delta\bar{\Phi}_0 - \frac{1}{hL} \int_0^L dz' \int_0^z dz'' [\delta\Omega_1(z'') - \delta\Omega_2(z'')]. \quad (9)$$

Equation (9) essentially supposes that the spatial average of $\Delta\Phi(z)$ is $\Delta\Phi_0 = \langle \Delta\Phi(z) \rangle_{\delta\Omega}$. Again, expressions for the recognition energies may be derived by substituting Eq. (9) into Eq. (4). We have refrained from giving these expressions in the main text since they are rather cumbersome and such an approximate theory does not yield good quantitative results upon comparing them to the approximation-free MC—results that we have obtained. However, the theory provides some qualitative insight into what happens when one changes L_{NH} .⁴⁹ Also, the perturbation result, which correctly reproduces the limiting behaviour for small L , could be used (as discussed later) as the basis of a more comprehensive theory, extending over the full range of L_{NH} and L_H and including torsional flexibility.

C. Torsionally flexible molecules

For torsionally flexible molecules, we consider the elastic rod model and introduce a torsional elastic term.^{32,43} The total pairing energy becomes

$$E_{\text{pair}} = \frac{C}{4} \int_x^L dz \left(\frac{d\Delta\Phi(z)}{dz} - \frac{\delta\Omega_1(z) - \delta\Omega_2(z)}{h} \right)^2 + E_{\text{int}}, \quad (10)$$

where C is the torsional elasticity of the molecules. For finite C , Eq. (2) is no longer obeyed. In fact, upon minimizing the energies given by the combination of Eqs. (4) and (10), $\Delta\Phi(z)$ now satisfies a sine–Gordon like non-linear equation,³² which has an inhomogeneous term that depends on the derivative of $\delta\Omega_1(z) - \delta\Omega_2(z)$ with respect to z (for details, see Refs. 32–34). Unfortunately, an exact analytical solution to this equation in terms of $\delta\Omega_1(z) - \delta\Omega_2(z)$ cannot be obtained, although some progress has been made by using path integral methods (see Appendix C of Ref. 50). Here, we use a much simpler approach. If the helix dependent interactions described by Eq. (4) are sufficiently strong, we may use a trial function of the form (originally suggested in Ref. 33)

$$\Delta\Phi = \Delta\Phi_0 + \frac{1}{2h} \int_{-\infty}^{\infty} \text{sgn}(z - z') \exp\left(-\frac{|z - z'|}{\lambda_h}\right) \times [\delta\Omega_1(z') - \delta\Omega_2(z')] dz' \quad (11)$$

and substitute this into Eqs. (4) and (10), performing the average over realizations of $\delta\Omega_1(z') - \delta\Omega_2(z')$. The parameters λ_h and $\Delta\Phi_0$ are chosen to minimize the average energy. The parameter λ_h is a correlation length, which we term as the adaptation length.^{33,47} Its value increases as we reduce the strength of helix dependent interactions, or if we increase the size of the torsional rigidity constant C . Here, for parallel homologous pairs, $\delta\Omega_1(z) = \delta\Omega_2(z - x)$ and for antiparallel homologous pairs $\delta\Omega_1(z) = \delta\Omega_2(x - z)$, and use Eq. (11) exclusively to describe the homologous regions. This approach was used in Ref. 43 for sliding fragments ($x \neq 0$) with parallel homology. Now, we consider the antiparallel case and test both these results against MC simulations that exactly average the minimum energy described by Eqs. (4) and (10). Expressions of the average energy and the self-consistent equations that both $\Delta\Phi_0$ and λ_h satisfy are given in Appendix C of the supplementary material⁴⁸ (and Ref. 43 for parallel homology).

In Eq. (11), we have set the limits of integration to infinity, which is justified provided that $L_H/\lambda_h \gg 1$. One of the key results is that the approximation (Eq. (11)) predicts that, in the limit $L_H \rightarrow \infty$, the recognition energy per base pair approaches zero for the antiparallel case. This is indeed confirmed by the simulations, for which there is good agreement with the approximation (see Fig. 3).

Unlike the rigid case, when $L_H/\lambda_h \gg 1$ there is effectively no difference in how we perform the order of averaging. Due to torsional flexibility, the correlation range λ_h is finite, whereas for rigid molecules, it is always infinite. The upshot is that over a length $|z - z'|$ that is larger than λ_h , the optimal phase difference $\Delta\Phi(z)$ (at z) loses memory of the pattern of base pair sequences at z' described by $\delta\Omega_1(z') - \delta\Omega_2(z')$. This is due to the ability of the DNA fragments to torsionally adapt, at the cost of torsional elastic energy, so as to reduce their electrostatic energy. It does not matter where accumulation of phase mismatch occurs, as it only accumulates over length scales smaller than the correlation length λ_h , not over the whole fragment, unlike the rigid case.

III. RESULTS

A. Rigid molecules

In all calculations, we use the parameter values $a_1 = 0.015 k_B T / \text{\AA}$ and $a_2 = 0.0035 k_B T / \text{\AA}$,⁵¹ which were employed in Ref. 42. We calculate the energy in units of $k_B T$ and the lengths are given in units of base pairs, of length $h = 3.4 \text{ \AA}$. Let us start by investigating the case where the torsional rigidity constant C is infinitely large, as discussed in the first part of Sec. II. In Fig. 2, plots (at $x = 0$) of the recognition energies of antiparallel homologous and parallel homologous pairs are provided. Upon ensemble averaging over different sequences before minimizing the free energy, we find that the agreement between the analytical form of the recognition energies (see Eqs. (7) and (8)) and those found from MC simulations is excellent, providing us with a useful test of the reliability of both. However, at present, only MC simulations can provide us with reliable quantitative results when we minimize the energy of a particular pairing of sequences before ensemble averaging over different sequence realizations. As discussed in Sec. II, it is indeed possible to develop a perturbation expansion, for which we have calculated the leading order term (Eq. (9)). Unfortunately, there is little quantitative agreement between the approximation and simulation for the values of L and L_H considered here. The perturbation expansion, however, provides the correct qualitative trends. Therefore, we have only included this analytical result as part of the supplementary material.⁴⁸

The recognition energies between fragments containing parallel homologous segments and between antiparallel homologous segments are somewhat similar for homology lengths less than 250 base pairs. For larger homology lengths, e.g., at 2000 base pairs, the difference between these recognition energies becomes rather large. Indeed, upon examining Eq. (7), the large limit of increasing L_H , the recognition energy for parallel homology grows linearly with L_H . However, for the case of antiparallel homology (Eq. (8)), it approaches a constant value.

Upon minimizing the interaction energy with respect to pair orientation before ensemble averaging over different sequence realizations (different values of $\Delta\Phi_0$ for each), we find that the magnitude of the recognition energy for parallel homologous segments is somewhat reduced but increases slightly for the antiparallel case.

B. Flexible molecules

We now consider the molecules as torsionally flexible so that C is finite. In the results that we present, we use a value of the torsional persistence length $C/k_bT = 725$ Å.⁵² This choice is in good agreement with the value obtained from the largest DNA length considered in the simulations of Ref. 53.

In Figure 3, we plot the recognition energy for both the pairs containing parallel and antiparallel homology as a function of x , the shift between the two fragments making up a particular pair. We find a clear difference between the recognition energy wells for antiparallel homology pairing and those with parallel homology pairing. The well for the antiparallel case is much shallower and changes little with increasing homology length, whereas for the parallel pairing, it becomes much deeper with increasing homology length, L_H . We find, as in previous studies,^{33,43} that introducing torsional flexibility reduces the depth of the recognition well (to see this compare Figure 2 with Figure 3 at $x = 0$). This is due to the torsional adaptation of the molecules, for which the energy for uncorrelated sequences now scales as L_H , as opposed to $\lambda_c^{(0)}$ for the helix structure dependent energy terms (which would be the case for long rigid fragments), hence reducing the pairing

energy for completely random pairs. Consequently, since the fragments can adapt to each other, the difference between the total energy of homologous pairs and nonhomologous, uncorrelated pairs will be less pronounced. But, again, because such adaptation requires torsional elastic energy, the recognition energy, i.e., the difference in the energies between homologous and nonhomologous pairs, survives.

Furthermore, the width of the recognition well, for parallel homology, grows narrower as compared to that found for rigid molecules, originally considered in Ref. 42. As discussed previously,⁴³ one would expect that, for torsionally flexible molecules, the width of the well is effectively determined by the shorter correlation/adaptation length λ_h (at $x = 0$), instead of the longer helical coherence length $\lambda_c^{(0)}$. As fragments slide along each other beyond the correlation length/adaptation length, one would expect that any information about the homology between them will effectively get washed out due to torsional adaptation of the molecules.

Next, we examine the recognition energy per base pair at $x = 0$ as a function of homology length L_H . Indeed, the theory [Eq. (11)] predicts that the recognition energy per base pair in the case of antiparallel homology should go to zero as $L_H \rightarrow \infty$, whereas for parallel homology it should tend to a constant value (see Appendix C of supplementary material⁴⁸). This indeed what is seen from the plots in Figure 4. A good fit to the MC numerical data of the recognition energy per base pair is given by the empirical formula $\varepsilon_{\text{rec}}(L_H) = (A + BL_H)^{-1}$.

We should add that in Appendix C of the supplementary material,⁴⁸ we included thermal fluctuations and have calculated the recognition well for the same parameter values, using also the values $L_H = 6800$ Å (corresponding to 2000

FIG. 2. The recognition energy as a function of the homology length. The MC simulation data are shown as points, whereas theoretical curves for averaging before minimizing (from Eqs. (7) and (8)) are presented by solid lines (all for $L_{NH} = 0$). In the figure, E_p , E_a , and E_n correspond to the pairing energies of the fragments containing parallel homologous segments, antiparallel homologous segments, and completely uncorrelated sequences. Here, in all cases, the top red lines/points correspond to the energy difference between pairs containing homologous segments in antiparallel alignment and completely uncorrelated sequences, while the lower black ones correspond to the difference between pairs that contain homologous sequences in parallel alignment and random pairs. For the simulation data, the squares and circles correspond to minimizing the energy after ensemble averaging over different sequences, while the diamonds and triangles correspond to minimizing the energy before sequence averaging. This difference between these is explained in the main text, along with values given for the parameters used.

FIG. 3. The recognition wells for various homology lengths (the value of L_H (in units of base pairs of length $h = 3.4$ Å) is given in the top right corner of each panel, $L_{NH} = 0$) plotted as function of the shift, x (offset for clarity's sake). The recognition energy for parallel homologous sequences is provided in black, whereas that for fragments containing antiparallel homologous sequences is provided in red. The points were found via MC simulations. For pairs containing antiparallel homology, the large L_H theory predicts that the recognition energy is zero, shown as a flat dashed line. In the case for pairs with parallel homology, the theoretical calculation of the recognition energy is given by the dashed line that passes through the simulation data, demonstrating good agreement with the MC simulation data. The MC data for antiparallel sequences, in red, were fit with a Lorentzian function, tracing out a very shallow well, to help guide the eye. The values of the parameters are given in the main text.

FIG. 4. The recognition energy per base pair at the bottom of the well ($x = 0$, no shift) as a function of homology length. Here, the minimum value of the recognition well (at $x = 0$) is shown as a function of the length of homology L_H for both the case of fragments containing parallel (lower black) and antiparallel (upper red) homologous segments (when nonhomologous segments flanking the homologous part are absent, $L_{NH} = 0$). The MC numerical data are fit very well by an empirical formula given in the text. The approximation for the recognition energy for the parallel homologous pairs, as described by Eq. (11), in the $L_H \rightarrow \infty$ limit is also provided in the plot (by the dashed black line). For antiparallel molecules, this approximation predicts zero recognition energy per base pair at infinite homology length, conforming to the trend seen in the simulations.

base pairs) and $L_{NH} = 0$. We find that the recognition well seems to be affected very slightly when thermal fluctuations are included.⁴⁸

IV. DISCUSSION

Intuitively, within the context of the recognition model, it was thought that DNA fragments containing homologous sequences in antiparallel alignment had the same average pairing energy as those with completely uncorrelated base pair tracts.⁴² Naively, it was believed that since homologous fragments in antiparallel alignment cannot maintain commensurability in the strand-groove alignment along their juxtaposition length, there would be no recognition energy. However, as there are still weak correlations between the azimuthal orientations of the two antiparallel helices, we realized that this may not be the case. This issue was a core motivation for this study, and it soon became clear that rigorous analysis was needed. We have found that there is, indeed, a minor difference in pairing energy between antiparallel aligned homologous sequences and nonhomologous sequences, resulting in a small value of recognition energy (the difference between homologous and completely uncorrelated pairing energies). In the rigid case, for homology lengths fewer than 250 base pairs, the recognition energy at antiparallel orientation is of comparable magnitude to that of the parallel one.⁵⁴ However, the general trend for both rigid and flexible molecules is that the antiparallel recognition energy per base pair disappears with increasing homology length, while for the parallel one it tends towards a finite constant value.

In devising this study, we also sought to test the analytical approximations used in previous studies^{42,43} against MC simulations. When optimizing the azimuthal orientations of

the fragments after minimizing the energy for rigid molecules, we have found excellent agreement between the two. We have also investigated whether the order of averaging over different realizations of base pair sequence dependent distortions and optimizing the angle of rotation of the fragments about their long axis mattered. For rigid fragments, the base pair helix disorder should be considered as quenched. This means that azimuthal orientations of the DNA fragments about their long axes should be optimized for each realization of the base pair dependent helix distortions first. The evaluation of the ensemble average of the energy should be considered next. However, swapping the correct order leads to simple analytical expressions, and we wanted to check the inaccuracy of this approximation. We have found only a slight difference in recognition energy. Moreover, this difference vanished when considering long torsionally flexible molecules. In this case, we found excellent agreement between the analytical result and MC simulation. This suggests that the effect of changing the order of operations is, at most, significant only for homology lengths smaller than the adaptation length λ_h (only DNA fragments smaller than λ_h can be considered as being effectively rigid^{33,55}). Although we have analytical approximations for both long and short homology lengths (rigid molecules), we have yet to build an approximation that interpolates between the two for the recognition well. This might be the subject of future work.

Through MC simulations, we are able to confirm that torsional flexibility reduced the depth of the recognition well as well as making it narrower. We also found that introducing thermal fluctuations⁴⁸ did little to change the shape of the well. Still the depth of the recognition well for long parallel homology fragments is much larger than $k_B T$ and the well is much wider than the Debye screening length in physiological solution (both as in previous works^{42,43}), which one might naively expect from electrostatic interactions between DNA. If the recognition well was indeed too narrow or too shallow, one would expect that thermal fluctuations could disrupt an alignment of homologous genes. A very narrow recognition well might lead to a long initial homology search or missing the target, both impeding the whole process of homology recognition.⁵⁶ On the other hand, if the recognition well was too wide, homologous genes may not reach ideal alignment, due to large thermal fluctuations about their optimal alignment; this could slow down the rate of homologous recombination and might reduce its fidelity (but there however might be the possibility of additional fine graining, discussed below).

We should point out that *in-vivo* there may be additional complicating factors in a recognition search.⁵⁷ Nevertheless, our calculations suggest that natural sequence-related helix distortions could provide a reliable initial coarse-grained homology search by providing a large target for recognition, relying also on homology between sequences much larger than 10 base pairs. For sequences containing 10 base pairs or fewer, our calculations suggest that this initial recognition step would be washed out by thermal fluctuations. This might prevent genes containing such small lengths of homology from efficiently recombining, except under specific conditions.⁵⁸

Global homology recognition through helix distortions may not be the complete search mechanism, even in the

absence of strand breakage.⁵⁹ There is some evidence⁴⁶ to suggest the importance of an interaction that recognizes homology at the local level of individual base pairs, yet the nature of this interaction remains unclear. However, as homologous pairing was shown not to depend on the exact nature of the base pair text,⁴⁶ those results seem to rule out the stem loop kissing mechanism.^{13–16} The pairing energy from such a local base pair specific interaction, for double helices, would also be affected by helix distortions, as the base pairs would have to align in an optimal orientation for this type of pairing interaction to be sufficiently strong. The theoretical implications of such a contribution to homology recognition have yet to be investigated, and may be the topic for further work.⁵⁹ However, the contribution of such an interaction alone would likely yield a very narrow recognition well, which would be determined by short range interactions. Hence, this mechanism alone would probably engender a long homology search.⁵⁹ Thus, we think that the global recognition mechanism due to base-pair specific helix distortions is important in providing a fast initial recognition step.

On the experimental side, one of us is currently investigating differences between supercoiled DNA *plasmids* that contain homologous segments, in either parallel or antiparallel alignment, and plasmids in which the sequence texts are completely uncorrelated. Indeed, if interaction forces that depend on DNA helix structure are indeed important, on the basis of the calculations presented here, we would expect a significant difference in conformation and dynamic properties (assessed by different experimental methods) between plasmids with parallel homology and plasmids with completely uncorrelated sequences, but not between plasmids containing antiparallel homologous segments and completely uncorrelated ones. In developing more accurate models to describe the difference between such plasmids, we hope to take account of the braiding of DNA through closed supercoiling. Currently, a statistical mechanical model for closed loop supercoiling, including interactions that depend on helix structure, is being developed.⁶⁰ This model is based on the previous work of Refs. 35, 50, and 61–63. We hope to use this model to build upon this investigation, further rationalizing the experiments underway. There are also further experiments that can be performed. The experiments of Ref. 46 were performed *in vivo*, although in the absence of the ME13 protein, an analogue of RecA specific to the species considered there (suggesting that single strand breakage was not important in the repeat point mutation process under study). Thus, future experimental work, *in vitro* in a protein free environment, should also consider building DNA pairs that contain special periodic arrays of sequence homology similar to those of Ref. 46. Such specially designed experiments, in controlled conditions, may help to further elucidate the relative importance of various homology pairing mechanisms.⁵⁹

V. CONCLUDING POINTS

In agreement with earlier expectations, we have shown that there is little difference between the interaction energy of a pair of nonhomologous sequences and a pair of antiparallel homologous ones. This fact is important in designing single

molecule experiments. But in addition, our analysis suggests that the homology recognition mechanism described herein significantly favours the correct, parallel alignment of homologous DNA.

We have discovered that a slight difference exists in the results for the calculated recognition energy for rigid molecules, depending on whether we optimized the azimuthal orientation of the molecules for each sequence dependent pattern of helix distortions, and then performed an ensemble average of the energy, as compared to averaging the energy before optimizing the azimuthal orientation. This difference, however, vanishes for long torsionally flexible molecules.

The MC simulations performed in this work confirm our previous results accounting for the estimated torsional flexibility of DNA, namely, that the recognition well grows shallower and narrower. This, however, does not preclude its biological significance.

ACKNOWLEDGMENTS

The authors would like to thank Sergey Leikin (NICHD, NIH Bethesda), Takashi Ohyama (Waseda University), Mara Prentiss (Harvard University), and Kenichi Yoshikawa (Doshisha University) for useful discussions. D.J.L. and A.A.K. would like to acknowledge the support of the HFSP grant “Chiral effects in DNA supercoiling” (Grant No. RGP0049/2010-C102). A.A.K. also acknowledges the JSPS short term award.

- ¹A. Barzel and M. Kupiec, *Nat. Rev. Genet.* **9**, 27 (2008).
- ²M. J. Neale and S. Keeney, *Nature* **442**, 153 (2006).
- ³D. R. F. Leach, *Genetic Recombination* (Blackwell Science, Oxford, 1996).
- ⁴R. D. Camerini-Otero and P. Hsieh, *Annu. Rev. Genet.* **29**, 509 (1995).
- ⁵B. Lewin, *Genes VI* (Oxford University Press, Oxford, 1997).
- ⁶H. M. Sobell, *Proc. Natl. Acad. Sci. U.S.A.* **69**, 2483 (1972).
- ⁷G. G. Doyle, *J. Theor. Biol.* **70**, 171 (1978).
- ⁸P. Hsieh, C. S. Camerini-Otero, and R. D. Camerini-Otero, *Proc. Natl. Acad. Sci. U.S.A.* **89**, 6492 (1992).
- ⁹D. Zickler and N. Kleckner, *Annu. Rev. Genet.* **33**, 603–754 (1999).
- ¹⁰D. Zickler, *Chromosoma* **115**, 158 (2006).
- ¹¹B. M. Weiner and N. Kleckner, *Cell* **77**, 977 (1994).
- ¹²K. A. Boateng, M. A. Bellani, I. V. Gregoret, F. Pratto, and R. D. Camerini-Otero, *Dev. Cell* **24**, 196 (2013).
- ¹³R. E. Wagner, Jr. and M. Radman, *Proc. Natl. Acad. Sci. U.S.A.* **72**, 3619 (1975).
- ¹⁴J. Tomizawa, *Cell* **38**, 861 (1984).
- ¹⁵D. R. Forsdyke, *Mol. Biol. Evol.* **12**, 949 (1995).
- ¹⁶D. R. Forsdyke, *Evolutionary Bioinformatics* (Springer, New York, 2006).
- ¹⁷R. E. Dickerson and H. R. Drew, *J. Mol. Biol.* **149**, 761 (1981).
- ¹⁸R. E. Dickerson, *Methods Enzymol.* **211**, 67 (1992).
- ¹⁹F. Battistini, C. A. Hunter, I. K. Moore, and J. Widom, *J. Mol. Biol.* **420**, 8 (2012).
- ²⁰D. J. Gaffney, G. McVicker, Y. N. Fondufe-Mittendorf, N. Lewellen, K. Michelini, J. Widom, Y. Gilad, and J. K. Pritchard, *PLoS Genet.* **8**, 1003036 (2012).
- ²¹W. K. Olson, A. A. Gorin, X.-J. Lu, L. M. Hock, and V. B. Zhurkin, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 11163 (1998).
- ²²W. K. Olson, N. L. Marky, R. L. Jernigan, and V. B. Zhurkin, *J. Mol. Biol.* **232**, 530 (1993).
- ²³A. G. Cherstvy, *J. Phys. Chem. B* **113**, 4242 (2009).
- ²⁴R. Rohs, S. M. West, A. Sosinsky, P. Liu, R. S. Mann, and B. Honig, *Nature* **461**, 1248 (2009).
- ²⁵A. A. Kornyshev, D. J. Lee, S. Leikin, and A. Wynveen, *Rev. Mod. Phys.* **79**, 943 (2007).
- ²⁶A. A. Kornyshev, *Phys. Chem. Chem. Phys.* **12**, 12352 (2010).
- ²⁷D. J. Lee, *J. Phys.: Condens. Matter* **23**, 105102 (2011).
- ²⁸M. Kanduć, J. Dobnikar, and R. Podgornik, *Soft Matter* **5**, 868 (2009).
- ²⁹A. Leforestier and F. Livolant, *Proc. Natl. Acad. Sci. U.S.A.* **106**, 9157 (2009).

- ³⁰A. A. Kornyshev, D. J. Lee, S. Leikin, A. Wynveen, and S. Zimmerman, *Phys. Rev. Lett.* **95**, 148102 (2005).
- ³¹A. Wynveen, D. J. Lee, A. Kornyshev, and S. Leikin, *Nucleic Acids Res.* **36**, 5540 (2008).
- ³²A. A. Kornyshev and S. Leikin, *Phys. Rev. Lett.* **86**, 3666 (2001).
- ³³A. G. Cherstvy, A. A. Kornyshev, and S. Leikin, *J. Phys. Chem. B* **108**, 6508 (2004).
- ³⁴A. A. Kornyshev and A. Wynveen, *Phys. Rev. E* **69**, 041905 (2004).
- ³⁵R. Cortini, A. A. Kornyshev, D. J. Lee, and S. Leikin, *Biophys. J.* **101**, 875 (2011).
- ³⁶A. A. Kornyshev and S. Leikin, *Phys. Rev. Lett.* **82**, 4138 (1999).
- ³⁷S. Inoe, S. Sugiyama, A. A. Travers, and T. Ohyama, *Biochemistry* **46**, 164 (2007).
- ³⁸G. S. Baldwin, N. J. Brooks, R. E. Robson, A. Wynveen, A. Goldar, S. Leikin, J. M. Seddon, and A. A. Kornyshev, *J. Phys. Chem. B* **112**, 1060 (2008).
- ³⁹J. Nishikawa and T. Ohyama, *Nucleic Acid Res.* **41**, 1544 (2013).
- ⁴⁰C. Danilovic, C. H. Lee, K. Kim, K. Hatch, V. W. Coljee, N. Kleckner, and M. Prentiss, *Proc. Natl. Acad. Sci. U.S.A.* **106**, 19824 (2009).
- ⁴¹X. Wang, X. Zhang, C. Mao, and N. C. Seeman, *Proc. Natl. Acad. Sci. U.S.A.* **107**, 12547 (2010).
- ⁴²A. A. Kornyshev and A. Wynveen, *Proc. Natl. Acad. Sci. U.S.A.* **106**, 4683 (2009).
- ⁴³D. J. Lee and A. A. Kornyshev, *J. Chem. Phys.* **131**, 155104 (2009); Erratum: *J. Chem. Phys.* **131**, 219901 (2009).
- ⁴⁴Within the same organisms there is only a tiny difference at most between the identical DNA texts and homologous ones (the same type of genes, responsible for the same encoding function).
- ⁴⁵M. E. Hood, M. Katawicz, and T. Giraud, *Genetics* **170**, 1081 (2005).
- ⁴⁶E. Gladyshev and N. Kleckner, *Nat. Commun.* **5**, 3509 (2014).
- ⁴⁷A. A. Kornyshev, D. J. Lee, S. Leikin, and A. Wynveen, *Nucleic Acids Res.* **39**, 7289 (2011).
- ⁴⁸See supplementary material at <http://dx.doi.org/10.1063/1.4905291> for details of the derivations (Appendixes A and B for rigid molecules and Appendix C for flexible molecules), and how the recognition energies depend on L_{NH} for rigid molecules (Appendix D), as well as the effect of thermal twisting fluctuations for flexible molecules (Appendix E).
- ⁴⁹As L_{NH} , the length of the non-homologous sections, increases, the perturbation theory suggests that the net value of the recognition energy decreases. This is different from the recognition energy per base pair, which always decreases with increasing L_{NH} , no matter which way we do the averaging. The former effect occurs because the nonhomologous sections will tend to find their optimal azimuthal alignments which may not allow to align optimally the homologous section. This can be seen from Eq. (9), where the optimum phase $\Delta\Phi_0$ depends on the spatial average of the accumulated mismatch, the integral of $\delta\Omega_1(z') - \delta\Omega_2(z')$. For $L_{NH} \ll L_H$, $\Delta\Phi_0$ lies close to the value that minimizes the energy for the homologous segments, and we would expect a significant recognition energy. However, as L_{NH} increases, $\Delta\Phi_0$ will change to a value closer to one that minimizes the energy of the uncorrelated sequence segments, away from the optimum alignment for the homologous segments; therefore, the magnitude of the recognition energy should decrease. This prediction is confirmed in the MC simulations although this happens at a much slower rate than suggested by the perturbation theory (for the plots see Appendix D of the supplementary material⁴⁸).
- ⁵⁰D. J. Lee, e-print [arxiv:1312.1568](http://arxiv.org/abs/1312.1568), <http://arxiv.org/abs/1312.1568>.
- ⁵¹In the mean field electrostatic model of charged interacting helices [A. A. Kornyshev and S. Leikin, *J. Chem. Phys.* **107**, 3656 (1997)], these correspond to the parameter values $\theta = 0.85$, $f_1 = 0.35$ and $f_2 = 0.65$ (for details and formulas see Ref. 25) with inter-axial separation $R = 30$ Å.
- ⁵²This value represents the mid-point of experimental values presented in [A. A. Kornyshev, D. J. Lee, S. Leikin, and A. Wynveen, *Rev. Mod. Phys.* **79**, 943 (2007)]. In later works we have sometimes used a higher value of $C/k_B T = 1000$ Å, reported in [J. Lipfert, J. W. J. Kerssemakers, T. Jager, and N. H. Dekker, *Nat. Methods* **7**, 977 (2010)]. This different choice does not change much the results presented here.
- ⁵³A. Noy and R. Golestanian, *Phys. Rev. Lett.* **109**, 228101 (2012).
- ⁵⁴This result has some implication for experiments similar to those of Baldwin *et al.* (Ref. 38). In these experiments two different sequences of DNA were used, labelled by different colour fluorophores and were observed for fragments of 298 base pairs, within cholesteric spherulites formed at mild osmotic stress. In this case, and for shorter fragments, the DNA can be approximated as rigid. Due to the nature of the labelling, these segregation experiments cannot distinguish between anti-parallel and parallel alignments. Based on our results, we might reasonably expect some phase segregation could happen for fragments with less than 250 base pairs. However, each region, containing only one of the types of base pair sequence, could consist of either mixtures of anti-parallel and parallel aligned homologous molecules or smaller domains of parallel aligned sequences orientated in either two directions. The study of this situation poses an interesting problem in statistical mechanics. Though, to detect these alignments requires different experiments.
- ⁵⁵D. J. Lee and A. Wynveen, *J. Phys.: Condens. Matter* **18**, 787 (2006).
- ⁵⁶If the recognition well is sufficiently deep it should trap the homologous DNA in optimal register regardless of width. However, if one considers the dynamics, one can imagine thermal “collisions” that bring parts of the DNA close each other due to Brownian motion. The narrower the recognition well, the more likely the collision will miss the target (not have significant recognition energy) and no pairing between homologs will be initiated.
- ⁵⁷A. G. Cherstvy, *J. Mol. Recognit.* **24**, 283 (2011).
- ⁵⁸The recognition of some small identical base tracts is known to happen in site specific recombination processes [N. D. F. Grindley, K. L. Whiteson, and P. A. Rice, *Annu. Rev. Biochem.* **75**, 567 (2006)]. This recognition is facilitated by proteins that recognize a particular base pair sequence.
- ⁵⁹M. Prentiss, private communication (2014).
- ⁶⁰D. J. Lee, e-print [arxiv:1411.4844](http://arxiv.org/abs/1411.4844), <http://arxiv.org/abs/1411.4844>.
- ⁶¹R. Cortini, D. J. Lee, and A. Kornyshev, *J. Phys.: Condens. Matter* **24**, 162203 (2012).
- ⁶²J. F. Marko and E. D. Siggia, *Phys. Rev. E* **52**, 2912 (1995).
- ⁶³J. Ubbink and T. Odijk, *Biophys. J.* **76**, 2502 (1999).