Sequence Recognition in the Pairing of DNA Duplexes

A. A. Kornyshev
Research Center "Jülich," D-52425 Jülich, Germany

S. Leikin*
Laboratory of Physical and Structural Biology, National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, Maryland 20892

(Received 20 July 2000)

Pairing of DNA fragments with homologous sequences occurs in gene shuffling, DNA repair, and other vital processes. While chemical individuality of base pairs is hidden inside the double helix, x ray and NMR revealed sequence-dependent modulation of the structure of DNA backbone. Here we show that the resulting modulation of the DNA surface charge pattern enables duplexes longer than ~50 base pairs to recognize sequence homology electrostatically at a distance of up to several water layers. This may explain the local recognition observed in pairing of homologous chromosomes and the observed length dependence of homologous recombination.

DOI: 10.1103/PhysRevLett.86.3666 PACS numbers: 87.14.Gg

Homologous recombination (exchange of genetic material) between two parental copies of DNA allows cells, e.g., to repair damaged DNA and to shuffle genes. It lies at the very heart of storage, processing, and transfer of genetic information. It has several pathways, involves many steps, and requires a number of specialized proteins [1]. Its initial step is sequence recognition and alignment of homologous fragments (identical genes) on parental copies of DNA. This is the key to avoiding recombination mistakes (rare but often devastating, e.g., carcinogenesis). Nevertheless, sequence recognition is probably the least understood step in the whole process.

According to textbooks "we know only one mechanism for nucleic acids to recognize one another on the basis of sequence: complementarity between *single strands*" [2]. However, recent observations suggest that homologous recombination is preceded by recognition and local pairing of *intact, duplex* (double stranded) DNA fragments [3]. This pairing does not involve known recombination proteins. It has been attributed to direct DNA-DNA interactions whose physical origin has not been understood [3,4]. In the present study we suggest a possible explanation by showing that two homologous duplexes can recognize each other electrostatically.

The essence of the idea is illustrated in Fig. 1. Two DNA fragments with homologous sequences can form an electrostatically favorable alignment with negatively charged strands facing positively charged grooves [5] over a large juxtaposition length [Fig. 1(b)]. Long DNA fragments with unrelated sequences cannot establish such alignment because uncorrelated sequence-dependent modulations of the helical pitch disrupt the strand-groove register [Fig. 1(c)].

The model.—The strength and accuracy of recognition clearly depend on the length of DNA fragments in juxtaposition. This dependence determines the viability of the recognition mechanism. To find it, consider interaction between two parallel DNA duplexes A and B. To take into account realistic, helical patterns of their surface charges, let us use the model shown in Fig. 1(a). We describe the alignment of helices in terms of $\delta\phi(z) = \phi_A(z) - \phi_B(z)$, where $\phi_A(z)$ and $\phi_B(z)$ are the azimuthal orientations of the middle of the minor groove at the axial position z.

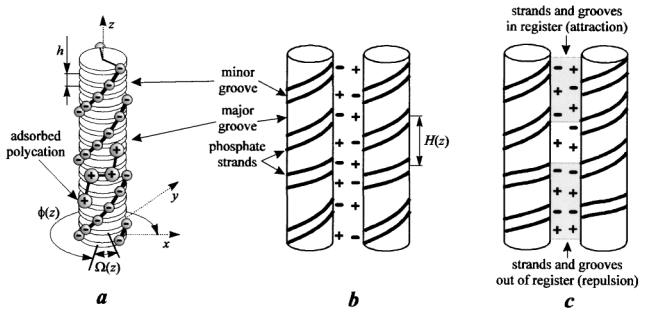

FIG. 1. (a) B-DNA schematically drawn as a stack of base pairs (disks). Each base pair has two negatively charged phosphate groups. We describe base pair orientation at the axial position z by the azimuthal angle $\phi(z)$ of the middle of the minor groove. Each combination of adjacent base pairs has a preferred twist angle $\Omega = \langle\Omega\rangle \pm \Delta\Omega$, where $\langle\Omega\rangle = 34°-35°$ and $\Delta\Omega = 4°-6°$ [7,8]. In the lowest energy conformation, $h[d\phi(z)/dz] = \Omega(z)$. Deviations from this conformation are described by Eq. (6). To calculate the energy of interaction between two parallel DNA, we describe DNA core as a cylinder with a low dielectric constant and phosphate strands as negatively charged spiral lines whose geometry is determined by $\phi(z)$. We account for adsorbed polycations (that tend to bind to DNA in cells) by introducing an excess positive surface charge density in the middle of each groove [5]. (b) The sequence-dependent twist modulation, $\Omega(z)$, leads to axial variation of the local helical pitch, $H(z)$. As a result, only DNA with homologous sequences can have negatively charged strands facing positively charged grooves over a large juxtaposition length. [To improve visual perception, the variation of $H(z)$ is strongly exaggerated.] (c) Molecules with unrelated sequences have uncorrelated $H(z)$ resulting in the loss of register between opposing strands and grooves and in higher juxtaposition energy.

 0031-9007/01/86(16)/3666(4)$15.00

For DNA fragments whose length L is much larger than their helical pitch, $H \approx 34$ Å, the dependence of the electrostatic interaction energy on L and $\delta\phi(z)$ is given by [6]

$$E_{\text{int}} \approx \int_0^L \{a_0 - a_1 \cos[\delta\phi(z)] + a_2 \cos[2\delta\phi(z)]\}\, dz\,, \tag{1}$$

where, for a given interaxial distance R between the duplexes, a_n are just numerical coefficients.

These coefficients can be calculated from [6]

$$a_0 = \frac{8\pi^2\sigma^2}{\varepsilon}\left\{\frac{(1-\theta)^2 K_0(\kappa R)}{\kappa^2[K_1(\kappa r)]^2} - \sum_{n,j=-\infty}^{\infty} \frac{[f(n,\theta)]^2}{\kappa_n^2}\left[\frac{[K_{n-j}(\kappa_n R)]^2 I_j'(\kappa_n r)}{[K_n'(\kappa_n r)]^2 K_j'(\kappa_n r)}\right]\right\}, \tag{2}$$

$$a_{n=1,2} = \frac{16\pi^2\sigma^2}{\varepsilon}\,\frac{[f(n,\theta)]^2}{\kappa_n^2}\,\frac{K_0(\kappa_n R)}{[K_n'(\kappa_n r)]^2}\,, \tag{3}$$

$$f(n,\theta) = f_1\theta + f_2(-1)^n\theta - (1 - f_3\theta)\cos(0.4n\pi)\,, \tag{4}$$

$$\kappa_n = \sqrt{\kappa^2 + n^2\left(\frac{2\pi}{H}\right)^2}\,, \tag{5}$$

where $I_n(x)$, $K_n(x)$, $I_n'(x)$, and $K_n'(x)$ are the modified Bessel functions and their derivatives, respectively; $r \approx 9$ Å is the radius of the cylindrical surface formed by centers of phosphates; $\sigma \approx 16.8\ \mu\text{C/cm}^2$ is the surface charge density of phosphates; θ is the fraction of phosphate charge neutralized by bound counterions; f_i are the fractions of counterions bound in the minor groove (f_1), in the major groove (f_2), and on the phosphate strands (f_3), $f_1 + f_2 + f_3 = 1$; $\varepsilon \approx 80$ is the dielectric constant of water; κ^{-1} is the Debye screening length (≈ 7 Å in physiological solution) [6].

Let us now take into account the fact that DNA consists of base pairs (bp) stacked with the axial step $h \approx 3.4$ Å/bp and twisted with respect to each other [Fig. 1(a)]. The preferred twist angle between adjacent base pairs $\Omega(z)$ is a "fingerprint" of the sequence [7,8]. The actual twist angle, $h[d\phi(z)/dz]$, may differ from $\Omega(z)$, because intermolecular interaction and thermal fluctuations may cause torsional deformation of DNA. To find $\phi(z)$ we can use the torsional energy

$$E_t = \frac{1}{2}\,C\int_0^L\left[\frac{d\phi(z)}{dz} - \frac{\Omega(z)}{h}\right]^2 dz\,, \tag{6}$$

where $C \approx 3 \times 10^{-19}$ erg cm [9] is the torsional rigidity modulus. The most favorable alignment of two DNA in close juxtaposition minimizes the sum of the torsional energy and the interaction energy given by Eq. (1). Hence, this alignment satisfies the following equation:

$$\lambda_t^2\,\frac{d^2[\delta\phi(z)]}{dz^2} - \sin[\delta\phi(z)](1 + b\sin^2[\delta\phi(z)/2]) = \frac{\lambda_t^2}{h}\,\frac{d[\delta\Omega(z)]}{dz}\,. \tag{7}$$

Here $b = 8a_2/(a_1 - 4a_2)$ and $\lambda_t = \sqrt{C/2(a_1 - 4a_2)}$.

Note that at large distances between DNA, $a_2 \ll a_1$ and $b \ll 1$ and Eq. (7) reduces to a time-independent sine-Gordon equation in a random field. It has a variety of solutions, including stable nonlinear waves (solitons). Combining statistics of solitons and statistics of small perturbations, one can calculate the partition function and determine the sequence-dependent free energy of interaction between two opposing DNA from Eqs. (1)–(7).

Here, we consider only the simplest case when we can assume that $d\phi(z)/dz \approx \Omega(z)/h$. This approximation should work reasonably well for DNA fragments shorter than $\sim$200 bp ($L < 700$ Å), since for such fragments $L < \lambda_t$ and $L < \lambda_p = C/k_BT$ ($\lambda_t > 500$–700 Å, $\lambda_p \approx 750$ Å).

Within this approximation, the interaction energy at the most favorable alignment is

$$E_{\text{int}} \approx [a_0 - \nu_1(L)a_1 + \nu_2(L)a_2]L\,, \tag{8}$$

where we introduced the *recognition coefficients*

$$\nu_n(L) = \frac{1}{L}\int_0^L \cos\left[\frac{n}{h}\int_0^z [\Omega_A(z') - \Omega_B(z')]\,dz'\right]dz\,. \tag{9}$$

For identical sequences, $\delta\Omega(z) = \Omega_A(z) - \Omega_B(z) = 0$ and $\nu_n(L) = 1$. For unrelated sequences, $\delta\Omega(z)$ is a "random walk" with uncorrelated steps [10]. Using the Gaussian statistics with $\langle\delta\Omega^2\rangle = 2\Delta\Omega^2$ ($\Delta\Omega \approx 0.07$–$0.1$ rad [7,8] is the rms variation in Ω for each sequence), we find

$$\nu_n(L) = F\left(n^2\,\frac{L}{\lambda_c}\right), \qquad F(x) = \frac{1 - e^{-x}}{x}\,, \tag{10}$$

where λ_c is a *helical coherence length* of DNA,

$$\lambda_c = \frac{h}{\Delta\Omega^2} \approx 300\text{–}700\ \text{Å}\,. \tag{11}$$

From Eqs. (8)–(11), we find the difference in the energy of juxtaposition between fragments with unrelated sequences and between fragments with identical sequences,

$$\Delta E \approx \{[1 - F(L/\lambda_c)]a_1 + [F(4L/\lambda_c) - 1]a_2\}L\,, \tag{12}$$

that one can refer to as the sequence recognition energy.

Recognition efficiency.—Let us calculate ΔE using parameters that mimic biologically relevant interactions ($\kappa^{-1} = 7$ Å, $\theta = 0.8$, $f_1 = 0.3$, $f_2 = 0.7$, and $f_3 = 0$ [5,11]). At 10 Å surface-to-surface separation ($R = 30$ Å), we find $a_0 \approx 3.0 \times 10^{-8}$ erg/cm, $a_1 \approx 6.0 \times 10^{-8}$ erg/cm, and $a_2 \approx 1.4 \times 10^{-8}$ erg/cm. Figure 2 illustrates the corresponding dependence of E_{int} and ΔE (inset) on L. The recognition energy exceeds the thermal energy k_BT for DNA fragments longer than 100 bp. Taking into account that the strength of interaction increases

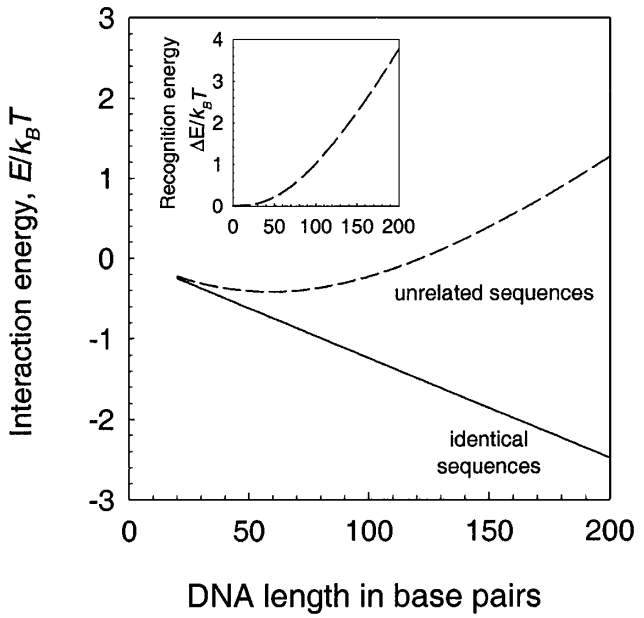

FIG. 2. Interaction energy (in units of thermal energy, k_BT) versus length of two parallel DNA fragments at 10 Å surface separation ($R = 30$ Å) calculated from Eqs. (8) and (10) at $\Delta\Omega = 6°$ ($\lambda_c = 310$ Å). Negative values indicate favorable juxtaposition compared to infinite separation. Inset: The dependence of the recognition energy (the energy gain from juxtaposition between homologous fragments compared to unrelated fragments) on the fragment length calculated from Eq. (12).

exponentially upon decreasing separation (approximately threefold for each 3–5 Å) [6,12], we estimate that at 5–10 Å surface separations ~50–200 bp DNA fragments should recognize each other with the energy gain $\sim(1-15)k_BT$.

This is exactly what one would expect for recognition energy in biological reactions. For instance, regulatory proteins bind to their target sites on DNA with 10^2-10^7 times higher probability than to nonspecific sites [13]. Since every factor of 10 corresponds to $2.3k_BT$, their recognition energy is $(5-16)k_BT$. Such energy is optimal for combining efficient recognition with rapid, on-demand search capabilities (it is larger than k_BT but not too large to prohibit dissociation). This is particularly important in the dense environment of a cell nucleus.

Interaction between nucleic acids in vitro.—We are not aware of *in vitro* sequence recognition measurements. Thus, our results should be viewed as conjectural. Still, many of the features of interaction between DNA duplexes built into our model have been experimentally observed.

Specifically, alignment of opposing strands and grooves was observed in oligonucleotide crystals [14], in hydrated, quasicrystalline fibers of natural DNA [15], and in crystals of nucleosome core particles [16]. The alignment becomes particularly favorable in the presence of cations capable of specific binding to nucleic acids, because such cations produce larger excess positive charge in grooves. The resulting attraction between opposing strands and grooves can overwhelm the repulsion between phosphate strands and lead to spontaneous aggregation of homopolymeric (or homologous) helices. Our predictions for such mechanism of aggregation [6] were confirmed experimentally for homopolymeric guanosine helices [17].

Although counterion-induced aggregation of homologous DNA has not been studied, it is known that a sufficiently high concentration of polyamines and several other ions causes spontaneous aggregation of even random DNA fragments [18]. We believe that this occurs when the strand-groove attraction becomes strong enough not only to compete with the repulsion between strands but also to induce torsional deformation and establish the strand-groove register. However, non-sequence-specific forces could contribute to this phenomenon as well [19].

Phosphate backbone structure and intermolecular interaction.—Torsional deformation is observed when DNA fragments with *unrelated sequences* are forced into close juxtaposition (≤10 Å surface separation) in hydrated aggregates. DNA duplexes have incommensurate $\Omega(z)$ ($\langle\Omega\rangle \approx 34°$) in solution and constant $\Omega = 36°$ (10.0 bp/turn) in aggregates [20]. In contrast, oligonucleotides with *identical sequences* crystallize in close juxtaposition without significant distortion of their base pair twist angles (as indicated by similar values of preferential twist angles deduced from x-ray crystal structures and from solution data; see [8]).

In other words, pairing of helices with uncorrelated base pair sequences requires DNA backbone overwinding [that can cost up to $\sim 0.5k_BT$ per base pair, Eq. (6)]. Pairing of helices with identical sequences does not require backbone deformation so that it is more energetically favorable. While one may argue that it is not a direct proof, this deduction suggests that sequence homology recognition does occur between duplex DNA *in vitro*.

Duplex-duplex recognition in vivo.—In cells, DNA bending and proteins may modify the interaction so that our equations may no longer work, but the same physical principles will still apply. It was suggested, e.g., that DNA pairing in eukaryotes involves homology-dependent local contacts between duplexes in nucleosome-free regions [21]. Electrostatic recognition could cause (or contribute to) this pairing. It could pair nucleosomal DNA as well (additional twist modulation of DNA in nucleosomes may even enhance the recognition efficiency).

Such a hypothesis can explain local pairing of duplex DNA on homologous chromosomes observed [3,4] in yeast. It can also explain why 50–200 bp homology is necessary for efficient homologous recombination in bacteriophages [22], bacteria [23], and mammalian cells [24]. This is the requirement for the electrostatic pairing of duplex DNA.

Homologous recombination.—Finally, let us point out that homologous recombination is commonly believed to be initiated by breaks in duplex DNA. Specialized proteins of RecA family coat single strands produced at such breaks and promote their association with homologous duplex DNA fragments [1,25]. However, this recognition requires only eight base pairs [26] and identical 8 bp sequences

must occur at least every $4^8 = 65\,536$ bp, e.g., any 8 bp sequence is repeated $\sim 100\,000$ times within the human genome. If this were the only sequence recognition mechanism in homologous recombination, frequent recombination mistakes would be inevitable.

The requirement of 50–200 bp homology is critical for avoiding mistakes [22–24], but such recognition cannot be precise since the corresponding energy per base pair is only 0.5%–2% of the total energy. Hence, a limited number of base pair mismatches should be tolerated. This could also lead to recombination mistakes if no additional control were present. A logical solution of this dilemma seems to be a two step recognition — first a coarse-grain alignment of fragments with 50–200 bp homology and only then a precise match of ~10 bp. The observed local pairing of intact DNA duplexes that precedes homologous recombination [3] is likely to be the first step. Our model of electrostatic duplex-duplex recognition may explain its mechanism. The synaptic complex formation between a single strand and duplex DNA mediated by RecA family proteins is likely to be the second step.

The authors are grateful to Per Hansen, Hartmut Loewen, Adrian Parsegian, Donald Rau, and Victor Zhurkin for many stimulating discussions and critical remarks. A. A. K. acknowledges the support of this work by the Deutsche Forschungsgemeinschaft, Grant No. KO 1391/4-1, and the financial support of his visits to Bethesda by the National Institute of Child Health and Human Development, NIH.

*Corresponding author.
Email address: leikin@helix.nih.gov

[1] D. R. F. Leach, *Genetic Recombination* (Blackwell Science, Oxford, 1996).
[2] B. Lewin, *Genes VI* (Oxford University, Oxford, 1997).
[3] S. M. Burgess, N. Kleckner, and B. M. Weiner, Genes Dev. **13**, 1627 (1999).
[4] B. M. Weiner and N. Kleckner, Cell **77**, 977 (1994).
[5] Cells contain many polycations that tightly bind to DNA and are likely to be responsible for neutralizing most of the DNA charge. For instance, polyamines (spermine and spermidine) are present in all cells in multimillimolar concentrations. They stimulate transcription and translation and they are essential for sperm formation [see, e.g., P. Coffino, Proc. Natl. Acad. Sci. U.S.A. **97**, 4421 (2000)]. Polyamines form hydrogen bonds with phosphate oxygens and bases and bind in DNA grooves, particularly in the major one [see, e.g., X. Shui, L. McFail-Isom, G. G. Hu, and L. D. Williams, Biochemistry **37**, 8341 (1998); M. Yuki, V. Grukhin, C.-S. Lee, and I. S. Haworth, Arch. Biochem. Biophys. **325**, 39 (1996)]. This creates surface charge separation with excess negative charge on phosphate strands and excess positive charge in grooves (Fig. 1). Note that even mobile counterions, such as Na^+, tend to condense in DNA grooves because of the combination of hard core and electrostatic interactions [see, e.g., L. McFail-Isom, C. S. Sines, and L. D. Williams, Curr. Opin. Struct. Biol. **9**, 298 (1999); T. K. Chiu and R. E. Dickerson, J. Mol. Biol. **301**, 915 (2000)].
[6] A. A. Kornyshev and S. Leikin, J. Chem. Phys. **107**, 3656 (1997); Phys. Rev. Lett. **82**, 4138 (1999).
[7] W. Kabsch, C. Sander, and E. N. Trifonov, Nucl. Acid Res. **10**, 1097 (1982); A. A. Gorin, V. B. Zhurkin, and W. K. Olson, J. Mol. Biol. **247**, 34 (1995).
[8] W. K. Olson, A. A. Gorin, X.-J. Lu, L. M. Hock, and V. B. Zhurkin, Proc. Natl. Acad. Sci. U.S.A. **95**, 11 163 (1998).
[9] D. M. Crothers, J. Drak, J. D. Kahn, and S. D. Levene, in *Methods in Enzymology,* edited by D. M. J. Lilley and J. E. Dahlberg (Academic Press, San Diego, 1992), Vol. 212 B, p. 3.
[10] Note that we do not assume that either $\Omega_A(z)$ or $\Omega_B(z)$ are random. However, if the sequences are unrelated, $\Omega_A(z) - \Omega_B(z)$ becomes a "random walk" with uncorrelated steps. For sequences ~100 bp (i.e., ~100 steps) we are interested in, Gaussian statistics should work very well.
[11] The ratio of the width of minor and major grooves is $\approx 0.4/0.6$. We use $f_1/f_2 = 0.3/0.7$ rather than $0.4/0.6$ to account for preferential adsorption of biologically relevant cations in the major groove.
[12] D. C. Rau, B. Lee, and V. A. Parsegian, Proc. Natl. Acad. Sci. U.S.A. **81**, 2621 (1984).
[13] L. Jen-Jacobson, Biopolymers **44**, 153 (1997).
[14] See, e.g., Y. Timsit and D. Moras, Methods Enzymol. **211**, 409 (1992). To avoid confusion, note that in some structures discussed in this paper oligonucleotides are not parallel, so that ridges of phosphate strands can fit into grooves. Such strand-groove alignment is, however, irrelevant for homology recognition at the scale of ~100 bp that is the subject of our work. This recognition requires ~100 bp juxtaposition and, therefore, parallel orientation of molecules.
[15] R. Langridge, H. R. Wilson, C. W. Hooper, M. H. F. Wilkins, and L. D. Hamilton, J. Mol. Biol. **2**, 19 (1960).
[16] K. Luger, A. W. Mader, R. K. Richmond, D. F. Sargent, and T. J. Richmond, Nature (London) **389**, 251 (1997).
[17] P. Mariani, F. Ciuchi, and L. Saturni, Biophys. J. **74**, 430 (1998).
[18] V. A. Bloomfield, Curr. Opin. Struct. Biol. **6**, 334 (1996).
[19] R. Marquet and C. Houssier, J. Biomol. Struct. Dyn. **9**, 159 (1991); J. Ray and G. Manning, Langmuir **10**, 2450 (1994); I. Rouzina and V. Bloomfield, J. Phys. Chem. **100**, 9977 (1996); N. Gronbech-Jensen, R. J. Mashl, R. F. Bruinsma, and W. M. Gelbart, Phys. Rev. Lett. **78**, 2477 (1997); B.-Y. Ha and A. J. Liu, Phys. Rev. Lett. **79**, 1289 (1997); **81**, 1011 (1998).
[20] D. Rhodes and A. Klug, Nature (London) **286**, 573 (1980); S. B. Zimmerman and B. H. Pfeiffer, Proc. Natl. Acad. Sci. U.S.A. **76**, 2703 (1979).
[21] S. Keeney and N. Kleckner, Genes Cells **1**, 475 (1996).
[22] B. S. Singer, L. Gold, P. Gauss, and D. H. Doherty, Cell **31**, 25 (1982).
[23] V. M. Watt, C. J. Ingles, M. S. Urdea, and W. J. Rutter, Proc. Natl. Acad. Sci. U.S.A. **82**, 4768 (1985).
[24] J. Rubnitz and S. Subramani, Mol. Cell Biol. **4**, 2253 (1984).
[25] R. D. Camerini-Otero and P. Hsieh, Annu. Rev. Genet. **29**, 509 (1995).
[26] P. Hsieh, C. S. Camerini-Otero, and R. D. Camerini-Otero, Proc. Natl. Acad. Sci. U.S.A. **89**, 6492 (1992).