Inhomogeneous DNA: Conducting exons and insulating introns

A. A. Krokhin,1,⇑  V. M. K. Bagci,1  F. M. Izrailev,2  O. V. Usatenko,3  and V. A. Yampol’skii3

1Department of Physics, University of North Texas, P.O. Box 311427, Denton, Texas 76203, USA

2Instituto de Fisica, Universidad Autónoma de Puebla, Apartado Postal J-48, Puebla 72570, Mexico

3A. Ya. Usikov Institute for Radiophysics and Electronics, Ukrainian Academy of Science, 12 Proskura Street, 61085 Kharkov, Ukraine

(Rceived 18 May 2009; revised manuscript received 8 July 2009; published 17 August 2009)

Parts of DNA sequences known as exons and introns play very different roles in coding and storage of genetic information. Here we show that their conducting properties are also very different. Taking into account long-range correlations among four basic nucleotides that form double-stranded DNA sequence, we calculate electron localization length for exon and intron regions. Analyzing different DNA molecules, we obtain that the exons have narrow bands of extended states, unlike the introns where all the states are well localized. The band of extended states is due to a specific form of the binary correlation function of the sequence of basic DNA nucleotides.

DOI: 10.1103/PhysRevB.80.085420

PACS number(s): 78.67.–n, 87.14.gk, 75.75.+a

I. INTRODUCTION

A DNA molecule is an exciting example of a natural complex system with intriguing properties. Many of these properties remain unexplained and need new approaches for further analysis. One of the fundamental questions is how information is transferred along a sequence of nucleotides. For example, if a mutation occurs in the sequence, it is usually healed. This means that some of physical parameters of the DNA molecule are sufficiently sensitive to detect this mutation. The length of a mutation is relatively short [2–10 base pairs (bps)] as compared to the length of a gene (103–106 base pairs). Because of small statistical weight of a mutation, the mechanical and thermodynamic characteristics are not sensitive enough for its robust detection. Unlike this, the electrical resistance of a DNA molecule strongly fluctuates even if a single nucleotide in a long sequence is replaced (or removed).1,2 This property is a signature of coherent transport that gives rise to universal fluctuations of conductance in mesoscopic samples.3

In a DNA molecule the charge carriers move along a double helix formed by two complementary sequences of four basic nucleotides: A, T, G, and C. A conduction band would form if the DNA texts would exhibit some periodicity.4 However, many studies of the DNA texts have revealed rich statistical properties but not the periodicity. One of the suggestions is that a DNA molecule is a stochastic sequence of nucleotides, the main feature of which is long-range correlations.5 Therefore a popular method of detection of correlations is mapping of a DNA sequence onto a random walk. Long-range correlations are manifested then in an anomalous scaling of the generated classical diffusion.6

Quantum transport through a DNA molecule is also strongly affected by the correlations. An uncorrelated sequence of nucleotides localizes all quantum electron states, as occurs in any one-dimensional (1D) white-noise potential, making impossible charge transfer at distances longer than the localization length \( l(E) \). However, since most of the mutations in DNA are successfully healed, one may assume the existence of charge transport7 through delocalized states that are responsible for the transfer of information at much longer distances. Such delocalized states are expected to exist within exons—the coding regions where the genetic information is stored. An important feature of charge transfer in carrying mutations exons was reported in Ref. 8. It was shown that cancerous mutations usually produce much less variation in the resistance than noncancerous ones. This apparent distinction shed light on the problem of survival of cancerous mutations. The healing of a mutation occurs only if it is detected by base excision repair enzymes. Since the detection of the mutation is most likely due to DNA-mediated charge transport,9 it is clear that cancerous mutations, being “electrically masked,” are very unlikely to be detected and then repaired.

On the other hand, the introns—the long segments of DNA that apparently do not carry genetic code—may not contain delocalized states in the energy spectrum, thus remaining insulators. In this paper we give evidence for the validity of this hypothesis using a theoretical approach based on the results of electron localization in correlated disordered potentials. Our study of various DNA molecules shows that the energy spectrum of the exons indeed contains practically delocalized states. Unlike this, the electron wave functions are well localized within the introns.

II. TWO-STRANDED MODEL OF DNA

Let us first consider the widely used model for electron transport in DNA molecules, which is a discrete lattice with random on-site potential \( \epsilon_n \) and site-independent nearest-neighbor hopping amplitude \( t \).

\[
t(\psi_{n+1} + \psi_{n-1}) = (\epsilon_n - E) \psi_n. \tag{1}
\]

The energies \( \epsilon_n \) are the ionization energies of the four nucleotides, \( \epsilon_A=8.24, \epsilon_T=9.14, \epsilon_G=8.87, \) and \( \epsilon_C=7.75 \) eV, and the hopping amplitude \( t \) may vary from 0.1 to 1 eV.10 Although the on-site energies in a sequence of coupled nucleotides do not coincide exactly with their ionization potentials, one may neglect this difference as it plays a minor role in our consideration. The regular periodic potential \( \epsilon_n=V_0 \) in Eq. (1) gives rise to Bloch functions \( \psi_n \propto \exp(i\mu n) \) with dispersion relation \( E-V_0=2t \cos \mu n \). The allowed energies of
these extended states lie in a single band of width $2t$, $|E - V_{0i}| < t$. In the opposite case of a white-noise potential, where $\langle \epsilon_i \epsilon_0 \rangle = 0$ and $\langle \epsilon_i \rangle = 0$, all the states are localized. For weak fluctuations, $\epsilon_i \approx t^2$, the Lyapunov exponent (inverse localization length) in the Born approximation is given by:

$$\gamma_0(E) = \frac{1}{I_0(E)} = \frac{\epsilon_0^2}{8t^2 \sin^2 \mu}. \quad (2)$$

In this approximation, the wave function extends over many sites, i.e., $I_0(E) \gg 1$ and the dispersion relation remains the same as for the regular potential, $E = 2t \cos \mu$.

Most of the existing random potentials are neither ideally periodic nor ideally disordered (white-noise potential). They form a wide class of so-called correlated disordered potentials. A generalization of Eq. (2) for this class of potentials was obtained in Ref. 12,

$$\gamma(E) = \gamma_0(E) \varphi(2\mu), \quad \varphi(\mu) = 1 + 2 \sum_{k=1}^{\infty} \xi(k) \cos(\mu k). \quad (3)$$

Here $\xi(k) = (\epsilon_{2n+k}^2) / \epsilon_0^2$ is the normalized binary correlator of the potential. Because of the correlations the energy spectrum may contain localized as well as extended states. In a first approximation over $\epsilon_0$, the extended states occupy the intervals where the function $\varphi(\mu)$ in Eq. (3) vanishes. The regions of localized and extended states are separated by a “mobility edge.” For example, a sharp vertical mobility edge at $\mu = \pi/3(E = t)$ appears if the correlation function decays slowly and oscillates: $\xi(k) = (3/2 \pi k) \sin(2 \pi k / 3)$. In Ref. 13 this type of sharp mobility edge was observed in the transmission and reflection spectra of a microwave waveguide with specially designed correlated scatterers. Power-law decay and oscillations of $\xi(k)$ are the necessary (although not sufficient) attributes of a sharp mobility edge in the energy spectrum. We studied the correlation function of many different DNA sequences and all of them exhibit slow decay and oscillations. A typical correlation function is shown in Fig. 1. A different approach to the Anderson transition in 1D potentials with long-range correlations has been developed by de Moura and Lyra. It is based on the method of generation of a random correlated sequence that is adopted from the theory of fractional Brownian motion. The presence or absence of a mobility edge is determined by the scaling properties of power spectrum of binary correlation function but not by the binary correlator itself.

Although Eq. (3) correctly accounts for the correlations in a single-channel random potential, it is not appropriate for the analysis of electron localization in real DNA. A DNA molecule is a two-stranded sequence of nucleotides, i.e., there are two conducting channels. It is known that the localization length strongly depends on the number of channels in disordered chains. Our case is even more specific since two strands, being random in the longitudinal direction, exhibit regular A-T and C-G matching in the transverse direction. This key-lock matching between the strands strongly affects electron transport in DNA.

To date, there have been a number of studies of the localization length in DNA molecules. In Ref. 18 an attempt was done to obtain numerically a localization-delocalization transition in a single-stranded binary artificial DNA sequence with a special kind of slowly decaying correlations. However, since in the thermodynamic limit the proposed sequence turned out to be a regular one, the problem of extended states remains open. In the numerical study an unexpected tendency to delocalization with an increase in nonperturbative disorder in an uncorrelated single-stranded DNA was observed. Recently it was claimed that the transverse key-to-lock base pairing by itself gives rise to a band of extended states even if the longitudinal correlations are ignored. This numerical result has since been criticized using analytical argumentation. Thus, it is now clear that for a correct evaluation of the localization length in DNA one has to (i) use the two-stranded model; (ii) avoid simplification of the four-letter DNA alphabet to a binary sequence; and (iii) account for the longitudinal correlations in both strands as well as transversal base pairing.

In our study we use a two-channel model where the Schrodinger equation reads:

$$t(\psi_{1,n+1} + \psi_{1,n-1}) + h \psi_{2,n} = (E - \epsilon_{1,n}) \psi_{1,n},$$

$$t(\psi_{2,n+1} + \psi_{2,n-1}) + h \psi_{1,n} = (E - \epsilon_{2,n}) \psi_{2,n}. \quad (4)$$

Here $\psi_{1,n}$ and $\psi_{2,n}$ are the on-site wave functions in the first and second chains, respectively, and $\epsilon_{1,n}$ and $\epsilon_{2,n}$ are the on-site potentials. The hopping parameters $t$ and $h$ determine the interstrand and intrastrand couplings.

In the case of a periodic potential there are two momenta $\mu_1$ and $\mu_2$ for each energy $E$. They are given by two dispersion relations, $E = 2t \cos \mu_1 \pm h$. Here we consider the case of a band structure when the two propagating channels overlap. This happens if $h < 2t$. The band of allowed energies spreads from $-2t - h$ to $2t + h$. Both channels are propagating (i.e., $\mu_1$ and $\mu_2$ are real) for the energies $|2t - h| < E$.

### III. LOCALIZATION LENGTH

For calculation of the localization length we use the perturbation-theory approach developed for two- and three-
channel waveguides in Ref. 16. The localization length is defined by the following formula:

$$\gamma(E) = \Gamma^{-1}(E) = -\lim_{N \to \infty} \frac{1}{2N} \ln \text{Tr}(\hat{t}^2)$$

where $\langle \ldots \rangle$ denotes averaging over disorder and $\hat{t}$ is a $2 \times 2$ transmission matrix. The transmission matrix $\hat{t}$ that enters into the Landauer formula $g = (2e^2/h)\ln \text{Tr}(\hat{t}^2)$ is calculated as a product of $N$ on-site transfer matrices. The transmission matrix is calculated in the linear (Born) approximation over weak disorder $\langle \epsilon_1^2 \rangle, \langle \epsilon_2^2 \rangle \ll \hbar^2$. The results reported here are based on the following formula for the Lyapunov exponent:

$$\gamma(E) = \frac{\epsilon_1^2}{64\hbar^2} \left[ \frac{\varphi_{11}(2\mu_1)}{\sin^2 \mu_1} + \frac{\varphi_{11}(2\mu_2)}{\sin^2 \mu_2} + \frac{2\varphi_{11}(\mu_1 + \mu_2)}{\sin \mu_1 \sin \mu_2} \right]$$

$$+ \frac{\epsilon_2^2}{64\hbar^2} \left[ \frac{\varphi_{22}(2\mu_2)}{\sin^2 \mu_2} + \frac{\varphi_{22}(2\mu_1)}{\sin^2 \mu_1} + \frac{2\varphi_{22}(\mu_1 + \mu_2)}{\sin \mu_1 \sin \mu_2} \right]$$

$$+ \frac{\epsilon_{12}}{32\hbar^2} \left[ \frac{\varphi_{12}(\mu_2)}{\sin^2 \mu_2} + \frac{\varphi_{12}(\mu_1)}{\sin^2 \mu_1} + \frac{2\varphi_{12}(\mu_1 + \mu_2)}{\sin \mu_1 \sin \mu_2} \right].$$

In the two-channel model electron localization occurs due to backscattering processes in both channels with intrachannel momenta transfers $2\mu_1$ and $2\mu_2$. There is also interchannel scattering with momentum transfer $\mu_1 + \mu_2$. Accordingly, there are terms $\varphi_{11}(2\mu_1), \varphi_{22}(2\mu_2)$, and $\varphi_{12}(\mu_1 + \mu_2)$ in Eq. (6). The functions $\varphi_{ij}$ are expressed through three binary correlators $\xi_{ij}$, similarly to Eq. (3),

$$\varphi_{ij}(\mu) = 1 + 2 \sum_{k=1}^{\infty} \xi_{ij}(k) \cos(\mu k) \quad i,j = 1,2.$$

These functions, $\xi_{11}, \xi_{22},$ and $\xi_{12}$ characterize the intrachannel and interchannel correlations, respectively,

$$\langle \epsilon_{1,1} \epsilon_{1,k} \rangle = \xi_{11}^2(k), \quad \langle \epsilon_{2,2} \epsilon_{2,k} \rangle = \xi_{22}^2(k),$$

$$\langle \epsilon_{1,1} \epsilon_{2,k} \rangle = \xi_{12} \xi_{12}^2(k).$$

Here the mean value $\xi_{12} = \langle \epsilon_{1,1} \epsilon_{2,k} \rangle$ can be either positive or negative, unlike always positive variances $\xi_{1,1}^2$. Equation (6) is valid if both channels are propagating, i.e., the wave numbers $\mu_1$ and $\mu_2$ are real. If one of the channels becomes evanescent it is replaced by Eq. (33) from Ref. 22.

$$\gamma(E) = \frac{1}{32 \sin^2 \mu_1} \left[ \xi_{11}^2 \varphi_{11}(2\mu_1) + \xi_{22}^2 \varphi_{22}(2\mu_1) + 2 \xi_{12} \varphi_{12}(2\mu_1) \right].$$

At the transition points when $E = E_c = 2|t - \hbar|$, one of the denominators in Eq. (6) vanishes ($\sin \mu_{1,2} = 0$) and the Born approximation fails.

We apply Eq. (6) to a two-stranded DNA molecule. Among a huge number of chemical and physical characteristics of a DNA molecule, we need here only the ionization potentials for each nucleotide, $E_t, E_t, E_c,$ and $E_c$, and two hopping amplitudes, $t$ and $\hbar$. Unlike previous studies, we develop here an analytical two-channel approach, which accounts for intrachannel and interchannel correlations. Therefore, we do not simplify a two-stranded DNA sequence to a binary sequence, using a coarse-graining procedure. From this point of view, our approach is much more close to reality.

The length of a DNA sequence may reach $10^6 - 10^9$ base pairs. In such a long disordered chain all the states are localized and a DNA molecule does not conduct. Much shorter segments may, however, exhibit very different behavior.\(^\text{1,10}\)

This means that the conducting properties of a DNA molecule vary along the sequence of nucleotides and explains a wide spectrum of conducting properties obtained in experiments; see Ref. 10. The physical characteristics, such as the ionization potential and hopping amplitudes, are independent of the position of a given nucleotide in the sequence. The only characteristics which may change along the sequence are the correlation functions. The exons and introns store different kind of information and this affects the correlators. Thus, the localization length and the conductance of a given segment of a DNA molecule are directly related to the genetic information stored in this segment. Equation (6) is a mathematical manifestation of this fact.

### IV. NUMERICAL RESULTS FOR LOCALIZATION LENGTH

In order to demonstrate the inhomogeneities in the conductivity of the DNA molecules we studied the localization length along the exons and introns. Exons are the parts where the genetic information is written and introns are the parts without apparent information for protein synthesis. The introns occupy a larger part of the DNA sequence of higher eukaryotes than do the exons. For procaryotes the situation is the opposite. From the point of view of “quality” of the carried information the introns and exons are the most different segments and one may expect very distinct localization properties to exist in these segments. It was recently shown that the melting of exons and introns also occurs in a different way.\(^\text{23}\)

We use Eqs. (6) and (9) for numerical calculation of the localization length. Here we give the results for the following human DNA molecules: BRCA, ADAM10, SNAP29, and SUHW1. The results are shown in Figs. 2–5 where we plot the localization length vs electron energy for the exon and intron segments. The parameters of nucleotide site energies and the hopping amplitudes are the same for all these figures. Thus, the very different patterns shown in the figures represent different information codes in different DNAs.

For most of the energies the localization length inside the exon region exceeds by order of magnitude the localization length inside the intron region. This confirms, by implication, the fact that very different kinds of information are coded in these regions. The vertical axis for each figure is cut off approximately at the length of the corresponding exon region. There are many peaks in the exon regions with the height that exceeds much the vertical scale, i.e., the states within these peaks are extended. Unlike this, in the intron regions all the states are well localized. The density of the peaks in Figs. 2 and 3 is much higher than that in Figs. 4 and
5. Most of the peaks are situated in the region of energy where one of the channels is evanescent. Similar sharp peaks in the transmission of the exon regions of Y3 DNA have been numerically obtained in Ref. 24 for a single-stranded DNA. It turns out that this feature is very robust since in that study a single-stranded model of DNA was used.

The fine structure of one of the peaks is shown in the inset of Fig. 3. Since the peaks are of a finite width (~20 meV), they are narrow bands of extended states but not the discrete resonant states predicted and observed in random dimers. The nature of resonant tunneling in random dimers is due to short-range correlations in contrast with specific long-range correlations which are necessary for existence of a continuous band of extended states. In the case of a single channel the width of the band of extended states can be controlled by the parameters of the binary correlator \( \xi(k) \) in Eq. (3). In particular, wide and narrow bands of the extended states have been observed in the experiments with single-mode microwave waveguides. For the two-channel system the relation between the positions of the mobility edges and the explicit form of the binary correlator is not known. One may expect that such relation is determined by the relative phase shifts between the Fourier components of the oscillatory correlators \( \xi_{ij}(k) \). It is worth mentioning that short- and long-range correlations lead not only to different localization properties but also to very different classical as well as quantum diffusion in DNA.

A pattern \( l(E) \) is a particular fingerprint of a given DNA sequence and it can be used, in principle, for classification of DNA molecules. In the previous studies (see, e.g., Refs. 6 and 27) the DNA sequences have been characterized by scaling exponent of the corresponding random walk. We consider that the inverse localization length [Eq. (6)] is more convenient since it characterizes a well-defined physical property—electrical resistivity. Moreover, Eq. (6) establishes a qualitative relation between the localization length and the informational characteristic (binary correlators) of the DNA sequence. At the same time the binary correlators by themselves are not very illustrative. In particular, the plots of the correlators for exon and intron regions look very similar, see Fig. 6, although these plots, of course, contain the same information about the DNA sequence as the plots for the

FIG. 2. (Color online) Localization length vs energy for the human BRCA gene measured in the number of base pairs. The length of the exon (intron) is 2120 (10 421) bps. The results for exon and intron are shown by black and gray (red) lines, respectively. The values of the hopping parameters are \( h=0.5 \) eV and \( t=1 \) eV. The two channels are propagating if \( 6.6<E<10.4 \) eV. One of the channels becomes evanescent in two symmetric regions, \( 10.4<E<11.4 \) eV and \( 5.6<E<6.6 \) eV, of the width of \( 2h=1 \) eV.

FIG. 3. (Color online) Localization length for the human ADAM10 gene. The length of the exon (intron) is 1030 (31 752) bps. Inset shows the fine structure of one of the peaks.

FIG. 4. (Color online) Localization length vs energy for the human SNAP29 gene. The length of the exon (intron) is 2141 (21 701) bps.

FIG. 5. (Color online) Localization length vs energy for the human SUHW1 gene. The length of the exon (intron) is 1963 (4405) bps.
Lyapunov exponents do. It is clear that the presence or absence of the bands of the extended states is determined by subtle interference among the Fourier harmonics of the functions \varphi_j(\mu) given by Eq. (7).

The Lyapunov exponent (\alpha) depends on \epsilon_t, \epsilon_r, \epsilon_C, and \epsilon_i as well as on the hopping amplitudes \tau and \hbar. Since the values of \tau and \hbar are not well established experimentally, we repeated the calculations for different values of the hopping amplitudes, 0.1 < \hbar < 0.5 and 0.7 < \tau < 1. Since our analytical approach is valid only in the region where specific long-range correlations between the nucleotides and not by a particular choice of control parameters of the model.

The conducting properties of DNA have attracted much attention since DNA may be used in electronic devices. We hope that our approach and results can be very useful for further theoretical and experimental studies of the electrical and optical properties of DNA.

V. CONCLUSIONS

In our study of the double-stranded model of DNA we observed much longer localization length in exon than in intron regions for practically all the allowed energies and for all randomly selected DNA sequences. Through statistical correlations of the nucleotide sequence making up a DNA molecule, we relate this persistent difference to qualitatively different information stored by exons and introns.

For each DNA the pattern \Gamma(E) is unique fingerprint and can be used for identification of DNA's. All presented results confirm the suggestion that the localization length in DNA is determined by specific long-range correlations between the nucleotides and not by a particular choice of control parameters of the model.

ACKNOWLEDGMENT

This work is supported by the U.S. Department of Energy under Grant No. DE-FG02-06ER46312.