Intrinsic Noise from Neighboring Bases in the DNA Transverse Tunneling Current

Jose R. Alvarez,[1] Dmitry Skachkov,[1] Steven E. Massey,[2] Junqiang Lu,[3] Alan Kalitsov,[1,4,*] and Julian P. Velev[1,5,†]

[1]*Department of Physics, Institute for Functional Nanomaterials, University of Puerto Rico, San Juan, Puerto Rico 00931-3344*

[2]*Department of Biology, University of Puerto Rico, San Juan, Puerto Rico 00931*

[3]*Department of Physics, Institute for Functional Nanomaterials, University of Puerto Rico, Mayaguez, Puerto Rico 00981*

[4]*Materials for Information Technologies Center, University of Alabama, Tuscaloosa, Alabama 35487, USA*

[5]*Department of Physics, University of Nebraska, Lincoln, Nebraska 68588-0111, USA*

(Received 3 January 2014; revised manuscript received 26 February 2014; published 15 April 2014)

Nanopore DNA sequencing holds great promise for producing long read lengths from small amounts of starting material, however, high error rates are a problem. We perform nonequilibrium electron transport calculations within an effective tight-binding model of the DNA molecule to study the intrinsic structural noise in DNA sequencing via transverse current in nanopores. The structural noise arises from the effect of neighboring bases on the tunneling current. We find that it could be comparable to the environmental noise, which is caused by changes of the position of the molecule with respect to the electrodes in the nanopore. Moreover, while the environmental noise can be reduced by continuous measurement and by improving the measurement setup, the structural noise is intrinsic. With the help of our methodology we optimize the dependence of the structural noise on the measurement parameters, such as the type of the electrodes and the applied bias. We also propose a statistical technique, utilizing not only the currents through the nucleotides but also the correlations in the currents, to improve the fidelity of the sequencing.

DOI: 10.1103/PhysRevApplied.1.034001

I. INTRODUCTION

The deoxyribonucleic acid (DNA) molecule encodes, in a unique sequence of four bases, adenine (A), guanine (G), thymine (T), and cytosine (C), the genetic information concerning the structure and function of all living organisms. For that reason tremendous effort has been invested into developing methods to determine the order of the bases in a DNA molecule. The original DNA sequencing method, developed by Sanger *et al.* [1], is a complicated biochemical process of DNA fragmentation, amplification and chain termination, optical detection, and computer-based sequence determination. In recent years, a variety of next generation sequencing (NGS) techniques have been developed which have vastly improved the output and reduced the cost of obtaining genome sequences [2,3]. However, current technologies remain error prone and limited in read length [4].

A third generation of sequencing methods, exploiting physical methods for DNA detection, is under development, which methods have the potential of dramatically lowering the cost and time necessary for DNA sequencing. DNA detection on parallel arrays of field-effect biosensors is an example [5,6]. Nanopore sequencing is at the forefront of these methods [7–10]. Single-stranded DNA (ssDNA) suspended in an electrolyte solution can be driven through a nanoscale pore and the variations of the ionic current as the nucleotides block the pore channel can be correlated to the type of the base in the nucleotide [11]. The approach promises to significantly extend read length while reducing the amount of starting material. Moreover, base modifications could be detected using this strategy. The first experiments utilized protein (α-hemolysin) nanopores [11], however, later effort switched to solid-state nanopores [12] which offer better control of the size and shape of the nanopore. The ability to detect DNA bases by ionic current in nanopore is demonstrated in both protein [13] and solid-state nanopores [14,15]. The advantages of this method over biochemical methods are tremendous in terms of cost and speed, because the sample requires minimum preparation and long DNA strands can be sequenced. However, the accuracy is relatively poor because the pores are fairly large and several nucleotides can block the current. Also, the bases are fairly similar in size and shape which, combined with the environmental effects, can make their signature difficult to distinguish. Different methods have been proposed to deal with these issues such as base immobilization in the nanopore [16], fitting an adaptor in the pore [17,18], or tagging the bases [19], as well as statistical processing to improve the accuracy [20,21].

In a variation of the nanopore technique it is proposed that the pore is retrofitted with electrodes and the transverse

*Corresponding author.
kalitsov@yahoo.com
†Corresponding author.
jvelev@gmail.com

current through the nucleotides is measured as they move through it [22–26]. Identifying individual nucleotides by tunneling current is demonstrated in principle by contacting a single nucleotide between an electrode and an STM tip [27–29]. Moreover, it is shown that electrodes can be manufactured on solid-state nanopores [30,31] and eventually base identification by tunneling current has been demonstrated [23,32–34]. Recently, graphene nanopores have been proposed, since graphene is an atomic layer thick single-base resolution can be achieved [35–37]. Noise, nevertheless, remains an issue. One problem is that the bases are not chemically attached to the electrodes and their temperature- and environment-driven transpositions strongly modify the tunneling current [24,25,32,33]. It is shown, by studying the translocation of homopolymers, that even though a single measurement of the current is not enough to distinguish the different nucleotides, the distribution of the values of electron current for each base is different [23,24,34]. As a result, the bases can be identified by the mean of the current distributions. Thus, statistically, the error due to environmental noise can be brought under any specified value by performing a larger number of readings of the same base [23,24]. Furthermore, physically the environmental noise can be minimized by reducing the degrees of freedom of the base inside the nanopore by either chemical modification of the electrodes and/or nanofabrication to restrict the conformations of the molecule in the nanopore [17,29,33,34].

Another, and so far overlooked, source of noise is the structural noise resulting from the random neighbors around each base. The current through identical bases will be different depending on the distribution of the neighboring bases in the sequence because of the electron dispersion longitudinally along the DNA chain. Earlier theoretical calculations suggested that this noise should be rather small, however, the study involved only very short sequences (triples) for small voltages and the data actually showed very significant changes of the current [22]. This noise is intrinsic and cannot be controlled by improved control of the environment. Moreover, multiple measurements of the same base will do nothing to alleviate the problem. In this work, we concentrate on the study of the intrinsic structural noise resulting from the influence of neighboring bases. We demonstrate that this noise is important and should be taken into account in order to reduce error rates. We show how the noise can be reduced by an appropriate choice of the electrodes and the applied bias. We also introduce a statistical procedure based on Bayesian inference to improve the fidelity of the readout.

II. METHODOLOGY

A. Tight-binding model

The DNA nanopore sequencing geometry is schematically illustrated in Fig. 1. A ssDNA translocates between two tapered metal electrodes which make contact with one base at a time. To represent ssDNA we adopt the ladder model for longitudinal DNA transport [38–43]. This is an effective tight-binding model which represents the DNA molecule with two sites: P for the phosphate backbone and X for the base ($X = A, G, T, C$). On each site there is a single energy level corresponding to the molecular level closest to the Fermi energy. Since the energy difference between the molecular orbitals is of the order of tenths of eV this model is a good approximation for small bias. The tight-binding parameters are fitted from first-principles calculations to correctly describe the π-π overlap between the bases [43]. This model is appropriate, because in order to account for the influence of the neighboring bases, it is important to describe the longitudinal transport correctly. Moreover, such a simplified model enables us to handle long DNA strands and to accumulate large statistics, which would be impossible from first-principles calculations.

The Hamiltonian of the system has the form $H = H_{\mathrm{DNA}} + H_L + H_R + H_{\mathrm{cpl}}$ where $H_{L/R}$ is the Hamiltonian of the left and right electrode and H_{cpl} is the coupling between the molecule and the electrodes. The Hamiltonian of the uncoupled DNA is given by

$$H_{\mathrm{DNA}} = \sum_i \varepsilon_{X_i} c_i^\dagger c_i + \sum_{i,j} t_{X_i X_j} c_i^\dagger c_j, \tag{1}$$

where ε_{X_i} is the on-site energy of base i, $t_{X_i X_j}$ is the hopping integral between bases i, and j, $c_i^\dagger(c_i)$ is the electron creation (annihilation) operator on base i. Photoemission data [44,45] and density functional theory (DFT) calculations [43,46] indicate that DNA is a large band-gap

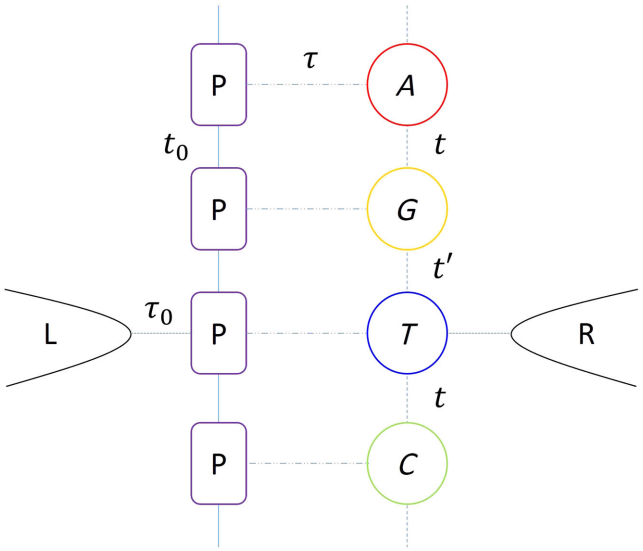

FIG. 1. Ladder model of a ssDNA molecule translocating in a nanopore between left (L) and right (R) electrodes. P indicates the phosphate backbone; A, G the purine; and T, C the pyrimidine bases. The different hopping matrix elements (t, τ) are also indicated.

semiconductor (with a π-π gap ~ 4 eV). In the effective model the on-site energies of the bases and the backbone are taken to be the ionization energies. A similar Hamiltonian can be written for each of the electrodes.

$H_{L(R)} = \sum_\mu \varepsilon_\mu c_\mu^\dagger c_\mu + \sum_{\mu,\nu} t_{\mu\nu} c_\mu^\dagger c_\nu$, where the on-site energy ε_μ of the metallic electrode is taken to be the metal work function [47]. Finally, the coupling between the electrodes and a particular nucleotide k of the DNA molecule is $H_{\rm cpl} = \sum_{\alpha=\{L,R\}} t_{X_k\alpha} c_k^\dagger c_\alpha + \text{H.c.}$.

B. Model parametrization

The most important parameters in the model are the on-site energies. There is quite a bit of discrepancy in the literature between the energy values calculated by different first-principles methods [43,46–50]. Since charge transport depends on the relative positions of the highest occupied molecular orbital (HOMO) of the DNA and the metal work functions of the electrodes, mixing values obtained using different methods or codes can lead to errors. To avoid this problem we perform first-principles calculation to obtain all on-site energies on the same level of theory. We use the Vienna *ab initio* simulation package (VASP) [51] with the Perdew-Burke-Ernzerhof exchange and correlation functional. The obtained values are listed in Table I [52]. Overall, the calculated DNA HOMO levels and metal work functions are consistent with previous DFT calculations [46,47]. Also, the alignment of the DNA levels with the metal work functions agrees well with photoemission data [45].

The overlap integrals do not suffer from the same issue as the on-site energies. Moreover, they are less crucial because, although they affect the energy level separation, their contribution is small compared to the on-site energy differences [52]. For these reasons we adopt the well-established and widely used parametrization available in the literature (Table I) [42–43]. Between base pairs which are structurally similar we use the same value $t = t_{XX} = t_{YY}$ and another value $t' = t_{XY}$ between dissimilar pairs (here $X = A$, G labels the purine and $Y = T$, C the pyrimidine bases). The hopping between the backbone and all the bases $\tau = t_{PX} = t_{PY}$ is taken to be the same because it is effectuated through the same C-N bond. Finally, $t_0 = t_{PP}$ is the hopping along the backbone.

TABLE I. Parametrization of the DNA ladder model: DNA on-site energies, hopping integrals, and metal work functions (in eV).

$\varepsilon_{\rm DNA}$					ε_M		
A	G	T	C	P	Al	Au	Pt
−5.54	−5.15	−5.99	−5.82	−5.92	−4.25	−5.22	−5.70

t_{XY}			
	A, G	T, C	P
A, G	0.35	0.17	0.70
T, C		0.35	0.70
P			0.15

C. Environmental noise model

Another advantage of the effective model is that it allows for the simple inclusion of environmental noise [39,53]. We allow for the base-backbone pair to vibrate and rotate around some equilibrium position in the pore. We assume that at temperature T the rotation angle follows a normal distribution with mean $\langle \Delta\theta_k \rangle = \theta_0$ and a temperature-dependent standard deviation $\sqrt{\langle \Delta\theta^2 \rangle} = \alpha T$, where $\alpha = \frac{1}{300}$ K^{-1} [39,54]. Similarly, the position of the pair center follows a normal distribution with mean $\langle \Delta d_k \rangle = d_0$ and a standard deviation $\sqrt{\langle \Delta d^2 \rangle} = \beta T$, where $\beta = \frac{1}{100}$ Å K^{-1} [22,55]. The coupling between the DNA and the electrodes $t_{X_k\alpha}$ depends on the wave function overlap and thus on the distance between the base or backbone and the electrode [56]. In our model the hopping between the electrode and the backbone or base is chosen to be the same $\tau_0 = t_{X\alpha} = t_{P\alpha}$ because the DNA molecule is only weakly bonded to the electrode. For the same reason the distance between the nucleotide and the electrode will vary substantially and the change of τ_0 is the leading source of the variations of the current [52].

D. Transverse charge current

Finally, we calculate the transverse current through base k using the Green's function (GF) method [57]

$$I^k = \frac{2e}{h} \int dE [f_L(E) - f_R(E)] \mathrm{Tr}[\Gamma_L^k G \Gamma_R^k G^\dagger], \quad (2)$$

where semi-infinite left and right (L/R) electrodes are in equilibrium with chemical potentials μ_L and μ_R and $f_{L/R}$ are the Fermi-Dirac distribution functions, $\Gamma_{L/R}^k$ are the escape rates to the electrodes when connected to base k, and G is the retarded GF of the DNA molecule connected to the electrodes. The calculation is performed in real space.

To calculate the current we diagonalize $H_{\rm DNA}$, Eq. (1), to obtain the GF of the uncoupled DNA molecule g. Next we find the GF of the DNA coupled to the electrodes by solving the Dyson equation $G = g + g\Sigma G$, where $\Sigma = \Sigma_L + \Sigma_R$ is the self-energy due to the connection to the electrodes. Since the molecular levels are discrete, while the typical metal band is several eVs wide, we can treat the electrodes on the level of the wide-band approximation. This implies that in the vicinity of the molecular level the density of states (DOS) of the electrode is essentially constant. Thus, within the wide-band approximation $\Sigma_{L/R}$ and $\Gamma_{L/R} = -2\,\mathrm{Im}[\Sigma_{L/R}]$ are independent of the energy. Since the DNA molecule is not chemically bonded to the electrodes, the contact is weak and we use $\Gamma_{L/R} = 10^{-3}$ eV in the calculations. The exact value of $\Gamma_{L/R}$ is not very

important because its effect is to scale the current, but it does not change the current distribution.

III. RESULTS AND DISCUSSION

A. Environmental noise (homopolymers)

First, we use the developed formalism to investigate the effect of the environmental noise on the current distribution. In order to isolate the environmental from the structural noise we consider ssDNA homopolymers, poly(X) where $X = A$, G, T, or C. In order to collect sufficient statistics we take molecules consisting of 1000 nucleotides and impose on them the displacements resulting from the environmental factors. Then we calculate the current through each base. The process is repeated 20 times. From the resulting 20 000 observation a nonparametric probability distribution function (PDF) is constructed by making a fine histogram and interpolating it with a smooth function. The PDF $P_X(I)$ has the meaning of the probability of measuring current I through base X. There are four PDFs for the current distribution through each nucleotide.

The resulting PDFs for 0.1 V applied bias are shown in Fig. 2 for two different temperatures and electrodes. For the electrodes we consider two typical cases: Al the chemical potential of which is well in the gap of the DNA and Au the chemical potential of which is aligned with the DNA HOMO levels (Table I). The current in the two cases is different by an order of magnitude and they clearly represent distinct transport regimes. For small bias in the Al case the current is a pure tunneling current, while in the Au case molecular orbitals fall in the bias window and the current has a resonant character.

At very low temperatures (~0 K) the nucleotides are at a particular fixed position with respect to the electrodes and the current through all the nucleotides is the same (indicated by vertical lines in Fig. 2). For Al the currents through the different bases are very distinct as the tunneling current depends exponentially on the barrier height. In the case of Au the C and T currents are very close because their on-site energies are close and they both produce resonant levels in the bias window. At room temperature (300 K) the DNA molecule starts vibrating and rotating within the pore. This produces displacements in the order of Å which strongly affect the contact with the electrodes. Thus, although the PDFs are centered close to the 0 K current value, for a particular nucleotide the current varies by orders of magnitude. Nevertheless, the PDFs are still statistically distinct. Despite its simplicity the model gives qualitatively very similar results to first-principles calculations [23–25].

B. Structural noise (random sequences)

In order to study the structural noise we generate twenty 1000-nucleotide long DNA sequences in which each of the bases appears randomly with the same probability, which is a good representation of natural DNA. Here the fact that we consider very long molecules is crucial because it allows each nucleotide to have essentially an unlimited number of neighbors. Similarly, at finite temperature we impose the nucleotide displacements and rotations resulting from environmental factors. Then we calculate the current through each base and use the obtained values to construct the current PDFs. The PDFs in the case of Al and Au electrodes at 0 and 300 K are plotted in Fig. 3. The first striking observation is that at 0 K, without any environmental noise, we still have a very wide current distribution.

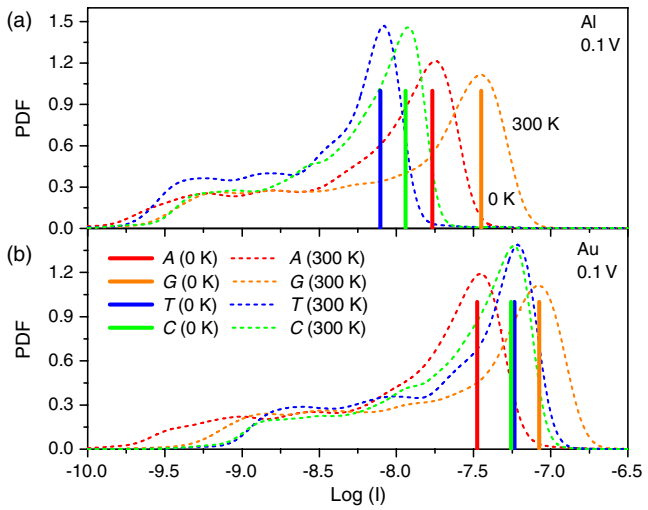

FIG. 2. Tunneling current probability distribution function (not normalized) for poly(X) DNA chains ($X = A$, G, T, C) for low (0 K) and high temperatures (300 K) at 0.1 V applied bias for (a) Al and (b) Au electrodes.

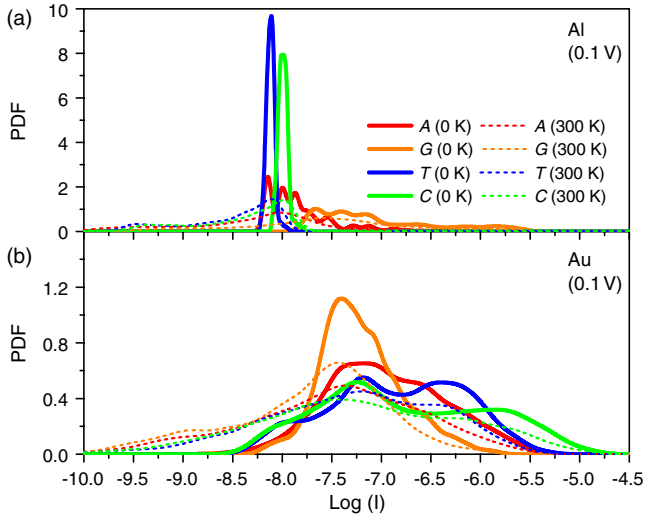

FIG. 3. Tunneling current probability distribution function (not normalized) for a random DNA chain for low (0 K) and high temperatures (300 K) at 0.1 V applied bias for (a) Al and (b) Au electrodes.

In fact, in the case of Au electrodes the PDFs almost completely overlap even at 0 K. This noise comes entirely from the lateral dispersion of the electron along the molecule.

Comparing the Al and Au electrodes we find that the 0 K noise is much smaller in the tunneling current. This fact can be explained as follows: a ssDNA molecule with n bases will have $2n$ energy levels many of which will be very close to the energy level of the contacted nucleotide [52]. If the level is in the proximity of the bias window the neighboring levels will also give resonant tunneling contributions to the current on the same footing, which leads to the almost complete smearing out of the current. In the tunneling regime the energy difference between the nucleotide and the satellite levels will translate into smaller satellite contributions to the current. The standard deviation of the structural noise PDFs is comparable to that of the environmental noise. Therefore, even after adding the environmental noise at 300 K the overall noise is still substantially influenced by the structural component. In particular in the case of Au electrode the thermal noise practically leaves the PDFs unchanged.

Recently, graphene nanopores have been proposed for DNA sequencing. The work function of graphene is ~4.66 eV and the work functions of graphene nanoribbons have been reported to be slightly lower ~4.58 eV [58]. Thus, the graphene Fermi level is halfway between Al and Au, just above the nucleotide resonant levels. Thus, we expect that as small bias the current will be in the tunneling regime. However, the work function of graphene can be substantially shifted by the contact with metallic electrodes, which could easily shift the transverse current in the resonant regime [59].

The effect of bias is to increase the structural noise while leaving the thermal noise unchanged. In the case of resonant tunneling (Au), as the bias window increases, more satellite levels fall in the window smearing the current even more. In the case of pure tunneling (Al) the size of the barrier decreases and the satellite contributions are felt more. At the same time the environmental noise is governed by the temperature-driven variations of the coupling with the electrodes which does not depend on the bias.

The above observations show that, while the thermal noise can be controlled by improving the setup to reduce the possible conformations of the molecule in the pore or by continuous measurement and subsequent averaging, the structural noise is intrinsic. It can be limited by keeping the transport in the tunneling regime by an appropriate choice of the electrodes and the bias (i.e., electrode Fermi energy close to the midgap of DNA and small bias). Despite that, unlike the environmental noise, it cannot be completely removed even in theory. Therefore, statistical techniques will be essential to reducing the error rates.

C. Fidelity improvement (Bayesian inference)

Statistical procedures, especially Bayesian hidden Markov models [60], have played an important role in genomics both in obtaining [21,61,62] and analyzing DNA sequences [63,64]. Here we develop a different, Bayesian-inspired statistical procedure to deal with the structural noise in DNA sequencing. In the first step, we calibrate the method by collecting a large number of current readings (20 000) for known DNA sequences (randomly generated). From these data we construct the joint distribution PDFs $P_X, P_{XY}, P_{XYZ}, \ldots$ with $X, Y, Z \in (A, G, T, C)$. The single PDF $P_X(I_X)$ has the meaning of the probability of measuring a current I_X through a base X. Joint PDFs, such as $P_{XYZ}(I_X, I_Y, I_Z)$, give the joint probability of measuring the currents I_X, I_Y, I_Z through the three neighboring bases X, Y, Z, respectively. In principle, we can compute joint probability functions to any order with large enough statistics.

The simplest measurement procedure uses only P_X ignoring correlations between the currents on neighboring bases. In this case a current I_k is read through base k of an unknown DNA sequence and a base is assigned based on the maximum probability $\tilde{X}_k = \max_X P_X(I_k)$ of measuring this current through any of the bases. As a metric of the success rate of such a procedure we define the fidelity

$$f = \frac{1}{N}\sum_{k=1}^{N}(\tilde{X}_k == X_k), \tag{3}$$

where $\tilde{X}_k$ is the guess, X_k is the actual base, and the sum runs over all the N bases in the DNA sequence. Each correct identification $\tilde{X}_k == X_k$ adds 1 to the sum. The normalization limits the fidelity in the interval from 0 to 1. This measure is complementary to the error rate. The calculated fidelity for the simple procedure is given in Table II for a 1000 base long DNA molecule for several electrodes and biases. In the tunneling regime (Al) the fidelity is less that 80% (i.e., error rates of more than 20%). At the same time in the resonant regime (Au) the fidelity is the dismal 37% (63% error rate). This is very low given that a completely random choice will produce a fidelity of 25%. In general, in the tunneling regime the fidelity decreases with the bias. In the resonant regime the bias trend is more difficult to identify because it depends on the exact way resonant levels enter and leave the bias window.

TABLE II. Fidelity at low temperature for a 1000 base random DNA sequence for several electrodes and values of the applied bias using the single and triple PDFs.

Bias (V)	Al		Au		Pt	
	P_X	P_{XYZ}	P_X	P_{XYZ}	P_X	P_{XYZ}
0.1	0.794	0.995	0.364	0.510	0.901	0.991
0.5	0.795	0.985	0.569	0.668	0.757	0.978

TABLE III. Partial fidelity for homotriplets in a 1000 base random DNA sequence for several electrodes and 0.1 V applied bias using the single and triple PDFs.

Triplet	Al		Au		Pt	
	P_X	P_{XYZ}	P_X	P_{XYZ}	P_X	P_{XYZ}
AAA	0.208	0.995	0.125	1.000	1.000	1.000
GGG	1.000	1.000	0.333	0.500	1.000	1.000
TTT	0.974	1.000	0.307	0.641	1.000	1.000
CCC	1.000	1.000	0.200	0.333	0.200	0.733

Of course, the currents through neighboring bases are not independent. To account for the correlations we use the joint PDFs in the following Bayesian-inspired procedure. We use the DNA sequence inferred from P_X as a starting estimate and consequently use the joint distributions P_{XY}, P_{XYZ} to improve it [52]. For example, to include second order correlations, we start with the estimated sequence $\tilde{X}_k^{(0)} = \max_X P_X(I_k)$. Then we check it for consistency with the joint PDFs. Given $\tilde{X}_{k-1}^{(0)}$ and $\tilde{X}_{k+1}^{(0)}$ we calculate $\tilde{X}_k^{(1)} = \max_X\{[P_{X\tilde{X}_{k+1}^{(0)}}(I_k, I_{k+1}) + P_{\tilde{X}_{k-1}^{(0)}X}(I_{k-1}, I_k)]/2\}$. If $\tilde{X}_k^{(0)} = \tilde{X}_k^{(1)}$ we assume it is certain. Then we update our probability distribution functions accordingly for this base: $P_{X\neq\tilde{X}_k}(I_k) = 0$ and $P_{X_{k-1}\neq\tilde{X}_{k-1}^{(0)},X}(I_{k-1}, I_k) = P_{X,X_{k+1}\neq\tilde{X}_{k+1}^{(0)}}(I_k, I_{k+1}) = 0$. Then with the modified PDFs we recalculate the sequence, which thus becomes second-order consistent. Similarly, we can add third-order correlations by checking the consistency with P_{XYZ}, i.e., given $\tilde{X}_{k-1}^{(1)}$ and $\tilde{X}_{k+1}^{(1)}$ we calculate $\tilde{X}_k^{(2)} = \max_X[P_{\tilde{X}_{k-1}^{(1)}X\tilde{X}_{k+1}^{(1)}}(I_{k-1}, I_k, I_{k+1})]$, determine the certain bases, and recalculate the sequence. Since the calibration is done once and the influence of the further neighbors is bound to be smaller, it is feasible to construct enough higher order PDFs to reduce the error rates below a desired threshold.

The fidelity for the random DNA readout accounting for up to third-order correlations is listed in Table II. We notice that in all cases the fidelity dramatically increases. In the tunneling case (Al) the fidelity reaches close to 99%. In the resonant case (Au) the fidelity also increases, however, the error rate is still unacceptably large indicating that Au is a poor choice for electrodes. For Al and Pt the fidelities larger than 99% are comparable to that of the Roche 454 pyrosequencer [65].

A number of NGS technologies are prone to errors in differentiating repeating, $XX\ldots X$, sequences embedded in the DNA chain. In fact, transverse current sequencing is also prone to these errors, as it can be seen from Table III for the single PDF. Including correlations essentially resolves this issue. We also observe that the electrode level positioning with respect to the base level has a dramatic influence on the occurrence of this type of error. A larger bias window generally alleviates the problem.

IV. CONCLUSIONS

We performed an extensive statistical investigation of the transverse current through individual nucleotides in long ssDNA molecules. We demonstrated that in addition to the environmental noise, arising from random displacements of the nucleotide inside the nanopore, an intrinsic structural noise is present resulting from the influence of the random neighboring nucleotides on the current as the electron disperses laterally in the molecule. The standard deviation of the current resulting from the structural noise could be comparable to that of the environmental noise and could lead to large error rates. Moreover, unlike the environmental noise it cannot be reduced by continuous measurement. We showed that the structural noise can be overcome by an iterative improvement statistical procedure using higher order correlations between the currents through neighboring bases. Taking into account correlations between the currents also solves the persistent problem of base misidentification due to sequence repeats.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation (Grants No. EPS-1002410, No. EPS-1010094, and No. DMR-1105474) and the U.S. Department of Energy (Grant No. DE-FG02-08ER46526).

[1] F. Sanger, S. Nicklen, and A. R. Coulson, DNA sequencing with chain-terminating inhibitors, Proc. Natl. Acad. Sci. U.S.A. **74**, 5463 (1977).

[2] J. Shendure and H. Ji, Next-generation DNA sequencing, Nat. Biotechnol. **26**, 1135 (2008).

[3] M. L. Metzker, Sequencing technologies—the next generation, Nat. Rev. Genet. **11**, 31 (2009).

[4] C. W. Fuller, L. R. Middendorf, S. A. Benner, G. M. Church, T. Harris, X. Huang, S. B. Jovanovich, J. R. Nelson, J. A. Schloss, D. C. Schwartz, D. V. Vezenov, The challenges of sequencing by synthesis, Nat. Biotechnol. **27**, 1013 (2009).

[5] M. H. Abouzar, A. Poghossian, A. G. Cherstvy, A. M. Pedraza, S. Ingebrandt, and M. J. Schöning, Label-free electrical detection of DNA by means of field-effect nanoplate capacitors: Experiments and modeling, Phys. Status Solidi A **209**, 925 (2012).

[6] A. G. Cherstvy, Detection of DNA hybridization by field-effect DNA-based biosensors: mechanisms of signal generation and open questions, Biosens. Bioelectron. **46**, 162 (2013).

[7] D. Branton, D. W. Deamer, A. Marziali, H. Bayley, S. A. Benner, T. Butler, M. Di Ventra, S. Garaj, A. Hibbs, X. Huang, S. B. Jovanovich, P. S. Krstic, S. Lindsay, X. S. Ling, C. H. Mastrangelo, A. Meller, J. S. Oliver, Y. V. Pershin, J. M. Ramsey, R. Riehn, G. V. Soni, V. Tabard-Cossa, M. Wanunu, M. Wiggin, and J. A. Schloss, The potential and challenges of nanopore sequencing, Nat. Biotechnol. **26**, 1146 (2008).

[8] B. M. Venkatesan and R. Bashir, Nanopore sensors for nucleic acid analysis, Nat. Nanotechnol. **6**, 615 (2011).
[9] R. H. Scheicher, A. Grigoriev, and R. Ahuja, DNA sequencing with nanopores from an ab initio perspective, J. Mater. Sci. **47**, 7439 (2012).
[10] M. Di Ventra, Fast DNA sequencing by electrical means inches closer, Nanotechnology **24**, 342501 (2013).
[11] J. J. Kasianowicz, E. Brandin, D. Branton, and D. W. Deamer, Characterization of individual polynucleotide molecules using a membrane channel, Proc. Natl. Acad. Sci. U.S.A. **93**, 13770 (1996).
[12] J. Li, D. Stein, C. McMullan, D. Branton, M. J. Aziz, and J. A. Golovchenko, Ion-beam sculpting at nanometre length scales, Nature (London) **412**, 166 (2001).
[13] M. Akeson, D. Branton, J. J. Kasianowicz, E. Brandin, and D. W. Deamer, Ion-beam sculpting at nanometre length scales, Biophys. J. **77**, 3227 (1999).
[14] M. Wanunu, J. Sutin, and A. Meller, DNA profiling using solid-state nanopores: Detection of DNA-binding molecules, Nano Lett. **9**, 3498 (2009).
[15] G. M. Skinner, M. van den Hout, O. Broekmans, C. Dekker, and N. H. Dekker, Distinguishing single and double-stranded nucleic acid molecules using solid-state nanopores, Nano Lett. **9**, 2953 (2009).
[16] R. F. Purnell, K. K. Mehta, and J. J. Schmidt, Nucleotide identification and orientation discrimination of DNA homopolymers immobilized in a protein nanopores, Nano Lett.. **8**, 3029 (2008).
[17] H. He, R. H. Scheicher, R. Pandey, A. R. Rocha, S. Sanvito, A. Grigoriev, R. Ahuja, S. P. Karna, Functionalized nanopore-embedded electrodes for rapid DNA sequencing, J. Phys. Chem. C **112**, 3456 (2008).
[18] J. Clarke, H.-C. Wu, L. Jayasinghe, A. Patel, S. Reid, and H. Bayley, Continuous base identification for single-molecule nanopore DNA sequencing, Nat. Nanotechnol. **4**, 265 (2009).
[19] A. Singer, M. Wanunu, W. Morrison, H. Kuhn, M. Frank-Kamenetskii, and A. Meller, Nanopore-based sequence-specific detection of duplex DNA for genomic profiling, Nano Lett. **10**, 738 (2010).
[20] D. Stoddart, G. Maglia, E. Mikhailova, A. J. Heron, and H. Bayley, Multiple base-recognition sites in a biological nanopore: Two heads are better than one, Angew. Chem. **122**, 566 (2010).
[21] W. Timp, J. Comer, and A. Aksimentiev, DNA base-calling from a nanopore using a Viterbi algorithm, Biophys. J. **102**, L37 (2012).
[22] M. Zwolak and M. Di Ventra, Electronic signature of DNA nucleotides via transverse transport, Nano Lett. **5**, 421 (2005).
[23] J. Lagerqvist, M. Zwolak, and M. Di Ventra, Fast DNA sequencing via transverse electronic transport, Nano Lett. **6**, 779 (2006).
[24] J. Lagerqvist, M. Zwolak, and M. Di Ventra, Influence of the environment and probes on rapid DNA sequencing via transverse electronic transport, Biophys. J. **93**, 2384 (2007).
[25] M. Krems, M. Zwolak, Y. V. Pershin, and M. Di Ventra, Effect of noise on DNA sequencing via transverse electronic transport, Biophys. J. **97**, 1990 (2009).
[26] M. Zwolak and M. Di Ventra, Colloquium: Physical approaches to DNA sequencing and detection, Rev. Mod. Phys. **80**, 141 (2008).
[27] E. Shapir, H. Cohen, A. Calzolari, C. Cavazzoni, D. A. Ryndyk, G. Cuniberti, A. Kotlyar, R. Di Felice, and D. Porath, Electronic structure of single DNA molecules resolved by transverse scanning tunnelling spectroscopy, Nat. Mater. **7**, 68 (2007).
[28] J. He, L. Lin, P. Zhang, Q. Spadola, Z. Xi, Q. Fu, and S. Lindsay, Transverse tunneling through DNA hydrogen bonded to an electrode, Nano Lett. **8**, 2530 (2008).
[29] S. Chang, S. Huang, J. He, F. Liang, P. Zhang, S. Li, X. Chen, O. Sankey, S. Lindsay, Electronic Signatures of all four DNA nucleosides in a tunneling gap, Nano Lett. **10**, 1070 (2010).
[30] T. S. Maleki, S. Mohammadi, and B. Ziaie, A nanofluidic channel with embedded transverse nanoelectrodes, Nanotechnology **20**, 105302 (2009).
[31] M. Tsutsui, S. Rahong, Y. Iizumi, T. Okazaki, M. Taniguchi, and T. Kawai, Single-molecule sensing electrode embedded in-plane nanopore, Sci. Rep. **1**, 46 (2011).
[32] M. Tsutsui, M. Taniguchi, K. Yokota, and T. Kawai, Identifying single nucleotides by tunnelling current, Nat. Nanotechnol. **5**, 286 (2010).
[33] A. P. Ivanov, E. Instuli, C. M. McGilvery, G. Baldwin, D. W. McComb, T. Albrecht, and J. B. Edel, DNA tunneling detector embedded in a nanopore, Nano Lett. **11**, 279 (2011).
[34] T. Ohshiro, K. Matsubara, M. Tsutsui, M. Furuhashi, M. Taniguchi, and T. Kawai, Single-molecule electrical random resequencing of DNA and RNA, Sci. Rep. **2**, 501 (2012).
[35] H. W. Ch. Postma, Rapid sequencing of individual DNA molecules in graphene nanogaps, Nano Lett. **10**, 420 (2010).
[36] T. Nelson, B. Zhang, and O. V. Prezhdo, Detection of nucleic acids with graphene nanopores: Ab initio characterization of a novel sequencing device, Nano Lett. **10**, 3237 (2010).
[37] K. K. Saha, M. Drndić, and B. K. Nikolić, DNA base-specific modulation of microampere transverse edge currents through a metallic graphene nanoribbon with a nanopore, Nano Lett. **12**, 50 (2012).
[38] G. Cuniberti, L. Craco, D. Porath, and C. Dekker, Backbone-induced semiconducting behavior in short DNA wires, Phys. Rev. B **65**, 241314 (2002).
[39] S. Roche, Sequence dependent DNA-mediated conduction, Phys. Rev. Lett. **91**, 108101 (2003).
[40] S. Roche, D. Bicout, E. Maciá, and E. Kats, Long-range correlation in DNA: Scaling properties and charge transfer efficiency, Phys. Rev. Lett. **91**, 228101 (2003).
[41] G. Cuniberti, E. Maciá, A. Rodríguez, and R. A. Römer, in *Charge Migration in DNA* (Springer, Berlin, Heidelberg, New York, 2007), pp. 1–20.
[42] C.-T. Shih, S. Roche, and R. A. Römer, Point mutations effects on charge transport properties of the tumor-suppressor gene p53, Phys. Rev. Lett. **100**, 018105 (2008).
[43] L. G. D. Hawke, G. Kalosakas, and C. Simserides, Tight-binding parameters for charge transfer along DNA, Eur. Phys. J. E **32**, 291 (2010).
[44] A. B. Trofimov, J. Schirmer, V. B. Kobychev, A. W. Potts, D. M. P. Holland, and L. Karlsson, Photoelectron spectra of

the nucleobases cytosine, thymine and adenine, J. Phys. B **39**, 305 (2006).

[45] Y. Lee, H. Lee, S. Park, and Y. Yi, Energy level alignment at the interfaces between typical electrodes and nucleobases: Al/adenine/indium-tin-oxide and Al/thymine/indium-tin-oxide, Appl. Phys. Lett. **101**, 233305 (2012).

[46] Y. Maeda, A. Okamoto, Y. Hoshiba, T. Tsukamoto, Y. Ishikawa, and N. Kurita, Effect of hydration on electrical conductivity of DNA duplex: Green's function study combined with DFT, Comput. Mater. Sci. **53**, 314 (2012).

[47] N. E. Singh-Miller, and N. Marzari, Surface energies, work functions, and surface relaxations of low-index metallic surfaces from first principles, Phys. Rev. B **80**, 235407 (2009).

[48] D. Roca-Sanjuán, M. Rubio, M. Merchán, and L. Serrano-Andrés, Ab initio determination of the ionization potentials of DNA and RNA nucleobases, J. Chem. Phys. **125**, 084302 (2006).

[49] D. Roca-Sanjuán, M. Merchán, L. Serrano-Andrés, and M. Rubio, Ab initio determination of the electron affinities of DNA and RNA nucleobases, J. Chem. Phys. **129**, 095104 (2008).

[50] C. Faber, C. Attaccalite, V. Olevano, E. Runge, and X. Blasé, First-principles GW calculations for DNA and RNA nucleobases, Phys. Rev. B **83**, 115123 (2011).

[51] G. Kresse and J. Furthmuller, Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set, Phys. Rev. B **54**, 11169 (1996).

[52] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevApplied.1.034001 for details on the first-principles calculations, the environmental noise modeling, and the statistical procedure. Additional results for the all electrodes are included. A discussion of the transmission through a single nucleotide and short nucleotide chains is included to elucidate the origin of the structural noise.

[53] E. Maciá, Electrical conductance in duplex DNA: Helical effects and low-frequency vibrational coupling, Phys. Rev. B **76**, 245123 (2007).

[54] Z. G. Yu and X. Song, Variable range hopping and electrical conductivity along DNA double helix, Phys. Rev. Lett. **86**, 6018 (2001).

[55] M. A. Young, G. Ravishanker, and D. L. Beveridge, A 5-nanosecond molecular dynamics trajectory for B-DNA: analysis of structure, motions, and solvation, Biophys. J. **73**, 2313 (1997).

[56] S. Obara and A. Saika, Efficient recursive computation of molecular integrals over Cartesian Gaussian functions, J. Chem. Phys. **84**, 3963 (1986).

[57] Y. Meir and N. Wingreen, Landauer formula for the current through an interacting electron region, Phys. Rev. Lett. **68**, 2512 (1992).

[58] Q. Yan, B. Huang, J. Yu, F. Zheng, J. Zang, J. Wu, B.-L. Gu, F. Liu, and W. Duan, Intrinsic current-voltage characteristics of graphene nanoribbon transistors and effect of edge doping, Nano Lett.. **7**, 1469 (2007).

[59] G. Giovannetti, P. A. Khomyakov, G. Brocks, V. M. Karpan, J. van den Brink, and P. J. Kelly, Doping graphene with metal contacts, Phys. Rev. Lett. **101**, 026803 (2008).

[60] B. J. Yoon, Hidden Markov models and their applications in biological sequence analysis, Curr. Genomics **10**, 402 (2009).

[61] G. A. Churchill and B. Lazareva, Bayesian restoration of a hidden Markov chain with applications to DNA sequencing, J. Comput. Biol. **6**, 261 (1999).

[62] K. C. Liang, X. Wang, and D. Anastassiou, Bayesian basecalling for DNA sequence analysis using hidden Markov models, IEEE ACM Trans. Comput. Biol. Bioinf. **4**, 430 (2007).

[63] J. V. Braun and H.-G. Muller, Statistical methods for DNA sequence segmentation, Stat. Sci. **13**, 142 (1998).

[64] R. Boys and D. Henderson, A Bayesian approach to DNA sequence segmentation, Biometrics **60**, 573 (2004).

[65] S. Balzer, K. Malde, A. Lanzen, A. Sharma, and I. Jonassen, Characteristics of 454 pyrosequencing data-enabling realistic simulation with flowsim, Bioinformatics **26**, i420 (2010).