Detection of base-pair mismatches in DNA using graphene-based nanopore device

Sourav Kundu and S N Karmakar

Condensed Matter Physics Division, Saha Institute of Nuclear Physics, 1/AF, Bidhannagar, Kolkata 700 064, India

E-mail: sourav.kundu@saha.ac.in and sachindranath.karmakar@saha.ac.in

Received 12 August 2015, revised 31 January 2016
Accepted for publication 1 February 2016
Published 22 February 2016

Abstract
We present a unique way to detect base-pair mismatches in DNA, leading to a different epigenetic disorder by the method of nanopore sequencing. Based on a tight-binding formulation of a graphene-based nanopore device, using the Green’s function approach we study the changes in the electronic transport properties of the device as we translocate a double-stranded DNA through the nanopore embedded in a zigzag graphene nanoribbon. In the present work we are not only successful in detecting the usual AT and GC pairs but also a set of possible mismatches in the complementary base pairing.

Keywords: DNA sequencing, graphene nanopore, base-pair mismatch

(Some figures may appear in colour only in the online journal)

1. Introduction

Base-pair mismatches in DNA are one of the major reasons behind several mutagenic disorders that may lead to different genomic instabilities, development of cancer [1], and other degenerative diseases. A mismatch in DNA base pairs occurs mainly due to misincorporation of nitrogen bases during DNA replication, oxidative or chemical damages, and ionizing radiations. In spite of dramatic advancements in medical science, many crucial issues, such as how DNA detects and repairs damages, individual mismatches, or what is the most accurate observable physical parameter to detect a base-pair mismatch, still remain clouded. Apart from the traditional fluorescence-based sequencing technique [2], several other methods are also applied to detect mismatches. Some examples are magnetic signatures [3], longitudinal electronic transport [4, 5], thermodynamic properties of base-pair mismatches [6], and study of stretched DNA using AFM [7], but no conclusive results appear. However, with the advent of nanopore-based sequencing [8–13] a new pathway is opened for marker-free gene testing. In the early days of nanopore sequencing people mostly used biological nanopores (α-Haemolysin), detecting the changes in ionic current as a single-stranded DNA (ss-DNA) passes through the pore [8–11]. With time, usage of nanopore materials also evolved from biological to solid-state nanopores. The latter type overcomes many drawbacks of biological nanopores, e.g., poor mechanical strength [14] and problems of integration with on-chip electronics [15]. Solid-state nanopores also provide some other advantages like multiplex detection [16] and different detectable physical parameters other than ionic current [12, 13, 17–22]. Although it provides so many advantages, it lacks in an important case: the average thickness of the synthetic membranes used for molecular detection is of the order of 10 nm, which will occupy several nucleobases at a time (distance between two consecutive nitrogen bases in a DNA chain is 0.34 nm), jeopardizing single-molecule base-specific detection. Graphene, a single layer of graphite [23], provides a solution to this problem. As the single-layer thickness is of the order of the distance between two consecutive bases in DNA and with various advantageous properties [24], it is the ideal candidate for sequencing applications. (Recently other two-dimensional (2D) materials such as silicene also have been studied for the purpose of DNA sequencing [25].) Graphene also provides several ways of sequential detection, e.g., nanoribbon conductance [26–28] and transverse tunnelling [29, 30]. Readers can consult some review articles [31–34] for a detailed description of nanopore-based sequencing techniques.

In this work we present a theoretical study to detect base-pair mismatches in DNA using the method of nanopore sequencing. Though several studies on ss-DNA sequencing...
al ready exist in the literature [26, 28, 29, 35, 36], there is no such report on double-stranded DNA (ds-DNA). We use a graphene nanopore-based sequencing device that is created on a single-layer zigzag graphene nanoribbon (zgnr) [28]. Using Landauer–Büttiker formalism, we study the changes in the electronic transport properties of the device as a ds-DNA (which also contains base-pair mismatches) translocates through the nanopore. Distinct features have been observed in transmission probability and, to some extent, in \( I-V \) response also for the canonical Watson–Crick pairs and for the four different types of possible mismatches. Study of local density of states (LDOS) also provides applicable insight. Our results open a new pathway for reliable detection of base-pair mismatches in DNA, a highly important diagnostic for genetic disorders.

2. Theoretical formulation

To perform numerical study on the sequential determination of base-pair mismatches in DNA, we use zigzag graphene nanoribbon with a pore created at the centre of it. We preserve the two-sublattice symmetry of graphene while creating the nanopore [37]. The whole zgnr system can be presented by an effective Hamiltonian (see figure 1):

\[
H_{\text{zgnr}} = \sum_{i=1}^{N} (\epsilon c_{i}^{\dagger} c_{i} + t c_{i}^{\dagger} c_{i+1} + \text{H.c.})
\]

where \( \epsilon \) is the site-energy of each carbon atom in zgnr, and \( t \) is the nearest-neighbour hopping amplitude. \( c_{i} \) and \( c_{i}^{\dagger} \) create or annihilate an electron at the \( i \)th site, respectively. For calculation of transport properties we also use semi-infinite zgnr as electrodes [28]. Thus the total Hamiltonian of the system can be written as \( H_{\text{tot}} = H_{\text{zgnr}} + H_{\text{leads}} + H_{\text{tun}} \), where \( H_{\text{tun}} \) represents a tunneling Hamiltonian between the nanopore device and electrodes. In our calculations we scale energy in terms of \( t \), i.e., we set \( t = 1.0 \text{ eV} \). The Hamiltonian of a ds-DNA can be expressed as

\[
H_{\text{DNA}} = \sum_{i=1}^{N} \sum_{j=1}^{L_{d}} (\epsilon_{ij} c_{ij}^{\dagger} c_{ij} + t_{ij} c_{ij}^{\dagger} c_{i+1,j} + \text{H.c.})
\]

\[
+ \sum_{i=1}^{N} (\epsilon_{id} c_{id}^{\dagger} c_{id} + \text{H.c.})
\]

where \( \epsilon_{ij} \) and \( \epsilon_{id} \) are the electron creation and annihilation operators at the \( i \)th nucleotide of the \( j \)th strand, \( t_{ij} \) = nearest-neighbour hopping amplitude between nucleotides along the \( j \)th strand, \( \epsilon_{ij} \) = on-site energy of the nucleotides, and \( v \) = interstrand hopping between the nucleobases.

Green’s function formalism is used for both the LDOS and transport properties calculations. The transmission probability of an electron with an energy \( E \) is given by

\[
T(E) = \text{Tr}[iG^{\dagger}T_{g}G^{\dagger}]
\]

where \( G' = [G_{\text{eff}}] \) and \( \Gamma_{\text{L(R)}} = i[d_{\text{L(R)}} - \Sigma_{\text{L(R)}}] \). \( G' = [E - H_{\text{zgnr}} - \Sigma'_{\text{L}} - \Sigma'_{\text{R}} + i\eta]^{-1} \) is the single-particle retarded Green’s function for the entire system at an energy \( E \), where \( \Sigma'_{\text{L(R)}} = H_{\text{zgnr}}G_{\text{L(R)}}H_{\text{zgnr}} \) represents the retarded (advanced) self energies of the left (right) zgnr electrodes, which is calculated following the recursive Green’s function technique [39, 40]. \( G_{\text{D}}(E) \) is the retarded (advanced) Green’s function of the left (right) lead. At absolute zero temperature, using the Landauer formula, the current through the nanopore device for an applied voltage \( V \) is given by \( I(V) = \frac{2e}{h} \int_{E_F}^{E_F+eV/2} T(E)dE \), \( E_F \) being the Fermi energy. Here we assume that there is no charge accumulation within the system. The LDOS profiles of the base pairs trapped inside the nanopore are given by \( \rho(E, \eta) = -\frac{1}{\pi} \text{Im}[G_{\eta}(E)] \), where \( G(E) = (E - H + i\eta)^{-1} \) is the Green’s function for the zgnr system, including the base pairs with electron energy \( E \) as \( \eta \to 0^{+} \), \( H \) = the Hamiltonian of the zgnr-nanopore, and \( \text{Im} \) represents the imaginary part of \( G_{\eta}(E) \). \( G_{\eta}(E) \) is the diagonal matrix element \( \langle [iG(E)]_{ij} \rangle \) of the Green’s function, \( |i> \) being the Wannier state associated with the trapped nucleotide.

3. Results

For the purpose of numerical investigation we use the ionization potentials of the nitrogen bases as their site energies, which are extracted from the \textit{ab-initio} calculations [41]: \( \epsilon_{G} = 8.178 \), \( \epsilon_{A} = 8.631 \), \( \epsilon_{C} = 9.722 \), and \( \epsilon_{T} = 9.464 \); all units are in eV. Then we shift the reference point of the energy to the average of the ionization potentials of the nucleobases, which is 8.995 eV, and with respect to this new origin of energy, the on-site energies for the bases G, A, C, and T become \(-0.82 \), \(-0.37 \), \(0.72 \), and \(0.47 \) eV, respectively. This is valid for model calculations, as it will not do any qualitative damage to the results. Similar methods have previously been employed where the average of the ionization potential was set as the backbone site energy [42].

In figure 2 we show the LDOS profiles for the four different nitrogen bases. We study this LDOS response of the bases as a part of the Watson–Crick base pairs not as individual; i.e., we trap the AT and GC pairs inside the nanopore.
and study the LDOS profile of the respective bases. The positions of different peaks in the LDOS are different, close to the characteristic site energies of the different nucleotides, and the peak values are also different. These relative differences in LDOS patterns present a chance to detect the base pairs using the angle resolved photo electron spectroscopy (ARPES) technique by trapping them inside the nanopore. As the LDOS behaviour is mostly dominated by the nitrogen bases and not by the backbones [35], this also provides a new way of biomolecular detection.

In figure 3 we plot the variation in transmission probability for different cases. The coupling parameter between the boundary sites of the zgen-nanopore and DNA base is set to 0.2 eV. The intrastrand hopping parameter between identical bases in the DNA chain is taken as $t_{ij} = 0.35$ eV and for different bases $t_{ij} = 0.17$ eV, whereas interstrand hopping between nucleobases is taken as $v = 0.035$ eV, one order of magnitude less than the intrastrand hopping. These values are consistent with previous reports [42–46]. Figure 3(a) shows the comparison between a bare nanopore and a DNA base-pair (GC pair) trapped into the nanopore. The changes in transmission spectra are clearly distinguishable. There are characteristic peaks in the profile both at the $+ve$ and $-ve$ energy range. Both the curves for the bare nanopore and the GC-nanopore are symmetric with respect to zero of energy, as the two-sublattice symmetry of the graphene nanopore is preserved in both the cases. It was violated in the case of ss-DNA sequencing [37]. Figure 3(b) shows the difference between the characteristic features of two Watson–Crick pairs AT and GC. Distinct peaks are present in the transmission profile at and around the characteristic site energies of the respective nucleobases. In figure 3(c) we show the relative changes in the transmission profile for four different types of base-pair mismatches. Each of the mismatches has a distinct response at and around their respective site energies. Variations are quite similar in the $+ve$ and $-ve$ energy range. They are clearly distinguishable at low energy, and the characteristic features die down as we move towards higher energy values. This is due to the fact that as we go to higher energy, we are moving away from the characteristic site energies of the nucleobases. In figure 3(d) we zoom in a small energy window of figure 3(c) for better visualization. The TC mismatch has a distinct peak around 0.3 eV. The GT and AC mismatches become clearly distinguishable between 0.4–0.45 eV and 0.6–0.65 eV, respectively, whereas the AG becomes visibly distinct in the energy range 0.8–0.9 eV.

In figure 4(a) we show changes in the I-V characteristics for a bare nanopore and a GC-nanopore. The effect of the base-pair inside the nanopore becomes prominent at considerable bias; inset shows a specific high-voltage range of the curves where they are clearly distinguishable. Figure 4(b) shows the variation in the current response between the two Watson–Crick pairs AT and GC. They also become differentiable at a high-voltage range between 1.7–2.0 V. The AT pair produces a higher current than the GC pair, which reflects their different electronic structure, as this current response depends on how the local charge density profile is modified due to the insertion of the DNA bases [26]. Figure 4(c) shows the relative differences between four possible mismatches of base pairs. At low bias, differences between them are very faint; they gradually become differentiable as we increase the bias. Insets of figures 4(b) and (c) show specific voltage windows within which the mutual separation between the base pairs is larger than elsewhere.

In figure 5 we finally show the sequencing application to detect base-pair mismatches along with the two canonical pairs AT and GC. We take a 30 base-pair-long random ATGC chain, translocate it through the nanopore, and record the characteristic current signals corresponding to the different base pairs, including the mismatches. During this translocation, bias is kept at 1.72 V; this voltage gives the maximum possible relative separation between the characteristic currents of different base pairs (see insets of figures 4(b) and (c)). Separation between a canonical pair GC and a mismatch TC is maximum, whereas that between AT and AC is minimum. The reason behind this is that G and T are from a different group: G is from a purine group, and T is from pyrimidine; the electronic structures of them are also quite different. So when the pairing changes from GC to TC, the corresponding change in current response is also big, while for AT and AC, both T and C are from the same pyrimidine group; hence the relative changes in the response are also much smaller. These relative changes in the current response represent the differences in their electronic structures. If we define a new quantity to measure the sensitivity of this type of sequencing device, e.g., percentage separation = $(I_{\text{max}} - I_{\text{min}})/I_{\text{min}}$, it turns out to be that the maximum and minimum values of the percentage separation achieved are 17.30% and 3.23%, which implies that the current signals for the respective base pairs can be detected with much more reliability. We also plot a separate figure (see right panel of figure 5) for a normal ds-DNA chain without any mismatches for better understanding of the effect of mismatches on the current response of the device. It is also important to mention that though we have presented current in arbitrary units, if we put the numerical values of various constants, e.g., h, e, and $h$, it turns out that the currents are of the order of 10 $\mu$A, which is much higher than previous reports on ss-DNA sequencing as well as much greater than the noise level of this type of device, which is of
the order of nA [28]. Very recently a report by Feliciano et al [47] on the dynamical effects of environment on operation of graphene-based sequencing devices showed that fluctuations of the nucleotides inside the nanopore may change the conductance of the devices relying on tunneling mechanism, though they conclude that these effects would not be very important for the devices, which rely on transverse conductance with a larger transmission probability. As our proposed device relies on transverse conductance and produces greater current output, the effect of these types of noises will be much lesser, whereas another study by Krems et al [48] in 2009 dealing with different types of noises that may occur in actual sequencing experiments showed that these environmental effects do not strongly influence the current distributions and working efficiency of these devices. However, based upon these results we can say that the overall sensitivity of our device will not be hampered too much, but there will always be sources of noise in actual experimental condition due to environmental fluctuations, the presence of water, and counterions, which can affect the device operation. It is also important to note that this is one of the early attempts to detect base-pair mismatches by means of nanopore sequencing, and the results given in this work are open to improvement in different ways. One example is by functionalization of the edge atoms of the nanopore, which can significantly enhance nucleobase–pore interaction, thus reducing the structural noise by enhancing the graphene–nucleobase electronic coupling [49, 50]. Different types of groups can be used for functionalization (e.g., hydroxide [51], amine, or nitrogen [28]) to provide a custom-made solution to overcome noise in electrical DNA sequencing techniques. It is also true for the devices relying on transverse conductance that most of the current passes through the edges of the nanoribbon, which is one of the reasons for the poor sensitivity of these types of devices, but this can be controlled with accurate engineering of the nanopore device dimension. See the appendix section for more details.

4. Conclusion

In summary, we present an effective and sound technique to detect base-pair mismatches in a given DNA sample. We analyzed different properties from LDOS to \( I-V \) response in connection with sequential determination and found distinguishable signatures in most of the cases. Most of the earlier results on DNA sequencing use ss-DNA, which neglects the basic problem of base-pair mismatch, leading to different neuro-degenerative diseases. As the different genetic diseases occur due to a mismatch of the base pair i.e., when a nitrogen base in a DNA double-helix is paired up with another base that is not the its complementary pair, sequencing of ss-DNA cannot provide this information. On the other hand, previous attempts to detect base-pair mismatches do not provide any
decisive results. With time both medical science and genetic research progress, and the reasons behind different genetic disorders, including neuro-degenerative ones (like Parkinson’s, Alzheimer’s, etc), are becoming more and more transparent. With this progress the need for low-cost and reliable DNA sequencing also increases, which should also provide the necessary technique for proper medical applications. In these circumstances we present a reliable tight-binding scheme to detect base-pair mismatches in DNA with much better accuracy than previous studies [52]. At the same time, we also understand that the proposed technique needs more improvements for actual application in a real environment and hope it will soon be tested with further modifications.

Figure 4. Current–voltage response of the active nanopore device for different cases. (a) Comparison of the current responses between a bare nanopore and a GC-nanopore. (b) Difference between characteristic current amplitudes of two Watson–Crick base pairs AT and GC. (c) Attributes of four different mismatches (AG, AC, GT, TC). Insets show selective voltage ranges for better visualization. $E_F = 0$ eV represents Fermi energy.

Figure 5. Left panel shows the stop-and-go translocation of a random ATGC ds-DNA chain through the nanopore, while bias across the device is fixed to a specific value ($1.72$ V), which gives maximum separation in current responses for different base pairs. We record the characteristic current output for the bases as they translocate through the nanopore. The respective base pairs and mismatches are indicated in the figure with their usual symbols (AT, GC, etc.). Right panel shows the same variation for a ds-DNA chain with no base-pair mismatch for better understanding of the left-panel figure. Though the current is presented in arbitrary units as we report a model calculation, if we put in the exact numerical values of different constants like $h$, $e$, and $\hbar$, it turns out of the order of $10 \mu$A.
Appendix

In this section we provide some additional information on base-pair detection of DNA. In figure 6 we plot the variation in the current response of our proposed device for AT and TA base pairs, both being the Watson–Crick pair. Now for the previous calculation we preserve the two-sublattice symmetry of graphene by symmetrically connecting the nucleotides with edge atoms; in this configuration it is hard to distinguish AT and TA separately. For better detectability we destroy the two-sublattice symmetry and find distinct responses. The same also has been done for detection between GC and CG. We want to mention that we checked all our results with the broken sublattice symmetry but find no significant changes for the results presented in the earlier sections. The percentage separation between AT and TA (GC and CG) is relatively small (1.5%), which implies that the proposed device is not effective in the same way as it is for base-pair mismatches.

We also check the sensitivity of the device on the nanoribbon width. To investigate this we make the zgnr width double of the previous results but keep the pore size fixed. In figure 7 we plot the sequential determination, i.e., the stop-and-go translocation of a ds-DNA chain containing mismatches, through the zgnr-nanopore with increased width. With increasing width the current output increases, which is trivial, as the width increases conductance of the device will also increase and so the current. But the sensitivity decreases to some extent. As we keep the pore size fixed, the fraction of the current passing around the pore will decrease with increasing nanoribbon width and the signature of the base pairs will also die out as the presence of the base pairs modifies this current only, which is then detected by the device. For the previous case (figure 5) the range of current variation is 0.09 (arbitrary unit) for different base pairs, which reduces to 0.06 (arbitrary unit) as we double the width of the zgnr.

Following the above results (figure 7), we can say that there are several issues competing in the sequential detection technique. First is that to get higher current output from the device, one has to increase the ribbon width, but this will also hamper device sensitivity to some extent. In order to maintain the desired accuracy one has to increase the nanopore dimension with increasing ribbon width. Increasing the pore size will increase the fraction of current passing around the pore, and the effect of the base pairs will also become more vivid, because only the changes in the current passing around the nanopore due to the presence of the base pair are detected by the device. To reduce the fluctuations of the base pairs inside the nanopore during translocation, the edge atoms of the nanopore have to be functionalized with different groups [28, 32, 51], as discussed in the earlier section. Thus, in the case of the sequential determination process of DNA or biomolecules, there are several parameters which have to be optimized accordingly for accurate and precise measurement.

References

[1] Loft S and Poulsen H E 1996 J. Mol. Med. 74 297
[2] Sanger F, Nicklen S and Coulson A R 1977 Proc. Natl Acad. Sci. USA 74 5463
