PCCP

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue



Cite this: *Phys. Chem. Chem. Phys.*, 2023, **25**, 7750

# Charge transport properties of ideal and natural DNA segments, as mutation detectors

Marilena Mantela, Konstantinos Lambropoulos and Constantinos Simserides *

DNA sequences of ideal and natural geometries are examined, studying their charge transport properties as mutation detectors. Ideal means textbook geometry. Natural means naturally distorted sequences; geometry taken from available databases. A tight-binding (TB) wire model at the base-pair level is recruited, together with a transfer matrix technique. The relevant TB parameters are obtained using a linear combination of all valence orbitals of all atoms, using geometry, either ideal or natural, as the only input. The investigated DNA sequences contain: (i) point substitution mutations – specifically, the transitions guanine (G) ↔ adenine (A) – and (ii) sequences extracted from human chromosomes, modified by expanding the cytosine–adenine–guanine triplet [$(CAG)_n$ repeats] to mimic the following diseases: (a) Huntington's disease, (b) Kennedy's disease, (c) Spinocerebellar ataxia 6, (d) Spinocerebellar ataxia 7. Quantities such as eigenspectra, density of states, transmission coefficients, and the – more experimentally relevant – current–voltage ($I$–$V$) curves are studied, intending to find adequate features to recognize mutations. To this end, the normalised deviation of the $I$–$V$ curve from the origin (NDIV) is also defined. The features of the NDIV seem to provide a clearer picture, being sensitive to the number of point mutations and allowing to characterise the degree of danger of developing the aforementioned diseases.

Received 17th January 2023,
Accepted 2nd February 2023

DOI: 10.1039/d3cp00268c

rsc.li/pccp



## 1 Introduction

DNA is the repository of genetic information in organisms. Each DNA double helix strand consists of a sugar-phosphate backbone connecting bases: purines [adenine (A), guanine (G)] and pyrimidines [thymine (T), cytosine (C)]. A purine (G or A) of one strand is joined with a pyrimidine (C or T) of the other strand *via* (three or two) hydrogen bonds, respectively. G is normally bonded with C, and A with T, unless a mutation occurs, *e.g.*, a transition (interchange of purines, A ↔ G, or of pyrimidines, C ↔ T) or a transversion (interchange purine ↔ pyrimidine).[1–4] The structure of DNA favours the overlap of the electron density of adjacent bases, which, besides stabilising the double helix, creates a nearly one-dimensional π-pathway along which charge transfer and transport are possible. The term transfer implies that a carrier, created (*e.g.* a hole by oxidation or an extra electron by reduction) or injected at a specific location, moves to more favourable sites, without application of external gradient (*e.g.* temperature gradient or voltage). The term transport implies the application of an external gradient. In this work, we consider electrodes between which voltage is applied. Experimental studies aiming to understand the properties and mechanisms of charge transfer and transport along DNA span over than two decades.[5–7] The most experimentally relevant quantities for charge transfer and transport are the transfer rates and the current–voltage curves, respectively. Charge movement along DNA is usually studied within two main frameworks: (i) incoherent or thermal hopping, (ii) coherent tunnelling or both.[8–10] The term tunnelling implies quantum mechanical tunnelling through the sequence. The coherent mechanism is expected to be dominant in the low temperature regime.[11] In thermal hopping, the carrier is localized and exchanges energy with the environment, hence it can travel longer than *via* the coherent mechanism. However, the presence of the two mechanisms is not mutually exclusive.[9,12,13] It is probable that long-range hole transfer in DNA proceeds *via* the incoherent multistep mechanism: a hole moves by hopping between neighbouring sites, but holes could be localized on a single base or spread over a number of bases, especially in sequences containing consecutive homobases.[14] In this work, coherent charge transport is studied.

Charge transfer and transport through the aromatic base-pair stack depends on the electronic coupling between adjacent bases. Therefore, *e.g.* distortions[15,16] affect charge transfer and transport. Also, deviations in that stacking, *e.g.*, through base modifications, insertions, or protein binding, can be electrically

*Department of Physics, National and Kapodistrian University of Athens, Panepistimiopolis, Zografos, GR-15784 Athens, Greece.*
*E-mail: mmantela@phys.uoa.gr, klambro@phys.uoa.gr, csimseri@phys.uoa.gr*

observed. DNA charge transfer and transport has been used to detect changes in DNA, like lesions, mismatches, mutations, binding proteins, protein activity, even reactions under weak magnetic fields.[17] Charge transfer and transport properties and long-range oxidation of DNA provide an understanding of its biological role and reveal potential nano-applications, such as nanosensors, nanocircuits, and molecular wires.[18–20] In the context of biomedicine, these properties can be used to detect pathogenic mutations at early stage. For example, the pairing of non-complementary bases leads to point mutations which are potentially harmful to the development of organisms (carcinogenesis). Each DNA sequence has a unique electronic signature, which may be useful for identifying a mutant DNA molecule.[21,22] Thus, charge transfer and transport can bring valuable information about sequencing. It is expected that these properties can be further employed to design electronic circuits as diagnostic tools.

Considering the above, this work focuses on charge transport along DNA molecules, using the Tight-Binding (TB) method, together with the transfer matrix technique, to solve the time-independent Schrödinger equation and finally obtain *I–V* curves. We study double-stranded DNA molecules, the ends of which are connected to electrodes, focusing on: (1) both ideal and natural geometries. (2) Two types of mutations: (i) point substitution mutations, specifically, transitions G ↔ A, and (ii) sequences extracted from segments of human chromosomes, modified by expanding the CAG triplet to mimic the following diseases: (a) Huntington's disease, (b) Kennedy's disease, (c) Spinocerebellar ataxia 6, (d) Spinocerebellar ataxia 7. Physical quantities such as eigenspectra, density of states, transmission coefficients, and current–voltage curves are obtained. The parameters used to describe the molecular electronic structure of nucleic acid bases and extract the on-site energies and the interaction integrals used in the recruited TB wire model were obtained from the linear combination of atomic orbitals (LCAO) method, considering the molecular wave function as a linear combination of all valence orbitals of all atoms, *i.e.*, 2s, $2p_x$, $2p_y$, $2p_z$ orbitals for C, N, and O atoms and 1s orbital for H atoms.

The novel features of this work compared to state of the art include the following: (1) Ideal and natural DNA geometries are compared. (2) Known mutations are examined, and mutated sequences, either containing point substitution mutations (G ↔ A transitions) or extracted from human chromosomes and modified by expanding the CAG triplet to mimic diseases, are compared to unmutated ones. (3) The potential use of physical quantities related to charge transport as mutation detectors is investigated. (4) The normalised deviation of the *I–V* curve from the origin (NDIV), which seems to be a useful quantity for that purpose, is defined.

The rest of this article is organized as follows: Section 2 includes a description of the employed methods; the studied sequences and genetic disorders are listed in Section 3; in Section 4, results for various physical quantities are presented and discussed; finally, Section 5 contains our conclusions and some reflections on perspectives.

# 2 Methods

TB is an approach to the calculation of the electronic structure of materials which assumes that the systems orbitals are tightly bound at the sites they belong. When the overlap of neighbouring sites' orbitals is small, the system's state can be expressed as a superposition of the states corresponding to isolated sites. The system's eigenspectra are distributed relatively close to the opposite of the ionization energies of isolated sites because the interaction with neighbouring sites is relatively limited. TB is semi-empirical; specifically, two sets of parameters are needed for its application: (a) the on-site energies of the orbitals at each site, and (b) the interaction integrals that quantify the coupling between orbitals belonging to neighbouring sites. Although the terms hopping or transfer integrals or parameters are usually used, the term interaction integral or parameter is generic, not reminiscent of the nature of carrier movement or specific problem. The simplicity of TB makes it appropriate to obtain analytical and numerical results with minimal computational cost, in contrast to *ab initio* techniques. Additionally, given an appropriate parametrization, TB can handle the effects induced by the variation of DNA sequences, hence any given segment can be treated within TB. On the other hand, *ab initio* methods such as Density Functional Theory (DFT) and Hartree–Fock Crystal Orbital (HFCO) theory, cannot do so in a straightforward manner, due to computational limitations[23] and underlying theoretical assumptions,[24,25] respectively.

In the present work, the so-called ''wire model'' variant of the TB method is employed. For double-stranded DNA, the wire model is essentially a description at the base-pair level, *i.e.*, the DNA polymer is considered as a wire, composed of successive base pairs (or monomers). The parameters required for the wire model description are the on-site energies of the base pairs and the interaction integrals between successive base pairs. In order to produce the required on-site energies, the Linear Combination of Atomic Orbitals (LCAO) method was employed, considering the molecular wave function as a linear combination of all valence orbitals, *i.e.*, of the 2s, $2p_x$, $2p_y$, $2p_z$ orbitals for C, N, and O atoms and the 1s orbital for H atoms. A novel parameterization was used, initially introduced in ref. 26. As for the interaction integrals, a Slater–Koster two-centre interaction form[27] was employed, using Harrison-type expressions,[28,29] with slightly modified factors relative to the original ones.[26] These parameters have been calibrated by comparing our LCAO predictions for the ionization and excitation energies of heterocycles with those obtained from the Ionization Potential Equation of Motion Coupled Cluster with Singles and Doubles (IP-EOMCCSD)/aug-cc-pVDZ level of theory and the Completely Renormalized Equation of Motion Coupled Cluster with Singles, Doubles, and non-iterative Triples (CR-EOMCCSD(T))/aug-cc-pVDZ level of theory, respectively (vertical values), as well as with experimental data.[30]

The problem to be solved, *i.e.*, the time-independent or time-dependent Schrödinger equation of the polymer, is reduced to a system of coupled algebraic equations or differential equations of first order, respectively. For example, the time-independent

TB system of equations, from which the eigenenergies, $E$, are obtained, for a DNA segment within the wire model, reads

$$E\psi_n = E_n\psi_n + t_{n-1,n}\psi_{n-1} + t_{n,n+1}\psi_{n+1}, \quad \forall n = 1, 2, \ldots, N, \tag{1}$$

where $E_n$ is the on-site energy of monomer $n$, $|\psi_n|^2$ is the occupation probability at monomer $n$, and $t_{n,n+1} := t_n$ is the interaction integral between neighbouring monomers $n$ and $n+1$. In the rest of this article, the 5′–3′ strand is used to denote the DNA segments. Therefore, the notation GG implies two base pairs of GG bases in the 5′–3′ strand and their complementary ones, CC, in the 3′–5′ strand. Eqn (1) can alternatively be written in the matrix form

$$\begin{pmatrix} \psi_{n+1} \\ \psi_n \end{pmatrix} = \begin{pmatrix} \frac{E-E_n}{t_n} & -\frac{t_{n-1}}{t_n} \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \psi_n \\ \psi_{n-1} \end{pmatrix} = P_n(E) \begin{pmatrix} \psi_n \\ \psi_{n-1} \end{pmatrix}, \tag{2}$$

and solved using the transfer matrix method. $P_n(E)$ is called the transfer matrix of monomer $n$. The product

$$M_N(E) = \prod_{n=N}^{1} P_n(E). \tag{3}$$

is the global transfer matrix (GTM) of the sequence, and contains all the information about its energetics.

## 3 Studied sequences and genetic disorders

Genetic disorders occur when a mutation affects an organism's genes or when the organism has the wrong amount of genetic material. Carrying the mutation does not necessarily mean that the organism will end up with a disease. There are many types of disorders, including single-gene, multifactorial and chromosomal ones. This work focuses on single-gene genetic disorders, *i.e.*, changes or mutations that occur in the DNA sequence of a single gene and are otherwise known as monogenetic disorders.

Point substitutional mutations are common; the G-T mismatch mutation alone occurs about once in every $10^4$–$10^5$ base pairs. Cell viability and health are highly dependent on keeping the mutation rate small. The high fidelity of DNA replication is established and secured by an enzyme, the replicative polymerase, though several mechanisms: (1) sensing proper geometry of the correct base pair, (2) slowing down catalysis in case of a mismatch, and (3) partitioning the mismatched primer to exonuclease active site.[31] However, the performance of polymerases is not error-free: it is estimated[31–33] that, even after proofreading, the overall fidelity of DNA synthesis lays in the range of one wrong nucleotide incorporated per $10^3$–$10^5$. Besides, DNA replication is constantly challenged by internal and external factors, non-canonical DNA structures, and complex DNA sequences.[31]

Another category of DNA mutations related to several diseases is the short tandem repeat (STR) expansions or microsatellites.[34,35] These are small sections of DNA, usually 2–6 nucleotides long, repeated at a defined region. At least 6.77% of the human genome is comprised of these repetitive DNA sequences.[35] Large STR expansions are potentially pathogenic, setting the ground for several neurological diseases. In fact, 37 of the already known STR genes that can cause disease when expanded, exhibit primary neurological presentations.[35] In neurological STR diseases, 'CAG' repeat expansions code for the amino acid glutamine. When expanded, they create polyglutamine tract expansions, which are thought to alter and expand the transcribed protein, creating insoluble protein aggregates within neuronal cells. This can cause perturbations in intracellular homeostasis and cell death.[36]

Two categories of DNA polymers are examined in this work: (i) sequences that contain point substitution mutations (specifically, transitions involving G ↔ A exchange), of both ideal and natural geometries, replacing the out-of-ring atoms that are different between A and G, while ensuring that the number of hydrogen bonds is correct, and (ii) sequences of ideal geometry extracted from segments of human chromosomes, subsequently modified by a CAG triplet expansion [$(CAG)_n$ repeats], to mimic four selected STR diseases, namely, (a) Huntington's disease, (b) Kennedy's disease, (c) Spinocerebellar ataxia 6, (d) Spinocerebellar ataxia 7. The number of pathogenic repeats, *i.e.*, CAG triplets, in Huntington's disease, is $n_p$ = 36–250, located in exon 1 of HTT gene, chromosome 4.[35,37,38] In spinal and bulbar muscular atrophy of Kennedy (Kennedy's disease), $n_p$ = 38–68, located in exon 1 of AR gene, chromosome X.[35,39–41] In spinocerebellar ataxia 6, $n_p$ = 19–33, located in exon 47 of CACNA1A gene, chromosome 19.[35,42,43] In spinocerebellar ataxia 7, $n_p$ = 34–460, located in exon 1 of ATXN7 gene, chromosome 3.[35,44,45]

In summary, the protocol to implement mutations in this study is the following: (i) for transitions involving G ↔ A exchange, of both ideal and natural geometries, the out-of-ring atoms that are different between A and G are replaced, while ensuring that the number of hydrogen bonds is correct. (ii) For all diseases with the same triplet motif, $(CAG)_n$, we keep 9 base pairs at the start and 9 base pairs at the other end of the sequence (primers). Of course, primers are different for each disease, but they only contain 18 base pairs altogether, which is not a large number when dealing with a sequence of 180 or 300 base pairs. Of course, the procedure used in the present work could be ameliorated in future studies.

## 4 Results and discussion

Section 4.1 is devoted to eigenenergies and densities of states, Section 4.2 to transmission coefficients and Section 4.3 to current ($I$)–voltage ($V$) curves, where the normalised deviation of the $I$–$V$ curve from the origin (NDIV) is introduced. In this work, we focus on charge transport through the highest occupied molecular orbitals (HOMOs).

For DNA segments of ideal geometry, the base pairs are not distorted; they are separated and twisted by 3.4 Å and 36°, respectively, relative to the double helix growth axis. The geometries of the natural sequences $G_{14}$ and $A_{15}$ have been extracted from Bioinformatics (RCSB) Protein Data Bank (**https://www.rcsb.org**) [accession numbers 4WZW and 6VAA, respectively], from the original ref. 46 and 47, respectively.

 

Table 1 HOMO ($E_H$) and LUMO ($E_L$) energies of ideal DNA bases, *i.e.*, A, T, G, C and base pairs, *i.e.*, A–T, G–C, A–C, obtained *via* LCAO with all valence orbitals.[26] Molecular orbitals are of π or π* character, unless otherwise stated. All values are given in eV

| Base or base pair | $E_H$ | $E_L$ |
|---|---|---|
| A | −8.50 | −4.19 |
| T | −9.12 | −4.30 |
| G | −8.31 | −4.12 |
| C | −8.67 | −4.43 ($\sigma^*$)<br>−4.11 |
| A–T | −8.49 | −4.31 |
| G–C | −8.30 | −4.43 ($\sigma^*$)<br>−4.14 |
| A–C | −8.43 | −4.43 ($\sigma^*$)<br>−4.23 |

The on-site energies and interaction integrals for all sequences were calculated using all valence orbitals of all atoms, according to the procedure described in ref. 26. For ideal sequences, the on-site energies are $E_{A–T} = -8.49$ eV for the A–T base pair, and $E_{G–C} = -8.30$ eV for the G–C base pair.[26] In Table 1, the energies of the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO) of the two B-DNA base pairs, A–T and G–C, are shown. The bases constituting these base pairs are slightly deformed relative to their geometry in gas phase. Table 1 also contains the HOMO and LUMO energies of these slightly deformed bases. These levels are of π and π* character, unless otherwise stated. The on-site energy of the mismatched A–C base pair was calculated, as well. Its HOMO value is $E_{A–C} = -8.43$ eV, *i.e.*, very close to the one of the A–T base pair. The HOMO and LUMO interaction integrals between successive base pairs of ideal geometry, without mismatches, calculated with the method described in ref. 26, using all valence orbitals of all atoms, are listed in Table 2. Mutations and distortions change the values of interaction integrals; this effect is included in present work, using the same method.[26]

*Observation*: the HOMO (LUMO) of a base pair is very close to the highest HOMO (lowest LUMO) of its constituent bases,[48] *cf.* Table 1. Hence, studying charge transport through HOMOs within the TB wire model (as done here), it is practically easier to examine purine substitution mutations; given that purines have higher HOMOs than pyrimidines, this substitution has a substantial effect on the base pair on-site energy. This generates an important diagonal disorder to the TB wire model Hamiltonian matrix, in addition to the always-present off-diagonal disorder caused by the modification of interaction parameters.

### 4.1 Eigenspectra and density of states

The term eigenspectra is used to describe the whole set of eigenenergies. The eigenspectra of the studied sequences were calculated by numerical diagonalisation of the Hamiltonian matrices, which, within the TB wire model, are real, tridiagonal and symmetric matrices.[48,49] A quantity used to describe the energy structure of a given system is the density of states (DOS), which shows the number ($N_E$) of states that can be occupied by electrons per energy ($E$) interval, namely, $\frac{dN_E}{dE}$. A useful equivalent definition is

$$g(E) = \sum_k \delta(E - E_k), \tag{4}$$

where the double spin degeneracy is not incorporated. The integrated density of states (IDOS) refers to the number of states that have energy smaller than $E$, and is defined as

$$\mathrm{IDOS}(E) = \int_{-\infty}^{E} g(E')\mathrm{d}(E'). \tag{5}$$

Table 2 Absolute values of interaction parameters between HOMOs (LUMOs), $|t_{LCAO}|$, for all possible combinations of successive base pairs, obtained *via* LCAO, using all valence orbitals, for close-to-ideal geometrical conformations.[26] All values are given in meV. The first row refers to dimers containing solely G–C monomers, the second row refers to dimers containing solely A–T monomers, the third row refers to dimers containing both G–C and A–T monomers, and the fourth row refers to dimers containing A–C mismatched monomers (denoted by Am) and G–C monomers. XY denotes the sequence in the 5′–3′ direction

| GG, CC | GC | CG |
|---|---|---|
| 116<br>(92($\sigma^*$), 2) | 10<br>(2($\sigma^*$), 19) | 75<br>(1($\sigma^*$), 9) |

| AA, TT | AT | TA |
|---|---|---|
| 38<br>(22) | 50<br>(1) | 37<br>(2) |

| AG, CT | TG, CA | AC, GT | TC, GA |
|---|---|---|---|
| 37<br>(11($\sigma^*$), 11) | 28<br>(2($\sigma^*$), 9) | 16<br>(1($\sigma^*$), 1) | 142<br>(3($\sigma^*$), 6) |

| GAm | AmG | AmAm |
|---|---|---|
| 130<br>(89($\sigma^*$), 8) | 31<br>(90($\sigma^*$), 20) | 36<br>(90($\sigma^*$), 25) |

**4.1.1 Unmutated sequences.** The eigenspectra and corresponding DOS of some representative examples of unmutated ideal and natural DNA homopolymers (or homo-oligomers), G… and A…, are depicted in Fig. 1. It is clear that the eigenstates of ideal G… and A… homopolymers are symmetrically positioned around the on-site energy of the ideal G–C or A–T base pair, respectively. Regarding natural sequences, although the eigenspectra are still close to the corresponding ideal G–C or A–T base pair on-site energy, they are no longer symmetrically positioned, due to the presence of diagonal and off-diagonal disorder. The corresponding normalised IDOS can be found in Fig. 2.

**4.1.2 Mutated sequences**

*(i) Point substitution mutations.* The transition G ↔ A occurs by introducing A instead of G in G… polymers: the pyrimidine

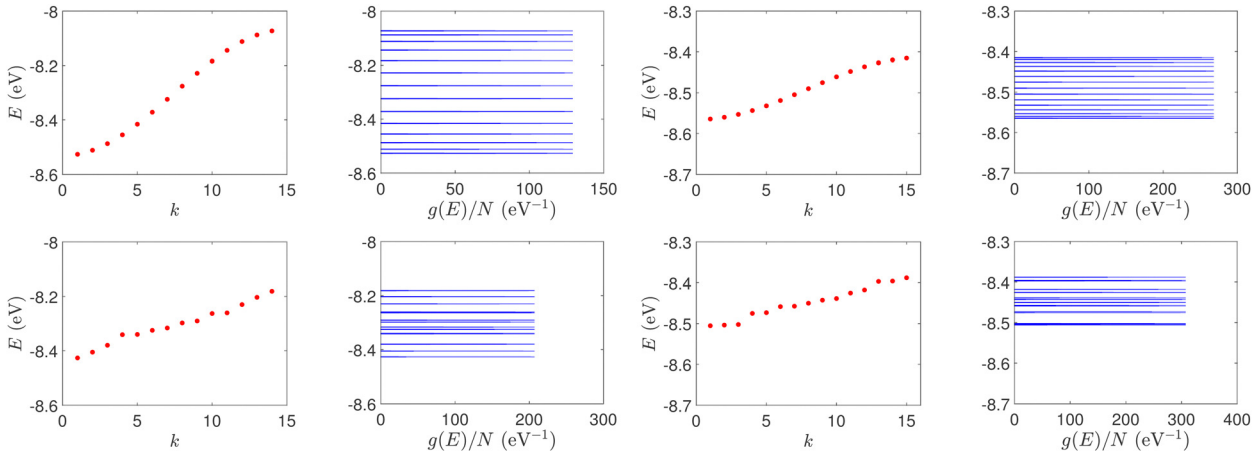


Fig. 1 Eigenspectra and DOS of unmutated DNA homopolymers. Upper panels: Ideal, lower panels: natural, left panels: $G_{14}$, right panels: $A_{15}$. The geometries of the natural sequences $G_{14}$ and $A_{15}$ have been extracted from the Bioinformatics (RCSB) Protein Data Bank (https://www.rcsb.org) [accession numbers 4WZW and 6VAA, respectively] from the original ref. 46 and 47, respectively. $k$ is the eigenenergy index.

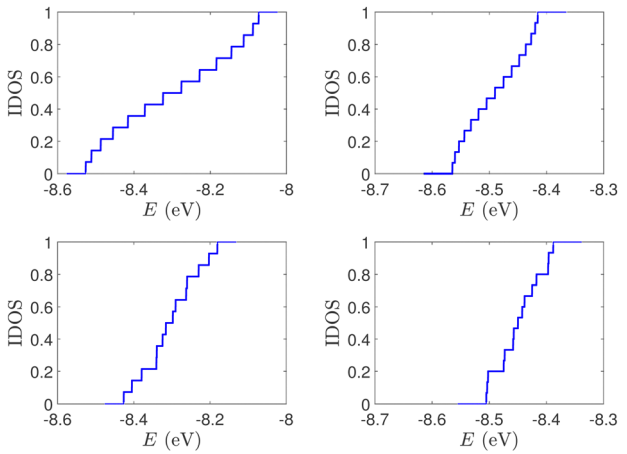


Fig. 2 Normalized IDOS of the unmutated DNA sequences depicted in Fig. 1. Upper panels: Ideal, lower panels: natural, left panels: $G_{14}$, right panels: $A_{15}$.

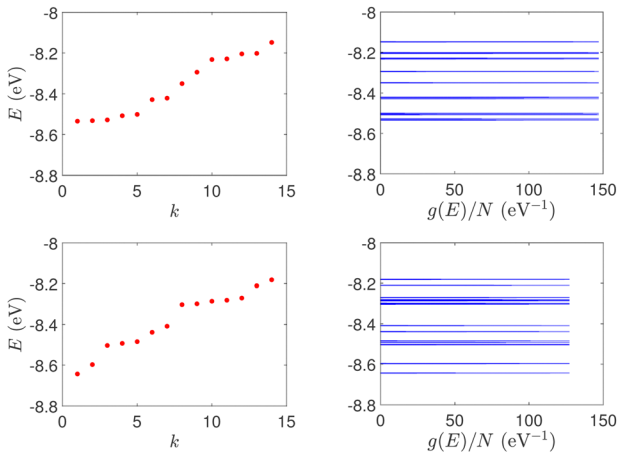


Fig. 3 Eigenspectra and DOS of $G_{14}$ sequences with 7 randomly positioned A–C mismatch mutations. The purine strand contains 7 G and 7 A, distributed randomly, while the pyrimidine strand contains 14 C. Upper panels: Ideal polymers, lower panels: natural polymers. This figure can be compared with the left part of Fig. 1.

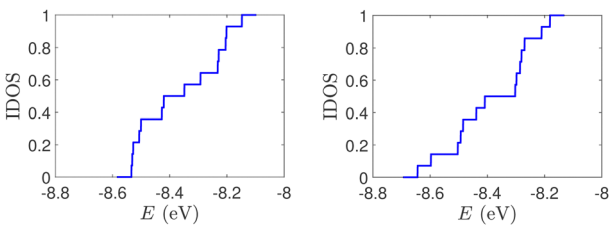


Fig. 4 Normalized IDOS of $G_{14}$ sequences with 7 randomly positioned A–C mismatch mutations. The purine strand contains 7 G and 7 A, distributed randomly, while the pyrimidine strand contains 14 C. Left panel: Ideal polymer, right panel: natural polymer. This figure can be compared with the left part of Fig. 2.

strand still contains only cytosines, but in the purine strand guanines are replaced by adenines. Hence, the replaced base pairs are A–C, instead of G–C. When such mismatches are introduced, the respective interaction integrals are modified, according the procedure described above, *i.e.*, using the input geometry and LCAO with all valence orbitals of all atoms.[26] The eigenspectra and the corresponding DOS for the studied ideal and natural $G_{14}$ sequences with 7 A–C mismatch mutations, randomly positioned in the sequence, are presented in Fig. 3. The corresponding normalised IDOS can be found in Fig. 4. In other words, the purine strand contains 7 G and 7 A randomly distributed, while the pyrimidine strand still contains 14 C. Comparing Fig. 1 with Fig. 3, it can be observed that, apart from the increased irregularity of the mutated sequences, there is roughly a movement of the mean value of eigenenergies from around $E_{G-C}$ towards around $E_{A-C}$ (*cf.* Table 1). It can be observed that the number of levels close to $E_{A-C}$ is increased. In particular, the natural mutated sequence displays a high density of levels closer to both $E_{G-C}$ and $E_{A-C}$. This can also be understood by inspecting the IDOS, *i.e.*, by comparing Fig. 2 with Fig. 4.

 

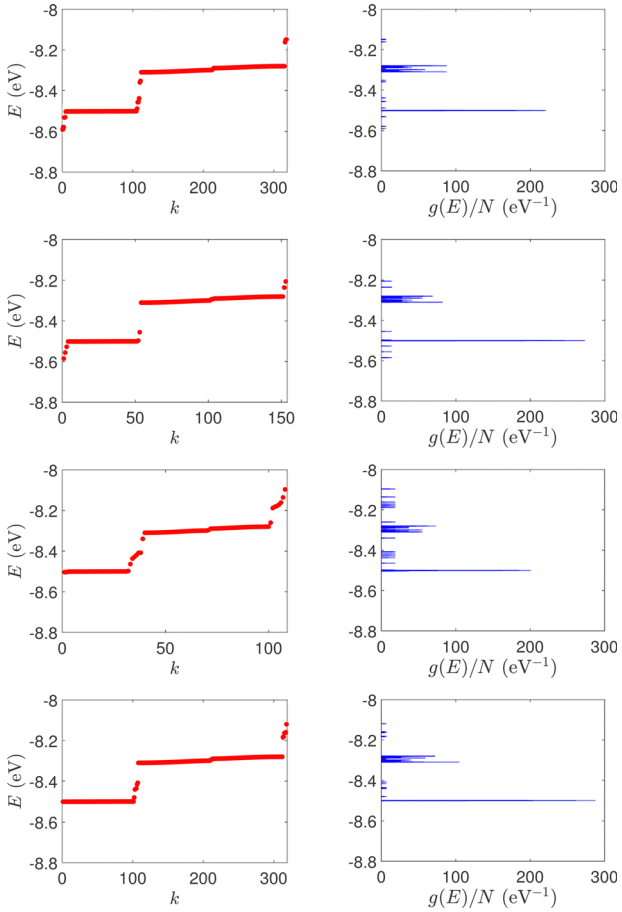


Fig. 5 Eigenspectra and DOS of the studied DNA sequences with STR expansion mutations. (a) First row: Huntington's disease with 100 STR expansions, (b) second row: Kennedy's disease with 45 STR expansions, (c) third row: Spinocerebellar ataxia 6 with 30 STR expansions, (d) fourth row: Spinocerebellar ataxia 7 with 100 STR expansions.

*(ii) Short tandem repeat (STR) expansions.* Four important cases of STR expansions are examined: (a) Huntington's disease, (b) Kennedy's disease, (c) Spinocerebellar ataxia 6, (d) Spinocerebellar ataxia 7. The complete sequences, including primers,[50] used in these examples are: (a) Huntington's disease: AAGTCCTTC$(CAG)_{100}$CAACAGCCG, (b) Kennedy's disease: CTGCTGCTG$(CAG)_{45}$CAAGAGACT, (c) Spinocerebellar ataxia 6: GGGCCCCCG$(CAG)_{30}$GCGGTGGCC, (d) Spinocerebellar ataxia 7: GCCGCCCGG$(CAG)_{100}$CCGCCGCCT. The eigenspectra and corresponding DOS for the studied sequences with STR expansion mutations are presented in Fig. 5. Two subbands occur around the on-site energies, $E_{G-C} = -8.30$ eV and $E_{A-T} = -8.49$ eV, in addition to scattered features, due to the presence of primers. The corresponding normalised IDOS can be found in Fig. 6.

### 4.2 Transmission coefficient

Charge transport properties are studied under the assumption that the DNA sequences of interest lie on sites $n = 1, 2, \ldots, N$, and are connected with two semi-infinite homogeneous metallic electrodes (leads), acting as carrier baths, which lie on sites $(-\infty, 0] \cup [N+1, +\infty)$. The leads are described by properly

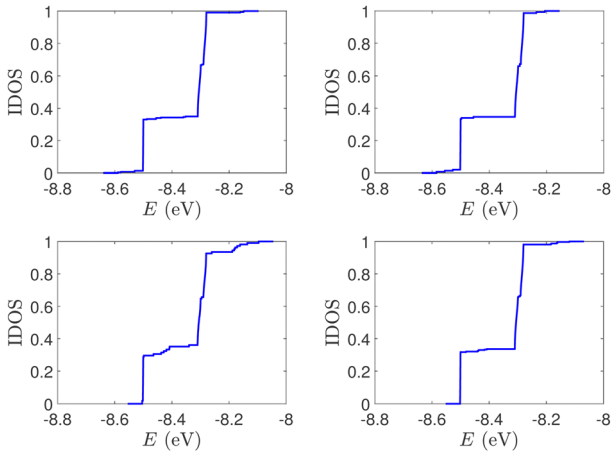


Fig. 6 IDOS of the studied DNA sequences with STR expansion mutations. (a) Upper left: Huntington's disease with 100 STR expansions, (b) upper right: Kennedy's disease with 45 STR expansions, (c) left lower: Spinocerebellar ataxia 6 with 30 STR expansions, (d) lower right: Spinocerebellar ataxia 7 with 100 STR expansions.

chosen on-site energies and interaction parameters. In this work, left and right electrodes are assumed to be identical. The electrodes energy spectrum is given by the dispersion relation[51] $E = E_m + 2t_m\cos(qa)$, where $E_m$ is the on-site energy of the electrodes, $t_m$ is the interaction integral between the electrodes sites, and $a$ is the lattice constant. The electrodes' band lies in the energy interval $[E_m - 2|t_m|, E_m + 2|t_m|]$. Hence, the energy center and bandwidth of the electrodes are $E_m$ and $4|t_m|$, respectively. It should be noted that the nature of the mobile charges (holes or electrons) depends upon factors such as the availability of states, their filling, and the alignment to the Fermi levels of the leads.[52] Here, the nature of the leads is not specified. However, we use HOMO levels, and it is assumed that $E_m$ is either equal to $E_{G-C}$ or $E_{A-T}$: for ideal or natural G…, $E_m = E_{G-C}$; for ideal or natural A…, $E_m = E_{A-T}$; for G… sequences containing A–C mutations, $E_m = E_{G-C}$; for diseases, the value $E_m = E_{G-C}$ is still used. If the lead is modelled as a homogeneous system with one electron per site, then the band is half-filled, the electrodes are metallic, and the Fermi level of the electrodes, $E_m^F$, is identified with the on-site energy of the electrodes, $E_m$. In this article, the value $|t_m| = 0.5$ eV is assumed, so that the leads' bandwidth contains all the eigenstates of the studied sequences; *cf.*, Fig. 1–6.

The details of the lead-DNA interface are rather complex;[11] the coupling of the sequence with the edge electrode sites is described by the effective interaction integrals $t_{cL(R)}$. The choice of appropriate parameters is important, since the quality of the contact plays a crucial role in charge transport;[53] in fact, it defines the optimum transport profile.[11,54] For periodic sequences, the coupling strength $\omega = \frac{t_m t_N}{t_{cR} t_{cL}}$ and coupling asymmetry $\chi = \frac{t_{cL}}{t_{cR}}$ have been defined previously.[54] The ideal coupling condition, which is definable only in periodic cases, is $|\omega| = 1$. The symmetric coupling condition is $|\chi| = 1$. In periodic cases, the ideal and symmetric coupling condition, $\omega = 1 = \chi$, leads to the most enhanced transmission.[54] Here, $t_{cL}$ and $t_{cR}$ are chosen

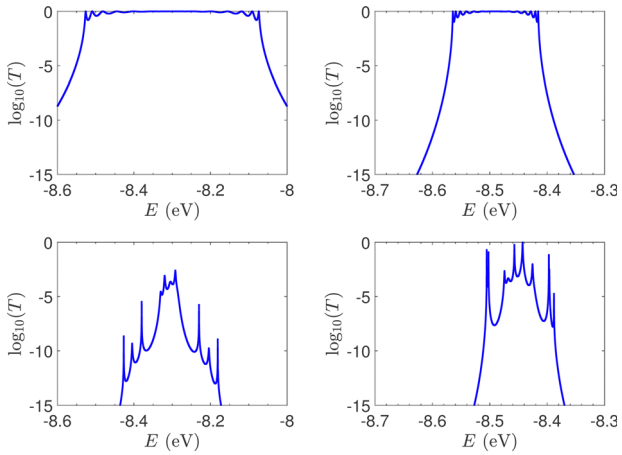


Fig. 7 The $\log_{10}(T(E))$ of the studied unmutated DNA sequences. Upper panels: Ideal polymers. There are $N-1$ peaks with full transmission (left) and $N$ peaks with full transmission (right). Theory[54] guarantees at least $N-1$ peaks. Lower panels: Natural polymers. Left: $G_{14}$, right: $A_{15}$.

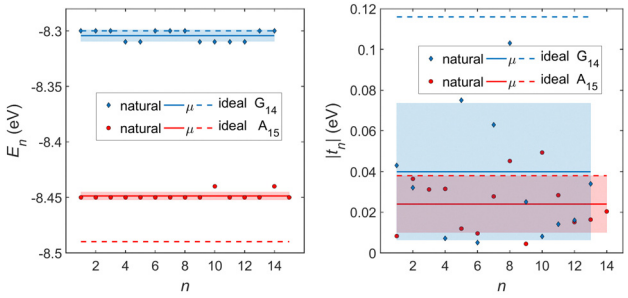


Fig. 8 TB parameters for the natural $G_{14}$ and $A_{15}$ polymers whose transmission is shown in the lower panels of Fig. 7. Left: On-site energies, $E_n$. Right: Absolute values of interaction parameters, $|t_n|$. Blue ($G_{14}$) diamonds and red ($A_{15}$) circles represent the values of the parameters at each site, continuous lines their mean values, $\mu$, and shaded areas include the region $\mu \pm \sigma$, where $\sigma$ is the standard deviation. The values of parameters for ideal polymers are shown in dashed lines, for reference.

from the ideal and symmetric coupling conditions of periodic cases of ideal homopolymers, G… and A…, *i.e.*, when dealing with G… or A…, $t_N$ is chosen as equal either to $t_{GG} = 0.116$ eV or to $t_{AA} = 0.038$ eV, according to our TB parametrization.[26] This procedure results in $t_{cL} = t_{cR} = 0.24$ eV for G… and 0.14 eV for A…. In natural homopolymers G…, the value $t_{cL} = t_{cR} = 0.24$ eV is still used. In natural homopolymers A…, the value $t_{cL} = t_{cR} = 0.14$ eV is still used. For A–C mismatches in G… as well as for diseases, the value $t_{cL} = t_{cR} = 0.24$ eV is still used.

The transmission coefficient at zero bias, $T(E)$, is a useful quantity for the description of charge transport properties; it refers to the probability that a carrier transmits through the sequence's eigenstates. To compute $T(E)$, a transfer matrix formalism[51,54,55] is used. After some manipulations, the analytical form of $T(E)$ can be expressed as:[51]

$$T(E) = \frac{4\sin^2(qa)}{\mathrm{Tr}(\tilde{M}_N)^2\sin^2(qa) + \left[\tilde{M}_N^{(12)} - \tilde{M}_N^{(21)} + (\tilde{M}_N^{(11)} - \tilde{M}_N^{(22)})\cos(qa)\right]^2}, \quad (6)$$

where

$$\tilde{M}_N = P_R M_N P_L, P_R = \begin{pmatrix} 1 & 0 \\ 0 & \frac{t_{cR}}{t_m} \end{pmatrix}, P_L = \begin{pmatrix} \frac{t_m}{t_{cL}} & 0 \\ 0 & 1 \end{pmatrix}, \quad (7)$$

Tr denotes the matrix trace and $M_N^{(ij)}$ are the elements of the GTM.

$T(E)$ for the studied ideal and natural unmutated DNA sequences, $G_{14}$ and $A_{15}$, are presented in Fig. 7. In ideal periodic segments, it is expected from theory[54] that full transmission ($T(E) = 1$) occurs at specific energies, at least $N-1$ in number. This is actually the case in the upper panels of Fig. 7 (not all peaks are seen clearly at this scale). Natural sequences have a significantly less symmetric profile, and significantly reduced overall transmission, as expected, since in this case neither the on-site energies nor the interaction integrals are identical, *i.e.*, in natural homopolymers, both diagonal and off-diagonal disorder are present.

In Fig. 8, the on-site energies (left) and absolute values of the interaction integrals (right) are depicted, together with their mean values, $\mu$, and standard deviations, $\sigma$, for the natural $G_{14}$ and $A_{15}$ sequences whose transmission is shown in the lower panels of Fig. 7. The corresponding values of ideal sequences are also shown, for reference. The mean values and standard deviations ($\mu$, $\sigma$) of the on-site energies, which account for diagonal disorder, are $\approx$(−8.304 eV, 0.005 eV) for $G_{14}$ and (−8.449 eV, 0.004 eV) for $A_{15}$, while, those of the magnitude of the interaction integrals, which account for off-diagonal disorder, are (0.040 eV, 0.034 eV) for $G_{14}$ (0.024 eV, 0.014 eV) for $A_{15}$. In terms of coefficients of variation, $\mathrm{CV} = \frac{|\mu|}{\sigma}$, diagonal disorder is small and of comparable magnitude between $G_{14}$ and $A_{15}$, *i.e.*, $\approx$0.06% and 0.05%, respectively. On the other hand, off-diagonal disorder is much larger, *i.e.*, $\approx$85.00% and 58.33%, respectively. Clearly, off-diagonal disorder is more pronounced in $G_{14}$. This explains qualitatively the smaller transmission peaks $G_{14}$ displays compared to $A_{15}$ (*cf.*, bottom panels of Fig. 7). Notice that $|t_n|$ was used to assess the off-diagonal disorder, since the spectrum of tridiagonal, irreducible, real, symmetric matrices (as all matrices studied here are, within the wire model) does not depend on the signs of their off-diagonal entries.[56]

In Fig. 9, the effect of including zero, one, and two A–C mismatches, randomly distributed in the sequence, is shown, for ideal and natural $G_{14}$ segments. Transmission coefficients, $T(E)$, are depicted in log-scale. The randomly positioned mismatches are placed at the same sites for both ideal and natural sequences. [$\log_{10}(T(E))$ for zero A–C mismatch mutations is also shown in the left panels of Fig. 7.] The values of $\int_{-\infty}^{+\infty} dE\ T(E)$ (which act as a measure of the overall transmission) for the three ideal cases are: 0.3856 eV, 0.0430 eV, and 0.0237 eV for 0, 1, and 2 (A–C) mismatches, respectively. Hence, in ideal cases,

 

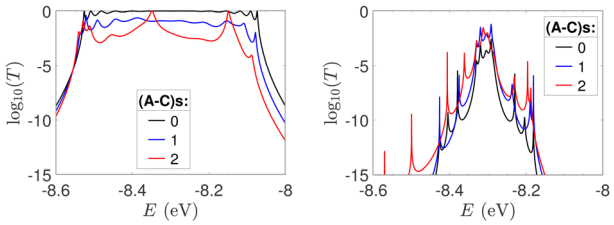


Fig. 9 $\log_{10}(T(E))$ of the studied $G_{14}$ sequences with zero, one, and two randomly positioned A–C mismatch mutations. Left: Ideal polymers, right: natural polymers. Mutations are placed at the same sites for both ideal and natural sequences.

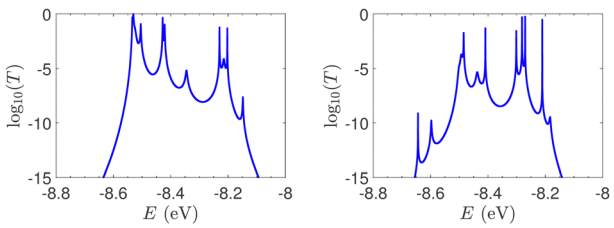


Fig. 10 $\log_{10}(T(E))$ of the studied $G_{14}$ sequences with seven randomly positioned A–C mismatch mutations. Left: ideal polymers, right: natural polymers. Mutations are placed at the same sites for both ideal and natural sequences.

inclusion of more mismatches decreases transmission, because the sequence homogeneity – in terms of on-site energies and interaction integrals – is deteriorated. On the other hand, for the three natural cases, the values of the integrals are $1.0651 \times 10^{-5}$ eV, $3.0247 \times 10^{-4}$ eV, and $2.2113 \times 10^{-4}$ eV, respectively. However, a natural sequence with no mismatches, is already inhomogeneous; there is no homogeneity to be lost by inserting (A–C) mismatches, since the sequences are already disordered. Therefore, it is difficult to characterize natural sequences based only upon $T(E)$.

$T(E)$ for ideal and natural $G_{14}$ sequences with seven randomly positioned A–C mismatch mutations are presented in Fig. 10. Fig. 10 can therefore be compared with the left column of Fig. 7. When seven mutations are included, *i.e.*, 50% of the total number of monomers, the polymer becomes a random binary sequence. The influence of the inclusion of A–C monomers can be observed; there are some lightly conducting states closer to $E_{A-C}$; *cf.*, Table 1. However, since $E_m$ is aligned with $E_{G-C}$, this effect is small.

$T(E)$ for the studied DNA sequences of ideal geometries with STR expansion mutations are presented in Fig. 11. As their DOS suggest (*cf.* Fig. 5), these sequences display narrow regions close to $E_{G-C}$ and $E_{A-T}$ within which transmission is allowed. The relative contribution of each region, as well as the overall transmission profile, are different for each sequence, allowing for distinct current–voltage curves, as it is shown below.

### 4.3 Current–voltage curves

The *I–V* curves of the DNA sequences have been calculated using the Landauer–Büttiker formalism.[57–59] A constant bias

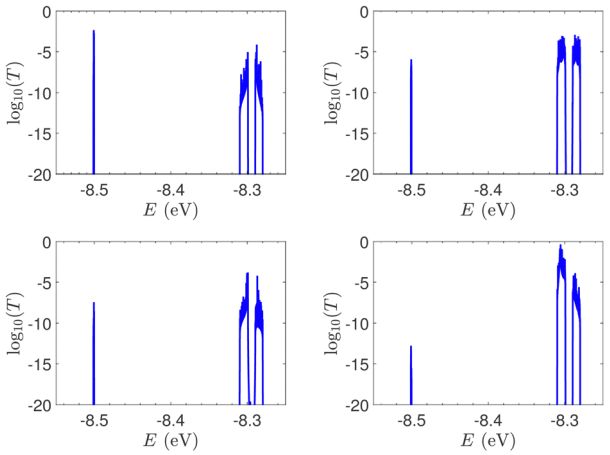


Fig. 11 The $\log_{10}(T(E))$ of the studied sequences (ideal geometries) with STR expansion mutations. Upper left: Huntington's disease with 100 STR expansions, upper right: Kennedy's disease with 45 STR expansions, left lower: Spinocerebellar ataxia 6 with 30 STR expansions, lower right: Spinocerebellar ataxia 7 with 100 STR expansions.

voltage was applied, $V_b$, between the leads, which induces a linear potential energy drop, $U_b = -eV_b$, from one to the other edge of the DNA sequence. Thus, the transmission coefficient becomes bias-dependent. The leads' chemical potential takes the form

$$\mu_{L(R)} = E_m \pm \frac{U_b}{2}. \quad (8)$$

The energy region between them defines the conductance channel. At zero temperature, the Fermi–Dirac distributions are Heaviside step functions, so the electrical current can be computed as

$$I(V) = \frac{2e}{h}\int_{\mu_R}^{\mu_L} dE\ T(E, U_b). \quad (9)$$

In Fig. 12, the absolute value of the current in logarithmic scale, $\log_{10}|I|$, is demonstrated as a function of both the leads on-site energy, $E_m$, and the applied voltage between the leads, $V$, for both ideal (left) and natural (right) $G_{14}$ polymers. It is evident that the electrode's on-site energy plays a crucial role in the shape and magnitude of the current–voltage curves. A general trend for homopolymers is that larger currents occur when $E_m$ is closer to the monomer's on-site energy. The *I–V*

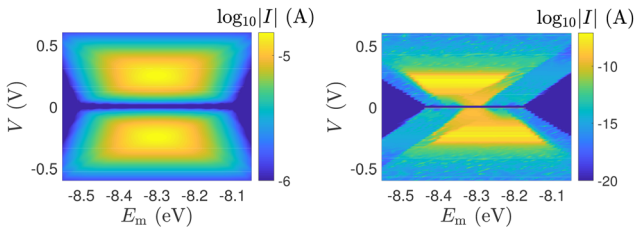


Fig. 12 $\log_{10}|I|$, *i.e.*, the absolute value of the current in logarithmic (colour) scale, as a function of both the leads on-site energy, $E_m$, and the applied voltage between the leads, $V$, for both ideal (left) and natural (right) $G_{14}$ polymers.

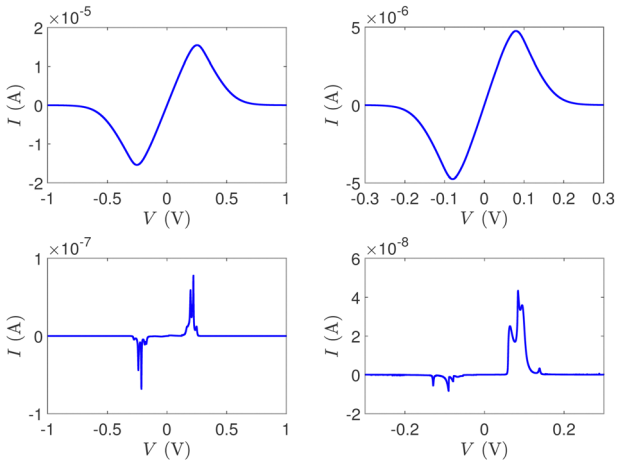


Fig. 13 The $I$–$V$ curves of the studied unmutated DNA sequences. Upper: Ideal polymers, lower: natural polymers, left: $G_{14}$ (with $E_m = E_{G-C}$), right: $A_{15}$ (with $E_m = E_{A-T}$).

curves of the studied ideal and natural unmutated DNA sequences are shown in Fig. 13, assuming $E_m = E_{G-C}$ for $G_{14}$ and $E_m = E_{A-T}$ for $A_{15}$. The left panels of Fig. 13 are a subset of Fig. 12, for $E_m = E_{G-C} = -8.3$ eV.

The order of magnitude of the $I$–$V$ curves and their shape varies dramatically when many mutations are included. Hence, another physical magnitude that could be used to characterise the $I$–$V$ curves was devised; this is the normalised deviation of the $I$–$V$ from the origin, defined as

$$\mathrm{NDIV}^{+} = \frac{\int_0^{\infty} \mathrm{d}V\, I(V)\, V}{\int_0^{\infty} \mathrm{d}V\, I(V)}, \tag{10}$$

for the positive $V$ regime and as

$$\mathrm{NDIV}^{-} = \frac{\int_{-\infty}^{0} \mathrm{d}V\, I(V)\, V}{\int_{-\infty}^{0} \mathrm{d}V\, I(V)}, \tag{11}$$

for the negative $V$ regime. Then, NDIV is defined as

$$\mathrm{NDIV} = \frac{|\mathrm{NDIV}^{+}| + |\mathrm{NDIV}^{-}|}{2}. \tag{12}$$

Fig. 14 and 15 display $I$–$V$ related diagrams of the studied ideal and natural $G_{14}$ DNA sequences, respectively, with one A–C mismatch mutation of varying position (left columns) and with varying number of randomly distributed A–C mismatch mutations (right columns). The rows contain the $I$–$V$ curves, the $\log_{10}|I| - V$ curves (*i.e.*, in logarithmic $|I|$ scale), and the newly introduced quantity, *i.e.*, the normalised deviation of the $I$–$V$ from the origin, NDIV. It can be seen that, for ideal segments, generally the $I$–$V$ curves do not vary significantly with the position of a single A–C mismatch in the sequence (≈half an order of magnitude); for natural segments, the position of the mismatch affects the current more significantly (some orders of magnitude). The variation of the $I$–$V$ curves becomes much more significant with increasing the number of A–C mismatches (many orders of magnitude). As a particular example, the $I$–$V$ curves of the studied ideal (left) and natural (right) DNA

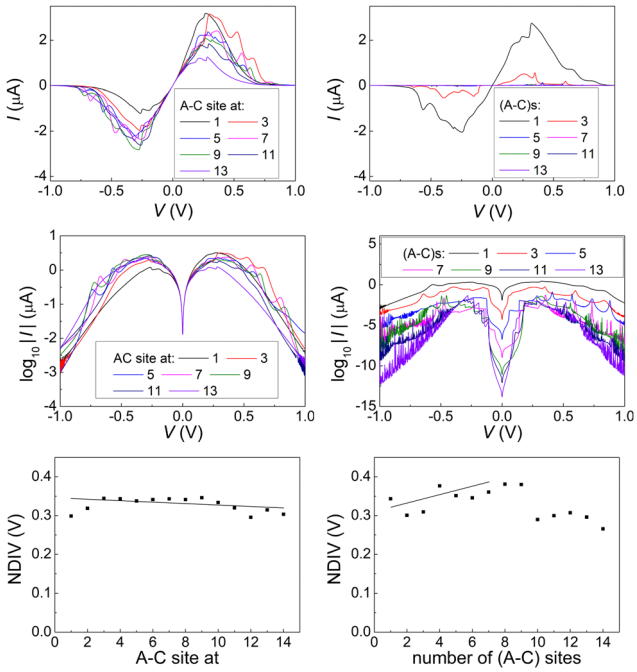


Fig. 14 The $I$–$V$ related diagrams of the studied $G_{14}$ DNA sequences of ideal geometry with one A–C mismatch mutation of varying position in the sequence (left panels) and varying number of A–C mismatch mutations randomly inserted in the sequence (right panels). First row: $I$–$V$ curves, second row: $\log_{10}|I|$–$V$ curves, *i.e.*, in logarithmic scale, third row: NDIV, *i.e.*, normalised deviation of the $I$–$V$ curve from the origin.

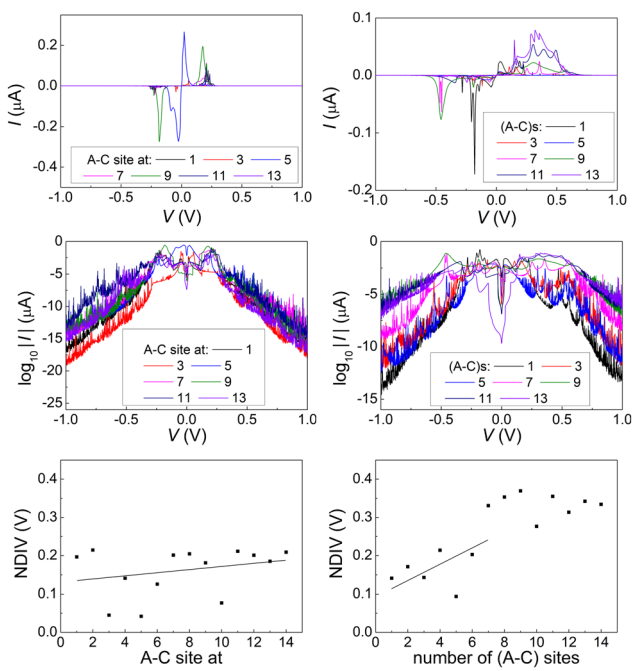


Fig. 15 The $I$–$V$ related diagrams of the studied $G_{14}$ DNA sequences of natural geometry with one A–C mismatch mutation of varying position in the sequence (left panels) and varying number of A–C mismatch mutations randomly inserted in the sequence (right panels). First row: $I$–$V$ curves, second row: $\log_{10}|I|$–$V$ curves, *i.e.*, in logarithmic scale, third row: NDIV, *i.e.*, normalised deviation of the $I$–$V$ curve from the origin.

 

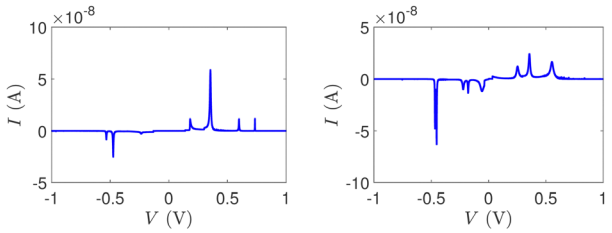


Fig. 16 The $I$–$V$ curves of the studied DNA sequences, initially $G_{14}$, but with 7 A–C mismatch mutations, randomly inserted in the sequence. Left: Ideal polymers, right: natural polymers.

sequences with 7 randomly positioned A–C mismatch mutations, are presented in Fig. 16.

In ideal sequences with one A–C mismatch of varying position, the NDIV remains almost constant; its slope *versus* the A–C site position is close to zero. Hence, the NDIV is insensitive to the position of a point substitution mutation. However, in ideal sequences with increasing number of A–C mismatch mutations, the NDIV does not remain constant; its slope *versus* the number of A–C mismatch mutations is positive, until the number of (A–C)s becomes equal to the number of (G–C)s. After that point, the number of (A–C)s becomes larger than the number of (G–C)s, *i.e.*, mutations become dominant, and a further increase of the number of (A–C)s stabilises the NDIV. Hence, the NDIV is sensitive to the increase of the number of point substitution mutations. In natural sequences, the situation is similar, but with pronounced slopes, especially when an increasing number of A–C mismatch mutations is introduced. Therefore, the NDIV is a useful quantity to characterise these sequences.

In Fig. 17 and 18, the $I$–$V$ related diagrams of the studied DNA sequences (ideal geometry) with STR expansion mutations are presented. For all studied cases, changes in the $I$–$V$ curves become more pronounced with increasing the number of STR expansion mutations (*i.e.*, the number of CAG repeats). The respective NDIV display significant but almost monotonous

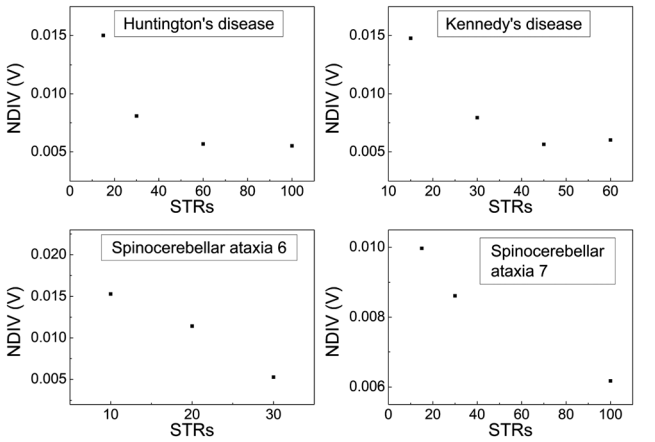


Fig. 18 Normalised deviation of the $I$–$V$ curve from the origin (NDIV) as a function of the number of (CAG) repeats aka short tandem repeat (STR) expansions. Upper left panel: Huntington's disease, upper right panel: Kennedy's disease, left lower panel: Spinocerebellar ataxia 6, lower right panel: Spinocerebellar ataxia 7.

variations, and can, therefore, be used to evaluate the number of (CAG) repeats in a sequence. This behaviour of the NDIV *versus* the number of CAG repeats suggests that it can be used to characterise the grade of danger for developing the studied diseases.

## 5 Conclusion and perspectives

The present work focused on to the effect of two types of mutations [(i) point substitution transitions, particularly guanine ↔ adenine, and (ii) cytosine–adenine–guanine triplet expansions], on the charge transport properties of DNA sequences. The following physical quantities were studied: eigenspectra, density of states, transmission coefficients, current–voltage curves. Both ideal (textbook geometry) and natural (naturally distorted, geometry from databases) sequences were

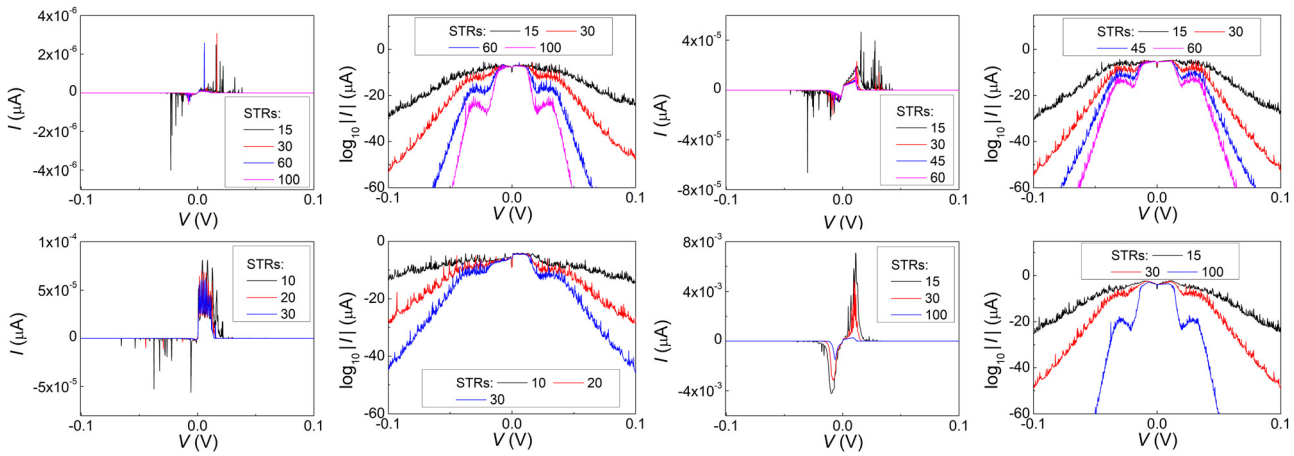


Fig. 17 The $I$–$V$ related diagrams ($I$–$V$ curves and $\log_{10}|I|$–($V$) curves, *i.e.*, in logarithmic scale) of the studied DNA sequences (ideal geometry) with varying STR expansions, *i.e.*, with different number of (CAG) triplets. Upper left two panels: Huntington's disease, upper right two panels: Kennedy's disease, left lower two panels: Spinocerebellar ataxia 6, lower two right panels: Spinocerebellar ataxia 7.

considered. A tight-binding wire model was recruited for charge transport, in conjunction with a transfer matrix technique. However, the on-site energies and interaction parameters for that TB wire model were obtained by another TB LCAO utilizing all valence orbitals, capable of treating any given geometry. Our results point to the following conclusions:

(1) All the aforementioned physical quantities possess interesting features that allow to distinguish between mutated and unmutated sequences.

(2) However, the most experimentally relevant quantities are the *I–V* curves. Their characteristics are significantly altered when mutations are introduced, and conclusions cannot be drawn in a straightforward manner.

(3) Since both the order of magnitude and the shape of the *I–V* curves varies when mutations are introduced, another physical quantity to characterise the *I–V* curves was introduced. *i.e.*, the normalised deviation of the *I–V* from the origin (NDIV).

(4) In ideal sequences with one A–C mismatch of varying position, the NDIV remains almost constant: its slope *versus* the mismatch position is close to zero.

(5) In ideal sequences with increasing number of A–C mismatch mutations, the NDIV does not remain constant: its slope *versus* the number of A–C mismatch mutations is positive, until the number of (A–C)s becomes equal to the number of (G–C)s. After that point, since the number of (A–C)s becomes larger than the number of (G–C)s, a further increase of the number of (A–C)s stabilises the situation, as expected.

(6) In natural sequences, the NDIV is similar but with pronounced slopes, especially when an increasing number of A–C mismatch mutations is introduced. Hence, NDIV is a useful quantity to characterise these sequences.

(7) Although dramatic changes in the *I–V* curves occur for all studied cases of STR expansions, as the number of CAG repeats increases, NDIV shows significant but almost monotonous variation, and can, therefore, be used to evaluate the number of (CAG) repeats in the sequence. Therefore, the NDIV can be used to characterise the grade of danger for developing the studied diseases.

(8) Overall, the NDIV is generally insensitive to the position of a point mutation, but rather sensitive to the number of point mutations and STR expansion mutations.

The recruitment of specific natural geometries in this work for type (i) sequences does not imply their association with the studied diseases *i.e.* with type (ii) sequences. These geometries were used to demonstrate the expected differences between the electrical behaviour of ideal and distorted conformations. However, it should be noted that the TB protocol with all valence orbitals used here[26] can be easily employed to sequences of arbitrary geometry.

Transitions with C ↔ T exchange have not been included in this work, because their effects could not be properly grasped within the wire model. These mutations could be studied within the extended ladder model,[48] *i.e.* a TB description at the single-base level; this will be hopefully done in the future.

Transitions are more likely than transversions (purine ↔ pyrimidine interchange), because it is easier to substitute a single ring by another single ring than a double ring for a single ring or *vice versa*. Hence, careful geometry optimization is necessary to study transversions. This will also be hopefully done in the future.

Another category of mutations is germline mutations,[60] *i.e.* a gene change in a reproductive cell that becomes incorporated into the DNA of every cell in the body of the offspring. This way, the mutation can be passed from parent to offspring, and is, therefore, hereditary. This could be also the subject of a future study.

The vibrational and large-scale dynamical flexibility of DNA has not been taken into account in this work. However, the presented methodology could be applied in a straightforward manner on top of snapshots extracted by Molecular Dynamics simulations.[16] This could be another future perspective.

All diseases studied here have the same triplet motif, *i.e.*, $(CAG)_n$, between 9 base pairs at the start and 9 base pairs at the other end of the sequence (primers). Of course, primers are different for each disease, but they only contain 18 base pairs altogether, which is not a large number when dealing with a sequence of 180 or 300 base pairs. Under these conditions, it is not safe to draw direct conclusions regarding the identity of the disease. This issue could be possibly tackled by including larger primers, which would produce much more distinctive DOS or IDOS features and allow for sequence recognition. However, this exceeds the scope of the present study, which has the aim to demonstrate the possibility of mutation detection using the TB model. The treatment discussed above could be included in a future work.

Another perspective would be to consider, *e.g.*, a 300 base-pair sequence and change both the number of repetitions $n$ and the length of primers, while keeping the number of total base pairs (*e.g.*, 300) constant. This is a different perspective that will hopefully be included in a future work. Such an alternative perspective has been previously adopted to study the Huntington's disease within a TB model, but off-diagonal disorder (the fact that interaction integrals are not identical) was not taken into account.[61]

Finally, a general comment: the interplay between periodicity and aperiodicity in biology[62] is a vast area of extreme interest to us; novel methods must be devised to explore it.

## Author contributions

Conceptualization, C. S.; methodology, C. S., K. L., M. M.; software, M. M., K. L.; validation, C. S., K. L., M. M.; formal analysis, M. M., K. L.; investigation, C. S., M. M., K. L.; resources, C. S., M. M., K. L.; data curation, M. M.; writing – original draft preparation, M. M.; writing – review and editing, C. S., K. L., M. M.; visualization, M. M., K. L.; supervision, C. S.; project administration, C. S. All authors have read and agreed to the published version of the manuscript.

## Conflicts of interest

There are no conflicts to declare.

 

## Acknowledgements

We thank Dr Georgia Thodi of Laboratory of Prenatal and Neonatal Screening, Neoscreen Ltd, Athens, Greece, for providing information about the sequences utilized in this work. M. Mantela wishes to thank the State Scholarships Foundation (IKY): This research is co-financed by Greece and the European Union (European Social Fund-ESF) through the Operational Programme "Human Resources Development, Education and Lifelong Learning" in the context of the project "Strengthening Human Resources Research Potential *via* Doctorate Research" (MIS-5000432), implemented by the State Scholarships Foundation (IKY).

## References

1 Z. Yang and A. D. Yoder, *J. Mol. Evol.*, 1999, **48**, 274–283.
2 I. Keller, D. Bensasson and R. A. Nichols, *PLoS Genet.*, 2007, **3**, 1–7.
3 A. Stoltzfus and R. W. Norris, *Mol. Biol. Evol.*, 2015, **33**, 595–602.
4 D. M. Lyons and A. S. Lauring, *Mol. Biol. Evol.*, 2017, **34**, 3205–3215.
5 D. Porath, A. Bezryadin, S. de Vries and C. Dekker, *Nature*, 2000, **403**, 635–638.
6 J. C. Genereux and J. K. Barton, *Chem. Rev.*, 2010, **110**, 1642–1662.
7 K. L. Jiménez Monroy, N. Renaud, J. Drijkoningen, D. Cortens, K. Schouteden, C. van Haesendonck, W. J. Guedens, J. V. Manca, L. D. A. Siebbeles, F. C. Grozema and P. H. Wagner, *J. Phys. Chem. A*, 2017, **121**, 1182–1188.
8 D. Rawtani, B. Kuntmal and Y. Agrawal, *Front. Life Sci.*, 2016, **9**, 214–225.
9 B. Giese, J. Amaudrut, A.-K. Köhler, M. Spormann and S. Wessely, *Nature*, 2001, **412**, 318–320.
10 A. Landi, A. Capobianco and A. Peluso, *J. Phys. Chem. Lett.*, 2020, **11**, 7769–7775.
11 E. Maciá, F. Triozon and S. Roche, *Phys. Rev. B*, 2005, **71**, 113106.
12 M. Bixon, B. Giese, S. Wessely, T. Langenbacher, M. E. Michel-Beyerle and J. Jortner, *Proc. Natl. Acad. Sci. U. S. A.*, 1999, **96**, 11713–11716.
13 L. Xiang, J. L. Palma, C. Bruot, V. Mujica, M. A. Ratner and N. Tao, *Nat. Chem.*, 2015, **7**, 221–226.
14 A. Capobianco, T. Caruso, A. M. D'Ursi, S. Fusco, A. Masi, M. Scrima, C. Chatgilialoglu and A. Peluso, *J. Phys. Chem. B*, 2015, **119**, 5462–5466.
15 A. Capobianco, A. Velardo and A. Peluso, *J. Phys. Chem. B*, 2018, **122**, 7978–7989.
16 M. Mantela, A. Morphis, K. Lambropoulos, C. Simserides and R. Di Felice, *J. Phys. Chem. B*, 2021, **125**, 3986–4003.
17 T. J. Zwang, E. C. M. Tse and J. K. Barton, *ACS Chem. Biol.*, 2018, **13**, 1799–1809.
18 F. D. Lewis and M. R. Wasielewski, *Pure Appl. Chem.*, 2013, **85**, 1379–1387.
19 C. H. Wohlgamuth, M. A. McWilliams and J. D. Slinker, *Anal. Chem.*, 2013, **85**, 8634–8640.
20 R. G. Endres, D. L. Cox and R. R. P. Singh, *Rev. Mod. Phys.*, 2004, **76**, 195–214.
21 S. R. Rajski, B. A. Jackson and J. K. Barton, *Mutat. Res.*, 2000, **447**, 49–72.
22 C.-T. Shih, Y.-Y. Cheng, S. A. Wells, C.-L. Hsu and R. A. Römer, *Comput. Phys. Commun.*, 2011, **182**, 36–38.
23 Y. Liu, X. Ren and L. He, *J. Chem. Phys.*, 2019, **151**, 215102.
24 A. Bende, F. Bogár and J. Ladik, *Solid State Commun.*, 2011, **151**, 301–305.
25 F. Bogár, A. Bende and J. Ladik, *Phys. Lett. A*, 2014, **378**, 2157–2162.
26 M. Mantela, C. Simserides and R. Di Felice, *Materials*, 2021, **14**, 4930.
27 J. C. Slater and G. F. Koster, *Phys. Rev.*, 1954, **94**, 1498–1524.
28 W. A. Harrison, *Electronic Structure and the Properties of Solids: the Physics of the Chemical Bond*, Dover, New York, 2nd edn, 1989.
29 W. A. Harrison, *Elementary Electronic Structure*, World Scientific, River Edge, NJ, 1999.
30 M. Mantela, A. Morphis, M. Tassi and C. Simserides, *Mol. Phys.*, 2016, **114**, 709–718.
31 A. Bebenek and I. Ziuzia-Graczyk, *Curr. Genet.*, 2018, **64**, 985–996.
32 T. A. Kunkel and K. Bebenek, *Annu. Rev. Biochem.*, 2000, **69**, 497–529.
33 A. K. Showalter and M.-D. Tsai, *Biochemistry*, 2002, **41**, 10571–10576.
34 H. Dashnow, M. Lek, B. Phipson, A. Halman, S. Sadedin, A. Lonsdale, M. Davis, P. Lamont, J. S. Clayton, N. G. Laing, D. G. MacArthur and A. Oshlack, *Genome Biol.*, 2018, **19**, 121.
35 S. R. Chintalaphani, S. S. Pineda, I. W. Deveson and K. R. Kumar, *Acta Neuropathol. Commun.*, 2021, **9**, 98.
36 I. H. Kratter and S. Finkbeiner, *Neuron*, 2010, **67**, 897–899.
37 M. E. MacDonald, C. M. Ambrose, M. P. Duyao, R. H. Myers, C. Lin, L. Srinidhi, G. Barnes, S. A. Taylor, M. James, N. Groot, H. MacFarlane, B. Jenkins, M. A. Anderson, N. S. Wexler, J. F. Gusella, G. P. Bates, S. Baxendale, H. Hummerich, S. Kirby, M. North, S. Youngman, R. Mott, G. Zehetner, Z. Sedlacek, A. Poustka, A.-M. Frischauf, H. Lehrach, A. J. Buckler, D. Church, L. Doucette-Stamm, M. C. O'Donovan, L. Riba-Ramirez, M. Shah, V. P. Stanton, S. A. Strobel, K. M. Draths, J. L. Wales, P. Dervan, D. E. Housman, M. Altherr, R. Shiang, L. Thompson, T. Fielder, J. J. Wasmuth, D. Tagle, J. Valdes, L. Elmer, M. Allard, L. Castilla, M. Swaroop, K. Blanchard, F. S. Collins, R. Snell, T. Holloway, K. Gillespie, N. Datson, D. Shaw and P. S. Harper, *Cell*, 1993, **72**, 971–983.
38 P. McColgan and S. J. Tabrizi, *Eur. J. Neurol.*, 2018, **25**, 24–34.
39 P. Fratta, T. Collins, S. Pemble, S. Nethisinghe, A. Devoy, P. Giunti, M. G. Sweeney, M. G. Hanna and E. M. Fisher, *Neurobiol. Aging*, 2014, **35**, 443.e1–443.e3.
40 G. Kuhlenbäumer, W. Kress, E. B. Ringelstein and F. Stögbauer, *J. Neurol.*, 2001, **248**, 23–26.
41 A. R. L. Spada, E. M. Wilson, D. B. Lubahn, A. E. Harding and K. H. Fischbeck, *Nature*, 1991, **352**, 77–79.
42 J. Sequeiros, S. Seneca and J. Martindale, *Eur. J. Hum. Genet.*, 2010, **18**, 1188–1195.

43 O. Zhuchenko, J. Bailey, P. Bonnen, T. Ashizawa, D. W. Stockton, C. Amos, W. B. Dobyns, S. H. Subramony, H. Y. Zoghbi and C. C. Lee, *Nat. Genet.*, 1997, **15**, 62–69.
44 C. Cagnoli, G. Stevanin, C. Michielotto, G. Gerbino Promis, A. Brussino, P. Pappi, A. Durr, E. Dragone, M. Viemont, C. Gellera, A. Brice, N. Migone and A. Brusco, *J. Mol. Diagn.*, 2006, **8**, 128–132.
45 G. David, N. Abbas, G. Stevanin, A. Dürr, G. Yvert, G. Cancel, C. Weber, G. Imbert, F. Saudou, E. Antoniou, H. Drabkin, R. Gemmill, P. Giunti, A. Benomar, N. Wood, M. Ruberg, Y. Agid, J.-L. Mandel and A. Brice, *Nat. Genet.*, 1997, **17**, 65–70.
46 C. Qiu, K. L. McCann, R. N. Wine, S. J. Baserga and T. M. T. Hall, *Proc. Natl. Acad. Sci. U. S. A.*, 2014, **111**, 18554–18559.
47 R. Wang, S. Wang, A. Dhar, C. Peralta and N. P. Pavletich, *Nature*, 2020, **580**, 278–282.
48 K. Lambropoulos, M. Chatzieleftheriou, A. Morphis, K. Kaklamanis, R. Lopp, M. Theodorakou, M. Tassi and C. Simserides, *Phys. Rev. E*, 2016, **94**, 062403.
49 K. Lambropoulos, M. Chatzieleftheriou, A. Morphis, K. Kaklamanis, M. Theodorakou and C. Simserides, *Phys. Rev. E*, 2015, **92**, 032725.
50 N. H. G. R. Institute, *Primers*, **https://www.genome.gov/genetics-glossary/Primer**, A primer, as related to genomics, is a short single-stranded DNA fragment used in certain laboratory techniques, such as the polymerase chain reaction (PCR). In the PCR method, a pair of primers hybridizes with the sample DNA and defines the region that will be amplified, resulting in millions and millions of copies in a very short timeframe, Primers are also used in DNA sequencing and other experimental processes.
51 K. Lambropoulos and C. Simserides, *Phys. Rev. E*, 2019, **99**, 032415.
52 D. Porath, G. Cuniberti and R. Di Felice, in *Charge Transport in DNA-Based Devices*, ed. G. Schuster, Springer Berlin Heidelberg, Berlin, Heidelberg, 2004, pp. 183–228.
53 E. Macia Barber, *Aperiodic Structures in Condensed Matter: Fundamentals and Applications*, CRC Press, 1st edn, 2008.
54 K. Lambropoulos and C. Simserides, *J. Phys. Commun.*, 2018, **2**, 035013.
55 K. Lambropoulos and C. Simserides, *Phys. Chem. Chem. Phys.*, 2017, **19**, 26890–26897.
56 M. Mantela, K. Lambropoulos, M. Theodorakou and C. Simserides, *Materials*, 2019, **12**, 2177.
57 R. Landauer, *IBM J. Res. Dev.*, 1957, **1**, 223–231.
58 M. Buttiker, *IBM J. Res. Dev.*, 1988, **32**, 317–334.
59 S. Datta, *Electronic Transport in Mesoscopic Systems*, Cambridge University Press, 1995.
60 G. Thodi, F. Fostira, R. Sandaltzopoulos, G. Nasioulas, A. Grivas, I. Boukovinas, M. Mylonaki, C. Panopoulos, M. B. Magic, G. Fountzilas and D. Yannoukakos, *BMC Cancer*, 2010, **10**, 544.
61 R. G. Sarmento, R. N. O. Silva, M. P. Madeira, N. F. Frazao, J. O. Sousa and A. Macedo-Filho, *Braz. J. Phys.*, 2018, **48**, 155–159.
62 E. Maciá, *J. Phys.: Condens. Matter*, 2022, **34**, 123001.

 