Technical University of Denmark

DTU

Classification of DNA nucleotides with transverse tunneling currents

Pedersen, Jonas Nyvold; Boynton, Paul; Ventra, Massimiliano Di; Jauho, Antti-Pekka; Flyvbjerg, Henrik

Link back to DTU Orbit

DTU Library
Technical Information Center of Denmark

Classification of DNA nucleotides with transverse tunneling currents

Jonas Nyvold Pedersen[1,2]**, Paul Boynton**[3]**, Massimiliano Di Ventra**[3]**, Antti-Pekka Jauho**[1,2] **and Henrik Flyvbjerg**[1]

[1] Department of Micro- and Nanotechnology, Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark,
[2] Center for Nanostructured Graphene (CNG), DTU Nanotech, Department of Micro- and Nanotechnology, Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark.
[3] Department of Physics, University of California, San Diego, La Jolla, CA 92093-0319, USA

E-mail: `jonas.pedersen@nanotech.dtu.dk`, `henrik.flyvbjerg@nanotech.dtu.dk`

Abstract.
It has been theoretically suggested and experimentally demonstrated that fast and low-cost sequencing of DNA, RNA, and peptide molecules might be achieved by passing such molecules between electrodes embedded in a nanochannel. The experimental realization of this scheme faces major challenges, however. In realistic liquid environments, typical currents in tunnelling devices are of the order of picoamps. This corresponds to only six electrons per microsecond, and this number affects the integration time required to do current measurements in real experiments. This limits the speed of sequencing, though current fluctuations due to Brownian motion of the molecule average out during the required integration time. Moreover, data acquisition equipment introduces noise, and electronic filters create correlations in time-series data. We discuss how these effects must be included in the analysis of, e.g., the assignment of specific nucleobases to current signals. As the signals from different molecules overlap, unambiguous classification is impossible with a single measurement. We argue that the assignment of molecules to a signal is a standard pattern classification problem and calculation of the error rates is straightforward. The ideas presented here can be extended to other sequencing approaches of current interest.

1. Introduction

Identification and sequencing of single DNA, RNA, and peptide molecules is a key step in many diagnostic protocols. Electronic sequencing of nucleobases and nucleic acids with nanopores or nanogaps has received growing interest as an alternative to optical methods
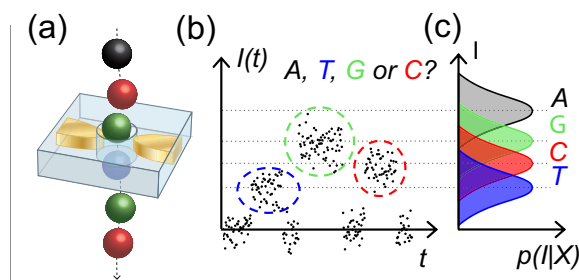
Figure 1. (a) Schematic of DNA passing through a nanopore with embedded electrodes forming a nanogap. (b) Electrons tunnel between the electrodes via the nucleotide in the gap and produce a nucleotide-specific current I versus time t. (c) Here $p(I|X)$ is the probability density for measuring the current value I, given that the nucleotide is X, where X is one of the four bases A, T, G, and C. Current signals from different nucleotides overlap, which prevents unambiguous classification [with a single current measurement].

in the last two decades [1–4]. Nanopore sequencing, as originally conceived, records the ionic current through a nanopore that is partially blocked by a nucleotide and attempts to identify that nucleotide from its degree of blocking. However, due to the thickness of the nanochannels employed and the longitudinal direction of the ionic current probe, single-base resolution is difficult to achieve with this approach [1, 2]. For this reason, a complementary concept ("quantum sequencing" [5]) has been suggested, based on the specific molecular fingerprints in the transverse tunneling current that passes through the nucleotide when the latter passes between two electrodes in a nanochannel [5–7], see figure 1(a).

With a break-junction as the electrode pair, single nucleotides have been identified experimentally by their respective transverse tunnelling currents [8, 9]. In addition, quantum sequencing has been used for identification of methylated DNA bases [10], for detection of post-translational modifications in single peptides [11], and for single-molecule spectroscopy of individual amino acids and peptides [12].

Current signals from single nucleotides have also been measured with a scanning-tunneling microscope (STM) [13]. With a functionalized STM tip, the individual nucleotides in a DNA oligomer have been read [14]. Nucleotides have also been identified with a fixed-gap device [15], and DNA molecules have been detected with nanowire-nanopore field-effect transistor sensors [16]. In all cases, the current signal was noisy and with step-like features, and a statistical analysis was required to get the actual sequence information, to determine the type of nucleotide, or just to detect a translocation event [17, 18].

In addition to these experimental efforts, simulations were found useful for testing alternative realizations of electronic nucleotide identification and nanopore sequencing [6, 7, 19–28]. One such alternative, e.g., measures changes in the current in a graphene nanoribbon while a DNA string passes through a hole in the ribbon [22, 27].

The experimental relevance of these simulations depends on the magnitude of

the currents that can be measured experimentally—specifically, it depends on the integration time (bandwidth) required to obtain a signal that stands out well enough over noise and filtering effects to distinguish between different nucleotides. This is a critical issue for any type of sequencing protocol that employs either transverse tunnelling or longitudinal ionic currents.

The present article discusses the subtleties related to the connection between theoretical ideas and simulations with actual experiments. In section 2, we describe how the transverse current through individual nucleotides is simulated. Then we discuss the magnitude of the average current, the amplitude of current fluctuations, and the correlation time of current fluctuations. The correlation time is, interestingly, even shorter than the average waiting time between electrons tunneling through the nucleotide.

Tunneling currents are typically very small so that long integration times are needed to measure them in actual experiments. The reason is charge quantization: A current of $1\,\text{pA}$ amounts to six electrons per microsecond, on average. Consequently, narrowly defined current values can be measured only with integration times much longer than microseconds. This limits the time resolution of current measurements, which can be ameliorated by multiplexing with several pairs of electrodes [29].

As a result, the integration time of data acquisition in a realistic experiment is long enough that current fluctuations due to thermal motion of nucleotides average out in a realistic recorded signal (section 3). Electronic noise, however, broadens the distribution of currents recorded for a given nucleotide, so current distributions for different nucleotides overlap (see figure 1(b,c) and section 4).

Electronic filters in the data acquisition system also affect the distribution of recorded currents and autocorrelate the time series of recorded currents (section 5). We show in section 6 how to assign a nucleotide to a current signal and that the autocorrelations play an important role in the assignment. Finally, in section 7 we compare the error rates of nucleotide assignment for simulated data with and without autocorrelations.

Throughout this article we consider only simulations of the transverse tunneling current through the four nucleotides A, T, G, and C. The analysis presented here is nevertheless also valid for other types of sensors that produce weak, overlapping current signals.

2. Magnitude and correlations of simulated current values

Nanopore experiments take place in a liquid environment at ambient temperature [5]. These conditions make simulations of the current through a single nucleotide both time consuming and computationally expensive [7] as they do not only involve the nucleotide of interest, but also the degrees of freedom of the surrounding molecules of the liquid. In previous work by one of us (MDV), the following protocol was used for simulating the transverse current through a single nucleotide as it passes through a nanopore [7,20]: The

molecule is driven by a driving field into the nanopore where the electrodes are placed. Then the driving field is reduced and the transverse field is turned on. The molecule moves due to the electric fields and the thermal motion caused by interactions with the surrounding water molecules. This motion is described by molecular dynamics (MD) simulations with a time resolution of 1 fs. The femtosecond timescale is also the timescale for a typical electron transport time through the trapped molecule. Each picosecond the motion is frozen and a tight-binding Hamiltonian is set up which describes the coupling between the electrodes, the liquid and the DNA molecule. The steady-state current is calculated using a single-particle scattering approach with an applied bias of less than 1 V. Then the molecule is released for another time interval of one picosecond and the procedure is repeated many times (on the order of 4000 to 5000 times).

Figure 2 shows an example of a current trace for the nucleotide A, and histograms of the current values for all four nucleotides are shown in figure 3 as obtained in reference [29]. We here plot the log-current probability distributions $p(\tilde{I}|X)$ with $\tilde{I} = \log_{10}(I/\text{Amp})$, and where $X \in \{A, T, G, C\}$ denotes the four types of nucleotides. That is, the probability distributions for the current I is $p(I|X) = (d\tilde{I}/dI)p(\tilde{I}|X) = p(\tilde{I}|X)/(I \ln 10)$.

Notice that the current distributions span six orders of magnitude; from 10^{-15} Amp to 10^{-9} Amp (see figure 3). Table 1 shows the corresponding expected values μ_X and standard deviations σ_X for the current probability density distributions of figure 3. In experiments with mechanically controlled break-junctions, the transverse current signal from individual nucleotides was in the range ~1–100 pA [8] and thus comparable to the expected values of the simulated currents, such as those shown in figure 2‡.

In the simulations, the contacts to the nucloetides are modelled as gold electrodes [7,20]. Due to the presence of water, the tunneling barrier is considerably reduced: to about 1 eV from the gold work function of about 4.5 eV. Other electrodes, such as Pt, can be (and are currently) used in experiments without much qualitative change in the distributions. For a detailed discussion of the current calculations, see references [6,7,20].

We next take advantage of the simulation times up to 1500 ps in the simulations of individual nucleotides in the nanopore. Although it is not possible to reach the experimentally relevant sampling times, which are of the order of micro- or miliseconds (see below), we can extract the relevant time scales without approximate solutions for times longer than picoseconds [30].

Current values calculated at different time points are not independent, and the correlations in the signal are quantified with the autocovariance $R_X^{\text{curr}}(k, \ell) \equiv \langle (I_k^X - \mu_X)(I_\ell^X - \mu_X) \rangle$, where I_k is the simulated current at the time point $t_k = k\Delta t_{\text{curr}}$ with $\Delta t_{\text{curr}} = 1$ ps §. Figure 4 shows the autocovariance for the nucleotide A. The

‡ In typical measurements of the ionic current through a nanopore, the current is in the range of hundreds of nA.
§ See SI for how to calculate the autocovariance from data.
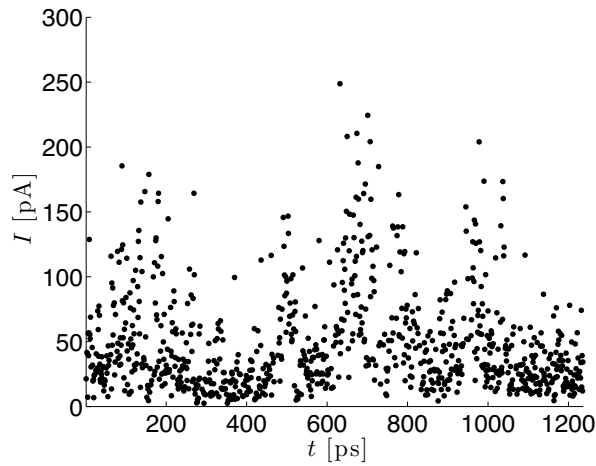
Figure 2. Current value as function of time for the nucleotide A. Note the large range of values. The current through the nucleotide is calculated each picosecond, but some data points are missing due to lack of convergence in the calculation (see SI for details).
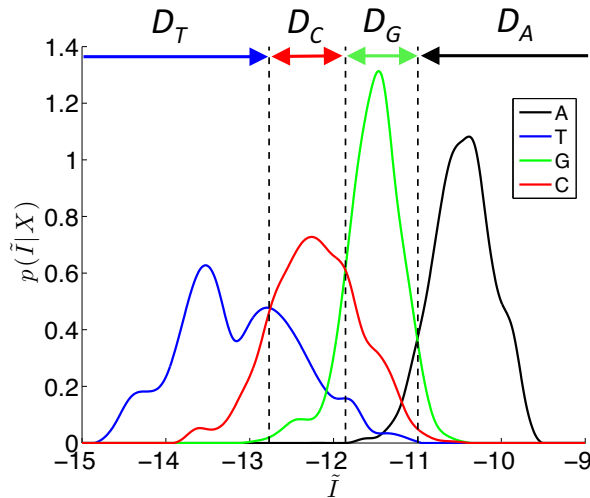
Figure 3. Histograms of the probability distributions $p(\tilde{I}|X)$ for the log-current values $\tilde{I} = \log_{10}(I/\text{Amp})$ for the four different nucleotides (same as figure 2 in reference [29]). Dashed vertical lines mark points on the current axis where one distribution replaces another at being the one with the highest probability density. The colored arrows show the ranges, D_X ($X \in \{A, T, G, C\}$), of current values in which nucleotide X is indicated by a single measured current value ($m = 1$, where m is the number of current measurements).

Figure 4. Autocovariance of the current values shown in figure 2. The black curve shows the exponential decrease for time lags τ larger than 1 ps. Its characteristic time is $\tau_A = 44$ ps. Notice that the black curve is not a fit to the data shown, because these data values are autocorrelated. Instead, the parameter τ_A of the exponential autocorrelation function was determined by fitting the Fourier transform of the autocorrelation function to the power spectrum of the data shown here (Wiener-Khinchin theorem; see SI for details).
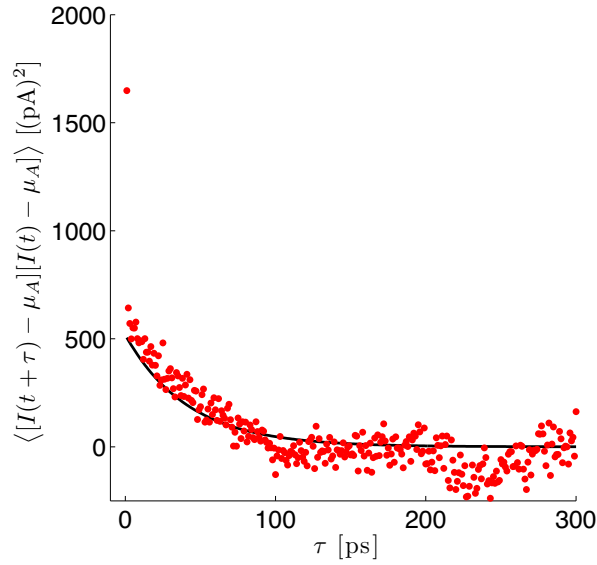
autocovariance is consistent with a process with two time-scales‖,

$$
\begin{aligned}
R_X^{\mathrm{curr}}(k,\ell) = \sigma_X^2 \big(w_{0,X} \delta_{k,\ell} \\
+ \big[1 - w_{0,X}\big] e^{-|k-\ell|\Delta t_{\mathrm{curr}}/\tau_X} \big),
\end{aligned} \tag{1}
$$

where σ_X^2 is the total noise-variance. The first term in equation (1) describes the total contribution from all processes with correlation times much shorter than the time between recordings, $\Delta t_{\mathrm{curr}} = 1$ ps, i.e., correlation times too short to be resolved. The second term is exponentially decreasing with a characteristic time-scale τ_X. Fitted values for the parameters of $R_X(k,\ell)$ are given in table 1 for all four nucleotides. The parameter $w_{0,X}$ is the weight factor for processes with correlation times too short to be resolved. It falls in the range from 0.70 to 0.94. Thus most correlations are too brief to be resolved, probably due to reorientation of the water molecules in the solvent, which happens on a time scale of tens of femtoseconds. The longer-lasting correlations decrease exponentially in time with a characteristic time scale τ_X in the range 14–80 ps. Correlations in the current on the longer time-scale are most likely caused by the motion of the nucleotide between the electrodes.

Figure 5 shows a schematic of the time scales in the simulation. These are: the time step in the MD-simulations, $\Delta t_{\mathrm{MD}} = 1$ fs, the time interval between consecutive

‖ The black curve in figure 4 is not obtained from a fit with the expression in equation (1), but from a fit to the corresponding power spectrum (see the SI for details).
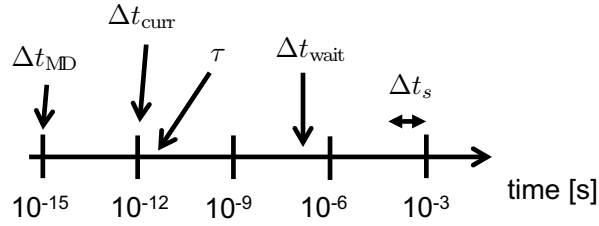
Figure 5. Schematic of the various time scales in the simulations of transverse tunneling currents through nucleotides. The time scales are the time step in MD-simulations Δt_{MD}, the time interval between consecutive recordings of the current Δt_{curr}, the correlation times in current traces τ, the average waiting time between electron tunneling Δt_{wait}, and the sampling time in an experiment Δt_s.

Table 1. Expected values μ_X and standard deviations σ_X of the current for the four nucleotides $X \in \{A, T, G, C\}$ for the current distributions shown in figure 3. The correlation time and weight factors are from fits of the experimental periodigrams to the theoretical power spectrum corresponding to the autocovariance stated in equation (1). Error bars on $w_{0,X}$ are less than 5% of the fitted values and thus not stated.

X	μ_X [pA]	σ_X [pA]	τ_X [ps]	$w_{0,X}$
A	48	41	44 ± 5	0.70
T	0.30	0.73	80 ± 40	0.92
G	4.0	3.2	60 ± 20	0.85
C	1.3	2.0	14 ± 7	0.94

recordings of the current $\Delta t_{\mathrm{curr}} = 1\,\mathrm{ps}$, and the correlation times in current traces $\tau_X \sim 40$–70 ps. Furthermore, for a current of 1 pA, the average waiting time, Δt_{wait}, between electrons is $\sim 0.1\,\mu$s; more than a 1000 times longer than the correlation time. Consequently, the measured currents are not affected by the thermal motion of the molecule, and the correlations in the calculated current signal cannot be measured experimentally. We elaborate on this finding in the next section.

3. Connecting simulated current values with experimental recordings

A current measured experimentally cannot be detected instantaneously but requires that the number of electrons passing through a surface is recorded over a finite time interval. That is, the current $I^X_{\mathrm{mes},i}$ measured at discrete time points $t_i = i\Delta t_s$ is the number of electrons N passing through the nucleotide X from time $t_i - \Delta t_s$ to t_i divided by the length of the interval ($I^X_{\mathrm{mes},i} = N/\Delta t_s$). Here, we argue that the uncertainty in the measured current is caused by two effects. The shot noise due to the discreteness of electrons, and the correlation time between current values. For the simulations considered here, we demonstrate that the uncertainty in the measured current is dominated by shot noise.

First, the low current values (fA to nA) set a lower limit on the experimental sampling time Δt_s. With the assumptions of an ideal detector and no correlations between events of electron tunneling, the latter events satisfy Poisson statistics, so a recording with an expected value of $\langle N \rangle$ electrons in the time interval Δt_s will have a relative uncertainty on the number of measured electrons of $1/\sqrt{\langle N \rangle}$. This uncertainty is due to shot noise.

Suppose we aim for an uncertainty of $3\,\%$, which requires $\langle N \rangle = 1000$. A current signal of the order of picoamperes corresponds to an expected value of approximately 10^7 electrons passing through the nucleotide per second. Thus, a measurement time of approximately $10^{-4}\,\mathrm{s} = 0.1\,\mathrm{ms}$ is needed to detect 1000 electrons on average. A sampling time $\Delta t_s = 0.1\,\mathrm{ms}$ gives a sampling frequency $f_s = 1/\Delta t_s = 10\,\mathrm{kHz}$¶. Similarly, detection of currents in the nanoamp-regime requires sampling frequencies of at most MHz. Higher sampling frequencies require larger currents. Thus, it seems of questionable relevance to analyze simulated current spikes with durations down to a few picoseconds and a current signal in the nanoamp-regime. Increased sampling frequency also leads to increased thermal noise, as we discuss in section 4.

Next, we consider the uncertainty in the measured current due to its auto-correlated variation caused by the thermal motion of the nucleotide. Mathematically, the current value $I_{\mathrm{mes},i}^X$ recorded for nucleotide X and associated with the point in time $t_i = i\Delta t_s$ is

$$I_{\mathrm{mes},i}^X = \frac{1}{\Delta t_s} \int_{t_i - \Delta t_s}^{t_i} dt\, I^X(t), \tag{2}$$

where $I^X(t)$ is the steady-state current for the configuration of the system at time t (see above). As $I^X(t)$ is fluctuating, the measured current $I_{\mathrm{mes},i}^X$ is a stochastic variable. It can be characterized by its expected value and its standard deviation. The expected value of the measured current is $\langle I_{\mathrm{mes},i}^X \rangle = \langle I^X(t) \rangle = \mu_X$. The standard deviation of the measured current depends on the correlations in the current due to the dynamics of the molecule itself and the motion of the surrounding water molecules. With the autocovariance defined in equation (1), the variance of the measured current is (for details, see SI)

$$\begin{aligned}
\sigma_{\mathrm{mes},X}^2 &\equiv \left\langle \left(I_{\mathrm{mes},i}^X - \mu_X \right)^2 \right\rangle \\
&\simeq \sigma_X^2 \left[\frac{\Delta t_{\mathrm{curr}}}{\Delta t_s} w_{0,X} + (1 - w_{0,X}) \frac{2\tau_X}{\Delta t_s} \right] \\
&\simeq (1 - w_{0,X}) \frac{2\tau_X}{\Delta t_s} \sigma_X^2,
\end{aligned} \tag{3}$$

where we used in the last two lines that the sampling time is much longer than the correlation time $\Delta t_s \gg \tau_X \gg \Delta t_{\mathrm{curr}}$. For a sampling frequency of $10\,\mathrm{kHz}$ and a correlation time of, say, $50\,\mathrm{ps}$, the prefactor is $2\tau_X/\Delta t_s \sim 10^{-6}$. So the standard deviation of the measured current is $\sigma_{\mathrm{mes},X} \sim 10^{-3}\sigma_X$, which for the present data is of

¶ A sampling frequency of 10 kHz is ten times the sampling frequency in the break-junction experiments in reference [8].

the order of, or less than, femtoamps. That is, the relative uncertainty of the measured current due to the thermal motion of the nucleotide is $\sigma_{\text{mes},X}/\mu_X \sim 10^{-3}$, which is much lower than the relative uncertainty due to shot noise. Thermal motion of the nucleotide thus does not affect experimental measurements.

According to table 1, the minimum distance between the expected current values $|\mu_X - \mu_{X'}|$ for $X \neq X'$ is approximately 1 pA; much larger than $\sigma_{\text{mes},X}$, or, e.g., the 3 % relative uncertainty caused by shot noise for $\langle N \rangle = 1000$. Consequently, an *ideal* measurement could easily distinguish between the four types of nucleotides as the distributions of the measured currents are nonoverlapping. That is, neither the configurational changes of the nucleotide and the surrounding water molecule nor the shot noise can explain the overlapping distributions seen in experiments. Furthermore, an ideal experiment would only be able to estimate the expected values, μ_X, of the simulated current probability distributions in figure 3, not the actual shapes of the distributions.

Finally, we notice that even though the molecule in the simulation goes through many different configurations during a given measurement, we do not know how much of its phase space is sampled. The molecule could be trapped in a local minimum and only sample a fraction of all possible minima. Therefore simulations should be performed for different initial configurations, and the dependence on the initial conditions should be investigated.

In section 4 we discuss the role of the thermal noise and in section 5 how filters change the current distributions for the case where the width of the distributions are not made negligible by the time-averaging in equation (2).

4. Experimental noise

Noise is unavoidable in real measurements. It causes current distributions to overlap and must be accounted for in order to avoid ambiguous classifications of the signal. Previous work has characterized the noise in the ionic current through a solid-state nanopore in a SiN membrane [31] and through graphene nanopores [32]. Both cases show a $1/f$-distribution at low frequencies. Reference [33] characterized the noise in the voltage across a gold-wire break-junction in vacuum at room temperature. Both in the presence and absence of a molecule in the junction, at high frequencies the power spectrum of the voltage is identical to the spectrum of thermal (Johnson-Nyquist) noise.

Thermal noise is inevitable in electronic circuits and is due to the thermal voltage fluctuations in a resistor [34]. It causes a Gaussian distributed white noise with standard deviation

$$\sigma_{\text{th}} = \sqrt{\frac{4k_{\text{B}}T\Delta f}{R}}. \tag{4}$$

Here, Δf is the frequency bandwidth within which the current is measured, and R is the resistance of a load resistance put in series with the molecular junction. Notice
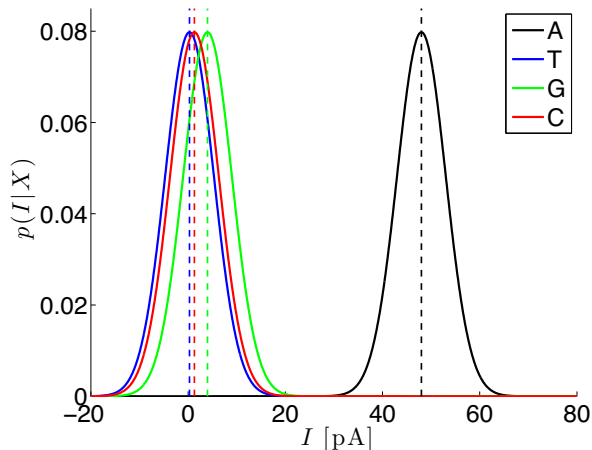
Figure 6. Illustration of the distributions of the measured currents for the four different nucleotides. The expected values are taken from table 1 and the widths are due to an added experimental noise with vanishing expected value and standard deviation $\sigma_{\text{noise}} = 5\,\text{pA}$. As σ_{noise} is larger than the expected value of the current for the nucleotides T, G, and C, negative current values occur for these nucleotides.

in particular how a decreased sampling time increases the thermal noise if the total measurement time t_{msr} is kept unchanged ($\Delta f = f_{\text{Nyq}} - 1/t_{\text{msr}} \simeq f_{\text{Nyq}} = 1/(2\Delta t_s)$). Equation 4 describes a system in equilibrium, while the noise increases if a DC voltage is applied. For measurements with nanogaps in a liquid environment, the standard deviation of the measured background signal was 10 pA for a load resistance of $10\,\text{k}\Omega$ and a bandwidth $\Delta f \simeq 1/(2\Delta t_s) = 0.5\,\text{kHz}$ [8]. Thus the estimate for the standard deviation of the thermal noise before filtering is $\sim 30\,\text{pA}$. Electronic lowpass filters reduce this noise amplitude, however (see section 5).

Figure 6 illustrates this situation with normal distributions with expected values given by μ_X in table 1 and with standard deviations $\sigma_{\text{noise}} = 5\,\text{pA}$. That is, we assume that the noise is normal distributed and added to the signal from the molecule. The distributions show clear overlaps for $X = T, G$, and C, as σ_{noise} is larger than the distance between the expected values. Current signals from the base A are well separated from the other values, making this nucleotide easily distinguishable. We use the distributions in figure 6 when we discuss nucleotide assignment and the corresponding error rates in sections 6 and 7, respectively.

5. Influence of electronic filters

Electronic lowpass filters are indispensable for measurements of small currents. They reduce the noise in measurements, but they also modify the shape of spikes in the signal. This effect is well-studied for the higher-order Bessel filters often used in patch-clamp techniques [35] and in measurements of the ionic blockade in nanopores [36] (see, e.g., references [35] and [37] for an introduction to random data and filters). Filters also

change the distribution of the measured current values, which must be considered when comparing measured and simulated currents (figure 7). Finally, filters introduce autocorrelations in the signal. An autocorrelated time series of current measurements contains less information than an uncorrelated series with the same variance, and thus gives higher error rates for the nucleotide assignment. The latter point is addressed in section 7.

Linear filters change an incoming signal by outputting a weighted sum over input values. Described in continuous time,

$$I_{\text{out}}(t) = \int_{-\infty}^{\infty} dt' h(t - t') I_{\text{in}}(t'), \tag{5}$$

where $I_{\text{in/out}}$ is the current before and after the filter, respectively, and the weight factor $h(t)$ is the filter's transfer function. For a causal system $h(t) = 0$ for $t < 0$. The Fourier transform of the transfer function is the frequency response function $H(f)$. Since a factor 2 is very nearly 3 dB, the frequency at which $|H(f)|^2 = 1/2$ is denoted by f_{3dB}. It is also called the critical frequency and denoted by f_c. In experiments, f_c-frequency is typically chosen as a fraction of the Nyquist frequency $f_{\text{Nyq}} = 1/(2\Delta t_s)$.

A discrete linear filter relates discrete inputs to outputs as

$$I_{\text{out},i} = \sum_{j=-\infty}^{\infty} h_{i-j} I_{\text{in},j} \ . \tag{6}$$

As an example, we here consider a simple first-order filter ($0 \leq \alpha \leq 1$),

$$I_{\text{out},i} = \alpha I_{\text{in},i} + (1 - \alpha) I_{\text{out},i-1} \ . \tag{7}$$

Here the output at a given point in time is the weighted sum of the simultaneous input, $I_{\text{in},i}$, and the output at the previous point in time, $I_{\text{out},i-1}$. Iteration of equation (7) gives the weight factors of the filter: $h_j = \alpha(1 - \alpha)^j = \alpha e^{j \ln(1-\alpha)} = \alpha e^{-j\Delta t_s/\tau_c}$ for $j \geq 0$ and zero otherwise, i.e., the output is an exponentially weighted superposition of the current and all past inputs. The characteristic time scale is $\tau_c = -\Delta t_s / \ln(1 - \alpha)$, and the characteristic frequency is $f_c = 1/(2\pi\tau_c)$.

Now consider an uncorrelated input signal with μ the average current and σ_{in}^2 the variance of the input signal, i.e., $\langle (I_{\text{in},i} - \mu)(I_{\text{in},j} - \mu) \rangle = \sigma_{\text{in}}^2 \delta_{i,j}$. With equation (6) and the definition of the exponential filter, the autocovariance of the output current follows,

$$R_{\text{out}}(i,j) \equiv \langle (I_{\text{out},i} - \mu)(I_{\text{out},j} - \mu) \rangle$$
$$= \sigma_{\text{out}}^2 e^{-|i-j|\Delta t_s/\tau_c} \ . \tag{8}$$

Here we have introduced $\sigma_{\text{out}}^2 = \sigma_{\text{in}}^2 \frac{\alpha}{2-\alpha}$. The first-order filter thus gives an exponentially decreasing correlation function and lowers the value of the total variance. We use this expression for the correlation function in section 7, where we calculate the error rates for nucleotide assignment for correlated data.

The distribution of the recorded output relative to the input is also changed by filters. Assume it were possible to measure the current values in figure 3 with a sampling
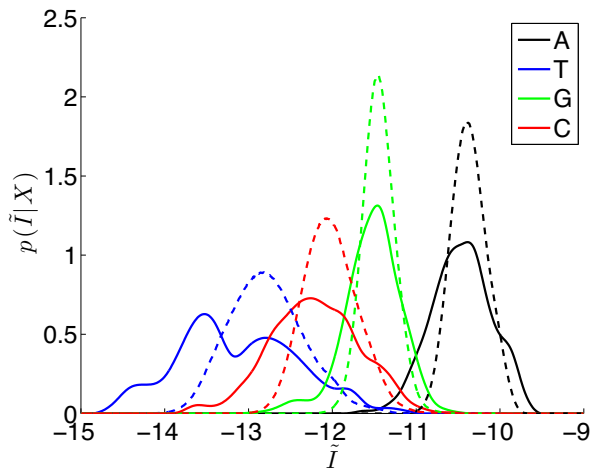
Figure 7. Effect of filtering. The continuous lines (reproduced from figure 3 for convenience) show probability distributions of simulated currents. The dashed lines show probability distributions of filtered simulated currents (first-order filter with $f_c = f_{\mathrm{Nyq}}/4$).

time as brief as the time between recordings, i.e., with $\Delta t_s = \Delta t_{\mathrm{curr}}$. Assume also absence of intrinsic correlations ($w_{0,X} = 1$) and a simple first-order filter with critical frequency $f_c = f_{\mathrm{Nyq}}/4$, i.e., with characteristic time scale $\tau_c \simeq 1.27\Delta t_s$ [+]. Then the distribution of the sampled current values would follow the distributions shown with dashed lines in figure 7. The filtered distributions are smoother than the original ones, and the standard deviations are reduced [see text below equation (8)]. In the limit of very long characteristic times, $\tau_c \gg \Delta t_s$, the distributions approach normal distributions by force of the central limit theorem. These effects are important to keep in mind when comparing simulation results with experimental data, as the comparison must take into account the distortion of experimental distributions by filters. This could be relevant, e.g., for simulations of the current through a nanoribbon with nucleotides passing through a hole in it. Simulations show an overlap for different nucleotides [22], but electronic filters will decrease these overlaps.

Finally, the autocovariance of experimental data is often affected both by the physical processes in the measured device and by filters in the data acquisition electronics [31, 32]. If the autocovariance can be determined experimentally, it can serve as input for the covariance matrix used when estimating the error rates.

6. Nucleotide assignment using maximum likelihood and error rates

Classification of output from biosensors (and sequencers) is often ambiguous because output values contain a stochastic element. When probability distributions for output values overlap, one cannot tell from a single measurement which input caused the

[+] For a discussion of filter design and of how to choose the critical frequency, see, e.g., [35].

output. For experimentally measured current signals the assignment is often further complicated due to, e.g., a varying background signal. The classification problem can then, e.g., be addressed by machine learning techniques, like Support Vector Machine (SVM) [30,38]. For simulated data with a stable background and with the current distributions for the different molecules available, we suggest to use the maximum likelihood decision rule for nucleotide assignment as it is a straightforward and standard procedure [39]. In addition, it is easy to simulate the corresponding error rates without any adjustable parameters. In the assignment procedure, the influence of time averaging, experimental noise, and correlations in the signal are included. We give here a basic vocabulary for the problem of how to assign a nucleotide to a given current signal; for a detailed introduction to pattern classification, see, e.g., reference [39].

As an example, we use the four different types of nucleotides $X \in \{A, T, G, C\}$ and their four associated distributions of values for the transverse tunnelling current. Let

$$\mathbf{I}_m^X = (I_1^X, I_2^X, \ldots, I_m^X) \tag{9}$$

denote the time series of m current measurements. All current values I_n^X stem from the *same* nucleotide, so we drop the superscript X from now on. Notice that it is assumed that the probability distribution of current values is known for each nucleotide. So given a current signal $\mathbf{I}_m = \mathbf{I}$ consisting of m measurements, the task is to give an algorithm for how to assign a specific type of nucleotide to the current signal and to determine the error rate, i.e., the relative frequency with which the assignment is incorrect.

The current signal $\mathbf{I}$ is our observation. It stems from one of the four types of nucleotides $X \in \{A, T, C, G\}$. The variable X denotes the '*state of nature*'. Let $P(X)$ denote the *a priori* probability for the nucleotide being X. How probable it is to observe the signal $\mathbf{I}$, will depend on the 'state of nature,' the value of X. So we introduce the *class-conditional probability distributions* $p(\mathbf{I}|X)$. For our problem, these functions are the probability distributions for values of currents (see figure 3), and they are known *a priori* from the simulations. If we assume that the priors $P(X)$ are also known, Bayes' formula states that the relation between the prior and the posterior probabilities, i.e., the probability that the 'state of nature' is X given the observation $\mathbf{I}$ is

$$P(X|\mathbf{I}) = \frac{p(\mathbf{I}|X)P(X)}{\sum_{X'} p(\mathbf{I}|X')P(X')}. \tag{10}$$

Notice the normalization condition $\sum_X P(X|\mathbf{I}) = 1$. Here, we also follow the convention in reference [39] and let the probability functions over discrete and continuous sets be denoted by upper-case P and lower-case p, respectively.

We need a decision rule to decide which 'state of nature' the system was in when it produced the current signal $\mathbf{I}$. It can be shown that the decision rule which minimises the error is Bayes' Decision Rule [39], which amounts to choosing the 'state-of-nature' X with the highest *a posteriori* probability $P(X|\mathbf{I})$. If we have no prior information about the molecules, it is reasonable to assume that they all have the same *a priori*

probability $P(X)$ for all X. This gives the maximum likelihood decision rule, which is to choose the X which maximizes the likelihood $p(\mathbf{I}|X)$, i.e.,

$$\text{decide } X \text{ if } p(\mathbf{I}|X) > p(\mathbf{I}|X') \text{ for all } X' \neq X. \tag{11}$$

This is the decision rule we will use below. Notice how the decision rule divides the m-dimensional space for the observable $\mathbf{I}$ into different domains D_X, where D_X is the domain where we choose X, i.e., $D_X = \{\mathbf{I} \,|\, p(\mathbf{I}|X) > p(\mathbf{I}|X') \text{ for all } X' \neq X\}$. This can also be expressed as an indicator function $\mathbf{1}_{D_X}(\mathbf{I})$ with the properties $\mathbf{1}_{D_X}(\mathbf{I}) = 1$ if $p(\mathbf{I}|X) > p(\mathbf{I}|X')$ for all $X' \neq X$ and 0 otherwise.

The different domains D_X are simple to illustrate for the probability distributions in figure 3 for the $m = 1$ case of a single measurement, see the horizontal arrows in figure 3. The vertical dashed lines mark the intersections between the distributions. For general probability density distributions, the partition of the space of possible current values may be more complicated.

So far we have not specified how to calculate the class-conditional probability density function $p(\mathbf{I}_m|X)$, but we return to this issue in section 7.

The easiest way to find the error rate is to calculate the probability $P_{\text{correct},m}^X$ of a correct assignment for the nucleotide X, and then find the error rate as $e_m^X = 1 - P_{\text{correct}}^X$. The probability of being correct can be expressed as the probability that the 'state of nature' is X and $\mathbf{I}$ is in D_X, i.e., [39]

$$
\begin{aligned}
P_{\text{correct}}^X &= \sum_X P(\mathbf{I} \in D_X|X)P(X) \\
&= \sum_X \int_{D_X} p(\mathbf{I}|X)P(X)\,d\mathbf{I} \\
&= \sum_X \int \mathbf{1}_{D_X}(\mathbf{I})p(\mathbf{I}|X)P(X)\,d\mathbf{I}
\end{aligned}
\tag{12}
$$

Here, $\mathbf{1}_{D_X}$ is an indicator function that is specified above for the maximum likelihood decision rule, although other possibilities exist [39].

Given a set of probability distributions $p(\mathbf{I}|X)$ and a partition D_X dividing the range of outcomes for $\mathbf{I}$, error rates can be calculated by direct evaluation of the m-dimensional integral in equation (12), e.g., by Monte Carlo integration [40]. Often it is much easier to Monte Carlo simulate the error rates, which is done separately for each type of nucleotide X_{chosen}. In case of m measurements, the procedure is:

(i) From the current probability distribution $p(\mathbf{I}_m|X_{\text{chosen}})$ draw m independent current values $\mathbf{I}_m$, (ii) calculate for all four nucleotides the conditional probability density $p(\mathbf{I}_m|X)$, (iii) assign to the current sequence $\mathbf{I}_m$ the nucleotide X_{assigned} with the highest conditional probability density $p(\mathbf{I}_m|X)$, and finally (iv) record whether the chosen nucleotide X_{chosen} is identical to the assigned nucleotide X_{assigned}. Steps (i)-(iv) are repeated many times.

The error rate e_m^X is simply the relative frequency with which a different nucleotide is assigned to a current sequence produced by the nucleotide X_{chosen}. The weighted

average of the error rates is

$$e_m = 1 - P_{\text{correct},m} = \sum_X e_m^X P(X), \tag{13}$$

where $P(X)$ is the prior for the nucleotide of type X.

In section 7 we demonstrate how to calculate the error rates of the nucleotide assignment for the distributions in figure 6 when the current measurements are correlated by first-order filtering.

7. Error rates for correlated data

Assignment of nucleotides and the corresponding error rates depend on the class-conditional probability density function $p(\mathbf{I}_m|X)$, i.e., the probability to measure the set of current values $\mathbf{I}_m$ for given nucleotide X. We argued above that both physical processes and electronic filters introduce correlations in the measured signal. We here demonstrate how the correlations influence the error rates for the nucleotide assignment.

For the sake of simplicity, we assume that the measurement noise is normally distributed as it is, e.g., for thermal noise. Then the probability density function $p(\mathbf{I}_m|X)$ is given by the multivariate normal distribution

$$p(\mathbf{I}_m|X) = \frac{1}{\sqrt{\det{(2\pi\mathbf{\Sigma}_X)}}}$$
$$\times \exp\left(-\frac{1}{2}\left[\mathbf{I}_m - \boldsymbol{\mu}_X\right]^T \mathbf{\Sigma}_X^{-1}\left[\mathbf{I}_m - \boldsymbol{\mu}_X\right]\right). \tag{14}$$

Here $\mu_{\mathbf{X}}$ is an m-dimensional vector with identical elements μ_X, and $\mathbf{\Sigma}_X$ is the (positive definite) $m \times m$-covariance matrix $\Sigma_{X,ij} = R(i,j)$, $i,j = 1, 2, \ldots, m$, where $R(i,j)$ is the autocovariance. Notice that if the current values are independent and identically distributed, the covariance matrix is a diagonal matrix with the variance of the distribution on the diagonal, $\Sigma_{X,ij} = \sigma_X^2 \delta_{ij}$. Then the expression in equation (14) reduces to the product form $p(\mathbf{I}_m|X) = \prod_{n=1}^m p(I_n|X) = \prod_{n=1}^m \frac{1}{\sqrt{2\pi\sigma_x^2}}\exp[-(I_n - \mu_X)^2/(2\sigma_X^2)]$.

As an example, we consider the case where the autocovariance is identical for all four nucleotides, and the autocoavariance matrix is $\Sigma_{X,ij} = \Sigma_{ij} = \sigma_{\text{noise}}^2 e^{-|i-j|\Delta t_s/\tau_c}$. This corresponds to the output from a first-order filter with a characteristic time scale τ_c, given a white-noise input. The characteristic time scale is again chosen such that it corresponds to a first-order filter with a critical frequency $f_c = \frac{1}{2\pi\tau_c} = f_{\text{Nyq}}/4$, i.e., $\tau_c \simeq 1.27\Delta t_s$. For the current distributions shown in figure 6, we then simulate the assignment of nucleotides as described above with the use of equation (14). Finally, we calculate the error rates for the individual nucleotides, e_m^X, and the average error rate, e_m, from equation (13)*. The error rates versus the number of measurement are shown as dashed lines in figure 8. Full lines are the results for independent measurements, all

---

* Multivariate normal distributions are built-in functions in, e.g., matlab.
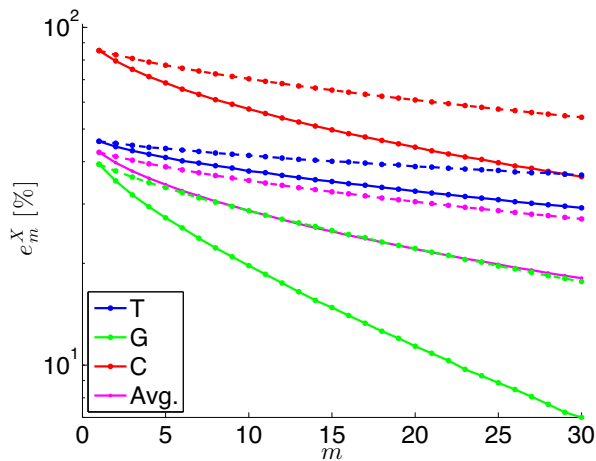
Figure 8. Error rates $e_m^X = 1 - P_{\text{correct}}^X$ versus the number of measurements m for the distributions T, G, and C in figure 6 (error rates for the nucleotide A are less than 0.01% for all m and thus not shown). Full lines show the error rates for uncorrelated data, while the dashed lines show error rates for data filtered through a first-order filter with a critical frequency $f_c = f_{\text{Nyq}}/4$. Notice that the total noise variance is $\Sigma_{ii} = \sigma_{\text{noise}}^2$ for both the correlated and uncorrelated data. The weighted average, e_m, of the error rate over all four nucleotides [equation (13) with $P(X) = 0.25$ for all X] is shown with magenta lines.

with the same total noise variance, i.e., $\Sigma_{ij} = \sigma_{\text{noise}}^2 \delta_{ij}$. Error rates are higher and decay slower for correlated than for independent measurements, since correlated data contain less independent information. Error rates for a Gaussian filter with the same critical frequency and using the same noise variance are found in SI. The results are very similar as those for a first-order filter with the same characteristic time-scale.

These findings stress the importance of including correlations in the algorithms for nucleotide assignment or step detection in experimental signals. The version of the step-finder algorithm CUSUM used for detection of multi-level events in nanopore translocation experiments [17] assumes a signal consisting of independent data points, although this condition is not fulfilled by the experimental data. The assumption might influence the results of the nucleotide assignment and the corresponding error rates; especially for high noise levels and small level separations of the expected current values for the different nucleotides.

The duration of the time a nucleotide spends between the electrodes determines the number of measurements done on it. Typically this cannot be easily controlled experimentally as the detachment of the nucleotides from the electrodes is a stochastic process, and the distribution of durations often is rather broad. For GMP molecules in a break-junction, the duration in the gap was in the interval from 1 to 100 ms and showed a dependence on the applied bias [8]. For a sampling frequency of 1 kHz, it corresponds to up to 100 measurements at the electrodes. The duration the target molecule spends at the electrodes can be increased by functionalizing the junction, which gives durations up to a second [13, 14, 38, 41]. Thus the relevance of theoretical proposals for sequencing

or biosensing depends both on the decrease of error rates with the number of measurements and on the four distributions of time spent by Molecule X between the electrodes.

8. Discussion

The present study emphasizes that the very weak transverse tunneling currents require experimental current measurements with long integration times, and it describes the consequences of a long integration time for the measured currents. These considerations are relevant not only for sequencing with fixed electrodes but also for simulations of nanopore sequencing of single-stranded DNA with graphene nanoribbons [27] and for recognition tunneling [30].

One consequence of the long integration time is that only the expected value of the current is probed experimentally, because the required integration time is very much longer than the autocorrelation time of current fluctuations caused by the nucleotide's thermal motion. Thus, a current measurement averages over so many different orientations of the nucleotide in the gap junction that the resulting current value is a thermal average with no dependence on nucleotide orientation. Consequently, different measurements with such long integration times should give very similar current values, i.e., values with a very narrow distribution on the current axis. Nevertheless, the full distributions of the simulated transverse tunneling currents are needed in order to determine their expected values. This is because the simulated current values for each nucleotide span almost three orders of magnitude due to the thermal fluctuations of the molecule in the nanogap. So it is not sufficient to calculate the tunneling current for only a few fixed configurations of a nucleotide. This can lead to incorrect values for the current's expected value.

Secondly, in the original simulations of transverse tunneling through nucleotides, the electron transport was described as coherent tunneling [6, 7]. A later simulation included dephasing of the tunneling electrons due to the fluctuations of the molecule and its environment. These effects changed the distribution of the simulated current values [20]. For experimentally relevant values of this dephasing, it caused a slight downward shift in the expected value of the current. It also slightly changed the shape of the current distribution. The shift might be detectable in experiments, but the change of shape is washed out by the long integration time required in real experiments.

We also addressed how to assign a nucleotide to a measured current signal with the maximum likelihood decision rule. The general challenge for the assignment is that the four different nucleotides have overlapping current distributions, broadened by electronic noise in the data acquisition system. Electronic sequencing would be easy without these overlaps: A single measurement of the transverse tunneling current would identify a nucleotide.

With some overlap, we can still distinguish between different nucleotides albeit with

non-zero error rate. We just need to repeat measurements on the individual nucleotide several times to obtain a reliable result. We must, however, consider that electronic filters in the data acquisition system produce autocorrelations in the filtered signal. So although electronic filters are indispensable for measurements of small currents, their effect on the recorded current signal must be included in the data analysis, since filtering reduce the information content in the signal relatively to a signal with the same number of measurements but with independent data points.

The maximum likelihood framework for nucleotide assignment is easily generalized to more complicated setups than just a single pair of electrodes (see, e.g., the setup in [29]), or extended to include other types of information than just the measured current values. Other aspects that could help the identification could be, e.g., the duration of current spikes, the time interval between spikes, and the fluctuations of currents within spikes [30]. This extra information can be exploited in the assignment of a molecule to a recorded signal, if correlations between the measured quantities—e.g., the duration of a spike and its height—are correctly accounted for in the analysis.

Recently, it was investigated theoretically by simulations whether the use of multiple electrode pairs coupled in series could improve identification of nucleotides [29, 42]. The advantage of multiple electrodes is an increased number of measurements for each nucleotide and, consequently, a lower error rate. If the distribution of current values measured with each electrode pair is known, then the assignment procedure described above can be applied directly.

9. Conclusion

We have demonstrated the importance of realistic experimental integration times, of autocorrelation times in simulated current values, and of electronic noise and filters. Simulations must relate to real experimental measurements, obviously, in order to access the feasibility of theoretical proposals for real experiments. When the probability distributions of current values are known, which is the case for simulated data, we recommend using the maximum likelihood decision rule for nucleotide assignment, but also account for the correlations in the measured signal in order not to underestimate the error rates for the assignment.

References

[1] Branton D, Deamer D W, Marziali A, Bayley H, Benner S A, Butler T, Di Ventra M, Garaj S, Hibbs A, Huang X, Jovanovich S B, Krstic P S, Lindsay S, Ling X S, Mastrangelo C H, Meller A, Oliver J S, Pershin Y V, Ramsey J M, Riehn R, Soni G V, Tabard-Cossa V, Wanunu M, Wiggin M and Schloss J A 2008 *Nat. Biotechnol.* **26** 1146–1153

[2] Venkatesan B M and Bashir R 2011 *Nat. Nanotechnol.* **6** 615–624

[3] Muthukumar M, Plesa C and Dekker C 2015 *Phys. Today* **68** 40–46

[4] Heerema S J and Dekker C 2016 *Nat. Nanotechnol.* **11** 127–136

[5] Di Ventra M and Taniguchi M 2016 *Nat. Nanotechnol.* **11** 117–126

[6] Zwolak M and Di Ventra M 2005 *Nano Lett.* **5** 421–424

[7] Lagerqvist J, Zwolak M and Di Ventra M 2006 *Nano Lett.* **6** 779–782

[8] Tsutsui M, Taniguchi M, Yokota K and Kawai T 2010 *Nat. Nanotechnol.* **5** 286–290

[9] Ohshiro T, Matsubara K, Tsutsui M, Furuhashi M, Taniguchi M and Kawai T 2012 *Sci. Rep.* **2**

[10] Tsutsui M, Matsubara K, Ohshiro T, Furuhashi M, Taniguchi M and Kawai T 2011 *J. Am. Chem. Soc.* **133** 9124–9128

[11] Ohshiro T, Tsutsui M, Yokota K, Furuhashi M, Taniguchi M and Kawai T 2014 *Nat. Nanotechnol.* **9** 835–840

[12] Zhao Y, Ashcroft B, Zhang P, Liu H, Sen S, Song W, Im J, Gyarfas B, Manna S, Biswas S, Borges C and Lindsay S 2014 *Nat. Nanotechnol.* **9** 466–473

[13] Chang S, Huang S, He J, Liang F, Zhang P, Li S, Chen X, Sankey O and Lindsay S 2010 *Nano Lett.* **10** 1070–1075

[14] Huang S, He J, Chang S, Zhang P, Liang F, Li S, Tuchband M, Fuhrmann A, Ros R and Lindsay S 2010 *Nature Nanotechnol.* **5** 868–873

[15] Pang P, Ashcroft B A, Song W, Zhang P, Biswas S, Qing Q, Yang J, Nemanich R J, Bai J, Smith J T, Reuter K, Balagurusamy V S K, Astier Y, Stolovitzky G and Lindsay S 2014 *ACS Nano* **8** 11994–12003

[16] Xie P, Xiong Q, Fang Y, Qing Q and Lieber C M 2012 *Nature Nanotechnol.* **7** 119–125

[17] Raillon C, Granjon P, Graf M, Steinbock L J and Radenovic A 2012 *Nanoscale* **4**(16) 4916–4924

[18] Plesa C and Dekker C 2015 *Nanotechnology* **26** 084003

[19] Zwolak M and Di Ventra M 2008 *Rev. Mod. Phys.* **80** 141–165

[20] Krems M, Zwolak M, Pershin Y V and Di Ventra M 2009 *Biophys. J.* **97** 1990–1996

[21] Nelson T, Zhang B and Prezhdo O V 2010 *Nano Lett.* **10** 3237–3242

[22] Saha K K, Drndić M and Nikolić B K 2012 *Nano Lett.* **12** 50–55

[23] Ahmed T, Kilina S, Das T, Haraldsen J T, Rehr J J and Balatsky A V 2012 *Nano Lett.* **12** 927–931

[24] Ahmed T, Haraldsen J T, Zhu J X and Balatsky A V 2014 *J. Phys. Chem. Lett.* **5** 2601–2607

[25] Farimani A B, Min K and Aluru N R 2014 *ACS Nano* **8** 7914–7922

[26] Kim H S and Kim Y H 2015 *Biosens. Bioelectron.* **69** 186–198

[27] Qiu H, Sarathy A, Leburton J P and Schulten K 2015 *Nano Lett.* **15** 8322–8330

[28] Qiu H, Girdhar A, Schulten K and Leburton J P 2016 *ACS Nano* **10** 4482–4488

[29] Boynton P, Balatsky A V, Schuller I K and Di Ventra M 2014 *J. Comput. Electron.* **13** 1–7

[30] Krstić P, Ashcroft B and Lindsay S 2015 *Nanotechnology* **26** 084001

[31] Smeets R M M, Keyser U F, Dekker N H and Dekker C 2008 *Proc. Natl. Acad. Sci. U.S.A.* **105** 417–421

[32] Heerema S J, Schneider G F, Rozemuller M, Vicarelli L, Zandbergen H W and Dekker C 2015 *Nanotechnology* **26** 074001

[33] Sydoruk V A, Xiang D, Vitusevich S A, Petrychuk M V, Vladyka A, Zhang Y, Offenhäusser A, Kochelap V A, Belyaev A E and Mayer D 2012 *J. Appl. Phys.* **112** 014908

[34] Kittel C and Kroemer H 1980 *Thermal physics* (San Francisco: W.H. Freeman)

[35] Colquhoun D and Sigworth F J 1995 Fitting and statistical analysis of single-channel records *Single-Channel Recording* ed Sakmann B and Neher E (Springer US) pp 483–587

[36] Pedone D, Firnkes M and Rant U 2009 *Anal. Chem.* **81** 9689–9694

[37] Bendat J S and Piersol A G 2010 *Random Data: Analysis and Measurement Procedures* 4th ed (Hoboken, N.J: Wiley)

[38] Chang S, Huang S, Liu H, Zhang P, Liang F, Akahori R, Li S, Gyarfas B, Shumway J, Ashcroft B, He J and Lindsay S 2012 *Nanotechnology* **23** 235101

[39] Duda R O, Hart P E and Stork D G 2000 *Pattern Classification (2Nd Edition)* (Wiley-Interscience)

[40] Press W H 1992 *Numerical Recipes in Fortran 77: The Art of Scientific Computing* 2nd ed (Cambridge England ; New York: Cambridge University Press)

[41] Lindsay S, He J, Sankey O, Hapala P, Jelinek P, Zhang P, Chang S and Huang S 2010 *Nanotechnology* **21** 262001

[42] Ahmed T, Haraldsen J T, Rehr J J, Di Ventra M, Schuller I and Balatsky A V 2014 *Nanotechnology* **25** 125705