The Effects of Diffusion on an Exonuclease/Nanopore-Based DNA Sequencing Engine

Joseph E. Reiner[1], Arvind Balijepalli[2,3], Joseph W.F. Robertson[2],

Daniel L. Burden[4], Bryon S. Drown[4], and John J. Kasianowicz[2]

1) Department of Physics, Virginia Commonwealth University, Richmond, VA 23284

2) Physical Measurement Laboratory, Semiconductor and Dimensional Metrology Division, CMOS and Reliability Group, National Institute of Standards and Technology, Gaithersburg, MD 20899.

3) Laboratory of Computational Biology, National Heart Lung and Blood Institute, National Institutes of Health, Rockville, MD 20892

4) Department of Chemistry, Wheaton College, Wheaton, IL 60187.

Editorial corresponding author:

John J. Kasianowicz
Physical Measurement Laboratory
NIST
Gaithersburg, MD 20899-8120

301-975-5853 (office)
john.kasianowicz@nist.gov

Keywords

nanopore, alpha hemolysin, DNA sequencing, exonuclease

Abstract

Over 15 years ago, the ability to electronically detect and characterize individual polynucleotides as they are driven through a single protein ion channel was suggested as a potential method for rapidly sequencing DNA, base-by-base, in a ticker tape-like fashion. More recently, a variation of this method was proposed in which a nanopore would instead detect single nucleotides cleaved sequentially by an exonuclease enzyme in close proximity to one pore entrance. We analyze the exonuclease/nanopore-based DNA sequencing engine using analytical theory and computer simulations that describe nucleotide transport. The available data and analytical results suggest that the proposed method will be limited to reading < 80 bases, imposed in part by the short lifetime each nucleotide spends in the vicinity of the detection element within the pore and the ability to accurately discriminate between the four mononucleotides.

Introduction

The ability to sequence DNA rapidly, at low cost, is of significant interest because of its potential positive impact on health care and in many other applications. Over four decades ago, Maxam and Gilbert devised a method for sequencing DNA in which the biopolymer is chemically modified with labels and subsequently cleaved at specific bases with an organic compound [1]. Sanger and colleagues developed another method based on dideoxynucleoside triphosphates as DNA chain terminators [2]. Modern variations of the Sanger method use contemporary technology to sequence DNA templates by capillary electrophoresis with optical detection of fluorescent bases, and requires time consuming procedures[3].

High-throughput methods based on sequencing by synthesis [4] were refined in the 1990s, and are typically performed with instruments that are relatively large and expensive to operate. The method was recently miniaturized via microtiter plate technology that reads DNA sequences optically from fluorescent tags[3,5] or electronically with CMOS/ISFET technology [6-8]. The latter technique is particularly significant because the device is massively parallel and label-free.

The next logical steps in genomic DNA sequencing include reducing the number of copies needed (ideally, the limit of detection would be a single molecule, to avoid the need for PCR amplification[9] with its attendant errors [10]), increasing the read length (to decrease the complexity of genome sequence reconstruction), increasing the measurement accuracy, and decreasing the time to read each base. Towards that goal, it was suggested that individual molecules of DNA might be sequenced in a ticker-tape fashion by threading them through a single nanopore and measuring the change each base makes to the pore's ionic current [11]. However, this conceptually simple approach has been elusive[12], in part because understanding the physics of polymer translocation through a single nanopore, and controlling the rate of polynucleotide translocation have both been challenging[13-18].

One variation on nanopore-based DNA sequencing suggested the use of an exonuclease attached to the cap domain of the α-hemolysin channel [19]. In this scheme, the enzyme would cleave nucleotides, one at a time, from duplex DNA. Conceptually, each base would migrate serially into the pore, bind to a detection element there, and be uniquely identified by the degree of ionic current blockade. Because β-cyclodextrin binds mononucleotides [20-22], and fits well inside the nanopore formed by *Staphylococcus aureus* α-hemolysin, it was a logical choice for use as DNA base detection element (Fig. 1A). Based on the degree by which the four DNA mononucleotides reduced the conductance of the α-hemolysin channel with a β-cyclodextrin adapter in it (Fig. 1B)[23], an excellent degree of separation of the four bases was achieved. The actual mean overlap between the peaks (8 %), determined using a fit to the data with Voigt functions[24] (Fig. 1B-E, *indigo*) is somewhat greater than the 1 % reported with a Gaussian mixture model fit (Fig. 1B-E, *orange*) [23]. Nevertheless, the relatively high degree of accuracy of the β-cyclodextrin adapter suggested that the concept has merit and deserves a detailed theoretical analysis. Here, we study the critical parts of this system's performance, determine the probability that a given length polynucleotide can be accurately sequenced with the technology, and suggest how the method might be implemented.

Model

One dimensional diffusion model and the probability of capture

First, we analyze theoretically the capture of single mononucleotides that are released near one entrance of a nanopore. The time dependent capture probability for an idealized diffusing particle is computed using solutions of the Fokker-Planck drift-diffusion equation [25-27]. Closed form analytical solutions for this equation can usually be found for geometries with simple boundary conditions. The more complicated case of a particle diffusing in a semi-infinite solution in the vicinity of a nanopore has no known closed form solution because it requires solving and matching boundary conditions between two distinct regions (i.e., the nanopore and the bulk). The problem was simplified recently by assuming that a one-dimensional line can represent the nanopore with a perfect absorber at one end (the particle detector/reader inside the nanopore) and a radiating boundary (which describes the escape of a particle from the nanopore to the bulk) at the other. [28,29]

Each particle is initially located at the entrance of the nanopore, and the Fokker-Planck equation is solved to extract the capture time probability distribution and from it the time-integrated capture probability. Figure 2 shows a schematic of the analytical model system. The diffusing particle propagator G under a biased flow in a one-dimensional nanopore is

$$(1) \qquad \frac{\partial G}{\partial t} + v_d \frac{\partial G}{\partial z} = D \frac{\partial^2 G}{\partial t^2}, z \in [0, L]$$

where v_d and D are the particle drift velocity and diffusion coefficient, respectively. The drift includes electrophoretic and electroosmotic contributions, which are functions of the applied transmembrane potential, V (*see Appendix*). A radiating boundary condition is applied at the pore exit ($z = L$) to model the loss of nucleotide particles into the bulk solution

$$(2) \quad \kappa G(L,t) = -D \left. \frac{\partial G(z,t)}{\partial z} \right|_{z=L} + v_d G(L,t)$$

where κ is an effective rate constant [28]. For a pore coupled to a three-dimensional bulk solution, $\kappa = 8D/\pi d$ [29] where d is the pore diameter. The location of the particle sink (the β-cyclodextrin adapter), which defines L, is assumed to be that defined in the original experiment[23].

The boundary condition for particle capture by the detector located at $z = 0$ is $G(z = 0, t) = 0$. We assume the particles are initially uniformly distributed at the nanopore entrance. The localization is further assumed to occur over a region of thickness δ, such that

$$(3) \quad G(z,0) = \begin{cases} 0, & z \in [0, L-\delta] \\ 1/\delta, & z \in [L-\delta, L] \end{cases}$$

The parameter of interest is the survival time probability density; $\phi(t) = D\left.\frac{\partial G}{\partial z}\right|_{z=0}$. G is solved from Eq. 1 by applying the absorption and radiating boundary conditions and working in Laplace space (*see Appendix*). This leads to the following expression for $\hat{\phi}(s)$ the Laplace transform of the survival time probability density,

$$\hat{\phi}(s) = \frac{\exp\left(\frac{\alpha}{2}\right)}{\cosh(y) + \left(\beta + \frac{\alpha}{2}\right)\frac{\sinh(y)}{y}} \tag{4}$$

where $\alpha = |v_d|\, L/D$, $y^2 = \alpha^2/4 + 2\tau s$, $\tau = L^2/2D$, and $\beta = \kappa L/D$.

Results and Discussion

The capture probability density function, $\phi(t)$, is estimated by numerically inverting Eq. 4[1]. Figure 3 shows a comparison between these capture probabilities and results from numerical simulations at applied potentials of $V = 0$ mV and 300 mV. The results show that an exonuclease-based nanopore sequencing device will be rate-limited by the turnover of exonuclease (~ 1 ms to 10 ms) and not the capture rate of each base by the nanopore (~ 10 ns to 100 ns). Therefore, we estimate the capture probability from the time-integrated capture probability density, which follows from Eq. 4 because $\hat{\phi}(s=0) = \int_0^\infty \phi(t)\,dt$,

$$\hat{\phi}(0) = P_{cap}(V) = \frac{\exp\left(\frac{\alpha}{2}\right)}{\cosh\left(\frac{\alpha}{2}\right) + \left(1 + \frac{2\beta}{\alpha}\right)\sinh\left(\frac{\alpha}{2}\right)}. \tag{5}$$

The two parameters that describe the total capture probability are the nanopore's aspect ratio (L/d) and the voltage-dependent drift velocity v_d. Figure 4 illustrates the voltage dependence of the time integrated capture probability estimated from the analytical model and numerical simulations with no free parameters. The results demonstrate the difficulty in capturing small diffusive particles. However, they also show that increasing the potential could increase the probability of capturing a given cleaved base. Thus, Fig. 4 suggests that a near certain capture probability can be achieved with transmembrane potentials that greatly exceed 1 *V*. Implicit in this discussion is the notion that the membrane is impervious to dielectric breakdown at such

---

[1] Hollenbeck, K.J. "INVLAP.M: A matlab function for numerical inversion of Laplace transforms by the de Hoog algorithm," http://mipav.cit.nih.gov/documentation/api/gov/nih/mipav/model/algorithms/InverseLaplace.html.

extreme voltages [30-33]. However, there are other experimental issues that preclude the use of relatively high voltages. For example, only a relatively small electrostatic potential drop (ca. 1% of the total field [34-36]) biases the random walk of each negatively charged base into the pore, and most of the potential will drop across the β-cyclodextrin adapter in the nanopore (see Fig. 2 in Clarke, et al. [23]). Thus, the corresponding mean residence time of each base on the detector will decrease if the applied potential is increased.

Assuming that the reaction between the mononucleotides and the molecular adapter proceed with first order kinetics, the mean residence time of the mononucleotides in the αHL nanopore should decrease exponentially with potential, i.e., $\tau(V) = \tau_0 e^{-V/V_c}$. That was indeed the case for thymine with applied potentials between 80 mV and 200 mV (τ_0 = 0.55 s and V_c = 47 mV)[23], but not for the other three mononucleotides (*see below*). Thus, for relatively small applied potentials, the probability density of the residence times will be $\rho_V(t';\tau(V)) = e^{-t'/\tau(V)}/\tau(V)$, for $t' \geq 0$ and where *t'* is the time from capture for a given nucleotide. We make the simplifying assumption that a given base cannot be read if it is bound to the adapter for times shorter than the inverse bandwidth (*B*) of the detection system. This absolute limit ignores the possible increase in noise with detector bandwidth that may further limit the probability of correctly reading a base. Nevertheless, the assumption is instructive and points to a limitation of increasing the capture probability through increased voltage. The total probability of successfully reading a base follows from the normalized probability density,

$$(6) \qquad P_{read}(V;B) = \int_{1/B}^{\infty} \rho_V(t';\tau(V))\,dt' = \exp\left(-\frac{\exp\left(-\frac{V}{V_c}\right)}{B\tau_o}\right)$$

Assuming that the processes governing the capture and read of individual nucleotides are independent, we can multiply Eqs. 5 and 6 to calculate the total probability of sensing a single base

$$(7) \qquad P(V;B) = \frac{\exp\left(\frac{\alpha}{2}\right)\exp\left(-\frac{\exp\left(-V/V_c\right)}{B\tau_o}\right)}{\cosh\left(\frac{\alpha}{2}\right) + \left(1+\frac{2\beta}{\alpha}\right)\sinh\left(\frac{\alpha}{2}\right)}$$

Figure 5 shows plots of the combined detection probability as a function of the transmembrane potential (Eq. 7) for three detection bandwidths (*B* = 5 kHz, 30 kHz and 500 kHz). The optimum probability, P_{max}, follows from Eq. 7 by setting $\partial P/\partial V = 0$ and is shown as a function of bandwidth in Fig. 5 (*inset*). The optimum read probability initially increases as a function bandwidth, but saturates at ca. 0.75. This result suggests that increasing the detection bandwidth beyond 500 kHz will have a limited impact on the efficiency of reading a given nucleotide and will most likely be worse given that detector noise will grow with increasing bandwidth. However, at high potentials, the mean residence time distributions for the negatively charged

nucleotides on the molecular adapter may only decrease linearly with voltage, not exponentially, if the potential drop across the adapter is much greater than the binding energy. The effect of this transition between the two regimes will cause the capture probability to decrease more slowly with increasing bandwidth than is shown in Fig. 5. Nevertheless, the probability of detecting a base will be limited in this extreme voltage regime.

In addition to the system bandwidth, which limits the fastest blockade events that can be accurately measured, the relatively small number of ions that probe each base during fast events will also be a limiting factor. For example, approximately 160 ions flow through the pore per microsecond under the experimental conditions reported elsewhere (400 mM KCl, 180 mV [23]), and therefore events that are measured at the 500 kHz bandwidth limit will have sufficiently large statistical counting error (ca. 8%).

Molecular adapter chemistry

The choice of β-cyclodextrin as a molecular adapter for detecting and discriminating the four DNA mononucleotides[23] was sensible because the two species have affinity for each other [20-22]. However, the question remains whether that particular adapter has precisely the right chemistry for sequencing DNA.

Experimentally, it was determined that 40 μM total bulk mononucleotide concentration causes ~ 30 blockades s^{-1} of the αHL channel modified with the β-cyclodextrin adapter [23]. An estimate of the collision rate between point particles in the bulk and pore mouth can be obtained using a relationship derived by Berg and Purcell[37]. That rate ($2 D d C_0$, where D, d, and C_0 are the mononucleotide diffusion coefficient, pore diameter, and mononucleotide concentration in the bulk, respectively) is ~ 58,000 blockades s^{-1}. Assuming that ~ 60 % of these particles reach the β-cyclodextrin adapter when V = 180 mV (Fig. 3), ~ 30,000 events per second should have been observed. Thus, one might conclude that only ~ 1 out of 1,000 particles that collide with the pore mouth reached the adapter. However, it is more likely that many blockade events are simply too short-lived to detect given the electronic measurement bandwidth in the nanopore apparatus (~ 10 kHz) and the additional digital filtering used to present the data (2 kHz) [23].

Indeed, the binding constants for the reactions between mononucleotides with β-cyclodextrin in the bulk are weak[20-22] (i.e., $K = k_{off}/k_{on}$ ~ 10 mM to 100 mM, where k_{off} and k_{on} are the dissociation and association rate constants, respectively). This suggests the reaction will have relatively short mean lifetimes ($\tau = 1/k_{off}$ ~ 0.1 μs to 1 μs, assuming $k_{on} \sim 10^9$ M^{-1} s^{-1}), consistent with experimentally measured values (τ = 0.75 μs to 2 μs[21,22]). If the binding constant for the complex in the pore is similar to that in the bulk, then most of the blockades would be too short-lived to resolve at 2 kHz bandwidth[23]. Because Clarke, et al. observed ca. 0.1 % of the events (*see above*), we estimate that the mean lifetimes are approximately 70-fold greater when the adapter is attached to the inside of the pore, assuming the residence times of the bases on the adapter are exponentially distributed. Finally, we note that the addition of one or more fixed positive charges on the adapter should facilitate considerably stronger binding with the mononucleotides, and thereby markedly increasing the observed apparent capture rate.

Analysis of the Exonuclease/nanopore-based DNA sequencing algorithm

Figure 5 illustrates the limitations on capturing and reading a single base cleaved by the exonuclease, but suggests that for a given electronic measurement bandwidth, we can select an optimum voltage to maximize the likelihood of correctly reading a single base with probability p. Assuming the probability of discriminating between the four bases is unity (but see Fig. 1) and that the system treats all bases alike, the probability of correctly sequencing the first m bases of a polynucleotide is p^m. However, because $p < 0.75$ (Fig. 5), accurately sequencing a single polynucleotide molecule in this manner is not practical. The statistics could of course be improved by reading many identical copies of the molecule by either the same nanopore or by an array of identical pores.

We next describe two algorithms to address the question of how many identical DNA copies must be measured to deduce the polynucleotide's sequence with a given degree of accuracy. We first adapt a "survivor" algorithm [38] in which a number of sequence measurements (i.e., subsequences), c, are aligned from one end, and an assignment of the sequence is based on the majority of assignments from each subsequence. On average, only cp subsequences survive the first "vote", and the remaining $c(1-p)$ subsequences are discarded. For the second letter in the sequence, the cp remaining subsequences will have, on average, cp^2 majority "votes" and the remaining $cp(1-p)$ subsequences are discarded. Therefore the average number of remaining subsequences after T "votes" is $N_T = cp^T$ and the maximum expected number of correctly sequenced letters $T_{\max}$ is

$$T_{\max} = \frac{\log\left(N_T / c \right)}{\log\left(p \right)} \quad . \tag{8}$$

Figure 6 shows the results of a numerical simulation using this sequencing algorithm. As an example, we generate N_{seq} subsequences, each with an average accuracy of 80 % ($p = 0.8$, assuming random deletions of letters, but no misreads) for a 100-mer DNA strand. We then calculate the probability of accurately reconstructing the original sequence using the algorithm described above. Finally, we fit the simulated results using eq. 8 with N_T as an adjustable parameter and a joint capture and discrimination probability of $p = 0.8$ (a slight overestimate of this sequencing engine's potential, Fig. 5).

The curve in Fig. 6 (*open squares*) highlights the drawbacks of discarding sequences, especially as the algorithm progresses. For 10^3 subsequences (obtained from reading identical copies using either a single nanopore or an array of nanopores), the amount of correctly read bases is only 25 %, which increases to 45 % and 80 % with the use of 10^5 and 10^8 nanopores, respectively (*not shown*). Given the increased uncertainty arising from discarding data, this method may not be amenable to accurately measuring long sequences. However, it may be possible to improve this approach by breaking the original sequence into a sufficient number of overlapping sequences of shorter lengths. These shorter strands could be sequenced and further analyzed with existing sequencing methodologies based on DNA subset overlapping.

The above algorithm is vastly improved by retaining information in the "discarded" subsequences. As before, if we assume that the errors in reading a given base are from deletions only (i.e., perfect discrimination between the bases), then at the end of one voting round we left-shift every subsequence that agrees with the majority and hold all other sequences at their previous positions. The results of numerical simulations show that this relatively simple modification permits the accurate reading of approximately 85 bases of a 100 mer's sequence (p = 0.8) with only ~ 200 subsequences (Fig. 6, *open circles*).

Less than perfect discrimination between nucleotides

Because the mononucleotide discrimination afforded by the β-cyclodextrin adapter is not perfect (Figs. 1A-1E), the probability of correctly reading DNA sequences with the exonuclease/nanopore-based technique would be proportionally lower than the estimates described above. How to achieve better discrimination with an internal molecular adapter is not yet clear. Further study into this problem would be helpful for both the separation of the four bases and the development of an adapter with sufficiently strong affinity for the mononucleotides.

Conclusions

An analysis of a proposed exonuclease/nanopore-based DNA sequencing engine illustrates that the system has significant merit, but several aspects of the technique need improvement. Our analysis shows that many bases will escape to the bulk before being detected in the proper sequence. The results also show that increasing the applied transmembrane potential, which also increases the potential just outside the pore, will increase the cleaved mononucleotide capture efficiency. However, the increased potential decreases the lifetime of a given mononucleotide on the molecular adapter, which will therefore decrease the probability of reading each base correctly (or even detecting them). The results also show that because of the relatively low probability of capturing a given base (~ 74 %), the number of DNA molecules that would need to be read scales exponentially with the number of bases. Based on the ability of a molecular adapter in the αHL nanopore [23] to discriminate between different mononucleotides (average accuracy = 92 %), the probability of correctly reading a given base is at best 68 %, which is likely insufficient for sequencing even moderate length strands of DNA.

Those conclusions notwithstanding, there are several possibilities for improving the exonuclease/nanopore-based DNA sequencing method. For example, the capture probability (P_{cap}) might be increased by orienting the exonuclease in such a manner that the cleaved mononucleotide is closer to the pore mouth. Conceivably, P_{cap} could also be increased by increasing the access resistance on the side that contains the exonuclease, but engineering that capability in a massively parallel array would need to be developed. Finally, the use of computer simulations may lead to the design of more accurate molecular adapters, if the required stereochemistry can be rationally designed and produced.

Acknowledgments

This work was supported in part by grants from NIST Office of Law Enforcement Standards (JJK, JWFR, JER), NRC/NIST-NIH Research Fellowship (AB) and the NSF (DLB). We thank Dr. James Russo at Columbia University for helpful comments on the manuscript.

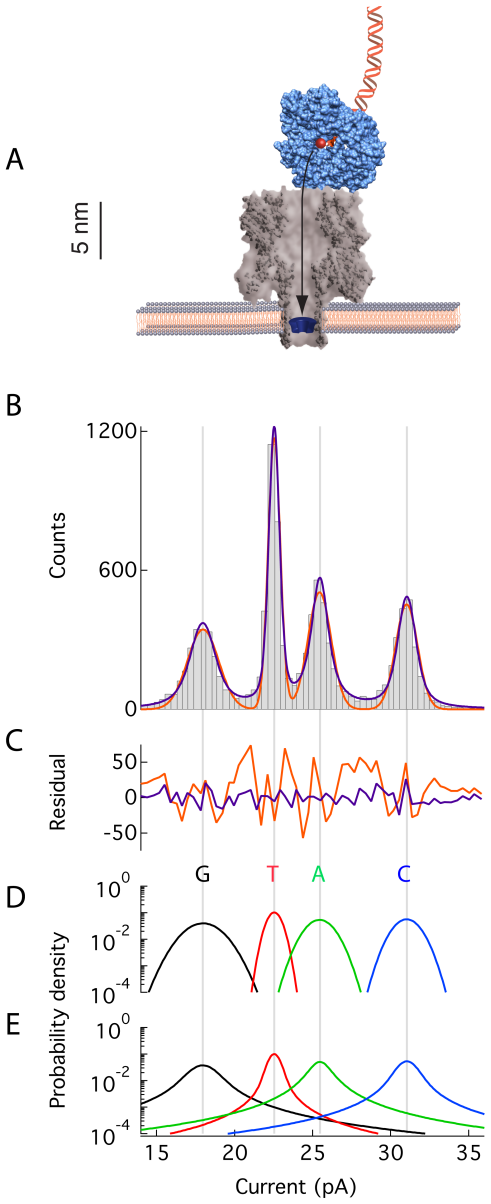

Figure 1. Electrical discrimination between mononucleotides by a β-cyclodextrin adapter in a nanopore [23]. **(A)** DNA sequencing engine [19] based on exonuclease attached to the cap domain of the α-hemolysin ion channel. The enzyme cleaves DNA (*red sphere*), one base at a time, near the pore mouth. In order to correctly sequence DNA, every cleaved base has to enter the pore sequentially, reach the base discrimination site (*dark blue*), and transport completely through the pore. **(B)** A histogram of the ionic current blockades caused by dGMP, dTMP, dAMP, and dCMP binding to a β-cyclodextrin adapter in the pore are fit with four Gaussian (*orange*) or Voigt (*indigo*) functions[24]. **(C)** The resultant fits demonstrate that the Voigt function better describes the data, and fully accounts for the overlap between neighboring peaks. The use of Gaussian functions overestimates the system's ability to separate the mononucleotides. The overlaps between neighboring fitted peaks for the Gaussian (**D**) and Voigt functions (**E**) are represented by their respective probability density functions on a log-linear scale. The fit of the Voigt function to the data shows that the accuracies of identifying the mononucleotides are G: 94 %, T: 93 %, A: 85 %, and C: 95 %, with an average accuracy of 92 %.

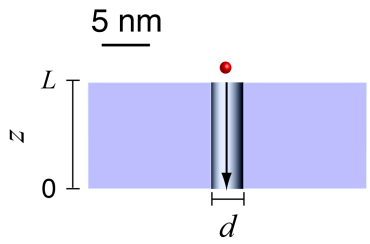

Figure 2. A schematic of the analytical model for the exonuclease/nanopore-based DNA sequencing engine. Each base (*red*) is released at a point above a pore of length L. The bulk solution is assumed to be infinite and the particle is captured at $z = 0$.

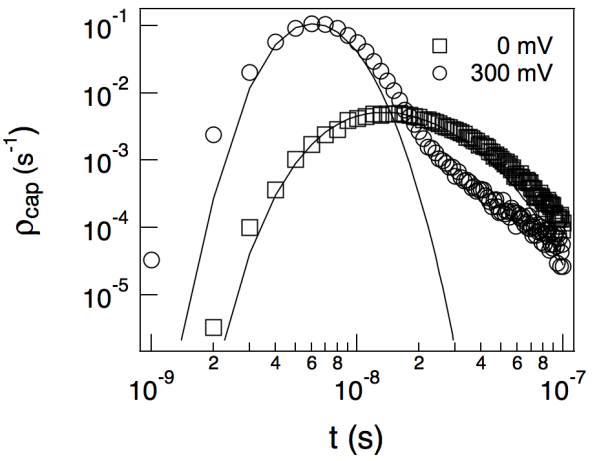

Figure 3. Distributions of capture times for an individual DNA base estimated from the inverse Laplace transform of eq. 4 (*solid lines*) and numerical simulations of a three-dimensional model of the system (*circles*, *squares*) for two different applied potentials, (V = 0 mV, *squares* and 300 mV, *circles*). The peak heights obtained with the analytical model were scaled to coincide with those estimated from the numerical simulations. Increasing the potential both increases the likelihood of capture and shifts the distribution to shorter capture times. A vast majority of capture events occur within the first 100 ns after cleavage of a base. The same parameters were used in both the analytical model (Eq. 4) and the numerical simulations. They are the base electrophoretic mobility ($\mu = -2.4 \times 10^{-4}$ cm^2/Vs), the distance between the pore entrance and the capture point (L = 6 nm), the single nanopore ionic conductance (g = 295 pS), the number of water molecules that hydrate mobile cations and anions (N_W = 10), the molar concentration of water ($c_W = 3.4 \times 10^{28}$ m^{-3}), the electronic charge ($e = 1.6 \times 10^{-19}$ C), the ratio of the nanopore's permeabilities to cations and anions ($P_+/P_- = 0.1$), the DNA base diffusion coefficient ($D = 3 \times 10^{-6}$ cm^2/s), and the nanopore diameter (d = 2.6 nm).

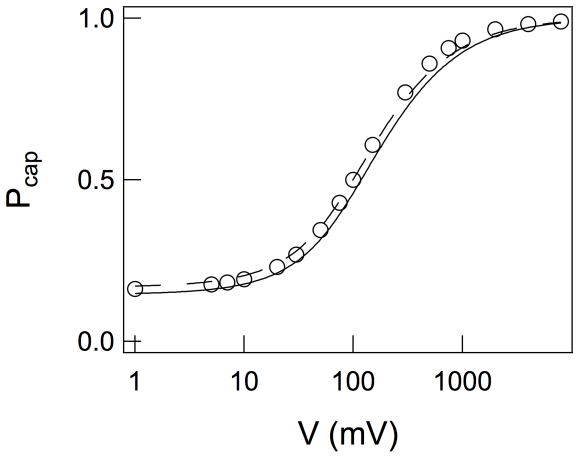

Figure 4. The capture probability as a function of the applied voltage estimated from simulations (*open circles*) and the analytical model (*solid line*), which has no adjustable parameters and uses the same values from Fig. 3 for the radiating boundary condition [29] $\kappa = 8D/\pi d; \beta = 8L/\pi d = 5.88$. A one-parameter least squares fit of Eq. 5 to the simulated results with $\beta = 4.7 \pm 0.1$ is also shown (*dashed line*). The slight disagreement between the simulation and theoretical model may result from the analytical model's assumption that the nanopore can be represented as a one-dimensional line.

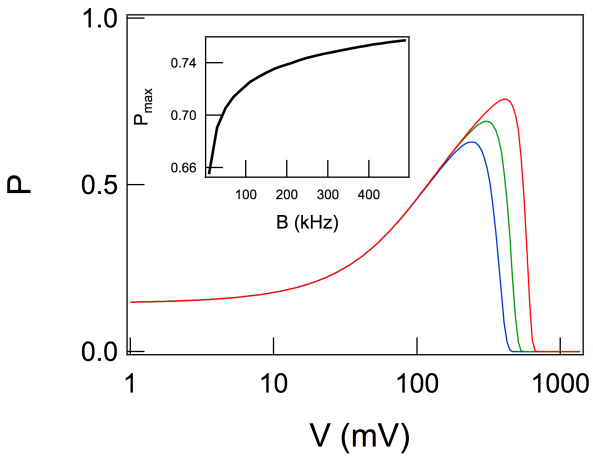

Figure 5. The probability of reading a single base correctly is dependent on the applied potential and the system bandwidth. At low applied potential, the capture process dominates the probability. At higher potential, the mean residence time of a base on the detector decreases exponentially, which significantly decreases the probability of identifying a given base. Increasing the system bandwidth (5 kHz *blue*; 30 kHz *green*; and 500 kHz *red*) increases the probability of correctly reading a base because it permits a greater potential to be used (which increases capture probability) and shorter-lived blockade events to be accurately measured. (*Inset*) Ignoring several sources of noise, the optimum read probability increases with increasing system bandwidth, but saturates at ca. 0.75.

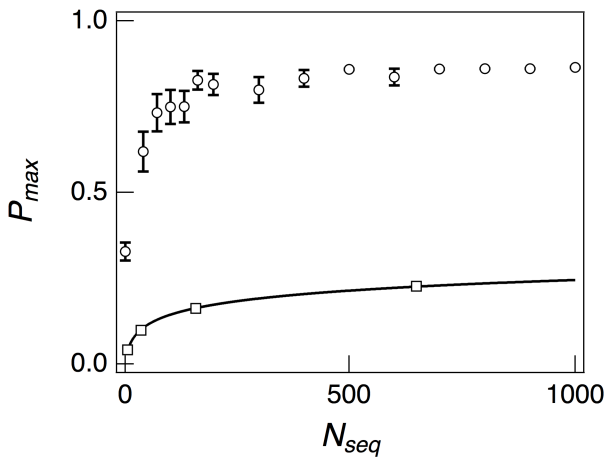

Figure 6. The probability of reading a DNA sequence correctly, P_{max}, as a function of the number of identical polynucleotides read, N_{seq}, evaluated using two algorithms described in the text. The results from the numerical simulations for the "survivor" algorithm (which discards a particular polynucleotide sequence measurement if its base assignment disagrees with that of the majority's consensus, *open squares*) and a "majority rules" algorithm (which makes use of the information in the minority strand sequence measurements, *open circles,* see text). In both cases, the probability of capturing a given nucleotide is assumed to be 0.8. The solid line is a least-square fit of eq. 8 (with $N_T = 4.3 \pm 0.2$) to the "survivor" algorithm simulation data. The "majority rules" simulation was performed 20 times for each N_{seq} to generate statistics. The error bars shown are the standard error.

Appendix

Drift velocity

The drift velocity of each base originates from electrophoretic and electroosmotic effects. The electrophoretic velocity of a charged particle in an electric field is $v_{ep} = \mu E$, where μ is the electrophoretic mobility.

A net electroosmotic velocity will occur in an ion selective nanopore. The net flux of ions through a nanopore under an applied potential is[39]

$$\text{(A1)} \qquad J_{\pm} = \frac{gV}{q}\left(\frac{P_+/P_- - 1}{P_+/P_- + 1}\right)$$

where $J_{\pm}$ is the net flux of cations or anions through the nanopore, g is the nanopore conductance, V is the applied transmembrane potential, q is the electric charge on each ion, and P_+/P_- is the nanopore ion selectivity ratio. Each ion is hydrated by a number of water molecules, N_W, as it flows through the pore. Therefore the flux of water molecules through the nanopore is given by $J_W = N_W J_{\pm}$ and the resulting net electroosmotic velocity v_{eo} of water through the nanopore $v_{eo} = J_w / c_w A_p$ where c_w is the molar concentration of water and A_p is the average cross-sectional area of the nanopore.

We assume the applied potential creates a uniform electric field inside the nanopore and zero field outside the pore,

$$\text{(A2)} \qquad V(z) = \begin{cases} \dfrac{Vz}{L}, & z = [0, L] \\ 0, & z > L \end{cases}$$

This condition is relaxed in another report, in which numerical simulations consider more complex electric field distributions. Nevertheless, the salient features of the electric field dependence can be understood here.

Combining the electroosmotic and electrophoretic terms leads to the following expression for the drift velocity in terms of the applied transmembrane

$$\text{(A3)} \qquad v_d = v_{ep} + v_{eo} = \left[\frac{\mu}{L} + \frac{gN_W}{c_W A_p e}\left(\frac{P_+/P_- - 1}{P_+/P_- + 1}\right)\right]V$$

For comparison to numerical simulations, we use the following parameters: $\mu = -2.4 \times 10^{-4}$ cm^2/Vs, $L = 7.5 \times 10^{-7}$ cm, $g = 2.95 \times 10^{-10}$ C/Vs, $N_W = 10$, $c_W = 3.4 \times 10^{22}$ cm^{-3}, $A_p = 5.3 \times 10^{-14}$ cm^2, $e = 1.6 \times 10^{-19}$ C, $P_+/P_- = 0.1$. Substitution of these values into eq. A3 leads to the following drift velocity voltage dependence,

$$\text{(A4)} \qquad v_d = (-328 \text{ cm/Vs})V$$

It is worth noting that the largest possible electroosmotic contribution that would enhance the drift of nucleotides through the nanopore occurs when $P_+/P_- = 0$ (i.e., purely anion selective). In this case, the prefactor in Eq. A4 would be slightly greater (-330 cm/Vs).

Details of the Model

$$\text{(A5)}\quad \hat{\phi}(s) = \int_0^\infty \exp(-st)\phi(t)\,dt = D\left.\frac{d\hat{G}(z,s)}{dz}\right|_{z=0}$$

In the given geometry, the drift velocity points towards $z = 0$ so $v < 0$. The Laplace transform of Eqs. 1 and 2 leads to

$$\text{(A6)}\quad s\hat{G}(z,s) - G(z,0) - |v_d|\frac{d\hat{G}(z,s)}{dz} = D\frac{d^2\hat{G}(z,s)}{dz^2}.$$

and

$$\text{(A7)}\quad -D\left.\frac{d\hat{G}(z,s)}{dz}\right|_{z=L} - |v_d|\hat{G}(L,s) = \kappa\hat{G}(L,s).$$

We solve for G and $\phi(s)$ subject to the absorption and radiating boundary conditions, setting $\hat{G}(L-\delta,s)_l = \hat{G}(L-\delta,s)_r$ and $\hat{G}'(L-\delta,s)_l = \hat{G}'(L-\delta,s)_r$ (where *l* and *r* refer to the limit from the left and right side of *L*-*s* and the prime refers to differentiation with respect to *z*), and taking the limit as $\delta \to 0$. This leads to the eq. 4 in the text.

Probability of discriminating mononucleotides

Each mononucleotide can be uniquely identified by the degree to which it blocks the ionic current in the pore (Fig. 1). The probability of identifying each nucleotide (A, T, G, or C) is calculated by fitting the data in the figure to four Voigt profiles (Igor Pro, WaveMetrics, Inc., Lake Oswego, OR, USA). Eqn. A8 gives the probability density of each peak, where *i* represents the blockade current, α is the amplitude, σ is the width, μ is the location of the peak and γ is a shape parameter[24].

$$\text{(A8)}\quad \Phi(i;\alpha,\sigma.\mu,\gamma) = \alpha\gamma\pi^{-3/2}\sqrt{\ln 2}\int_{-\infty}^{\infty}\frac{\exp(-t^2)}{\gamma^2 + \left[\sigma(i-\mu) - t\right]^2}\,dt$$

The probability of uniquely discriminating each base is given by the area under each peak in Fig 1B using the expression, $P_d = \int_{i_1}^{i_2}\Phi(i;\alpha,\sigma,\mu,\gamma)\,di$, with the limits of integration (i_1 and i_2) set to the points of intersection between neighboring peaks. The probabilities of uniquely identifying a mononucleotide are then calculated to be 0.94, 0.93, 0.85, and 0.95 for dGMP, dTMP, dAMP and dCMP, respectively. Finally, the average probability of uniquely discriminating a single mononucleotide is 0.92.

Numerical Simulations

The diffusion simulator uses bulk physical parameters (such as the analyte diffusion constant, electrophoretic mobility, nanopore selectivity, total transmembrane voltage, etc.) to perform a random walk in 3D, constrained by the protein and membrane volumes. No periodic boundary conditions are implemented, because each particle is allowed to translate until it is successfully captured, or a specified maximum run time is exceeded (typically 1 μs). Each diffusing species moves with a randomly determined cardinal direction and a step length determined by sampling from a Gaussian distribution with a mean of zero and a standard deviation (Eq. A9), *dx*, which is related to the characteristic diffusion coefficient [37].

(A9) $$dx = \sqrt{6Ddt}$$

As long as the simulation time step (*dt*) is small, translational motion inside the nanopore can be modeled accurately. A time step of 1 ps (step size = 40 pm) produces trajectories with adequate spatial resolution. Drift distance arising from electrophoretic and electroosmotic (EO) flow (Eq. A3) over the 1 ps time step is added to *dx* to give the particle displacement.

Nucleotides are simulated as point particles released from $z = 0$ (Figure 1) and reflect off the ion channel lumen walls. The top exterior of the nanopore extends radially to infinity and is also assumed to be reflective. Nucleotides freely migrate above the nanopore in bulk solution without geometric constraints. The capture probability is defined as the ratio of the number of successful nucleotide captures to the total number of nucleotides used in the simulation.

References

[1] A.M. Maxam and W. Gilbert, Proc. Natl. Acad. Sci. USA **74**, 560 (1977).
[2] F. Sanger, S. Nicklen, and A.R. Coulson, Proc. Natl. Acad. Sci. USA **74**, 5463 (1977).
[3] D.A. Wheeler, M. Srinivasan, M. Egholm, Y. Shen, L. Chen, A. McGuire, W. He, Y.-J. Chen, V. Makhijani, G.T. Roth, X. Gomes, K. Tartaro, F. Niazi, C.L. Turcotte, G.P. Irzyk, J.R. Lupski, C. Chinault, X.-Z. Song, Y. Liu, Y. Yuan, L. Nazareth, X. Qin, D.M. Muzny, M. Margulies, G.M. Weinstock, R.A. Gibbs, and J.M. Rothberg, Nature **452**, 872 (2008).
[4] J. Guo, L. Yu, N.J. Turro, and J. Ju, Acc. Chem. Res. **43**, 551 (2010).
[5] M. Margulies, M. Egholm, W.E. Altman, S. Attiya, J.S. Bader, L.A. Bemben, J. Berka, M.S. Braverman, Y.-J. Chen, Z. Chen, S.B. Dewell, L. Du, J.M. Fierro, X.V. Gomes, B.C. Godwin, W. He, S. Helgesen, C.H. Ho, C.H. Ho, G.P. Irzyk, S.C. Jando, M.L.I. Alenquer, T.P. Jarvie, K.B. Jirage, J.-B. Kim, J.R. Knight, J.R. Lanza, J.H. Leamon, S.M. Lefkowitz, M. Lei, J. Li, K.L. Lohman, H. Lu, V.B. Makhijani, K.E. McDade, M.P. McKenna, E.W. Myers, E. Nickerson, J.R. Nobile, R. Plant, B.P. Puc, M.T. Ronan, G.T. Roth, G.J. Sarkis, J.F. Simons, J.W. Simpson, M. Srinivasan, K.R. Tartaro, A. Tomasz, K.A. Vogt, G.A. Volkmer, S.H. Wang, Y. Wang, M.P. Weiner, P. Yu, R.F. Begley, and J.M. Rothberg, Nature **437**, 376 (2005).
[6] P. Bergveld, IEEE Trans. Bio-Med Eng. **BM17**, 70 (1970).
[7] P. Bergveld, Sensors Actuat. B **88**, 1 (2003).
[8] J.M. Rothberg, W. Hinz, T.M. Rearick, J. Schultz, W. Mileski, M. Davey, J.H. Leamon, K. Johnson, M.J. Milgrew, M. Edwards, J. Hoon, J.F. Simons, D. Marran, J.W. Myers, J.F. Davidson, A. Branting, J.R. Nobile, B.P. Puc, D. Light, T.A. Clark, M. Huber, J.T. Branciforte, I.B. Stoner, S.E. Cawley, M. Lyons, Y. Fu, N. Homer, M. Sedova, X. Miao, B. Reed, J. Sabina, E. Feierstein, M. Schorn, M. Alanjary, E. Dimalanta, D. Dressman, R. Kasinskas, T. Sokolsky, J.A. Fidanza, E. Namsaraev, K.J. McKernan, A. Williams, G.T. Roth, and J. Bustillo, Nature **475**, 348 (2011).
[9] K. Mullis, F. Faloona, S. Scharf, R. Saiki, G. Horn, and H. Erlich, Cold Spring Harb. Symp. Quant. Biol. **51 Pt 1**, 263 (1986).
[10] Y. Kong, J. Comput. Biol. **16**, 817 (2009).
[11] J.J. Kasianowicz, E. Brandin, D. Branton, and D. Deamer, Proc. Natl. Acad. Sci. USA **93**, 13770 (1996).
[12] J.J. Kasianowicz, J.W.F. Robertson, E.R. Chan, J.E. Reiner, and V.M. Stanford, Annu. Rev. Anal. Chem. **1**, 737 (2008).
[13] D. Lubensky and D. Nelson, Biophys. J. **77**, 1824 (1999).
[14] C.T.A. Wong and M. Muthukumar, J. Chem. Phys. **128**, 154903 (2008).
[15] M. Muthukumar, *Polymer Translocation* (CRC, Boca Raton, FL, 2011).
[16] C.T.A. Wong and M. Muthukumar, J. Chem. Phys. **133**, 045101 (2010).
[17] A.S. Panwar and M. Muthukumar, J. Am. Chem. Soc. **131**, 18563 (2009).
[18] T.Z. Butler, M. Pavlenok, I.M. Derrington, M. Niederweis, and J.H. Gundlach, Proc. Natl. Acad. Sci. USA **105**, 20647 (2008).
[19] D. Branton, D.W. Deamer, A. Marziali, H. Bayley, S.A. Benner, T. Butler, M. Di Ventra, S. Garaj, A. Hibbs, X. Huang, S.B. Jovanovich, P.S. Krstic, S. Lindsay, X.S. Ling, C.H. Mastrangelo, A. Meller, J.S. Oliver, Y.V. Pershin, J.M. Ramsey, R. Riehn, G.V. Soni, V. Tabard-Cossa, M. Wanunu, M. Wiggin, and J.A. Schloss, Nature Biotech. **26**, 1146 (2008).
[20] C. Formoso, Biochem. Biophys. Res. Comm. **50**, 999 (1973).
[21] P. BritzMcKibbin and D. Chen, J. Chromatogr. A **781**, 23 (1997).
[22] J.-R. Bae and C. Lee, Bull. Korean Chem. Soc. **30**, 145 (2009).

[23] J. Clarke, H.-C. Wu, L. Jayasinghe, A. Patel, S. Reid, and H. Bayley, Nature Nanotech. **4**, 265 (2009).
[24] B. Armstrong, J. Quantitat. Spectroscop. Rad. **7**, 61 (1967).
[25] S. Chandrasekhar, Rev Mod Phys **21**, 383 (1949).
[26] M. Weissbluth, *Photon-Atom Interactions* (Academic Press, 1989), pp. 1–37.
[27] H. Risken, *The Fokker-Planck Equation: Methods of Solution and Applications* (Springer, New York, NY, 1996).
[28] A.M. Berezhkovskii, A.V. Barzykin, and V.Y. Zitserman, J. Chem. Phys. **130**, 245104 (2009).
[29] S. Bezrukov, A. Berezhkovskii, M. Pustovoit, and A. Szabo, J. Chem. Phys. **113**, 8206 (2000).
[30] M. Winterhalter and W. Helfrich, Phys. Rev., A **36**, 5874 (1987).
[31] C. Wilhelm, M. Winterhalter, U. Zimmermann, and R. Benz, Biophys. J. **64**, 121 (1993).
[32] K. Klotz, M. Winterhalter, and R. Benz, Biochim. Biophys. Acta **1147**, 161 (1993).
[33] M. Winterhalter, K.H. Klotz, R. Benz, and W.M. Arnold, IEEE Trans. on Ind. Applicat. **32**, 125 (1996).
[34] J.E. Hall, J. Gen. Physiol. **66**, 531 (1975).
[35] S.M. Bezrukov, I. Vodyanoy, R. Brutyan, and J.J. Kasianowicz, Macromolecules **29**, 8517 (1996).
[36] A. Aksimentiev and K. Schulten, Biophys. J. **88**, 3745 (2005).
[37] H. Berg and E. Purcell, Biophys. J. **20**, 193 (1977).
[38] V.I. Levenshtein, J. Combin. Theory A **93**, 310 (2001).
[39] L.-Q. Gu, S. Cheley, and H. Bayley, Proc. Natl. Acad. Sci. USA **100**, 15498 (2003).