

# TRANSPORT OF DNA THROUGH A SINGLE NANOMETER-SCALE PORE: EVOLUTION OF SIGNAL STRUCTURE

Vincent M. Stanford<sup>1</sup> and John J. Kasianowicz<sup>2</sup>

<sup>1</sup>NIST, Information Access Division, 225/A231, Gaithersburg, MD 20899-8940

<sup>2</sup>NIST, Biotechnology Division, 227/A251, Gaithersburg, MD 20899-8313

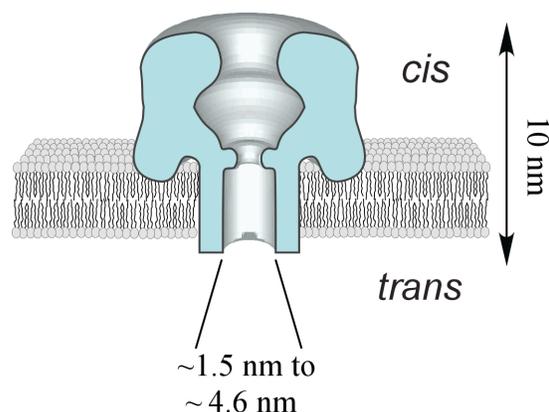
## ABSTRACT

Single-stranded DNA can be driven through a single nanometer-scale pore. This process causes the ionic current that otherwise flows through the pore to decrease for characteristic times that a polynucleotide and the pore interact. We previously reported a method for characterizing these signals using ergodic, but persistent Hidden Markov Models (HMMs). Gaussian mixture models (GMMs) were used as output distributions to obtain a maximum likelihood estimate of state sequence and lifetime. We show here that a more economical state description can be applied to signals from shorter lifetime events. The results are consistent with the known structure of the nanopore and may suggest approaches to more detailed interrogation of the information stored in DNA and other polymers.

## 1. INTRODUCTION

Nature has adapted a simple structural motif, a nanometer-scale pore, for transporting ions and molecules across the membranes of cells and organelles [1]. These “ion channels” and similar structures perform a wide variety of tasks including inter- and intra- nerve communication, the control of muscle action, protein secretion, viral infection, and bacterial gene transduction.

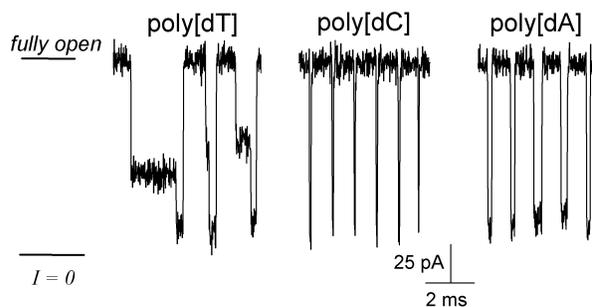
We have adapted a single ionic channel (see Fig. 1) as a model system for devising novel sensor technologies and to better understand how DNA is transferred between organisms. Towards these goals, we recently demonstrated that individual single stranded DNA molecules can be detected and characterized as they thread through a single channel. Research on this subject might aid the development of ultra-rapid DNA sequencing technology, of more effective antiviral agents and genetic therapies.



**Figure 1.** Schematic representation of the  $\alpha$ -hemolysin ion channel ( $\alpha$ HL) that spontaneously forms a nanometer-scale pore across a lipid bilayer membrane. The *cis* and *trans* sides denote the two different domains of the channel.

## 2. DNA TRANSPORT IN A SINGLE $\alpha$ HL PORE

The nanometer-scale pore formed by the bacterial toxin  $\alpha$ -hemolysin ( $\alpha$ HL) provides the basis for the design of modern biosensors with single molecule resolution [2,3]. More recently, it was shown that an electric field can drive single-stranded DNA through this pore [4]. The passage of individual molecules through the pore causes transient, but characteristic blockades in the ionic current (Fig. 2).

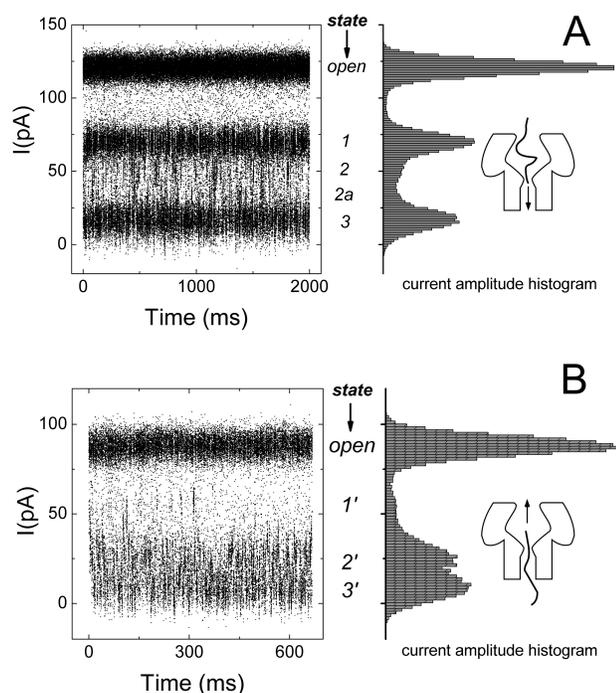


**Figure 2.** Blockades in the ionic current through the  $\alpha$ HL channel caused by individual molecules of 100 base long poly(thymine), poly(cytosine) and poly(adenosine). From [3].

The data in Fig. 2 suggest that the blockade lifetime, amplitude and patterns might be “fingerprints” for identifying nucleic acid polymers from these electronic signatures. We discuss here the statistical description of these blockades and show that for short lifetime events caused by one of these polymers (100 nucleotide-long poly[thymine], poly[dT]<sub>100</sub>), the number of states that describe the data is relatively small and may relate to the simple structure of the nanopore itself.

### 3. DESCRIPTION OF CURRENT BLOCKADES

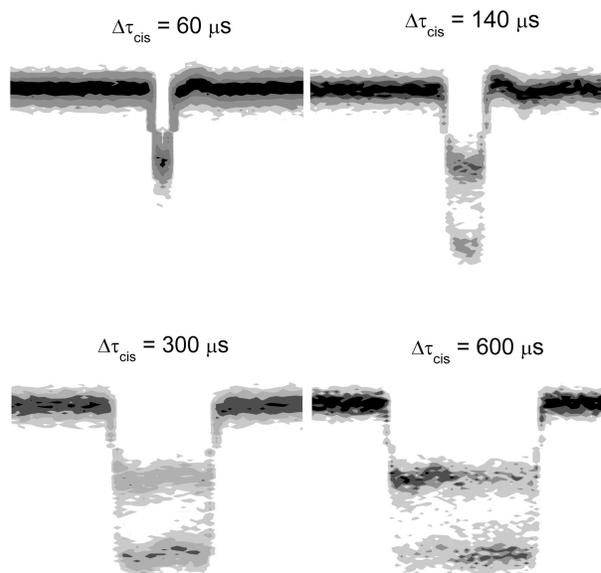
We previously reported a detailed statistical analysis of a large population of blockades caused by poly[dT]<sub>100</sub> [5]. The lifetimes of these events ranged from 40  $\mu$ s to 2 full seconds. The substate structure of this aggregated set required a large number of Gaussian mixture components for a full description. Here, we disaggregate the events into subsets with limited lifetime ranges. This reduces the number of substates to describe the signal amplitude distribution. For example, for events with lifetimes between 40  $\mu$ s to 2 ms, the blockades caused by the polymer transits can be described by 4 states (1 open and 3 occluded; Fig. 3A).



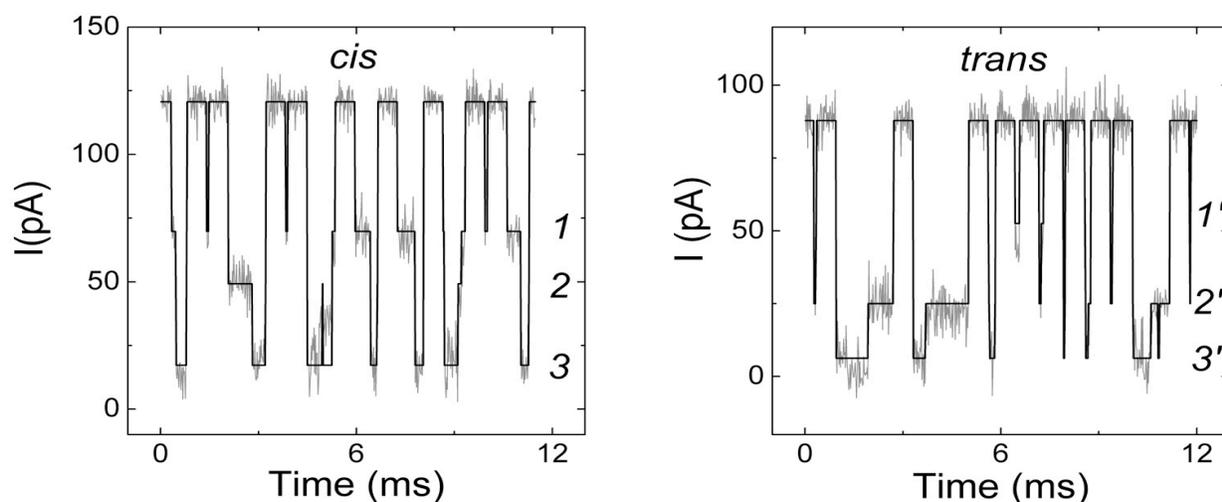
**Figure 3.** Time series for a set of many poly[dT]<sub>100</sub> transit events with lifetimes between 40  $\mu$ s to 2 ms (*left*) and current amplitude distribution (*right*) for polymers entering the pore from either the **A) cis** or **B) trans** pore entrances. The applied potential was -120 mV or +120 mV, respectively.

The condensed time-series representation in Fig. 3 obscures details of the poly[dT]-induced current blockades (Fig. 2), but makes it easier to visually detect the most probable occluded states. For example, the 3 darkest bands in the current recording (*left*) and the largest peaks in the current amplitude histogram (*right*) correspond to the fully open state and two most probable occluded current states (states 1 and 3). The limited number of occluded states may represent the polymer negotiating through regions of the pore with different diameters.

The ratios of the mean current values for each of the three most probable occluded states and the respective mean open channel current for poly[dT]<sub>100</sub> entering the pore from the *cis* side (Figure 3A) do not appear to correspond to those for the three most probable occluded states for polymers threading the pore in the opposite direction (Figure 3B).



**Figure 4.** The morphology of poly[dT]<sub>100</sub>-induced current blockades evolves with increasing event lifetime ( $\Delta\tau_{cis}$ ) for  $60 \mu$ s  $\leq \Delta\tau_{cis} \leq 600 \mu$ s. The single channel current time series for ensembles of events at four different lifetimes are aligned at the onset of each channel blockade. The shades of gray indicate the distribution of signal amplitude vs. time.

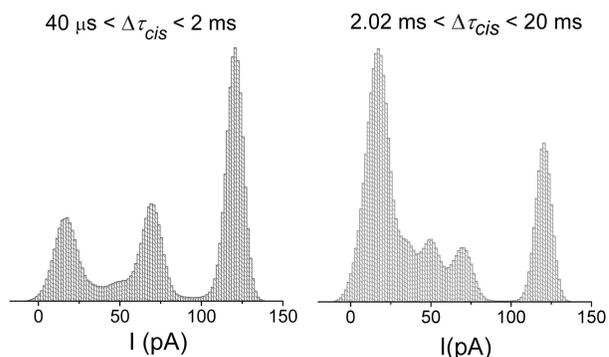


**Figure 5.** Subsets of single channel current blockades for poly[dT]<sub>100</sub> entering the *cis* and *trans*  $\alpha$ HL pore entrances and the maximum likelihood estimate of the amplitude state sequence (lines) superimposed on the data (gray). State sequences are determined from the data using Viterbi decoding [6] of the dwell times.

The poly[dT]<sub>100</sub> duration current blockade ensembles for  $\Delta\tau_{cis} \leq 600 \mu\text{s}$  evolve as follows. The 60  $\mu\text{s}$  duration events are described by a single occluded state (i.e., state 1, Figure 3A). The 140  $\mu\text{s}$  and 300  $\mu\text{s}$  lifetime event ensembles show predominantly a bimodal density of occluded states (i.e., states 1 and 3 in Figure 3A). Note that the channel blockades are virtually always either in either state 1 or state 3. The longer event duration ensemble (i.e.,  $\Delta\tau_{cis} = 600 \mu\text{s}$ ) also shows the bimodal pattern observed in the 140  $\mu\text{s}$  and 300  $\mu\text{s}$  long blockades. However, the evolution of a third event pattern, a transition from state 1 to state 3 within single events, is readily apparent.

For polymers entering the pore from the *cis* side of the  $\alpha$ HL channel (Figure 5A), the blockade events show the three common patterns shown in Figure 4, i.e., state 1 only, state 3 only, and intra-event transition from state 1 to state 3). There is also an example of a relatively rare event comprised of a transition from state 2 to state 3. A corresponding class of events occurs when the polymer enters the *trans* side (Figure 5B). However, note that there are transitions from a greater occluded state to a lesser occluded one (i.e., exactly opposite the two-step pattern observed when the polymer enters the pore from the *cis* entrance). This mirror-asymmetry of the step transition between two occluded states suggests that there is an asymmetry in polymer-pore interaction that is made manifest by the direction the polymer travels through the  $\alpha$ HL channel. The small number of current blockade states and patterns also suggests that a simple description of the blockade mechanism at short event lifetimes (i.e.,  $\Delta\tau_{cis} \leq 600 \mu\text{s}$ ) may be valid.

A large number ( $\sim 2 \times 10^5$ ) of poly[dT]<sub>100</sub> events, from both *cis* and *trans* directions provided an interesting and potentially important point of departure for the use of polymers as molecular rulers for the ion channel. The state structure of the amplitude distribution evolves with event lifetime (e.g., Fig. 6).



**Figure 6.** Event amplitude state structure varies with the event lifetime for poly[dT]<sub>100</sub> entering the *cis* side of the  $\alpha$ HL channel. Transit events exist over a wide range, from 20  $\mu\text{s}$  to over 2 s; with 99% percent of the event lifetimes < 5.5 ms. Note that longer lifetime events are in deeper blockade levels and that the deeper blockades require more Gaussian states to provide an adequate description of the data.

Understanding how polynucleotides interact with specific segments within a given nanopore suggests that the more full blockade states provide the best *coign of vantage* for higher resolution measurements of the ssDNA strands themselves. The *ansatz* here is that the deeper blockade

levels represent the interaction of the polymer with the narrowest constrictions in the nanopore channel.

#### 4. STATISTICAL METHODS

We previously demonstrated that statistical signal-processing can be used to estimate and measure substates and their lifetimes in the DNA-induced current blockades [4,5,7]. In many blockade events, the single channel current is piecewise stationary, with substates that overlap in amplitude but not in time. This signal structure prompted the use of an ergodic Hidden Markov Model (HMM) architecture for state sequence estimation. Statistical characterizations of the DNA-induced  $\alpha$ HL channel current blockades were based on Gaussian Mixture Models (GMM) and an Expectation Maximization (EM) procedure [8,9]. The latter provides a set of Gaussian components and mixture weights from a large population of events. The amplitudes of piecewise stationary segments, and their lifetimes, can subsequently be estimated with Viterbi decoding [6], which provides a maximum likelihood state sequence as the probable generating function for the observed time series.

While the state sequence itself is globally optimal, if given the correct output distributions, it is well known that the EM GMM estimator can arrive at local likelihood maxima that are globally suboptimal [10,11]. However, a good statistical criterion can mitigate the practical impact of this limitation, with random initial conditions being explored until no further improvements in the likelihood, or goodness of fit criteria are forthcoming. The well-known Kolmogorov-Smirnov (KS) statistic offers one such stopping criterion. The KS statistic is computed for each candidate amplitude mixture model, and additional components added if the p-value of the model fit could be rejected at the 0.05 level.

Because the GMM components usually overlap, simply assigning each point to the highest likelihood component results in many physically implausible transitions between states that overlap in amplitude but not time. To model the time coherence in the signal states implied by the passage of polynucleotides through the pore, and confirmed by observation, we employed the GMM components as the output distributions of an ergodic HMM but with a state transition matrix favoring state persistence, meaning it has a dominant diagonal with much smaller probabilities off-diagonal. Figure 5 illustrates some of the results from this decoding procedure and the identified state sequences. These state sequences also provide lifetime and amplitude measurements for further analyses, such as discriminators for molecule types, or automated extraction of structural information from individual molecules.

Also, an unconstrained maximum likelihood GMM can result in distribution states that are highly heteroscedastic (i.e., a set of statistical distributions having different variances), which can therefore result in unrealistic decoding of the state sequences. One solution is to constrain the variances to equality during the GMM EM estimation procedure. This offers a better approximation to the actual physical phenomenon of a nanometer-scale pore that is occluded by a comparable-sized polymer than does unconstrained variances for the states.

#### 5. REFERENCES

- [1] B. Hille. Ionic channels of excitable membranes, 2nd ed. Sinauer Associates, Sunderland, MA. (1992)
- [2] S.M. Bezrukov and J.J. Kasianowicz. "Current noise reveals protonation kinetics and number of ionizable sites in an open protein ion channel". *Phys.Rev. Lett.* **70**, 2352-2355. (1993)
- [3] J.J. Kasianowicz, S.E. Henrickson, H.H. Weetall and B. Robertson. "Simultaneous multianalyte detection with a nanopore". *Analytical Chemistry* **73**, 2268-2272. (2001)
- [4] J.J. Kasianowicz, E. Brandin, D. Branton and D.W. Deamer. "Characterization of individual polynucleotide molecules using a membrane channel". *Proc. Natl. Acad. Sci. (USA)* **93**, 13770-13773. (1996)
- [5] J.J. Kasianowicz, S.E. Henrickson, M. Misakian, H.H. Weetall, B. Robertson and V. Stanford. "Physics of DNA threading through a nanometer pore and applications to simultaneous multianalyte sensing". NATO Advanced Research Workshop. Structure and Dynamics of Confined Polymers. Kluwer Press. Eds. J.J. Kasianowicz, M.S.Z. Kellermayer and D.W. Deamer. pp. 141-163. (2002)
- [6] A.J. Viterbi. "Error bounds for convolutional codes and an asymmetrically optimum decoding algorithm". *IEEE Transactions on Information Theory* **IT-13**, 260-267. 1967.
- [7] V. Stanford and J.J. Kasianowicz. "On quantifying signatures from single-stranded DNA driven through nanometer-scale  $\alpha$ -hemolysin ion channels using Hidden Markov Models". *IEEE Workshop on Genomic Signal Processing and Statistics*. October 11-13, 2002. Raleigh, NC
- [8] R. Redner and H. Walker. "Mixture densities, maximum likelihood and the EM algorithm". *SIAM Review* **26**, 195-239. (1984)
- [9] L. Xu, and M. Jordan. "On convergence properties of the EM algorithm for Gaussian mixtures". *Neural Computation* **8**, 129-151. (1996)
- [10] L. Baum. "An inequality and associated maximization technique in statistical estimation of probabilistic functions of a Markov process". *Inequalities* **3**, 1-8. (1972)
- [11] L.E. Baum and T. Petrie. "Statistical inference for probabilistic functions of finite state Markov chains". *Annals of Mathematical Statistics* **37**, 1559-1563. (1966)