Nanopore detection using channel current cheminformatics

Stephen Winters-Hilt*[1,2]
[1]Department of Computer Science, University of New Orleans, New Orleans, LA 70148
[2]Research Institute for Children, Children's Hospital, New Orleans, LA 70118

ABSTRACT

A novel detector is used for analysis of single DNA molecules. The detector is based on current blockade measurements through a single, nanometer-scale, α-hemolysin ion channel. The biologically based α-hemolysin channel self-assembles in lipid bilayers, permitting highly reproducible experiments with Angstrom resolution. In previous work[1-3] the spectrum of dsDNA blockade states could be explained in terms of the dsDNA-protein binding kinetics, and dsDNA terminus fraying (bond dissociation) kinetics. Results presented here strengthen the hypothesis that conformational dynamics can be observed as well, when the channel-captured dsDNA end is in an unbound state. *Feature discovery methods:* include a time-domain finite state automaton (FSA), a wavelet domain FSA, and a Hidden Markov Model (HMM). *Classifier feature extraction methods:* includes a time-domain FSA for signal acquisition and a generalized HMM with EM for features extraction. *Classification method:* Support Vector Machines (SVMs) are used with novel kernel designs. *Kinetic feature extraction tool*: a time-domain FSA projects current observations to a (small) set of blockade states. Those states are provided by the generalized HMM analysis. Noise sources limit the resolution of the nanopore device, and its multiclass scaling capabilities, and this is discussed in the context of ongoing refinements to the device.

Keywords: nanopore detector, alpha-hemolysin ion channel, single molecule biophysics, hidden Markov model, support vector machine, automated kinetic analysis

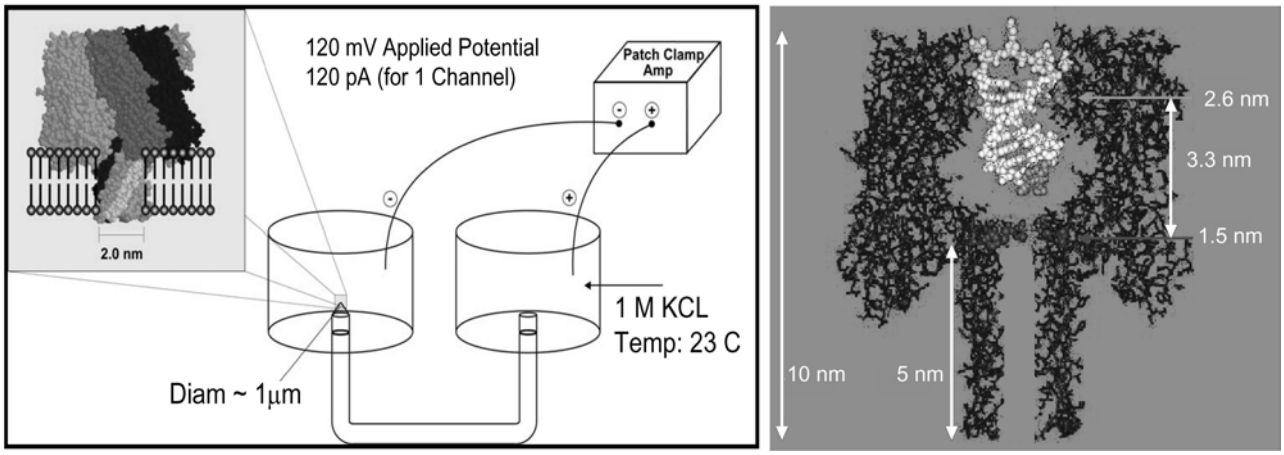
1. INTRODUCTION

Nanopore cheminformatics provides a new method for biophysical and biochemical analysis. Single biomolecules, and the ends of biopolymers such as DNA, can now be examined in solution with nanometer-scale precision[1-3]. This is done by using a nanometer-scale channel to associate ionic current measurements with single-molecule channel blockades. A biologically based (protein) channel, α-hemolysin, is used for this purpose because it consists of a solution soluble monomer that easily self-assembles in membranes as a heptamer channel[4,5]. This leads to an inexpensive and reproducible nanopore detector (see Fig. 1.a). α-hemolysin is also chosen because it is stable (e.g., non-gating) and has dimensions well suited to DNA/RNA measurement: ssDNA translocates while dsDNA does not, being held in the channel's *cis*-side vestibule instead. Figure 1.b shows the crystal structure[4], with the 1.5 nm limiting aperture (ringed by Glutamic Acids and Lysines) shown darker. The entry aperture on the *cis*-side is 2.6 nm in diameter (ringed by Threonines), which is large enough to admit (and capture) dsDNA.

Single channel observations have been performed by biophysicists and medical researchers since the 1970's using sensitive amplifiers to detect picoampere changes in ionic current[6,7]. Typically these single channel techniques are applied to low conductance, ion selective, channels, such as gated potassium channels, but they have also been applied to larger channels involved in metabolite and macromolecule transport. In 1994, Bezrukov and coworkers[8] broke new ground and used one of these large channels (alamethicin) to detect current impedance caused by a polymer (polyethylene-glycol). Their work proved that resistive pulse measurements, familiar from cell counting with a Coulter counter[9], could be reduced to the molecular scale and applied to polymers in solution. A seminal paper by Kasianowicz et al.[10] then showed that *individual* DNA and RNA polymers could be detected via their translocation blockade of a nanoscale pore formed by α-hemolysin toxin.

* winters@cs.uno.edu; http://www.cs.uno.edu/~winters; Children's Hospital, 200 Henry Clay Ave, Research & Education, New Orleans, LA 70118; (504) 896-2761; fax: (504) 894-5379

A Coulter Counter is a channel flow measuring device that is designed to count bacterial cells[9]. Transport of cells through the device is driven by hydrostatic pressure -- where interactions between the cells and the wall of the channel can be ignored. The α-hemolysin nanopore, on the other hand, strongly interacts with translocating biomolecules. The α-hemolysin channel is a protein heptamer, formed by seven identical 33 kD protein molecules secreted by *Staphylococcus aureus*. The total channel length is 10 nm and is comprised of a 5 nm *trans*-membrane domain and a 5 nm vestibule that protrudes into the aqueous *cis* compartment[4]. The narrowest segment of the pore is a 1.5 nm-diameter aperture[4]. By comparison, a single strand of DNA is about 1.3 nm in diameter. Given that water molecules are 0.15 nm in diameter, this means that one hydration layer separates ssDNA from the amino acids in the limiting aperture. This places the charged phosphodiester backbone, hydrogen bond donors and acceptors, and apolar rings of the DNA bases within one Debye length (3 Å in 1 M KCl) of the pore wall. Not surprisingly, DNA and RNA strongly interact with the α-hemolysin channel.

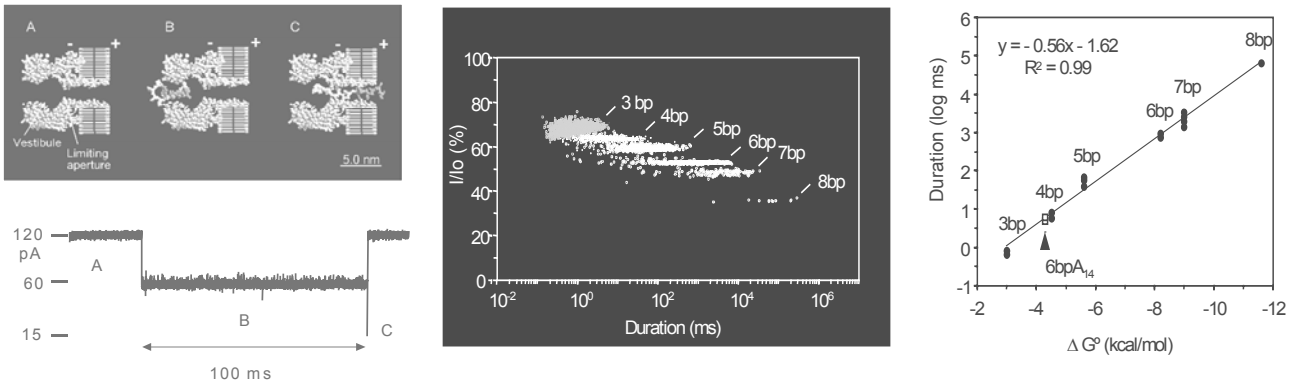


(a) **(b)**

Figure 1. The Nanopore Detector. Panel (a) shows the electrochemistry setup for the nanopore device. Panel (b) shows the crystallographic description of the α-hemolysin channel with a nine base pair DNA hairpin superimposed.

A nanometer-scale, α-hemolysin based, channel detector, or "nanopore detector," can be used to observe single ssDNA molecules during channel translocation[10-13], or to observe the ends of single dsDNA molecules captured by the pore[1-3]. In 1.0 M KCl (pH 8.0), a 120 mV applied potential (standard conditions) produces a steady open channel current (I_o) of 120 ± 5 pA at 23 °C (Fig. 2a).Translocation of single-stranded linear DNA (~ 1.3 nm diameter) reduces the current to I ≅ 14 pA (I/I_o = 12%) (Fig. 2a). Each monomer within single stranded DNA traverses the length of the 10-nm pore in 1 to 3 μs at ambient temperature. For the α-hemolysin nanopore detector, progress analyzing ssDNA translocations has been limited, however, due to the high speed of such translations. Lowering the applied potential to slow the ssDNA translocation doesn't trivially solve the problem either, since a minimal applied potential is required to draw molecules into the channel, i.e., a free energy barrier must be overcome (Meller and Hendrickson). Although it is possible to capture ssDNA at higher voltage, then translocated the ssDNA with a smaller applied voltage, this requires more sophisticated electronics[13] and will not be discussed further.

DNA and RNA hairpins were chosen as model duplexes because they can be formed from short, highly pure oligonucleotides that can be designed to adopt one base-paired secondary structure in 1.0 M salt at room temperature. The initial experiments involved a well-characterized DNA hairpin with a six-base-pair stem and a four-deoxythymidine loop. When captured within an α-hemolysin nanopore, this molecule caused a partial current blockade (or 'shoulder') lasting hundreds of milliseconds (Figure 2a, bottom panel) followed by a rapid downward spike. The shoulder-spike explanation was tested using a series of blunt-ended DNA hairpins with stems that ranged in length from 3 to 8 base-pairs. Each base pair addition resulted in a measurable increase in blockade shoulder lifetime that correlated with the calculated ΔG° of hairpin formation (Figure 2c). A downward trend in shoulder current amplitude was also observed from I/I_o equal to 68% for a 3 bp stem to I/I_o equal to 32% for a 9 bp stem. These results are consistent with greater obstruction of ionic current as the hairpin stem extends further into the vestibule with each additional base pair.

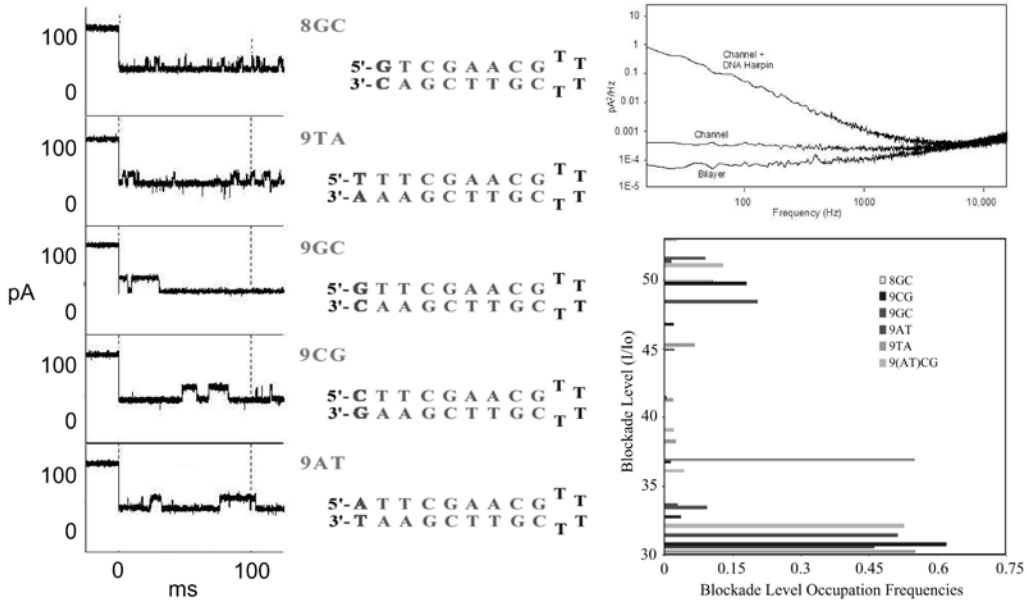


(a) **(b)** **(c)**

Figure 2. Panel (a). Impedance of ionic current through the α-hemolysin pore by a 6bp DNA hairpin. The current trace is shown in the lower panel. Each letter corresponds to a diagram in the top (a) sub-panel. A) Open channel current of ~120 pA. B) Capture of the hairpin molecule in the vestibule reduces the current to ~55 pA. The data indicates that the hairpin loop is perched at the mouth of the vestibule with the stem inside the vestibule (as shown). C) When the duplex stem dissociates, the applied electric field pulls the, now, single stranded DNA through the limiting aperture causing a transient spike to about 15 pA residual current. Panel (b) shows the influence of hairpin stem length on current impedance. In the panel (b) plot, each point represents the amplitude and duration for translocation of one DNA hairpin molecule. The duplex stems ranged from 3bp to 8bp. Panel (c) shows the average blockade durations plotted as a function of duplex hairpin stability in kcal mol. (Calculated using 'Mfold'[14]. ''6bpA14' is a 6 bp hairpin with an A•A mismatch.)

1.1 Channel Current Cheminformatics

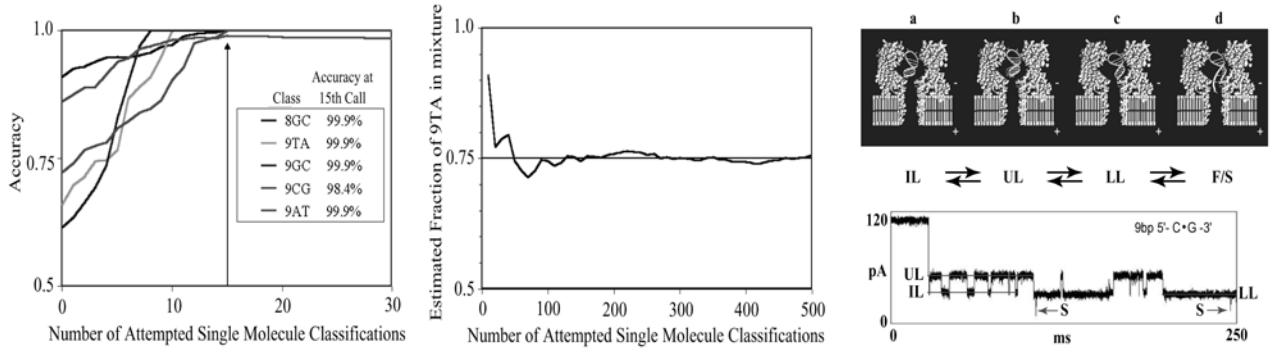
The cheminformatics classifiers train on channel blockade data, and do so without assuming anything about the nature of the channel (e.g., biological or semiconductor) other than it be a stable (non-gating) channel. As such the software solution is adaptable to different channel environments and can easily generalize to counter sufficiently slow drift in the channel configuration itself (for semiconductor channels).



(a) **(b)**

Figure 3. The channel current blockade signals observed in the prototype study. Panel (a) shows the five DNA hairpins, with sample blockades, that were used to test the sensitivity of the nanopore device. Panel (b), top, shows the power spectral density for signals obtained from the different nine base pair DNA hairpins, as well as the open channel. Panel (b), bottom, shows the dominant blockades, and their frequencies, for the different hairpin molecules in panel (a).

Five DNA hairpins were used in a prototype study to explore the sensitivity of the detector[2], as well as to probe the pore geometry (see Discussion). The DNA hairpins studied (see Fig. 3a) differed only in their terminal base pairs. HMM/EM characterization on the five classes of hairpin signatures revealed the existence of two major conductance blockade levels, one minor level intermediate between them, and one to three other statistically relevant levels depending on the hairpin (Fig. 3b, bottom). By examining the transition probabilities between the various levels it was found that blockades typically began in the less common intermediate level and from there almost always transitioned to the UL blockade level. Classification accuracy was 99.6% on average for the five DNA hairpins (see Fig. 4.a), and this was accomplished by the fifteenth classification attempt (6 seconds on average). The classification result for a mixture solution of 9TA and 9GC hairpin species (Fig. 4.b) is shown as the number of single molecule samplings is increased. The mixture was in a 3:1 ratio of 9TA:9GC, consistent with the 75% asymptote. Less than 1% error (on majority population size) was obtained by the hundredth observation. The molecular mechanism underlying the DNA hairpin's level transitions[3] is described in Fig. 4c



(a) **(b)** **(c)**

Figure 4. Panel (a) shows the single-species prediction accuracy as the number of signal classification attempts increases (allowing increase in the rejection threshold). Panel (b) shows the prediction accuracy on 3:1 mixture of 9TA to 9GC DNA hairpins. Panel (c) shows the nine base pair DNA hairpin blockade mechanism: When a 9bp DNA hairpin initially enters the pore, the loop is perched in the vestibule mouth and the stem terminus binds to amino acid residues near the limiting aperture. This results in the IL conductance level. When the terminal base pair desorbs from the pore wall, the stem and loop may realign, resulting in a substantial current increase to UL. Interconversion between the IL and UL states may occur numerous times, or UL may convert to the LL state. The LL state corresponds to binding of the stem terminus to amino acids near the limiting aperture but in a different manner from IL. From the LL bound state, the duplex terminus may fray with rapid extension and capture of one strand in the pore constriction (denoted F/S).

Angstrom precision structures for numerous DNA, RNA, and protein molecules have been revealed by X-ray diffraction analysis and NMR spectroscopy. These approaches rely upon average properties of very large numbers of molecules and are often biased towards crystallization and NMR conformer structures different from those present in solution under physiological conditions. With the introduction of single molecule analytical techniques in the early 1990's, however, new explorations into polymer structure and dynamics have begun. Atomic force microscopy and laser tweezers, in particular, have permitted three direct measures of the force at the single molecule level: (1) the force required to break A•T or G•C base pairs[15-17], (2) the force required to extend single or double stranded DNA through distinct structural conformations (e.g., B form to S form DNA, etc.[18,19]), and (3) the forces exerted by polymerases working on polynucleotides[20]. While single molecule measurements using molecule-sized nanopores have just begun, they have already demonstrated Angstrom scale resolving (99.9% accurate) capabilities[2,3]. The prospects are impressive: a nanopore-based assayer could *directly* measure molecular characteristics in terms of the blockade properties of individual molecules – due to the kinetic information that is embedded in the blockade measurements as a history of the adsorptions-desorptions of the molecule to the surrounding channel (as well as configurational changes in the molecule, see Results). Nanopore methods may even have potential for *single-molecule* sequencing at some point in the future. (The nanopore-based assayer can also *indirectly* measure molecular characteristics if it uses a reporter molecule that binds to the target molecule in solution, with subsequent distinctive blockade by the bound molecules. Such methods will not be discussed further in this paper.)

1.2 Channel Current Biophysics

As shown in Fig. 2b, residual current decreases as DNA hairpins increase their stem length from 3 to 8 base-pairs. For DNA hairpins with stems shorter than 8 base-pairs, multiple states were not clearly discernible by the prototype, presumably because the hairpins were too short to interact with the current/force constriction near the limiting aperture. For 9 base-pair hairpins, and longer, a clear 1/f noise (flicker noise) is discernible (Fig. 3b, top) – a preliminary indication of the single-molecule binding kinetics described in Fig. 4c. The mechanism[3,21] described in Fig. 4c hypothesizes that the upper level (UL) blockade state is unbound. A preliminary result (CL for detailed analysis), that strengthens the UL unbound hypothesis, is that the UL blockade level plateaus once the hairpin stem length reaches seven base pairs. This plateau occurs well before that of the other blockade levels (which is due to the hairpin stem extending beyond the pore vestibule entrance). The explanation for the UL plateau centers on the tight flow geometry between channel and captured hairpin. In such a geometry, much of the ionic flow is confined to be in or near the grooves of the captured DNA molecule. For the unbound molecule, this groove flow can be directed towards the limiting aperture by appropriate orientation of the hairpin molecule (which it is free to do since it is unbound). The unbound molecule can thus cause a "short circuit" effect, where the contribution to the ionic current is not significantly altered as the hairpin is extended (by base-pair addition), thus explaining the early plateau.

If the UL blockade state is unbound at its terminus there is the possibility that conformational kinetics might be observable at the pore-captured polymer end. This motivated examination of a set of dsDNA termini that had already been examined using NMR. Preliminary results are shown in Fig. 5 (further discussion Ref. 22), with indications that conformational kinetics are observable and consistent with NMR results via agreement on number of low energy conformational states.

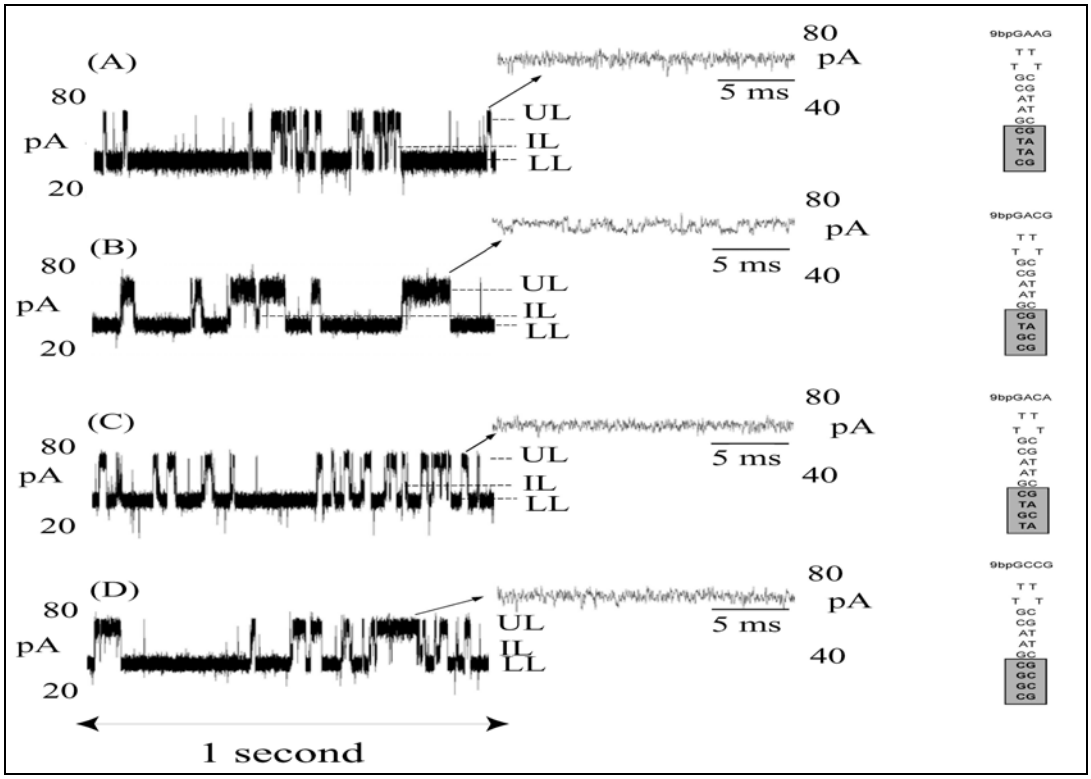


Figure 5. UL toggle correlates with NMR predicted conformational switching. Preliminary results add credence to the hypothesis that the UL state has an unbound terminus, and that while in that state conformational switching may be observable. For the molecules and blockades shown, three are found by NMR to have one dominant ground state (no switching), while one of the molecules is found by NMR to have two dominant low energy states (switching). This corresponds exactly with what has been observed in terms of the existence of fine structure (toggling) in the UL blockade state.

1.3 Patch clamp noise limitations

Nano-scale pores in bilayers are studied by measuring the ionic current induced by applying a potential difference across the bilayer. Polymer dynamics in nanometer-scale pores have been measured in timescales ranging from microseconds to seconds. The lower limit of this range, or accessible detector bandwidth, is determined by noise resulting from 1/f (flicker) noise, Johnson noise, Shot noise, and membrane capacitance noise. In Fig. 3b, the current spectral density is shown for the typical bilayer, an open α-hemolysin channel, and a channel with DNA hairpin blockade.

Noise (or signal) is 1/f-type if it has power spectrum S(f) proportional to $1/f^{\alpha}$ at low frequencies, with $0.5<\alpha<1.5$. Changes in channel configuration typically appear as 1/f noise in the channel current spectral analysis[23]. Channel gating is found to be the norm for biological channels, with α-hemolysin a notable exception (see Fig. 3b). The same is not true for observations of the channel during molecular-capture events. The thermal noise[24] contribution at the 1 GΩ channel resistance has an RMS noise current of 0.4 pA, consistent with Fig. 3b. Shot noise is the result of current flow based on discrete charge transport[25]. During nanopore operation with 120pA current (with 10KHz bandwidth) there is, similarly, about 0.6 pA noise due to the discreteness of the charge flow. As with Johnson noise, the Shot noise spectrum is white, consistent with Fig. 3b. The specific capacitance of lipid bilayers is approximately 0.8 $\mu F/cm^2$ (very large due to molecular dimensions), and the specific conductance is approximately $10^{-6}\ \Omega^{-1}cm^{-2}$ (see Ref. 26 for further details). In order for bilayer conductance to produce less RMS noise current than fundamental noise sources (under the conditions above), the leakage current must be a fraction of a pA. This problem is solved by reducing to less than a $500\mu m^2$ bilayer area[11,1,2], for which less than 0.6 pA leakage current results and for which total bilayer capacitance is at most 4pF. This indicates that a decrease in bilayer area by another magnitude is about as far as this type of noise reduction can go. Preliminary attempts to do this, however, indicate a less controllable toxin intercalation rate, among other difficulties.

A preliminary assessment of the bandwidth of the nanopore detector can be given: the bandwidth is high-pass limited at approximately the frequency where the blockade spectral density asymptotes to the open-channel spectral density. In Fig. 3b this occurs at approximately 10 kHz, indicating a usable detector bandwidth of approximately 10 kHz. Higher frequency information may still be accessible, however, via approaches akin to particle scattering methods, the idea being to inject high-energy bursts, using a laser perhaps, and to observe the ring-down, or broken pieces, of the excited analyte. Overall, the nanopore geometry, with it's associated gigohm range of resistances, appears to limit blockade analysis to frequencies below 100kHz, assuming bandwidth is restricted solely by fundamental noise sources (Johnson noise and shot noise) and not further restricted by membrane noise. Going the extra distance on bandwidth, from 10kHz to 100kHz, poses a challenging problem for managing the membrane noise. In fact, it may only be possible when there is no membrane, i.e., reaching 100kHz bandwidth may only be possible with solid-state nanopores where the mounting substrate has negligible noise contributions.

2. MATERIALS AND METHODS

2.1 Nanopore Implementation And DNA Hairpin Design

Each experiment was conducted using one α-hemolysin channel inserted into a diphytanoyl-phosphatidylcholine/hexadecane bilayer, where the bilayer was formed across a 20-micron diameter horizontal Teflon aperture[1]. The bilayer separates two seventy-microliter chambers containing 1.0 M KCl buffered at pH 8.0 (10 mM HEPES/KOH). The nine base-pair hairpin molecules examined share an eight base-pair hairpin core sequence, with addition of one of the four permutations of Watson-Crick base-pairs that may exist at the blunt end terminus, i.e., 5'-G•C-3', 5'-C•G-3', 5'-T•A-3', and 5'-A•T-3'. Denoted 9GC, 9CG, 9TA, and 9AT, respectively. The full sequence for the 9CG hairpin is 5' CTTCGAACGTTTTCGTTCGAAG 3', where the base-pairing region is underlined. An eight base-pair DNA hairpin with a 5'-G•C-3' terminus was also tested. The prediction that each hairpin would adopt one base-paired structure was tested and confirmed using the DNA mfold server (http://bioinfo.math.rpi.edu/~mfold/dna/form1.cgi), which is based in part on data from SantaLucia 1998 (Ref. 27). Further details on the nanopore construction and the DNA synthesis tools are in Ref. 2.

2.2 Signal processing methods and architecture

The software developed for analysis of stochastic sequential data is divided into four main divisions: feature identification, feature extraction, pattern recognition, and kinetics analysis. Different tools are generally employed at each stage in order to realize the most noise resistant tools for information extraction and classification. At the feature identification stage the main tool is a generalized HMM. An assortment of theoretical statistical methods are also used as needed, based on features extracted via time-domain and wavelet-domain analysis. At the feature extraction stage, the

main tools are HMMs, time-domain Finite State Automata (FSAs), and wavelet-domain FSAs. At the pattern recognition stage the main tool employed is a Support Vector Machine (SVM). SVMs draw upon variational methods in their construction and have efficient computational implementations. The SVM approach also encapsulates a significant amount of discriminatory information in the choice of kernel in the SVM (akin to a regularized distance measure in the space of the training data), and a number of novel kernels are used. At the kinetics analysis stage the goal is primarily to obtain a better biophysical understanding of the captured analyte. Direct application of kinetic information to the classification problem is unlikely since it requires much more signal to get a good kinetics profile than the brief 100ms of information used by the full HMM feature extraction.

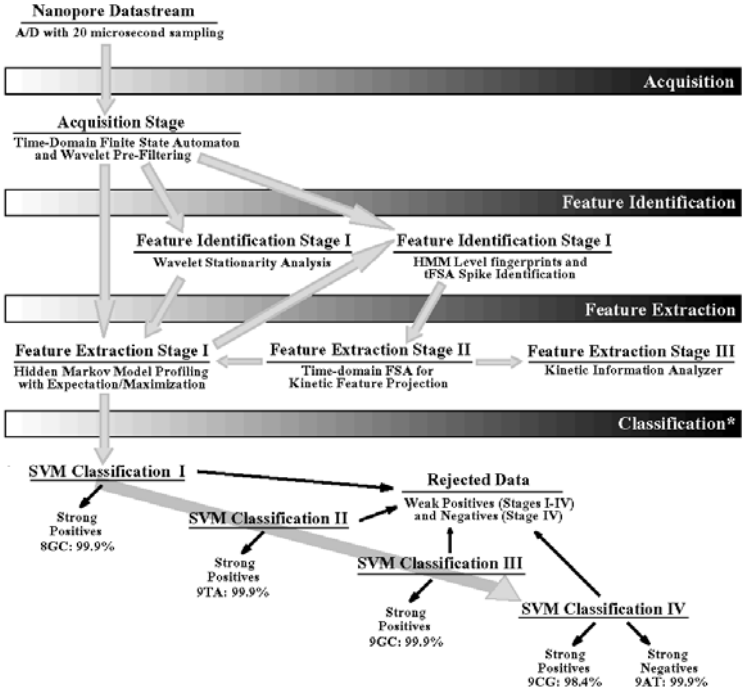


Figure 6. The signal processing architecture. Signal acquisition was performed using a time-domain, thresholding, Finite State Automaton. This was followed by adaptive pre-filtering using a wavelet-domain Finite State Automaton. Feature extraction on those acquired channel blockades was done by Hidden Markov Model processing[28-31]; and classification was done by Support Vector Machine[32,33]. The optimal SVM architecture is shown for classification of molecules 9CG, 9GC, 9TA, 9AT, and 8GC. The linear tree multi-class SVM architecture benefits from strong signal skimming and weak signal rejection along the line of decision nodes. Scalability to larger multi-class problems is possible since the main on-line computational cost is at the Hidden Markov Model feature extraction stage. The accuracy shown is for single-species mixture identification upon completing the 15th single molecule sampling/classification (in approx. 6 seconds). Off-line kinetic feature extraction was done at the newly added HMM Level and tFSA Spike Identification module and the time-domain FSA Kinetic Feature Projection module.

The solution sampling protocol used periodic reversal of the applied potential to accomplish the capture and ejection of single DNA molecules (added to the cis chamber in 20 μM concentrations). The current blockade data was filtered at 10 kHz bandwidth using an analog low pass Bessel filter and recorded at 20 μs intervals using an Axopatch 200B amplifier coupled to an Axon Digidata 1200 digitizer (Axon Instruments, Foster City, CA). A time-domain finite state automaton (FSA[34]) with eight states performed the generic signal identification/acquisition for the first 100 msec of blockade signal (Acquisition Stage, Fig. 6). The effective duty cycle for acquiring the desired 100 ms blockade measurements was one reading every 0.4 seconds. Further details on the voltage toggling protocol and the time-domain FSA are in Ref. 2.

The signal preprocessing makes use of an off-line wavelet stationarity analysis[35] (Wavelet Stationarity Analysis in Fig. 6). With completion of preprocessing, an HMM with EM[31] was used to remove noise from the acquired signals, and to extract features from them (Feature Extraction Stage, Fig. 6). 150 feature vector components were extracted from parameterized emission probabilities, a compressed representation of transition probabilities, and use of *a posteriori* information deriving from the Viterbi path solution. The resulting parameter vector, normalized such that vector

components sum to unity, was used to represent the acquired signal in discrimination at the Support Vector Machine stages. (Further details in Ref. 2)

Extraction of kinetic information begins with identification of the main blockade levels for the various blockade classes (off-line). This information is then used to scan through already labeled (classified) blockade data, with projection of the blockade levels onto the levels identified for that class of molecule. A time-domain FSA performs the above scan, and uses the information obtained to tabulate the lifetimes of the various blockade levels. Once the lifetimes of the various levels are obtained, information about a variety of kinetic properties are accessible. If the experiment is repeated over a range of temperatures, a full set of kinetic data can be obtained. This data can be used to calculate k_{on} and k_{off} rates for binding events, as well as indirectly calculate forces by means of the van't Hoff Arrhenius equation.

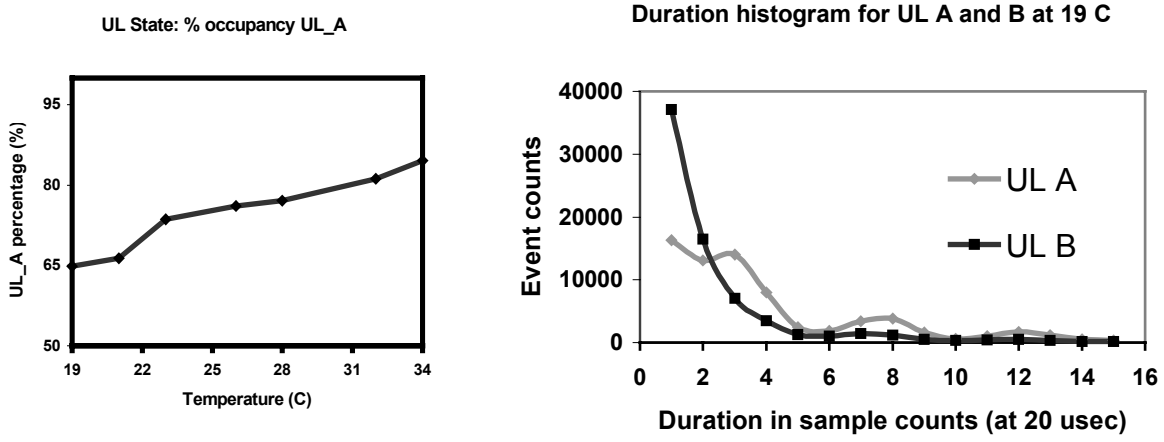
The normalized feature vectors obtained from the feature extraction stage are classified using binary Support Vector Machines (SVMs). Binary SVMs are based on a decision-hyperplane heuristic that incorporates structural risk management by attempting to obtain the greatest training-instance void, or "margin", around the decision hyperplane. Binary SVMs can be grouped into a classifier tree to perform multi-class discrimination (shown in classification stages I-IV in Figure 1). Tuning on the multi-class SVM architecture itself was done for performance optimization, and separate tuning was done on the polarization strength used in the data cleaning. Tuning was also done on the SVM internals, over families of kernels based on regularized distances[36] and regularized information divergences. In the former case, the squared Euclidean distance between feature vectors $\mathbf{x}$ and $\mathbf{y}$, $\mathbf{d^2(x,y) = \Sigma_k(x_k-y_k)^2}$, also known as the squared $\mathbf{l_2}$-norm on $\mathbf{(x-y)}$, $\mathbf{[l_2(x-y)]^2 = d^2(x,y)}$, is associated with the Gaussian kernel: $\mathbf{K_G(x,y) = exp(-d^2(x,y)/2\sigma^2)}$. The latter case represents a whole new class of kernels[2] based on information-theoretic measures of distance between probability vectors (discrete distributions). The information divergence (relative entropy) between probability vectors $\mathbf{x}$ and $\mathbf{y}$, $\mathbf{D(x||y) = \Sigma_k x_k log(x_k/y_k)}$, can be associated with the "Entropic kernel:" $\mathbf{K_E(x,y) = \tilde{exp}([D(x||y)+D(y||x)]/2\sigma^2)}$. The terminating SVM node of the classifier tree (stage IV in Figure 6) performed best with such an Entropic kernel. The other nodes of the classifier tree used a regularized-distance type kernel, the "Variation-distance kernel," based on the square root of the $\mathbf{l_1}$-norm, where $\mathbf{l_1(x-y) = \Sigma_k|x_k-y_k|}$, with kernel $\mathbf{K_I(x,y) = \tilde{exp}(\sqrt{l_1(x-y)}/2\sigma^2)}$.

The SVM discriminators were trained by solving their KKT relations using the Sequential Minimal Optimization (SMO) procedure[37]. A chunking[38,39] variant of SMO was employed to manage the large training task at each SVM node. The multi-class SVM training was based on over ten thousand blockade signatures for each DNA hairpin species. The data cleaning needed on the training data was accomplished by an extra SVM training round (further details on data cleaning in Ref. 2).

In the five DNA hairpin study, the test data consisted of over two thousand blockade signals for each DNA hairpin species and was drawn from experiments that were run on days (and nanopores) different from those used to acquire the training data. Testing on single-species mixture calling was done directly, with classification on observations from single-species solutions in the *cis* chamber. One goal of the study was to find how many classification attempts were required to classify the single-species solutions with very high confidence. Scoring was possible by tracking the known labels on the test data. For the mixture tests some of the train data was used for an added calibration. An extra calibration was required because true mixtures of hairpins are sensitive to the different (entropic) acceptance rates and (discriminator) rejection rates by the nanopore instrument for the different hairpin species.

3. RESULTS

Further information has been obtained that supports the hypothesis that the UL state is unbound at its terminus. The new results also indicate that conformational switching can be observed at the terminus of captured DNA hairpins (see Figures 7 and 8). The new results also suggest the possibility that the upper UL state (called UL_A for blockades with UL toggles) may not only be unbound at its terminus, but at its DNA loop as well. This is because a 80 μs period (with multiples) is observed in the % blockade in UL_A (see Fig. 8). In other words, the DNA hairpin might undergo one or more 360° "hops" while in its UL_A state, and partway through those hops it is unlikely that the hairpin can reseat itself (transition to UL_B), which leads to a boost in UL_A lifetime (compared to UL_B) at the 80 μs period and its multiples.



(a) **(b)**

Figure 7. Panel (a) shows a monotonically increasing growth in UL_A percentage (of UL) with temperature. Panel (b) shows the blockade counts for a range of level-lifetimes.

The latest results on conformational (and rotational) dynamics was made possible by data analysis using the new software additions in Fig. 6 at the feature extraction and feature identification stages. Other software developments include: (1) evidence that the kernels used appear to satisfy Mercer's conditions, allowing possible extension to other divergence-based kernels (see Disc. for more detail). (2) Preliminary HMM projection results (for use in the kinetic analysis instead of a FSA) indicate that highly accurate preservation of transitions can be obtained with an HMM with parameterized emission/transition variables that are driven (during EM cycling) to the dominant blockade levels. Although the transitions are well preserved, which maintains the desired lifetime kinetics information, the levels all drift to the overall signal average. This is done in a manner that preserves the well-ordering of the blockade levels, however, so the level/kinetic information should be fully recoverable with this approach, with benefit an HMM formalism for kinetic state-projection rather than an *ad hoc* FSA formalism.

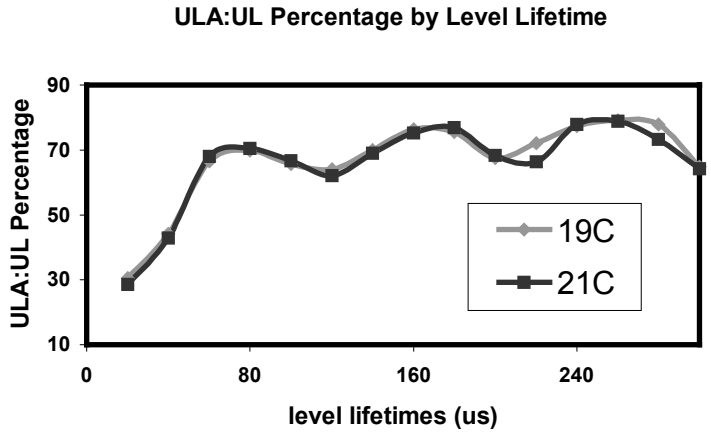


Figure 8. The percentage of time that the UL level spends in UL_A state is generally greater as the blockade durations grow longer. There is a noticeable periodic rise at 80 μs and its multiples, however, suggesting the possibility that the hairpin is performing 360 degree hops (see Disc. for further details.)

The latest biophysics results: using lambda exonuclease as a translocational brake for ssDNA appears to be possible. Conditions have been obtained where both the exonuclease retains function and the toxin self-assembles[40]. Force probing of the nanopore environment continues as well[41].

4. DISCUSSION

The use of large biological pores as polymer sensors is a relatively new possibility dating from the pioneering experiments of Bezrukov et al.[8] and Kasianowicz et al.[10]. Commercialization of these devices will probably require stabilization or replacement of lipid bilayers with more durable films. Protein channels in conventional bilayers, however, will continue to serve as prototypes for those devices. They can also serve as research tools for analyzing the structure and dynamics of solution polymers and pore-forming toxins themselves. At the moment, X-ray diffraction analysis and NMR spectroscopy form the foundation of DNA structure analysis at single base pair resolution. These techniques are extremely time-consuming, however, and relatively few structures have been resolved.

Nanopore-based cheminformatics offers an exciting new arena in which to develop and test the latest machine learning approaches. The cheminformatics work on α-hemolysin is also generally applicable to other channel blockade analysis problems (whether originating from biologically-based or semiconductor-based channels). For channel current cheminformatics in general, this permits the early exploration of cheminformatics software tools and methods.

4.1 Biophysics Research

Given the restricted flow geometry between protein channel and a captured DNA hairpin (shown in Fig. 1b), it is perhaps surprising that a number of unexpected nanomechanical and nanofluidic issues haven't arisen. So far, there is only the odd "short circuit" effect described earlier. In a forthcoming manuscript[41], a variety of DNA hairpins are used as probes of the α-hemolysin protein channel geometry. The same experiments also serve to reveal the forces at various points in the channel. Building on the work of Vercoutere et al.[1], a series of blunt-ended DNA hairpins have been used to probe the depth of the vestibule. The blockade signal exhibits a single blockade level for hairpins with stem lengths ranging from three base-pairs (3bps) to seven base-pairs (7bps). For the 8bp hairpin a telegraph signal appears, with the primary blockade level at the greater resistance. For 9bp hairpins, and those with longer stems, there appear to be three main levels. The geometric bottom of the vestibule is reached with a 9bp hairpin, ±1bp. Using the 9bp hairpin as a base, and taking into account the 3′-fraying/extension hypothesis (mentioned in the mechanism description in Fig. 4c), single-stranded DNA overhangs of varying length were added to the base at the 3′ terminus. This permits critical force/geometry probing of the trans-membrane part of the channel in a very controlled manner, by a single (captured) molecule event. Preliminary results indicate two significant trans-membrane constrictions, one at the limiting aperture, and one near the trans-opening. The resolving power of the limiting-aperture/trans-opening constrictions is of critical importance in DNA-sequencing and biosensor applications, and is undetermined as of yet.

The Sanger dideoxy method uses capillary gel electrophoresis with linear polyacrylamide as the separation matrix, is limited to a 1500 nucleotide read length, and requires substantial front-end sample preparation. Nanopore-based terminus classification, on the other hand, offers the prospect of the entire Sanger sequencing protocol being performed on a microchip-sized laboratory. Unlike fluorescence based devices, the efficiency of a nanopore sensor will increase with decreasing capillary size. For single nucleotide polymorphism (SNP) identification, a nanopore detector also offer the prospect that only small sample volumes need be used, such that PCR amplification may not even be needed. Non-PCR expression analysis, in general, may offer a new method for biological experimentation on live cells using patch-clamp methods. From the inception of the nanopore detector idea in the 1990s it has been also been suggested that a single-molecule approach to sequence individual DNA molecules might be made possible.

If ssDNA translocation through α-hemolysin can be slowed enough, by use of single-enzyme couplings or servo-electronics, then single-molecule DNA sequencing may prove possible as well. Polymerases and exonucleases typically add or remove nucleotides on the millisecond timescale, and they can do so at high fidelity for up to 1 megabase without dissociating (RNA polymerase citation form Stanford). Recently, conditions have been obtained where both lambda exonuclease retains function and the toxin self-assembles. Further experiments are being pursued along these lines. If a DNA brake is successfully employed, this would greatly relax the bandwidth requirement for the nanopore sensor. Polymerases may also find use in the nanopore sensor. Recent experiments demonstrate that a zero mode wave-guide can be used to detect incorporation of individual fluorescent nucleotides by a single T7 DNA polymerase[42].

For single-molecule sequencing to be successful, however, the deconvolution problem must be solved for the collection of bases at the main current restrictions. Deconvolution of base content from a single blockade signal may be possible if dominant contributions to resistance span only 20 Å or so (amounting to about three nucleotides length of ssDNA). Thus, single-molecule sequencing will require further progress in the force/geometry probing and the enzyme braking

efforts. Since dsDNA carries much more information than ssDNA (i.e., the molecular motions are much more constrained and readable), progress may eventually be made with easily formed synthetic/ssDNA chimeric molecules that are sized more like ssDNA, but have the richer bond-formation structure of dsDNA. Other work on methylated or dye-tagged ssDNA and dsDNA appears to offer significant new information as well (without laser excitation of dyes being introduced yet).

4.2 Machine Learning Research

The variation and entropic kernels probably satisfy Mercer's conditions since they can be described as metrics "regularized" by incorporation as positive arguments in a decaying exponential. A general formulation was used for feature extraction that was designed to arrive at probability vectors (i.e., discrete probability distributions) on a predefined, and complete, space of possibilities. (The different blockade levels, and their frequencies, for example.) This turns out to be a very general formulation, wherein feature extraction makes use of signal decomposition into a complete set of separable states. A probability vector formulation also provides a straightforward hand-off to the SVM classifiers since all feature vectors have the same length with such an approach. What this means for the SVM is that geometric notions of distance are no longer the best measure for comparing feature vectors. For probability vectors (i.e., discrete distributions), the best measures of similarity are the various information-theoretic divergences: Kullback –Leibler, Renyi, etc. By symmetrizing over the arguments of those divergences we obtain a rich source of kernels that might work well with the types of probabilistic data obtained. Thus far, only the Kullback –Leibler divergence has been extensively studied in this manner.

A single HMM/EM process was used to perform feature extractions in the experiments. If separate HMMs were used to model each species, the HMM/EM processing could also be operated in a discriminative mode. This requires multiple HMM/EM evaluations (one for each species) on each unknown signal as it is observed. Increased computational burden at the worst place in an on-line pattern recognition setting: the expensive feature extraction stage. The multiclass discriminator that was implemented used binary SVMs grouped in a decision tree architecture (see Fig. 6). Alternatively, a (single) multiclass SVM could have been implemented. The latter takes on a much more complicated form that appears much more susceptible to noise, however, and is much more difficult to train since larger "chunks" are needed to carry all the support vectors. Re-establishing the α-hemolysin channel on a day-to-day basis presents a major complication to the pattern recognition task. SVM classification in such circumstances faces weaker training convergence and poorer signal calling. For the five classes considered in the prototype, a passive stabilization approach was used that optimized the kernels for high rejection. More active (computationally-based) stabilization methods are being studied for larger multiclass problems and improved accuracy overall. See Ref. 21 for more detail.

4.3 Conclusion

Nanopore cheminformatics based on the α-hemolysin nanopore detector offers a new method for single molecule experimentation. It has already been established that molecules can be characterized in terms of their binding kinetics and dissociation kinetics[2,3]. Preliminary results shown here indicate that molecules may also be characterized in terms of their conformational kinetics (and possibly rotational kinetics). Biomolecular discrimination will generally not benefit from the new kinetic information extraction methods, however, since they are at millisecond time-scales. Even so, the methods may eventually play a discriminatory role once laser modulations are introduces into the system.

5. ACKNOWLEDGEMENTS

Special thanks to Mark Akeson for instructions on setting-up and operating a nanopore detector. Thanks again to Mark for directing my attention to the DNA hairpin thought to be undergoing conformational change. For their many helpful conversations and technical assistance, I thank my lab technician, Andrew Duda, and my colleagues at the University of California, Santa Cruz: Veronica DeGuzman, David Deamer, Wenonah Vercoutere, Andrea Solbrig, and Clarence Lee. I also thank my colleagues at the University of New Orleans and the Research Institute for Children, New Orleans: Mahdi Abdelguerfi and Seth Pincus. This work was funded by the Research Institute for Children, New Orleans, and the University of New Orleans.

REFERENCES

1. Vercoutere W., S. Winters-Hilt, H. Olsen, D.W. Deamer, D. Haussler, and M. Akeson, 2001. Rapid discrimination among individual DNA hairpin molecules at single-nucleotide resolution using an ion channel. Nat. Biotechnol. 19(3):248-252
2. Winters-Hilt, S., W. Vercoutere, V. S. DeGuzman, D.W. Deamer, M. Akeson, and D. Haussler, 2003. Highly Accurate Classification of Watson-Crick Base-Pairs on Termini of Single DNA Molecules. Biophys. J. 84 (2) 1-10.
3. Vercoutere, W., S. Winters-Hilt, V.S. DeGuzman, D.W. Deamer, S. Ridino, J.T. Rodgers, H. Olsen, A. Marziali, and M. Akeson. 2003. Discrimination Among Individual Watson-Crick Base-Pairs at the Termini of Single DNA Hairpin Molecules. Nucl. Acids. Res. 31, 1311-18.
4. Song L., M.R. Hobaugh, C. Shustak, S. Cheley, H. Bayley, and J.E. Gouaux, 1996. Structure of Staphylococcal Alpha-Hemolysin, a Heptameric Transmembrane Pore. Science 274 (5294):1859-1866.
5. Gouaux J.E., O. Braha, M.R. Hobaugh, L. Song, S. Cheley, C. Shustak, and H. Bayley, 1994. Subunit stoichiometry of staphylococcal alpha-hemolysin in crystals and on membranes: a heptameric transmembrane pore. Proc. Natl. Acad. Sci. USA 91:12828-12831.
6. Sakmann, B., E. Neher (eds). 1995. Single-Channel Recording, Plenum Press
7. Ashcroft, F. 2000. Ion Channels and Disease. Academic Press.
8. Bezrukov, S.M., I. Vodyanoy, V.A. Parsegian. 1994. Counting polymers moving through a single ion channel. Nature 370 (6457), pgs 279-281.
9. Coulter, W.H., 1957, High speed automatic blood cell counter and cell size analyzer: proceedings of the National Electronics Conference, p. 1034-1042.
10. Kasianowicz, J.J., E. Brandin, D. Branton, and D.W. Deamer, 1996. Characterization of Individual Polynucleotide Molecules Using a Membrane Channel. Proc. Natl. Acad. Sci. USA 93(24):13770-13773.
11. Akeson M, D. Branton, J.J. Kasianowicz, E. Brandin, D.W. Deamer. 1999. Microsecond Time-Scale Discrimination Among Polycytidylic Acid, Polyadenylic Acid, and Polyuridylic Acid as Homopolymers or as Segments Within Single RNA Molecules. Biophys. J. 77(6):3227-3233.
12. Meller A, L. Nivon, E. Brandin, J. Golovchenko, and D. Branton, 2000. Rapid nanopore discrimination between single polynucleotide molecules. Proc. Natl. Acad. Sci. USA 97(3):1079-1084.
13. Meller A, L. Nivon, and D. Branton, 2001. Voltage-driven DNA translocations through a nanopore. Phys. Rev. Lett. 86(15):3435-3438.
14. Mfold: http://bioinfo.math.rpi.edu/~mfold/dna/form1.cgi
15. Rief, M., H. Clausen-Schaumann, H. E. Gaub. 1999. Sequence dependent mechanics of single DNA molecules. Nature Struct Biology 6 (4), pg 346-349.
16. Clausen-Schaumann, H., M. Rief, C. Tolksdorf, H. E. Gaub, H.E. 2000. Mechanical stability of single DNA molecules. Biophys. J. 78 (4), pg 1997-2007.
17. EssevazRoulet, B., U. Bockelmann, F. Heslot. 1997. Mechanical separation of complementary strands of DNA. Proc. Nat. Acad. Sci USA 94 (22), pg 11935-11940.
18. Fisher, T.E., P.E. Marszalek, J.M. Fernandez. 2000. Stretching single molecules into novel conformations using the atomic force microscope. Nature Struct. Biology 7 (9), pg 719-724.
19. Smith, S.B., Y. Cui, C. Bustamante. 1996. Overstretching B-DNA: the elastic response of individual double-stranded and single-stranded DNA molecules. Science 271 (5250), pgs 795-9.
20. Wang, M.D., M.J. Schnitzer, H. Yin, R. Landick, J. Gelles, S.M. Block. Force and velocity measured for single molecules of RNA polymerase. Science 282 (5390), pgs 902-907.
21. S. Winters-Hilt, "Highly Accurate Real-Time Classification of Channel-Captured DNA Termini," Third International Conference on Unsolved Problems of Noise and Fluctuations in Physics, Biology, and High Technology, pg 355-368, 2003.
22. Winters-Hilt, S., Akeson, M., Nanopore Cheminformatics, Sub. to DNA and Cell Biology (MCBIOS), Feb. 2004.
23. Bezrukov, S.M. 2000. Ion Channels as Molecular Coulter Counters to Probe Metabolite Transport. J. Membrane Biol. 174, 1-13.
24. Nyquist, H. 1927. Thermal agitation in conductors. Phys. Rev. 29, 614. Nyquist, H. 1928. Thermal agitation of electrical charge in conductors. Phys. Rev. 32, 110-113.
25. Schottky, W. 1918. Uber spontane Stromschwankungen in verschiedenen Elektrizitatsleitern. Ann. Phys. 57, 541-567.
26. Hille, B. 1992. Ionic Channels of Excitable Membranes (2nd Ed.). Sinauer Associates, Sunderland, MA.

27. SantaLucia J. 1998. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. Proc. Natl. Acad. Sci. USA 95(4):1460-1465.
28. Chung, S.H., J.B. Moore, L. Xia, L. S. Premkumar, and P. W. Gage. 1990. Characterization of single channel currents using digital signal processing techniques based on Hidden Markov models. Phil. Trans. R. Soc. Lond. B 329. 265-285.
29. Chung, S-H., and P. W. Gage. 1998. Signal processing techniques for channel current analysis based on hidden Markov models. *In* Methods in Enzymology; Ion channels, Part B. P. M. Conn editior. Academic Press, Inc., San Diego. 420-437.
30. Colquhoun, D., and F. J. Sigworth. 1995. Fitting and statistical analysis of single-channel products. *In* Single-channel recording. B. Sakmann and E. Neher editors. Second edition. Plenum Publishing Corp., New York. 483-587.
31. Durbin R. 1998. Biological sequence analysis : probalistic models of proteins and nucleic acids. Cambridge, UK New York: Cambridge University Press. xi, 356 p.
32. Vapnik, V. N. 1999. The Nature of Statistical Learning Theory (2nd ed.). Springer-Verlag, New York.
33. Burges, C.J.C**.** 1998. A tutorial on support vector machines for pattern recognition. Data Min. Knowl. Discov., 2. 121-67.
34. Cormen, T.H., C. E. Leiserson, and R. L. Rivest. 1989. Introduction to Algorithms. MIT-Press, Cambridge, USA.
35. Diserbo M, P. Masson, P. Gourmelon, and R. Caterini, 2000. Utility of the wavelet transform to analyze the stationarity of single ionic channel recordings. J. Neurosci. Methods 99(1-2):137-141.
36. Jaakkola, T. S., and D. Haussler. 1998. Exploiting generative models in discriminative classifiers. *In* Advances in Neural Processing Systems 11. Cambridge, MA, 1999. MIT Press.
37. Platt, J. C. 1998. Fast Training of Support Vector Machines using Sequential Minimal Optimization. *In* Advances in Kernel Methods -- Support Vector Learning. B. Scholkopf, C. J. C. Burges, and A. J. Smola editors. MIT Press, Cambridge, USA. Ch. 12.
38. Osuna, E.; R. Freund, and F. Girosi. 1997. An improved training algorithm for support vector machines. *In* Neural Networks for Signal Processing VII. J. Principe, L. Gile, N. Morgan, and E. Wilson editors. IEEE, New York. 276-85.
39. Joachims, T. 1998. Making large-scale SVM learning practical. *In* Advances in Kernel Methods -- Support Vector Learning. B. Scholkopf, C. J. C. Burges, and A. J. Smola editors. MIT Press, Cambridge, USA. Ch. 11.
40. A. Solbrig, S. Winters-Hilt, V. Deguzman, M. Akeson, Using lambda exonuclease as a braking mechanism in ssDNA translocation through a nanopore. Preprint in preparation.
41. DeGuzman, V., S. Winters-Hilt, M. Akeson, Using DNA hairpins to probe the α-hemolysin geometry and force environment. Preprint in preparation.
42. Levene, M. J., J. Korlach, S. W. Turner, M. Foquet, H. G. Craighead and W. W. Webb, "Zero-mode waveguides for single molecule analysis at high fluorophore concentrations," *Science* **299**, 682-686, 2003