Single nucleotide polymorphism detection via cheminformatics analysis of nanopore-blockades engineered for event transduction

Stephen Winters-Hilt
Computer Science Department
University of New Orleans
New Orleans, LA 70148, USA
winters@cs.uno.edu

Abstract

The nanopore transduction detector is a unique platform for detection and analysis of single molecules. Proof-of-Concept experiments indicate a promising approach to single nucleotide polymorphism (SNP) detection in a clinical environment, via use of the channel-blockade signals produced by engineered event-transducers. The transducer molecule is a bi-functional molecule; one end is captured in the nanopore channel while the other end is outside the channel. This extra-channel end is engineered to bond to a specific target: the analyte being measured. When the outside portion is bound to the target, the molecular changes (conformational and charge) and environmental changes (current flow obstruction geometry and electro-osmotic flow) result in a change in the channel-binding kinetics of the portion that is captured in the channel. This change of kinetics generates a change in the channel blockade current which is engineered to have a signal unique to the target molecule; the transducer molecule is, thus, a bi-functional molecule which is engineered to produce a unique signal change upon binding to its cognate. This provides the basis for a highly sensitive and accurate biosensor.

1 INTRODUCTION & BACKGROUND

SNP detection offers the tantalizing prospect of medical diagnostics and cancer screening by assaying targeted regions of genomic variation. Common methods for SNP detection are typically PCR-based, thus inherit the PCR error rate (0.1% in some situations). The percentages of minority SNP population might be 0.1%, or less, in instances of clinical interest, thus the PCR error rate is critically limiting in the standard approach. Standard methods for SNP detection have high sensitivity, but typically lack high specificity and versatility. As will be shown, the Nanopore Transduction Detector is a unique platform for direct detection of SNPs with both high sensitivity and high specificity, and without use of PCR-amplification.

There are two approaches to utilizing a nanopore for detection purposes: translocation and/or dwell-time (T/DT) based approaches, which strongly relies on blockade dwell-times, and nanopore transduction detection (NTD) based approaches, which functionalizes the nanopore by utilizing an engineered blockade molecule with blockade features typically not including dwell-time.

Translocation/dwell-time methods introduce different states to the channel via use of the frequency of channel blockade events and their durations (the classic Coulter counter features) [1,3,4]. The strongest feature employed in translocation/dwell-time discrimination, and often the only feature, is the blockade dwell-time where the dwell-time is typically engineered to be associated with the lifetime until a specific bond failure occurs. Other feature variations include time *until* a bond-formation occurs, or simply measuring the approximate length of a polymer according to its translocation 'dwell'-time.

Transduction methods introduce different states to the channel via observations of changes in blockade statistics on a specially engineered, partially-captured, channel modulator, typically with a binding moiety for a specific target of interest linked to the modulator's extra-channel portion. The modulators 'state' changes according to whether its binding moiety is bound or unbound. For a comparative analysis, see Table 1 below.

(1) Feature Space. The T/DT approach typically has a single feature, the dwell time. Sometimes a second feature, the fixed blockade level observed, is also considered, but usually not more features sought (or engineered) than that. The NTD approach has multiple features, e.g., blockade HMM parameters, etc., with number and type according to modulator design objectives.
(2) Versatility. T/DT: highly engineered/pre-processed for detection application to a particular target. NTD: requires minimal preparation/augmentation to the transduction platform via use of separately provided binding moieties (antibody or aptamer, for example) for particular target or biomarker (which are then simply linked to modulator)
(3) Speed. T/DT: Slow: entire detection "process" is at the channel, and typically restricted on processing speed to the average time-scale feature (dwell-time) for the longest-lived blockade signal class. NTD: Fast: feature extraction not dependent on dwell-time. Very low probability to get a false positive.
(4) Multichannel. T/DT: Method not amenable to multichannel gain with single-potential platform (can't resolve single-channel blockade signal with multichannel noise). NTD: Have multichannel gain due to rich signal resolution capabilities of an engineered modulator molecule.
(5) Feature Refinement/Engineering. T/DT: No buffer modifications or off-channel detection extensions via introduction of substrates; the weak feature set limited to dwell-time doesn't allow such methods to be utilized. NTD: Have "lock-and-key" level signal resolution. The introduction of off-channel substrates in the buffer solution can increase sensitivity.
(6) Multiplex capabilities. T/DT: Each modified channel is limited to detect a single analyte or single bond-change-event detection, so no multiplexing without brute force production of arrays of T/DT detectors in a semiconductor production setting. NTD: Supports multi-transducer, multi-analyte detection from a single sample. Supports multichannel with a single aperture.

Table 1. Comparative analysis of the Translocation/Dwell-Time (T/TD) approach and the Nanopore Transduction Detection (NTD) approach.

The nanopore transduction detection (NTD) platform (Fig. 1) involves functionalizing a nanopore detector platform in a new way that is cognizant of signal processing and machine learning capabilities and advantages, such that a highly sensitive biosensing capability is achieved [21]. The

core idea in the NTD functionalization of the nanopore detector is to design a molecule that can be drawn into the channel (by an applied potential) but be too big to translocate, instead becoming stuck in a bistable 'capture' such that it modulates the ion-flow in a distinctive way (Fig. 2 shows some controls). An approximately two-state 'telegraph signal' has been engineered for a number of NTD modulators. If the channel modulator is bifunctional in that one end is meant to be captured and modulate while the other end is linked to an aptamer or antibody for specific binding, then we have the basis for a remarkably sensitive and specific biosensing capability. The biosensing task is reduced to the channel-based recognition of bound or unbound NTD modulators.

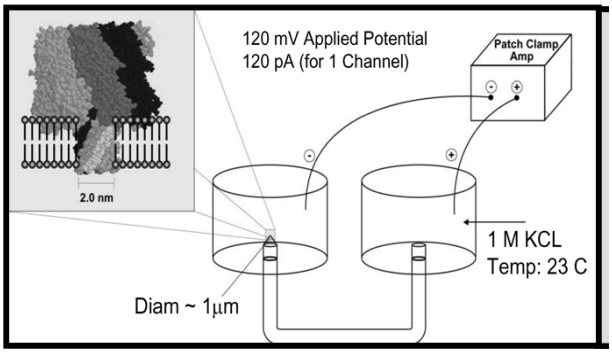

Figure 1. Schematic diagram of the nanopore transduction detector. The nanopore detector consists of a single pore in a lipid bilayer that is created by the oligomerization of the staphylococcal alpha-hemolysin toxin, and a patch clamp amplifier capable of measuring pico Ampere channel currents.

1.1 Nanopore Transduction Detection: a highly versatile platform for highly sensitive biosensing

The use of a channel modulator introduces significant, engineered, signal analysis complexity, that we resolve using artificial intelligence (machine learning) methods. The benefit of this complication is a significant gain in sensitivity over T/TD, that uses a 'sensing' moiety covalently attached to the channel itself, where they have a T/TD-type blockade 'lifetime' event, with minimal or no internal blockade structure engineered [9,10]. The NTD approach, on the other hand, has significant improvement in versatility, e.g., we can 'swap out' modulators on a given channel, in a variety of ways, since they are not covalently attached to the channel. This is possible because the modulators are drawn into the channel by an applied potential, so a quick reversal of applied potential is all that is needed to eject an electrophoretically captured molecule. Such voltage reversal is conditioned on pattern recognition informed sampling decisions in [2,8]. Modulator exchanges via buffer exchanges can be done using massive perfusions given the small volume (approximately 70 microliters) of the operational chamber (in Figure 1 this is the left chamber with the aperture). The improvements in sensitivity derive from the measurable stationary statistics of the channel blockades (and how this can be used to classify state with very high accuracy). The improvement in versatility is because all that needs to be redesigned for a different NTD experiment (or binding assay) is the linkage-interaction moiety portion of the bifunctional molecule. There is also the versatility that *mixtures* of different types of transducers can be used, a method that can't be employed in single-channel devices that use covalently bound binding moieties (or that discriminate by dwell-time in the channel) [9,10].

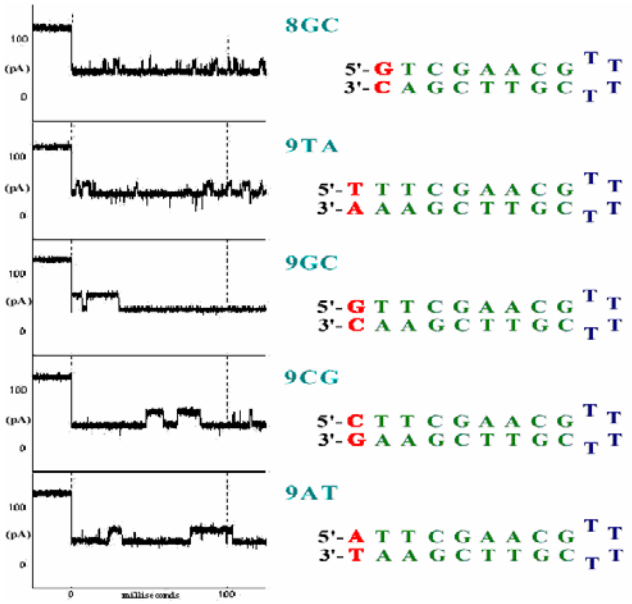

Figure 2. DNA hairpin controls and their diagnostic signals. The secondary structure of the DNA hairpins is shown on the right, with their highest scoring diagnostic signals shown on the left [22]. Each signal trace starts at approximately 120 pA open channel current and all blockades are in a range 40-60 pA upon "capture" of the associated DNA hairpin. Even so, the signal traces have discernibly different blockade structure, which is extracted using an HMM. The signals are aligned at their blockade starts and the demarked time-trace is for 100 ms.

At the nanopore channel one can observe a sampling of bound/unbound states, each sample only held for the length of time necessary for a high accuracy classification. Or, one could hold and observe a single bound/unbound system and track its history of bound/unbound states or conformational states. The *single* molecule detection, thus, allows measurement of molecular characteristics that are obscured in ensemble-based measurements. Ensemble averages, for example, lose information about the true diversity of behavior of individual molecules. For complex *bio*molecules there is likely to be a tremendous diversity in behavior, and in many cases this diversity may be the basis for their function. There can also be a great deal of diversity via post-translational modifications, as well, such as with heterogeneous mixtures of protein glycoforms that typically occur in living organisms (e.g., for TSH and

hemoglobin proteins in blood serum and red blood cells, respectively). The hemoglobin 'A1c' glycoprotein is a disease diagnostic (diabetes), and for TSH, glycation is critical component in the TSH-based regulation of the endocrine axis. Multi-component regulatory systems and their variations (often sources of disease) could also be studied much more directly using the NTD approach, as could multi-component (or multi-cofactor) enzyme systems. Glycoform assays, characterization of single-molecule conformational variants, and multi-component assays are significant capabilities to be developed further with the NTD approach, but in what follows we focus on the most wide-ranging, immediate-impact, area: NTD-based SNP assays.

1.2 Potential Impact

Nanopore transduction detection provides an inexpensive, quick, accurate, and versatile method for performing medical diagnostics. It is hypothesized that NTD biomarkers can be developed for early stage disease detection with femtomolar to attomolar sensitivity (see Table 2) for doing the standard clinical tests of the future. The potentially incredible sensitivity of the NTD targeting on biomarkers also provides a significant new tool for public health and biodefense in general.

In the preliminary results shown in Fig. 3, and detailed in [21], we show a 0.17 μM streptavidin sensitivity in the presence of a 0.5 μM concentration of detection probes with a 100 second detection window. The detection probe is a biotinylated DNA-hairpin transducer molecule (Bt-8gc) [21]. In repeated experiments we see the sensitivity limit ranging inversely to the concentration of detection probes. If taken to its limits, with established PRI sampling capabilities [2,10,39], and with stock Bt-8gc at 1mM concentration conveniently available, we believe it is possible to boost probe concentration almost three magnitudes. In doing so, we would boost sensitivity by similar measure, until the minimal observation time needed to reject limits this gain mechanism (see Table 2).

METHOD	SN
Low-probe concentration, 100s obs.	100 nM
High probe conc, 100s observation	100 pM
High probe conc, long observation (~1dy)	100 fM *
TARISA (conc. gain), 100s observation	100 fM
TERISA (enzyme gain), 100s obs.	100 aM **
Electrophoretic contrast gain, 100 s	1.0 aM

Table 2. Sensitivity limits for detection in the streptavidin-biosensor model system. *We have done 1 -1.5 day long experiments in other contexts, but not longer. Thus, current capabilities, with no modifications to the NTD platform for specialization for biosensing, can achieve close to 100 fM sensitivity by pushing the device limits and the observation window. **Only a slow enzyme turnover of 10 per second is assumed. Detection in the attomolar regime is critical for early discovery of type I diabetes destructive processes and for early detection of Hepatitis B. Early PSA detection currently has a 500 aM sensitivity. For some toxins, their potency, even at trace amounts, precludes their usage in the typical antibody-generation procedures (for mAb's that target that toxin). In this instance, however, aptamer-based NTD probes can still be obtained.

2 PRELIMINARY RESULTS

2.1 Model system based on streptavidin and biotin

A biotinylated DNA-hairpin that is engineered to generate two signals depending on whether or not a streptavidin molecule is bound to the biotin (see Figs. 3A, 3B & 2, preliminary work is described in [5] and [9]). In Figures 3A, 3B, 4B, and 7 that follow are depicted "Dot Plots" where each dot corresponds to a single channel blockade signal. The duration of channel blockade signal observed is 100 milliseconds in Fig.'s 3A, 3B, 4B, and for the small radii dots in Fig. 7. The large radii dots in Fig. 7 have duration approximately 5 seconds.

For given signal observation two simple statistics for the channel blockades can be used, based on the channel blockade signal's mean and standard deviation. (In the actual pattern recognition software the feature set is based on 150 HMM feature extracts [22]). If the channel is blockaded at a fixed level the standard deviation will describe only the fluctuation noise (typically Gaussian) of the blockade. If the channel is blockaded by a molecule at multiple levels during capture (e.g., a molecular capture with a collection of at least one transient bound state interaction) then the level switching makes much larger contribution to standard deviation evaluations (such as for the signals shown in Fig. 2). Channel modulators are much more easily discerned, as will be seen in what follows.

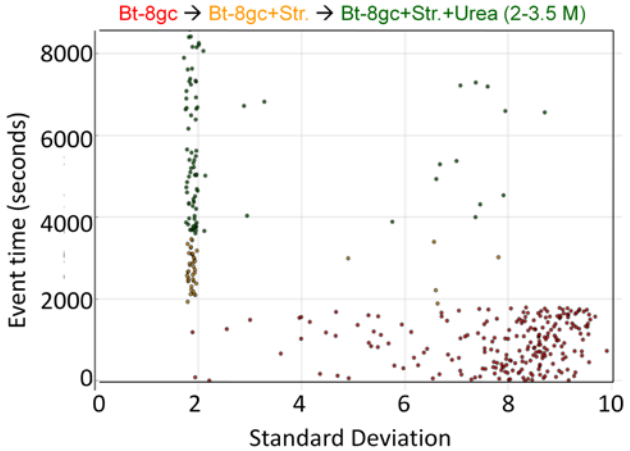

Figure 3A. Observations of individual blockade events are shown in terms of their blockade standard deviation (x-axis) and labeled by their observation time (y-axis). The standard deviation provides a good discriminatory parameter in this instance since the transducer molecules are engineered to have a notably higher standard deviation than typical noise or contaminant signals. At T=0 seconds, 1.0 μM Bt-8gc is introduced and event tracking is shown on the horizontal axis via the individual blockade standard

deviation values about their means. At T=2000 seconds, 1.0 µM Streptavidin is introduced. Immediately thereafter, there is a shift in blockade signal classes observed to a quiescent blockade signal, as can be visually discerned. The new signal class is hypothesized to be due to (Streptavidin)-(Bt-8gc) bound-complex captures.

In the Dot plot in Fig. 3A the modulator property of the unbound biotinylated DNA probe is shown as the signals with high standard deviation, where the abrupt transition to lower standard deviation on channel blockades is seen at T=2100 seconds, precisely when streptavidin is added such that the sampled biotinylated hairpins are bound (where the channel blockade of the biotinylated DNA with streptavidin bound is NOT modulatory). Results in Fig. 3B suggest that the new, bound, signal class is actually a racemic mixture of two hairpin-loop twist states. Another characteristic, apparent in Fig. 3A for hypothesized bound modulator signals after T=2100, and in Fig 3B for the two racemic trajectories, is that signal classes group as trajectories in the dot plots, as will become more apparent in deciphering the signals shown in Fig. 4B to come.

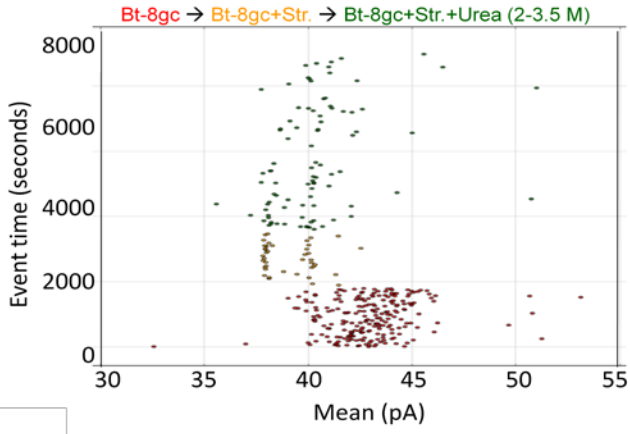

Figure 3B. As with Fig. 3A on the same data, a marked change in the Bt-8gc blockade observations is shown immediately upon introducing streptavidin at T=2000 seconds, but with the mean feature we clearly see two distinctive and equally frequented (racemic) event categories. Introduction of chaotropic agents degrades first one, then both, of the event categories, as 2.0 M urea is introduced at T=4000 seconds and steadily increased to 3.5 M urea at T=8100 seconds.

Figure 3C & 3D shows transduction of bound/unbound signals at three different transducer concentrations and a range of binding target (streptavidin) concentrations. The rescaling on counts, with the count of events at the 0.05 µM concentration Bt-8gC scaled up by a factor of 20 in Fig. 3C, for example, for comparing event rate observations at different concentrations. The good agreement of the curves in Fig. 3C strongly validates the NTD biosensing hypothesis and indicates a linear response where probe density can also provide a detector sensitivity gain in circumstances where pattern-recognition informed nanopore sampling/selection is enabled to effectively 'ignore' unbound probes.

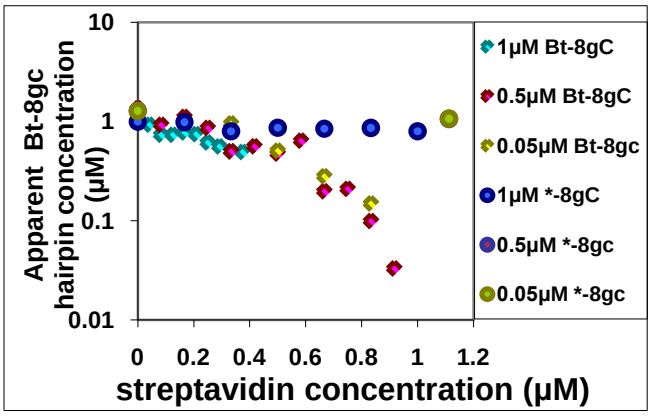

Figure 3C. The apparent Bt-8gc concentration upon exposure to Streptavidin. The vertical axis describes the counts on unbound Bt-8gc blockade events and the above-defined mapping to "apparent" concentration is used. In the dilution cases, a direct rescaling on the counts is done, to bring their "apparent" concentration to 1.0 µM concentration (i.e., the 0.5 µM concentration counts were multiplied by 2). For the control experiments with no biotin (denoted '*-8gc'), the *-8gc concentration shows no responsiveness to the streptavidin concentration.

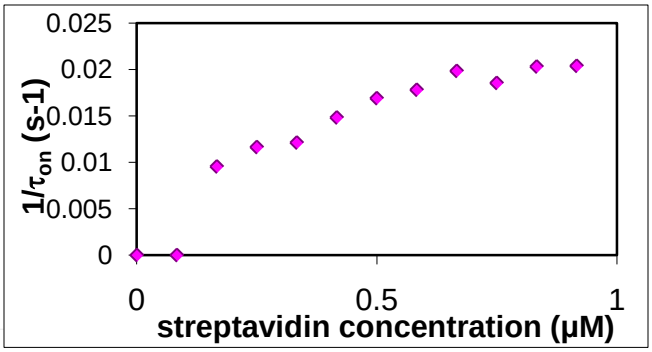

Figure 3D. The increasing frequency of the blockades of a type associated with the streptavidin-Bt-8gc bound complex. The background Bt-8gc concentration is 0.5 µM, and the lowest clearly discernible detection concentration is at 0.17 µM streptavidin.

2.2 SNP Detection – Proof of Concept

A unique, Y-shaped, NTD-aptamer is described in Fig. 4A. In this experiment a stable modulator is established using a Y-shaped molecule, where one arm is loop terminated such that it can't be captured in the channel, leaving one arm with a ssDNA extension for annealing to complement target.

A preliminary test of DNA annealing has been performed with the Y-shaped DNA transduction molecule indicated, where the molecule is engineered to have an eight-base overhang for annealing studies. A DNA hairpin with

complementary 8 base overhang is used as the binding partner. Figure 4B shows the binding results at the population-level (where numerous single-molecule events are sampled and identified), where the effects of binding are discernible, as are potential isoforms, and the introduction of urea at 2.0 M concentration is easily tolerated (a mild chaotrope) and actually helps in discerning collective binding interactions such as with the DNA annealing.

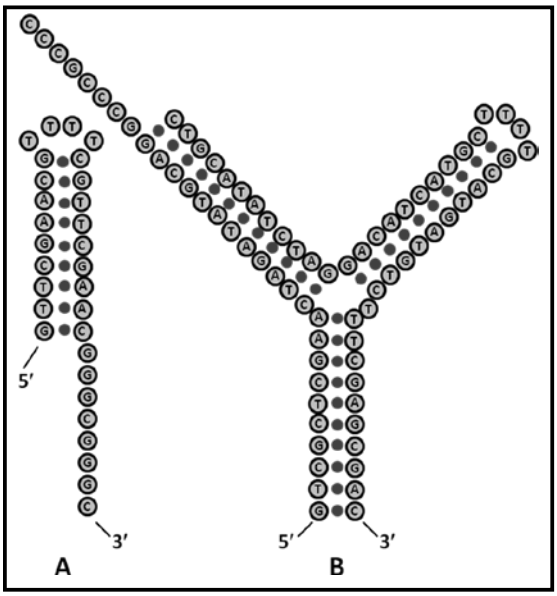

Figure 4A. The Y-Anneal transducer, and its annealing target.

Only a portion of a repetitive validation experiment involving the molecules in Fig. 4A. is shown in Fig. 4B, thus time indexing starts at the 6000th second. From time 6000 to 6300 seconds (the first 5 minutes of data shown) only the DNA hairpin is introduced into the analyte chamber, where each point in the plots corresponds to an individual molecular blockade measurement. At time 6300 seconds urea is introduced into the analyte chamber at a concentration of 2.0 M. The DNA hairpin with overhang is found to have two capture states (clearly identified at 2 M urea). The two hairpin channel-capture states are marked with the green and red lines, in both the plot of signal means and signal standard deviations. After 30 minutes of sampling on the hairpin+urea mixture (from 6300 to 8100 seconds), the Y-shaped DNA molecule is introduced at time 8100. Observations are shown for an hour (8100 to 11700 seconds). A number of changes and new signals now are observed: (i) the DNA hairpin signal class identified with the green line is no longer observed – this class is hypothesized to be no longer free, but annealed to its Y-shaped DNA partner; (ii) the Y-shaped DNA molecule is found to have a bifurcation in its class identified with the yellow lines, a bifurcation clearly discernible in the plots of the signal standard deviations. (iii) the hairpin class with the red line appears to be unable to bind to its Y-shaped DNA partner, an inhibition currently thought to be due to G-quadruplex formation in its G-rich overhang. (iv) The Y-shaped DNA molecule also exhibits a signal class (blue line) associated with capture of the arm of the 'Y' that is meant for annealing, rather than the base of the 'Y' that is designed for channel capture.

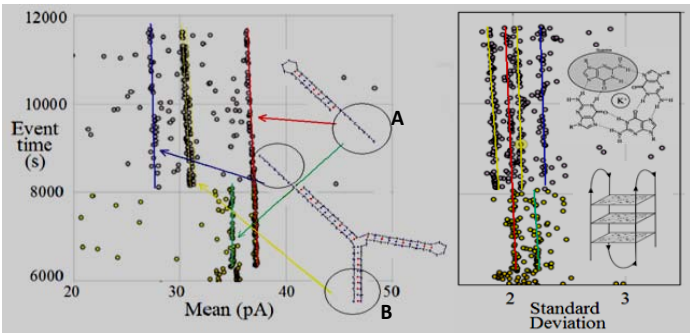

Figure 4B. Y-shaped DNA transducer with overhang binding to DNA hairpin with complementary overhang. In the Std. Dev. box are shown diagrams for the G-tetrad (upper) and the G-quadruplex (lower) that is constructed from stacking tetrads. The possible observation of G-quadruplex formation bodes well for use of aptamers in further efforts.

SNP variant detection is reduced to resolving the signals of two Y-shaped duplex DNA molecules, one with mismatch at SNP, one with Watson-Crick base-pairing match at SNP. In preliminary studies of Y-shaped DNA molecules, numerous Y-shaped DNA molecules were considered. Three variants successfully demonstrated the easily discernible, modulatory, channel blockade signals (one is shown in Fig. 5, see [9]). In those variants we considered the Y-nexus with and without an extra base (that is not base-paired), in a molecular engineering process [24].

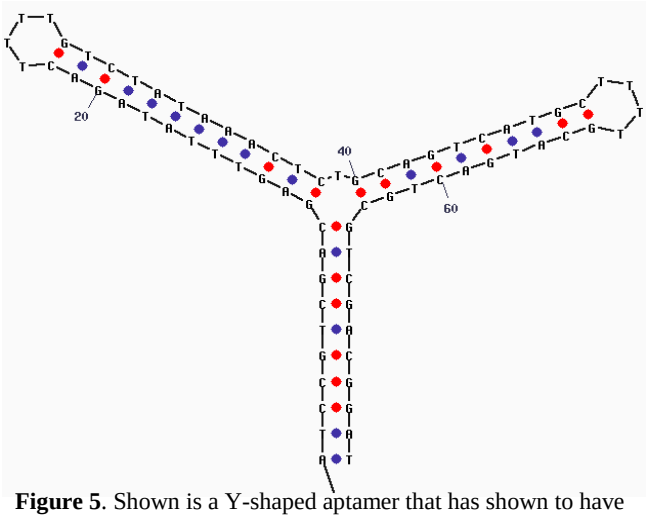

Figure 5. Shown is a Y-shaped aptamer that has shown to have capture states with the desired blockade toggling.

The determination of aptamers can be done (or initiated) via Systematic Evolution of Ligands by Exponential Enrichment (SELEX), as shown schematically on the left. What is proposed in [24] is a linkage to a

*na*nopore-detector *dir*ected (NADIR) search for aptamers that is based on bound-state lifetime measurements. NADIR complements and augments SELEX in usage: SELEX can be used to obtain a functional aptamer, and NADIR used for directed modifications (for stronger binding affinity, for example).

In using the NADIR refinement process to arrive at the Y-transducer used in the DNA annealing test in Fig. 5, we have demonstrated how *single-base insertions or modifications at the nexus of the Y-shaped molecule can have clearly discernible changes in channel-blockade signal*. We can leverage this capability using the NTD method to obtain a viable prospect for SNP variant detection to very high accuracy (possibly matching the greater than 99.999% accuracy with which the NTD can discern DNA control hairpins, shown in Fig. 2, that only differ in terminal base-pair. Classification of the DNA controls was shown to be possible with better than 99.9% accuracy in 2003 [22], where the limitation on stated accuracy is simply on acquiring large enough datasets to validate the accuracy on larger sets of data. SNP detection via *translocation*-based methods, on the other hand, must discern between two SNP variants according to the different dwell times of the complement-template annealed SNPs, until dissociation from the template allows translocation of the blockading dsDNA annealed conformation.

3 METHODS

3.1 NTD Cheminformatics Methods

A protocol has been developed for the discovery, characterization, and classification of localizable, approximately-stationary, statistical signal structures in channel current data, and changes between such structures. Along the lines of previous work in channel current cheminformatics [5,6,19,22], the protocol has three stages:

(Stage 1) primitive feature identification: this stage is typically finite-state automaton based, with feature identification comprising identification of signal regions (critically, their beginnings and ends), and, as-needed, identification of sharply localizable 'spike' behavior in any parameter of the 'complete' (non-lossy, reversibly transformable) classic EE signal representation domains: raw time-domain, Fourier transform domain, wavelet domain, etc. The FSA method that is primarily used in the signal discovery and acquisition is to identify signal-regions in terms of their having a valid 'start' and a valid 'end', with internal information to the hypothesized signal region consisting, minimally, of the duration of that signal (e.g., the duration between the hypothesized valid 'end and hypothesized valid 'start'). The FSA signal analysis methodology used here involves identifying anomalously long-duration regions, which is an extension of the ORF-finder anomalous duration reading-frame identification methods used in bioinformatics. Identification of anomalously-long duration regions in a more sophisticated Hidden Markov model (HMM) representation is possible but would require use of a HMM-with-duration and this is a much more computationally intensive method than the FSA-based approach, and typically unnecessary for purposes of simply *acquiring* the purported signal region in channel current cheminformatics [18,22].

(Stage 2) feature identification and feature selection: this stage in the signal processing protocol is typically Hidden Markov model (HMM) based, where identified signal regions are examined using a fixed state HMM feature extractor [19,22]. The Stage 2 HMM methods are the core methodology/stage in the CCC protocol in that the other stages can be dropped or merged with the Stage 2 HMM in many incarnations [7,14,18]. The HMM features, and other features (from neural net, wavelet, or spike profiling, etc.) can be fused and selected via use of Adaboost training [11]. After some tuning, the HMM-based feature extraction provides a well-focused set of 'eyes' on the data, no matter what its nature, according to the underpinnings of its Bayesian statistical representation. The key is that the HMM not be too limiting in its state definition, given the typical engineering trade-off on the choice of number of states (which impacts the order of computation via a quadratic factor on N).

(Stage 3) classification: SVMs are one of the strongest classification methods [13,15-17,22]. In part because they draw upon the powerful formalism of variational calculus that underpins much of physics and control theory, etc. If there are more classes than two, the SVM can either be applied in a Decision Tree construction with binary-SVM classifiers at each node [18,20,22], or the SVM can internally represent the multiple classes [17]. Depending on the noise attributes of the data, one or the other approach may be optimal (or even achievable). Both methods are explored in tuning, where a variety of kernels and kernel parameters are also chosen, as well as tuning on internal KKT handling protocols [18,22]. Simulated annealing and genetic algorithms have been found to be useful in doing the tuning in an orderly, efficient, manner [12,15,16]. Use of divergence kernels with probability feature vectors have been found to work well with channel blockade analysis [18,22].

Due to the molecular dynamics of the captured transducer molecule, a unique reference signal with stationary (or approximately stationary) statistics is engineered to be generated during transducer blockade, analogous to a carrier signal in standard electrical engineering signal analysis. The adaptive machine learning algorithms for real-time analysis of the stochastic signal generated by the transducer molecule offer a "lock

and key" level of signal discrimination. The heart of the signal processing algorithm is an adaptive Hidden Markov Model (AHMM) based feature extraction method, implemented on a distributed processing platform for real-time operation. For real-time processing, the AHMM is used for feature extraction on channel blockade current data while classification and clustering analysis are implemented using a Support Vector Machine (SVM). In addition, the design of the machine learning based algorithms allow for scaling to large datasets, real-time distributed processing, and are adaptable to analysis on any channel-based dataset, including resolving signals for different nanopore substrates (e.g. solid state configurations) or for systems based on translocation technology. The machine learning software has been integrated into the nanopore detector for "real-time" pattern-recognition informed (PRI) feedback [2,8]. The methods used to implement the PRI feedback include *distributed* HMM and SVM implementations, which enable the 100x to 1000x processing speedup that is needed (see Fig. 6).

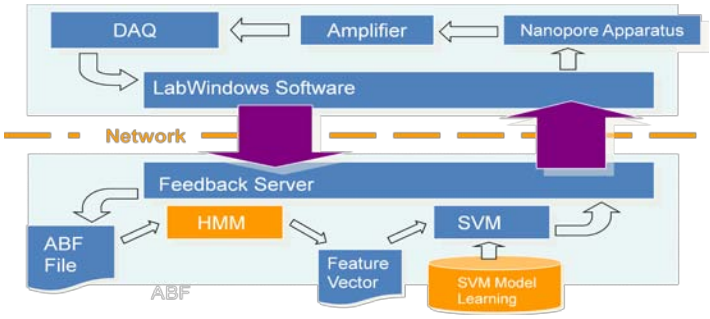

Figure 6. PRI Sampling Control (see [8] for specific details). Labwindows/Feedback Server Architecture with Distributed CCC processing. The HMM learning (on-line) and SVM learning (off-line), denoted in orange, are network distributed for N-fold speed-up, where N is the number of computational threads in the cluster network.

A mixture of two DNA hairpin species {9TA, 9GC} (from Fig. 2) is examined in an experimental test of the PRI system. In separate experiments, data is gathered for the 9TA and 9GC blockades in order to have known examples to train the SVM pattern recognition software. A nanopore experiment is then run with a 1:70 mix of 9GC:9TA, with the goal to eject 9TA signals as soon as they are identified, while keeping the 9GC's for a full 5 seconds (when possible, sometimes a channel-dissociation or melting event can occur in less than that time). The results showing the successful operation of the PRI system is shown in Fig. 7 as a 4D plot, where the radius of the event 'points' corresponds to the duration of the signal blockade (the 4th dimension). The result in Fig. 7 demonstrates an approximately 50-fold speedup on data acquisition of the desired minority species.

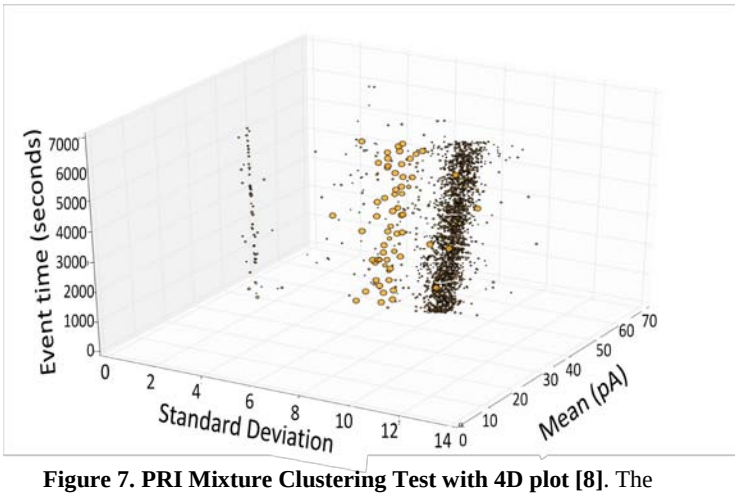

Figure 7. PRI Mixture Clustering Test with 4D plot [8]. The vertical axis is the event observation time, and the plotted points correspond to the standard deviation and mean values for the event observed at the indicated event time. The radius of the points correspond to the duration of the corresponding signal blockade (the 4th dimension). Three blockade clusters appear as the three vertical trajectories. The abundant 9TA events appear as the thick band of small-diameter (short duration, ~100ms) blockade events. The 1:70 rarer 9GC events appear as the band of large-diameter (long duration, ~ 5s) blockade events. The third, very small, blockade class corresponds to blockades that partially thread and almost entirely blockade the channel.

3.2 NTD Control Experiments

Nanopore experiments have been performed with decoy molecules, introduction of chaotropes, and introduction of a number of other buffer modifying constituents, to increase viscosity and other transport parameters as well as modify the usual pH, salt, temperature, and other standard biochemical features. In all of these efforts the strong signal resolution capabilities of a designed channel modulator were not only are retained, but the channel-captured nanopore modulator/transducer molecule is even found to stabilize the channel's integrity – the channel can't easily deform or collapse inward with the electrophoretically captured analyte in place. Impressively high concentrations of urea were observed, for example, in blockaded channel configurations that were not achieved with open channels (results not shown). Given the detection accuracy and robust detection mechanism, we believe that there is ample evidence to support a SNP detection experiment with a clinical sample.

4 CONCLUSIONS

Nanopore Transduction Detection (NTD) is a unique platform for detection and analysis of single molecules. Proof-of-Concept experiments indicate a promising approach to SNP detection in a clinical environment, via use of the channel-blockade signals produced by engineered event-transducers in the NTD approach. This provides the basis for a highly sensitive and accurate

biosensor. Given the robust selection/filtering operations possible with DNA molecules, this provides significant sensitivity for many of the DNA-based NTD transducers developed thus far, particularly in biosensing on SNP variants targeted for future diagnostic methods.

5 ACKNOWLEDGMENTS

The author would like to thank lab technicians Amanda Alba and Eric Morales for help performing the nanopore experiments and University of New Orleans students Ahmet Eren and Joshua Morrison for help with the channel current cheminformatics analysis. The author would like to thank the University of New Orleans, Children's Hospital -- New Orleans, NIH, NSF, NASA, and the Louisiana Board of Regents for research support. The author would also like to thank META LOGOS Inc., for research support and a research license. (META LOGOS was co-founded by the author in 2009 and has recently obtained exclusive license to the NTD and machine-learning based signal processing intellectual property.) The author would also like to thank Robert Adelman (CEO META LOGOS, Inc.), Andrew Peck (CEO PxBioSciences), and Mike Lewis (Professor, University of Missouri-Columbia), for insights into the potential impact of the NTD approach.

6 REFERENCES

[1] 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.

[2] Baribault, C and Winters-Hilt S. A novel, fast, HMM-with-Duration implementation -- for application with a new, pattern recognition informed, nanopore detector. BMC Bioinformatics 2007, 8 S7: S19.

[3] Bezrukov, S.M. 2000. Ion Channels as Molecular Coulter Counters to Probe Metabolite Transport. J. Membrane Biol. 174, 1-13.

[4] Bezrukov, S.M., I. Vodyanoy, V.A. Parsegian. 1994. Counting polymers moving through a single ion channel. Nature 370 (6457), pgs 279-281.

[5] Churbanov A and Winters-Hilt S. Clustering ionic flow blockade toggles with a Mixture of HMMs. BMC Bioinf. 9 S9: S13 (2008).

[6] Churbanov A, Baribault C, Winters-Hilt S. Duration learning for nanopore ionic flow blockade analysis. BMC Bioinf. 8 S7: S14 (2007).

[7] Churbanov, Alexander and S. Winters-Hilt. Implementing EM and Viterbi algorithms for Hidden Markov Model in linear memory. BMC Bioinformatics 2008, 9:228.

[8] Eren AM, Amin I, Alba A, Morales E, Stoyanov A, and Winters-Hilt S. Pattern Recognition Informed Feedback for Nanopore Detector Cheminformatics. Accepted paper in book "Advances in Computational Biology", Springer: Advances in Experimental Medicine and Biology, June 2010

[9] Gu, L-Q., Braha, O., Conlan, S., Cheley, S., H. Bayley. Stochastic sensing of organic analytes by a pore-forming protein containing a molecular adapter. Nature, vol 398, no. 6729, 1999.

[10] Howorka, S., S. Cheley, and H. Bayley, "Sequence-specific detection of individual DNA strands using engineered nanopores," Nat. Biotechnol., vol. 19, no. 7, pp. 636–639, July 2001.

[11] Iqbal R, Landry M, Winters-Hilt S: DNA Molecule Classification Using Feature Primitives. BMC Bioinformatics 2006, 7 S2: S15.

[12] Roux B and Winters-Hilt S. Hybrid SVM/MM Structural Sensors for Stochastic Sequential Data. BMC Bioinf. 9 S9, S12 (2008).

[13] Winters-Hilt S and Armond Jr. K. Distributed SVM Learning and Support Vector Reduction. Submitted to BMC Bioinformatics. (http://www.cs.uno.edu/~winters/ SVM_SWH_preprint.pdf)

[14] Winters-Hilt S and Jiang Z. A hidden Markov model with binned duration algorithm. IEEE Trans. on Sig. Proc., Vol. 58 (2), Feb. 2010.

[15] Winters-Hilt S and Merat S. SVM Clustering. BMC Bioinf. 8 S7: S18 (2007).

[16] Winters-Hilt S and Merat S. Unsupervised clustering using supervised support vector machines. Submitted to BMC Bioinformatics. (http://www.cs.uno.edu/~winters/ClustSVM_preprint.pdf)

[17] Winters-Hilt S, Yelundur A, McChesney C, Landry M: Support Vector Machine Implementations for Classification & Clustering. BMC Bioinformatics 2006, 7 S2: S4.

[18] Winters-Hilt S. Nanopore Cheminformatics based Studies of Individual Molecular Interactions. Ch. 19 in Y. Zhang and J. C. Rajapakse, editors, Machine Learning in Bioinformatics, John Wiley & Sons, 2009.

[19] Winters-Hilt S: Hidden Markov Model Variants and their Application. BMC Bioinf. 2006, 7 S2: S14.

[20] Winters-Hilt, S. and M. Akeson, "Nanopore cheminformatics," *DNA and Cell Biology, Vol. 23 (10), 2004.*

[21] Winters-Hilt, S. Single nucleotide polymorphism and pathogen detection via DNA molecules engineered to have different nanopore-blockade signal upon annealing to detection target. In preparation for submission to *BMC Biotechnology*.

[22] Winters-Hilt, S., W. Vercoutere, V. S. DeGuzman, D. Deamer, M. Akeson, and D. Haussler, "Highly Accurate Classification of Watson-Crick Base-Pairs on Termini of Single DNA Molecules," *Biophys. J.* Vol. 84, pg 967, 2003.

[23] Winters-Hilt, S., Pattern Recognition Informed (PRI) Nanopore Detection for Sample Boosting, Nanomanipulation, and Device Stabilization; and PRI Device Stabilization Methods in General. PATENT, UNO filing, August 2009.

[24] Winters-Hilt, S., Nanopore Detector based analysis of single-molecule conformational kinetics and binding interactions. BMC Bioinformatics 2006; 7 (Suppl. 2): S21.