12

Event transduction analysis of individual DNA interactions with TBP transcription-factor and HIV integrase

Stephen Winters-Hilt
University of New Orleans
USA

1. Introduction

The nanopore transduction detector is a unique platform for detection and analysis of single molecules. Proof-of-Concept experiments indicate a promising approach to pathogen, SNP, aptamer-based, and antibody-based detection, 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.

Nanopore cheminformatics binding observations for the biotinylated-DNA/straptavidin model system are shown in Sec. 2, along with preliminary SNP annealing-based detection test. 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 can be limiting in the standard approach. Standard methods for SNP detection have high sensitivity, but typically lack high specificity and versatility. 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.

Synthetic transcription factors (STFs) promise to offer a powerful new therapeutic against Cancer, AIDS, and genetic disease. A growing percentage of drugs are of this type, including salicylate and tamoxifen. STFs that can appropriately target their transcription

factor binding sites on genomic DNA provide a means to directly influence cellular mRNA production. An effective mechanism for screening amongst transcription factor (TF) candidates would itself be highly valued, and such may be possible with nanopore cheminformatics methods. To this end, TATA-binding protein (TBP) has been examined using a bifunctional "Y" shaped DNA duplex, with a TATA consensus sequence in one arm. Similarly, nanopore transduction detection of *terminal* (blunt-ended) dsDNA binding to protein, by HIV's DNA integrase, has been examined as well. For the latter, a bifunctional "Y" shaped DNA duplex with an HIV consensus terminus at one arm is exposed for interaction. Preliminary results for protein-DNA interactions, at terminal and non-terminal duplex DNA regions, are shown in Sec. 3. Further work on indivudual enzyme studies have begun via an extension of the DNA-terminus binding study in Sec. 3.2.2, where the protein is the HIV integrase (see Sec. 5.3 for discussion).

There are two approaches to utilizing a nanopore for detection purposes: translocation and/or dwell-time (T/DT) based approaches, that typically rely on blockade dwell-times; and nanopore transduction detection (NTD) based approaches, thst functionalize the nanopore by utilizing an engineered blockade molecule.

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) (Akeson et al., 1999; Bezrukov, 2000; Bezrukov et al., 1994). 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 the dwell time and a fixed blockade level. The NTD approach has multiple features, with number and type according to modulator design objectives.
(2) Versatility. T/DT: is highly engineered for detection application to a particular target, and can't be re-targeted. NTD: requires minimal preparation/augmentation to the transduction modulator platform via use of modulator-linked binding moieties (antibody or aptamer, for example) for particular target or biomarker (which are then simply linked to modulator)
(3) Speed. T/DT: Slow, especially if multiclass, since detection must differentiate by dwell-time. NTD: Fast: feature extraction not dependent on dwell-time.
(4) Multichannel. T/DT: 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.

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. 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. 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.

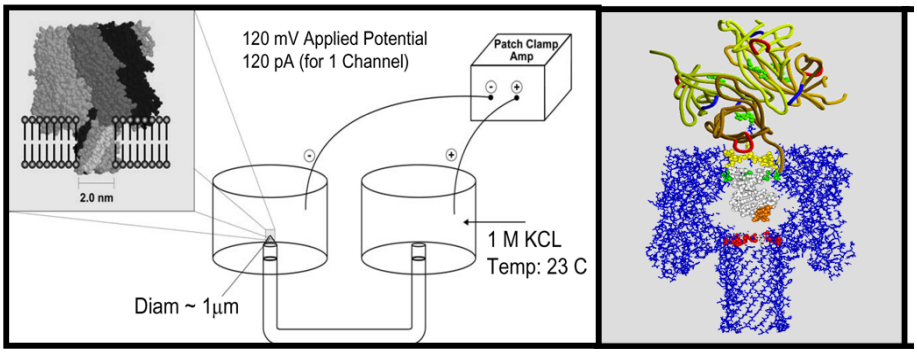

Figure 1. Left. Schematic diagram of the nanopore transduction detector. The nanopore detector consists of a single pore in a lipid bilayer that is created by the oligomerization of the staphylococcal alpha-hemolysin toxin, and a patch clamp amplifier capable of measuring pA channel currents. **Right.** Bound NTD modulator `captured' in the channel.

2. Background

2.1 Nanopore Transduction

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 those T/TD Methods that use 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 (or possible, given the covalent attachement) (Gu et al., 1999; Howarka et al., 2001). 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. 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).

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, and this is expected to be critically relevant in the next generation of multi-component, multi-cofactor, enzyme studies at the single-molecule level (see Sec. 5.3). 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), for example, 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) can also be studied much more directly using the NTD approach.

2.2 Model Nanopore Transduction Detection Systems

Three-way DNA junctions – "Y-aptamers" are used for simple nanopore-based examination with a single orientation of capture, when possible, and enhanced pattern recognition is done using channel current cheminformatics software.

The first NTD transducer examined with our Nanopore Detector (see Methods) is a bifunctional three-way DNA junction, or "Y-aptamer". Aptamers are nucleic acid species that have been engineered to bind to various molecular targets such as small molecules, proteins, and nucleic acids. Aptamers are advantageous for biotechnological applications, because they are readily produced by chemical synthesis and possess desirable storage properties. The critical bifunctionality of the NTD-aptamer is accomplished in part by certain steric constraints arising from its "Y" shape. According to our design, the blunt-ended terminus corresponding to the base of the Y will be captured and carefully perched over an internal limiting aperture in the channel detector (see Methods). The two remaining termini, the "arms" of the Y-shaped aptamer, are capped with thymine loops, preventing them from entering the narrow channel. The benefit of this particular arrangement is that now the detector can be operated in a way sensitive to binding events on the extremities of the captured DNA molecule. Thus, one or both arms can then be outfitted with a binding site to fulfill the molecule's additional function, which would allow an instance of non-terminal dsDNA binding to be appraised. In our case, a TATA box binding receptor is placed approximately at the mid-point of one of the aptamer arms (see Sec. 3.2.1). The

experiment is also repeated, with the receptor arm elongated several base pairs for more distal receptor placement from the channel environment, in order to ensure accommodation for the TATA binding protein (TBP), with similar indication of binding in our experiments. An investigation of this kind has relevance to protein-based dsDNA binding, such as for TBP and multi-component TFs in general.

For the other DNA-protein interaction studied we have placed the HIV consensus terminus at the end of the Y-aptamer arm (that was the first DNA molecule's TATA arm) – where it is exposed for binding to integrase. This allows direct examination of protein binding to the terminal DNA region, where co-factors were typically not present, thereby arresting the integrase's "clipping" function, obtaining, instead, bound integrase-DNA complexes.

In the preliminary results shown in Fig. 3, 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) (see Fig. 1)]. In repeated experiments we see the sensitivity limit ranging inversely to the concentration of detection probes (and in direct proportion to the detection window, an aspect of the stationarity of the statistrics observed). If taken to its limits, with established PRI sampling capabilities, with longer observation windows than 100 seconds, and with stock Bt-8gc at 1mM concentration, we can probe the sensitivities indicated in 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.

2.2.1 Model system based on streptavidin-biotin binding

The biotinylated DNA-hairpin (Fig. 2) is engineered to generate two signals depending on whether or not a streptavidin molecule is bound to the biotin (see Figs. 2A and 2B).

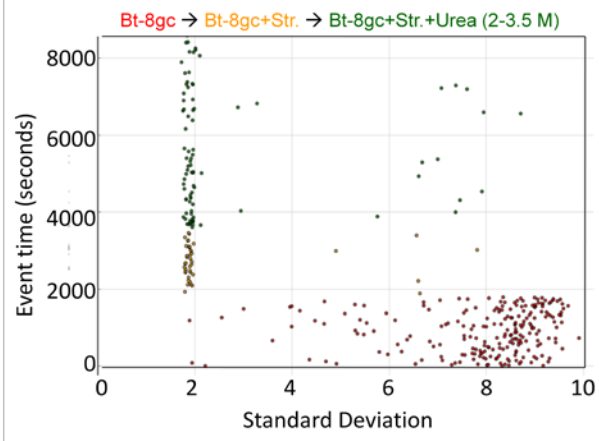

Figure 2.A**.** Observations of individual blockade events are shown in terms of their blockade standard deviation (x-axis) and labeled by their observation time (y-axis). The standard deviation provides a good discriminatory parameter in this instance since the transducer molecules are engineered to 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.

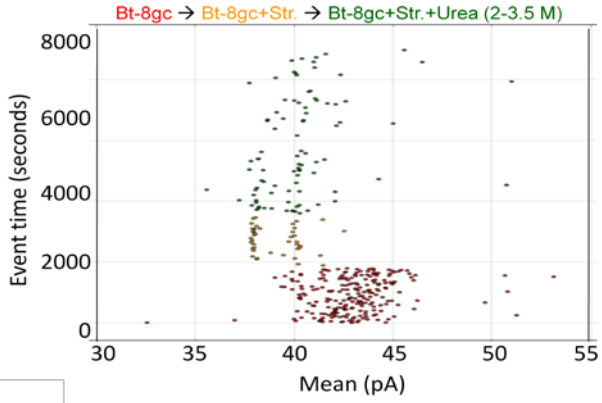

Figure2.B. As with Fig. 2A on the same data, a marked change in the Bt-8gc blockade observations is shown immediately upon introducing streptavidin at T=2000 seconds, but with the mean feature we clearly see two distinctive and equally frequented (racemic) event categories. Introduction of chaotropic 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.

Results in Fig. 2B suggest that the new signal class is actually a racemic mixture of two hairpin-loop twist states. At T=4000 urea is introduced at 2.0 M and gradually increased to 3.5 M at T=8,100.

2.2.2 Model system based on DNA annealing for Pathogen and SNP detection

A unique, Y-shaped, NTD-aptamer is described in Fig. 3A. 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.

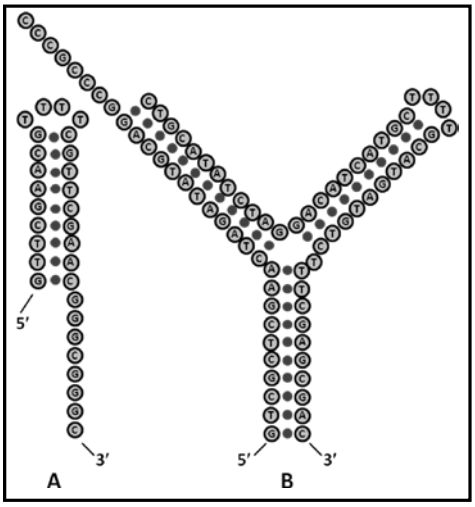

Figure 3A. The Y-Anneal transducer, and its annealing target.

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.

Only a portion of a repetitive validation experiment involving the molecules in Fig. 3A. is shown in Fig. 3B, 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. 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.

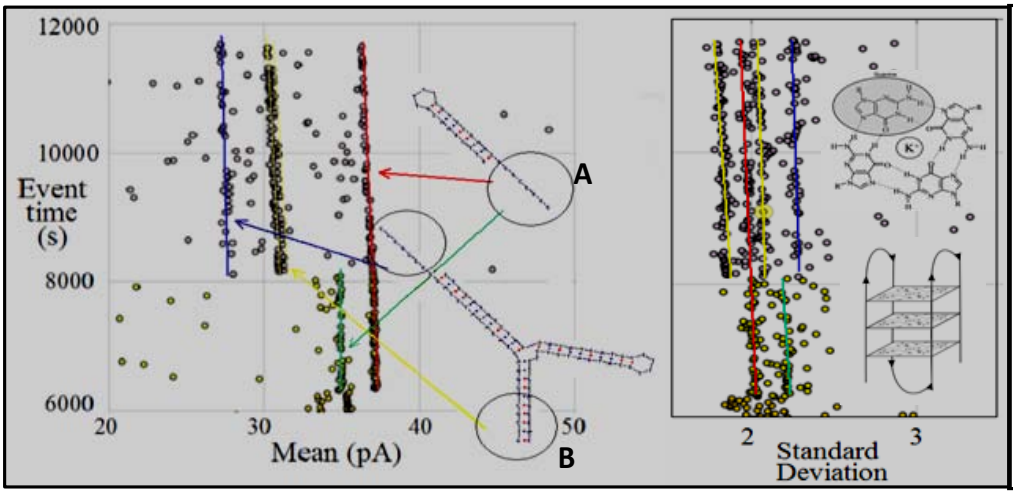

Figure 3B. Y-shaped DNA transducer with overhang binding to DNA hairpin with complementary overhang.

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. 4). In those variants we considered the Y-nexus with and without an extra base (that is not base-paired), in a molecular engineering process.

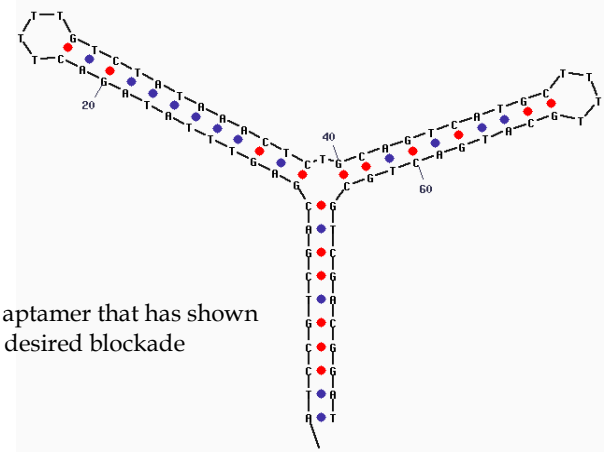

Figure 4. Shown is a Y-shaped aptamer that has shown to have capture states with the desired blockade toggling.

The determination of aptamers can be done (or initiated) via Systematic Evolution of Ligands by Exponential Enrichment (SELEX). With NTD, on the other hand, we have a *na*nopore-detector *di*rected (NADIR) search for aptamers 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. 4, 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 equaling the accuracy with which the NTD can discern DNA control hairpins that only differ in terminal base-pair, greater than 99.999%).

2.3 NTD Systems for Assaying TFs

TFs are proteins that regulate gene expression by binding to the promoter elements upstream of genes to either facilitate or inhibit transcription [10,11]. They are composed of two essential functional regions: a DNA-binding domain and an activator domain. The DNA-binding domain consists of amino acids that recognize specific DNA bases near the start of transcription. TFs are typically classified according to the structure of their DNA-binding domain, which are of one of the following types: zinc fingers, helix-turn-helix, leucine zipper, helix-loop-helix, and high mobility groups. The activator domains of TFs interact with the components of the transcriptional apparatus and with other regulatory proteins, thereby affecting the efficiency of DNA binding. A cluster of TFs, for example, is used in the preinitiation complex (PIC) that recruits and activates RNA polymerase. Conversely, repressor TFs inhibit transcription by blocking the attachment of activator proteins.

Therapeutic drugs based on STFs represent a compelling new approach to the regulation of Cancer, AIDS, and genetic diseases. The creation of STFs that can appropriately target their transcription factor binding sites on native genomic DNA provides a means to directly influence cellular mRNA production (e.g. to induce death or dormancy for cancer and AIDs cells, or restore proper cellular function in the case of genetic disease). Since the cognate TF for many binding sites remains unidentified, an automated method for screening among candidates would be a highly valuable contribution to the manufacture of medicinal TFs. Developed, as it is, to study single molecule interactions/blockades on a nanometer-scale, the nanopore detector is an ideal choice for such a task. Nanopore-based research of transcription factor binding could afford the means to quantitatively understand much of the Transcriptome. This same information, coupled with supplementary interaction information upon introduction of STFs, provides a very powerful, directed approach to drug discovery.

3. Preliminary Results

3.1 Nanopore Transduction Platform and Transduction "Carrier" Signal

Upon addition of the alpha-hemolysin monomers to the *cis*-well, according to the standard nanopore protocol, the toxin oligomerizes to form a water-filled transmembrane channel in the phospholipid bilayer. Next, the TY10T1-GC aptamer (see Fig. 5, Right) was applied through refluxing to this environment and began to engage the alpha-hemolysin channel. Upon capture of a single TY10T1-GC aptamer at the channel there is an immediate and overall current reduction. Thereafter, the steady flow of ions through the channel was alternately blockaded at levels corresponding to approximately 40% and 60% of baseline, hypothesized to correspond with the binding/unbinding of the aptamer's blunt-ended terminus to the surrounding vestibule walls. These fluctuations in ionic current were measured and recorded as a blockade pattern. The two-level dominant blockade signal is shown in Fig. 5 for T-Y10T1-GC.

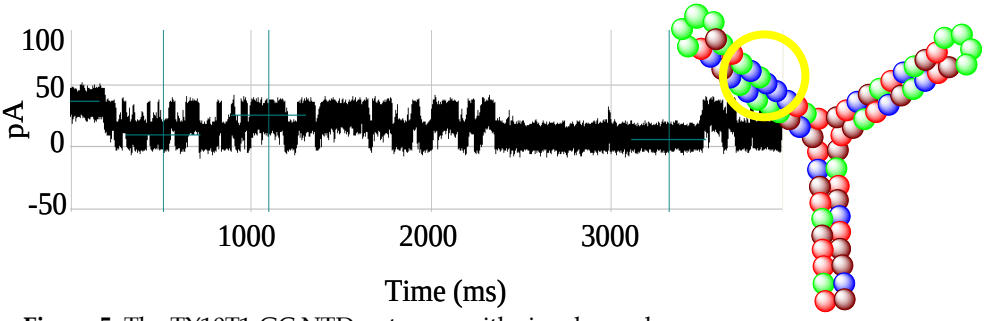

Figure 5. The TY10T1-GC NTD-aptamer, with signal sample.

3.2 Model systems for NTD Assaying

3.2.1 Aptamer-TBP model

In an attempt to demonstrate the nanopore detector's capacity for describing the transcription factor/transcription factor binding site interaction, we examined the TBP/TATA box complex following the nanopore protocol. TBP, a subunit of transcription factor TFIID, was selected for its broad commercial availability and nominal price. TFIID is the first protein to bind to DNA during the formation of the pre-initiation transcription complex of RNA polymerase II (RNA Pol II). The TATA box, located in the promoter region of most eukaryotic genes, assists in directing RNA Pol II to the transcription initiation site downstream on DNA. For our transduction molecular system, the TATA box is located on a 4dT-loop terminating arm of our Y-aptamer, which was prepared in the lab by annealing with two DNA hairpin molecules. The base stem of our bifunctional Y-aptamer is designed to target and bind the area around the limiting aperture of the alpha-hemolysin channel, while the arm containing the TATA box binds the TBP.

We find that some of the blockade signals are only seen after introduction of TBP, which is hypothesized to be the sought after indication of TBP/TATA Box complex formation. The automated signal analysis profiles for T-Y10T1-GC w/wo TBP are shown in Fig. 6. The experiment is also repeated (not shown), with the receptor arm elongated several base pairs for more distal receptor placement from the channel environment, in order to ensure accommodation for the TATA binding protein (TBP), with similar indication of binding in our experiments.

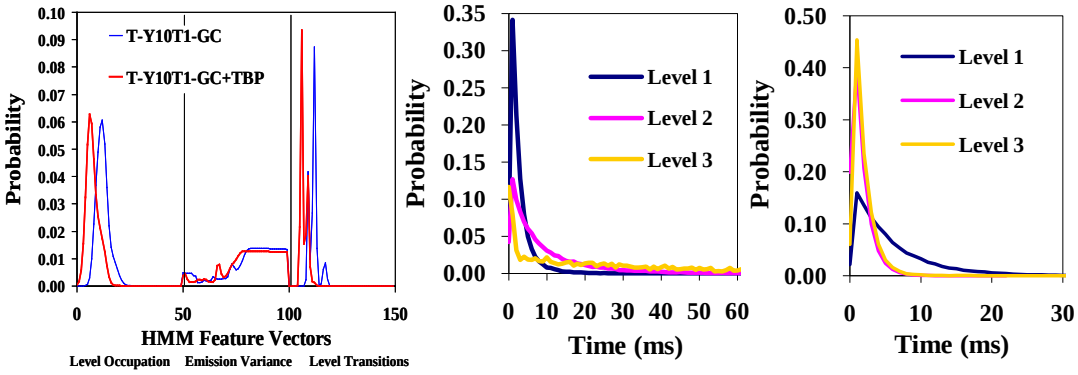

Figure 6. Left. Standard 150-component HMM-based feature extraction for collections of T-Y10T1-GC blockade signals, w/wo TBP. After the EM iterations, 150 parameters are extracted from the HMM. The 150 feature vectors obtained from the 50-state HMM-EM/Viterbi implementation in [5] are: the 50 dwell percentage in the different blockade levels (from the Viterbi trace-back states), the 50 variances of the emission probability distributions associated with the different states, and the 50 merged transition probabilities from the primary and secondary blockade occupation levels (fits to two-state dominant modulatory blockade signals). **Center.** Dwell Time at Each Level for T-Y10T1-GC (see Figure 1 to visually identify the three levels – with two dominating). **Right.** Dwell Time at Each Level for T-Y10T1-GC + TBP (sample signal blockade not shown).

3.2.2 HIV Integrase Binding with Consensus DNA Terminus Model

One of the most critical stages in HIV's attack is the binding between viral and human DNA, which is influenced by the dynamic-coupling induced high flexibility of a CA dinucleotide positioned precisely two base-pairs from the blunt terminus of the duplex viral DNA. The CA dinucleotide presence is a universal characteristic of retroviral genomes. In previous work, we observed unusual nanopore blockade activity for molecules with such a CA step, indicative of unusual binding or conformational activity. As mentioned in the Introduction, the other NTD modulator studied consists of the HIV consensus terminus at the end of the Y-aptamer arm with the TATA receptor – where it is exposed for binding to integrase. Since this molecule presents another blunt-ended dsDNA for capture, it is no surprise that such events occur. The signal analysis must separate between two classes of signal associated with these two dominant forms of capture -- associated with capture of the two blunt-ended DNA regions (at the base of the Y and at the end of the integrase-binding arm). With appropriate capture of the molecule at the base of the Y, this permits direct examination of protein binding to the terminal DNA region.

Preliminary binding observations are shown in Figures 7-9.

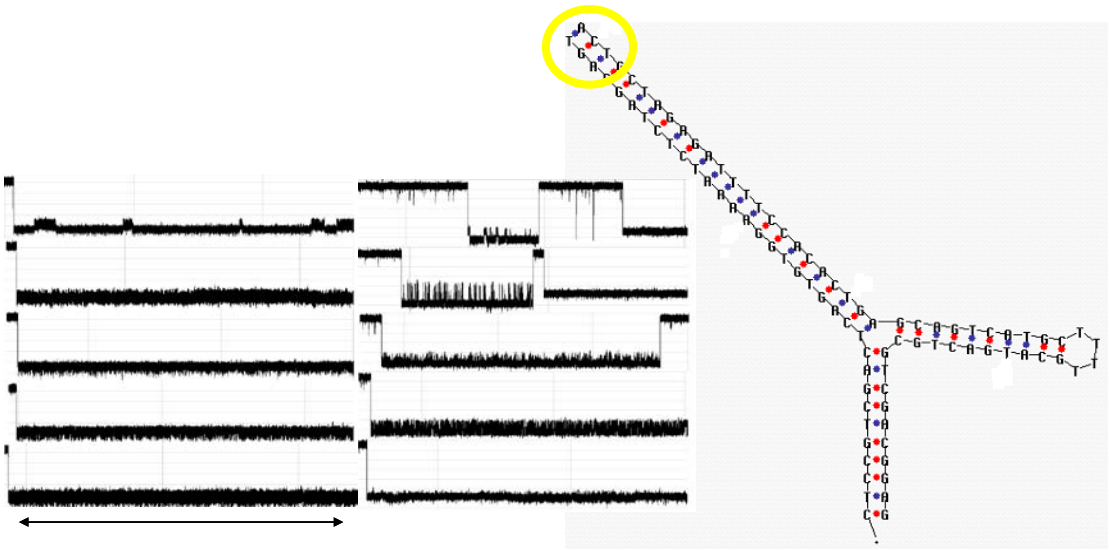

Figure 7. Right: the mfold secondary structure map of the Y-aptamer used in the integrase binding study. Integrase will bind to the blunt-ended arm shown in the yellow circle, where the HIV DNA Terminus consensus sequence has been place. **Left:** Blockade signals produced before (left) and after (right) introduction and possible binding of HIV Integrase to the HIV-corresponding terminus of one arm of a channel-captured y-shaped aptamer. The time elapsed during each frame is approximately three seconds.

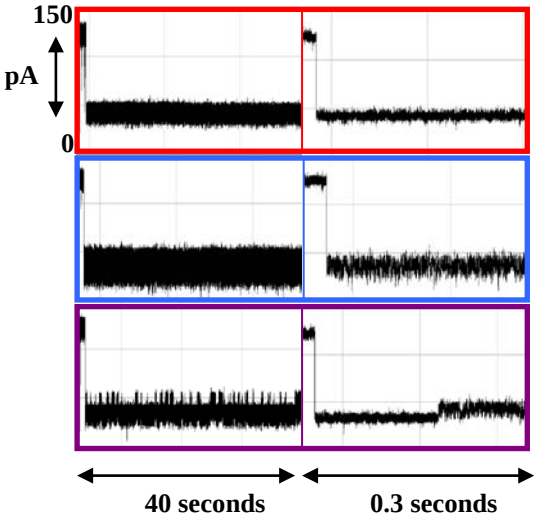

Figure 8. Three most common signal classes for HIV Y-aptamer shown in Fig. 7. Right and left boxes are identical signals shown at two different time scales.

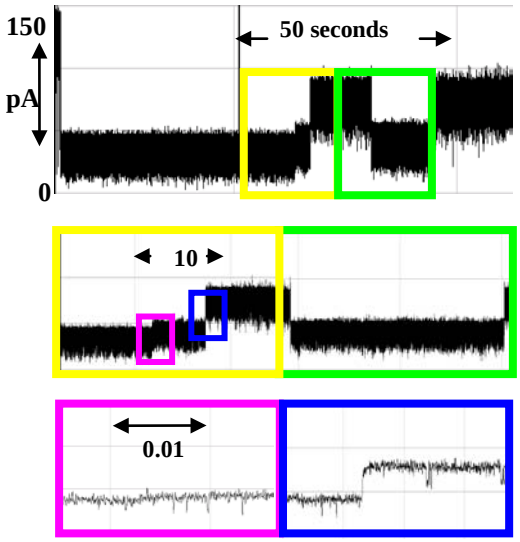

Figure 9. New type of HIV Y-aptamer Blockade Signal only seen after introduction of HIV Integrase to detector (with the Y-aptamer shown in Fig. 7 already present). Possible binding event observed at the change in signal pattern from fixed level that ends in the yellow box, (with actual end transition shown in the pink box).

4. Methods

4.1 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, 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.

(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. 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. The HMM features, and other features (from neural net, wavelet, or spike profiling, etc.) can be fused and selected via use of Adaboost training. 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. 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, or the SVM can internally represent the multiple classes. Depending on the noise attributes of the data, one or the other approach may be optimal (or even achievable). Both methods are explored in practice, where a variety of kernels and kernel parameters have been explored, as well as tuning on internal KKT handling protocols. Simulated annealing and genetic algorithms have been found to be useful in doing the tuning in an orderly, efficient, manner. Use of divergence kernels with probability feature vectors have been found to work well with channel blockade analysis.

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. 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. 10).

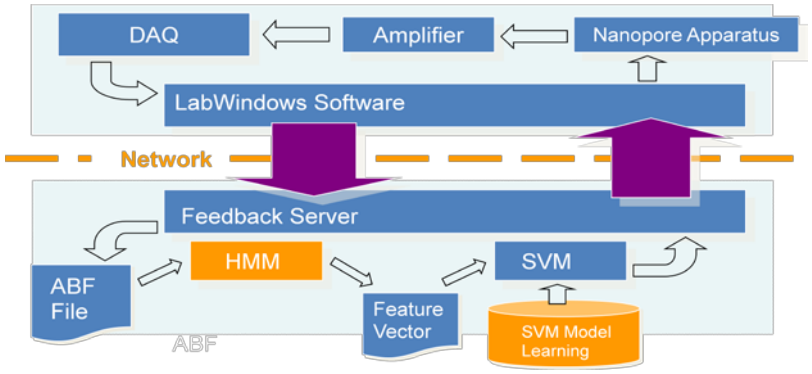

Figure 10. PRI Sampling Control. Labwindows/Feedback Server Architecture with Distributed CCC processing. The HMM learning (on-line) and SVM learning (off-line), denoted in orange, are network distributed for N-fold speed-up, where N is the number of computational threads in the cluster network.

A mixture of two DNA hairpin species {9TA, 9GC} (with identical nine base-pair stems, aside from their terminal base pairs, one with 5'-T | A-3' and one with 5'G | C-3', where " | " denotes a Watson-Crick base-pair) 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. 11 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. 11 demonstrates an approximately 50-fold speedup on data acquisition of the desired minority species. In effect, PRI offers a ‚Maxwell Demon' operational mode whereby the nanopore is always "non-blocking", e.g., the nanopore is always ready to receive, approximately, given its rapid recognition and ejection of captured analyte (100 ms or less).

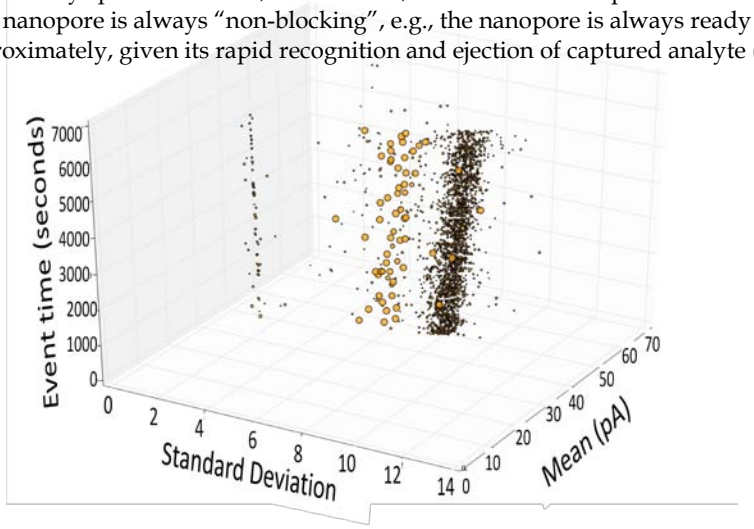

Figure 11. PRI Mixture Clustering Test with 4D plot. The vertical axis is the event observation time, and the plotted points correspond to the standard deviation and mean values for the event observed at the indicated event time. The radius of the points correspond to the duration of the corresponding 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.

4.2 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 channel modulator 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 NTD-based detection experiments with a clinical sample.

4.3 Nanopore Experiments

Each experiment is conducted using one alpha-hemolysin channel inserted into a diphytanoyl-phosphatidylcholine/hexadecane bilayer across a, typically, 20-micron-diameter horizontal Teflon aperture, as described previously [5,6]. The alpha-hemolysin pore has a 2.0 nm width allowing a dsDNA molecule to be captured while a ssDNA molecule translocates. The effective diameter of the bilayer ranges mainly between 5-25 µm (1 µm is the smallest examined). This value has some fluctuation depending on the condition of the aperture, which station is used (each nanopore station, there are four, has its own multiple aperture selections), and the bilayer applied on a day to day basis. Seventy microliter chambers on either side of the bilayer contain 1.0 M KCl buffered at pH 8.0 (10 mM HEPES/KOH) except in the case of buffer experiments where the salt concentration, pH, or identity may be varied. Voltage is applied across the bilayer between Ag-AgCl electrodes. DNA control probes are added to the *cis* chamber at 10-20 nM final concentration. All experiments are maintained at room temperature (23 ± 0.1 °C), using a Peltier device.

4.4 Control probe design

Since the five DNA hairpins studied in the prototype experiment have been carefully characterized, they are used in the antibody (and other) experiments as highly sensitive controls. The nine base-pair hairpin molecules examined in the prototype experiment 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. The eight base-pair DNA hairpin is identical to the core nine base-pair subsequence, except the terminal base-pair is 5'-G|C-3'. 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).

4.5 NTD-Aptamer Design

The Y-shaped NTD-aptamer molecule design we are currently using consists of a three-way DNA junction created: 5'-CTCCGTCGAC GAGTTTATAGAC TTTT GTCTATAAACTC GCAGTCATGC TTTT GCATGACTGC GTCGACGGAG-3'. Two of the junctions' arms terminate in a 4T-loop and the remaining arm, of length 10 base-pairs, is usually designed to be blunt ended (sometimes shorter with an overhang). The blunt ended arm has been designed such that when it is captured by the nanopore it produces a toggling blockade. One of the arms of the Y-shaped aptamer (Y-aptamer) has a TATA sequence, and is meant to be a binding target for TBP. A variant Y-shaped DNA aptamer has at one arm, instead of a 4dT loop, an HIV integrase consensusterminus sequence (for use in studies of HIV integrase inhibitors, for example). In general, any transcription factor binding site or DNA enzyme could be studied (or verified) in this manner. NADIR selection is constrained to the subset of aptamers that not only have desirable binding properties to their target, but that also have a common "base", partly consisting of a blunt 10base-pair dsDNA end that provides a sensitive "toggle" signal upon capture by the electrophoretic forces at the nanopore's limiting aperture. This is accomplished by having the terminal base-pair perch over the limiting aperture due to a Y-branching in the DNA molecule being held at the channel's outer (cis) mouth (performing a similar role to the hairpin loop in prior, DNA hairpin, studies).

In general, the determination of aptamers can be done (or initiated) via Systematic Evolution of Ligands by Exponential Enrichment (SELEX). This can be followed by a "NAnopore-detector DIRected" (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 many respects, the DNA hairpins used in previous studies], and now used as controls, can be viewed as "dumb" NTD-aptamers in that they are nucleic acids with no binding properties. For the aptamer binding studies, where the choice of DNA aptamer is under the experimenters discretion, bi-functional aptamers are described that provide the desired "toggling" blockade signal with their captured ends (like the controls), while their uncaptured regions are designed to have binding moieties for various targets, i.e., annealing of a duplex DNA overhang to its compliment.

4.6 Data acquisition

Data is acquired and processed in two ways depending on the experimental objectives: (i) using commercial software from Axon Instruments (Redwood City, CA) to acquire data, where current was typically be filtered at 50 kHz bandwidth using an analog low pass Bessel filter and recorded at 20 μs intervals using an Axopatch 200B amplifier (Axon Instruments, Foster City, CA) coupled to an Axon Digidata 1200 digitizer. Applied potential was 120 mV (*trans* side positive) unless otherwise noted. In some experiments, semi-automated analysis of transition level blockades, current, and duration were performed using Clampex (Axon Instruments, Foster City, CA). (ii) using LabViewbased experimental

automation. In this case, ionic current was also acquired using an Axopatch 200B patch clamp amplifier (Axon Instruments, Foster City, CA), but it was then recorded using a NI-MIO-16E-4 National Instruments data acquisition card (National Instruments, Austin TX). In the LabView format, data was low-pass filtered by the amplifier unit at 50 kHz, and recorded at 20 □s intervals.

4.7 Further Details on the HMM Cheminformatics Implementation

With completion of preprocessing, an HMM is used to remove noise from the acquired signals, and to extract features from them. The HMM is, initially, implemented with fifty states, corresponding to current blockades in 1% increments ranging from 20% residual current to 69% residual current. The HMM states, numbered 0 to 49, corresponded to the 50 different current blockade levels in the sequences that are processed. The state emission parameters of the HMM are initially set so that the state j, 0 <= j <= 49 corresponding to level L = j+20, can emit all possible levels, with the probability distribution over emitted levels set to a discretized Gaussian with mean L and unit variance. All transitions between states are possible, and initially are equally likely. Each blockade signature is de-noised by 5 rounds of Expectation- Maximization (EM) training on the parameters of the HMM. After the EM iterations, 150 parameters are extracted from the HMM. The 150 feature vectors obtained from the 50- state HMM-EM/Viterbi implementation are: the 50 dwell percentage in the different blockade levels (from the Viterbi trace-back states), the 50 variances of the emission probability distributions associated with the different states, and the 50 merged transition probabilities from the primary and secondary blockade occupation levels (fits to two-state dominant modulatory blockade signals).

5. Discussion and Conclusion

A new form of binding analysis/characterization is possible at the level of an individual molecular complex. This has direct relevance to biology and medicine since it can potentially provide a very direct form of binding analysis. Such a method is needed because information about molecular species, particularly their binding properties in the presence of cofactors and adjuvants, is what is often sought in drug discovery, a very complex and expensive design/discovery process. Nanopore transduction detection, thus, holds promise for helping to identify medicinal co-factors or adjuvants that lead to stronger immune response by aiding the design/selection for specific binding properties. In the results described here, TBP-TATA binding is examined (an instance of a TF/TFBS, or transcriptome, analysis), and Integrase-HIVDNA binding is examined.

Conclusion

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

NTD is also a unique platform for analysis and assaying of single molecules. Channel current based kinetic feature extraction not only appears to be practical in this regard, but

the next key step in the study of individual reaction histories. Nanopore detection promises to be a very precise method for evaluating binding strengths, assaying enzymatic activity, and observing single-molecule conformational changes. Preliminary work along these lines is shown here for integrase attachment to its DNA terminus consensus binding site. The activity of enzymes and TF binding is shown to be directly monitored, and anything interfering with the critical binding of that enzyme of TF to its substrate can, thus, be monitored, which permits new assaying capabilities.

6. Acknowledgement

The author would like to thank lab technicians Amanda Alba and Eric Morales for help performing the nanopore experiments and 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.

7. References

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.

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

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

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.

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.