SAND2007-3607P

Bayesian Analysis of Nanopore Detector Array Signals

Habib N. Najm*, José M. Ortega, Bert J. Debusschere,
Sandia National Laboratories, Livermore, CA 94550, USA.
and
Murugappan Muthukumar and Ryan Murphy,
University of Massachusetts, Amherst, MA 01003, USA.

DARPA MOLDICE Final Project Report

June 6, 2007

Abstract

Engineered membrane protein channels based on the nanopore formed by the bacterial toxin α-hemolysin (α-HL) are sensitive and selective detectors for metal ions and organic macromolecules such as proteins or DNA. The passage of the individual molecules or ions through the pore causes transient blockages in the ionic current when a voltage is applied across the membrane. The current signature (amplitude, duration of shut/open times, and frequency of blockages) provides information that can be used to identify the nature of the analytes and estimate their concentrations. The analysis of the experimental data poses a significant challenge: to predict the identity of the components of the sample (categorical prediction or classification) and their concentration (numeric prediction or regression) using large datasets of stochastic data from which the proper attributes have to be identified in order to extract useful information. We present an implementation of a Bayesian classification/regression approach for the identification of analytes and their concentration. Important properties that have been considered in the Bayesian procedure are robustness against high levels of noise and signal artifacts, detection of analyte mixtures, efficient feature extraction, and low computational cost. We also report findings from a computational study of DNA translocation through an α-HL nanopore.

Background

Nanopore detectors have generated increased interest in the biophysics and biotechnology fields for their potential applications in DNA identification and selective biosensing of heavy-metal contaminants, hazardous bioagents and toxins [3,4,6–8,11,14]. α-Hemolysin is a toxin secreted by Staphylococcus Aureus [5] that has the property of self-assembling in lipid bilayers to form stable nanopores with well defined and reproducible properties. The αHL pore is a heptameric structure with a large internal cavity that can remain open for extended periods, allowing continuous current to flow across [8]. The diameter of the pores is large enough so that, when a voltage is applied across the membrane, large macromolecules such as single stranded DNA can translocate, but not double-stranded DNA [12]. αHL nanopores also show limited ion selectivity [6], despite the fact that their size is larger than typical ion channels.

The natural biological channels can be engineered as nanopores with specific properties. Bayley *et* al. [2,4] altered the ion selectivity of α-HL by replacing residues within the lumen or by adsoring non-covalently bound internal surface layers. They showed that the ionic currents flowing through individual engineered pores are modulated by metal ion analytes, revealing both the concentration and identity of the ions. Specifically, the frequency of the individual blockages or events is related to the concentration, while the duration and

*Corresponding author: Sandia National Laboratories, 7011 East Avenue, MS 9051, Livermore, CA 94550, USA. Email: hnnajm@sandia.gov, phone: (925) 294-2054, fax: (925) 294-2595.

amplitude of blocked events are the signal attributes that can be used to identify the analyte [2,9]. Bayley *et al* [2,4] also indicate that since only one metal ion is bound and detected at any given moment, "stochastic sensing" can be used to detect mixtures of ions.

In this report we address the data analysis challenge for analyte/analyte-mixture detection using α-HL nanopores. This problem poses significant challenges in terms of extraction of information from the signal and its correlation with analyte identity and concentration, assignment of confidence and reliability measures on the interpretation of the results, and computational requirements. For example, standard approaches based on threshold crossings fail to robustly extract relevant information from the stochastic signal in the presence of multiple current levels, artifacts, or low signal to noise (SN) ratios. On the other hand, a more sophisticated feature extraction based on Hidden Markov Models [13] requires significant computational resources and presents training scalability issues.

If available, physical models can be used to correlate the attributes of the signal with parameters such as the identity of the analytes and their concentration, by providing a mathematical representation of the relationship between the properties of the stochastic ionic current signal and the molecular processes that take place as the expression of the physical and chemical properties of the analytes. The existence of such a model requires an understanding of the foundations of the biological processes during the translocation. Muthukumar *et al* [9], addressing some experimental results on stochastic sensing, computed simulations of polymer conformations and the ionic currents through protein channels using simultaneous Langevin dynamics and three-dimensional Poisson-Nernst-Planck (PNP) equations. Although the theoretical results from these simulations are valuable tools in the design of a nanopore detector, the applicability of the physical models for nanopore data analysis is, from the computational point of view, hindered by the large computational cost of the runs needed for model validation, parameter estimation, and uncertainty analysis.

To overcome these difficulties we carry out the data analysis using a Bayesian machine learning approach, i.e. an empirical training–classification scheme without the use of a physical model, and where a few attributes are selected based on their robustness, ease of extraction, as well as the information they provide. Here, the Bayesian definition of probability is used to represent the uncertainty about the relationship between the attributes of the signal and the properties of the analyte and the predictions. We explore some of the advantages of this theoretical framework, which allows for a probabilistic interpretation of the data analysis that assigns probabilities to propositions regarding the the presence of an analyte in the sample or, for continuous variables such as the concentration, probability density functions.

For two-class classifiers, the trade-off between probability of detection and probability of false alarms can be evaluated by means of Receiver Operating Characteristic (ROC) curves, which can be computed by evaluating the behavior of the Bayes factors or the ratios of posterior probabilities under different operational conditions and data.

Bayesian methods in machine learning and pattern recognition include applications such as speech and handwriting recognition, computer vision, text classification (spam filtering being the most widely recognized application), or information retrieval and data management of large datasets.

One of the main difficulties of a Bayesian solution is the need to compute multidimensional integrals, especially when they involve continuous variables. A common approach in machine learning algorithms is to discretize the continuous variables as a means of reducing the computational costs for the analysis of large data sets, as is the case in the nanopore problem.

Finally, we note that detailed computational modeling of the translocation of polymers through a nanopore can assist in both the understanding of the ongoing physical processes and the design and choice of detection attributes to be used in the Bayesian detection procedure. To this extent, we conducted detailed macromolecular computational studies of the translocation of DNA through an α-HL nanopore, and will present these results in this report.

In the following, we begin by a discussion of the underlying statistical model we use for assimilating the nanopore data. We then demonstrate the Bayesian machinery in the context of simulated nanopore data, before proceeding to the application of this construction to the analysis of experimental data for the translocation of polymers through a nanopore. We end by reporting results from our computational studies of DNA translocation through an α-HL nanopore.

1 Specification of the Statistical Model

The data analysis challenge is to extract the right information from the noisy stochastic signal and use it to detect the presence of toxic agents, identify them and estimate their concentrations. Application of Bayesian classification to achieve this goal requires the identification of specific attributes in the data that will be used to distinguish between different classes. To this end, we characterize the current signature by attributes such as the frequency of transitions from one current state to another, which indicate that a blockage is taking place, and the statistics of the durations where the current stays at a given current level, which are measures of the length of the blockage (shut time) and the time between blockages (open time). During data acquisition, the current is sampled at regular time intervals, converting the continuous signal to a discrete one. The output signal, a time series consisting of a large number of current time series values, is analyzed to detect changes in current level that indicate blockage start and end times. These events are used to characterize the frequency of transitions and the statistics of the shut and open times.

In the Bayesian context, probability distributions are used to represent what is known about each variable of interest. Bayesian inference then proceeds by formal application of methods for updating probability distributions. Bayes theorem, a statement about conditional probabilities, is central to the approach. Simply put, inferred information about unknowns of interest M (in the present context $M = \{\text{Agent(s)ID}, \text{Concentration(s)}\}$), given data, is given in terms of probability density functions (PDFs) by a straightforward application of Bayes Theorem [10]:

$$p(M|\text{data}) = \frac{p(\text{data}|M) \cdot p(M)}{p(\text{data})} \tag{1}$$

where $p(M)$ is the *prior*, encapsulating all prior information on M, $p(\text{data}|M)$ is the *likelihood* function, expressing the extent to which the model predictions fit the data, and $p(M|\text{data})$ is the *posterior* PDF, representing inferred information about M.

To construct the likelihood function $p(\text{data}|M)$ two different approaches can be considered. The first one relies on a model based on the understanding of the physics of the system. Such a mechanistic model provides a mathematical relationship between the stochastic properties of the signal and quantities of interest such as concentration, and requires an accurate detection and characterization of the events (amplitude, duration) both during the phase of parameter estimation and during the operation of the detector. Low Signal-to-Noise Ratio (SNR) and the presence of multiple current amplitudes within the same blockage period make it more challenging to extract the relevant attributes and can negatively affect its effectiveness. Hidden Markov Models have been proposed for feature extraction [13], but they tend to be extremely expensive computationally. The second approach, implemented here, is based on a (supervised) machine learning approach, where no physical model is used to relate the stochastic properties of the signal to physical properties such as concentration. This approach allows for the possibility of implementing signal processing algorithms that rely on indirect measurements of the attributes. While this method generally requires more training data, its main advantages are an improved robustness against high levels of noise and the ability to easily handle intermediate current levels that are challenging for a feature extraction approach.

We use a statistical model to characterize the frequency of the transitions and the duration of the shut/open times. Regarding the first attribute, consider a system with two well defined current levels associated with both states of the pore, open and shut. Taking as a reference one of the levels, let N be the maximum number of transitions that can take place for a dataset acquired at a given sampling rate. $N = N_s/2$ if the total number of current values sampled N_s is even and $(N_s - 1)/2$ otherwise. If n is the actual number of observed transitions, then the transition frequency is the ratio n/N We model the frequency n/N as a random variable whose behavior can be characterized by a binomial distribution with parameters N and θ

$$n \sim \text{Binomial}(n|\theta, N) \tag{2}$$

assuming that the transitions are statistically independent and the probability θ of a transition taking place is the same for all transitions.

The durations of the open and shut periods are continuous variables that are discretized in a manner that achieves efficient computation for fast inference and classification in situations with large data sets and real-time signal processing. The optimal number of bins to use in the discretization is a trade-off between computational efficiency and the ability to capture the relevant characteristics of the distribution. Further, the assumption of attribute independence of Naïve Bayesian Classifiers is used for probabilistic learning and inference. This assumption is justified since, intuitively, the time a molecule spends translocating through the nanopore (shut time) is not related to the time the nanopore is unoccupied (open time). We also assume that the pore is occupied by only one molecule at a time. Considering the assumption of attribute independence, and the evident computational advantages of discretization, we model the statistics of the shut/open intervals using the multinomial distribution:

$$\boldsymbol{n} \sim \text{Multinomial}(\boldsymbol{n}|\boldsymbol{\theta}) \qquad \boldsymbol{n} = \{n_1, n_2, \ldots, n_k\} \qquad \boldsymbol{\theta} = \{\theta_1, \theta_2, \ldots, \theta_k\} \tag{3}$$

where k is the number of intervals or bins, $\boldsymbol{\theta}$ the associated vector of probabilities and $\boldsymbol{n}$ the corresponding vector of counts.

Finally, given the assumption that the attributes are statistically independent, the likelihood function is the product of the individual predictive probabilities:

$$p(\text{data}|M) \equiv \prod_i p(\text{attribute}_i|M), \tag{4}$$

with the parameters of the distribution for each attribute obtained from trainig data.

Under this supervised learning approach, the predicted likelihood could be used to evaluate the posterior (discrete) probabilities of each of the labeled training classes for a given data set with an unknown analyte and concentration. The class with the highest posterior probability would be selected by a Maximum A Posteriori (MAP) classifier as the matching class. Further, in the present work, assuming the absence of any specific prior information, the prior probabilities will be set equal for all classes.

2 Bayesian Analysis of QuB Simulated Data

2.1 Classification

We tested our approach using simulated data that mimics the behavior of the binary mixture Zn-Co in an α-HL pore. This particular mixture has three characteristic levels in its current signature corresponding to the states where the pore is either open or occupied by Zn or Co. The simulated stochastic data was generated using the QuB Software Suite [1]. The training data consisted of simulations carried out for the values of concentration in the 2D 7×7 grid formed by the points Co={0.0, 0.3, 2, 4, 6, 8, 10} μM and Zn={0, 5, 30, 60, 90, 120, 150} nM. The attributes are the observed transitions from the upper and lower current level to any other level and the duration of the current gaps both at the bottom and the top. The characterization of the current signature using these attributes is simple and fast. A first test evaluated the performance of each of the attributes to infer the unknown concentration of one analyte in data runs where the concentration of the other analyte was known. The results indicate that, for Cobalt, the width and frequency of top-level gaps are the most useful attributes, while for Zinc the width and frequency of bottom-level gaps provide the most information.

We also evaluated the posterior PDF, $p([Co], [Zn]|data)$ in a test where the agent identities are known, but not their concentrations. Figure 1 shows the MCMC simulation of the posterior PDF plotted for test samples with pure Co, pure Zn and a mixture of Co and Zn. We use a uniform prior for log concentration. In all cases the PDF's are highly peaked and well centered on the true values, showing the validity of the approach. We tested the algorithms with increased noise (standard deviation increased by a factor of 1.5–3) in the stochastic signal by adding white noise to the original signal (in both training and testing). The results are also shown in Figure 1. The posterior PDF is less informative, as expected, but is still in the vicinity of the true values, which illustrates the robustness of the approach.

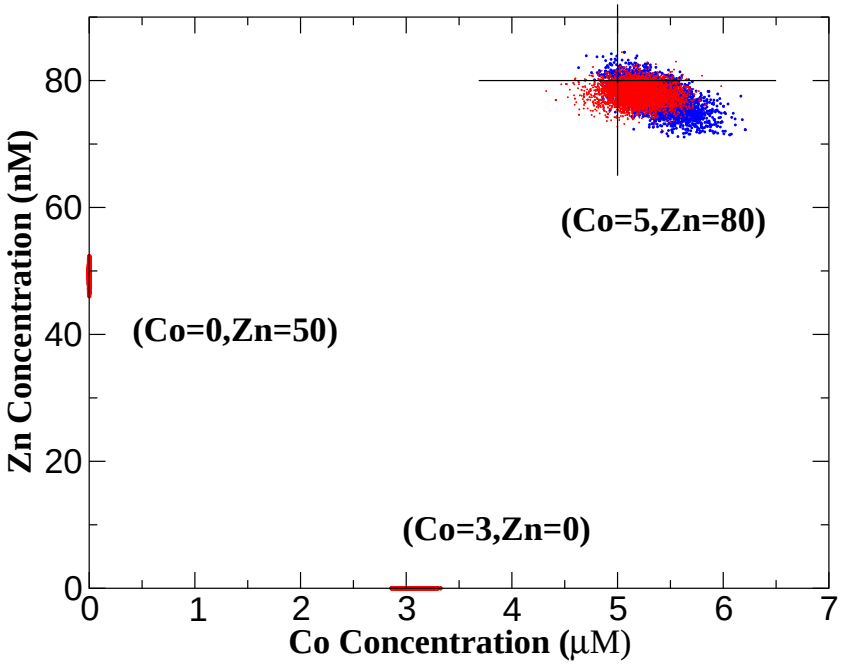

Figure 1: Scatter plots resulting from the MCMC procedure in the ([Co],[Zn]) plane for cases with pure Co or pure Zn, along with two mixture cases. Blue scatter points denote the mixture case with the higher noise level.

2.2 Analysis of the Detection Performance using ROC Curves

In signal detection theory, the Receiver Operating Characteristic (ROC) curve of a sensor is a quantitative tool used to analyze its key operational parameters. ROC curves are well known and widely used in medical decision making, although their origin can be traced back to early research on radar signals. The interpretation of the radar signals is a binary classification problem based on noisy data where the possible outcomes are

1. Stimulus present and positive response: hit
2. Stimulus present and negative response: miss
3. Stimulus absent and positive response: false alarm
4. Stimulus absent and negative response: correct rejection

The ROC curve of a sensor shows the trade-off between sensitivity (true positives) and specificity (false negatives) for different values of a threshold/cutoff value. This cutoff value is a control variable that must be properly selected in order to optimize the performance of the classification. There are two types of errors that can be made, false positives (false alarm) and false negatives (misses). Increasing sensitivity increases the probability of false positives, and at the same time decreases the probability of false negatives. On the other hand, too little sensitivity decreases the probability of false positives, but at the cost of increasing the probability of misses. Neither of these extremes are optimal. Response time is also reflected in the ROC curve, although not explicitly. Fast response times increase the probability of false alarms and/or misses because of the amount of data collected, technological factors, and computational requirements of signal processing algorithms. The ROC curve captures this trade off by representing the sensor sensitivity vs. probability of false alarm (1 - specificity) for a binary classifier system as its discrimination threshold is varied. One can construct such a ROC curve using experimental data from the sensor by plotting the fraction of true positives (sensitivity) vs. the fraction of false positives (1 - specificity).

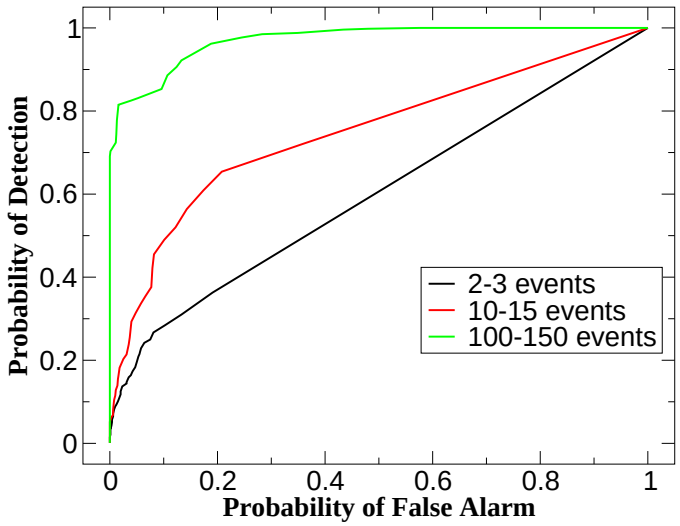

Figure 2: Discrimination between Co=3 μM, Zn=30 nM and Co=0.3 μM, Zn=5 nM.

ROC curves are used to determine visually the quality of the detector. The best possible scenario is a sensor with 100% sensitivity (all true positives are found) and 100% specificity (no false positives are found). The ROC curve would be represented by 3 points: the origin, the upper left corner $(0,1)$ and the upper right corner $(1,1)$, with the area under this curve being 1. On the other extreme, a completely unreliable detector would be represented by a straight line from $(0,0)$ to $(1,1)$, i.e. when the probabilities of true and false positives are identical independently of the discrimination threshold.

A natural choice to deal with the 2-class classification problem is to evaluate the log of the Bayes Factor or ratio of posterior probabilities. Let P and N be the two classes, and y_P and y_N the data we would observe if classes P or N are true respectively. For each of these situations, the log of the Bayes Factor is

$$\log B_{PP} = \log \frac{p(P|y_P)}{p(N|y_P)} \qquad \log B_{PN} = \log \frac{p(P|y_N)}{p(N|y_N)}$$

and the associated predictive densities, obtained by integrating over the sample space are

$$p\left(\log B_{PP}\right) = \int_{\mathbf{Y_P}} p\left(\log B_{PP}, y_P\right) dy_P$$
$$p\left(\log B_{PN}\right) = \int_{\mathbf{Y_N}} p\left(\log B_{PN}, y_N\right) dy_N$$

The objective is to design a detector that will perform well (high sensitivity, high specificity) by selecting the optimal operational and control variables, before the data is observed. By plotting $(1 - \texttt{Specificity})$ vs. $\texttt{Selectivity}$, where

$$\texttt{Specificity}(x) = \int_{-\infty}^{x} p\left(\log B_{PN}\right) \mathrm{d}\log B_{PN}$$
$$\texttt{Selectivity}(x) = \int_{x}^{\infty} p\left(\log B_{PP}\right) \mathrm{d}\log B_{PP}$$

for all values of x, with x being the detector threshold value, we obtain the ROC curve. This ROC curve is therefore a graphical tool to measure the performance of the detector under different conditions. We have computed the ROC curves for 2-class classifiers for different samples sizes and average number of transitions for the detection of different mixtures of (Co:Zn) against almost zero levels. The results, shown in Figure 2, indicate that on the order of 100 events are sufficient for excellent detector performance.

3 Nanopore Detection of Protein Complexes

Having demonstrated the efficacy of this approach with simulated data, we now analyze the data obtained from experiments conducted using an α-Hemolysin channel based detector.

3.1 Materials and Methods

3.1.1 Materials

The analytes are protein complexes with chains that have contour lengths longer than the beta-barrel of the α-hemolysin pore. All samples were prepared from polystyrene sulfonate (NaPSS) purchased from Scientific Polymer Products (Ontario, NY) with molecular weights $M_w = 1,600$ g mol^{-1}, $16,000$ g mol^{-1}, $57,500$ g mol^{-1}, $100,000$ g mol^{-1}, and $500,000$ g mol^{-1} with polydispersity indices of $1.12, 1.13, 1.10, 1.17$, and 1.24, respectively. All molecular weights were used as delivered. Samples were prepared using a 1.0M KCl/5mM HEPES buffer solution ($pH = 8$) and mixed well before being tested. Double deionized water with resistivity of 18MΩ cm (Millipore) was used as the solvent for all measurements. The lipid used was diphytanoyl-PC (Avanti Polar Lipids), while the protein used was α-hemolysin toxin from Staphylococcus aureus (Calbiochem) and both were used without any further purification.

3.1.2 Bilayer Formation and Translocation Protocol

Translocation experiments were done utilizing a horizontal bilayer apparatus similar to those used in [2–4]. The lipid membranes were formed across a $< 100\mu$m aperture that was mechanically punched into one end of a Teflon tube, known as the *cis* side. The bilayers formed had $> 200G\Omega$ resistance, and were stable for many hours at voltages up to 180mV. Using Ag-AgCl electrodes, a 5mV, 60Hz square wave seal test was applied across the membrane to test membrane stability. Once a stable membrane was formed, the protein was then introduced to the *cis* chamber (≈ 0.04 μg) and allowed to self assemble to form a pore. A single pore yields a current of 1pA/mV of applied voltage in the above-described buffer. Next, $2.5 - 15\mu$L of a 0.5 g/L NaPSS buffer stock solution was added to the *cis* chamber, and events were recorded with a 3μs sample rate, and filtered at 10kHz with a low-pass Bessel filter. The molecular weight mixture solutions where prepared at different weight fraction ratios, with the total weight of polymer in each sample being kept at 15mg. All experiments were carried out inside a Faraday cage in order to ensure low ambient noise. Ionic current was recorded using an Axopatch 200B integrating patch clamp amplifier (Axon Instruments, CA) in the voltage-clamp mode. Analysis was carried out using pClamp9.2 software (Axon Instruments, CA) in the single-channel mode.

3.2 Experimental Data

For this salt solution, the baseline current for a single open pore is 1pA/mV. The mixture data consists of a 1:1 mixture (0.25 g/L each MW) of 57 k and 500 k NaPSS, in the same buffer solution. The applied voltage is 150 mV. The voltage is slightly different across the experiments to ensure the stability of the protein complexes, as 10 mV can mean the difference between stable and unstable.

The baseline current for the open pore does vary from experiment to experiment. The open pore current should be about 1 pA/mV, thus 160 mV should correspond to a 160 pA open pore current. This does not always happen, however. The exact reason is unknown, but could be due to slight gating of the pore, solution conditions, etc. Due to this fact, the current data is normalized as I/I_o, where I is the current during the blockade, and I_o is the baseline current. Representing the data this way accounts for the changes in baseline current between different voltage experiments, as well as for the change in baseline current for the same voltage.

The data for single molecular weight (MW) samples is obtained for a concentration $0.5g/L$. For the binary mixture data, the concentration of the individual polymers is 0.25 g/L, combining for a total concentration of 0.5 g/L.

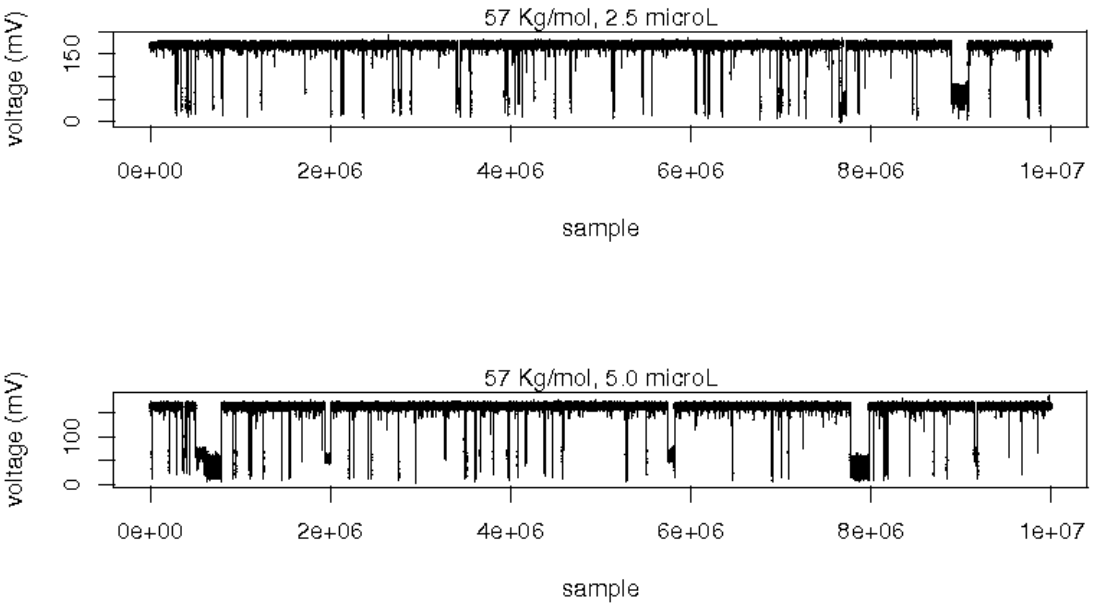

Figure 3: Examples of stochastic signals using a synthetic polymer (MW=57 Kg/mol), for concentrations 2.5 and 5.0 μL.

The experiments suffer from voltage/stability issues. The membranes are essentially spread over a 100 μm hole, made of Teflon. A long pretreatment step is necessary before any translocation experiment can be done in order to ensure that the lipid will bind to the Teflon. The pretreatment step consists of mixing the same lipid to be used to form the membrane with an organic solvent, applying a 15 μL concentration to the hole, blowing it away from the hole, then letting it dry for 1 - 2 hrs. Thus, some days the pretreatment is better than others. This is one of the main factors in the stability of these pores. As the applied voltage gets larger, there is a greater stress imparted to the membrane/pore complex. Thus, on a given day with a less than ideal pretreatment, the stability can suffer, even within 10mV difference in voltage, especially at higher voltages. There is also the possibility that the pore insertion is not ideally stable. If the protein complex has inserted enough to allow flux, but is still a bit askew or off, an increase in voltage will exploit this nature and cause larger signal noise. These are some of the biggest drawbacks to this method, and one of the main reasons why not many people are doing this experiment any longer. However, this pore complex does form a diameter smaller than any synthetic pore to date, so, as far as analyzing single stranded, flexible/semiflexible polyelectrolytes or small analytes, this pore has the upper hand. It is natural for the baseline to shift anywhere from 10-15% for a given voltage on different days.

3.3 Classification of Samples with Different Concentration

We now discuss the classification results for data obtained in the laboratory for the same polymer (Molecular Weight 57 Kg/mol) at two available concentrations, 2.5 and 5 μL. Sample data is shown in Figure 3.

Despite the large size of the available data files (a total of 4GB: 1.6GB for 2.5μL and 2.4GB for 5μL), the number of events is relatively low. The sampling frequency rate is 3μs, and the total duration of the experiments is on the order of 450 and 730 seconds, respectively. We assume that the experimental conditions at which this data is obtained are representative of all possible variability. The classifier is trained with part

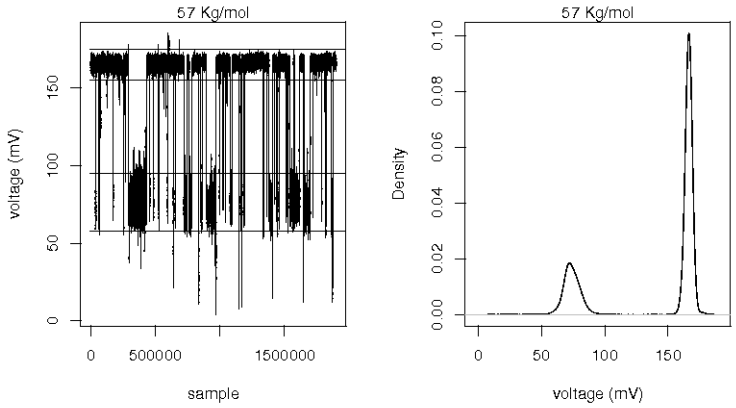

Figure 4: Typical time series signal (left), and probability density function for the signal amplitude (right).

of the data, and the rest is used for testing. To ensure that the variability is represented in the training set, the original data is divided into small chunks from which a fraction of the data is selected randomly. The testing is carried out with the original data, but without the subset that was used for training. For the results shown here, a third of the data available, for each concentration, is used for training . The testing is carried out by selecting a contiguous block (10% of the remaining data) for each concentration.

The attributes that we consider are the statistics of the gaps in the baseline (upper current level) and the statistics of the gaps of the predominant lower current level (i.e. the statistics of small current changes within each blockade), see Figure 4. Although the frequency of the events (blockades) is an attribute that is generally strongly related to the concentration, the number of events in these datasets is too small to obtain reliable statistics on this attribute.

For test samples taken from each concentration we compute the logarithm of the Bayes factor or ratio of posterior probabilities. This test is repeated 1000 times, selecting the training and the testing datasets randomly. The densities of the posterior odds are shown in Figures 5. The top row shows the densities of the log of the Bayes Factor or ratio of posterior probabilities between Class 1 (2.5 μL) and Class 2 (5 μL), i.e. B_{12}, when the data used for the classification is obtained from experiments where the true concentration is 2.5. The bottom row, similarly, shows the densities of the logarithm of ratio of posterior probabilities (Class 2 vs Class 1), i.e. B_{21}, when the testing data is from the experiments at a concentration $C = 5\mu L$ (i.e. Class 2). We can observe that the classification is not very reliable when only the statistics of the upper current level is used as attribute. On the other hand, the lower current level attribute provides enough information to discriminate between both classes: the logarithm of the Bayes Factor is in the $20 - 40$ range, showing strong evidence in favor of the true concentration in each case.

3.4 Classification of Samples with Different Polymers

Similarly we tested the performance of the training-classification with polymers of different molecular weight with identical concentration. The 4 classes are

1. Molecular Weight 57 Kg/mol
2. Molecular Weight 100 Kg/mol
3. Molecular Weight 500 Kg/mol
4. Mixture MW 57 Kg/mol + MW 500 Kg/mol (50% molar)

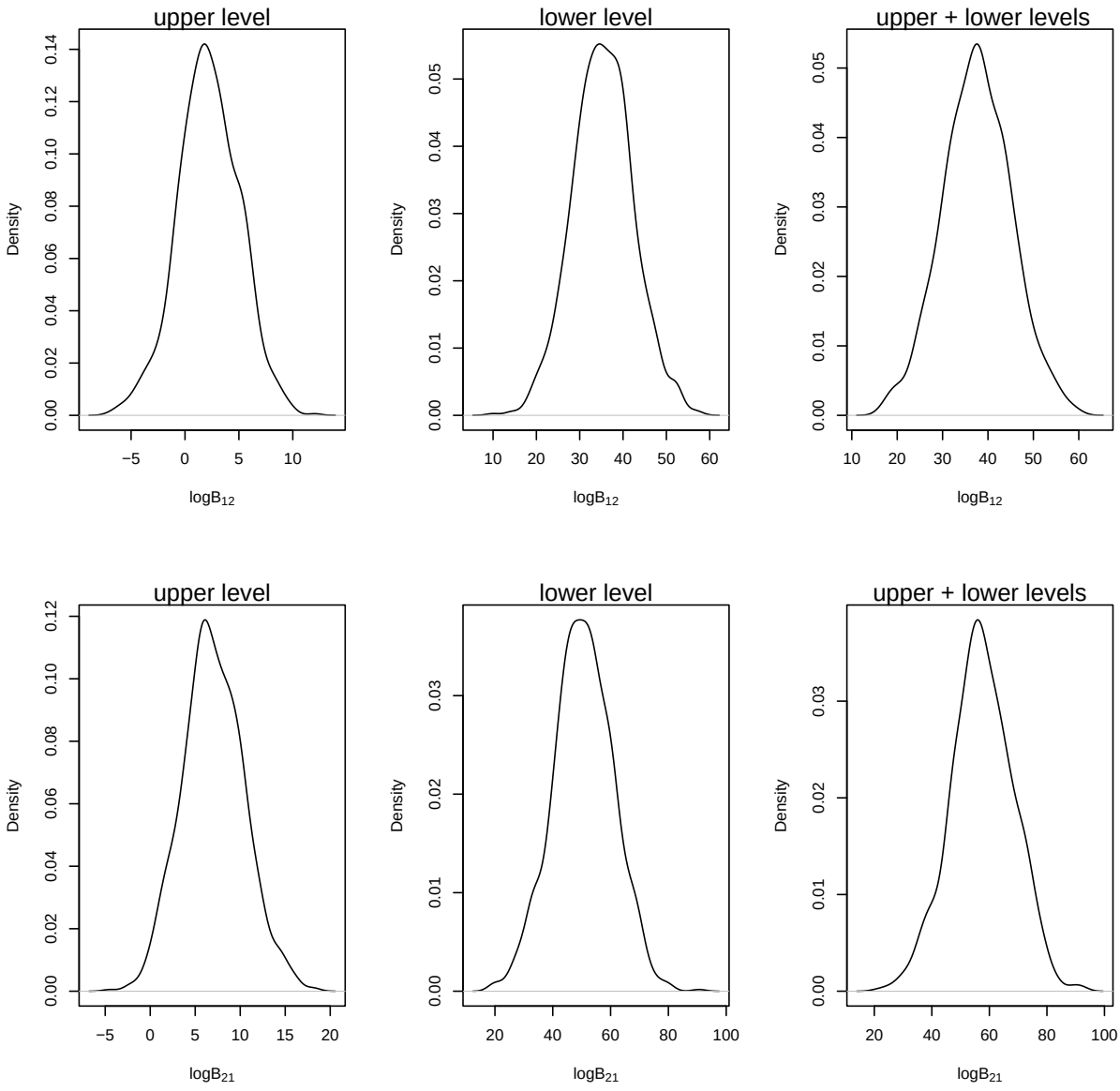

Figure 5: Probability density functions for the logarithms of the Bayes factors B_{12} (top row) and B_{21} (bottom row) to distinguish between different concentrations of the analyte. Each row illustrates the relative utility of the upper and lower level signal attributes, as well as the combination of the two attributes, for the effective discrimination between the two classes. The larger B_{12}, the higher is the confidence in the conclusion in favour of Class 1 versus Class 2, and similarly for B_{21}.

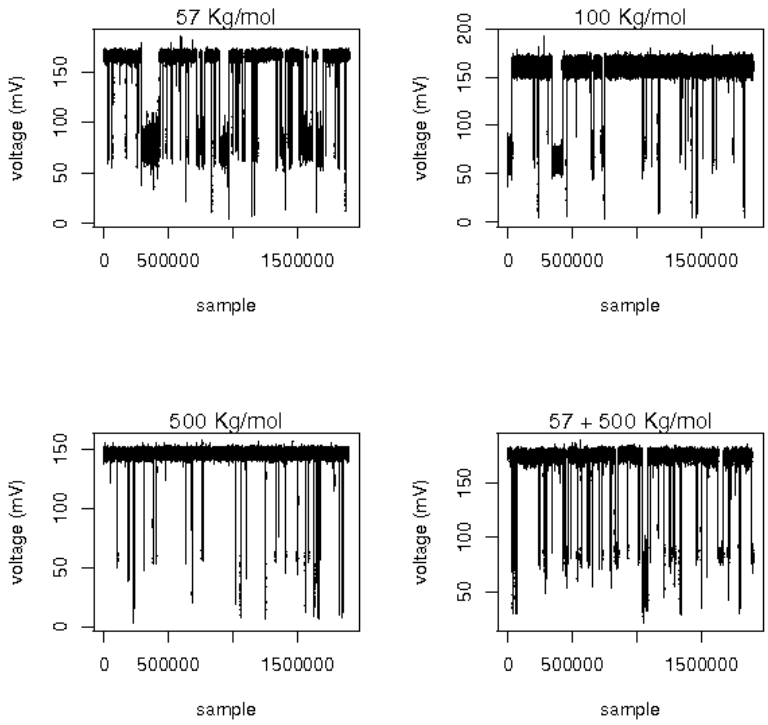

Figure 6: Examples of stochastic signals for each of the polymer samples. The number of data points shown is 2 million, which corresponds to approximately 6 seconds.

The available data (example shown in Figure 6) volumes and time spans are as follows :

- 57 Kg/mol: 2.6GB, 150 million points, 500 seconds
- 100 Kg/mol: 1.2GB, 65 million points, 216 seconds
- 500 Kg/mol: 1.6GB, 89 million points, 297 seconds
- Mixture 57+500 Kg/mol: 1.8GB, 99 million points, 330 seconds

As in the previous section, a third of the data is used for training purposes, and the classification test is carried out with 10% of the remaining available data, for each class. Figures 7, 8, 9, and 10 show the probability density of the logarithm of the Bayes factor for each pair of classes, after repeating 1000 times the classification test with arbitrarily training and test datasets. The attributes used to characterize the signal are again the statistics of the gaps in both upper and lower current levels. The range of values of the posterior odds show the strong evidence in favor always of the true class.

Modeling

We also conducted computational studies, focusing on macromolecular modeling of the translocation of single-stranded DNA molecules through an alpha-hemolysine (AHL) channel embedded across a phospholipid membrane, in an electrolyte solution under an externally applied electric field. We computed simultaneously

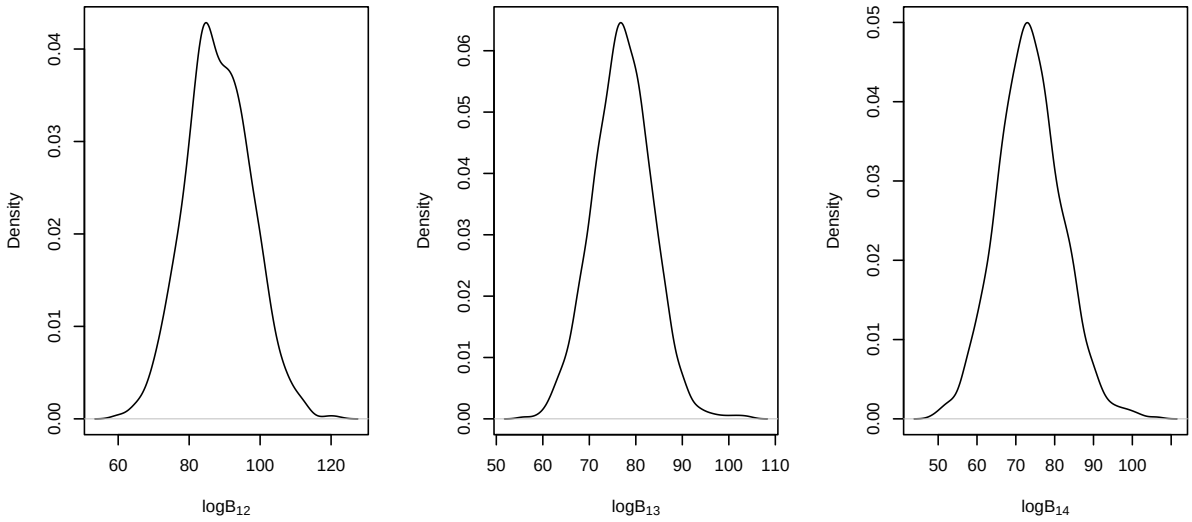

Figure 7: log of the Bayes factor computed with samples from the 57Kg/mol polymer (labeled as Class 1)

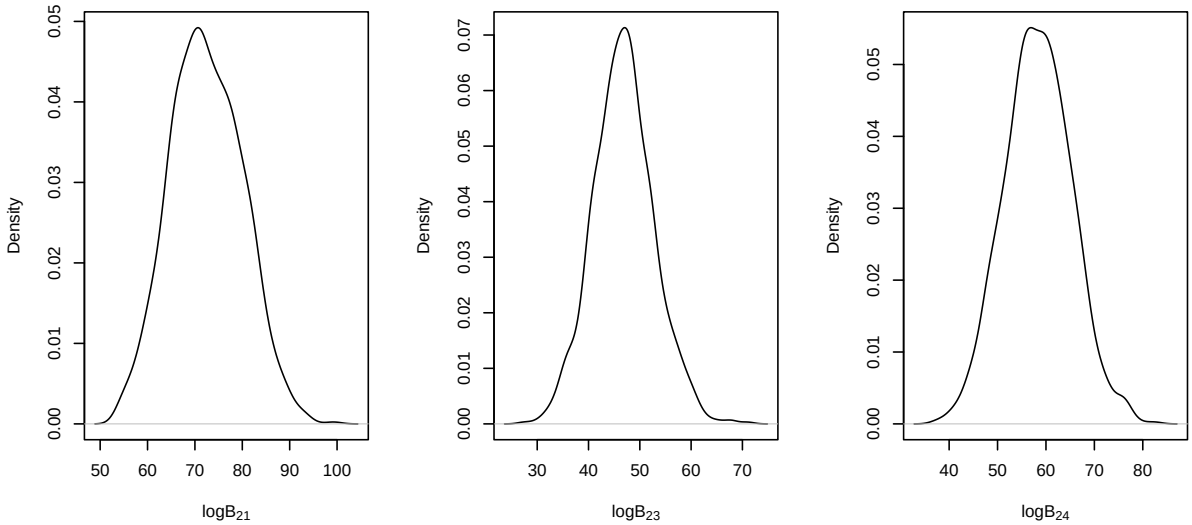

Figure 8: log of the Bayes factor computed with samples from the 100Kg/mol polymer (labeled as Class 2)

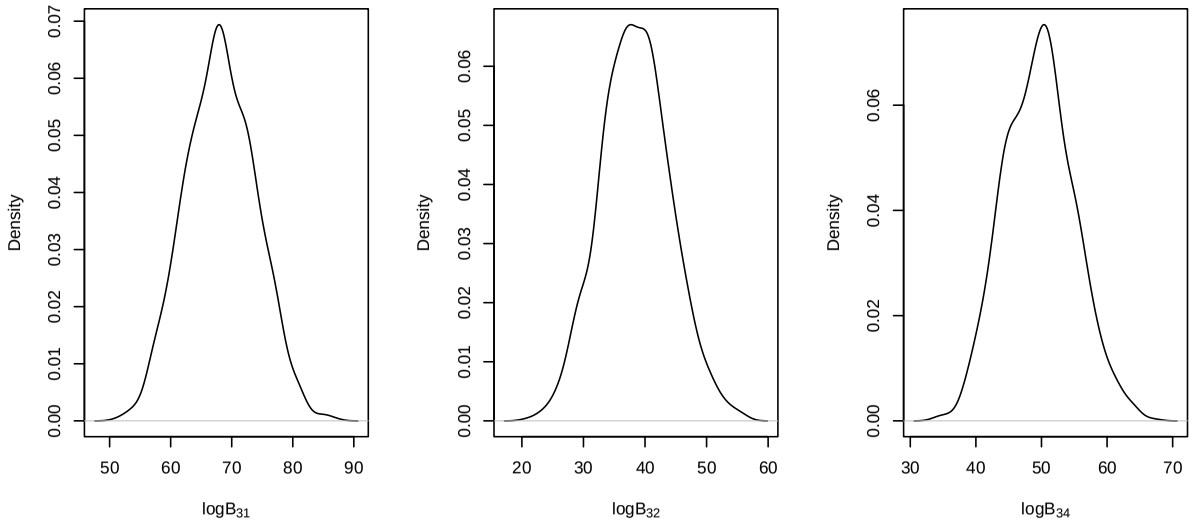

Figure 9: log of the Bayes factor computed with samples from the 500Kg/mol polymer (labeled as Class 3)

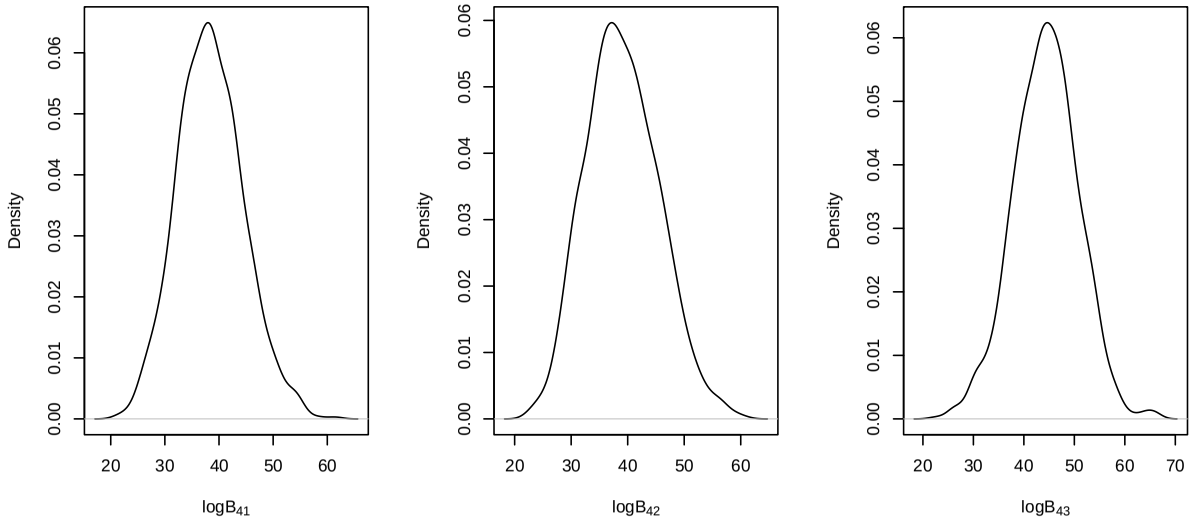

Figure 10: log of the Bayes factor computed with samples from the mixture (57+500) Kg/mol polymers (labeled as Class 4)

the conformations of the polymer during translocation and the accompanying signature in the ionic current trace. We have monitored both the translocation time and the blockade current. The primary goal was to investigate systematically the statistics of the frequency of occurrence of the translocation events in terms of the concentration of the polymer and the polydispersity of the chain. In the initial simulations, we placed ten chains of given length in the cis chamber, at a concentration, which is slightly less than the overlap concentration. Then we monitored how these chains entered the vestibule, which is a prerequisite for the translocation event. Due to entropic pressures between different chains, and the excluded volume interaction between the vestibule and the polymer, the molecules suffered from a very substantial barrier for penetration into the vestibule. Given this, we reduced the concentration to only four in the cis chamber. Now, the chains approached the mouth of the vestibule by the diffusion-limited mechanism. This result was consistent with the fact that the driving force from the electric field is operative only at the mouth of the beta barrel. Yet, only one molecule got into the vestibule. Attempts to capture more than one molecule inside the vestibule failed due to the inter-chain excluded volume effects. The major qualitative conclusion is that the arrival time statistics is basically the diffusion-limited capture of the polymer molecules by the mouth of the vestibule. Since the arrival of the polymer at the vestibule does not guarantee a successful translocation, systematic simulations are being conducted to investigate the role of the ratios of the feeding rate into the vestibule, rate of entropic trapping inside the vestibule, and the rate of actual translocation. This effort, which is a vital step in the fundamental understanding of ionic current signals, is demanding much more resources requiring analytical work and computational power.

Conclusions

This work shows that the Bayesian methodology is a powerful approach for stochastic nanopore data analysis. The construction of the likelihood function without the need for a physical model provides a flexible methodology that allows the implementation of efficient algorithms for feature extraction and inference, while keeping the computational cost low. The ROC curves based on Bayes factors for the 2-class classification problem show robust detection performance. Computational modeling of this system provided insights into the translocation process.

Acknowledgments

This work was supported by DARPA contract # 083050524. Sandia National Laboratories is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under contract DE-AC04-94-AL85000.

References

[1] Qub: a software package for markov analysis of single-molecule kinetics, especially ion channel records. http://www.qub.buffalo.edu/.

[2] H Bayley, O Braha, S Cheley, and LQ. Gu. *Engineered nanopores*, chapter 7, pages 93–112. Wiley-VCH Verlag, 2004.

[3] O Braha, L Gu, L Zhou, X Lu, S Cheley, and H Bayley. Simultaneous stochastic sensing of divalent metal ions. *Nature Biotechnology*, 18(9):1005–1007, 1997.

[4] O. Braha, B. Walker, S. Cheley, J. Kasianowicz, L. Song, J. Gouaux, and H. Bayley. Designed protein pores as components for biosensors. *Biosensors and Bioelectronics*, 1:iii–iv(1), 1997.

[5] E. Gouaux. alpha-hemolysin from staphylococcus aureus: An archetype of beta-barrel, channel-forming toxins. *Journal of Structural Biology*, 121:110–122(13), 1998.

[6] L Gu and H Bayley. Interaction of the noncovalent molecular adapter, beta-cyclodextrin, with the staphylococcal alpha-hemolysin pore. *Biophysical Journal Vol*, 79:1967–1975, 2000.

[7] L Gu, O Braha, S Conlan, S Cheley, and H Bayley. Stochastic sensing of organic analytes by a pore-forming protein containing a molecular adapter. *Nature*, (398):686–690, 1999.

[8] JJ Kasianowicz, E Brandin, D Branton, and DW Deamer. Characterization of individual polynucleotide molecules using a membrane channel. *Proc Natl Acad Sci USA*, 93(24):13770–73, 1996.

[9] M. Muthukumar and C. Y. Kong. Simulation of polymer translocation through protein channels. *PNAS*, 103(14):5273–5278, 2006.

[10] D.S. Sivia. *Data Analysis A Bayesian Tutorial*. Oxford Science, 1996.

[11] A Valeva, I Walev, M Pinkernell, B Walker, H Bayley, M Palmer, and S Bhakdi. Transmembrane beta-barrel of staphylococcal alpha-toxin forms in sensitive but not in resistant cells. *Proc Natl Acad Sci USA*, 94:11607–11611, 1997.

[12] W Vercoutere, S Winters-Hilt, H Olsen, DW Deamer, D Haussler, and M Akeson. Rapid discrimination among individual dna hairpin molecules at single-nucleotide resolution using an ion channel. *Nat Biotechnol*, 19(3):248–252, 2001.

[13] Stephen Winters-Hilt. Hidden markov model variants and their application. *BMC Bioinformatics*, 7(Suppl 2):S14, 2006.

[14] C. Ziegler and W. Gopel. Biosensor development. *Curr Opin Chem Biol*, 25:585–591, 1998.