TOWARDS A STATISTICAL DESCRIPTION OF EXPERIMENTAL DATA FOR DETECTION-ESTIMATION PROBLEMS IN DNA TRANSLOCATIONS THROUGH NANOPORES

S. Michelet[1], JP. Barbot[1], O. Français[1], PY. Joubert[1], P. Larzabal[1],
R. Kawano[2], H. Sasaki[2], T. Osaki[2], S. Takeuchi[2], B. Le Pioufle[1]
[1] *SATIE, ENS Cachan, CNRS, UniverSud, 61 Avenue du Président Wilson, Cachan, France*
[2] *KAST and IIS University of Tokyo, Japan*
stephane.michelet@ens-cachan.fr, bruno.lepioufle@satie.ens-cachan.fr

Keywords: Signal processing, stochastic framework, detection, estimation, characterization of DNA translocations

Abstract: This paper investigates the properties of DNA translocations signals in a stochastic framework. The considered signals are relative to the translocation of single strand DNA through natural nanopores, and are obtained using a planar patch clamp method. The stochastic signal analysis is carried out considering the statistical distribution of DNA translocation parameters, considered as random variables including the amplitude, the duration and the apparition of the DNA translocation events as well as the no-translocation signal features. For each of these variables, a distribution function is proposed and assessed using a Kolmogorov-Smirnov test, and their features are estimated. The DNA translocation signal stochastic analysis enables to characterize the detection and/or estimation performances of existing algorithms, such as a breackdown detection algorithm, in a stochastic framework. Moreover, it opens the way to the design of model based algorithms such as detection tests using a likelihood ratio or joint detection-estimation algorithms using a maximum likelihood approach, for an enhanced characterization of DNA translocations.

1 INTRODUCTION

In view of the DNA sequencing, a biochip dedicated to the DNA translocation through natural nanopores reconstituted on an artificial biomimetic membrane was designed in (Osaki et al., 2009). The biochip consists in a partition between a fluidic chamber and a channel, made with a thin film of parylen obtained by chemical vapor deposition, and micromachined through oxygen plama (see figure 1).

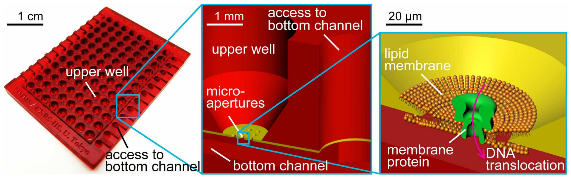

Figure 1: Presentation of the biochip used for DNA translocation detection.

The biomimetic artificial membrane is built-up on this partition, by the successive flow of lipids and buffer into the channel, as described in (Osaki et al., 2009) and the nanopore is created thanks to the insertion of an α-hemolysin natural membrane protein. The application of a voltage on both sides of the membrane induces the movements of ions, and therefore the apparition of a current through the channel. The DNA strand crossing through the membrane induces a current blockade, measured thanks to a patch clamp amplifier. The amplitude and duration of this blockade characterizes the DNA composition and length. The blockade current constitutes the informative signal which is sampled and digitalized by the experimental setup. In order to avoid aliasing during the acquisition process, a so called anti-aliasing low pass filter is used to process the experimental data.

Getting DNA translocation signals is a delicate experiment, since the obtained signals depend on many parameters, such as temperature, humidity, sealing of the artificial membrane or surface conditions of the electrodes.

In this study, in order to avoid repetitive experiments required to adjust the acquisition parameters left to the users and the dedicated data processing techniques, artificial signals are generated. The properties of these artificial signals are determined through the statistical investigations of actual biosignals. In section 2 the statistical properties of the signal are estimated, including the no-translocation current, the amplitude and duration of the DNA translocations events, and the delay between events. In section 3, corresponding artificial signals are generated and used to optimally design an amplitude-duration characterization algorithm based on breackdown detection approach, and used to evaluate the amplitude-duration characterization performances. In section 4, thanks to the proposed statistical framework,the relevance of model based approaches is pointed out, in order to develop i) a detection test using likelihood ratio, and ii) a joint detection-estimation algorithm based on a maximum likelihood method.

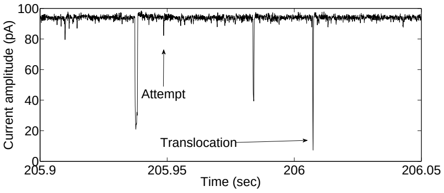

Figure 2: Examples of current blockades relative to DNA translocation and translocation attempts.

2 A STATISTICAL DESCRIPTION OF THE DNA TRANSLOCATION SIGNAL

In this study, the investigation is carried out in a stochastic framework for the current flowing through the nanopore. The considered experimental data are relative to the translocations of a 41mer ssDNA TTTTTTTTTCACTGACCTGGGGGAGTATTGCGGAGGAAGGT, the concentration of which is 45 μM in a 1.0 M KCl, 10 mM PBS, 1 mM EDTA buffer featuring pH=7.4. The DNA translocations are conducted thanks to a 80 mV voltage applied between both sides of the lipid bilayer.

The stochastic characterization of DNA signals consists in the evaluation of the statistical distribution of the amplitude, denoted AMP, the duration (DUR), the delay between translocation (DBT), and the no-translocation signal (NTS), which are defined in figure 3.

2.1 Properties of the Current Through Nanopore in Absence of Translocation

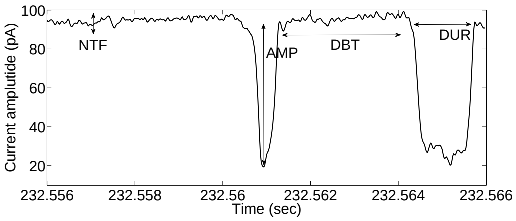

Figure 3: Features of the DNA translocation signal

Firstly we examine the statistical properties of the actual current flowing through an open α-hemolysin channel nanopore without any DNA stand translocation. An example of the current flowing through the nanopore is shown in figure 4.

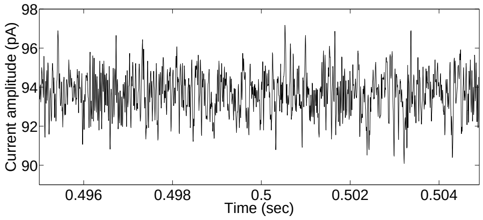

Figure 4: Real current variations through α-hemolysin channel without any translocation event.

An histogram of this no-translocation signal is shown on figure 5, which seems to exhibit a Gaussian distribution. The mean and standard deviation parameters of the distribution, respectively denoted μ and σ, are estimated using:

$$\mu = \frac{1}{n}\sum_{k=1}^{n} x[k] \tag{1}$$

$$\sigma^2 = \frac{1}{n-1}\sum_{k=1}^{n}(x[k]-\mu)^2 \tag{2}$$

where x is the signal and n the number of samples. Considering the available experimental data, the estimation using eq. (1) and eq. (2) leads to μ_{NTS} = 93.7 pA and σ_{NTS} = 1 pA.

In order to attest the assumed Gaussian distribution of the no-translocation signal, a Kolmogorov-Smirnov (KS) test was implemented. The KS test actually quantifies the distance between the cumulative distribution function (CDF) of the considered experimental data, denoted $F_n(x)$, and the CDF of a reference distribution denoted F(x) (Kendall and Stuart, 1979). This KS will be prefered to the Chi-2 test which is sensitive to a lack of data in the experimental histogram. The KS distance is expressed by:

$$D_n = \sqrt{n} \times \sup_x |F(x) - F_n(x)| \tag{3}$$

where n is the number of samples of the experimental data. If this distance D_n is greater than a predefined threshold, then the hypothesis according to which the experimental data distribution is close to the candidate reference distribution is rejected. The threshold is adjusted for a false reject rate of 1%.

Here, the KS test validates the normal distribution of the no-translocation current, as shown on figure 7, which exhibits the CDF $F_n(x)$ and $F(x)$.

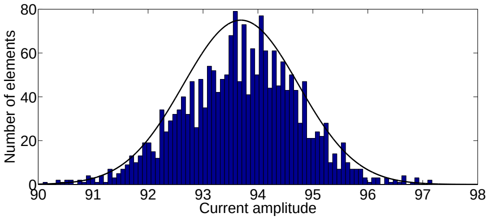

Figure 5: Histogram of no-translocation experimental current samples (n=1982 samples)

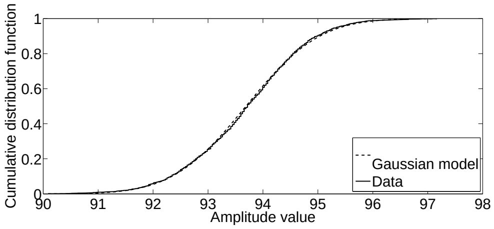

Figure 6: Cumulative distribution functions $F_n(x)$ (Gaussian) $F(x)$ (experimental data)

2.2 Statistical Properties of Translocations Events

In this section, we characterize the translocation events through their duration and amplitude distribution. Indeed, the translocation event provokes a current blockade featured by a duration and an amplitude which give biological information on the ssDNA. The amplitude vs time duration graph permits to determine the length of the DNA, and provides information about its composition, such as the discrimination between polyU, polyC or polyA, (Akeson et al., 1999).

As usually admitted (Kasianowicz et al., 1996), only translocations with a current amplitude decreasing more than 80% of the initial value correspond to complete translocations. Others are translocation attempts which are not considered here.

2.2.1 Amplitude and Duration of the Translocation Events

Thanks to equations (1) and (2) the amplitude distribution mean value and standard deviation of the translocation amplitude AMP can be estimated:

μ_{AMP} = 89.2 pA and σ_{AMP} = 7.33 pA.

In this study, the translocation current amplitude AMP is assumed to be normally distributed, and a KS test implemented has validated this assumption.

In (Meller et al., 2000), the distribution of the translocation duration was approximated using a mixture of a Gaussian law and an exponentially decaying law. Here, for tractability purposes, a Rayleigh law seems to be more adequate to fit the DUR actual distribution law, and will therefore be prefered. The KS test validates this distribution law.

For the duration distribution the Rayleigh parameter r is estimated according to equation (4).

$$r_{DUR} = \frac{1}{2n} \sum_{k \in [1,n]} (x[k])^2 = 924 \mu sec \tag{4}$$

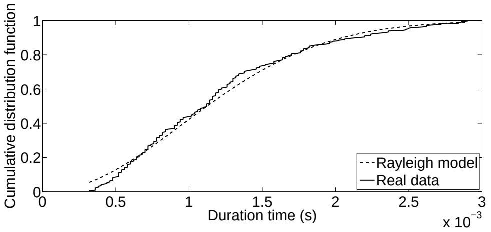

Figure 7: Cumulative distribution functions F_n and F relative to the Rayleigh distribution of DUR

2.2.2 Statistics of Delay Between Translocations Events

The distribution of the delay between translocations (DBT) is considered in this section and assumed to be a decreasing exponential, expressed by :

$$f(DBT) = \alpha \exp(-\alpha DBT) \tag{5}$$

where $\alpha = (17.6ms)^{-1}$.

2.2.3 Statistical Description of the DNA Translocation Signal

Finally, the distribution features of the random variables AMP, DUR, DBT and NTS, estimated from experimental DNA translocation signals are gathered in table 1.

Table 1: Distributions features.

	Dist	param. 1	param. 2
NTS	Gaus.	$\mu = 93.7$ pA	$\sigma = 1$ pA
AMP	Gaus.	$\mu = 89.2$ pA	$\sigma = 7.33$ pA
DUR	Ray.	r = 924 μs	
DBT	Exp.	α = (17.6 ms)$^{-1}$	

3 Performances of a no-parametric amplitude-duration estimation algorithm

In this section, the DNA translocation signal characterization results are used to evaluate, in a stochastic framework, the performances of an elementary translocation characterization algorithm. The considered algorithm is based on a breakdown detection technique, presented in (Osaki et al., 2010) which allows the amplitude and duration of translocation events to be characterized. To evaluate the performances of this characterization algorithm for various signal features, we build up artificial biomimetic signal considering the AMP, DUR, DBT and NTS distributions estimated in the previous section. Moreover, in order to take account for the possible experimental noise variations, we elaborate artificial signals featuring various signal to noise ratios (SNR) defined as:

$$SNR = 20log|\frac{\mu_{AMP}}{\sigma_{NTS}}| \qquad (6)$$

An example of a 319 translocation signal sequence featured by a 30 dB SNR is represented in figure 8 and figure 9 exhibits the detail of a single artificial translocation event. The implementation of the break down detection algorithm applied to this translocation sequence allows the AMP and DUR values of the 319 translocations to be estimated. The corresponding amplitude vs duration representation diagram is depicted in figure 10.

In order to quantify the characterization performances of the algorithm, we compute the true positive rate and the false positive rate of the characterization algorithm, considering SNRs ranging from 6 to 46 dB. The true positive rate (TPR) is computed as the rate of the estimated AMP-DUR values of each considered translocation event which are close to the actual values whithin a predifined distance ν. On the other hand, the false positive rate (FPR) is defined as the rate of the estimated AMP-DUR values which are at a distance of the actual AMP-DUR values higher than the predefined value ν. Then, the receiver operational characteristic (ROC) which plots the TPR as a function of FPR for various values of ν can be considered to quantify the characterization performances (Bradley, 1997). ROC curves obtained for the considered translocations data are presented in figure 11. One can note for example that a 90 % TPR is reached at the cost of a 0.01 % FPR considering a translocation sequence with SNR = 30 dB, and that the same 90 % TPR is reached at the cost of a 1 % FPR when the SNR falls down to 18 dB. An other means of quantifying the performance of the amplitude-duration estimation algorithm is to evaluate the mean square error (MSE) defined in equation (7), of the characterization as a function of the SNR of the translocation signal.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\sqrt{\left(\frac{\widehat{AMP_i} - AMP_i}{\mu_{AMP}}\right)^2 + \left(\frac{\widehat{DUR_i} - DUR_i}{\mu_{DUR}}\right)^2} \qquad (7)$$

where $\widehat{AMP}$ and $\widehat{DUR}$ are the estimated values of AMP and DUR respectively, n is the number of translocations equal to 319, and where the contribution of the amplitude and duration errors are normalized by their mean values in order to give them the same weight in the computation of the MSE. The MSE computed according to equation (7) and expressed in percent is represented in figure 12. One can note that the MSE falls from 30 % to 0.02 % when the SNR rises from 6 dB up to 46 dB, respectively.

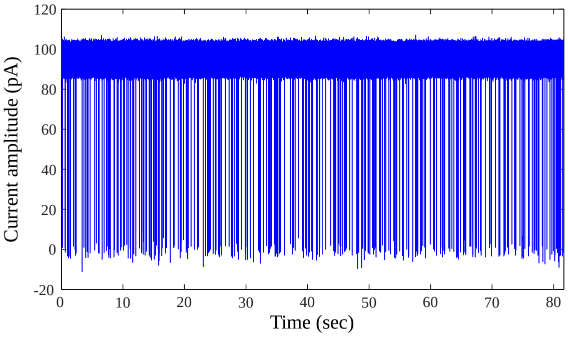

Figure 8: Generated artificial signal.

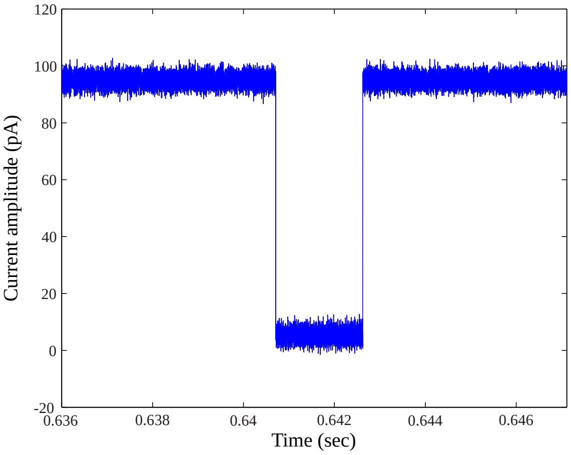

Figure 9: Single translocation event in the artificial signal.

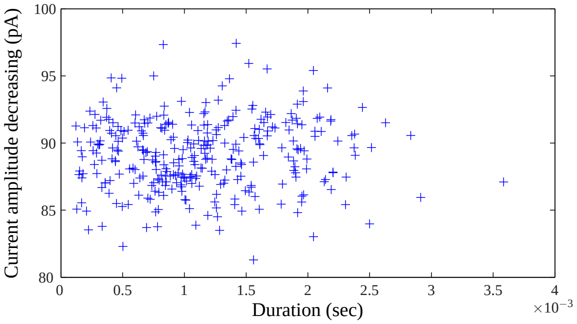

Figure 10: Diagram duration vs amplitude for the artificials translocations.

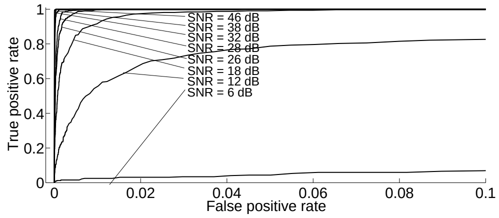

Figure 11: ROC of the duration-amplitude characterization.

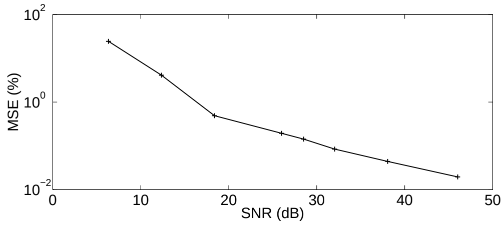

Figure 12: MSE of the amplitude-duration estimation.

4 DISCUSSION

We have proposed a statistical characterization of nanopore DNA translocation current allowing well known methods of amplitude/duration characterizations (Basseville and Nikiforov, 1993) to be implemented and evaluated using intensive computer simulations.

More challenging now is the use of this statistical characterization to optimally built new model based approach to improve the characterization performances. The proposed modelling of translocation signals opens this way. As a complete statistical characterisation of a steplike signal is now available, several ways of investigation are opened. Let us briefly point out two of them for further works:

1) a model based segmentation procedure which detects multiple change points in a steplike signal can be built on a generalized likelihood ratio test or on information theoretic criterion such as Akaike information criterion like tests. Moreover, since this segmentation technique considers the DNA translocation signal as a whole sequence, it avoids the well known drawbacks relative to sliding window data processing approaches.

2) a regularized maximum likelihood method can be built, looking for the unknown parameters $\boldsymbol{\theta}$ as :

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{argmin}\{\|s*(t) - s(\boldsymbol{\theta})\|_2^2 + \lambda\|\nabla s(\boldsymbol{\theta})\|_1\} \quad (8)$$

with

$\boldsymbol{\theta} = [t_1, t_2 ... t_N, a_1, a_2 ... a_N]^T$ where t_i are the step location parameters and a_i are the step amplitude parameters.

$\nabla s(\boldsymbol{\theta})$ is the gradient of the solution. As we are looking for a steplike signal, for regularization purposes a l_1 norm will be used for the gradient.

$s*(t)$ is the actual recorded signal and $s(\boldsymbol{\theta})$ is a candidate signal. Recent developments in convexe constraint optimisation open the way to an efficient optimisation of the criterium expressed in equation (8). λ is a paramater used to adjust the contribution of each terms of the regularization criterion.

This provides a statistical framework for DNA translocation characterisation.

ACKNOWLEDGEMENTS

This work was supported by grants of a joint CNRS-JST research program.

REFERENCES

Akeson, M., Branton, D., Kasianowicz, J. J., Brandin, E., and Deamer, D. W. (1999). Microsecond time-scale discrimination among polycytidylic acid, polyadenylic acid, and polyuridylic acid as homopolymers or as segments within single rna molecules. Biophysical Journal Volume 77.

Basseville, M. and Nikiforov, I. V. (1993). Detection of abrupt changes : theory and applications. Prentice-Hall, Englewood Cliff, NJ.

Bradley, A. P. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. Pattern Recognition Lett 30(7):1145159.

Kasianowicz, J. J., Brandin, E., Branton, D., and Deamer, D. W. (1996). Characterization of individual polynucleotide molecules using a membrane channel. Proc. Natl. Acad. Sci. USA 93.

Kendall, M. G. and Stuart, A. (1979). *The Advanced Theory of Statistics*, volume 2. Charles Griffin, 4th edition.

Meller, A., Nivon, L., Brandin, E., Golovchenko, J., and Branton, D. (February 2000). Rapid nanopore discrimination between single polynucleotide molecules. PNAS vol. 97.

Osaki, T., Barbot, J., Kawano, R., Sasaki, H., Franais, O., Pioufle, B. L., and Takeuchi, S. (September 2010). A rupture detection algorithm for the dna translocation detection though biological nanopore. accepted in Proc. Eurosensors XXIV.

Osaki, T., Suzuki, H., Pioufle, B. L., and Takeuchi, S. (2009). Multichannel simultaneous measurements of single-molecule translocation in α-hemolysin nanopore array. Analytical Chemistry 81.