Evaluation of Dynamic Time Warp Barycenter Averaging (*DBA*) for its Potential in Generating a Consensus Nanopore Signal for Genetic and Epigenetic Sequences *

Rachel S. L. Chan, Paul Gordon and Michael R Smith, *Senior Member, IEEE*

***Abstract*— Epigenetics is a chemical modification to DNA without changes in the base sequence. While it is known that epigenetic modifications have far reaching implications on how genes are expressed, it is difficult to identify what the modification is or where it can be found. A next-generation method of sequencing called nanopore sequencing may be the solution. Nanopore sequencing runs a voltage bias across the DNA sequence and outputs a unique electric response to each genetic unit. Epigenetic modifications may then be identified by their distinct electric response. In this paper we provide preliminary results of applying dynamic time warp Barycenter Averaging (*DBA*) to multiple noisy nanopore streams to generate a consensus signal that can be used to identify genetic sequences and their modifications. *DBA* convergence rates, time complexity, together with qualitative and quantitative metrics to compare the consensus signal with gold standards are evaluated.**

I. Introduction

Bioinformatics combines elements of biology and computer science to collect and analyze complex biological data; commonly macromolecular structures, genetic sequencing, and genomic experiments, or gene expression data [1]. Observing consistent deviations across comparable DNA sequences could indicate and identify potential epigenetic modification of the DNA. This is significant because these modifications may also be expressed through behavioural characteristics of an organism, such as susceptibility to disease or metabolic rate.

Nanopore sequencing is a next generation gene sequencing technique that involves applying a voltage bias across the DNA sequence and outputs a unique electric response to each genetic unit. The pico-amperage current characteristics from the stream passing through the pore are distinct for each base and any of its modifications will have a distinct amperage [2]. Currently, dynamic time warping (*DTW*) is used to compare the results from nanopore sequencing to some reference signal. *DTW* finds the optimal alignment between two time series. As *DTW* is not entirely dependent on synchronous timing, the time series may be warped non-linearly, their time axis stretched or shrunk, during the alignment process [3].

This sequencing method may be used in conjunction with powerful data mining algorithms in an attempt to identify and characterize the epigenetic modifications. While *DTW* is known to be robust and is the current industry standard, there are two main weaknesses when gene sequencing: *DTW* can only consider the minimum path between sequences and it can only compare two signals at the same time. Therefore, *DTW* alone is unable to generate a consensus signal from the multiple nanostreams that may available from a given sample.

DTW Barycenter Averaging (*DBA*) is a relatively new data mining algorithm which can combine an unlimited number of datasets to generate a consensus signal without averaging out individual dataset characteristics. *DBA* has been used to average financial time series [4] to develop alpha curves visualizing a balance between risk and return elements by tracking error against added return over a benchmark. Central to *DBA* is its ability to generate an average of multiple data streams while preserving key features. We suggest that the usefulness of this property is as important to geneticists requiring a consensus signal as it is to traders wanting to create valid alpha curves. We believe that this is the first investigation of using *DBA* to generate a consensus signal that preserves the key features of multiple nanopore streams while avoiding a smoothed result that is characteristic of most averaging methods.

The paper is organized as follows: Section II covers a short discussion comparing the characteristics of *DTW* and *DBA*. Section III covers the simulation approach used in this preliminary investigation of the convergence rate of *DBA* to form a consensus signal from a series of nanopore streams that are individually distorted versions of a known "gold-standard" DNA sequence. Section IV covers the results of applying the *DBA* analysis to simulated nanopore streams. A conclusion follows to summarize the key features of this paper.

II. Relationship between *DTW* and *DBA*

Nanopore sequencing is a fast, portable, and relatively inexpensive method of genetic sequencing [2]. A small voltage bias is generated across a nano-metre sized pore immersed in an aqueous electrolyte [5]. A DNA strand is electrophoretically driven through the nanopore causing the nucleobases of that strand to modulate an ionic current that

*Research supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada and the University of Calgary

R. S. Chan is with the Department of Electrical and Computer Engineering, University of Calgary, Alberta, Canada T2N 1N4 (e-mail: rschan@ucalgary.ca).

P. Gordon is with Bioinformatics, Alberta Childrens' Hospital Research Institute, University of Calgary, Alberta, Canada T2N 1N4 (e-mail: gordonp@ucalgary.ca).

M. R. Smith is with the Department of Electrical and Computer Engineering, University of Calgary, Alberta, Canada T2N 1N4 (Corresponding author: phone: +1-403-220-6142; e-mail smithmr@ucalgary.ca).

can be translated into a pico-amperage signal. This signal can be segmented, to generate a current squiggle [6], as shown schematically in Fig. 1. The squiggle characteristics of the ionic current depends on the genotype as well as on the epigenetic modifications that affect the phenotype. Though epigenetic modifications do not affect the base sequence, they do generate different current squiggles. This difference can be identified by comparing the nanopore squiggle with a "gold standard" reference converted into a hypothetical current squiggle [6].

Dynamic time warping (*DTW*) can be used to compare the results from nanopore sequencing to some reference signal. *DTW* finds the optimal alignment between two time series, without being dependent on synchronous timing [3]. As a result, the time series may be warped non-linearly, their time axes stretched or shrunk. This is an improvement over traditional Euclidean method where distance measurements are used to compare the similarity between two signals. Moreover, points of the signals are not compared on a one-by-one basis, meaning a single point on one signal can map to several points on another. This means that *DTW* recognizes two identical signals even if they were not already lined up on the time axis.

While the *DTW* algorithm can generate a signal that represents the key joint features of the two compared signals, it is not suitable for combining those key features across many data streams. This inability to generate a consensus makes *DTW* vulnerable to false negatives and false positives since epigenetic modifications generate differences in current squiggles comparable to the electrical noise levels that are experimentally present on the nanostream currents.

DTW Barycenter Averaging (*DBA*) is an iterative algorithm that applies dynamic time warping, but aligns all the series with an evolving average. Applying this concept, signal averaging will not simply depend on the average of the input signal, but by the magnitudes of each point, the 'weight', in the signal [4].

This approach has proven to be a robust method of data mining over a changing average in the context of financial data series [4]. In the next sections we explore whether this approach is appropriate for combining multiple nanopore streams into a consensus signal. We need to determine the extent that this consensus signal has improved signal-to-noise and time dilation characteristics compared to an individual nanopore stream while retaining the features of the known "gold standard" from which each nanostream was experimentally derived.

III. Nanopore Stream *DBA*-Consensus Evaluation

In this section, we detail the simulation of the hypothetical pico-current squiggle from a "known gold standard" reference. A method is described for simulating the squiggles derived from nanocurrents generated from the multiple pores of a device mimicking the Oxford MinIon gene sequencer [5]. Metrics to allow comparison of the *DBA* consensus signal with the original gold standard are described.

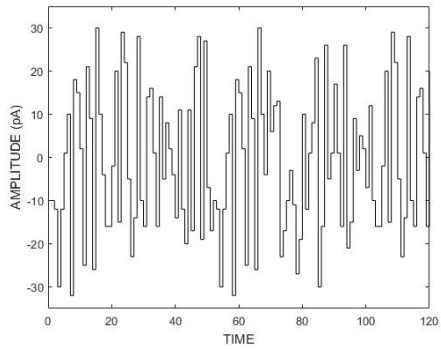

Fig. 1. The known gold standard is converted into a simulated nanopore stream that has been converted into squiggle space.

A. *DTW and DBA Implementation*

All simulations were performed in *MATLAB R2017A* on an *Intel i5-4570* core clocked at 3.2 GHz. The *DBA MATLAB* algorithm by F. Petitjean and I. Paparrizos [7, 8] distributed under the GNU general public license, was used. An approximately 10-fold speed gain was gained by using the *DTW* code provided within the *MATLAB* tool-boxes rather than the *DTW* provided with the *DBA* package.

B. *Gold Standard Simulation*

The pico-amperage generated by a specific base passing through a nanopore simulation is not constant but is impacted by the other nucleotides in the local sequence. A total of $L = 4 \wedge k$ unique pico-current levels are present in a k-mer system. A simulated squiggle space signal of length S from a known gold standard was generated by placing S uniformly random values across the L k-mer levels.

Algorithms involving *DTW* calculation have a time complexity $Order(S^2)$. To keep simulation computation time within reasonable bounds, this study simulated a *3*-mer system, with *64* levels, with stream lengths in the range $100 \le S \le 1600$. Fig. 1 shows a section of a simulated gold-standard nanostream containing $S = 150$ events.

C. *Individual nanostream simulation*

To simulate the variation introduced as each stream passes through P pores, $P \le 512$, of the gene-sequencer, the gold-standard stream was modified according to the following rules.

- Any particular k-mer had a $D\%$ chance of being dropped, i.e. skipped over, and not recognized as participating in the nano-stream measured at a particular pore
- Mis-segmentation of the noisy pico-current stream into events was simulated by providing any particular k-mer an $A\%$ chance of generating an additional event.
- Any particular nano-stream may be delayed relatively to another stream by $\pm$ T time units.

The lower khaki lines in Fig. 2 show the generation of 4 simulated distorted nano-streams using parameters: event drop probability $D = 7.5\%$, event addition probability $A = 7.5\%$, and time delay range $T = \pm$ 8 T.U. It can be seen that

such parameter values introduce considerable changes in an individual stream's appearance from the black gold-standard..

Noise was added to the pico-signals to better match the expected experimental uncertainty in correctly identifying the *k*-mer pico-current level.

- The uncertainty in the measurement of each *k*-mer level was set equivalent to the introduction of a Gaussian white-noise with a standard deviation of *N* levels corresponding to 0.01*N / K^4%* of the pico-current range.

IV. Results and Discussion

In this section we explore *DBA* convergence rates and time complexity, and examine several qualitative and quantitative metrics for comparing the consensus signal derived from several noisy nanostream with the original gold standards.

A. *DBA convergence rate and time complexity*

The *DBA* algorithm uses iteration to generate a consensus signal by repeatedly applying *DTW* to analysis an evolving average with *P* nanostream signals. As *DTW* already involves a time complexity of $Order(S^2)$, it is important to identify the number of iterations necessary to achieve convergence to a consensus. We propose the following measure for the *DBA* convergence of the *P* individual streams to a consensus signal.

DBA convergence can be recognized when the DTW distance between the I^{th} and $(I+1)^{th}$ estimate of the consensus signal becomes close to 0.

Experimentally we found that this measure indicated consensus convergence after 14 to 16 iterations for $P < 256$; essentially the default iteration setting for the Petitjean and Paparrizos *DBA* algorithm [7, 8].

The execution time complexity for the *DBA* algorithm to form a consensus signal from *P* streams was experimentally found to be $Order(P^{1.5})$ for the current *DBA* implementation. Although this is additional time compared to a single *DTW* comparison, the number of streams available for combination is small, typically $P < 256$ for the current nanopore technology, unlike *S*, the total number of events in a stream.

B. *Qualitative DBA consensus metrics*

Qualitative measures on how well the consensus signal from multi-nanopore stream matches the original gold-standard can be obtained through two approaches that visualize the results of performing a *DTW* comparison between the two signals. In Fig. 3, the blue lines join gold standard and consensus features that have been matched via *DTW*.

Fig. 3A provides a comparison between the gold standard and the *DBA* consensus signal in a normal linear space. An initial observation would indicate that certain sections of the consensus signal are a better match to the original gold standard than others. This indicates potential new computation paths to explore to improve *DBA's* efficacy in this new application area.

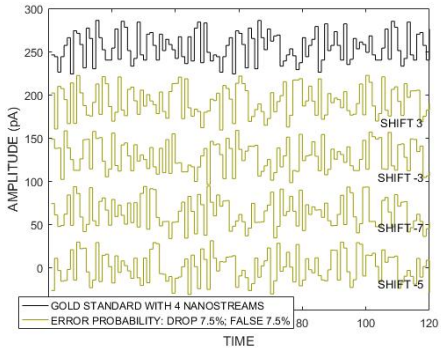

Fig. 2. The black curve shows the simulated pico-current from this gold standard and consists of 150 squiggles spread across 64 3-mer levels. The khaki lines show the generation of 4 simulated distorted nanostreams using parameters: event drop probability *D* = 7.5%, false event addition probability *A* = 7.5%, time delay range *T* = ± 8 T.U.

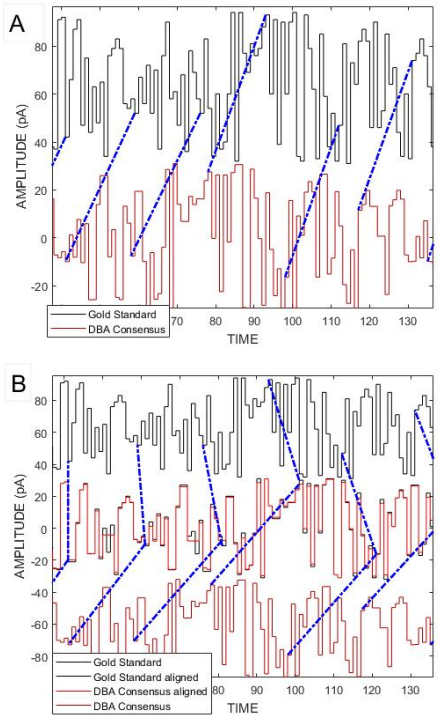

Fig. 3. Qualitative comparisons between the gold standard and the *DBA* consensus signals in A) normal linear space or B) in the warped *DTW* space. The blue lines join components matched using a *DTW* analysis

Fig. 3B provides an alternative qualitative comparison that shows where similarities and differences of gold standard and the *DBA* consensus signals exist in the time warped *DTW* space.

A. *Qualitative DBA consensus metrics*

The following quantitative metrics were used to explore gold standard and consensus signal differences at nanostream defect insertion probabilities of 3.75%, 7.5% and 15%, with relative, random, time shifts in range -8 to +8 time units

- A self-consistency measure based on the *DTW* distance between the gold standard and the consensus signal.
- An external consistency measure of the respective *DTW* means and standard deviations from comparing all nanostreams with the gold standard or the consensus signal.

Table 1: Quantitative comparison using *DTW* distances between the gold standard and the *DBA* consensus, and comparing the gold standard and the *DBA* consensus with individual nanostreams contaminated with Gaussian noise; standard deviation of 1 level in 64 (1.5% of maximum signal amplitude) for various defect introduction rates. There were 400 events present in the gold standard. Nanostreams had relative time shifts in the range -8 to +8.

	DEFECT RATE: *D = A = 3.75%*			**DEFECT RATE *D = A = 7.5%***			**DEFECT RATE *D = A = 15%***		
NUMBER STREAMS	**GOLD v DBA**	**GOLD v STREAMS**	**DBA v STREAMS**	**GOLD v DBA**	**GOLD v STREAMS**	**DBA v STREAMS**	**GOLD v DBA**	**GOLD v STREAMS**	**DBA v STREAMS**
4	1.70+-0.07	1.73+-0.66	1.95+0.46	1.91+-0.06	1.97+-0.39	2.14+-0.64	4.76+-0.04	3.70+-0.17	4.35+-1.03
8	1.64+-0.07	1.69+-0.49	2.21+0.39	1.83+-0.06	1.86+-0.29	2.37+-0.34	4.17+-0.04	3.63+-0.53	4.58+-0.77
16	1.57+-0.07	1.73+-0.43	1.92+0.27	1.65+-0.06	1.85+-0.35	2.42+-0.33	4.12+-0.04	3.53+-0.51	4.94+-0.67
32	1.54+-0.07	1.77+-0.42	1.94+0.22	1.66+-0.06	1.88+-0.36	2.54+-0.28	3.85+-0.04	3.43+-0.48	4.90+-0.54
64	1.19+-0.07	1.71+-0.41	2.06+0.28	1.73+-0.06	1.91+-0.38	2.46+-0.27	3.79+-0.04	3.40+-0.53	4.84+-0.45
128	1.53+-0.07	1.70+-0.43	1.94+0.21	1.70+-0.06	1.97+-0.39	2.46+-0.25	3.86+-0.04	3.48+-0.56	4.92+-0.40

Table 2 – Quantitative comparison using *DTW* distances between the gold standard and the *DBA* consensus, and comparing the gold standard and the *DBA* consensus with individual nanostreams contaminated with Gaussian noise; standard deviation of 4 levels in 64 (6% of maximum signal amplitude) for various defect introduction rates. There were 400 events present in the gold standard. Nanostreams had relative time shifts in the range -8 to +8.

	DEFECT RATE: *D = A = 3.75%*			**DEFECT RATE *D = A = 7.5%***			**DEFECT RATE *D = A = 15%***		
NUMBER STREAMS	**GOLD v DBA**	**GOLD v STREAMS**	**DBA v STREAMS**	**GOLD v DBA**	**GOLD v STREAMS**	**DBA v STREAMS**	**GOLD v DBA**	**GOLD v STREAMS**	**DBA v STREAMS**
4	2.03+-0.07	3.01+-0.32	2.75+-0.45	2.85+-0.18	3.58+-0.37	3.23+-0.61	3.82+-0.25	4.52+-0.83	4.36+-1.10
8	2.15+-0.07	3.14+-0.42	3.10+-0.32	2.47+-0.18	3.61+-0.45	3.72+-0.49	3.34+-0.25	4.50+-0.70	5.75+-0.89
16	1.50+-0.07	3.09+-0.48	3.22+-0.37	2.60+-0.18	3.60+-0.38	3.98+-0.40	3.31+-0.25	4.45+-0.51	5.02+-0.63
32	1.33+-0.07	3.16+-0.47	3.29+-0.36	2.46+-0.18	3.58+-0.37	3.91+-0.31	3.27+-0.25	4.57+-0.59	5.15+-0.51
64	1.33+-0.07	3.24+-0.50	3.41+-0.39	1.94+-0.18	3.56+-0.35	3.76+-0.31	3.14+-0.25	4.57+-0.60	5.19+-0.46
128	1.44+-0.07	3.24+-0.44	3.23+-0.29	1.97+-0.18	3.59+-0.36	3.85+-0.30	3.26+-0.25	4.60+-0.58	5.18+-0.35

In Table 1, these metrics are evaluated when the individual nanostreams are assumed to be experimentally distorted with Gaussian white noise with a standard deviation of 1 level in 64 (1.5% of maximum pico-current amplitude). In Table 2, the noise has a standard deviation of 4 levels in 64 (4% of maximum pico-current amplitude).

The external *DTW* consistency measure is an indicator of how well the gold standard and consensus signal individually describe the underlying characteristics of the distorted nanostreams. For low defect rates, *D = A = 3.75%,* the *DTW* means and standard deviation measures indicate that, within experimental error, both signals are equivalent indicators of the underlying properties of the distorted streams. At higher distortions levels this measure indicates that the gold standard is consistently a better descriptor, a trend weakly seen at the lower distortion levels.

Evidence that increased *DBA* averaging leads to an improved consensus signal that better approximates the original gold standard signal is present in both Tables. The *DTW* distance between the gold standard and consensus signal decreases with the number of stream processed. However, this measure provides weaker evidence of convergence at higher defect introduction rates.

For higher noise levels we consider that decreases in the standard deviation of the *DTW* distance between the *DBA* consensus signal and all distorted nanostreams provides a better indicator of convergence.

We believe the decrease in *DTW* distance values for high defect rates, *D = A = 15%,* as the noise level increases is related to the *3*-mer simulation parameters. At low noise levels, when the noise standard deviation is of the order of the distance between two measurement levels, the signal will be forced to one of the quantized *k*-mer levels. This introduces a systematic bias rather than converging to a true consensus.

V. Conclusion

There is considerable potential in marrying nanopore sequencing and *DTW* Barycenter Averaging (*DBA*). In this preliminary study, we have shown experimentally that the consensus signal becomes a better match to the underlying common characteristics of the distorted nanostreams as the number of streams *DBA* averaged increases. Further work is needed to better characterize the simulated defect insertion rates with experimental stream distortions, and to determine the impact that the use of *DBA* consensus signals has on reducing error rates further down the analysis chain.

References

[1] DNA Sequencing Fact Sheet", *National Human Genome Research Institute (NHGRI)*, 2017. [Online]. Available: https://www.genome.gov/10001177/dna-sequencing-fact-sheet/. [Accessed: 27- Aug- 2017]

[2] D. Branton, D. Deamer, A. Marziali, et al., "The potential and challenges of nanopore sequencing", *Nature Biotechnology*, vol. 26, no. 10, pp. 1146-1153, 2008.

[3] A. Laszlo, I. Derrington, B. Ross, et al., "Decoding long nanopore sequencing reads of natural DNA", *Nature Biotechnology*, vol. 32, no. 8, pp. 829-833, 2014.

[4] Reverse Engineering "Alpha Curves": DTW Barycenter Averaging", *QUSMA*, 2014. [Online]. Available: http://qusma.com/2014/03/26/dtw-barycenter-averaging/. [Accessed: 22- Sep- 2017]

[5] M. Jain, H. Olsen, B. Paten and M. Akeson, "The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community", *Genome Biology*, vol. 17, no. 1, 2016.

[6] "Dynamic Time Warping averaging of time series allows faster and more accurate classification", *the morning paper*, 2016. [Online]. Available: https://blog.acolyer.org/2016/05/13/dynamic-time-warping-averaging-of-time-series-allows-faster-and-more-accurate-classification. [Accessed: 20- Sep- 2017].

[7] F. Petitjean, A. Ketterlin and P. Gançarski, "A global averaging method for dynamic time warping, with applications to clustering", Pattern Recognition, vol. 44, no. 3, pp. 678-693, 2011

[8] F. Petitjean and I. Paparrizos, "DBA: Averaging for Dynamic Time Warping", https://github.com/fpetitjean/DBA; Accessed November, 2017.