Athens Institute for Education and Research
ATINER

ATINER's Conference Paper Series
ENG2016-2079

# Optimizing DNA-Alignment Algorithms for an Embedded SoC/FPGA Platform

**Christian Rohrandt**
**PhD Student**
**Kiel University of Applied Sciences**
**Germany**

**Pay Giesselmann**
**Master Student**
**Kiel University of Applied Sciences**
**Germany**

**Ulrich Jetzek**
**Professor**
**Kiel University of Applied Sciences**
**Germany**

**Franz-Josef Mueller**
**Head**
**Kiel Centre for Integrative Psychiatry**
**Germany**

ATINER CONFERENCE PAPER SERIES No: ENG2016-2079

# An Introduction to ATINER's Conference Paper Series

ATINER started to publish this conference papers series in 2012. It includes only the papers submitted for publication after they were presented at one of the conferences organized by our Institute every year. This paper has been peer reviewed by at least two academic members of ATINER.

Dr. Gregory T. Papanikos
President
Athens Institute for Education and Research

This paper should be cited as follows:

**Rohrandt, C., Giesselmann, P., Jetzek, U. and Mueller, F. J.** (2016). **"Optimizing DNA-Alignment Algorithms for an Embedded SoC/FPGA Platform",** Athens: ATINER'S Conference Paper Series, No: **ENG2016-2079.**

Athens Institute for Education and Research
8 Valaoritou Street, Kolonaki, 10671 Athens, Greece
Tel: + 30 210 3634210 Fax: + 30 210 3634209 Email: info@atiner.gr URL: www.atiner.gr
URL Conference Papers Series: www.atiner.gr/papers.htm

ISSN: 2241-2891
08/12/2016

# Optimizing DNA-Alignment Algorithms for an Embedded SoC/FPGA Platform

**Christian Rohrandt**
**Pay Giesselmann**
**Ulrich Jetzek**
**Franz-Josef Mueller**

## Abstract

Massively parallel sequencing technologies – often referred to as next-generation sequencing (NGS) technologies – have recently revolutionized the field of biomedicine. Pressing issues in human development, disease, and recovery, which seemed to be irresolvable a few years ago, can now be answered with widely available and affordable sequencing datasets. Currently, NGS technologies are operated in specialized facilities requiring significant investments in staff, equipment, reagents and compute resources. With the, as of today, still exponentially increasingly generation of NGS data sets, the computational challenges associated with the processing of NGS data is a bottleneck limiting scientific progress. To be ability to process NGS datasets today at least a powerful desktop computer is needed. For processing more than a few samples, the researcher needs to have access to a computer cluster and tera- to petabyte storage systems. At the same time, future use cases for NGS technologies may even be the use of sequencing in the field of humanitarian crises following armed conflicts, natural or technological disasters, for example, to detect, identify and treat Ebola strains in a region where infrastructure resource such as electricity has completely broken down. Up to now, in such life threatening situations DNA samples cannot be analyzed locally, but need to be taken to a place offering the aforementioned computer infrastructure. As a consequence potentially life-saving measures will be severely delayed. As DNA analysis becomes a more and more commonly applied diagnostic method for diseases the demand for a field-deployable, portable and ideally handheld analysis device is increasing. For enabling such a futuristic technology, a far more efficient ways to align sequencing reads to a reference genome without the availability of large-scale computer resources have to be developed. Consequently, the development of a hardware platform and the optimization of the software running on it is the next step. This hardware platform will incorporate a mobile processor system, presumably based on the ARM architecture, paired with an FPGA. This way the energy efficiency of an embedded mobile platform can be combined with the flexibility and computational performance of programmable logic.

**Keywords:** ARM, DNA sequence alignment, FPGA, Next Generation Sequencing, System-on-Chip.

**Acknowledgments:** We are in dept to Benedicte Willert and Prof. Dr. Ole Ammerpol at the Institute for Human Genetics at the University Hospital Schleswig Holstein Campus Kiel, Kiel, Germany for conducting the initial nanopore sequencing experiments in their lab. This work was supported by the German Federal Ministry of Education and Research (BMBF) through the "PluriTest2" project, grant nr. 13GW0128 (F.-J. M.), and the German Research Foundation (DFG) through grant MU 3231/3-1 (F.-J. M.).

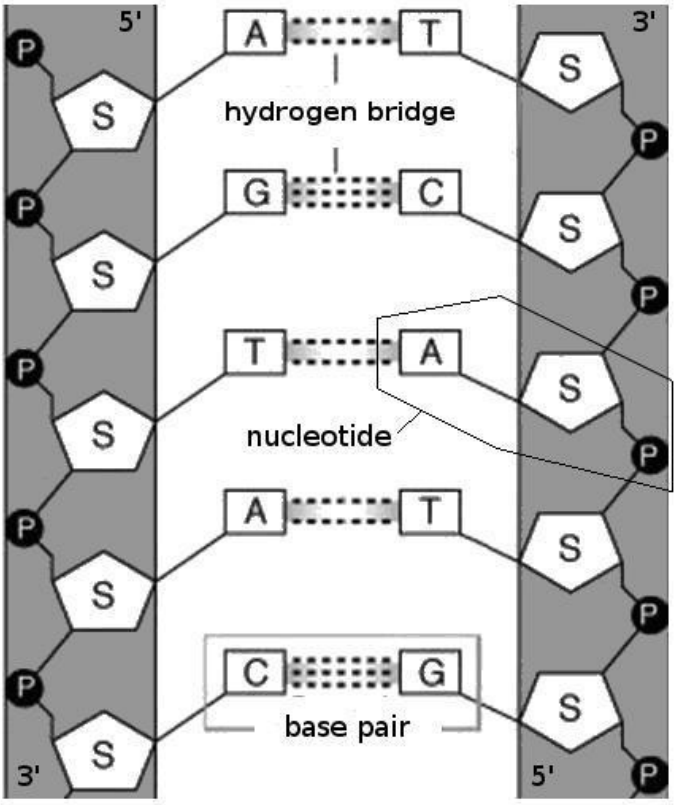
## Introduction

Deoxyribonucleic acid (DNA) is considered to be the source code of life on earth because most of the genetic information in animals, plants and single cell organisms such as bacteria as well as most viruses is encoded in DNA sequences. DNA is a large polymer molecule consisting of a backbone, formed by a type of sugar (deoxyribose), phosphate groups and four different nucleobases. The monomer units of DNA are termed nucleotides, and the polymer is referred to as a "polynucleotide". The most important nucleotides are guanine (G), adenine (A), thymine (T) and cytosine (C).

Each of these nucleotides connects with a deoxyribose molecule and a phosphate group to another nucleotide (Figure 1). Deoxyribose consists of five carbon atoms, each atom identified by a number. A single, linear DNA-strand possesses directionality as indicated by the 3$^{rd}$ deoxyribose carbon atom which connects one nucleotide through a chemical bond to the phosphate attached to the next nucleotide in the DNA sequence. The DNA direction indicated on one end of a DNA molecule by the 3$^{rd}$ deoxyribose carbon atom is consequently termed 3' (pronounced:"three prime end"). On the other end, the 5$^{th}$ carbon atom connects one nucleotide to the next and is consequently called 5' end (pronounced: "five prime end").

Nucleotides linked together forming a chain are the building blocks of a DNA strand. Francis Crick and James Watson made the famous discovery from data generated in another lab that this single strand usually binds to another, complementary strand of DNA through the unique pairing of each nucleotide with a single other nucleotide [1]. "Unique" in this context also means that A can only bind to T and C to G through electrostatic attraction, termed hydrogen bonds (Figure 1). This complementary binding pattern results in both strands to bind to each other in opposite directions. As a consequence, the nucleotide sequence on one strand from 5' to 3' encodes the same sequence information as in the opposite strand in the 3' to 5' direction (Fig. 1). For reference genomes, one of the two strands is often referred to as a 'forward strand' and the other 'reverse strand'. This designation is made arbitrarily, as in complex genomes half of the genes are located on the forward and the other half on the reverse strand. This biochemical structure is responsible that each DNA molecule with both strands contains two copies of the same information.

**Figure 1.** *Schematic Illustration of Complementary DNA Strands Backbone with the Hydrogen Bonds of the Base Pairs (Cornelius Courts, "Basics – Die DNA," blooDNAcid. Mai-2011)*




DNA sequencing' means determining the order and identity of each nucleotide (A, C, G and T) in the DNA polymer chain. By convention, a DNA sequence is recorded as the forward strand sequence from the 5' to the 3' end. A substring of this sequence of length k is called k-mer.

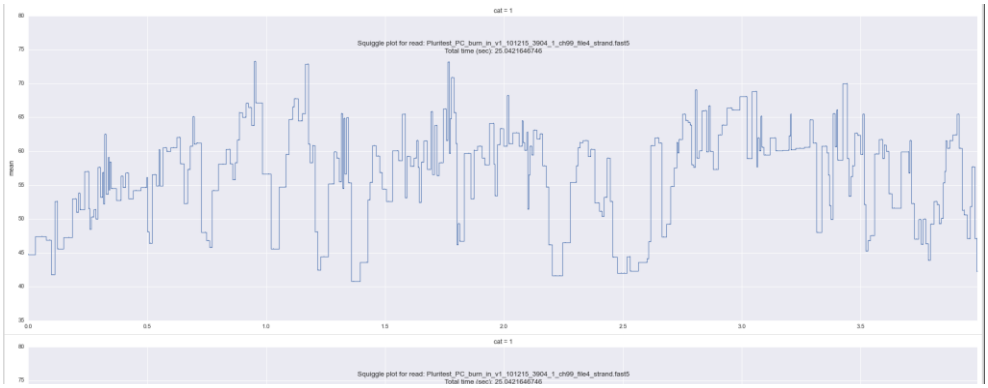
DNA sequencing enables numerous important applications in the biomedical field, from the identification of pathogens to selecting individualized cancer therapies.

A new method for DNA sequencing has become commercially available in 2014 when the British company Oxford Nanopore Technologies (ONT) introduced their MinION DNA sequence [2]. It is a small and portable device powered by a USB interface which enables the user to perform DNA sequencing almost anywhere anytime, i.e. to generate the DNA sequence data information based on A, C, G and T from some biological tissue. The technology used in the MinION is nanopore sequencing. A nanopore is a small protein pore in the size of nanometers with a hole approximately 1 nanometer in diameter in its center. Positioning the pore into an insulating membrane allows putting an electrical voltage to the pore. With this voltage across the pore, a positively charged DNA strand fed through the hole will result in a change of electrical current across the membrane while the different nucleotides traverse the pore. As these current

changes are characteristic for each nucleotide, this technology provides a way of directly reading the sequence of a DNA molecule. In the MinION, six nucleotides are inside the pore at one point of the current sample. So a signal for each nucleotide is generated six times in the context of different 6-mers. To be able to feed the DNA molecule to the nanopore first an adaptor protein, also referred to as 'motor protein' needs to be attached to the DNA sample, which in turn attaches and then guides the DNA strand through the nanopore.

The MinION has a total number of 512 nanopores in one so-called flowcell. In the configuration available the pore R7 is used. With this setting, each nanopore can read 70 nucleotides per second. A new generation of nanopores will be available with R9 in 2016 with a sequencing speed of approximately 250 to 500 nucleotides per second. DNA samples from all animals and plants are too large for being read by one pore as one strand. For example, the human genome encompasses of approximately three billion base pairs. Hence, a DNA sample usually is first broken randomly down into shorter sequences with enzymes or physical shear forces. Afterward, these smaller nucleotide sequences can be sequenced by several nanopores in parallel. As the DNA disruption is a stochastic and random process, this type of sequencing is often referred to as a 'shotgun sequencing'. The Gaussian distributed average length of DNA sequences suitable for sequencing with the ONT R9 nanopore is approximately 20.000 bases per read.

**Figure 2.** *Squiggle-Plot of a MinION Read using Poretools [8]*



To read out the sequence of the DNA sample, it would be enough to only read one of the strands, because the other strand is the complement and hence redundant. With errors introduced by the sequencing method and chemistries used to prepare the sample, the redundancy can be used to lower the error rate of the overall sequencing result. Oxford Nanopore Technologies refers to this as a 2D-read. For this, the two strands are separated in the preparation. Afterward, the motor protein binds to the 5' end of the forward strand and a so-called hairpin adapter, a known sequence of nucleotides, is attached to the 3' end. To this hairpin adapter, the reverse strand is attached. If the whole chain of molecules now gets fed to the nanopore first the sequence of the forward strand will be read, then the known hairpin adapter sequence and afterward the reverse strand. In the

next step, the two strands can get aligned to each other to have two readouts per nucleotide for the final sequence determination. Consequently, in the preparation, not all samples are assembled to 2D-reads. A mixture of forward strands, 2D-reads, and single reverse strands get sequenced by the MinION.

The MinION itself does not generate a string of characters but produces a series of events, to be more precise a sequence of changing current values, called squiggles (Figure 2). One critical step in sequencing biological tissue is the mapping of squiggles towards a corresponding DNA element (A, C, G or T). Oxford Nanopore Technologies offers a cloud instance called Metrichor as a solution for interpreting the raw data as a nucleotide sequence. This process is termed 'base calling', as Metrichor maps one of the four bases (A, C, G or T) onto the squiggle data. After the completion of this time consuming step, the sequence is downloaded back to the computer that operates the MinION.

One of the unique features of the MinION technology is the ability to terminate the read process by reversing the applied electrical voltage of the nanopore. As a result, the DNA molecule inside the nanopore is ejected. After the ejection, the motor protein falls off the DNA polymer. As a result, the same DNA molecule will not be fed to any nanopore afterward. This feature enables to only sequence sample parts that are of interest, by determining based on the information provided by the first few sequenced nucleotides if a sequence is of interest to the biological question and decided if the sequencing should progress beyond the fist nucleotides. For example, if there are impurities in the sample mixture, such as fragments of the DNA from other, contaminating organisms, these can be rejected from being sequenced. Thus, salient features of the nanopore technologies can be utilized to answer the biological question in the shortest possible time. Up to this day, ONT has not made this feature publicly accessible, but it is commonly referred to as "Read-Until" [4].

To be able to distinguish if the read inside the nanopore is of interest or not the sequence has to be looked up in a reference genome. Such a reference is usually a DNA sequence characteristic for the organism to be analyzed. For many organisms, these references are publicly available. For the biomedical research community, the Human Genome Project assembled the human reference genome. The first draft sequence was published in 2001 [3] and the project completed in 2003.

The key disadvantage of the current nanopore sequencing technology we aim to improve is that the sequence of a read from the MinION is only available after processing by Metrichor and then downloaded to the user's computer. So the only data that is usable off-line is the squiggle data. This off-line domain will be referred to as the squiggle space. In the squiggle space, the mean values and standard deviations for each k-mer are recorded as summary statistics representing the raw electrical current measurements, which are sampled at a rate of 3 kHz. As the currents over a single pore are in the range of pico-Ampere, mapping of the raw data in the squiggle space is complicated by three different technical noise factors: shift, scale, and drift.

Shift refers to the effect that during sequencing the voltage over the pore will decrease with the continuous wear of the proteinaceous pore itself. Hence, the current values of the squiggle will also decrease over time. Scale describes the variability of the current's amplitude of the squiggle over time. Lastly, drift represents differences in length of each single k-mer mean value because of continuously changing speeds of the molecule traversing the pore. Causes for this are assumed to be complex 'wound up' 3D patterns inherent to a long DNA molecule, which needs to be straightened out to fit in the pore.

Additionally complicating are inserted or deleted bases in the read squiggle referred to the reference sequence. One reason for such differences can be technical effects caused by the preparation chemistries or inaccuracies in the base calling step. The scientifically most important reason is biological differences between the sample DNA and the reference DNA. Every human possesses a unique, individual DNA sequence. In humans, 99.8% of any given individual genome is identical to those of other humans. The 0.2% difference results in approximately 6 million bases which are variable from human to human. On the technical level, for sequence alignment, variable, inserted or deleted bases need to be accounted for.

The first available system using "Read-Until" was described by Matt Loose and his colleagues in 2016 [4]. The researchers employed the Dynamic Time Warping (DTW) algorithm [5] to align squiggle data to a small reference genome. DTW is a well-known algorithm that is used in several speech and handwriting recognition tasks. It is a dynamic programming algorithm able to deal with insertions and deletions. Also, small differences in the amplitude of two signals will result in a high similarity score.

As "Read Until" and the MinION technologies are very recent developments and are rapidly evolving, this paper aims at a critical assessment of the DTW algorithm used for the above-outlined use case. We will review previous work in this field and expand on the publicly available source codes of the "Read Until" [6] to enable testing and evaluation of DTW’s efficiency.

To be able to parallelize "Read Until" to a system in which every nanopore can have its own "Read Until"-core distinguishing if the currently sequenced read is of interest or not, the system shall be transferable to an FPGA. In 2010 Sart and colleagues investigated the performance of DTW on different hardware [7] and showed that meaningful increases in performance are expected in FPGA environments. In VHDL, a hardware description language used in programming FPGAs, floating point operations are very cumbersome and tend to produce very complex implementations. To have a lower complexity of a DTW implementation on FPGAs, it should ideally only contain integers and thus simple bit operations. Hence, we tested the effect of a conversion of floating point values to integer values of a million times the float value on the DTW performance.

## Methods

As the basis for our study, we used a 2D dataset consisting of 10922 reads generated in a house with the MinION using the currently available R7 pore generation. The data generated was part of an initial experiment mandatory for each MinION device, called 'burn-in' experiment.

Briefly, lambda-phage DNA provided by ONT was processed with the ONT Nanopore Sequencing Kit (R7 version, SQK-MAP006) following the manufacturer's instructions and nanopore sequenced on a MinION Flow Cell Mk I (R7 version, FLO-MAP103).

Our dataset includes the squiggle as well as the base-called data. Out of all reads, only 5614 reads were base-called and subjected to the quality checks of the metrichor cloud instance. That results in 51.4% of all reads generated passing ONT's quality control. Reads not passing the quality filters consisted either only of one DNA strand, or the base calling error rate was too high. Unfortunately at this moment, no quality control information is provided in the squiggle space.

From each read, 250 bases from the beginning were extracted and were aligned to the reference genome using the Smith-Waterman algorithm (SW). SW is a computationally expensive, dynamic programming algorithm used to create accurate alignments of character-based DNA sequences. SW can handle insertions and deletions of bases. As a result, we computed a high-quality reference alignment to benchmark the DTW alignment.

Next, we translated the character based reference of the lambda-phage virus into a squiggle space utilizing a k-mer model provided by Metrichor. This generated reference is called squiggle reference. Afterward, we aligned 250 squiggles from the beginning of every read in the dataset to the squiggle reference with the DTW algorithm. After this step for every read in the dataset there is an alignment made with SW and one with DTW, assumed the read is able to align to the reference. To lower the impact of the shift and scale effect in the reads, they need to be z-score normalized as stated in [4]. This circumstance has been considered in the implementation. Regarding the drift effect the DTW itself accounts for it by handling insertions and deletions. As the last step, we developed a summary statistic to assess the accuracy of the DTW results in comparison to the alignment made with SW.

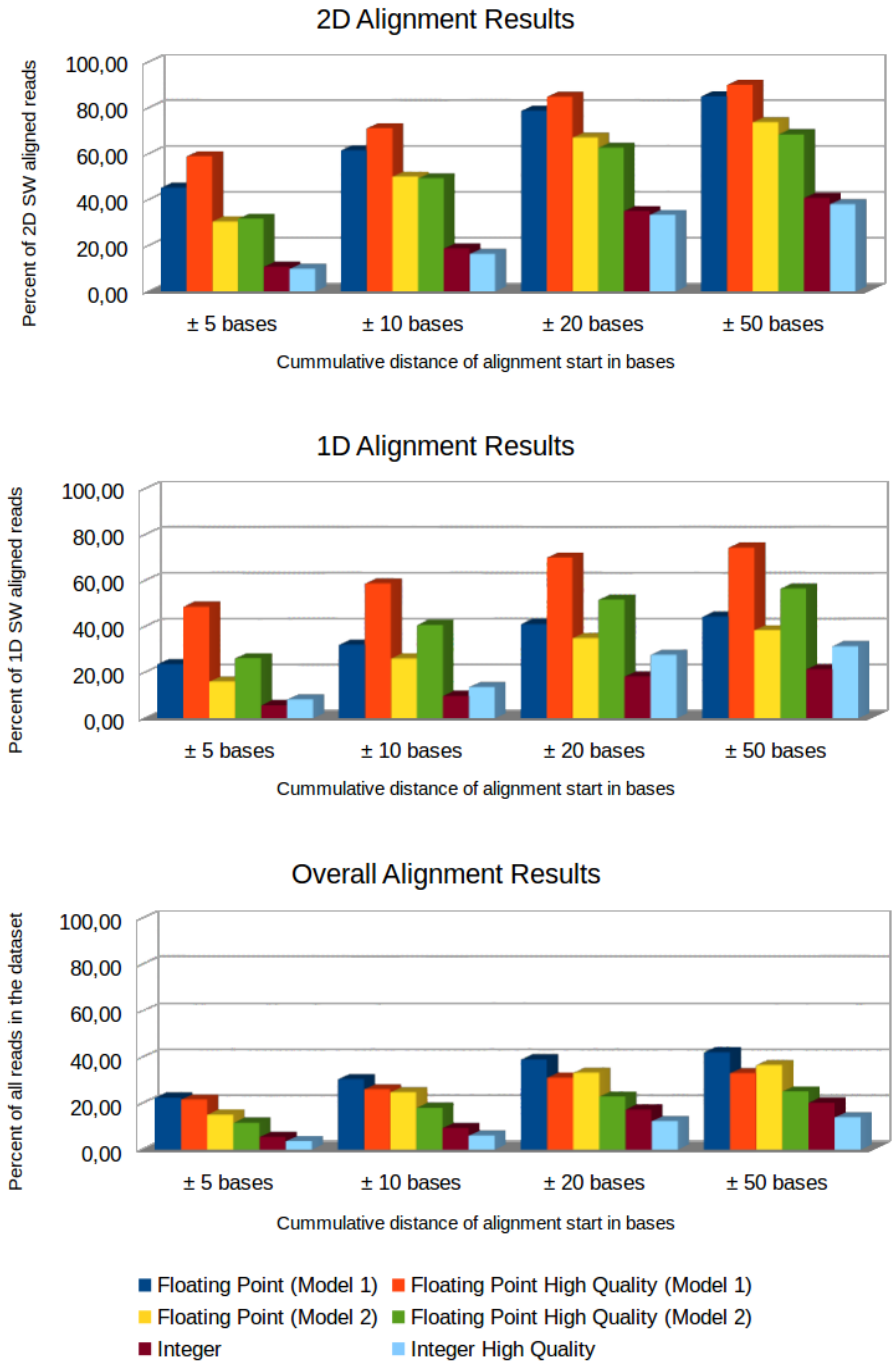
## Results

Our experiments revealed several insights: For one, the base-calling model currently provided by Metrichor has an enormous impact on the overall DTW alignment result. In our in-house generated dataset Metrichor employed two different models for reasons unclear to us. For another, the filtering introduced by Metrichor's quality control makes it difficult to compare the DTW and SW alignments. Only 51.4% of the reads passed the Metrichor filter, but the other 49.6% may also generate SW alignments, since they were still base-called. If the basecalling result is incorrect due to a poor read quality the SW alignment of the character string will also lead to an erroneous result. As a consequence the results for the overall and 1D

alignments incorporate an uncertainty and only the high quality reads may be taken as fully comparable. Unfortunately this uncertainty is not easily to evaluate for the user as the basecalling process is closed source and proprietary to ONT.

Our results are summarized in Figure 3 and reported with three distinct metrics:

**Figure 3.** *Alignment Results of the Experiments Separated into Three Groups: Reads that Have Two SW Aligned Strands (2D Reads), Reads that Have Exactly One SW Aligned Strand (1D Reads), and All Reads whether they Were SW Alignable or Not*



The first metric is the percentage of reads that were aligned compared to the total amount of reads.

The second compares to the amount of reads that had at least one strand aligned with the SW and DTW algorithm, the 1D reads and the third metric shows the amount of reads fully able to align, the 2D reads that were aligned correctly.

Furthermore, different quality groups are separated. The best results lay in a distance of ± five bases. That means that the DTW alignment start, about the reference, is only five bases away from the alignment start of the SW alignment. The next groups are ± 10, ± 20 and ± 50 bases. Even if the alignment is 50 bases away, the result could help distinguish if the read is in an area that is of interest or not. The difference as such is a result of DTW and SW handling insertions and deletions. They may be handled differently by the algorithms.

Altogether the experiments were conducted with two reference sequences translated from the two different base-calling models provided within our in-house generated dataset. A third test was made with all values in integer representation. Each experiment is compared to a reference alignment consisting of all reads in the test dataset and an alignment only composed of the reads that passed the quality filter of Metrichor. The latter metric is referred to as "High Quality" in Fig. 3.

The results show that 90% of the DTW alignments lay only 50 bases away from the reference alignment, if the 2D reads are taken into account that passed the Metrichor quality filter. That means only 10% of the reads with high quality which lay outside of the region of interest get sequenced. Consequently the time needed to answer a biological question can be reduced considerably.

We also performed a test with all reads of the dataset to get an overall accuracy of the DTW. As compared to a random hit, which would have a hitting probability of 25% for the 4 DNA-elements A, C, G and T (under the assumption that all elements are equally distributed), our results proved a hitting rate of roughly 42% of reads with the DTW alignment laying only 50 bases away from the reference alignment. This is if the overall results are taken into consideration, i.e. the DTW aligned reads are compared to the total amount of reads in the dataset. These results show that the DTW can detect the existence and position of a short sample inside a reference even in the presence of errors.

The integer representation implementation shows a clearly worse result. Only 20% of the DTW alignments get reported to be only 50 bases away from the reference alignment when compared to the total amount of reads. Using only high quality reads the integer DTW implementation can align 40% of the reads inside the 50 bases range. These results are less than half of the floating point DTW implementation, and the current implementation is not viable in a production system.

## Conclusions

The results demonstrate that DTW alignment of DNA sequence data is possible in the ONT squiggle space. With the new generation R9 of the

nanopore, the overall 1D accuracy is expected to further improve and the DTW alignment in squiggle space can then be used even for 1D sequencing. The preparation time of the sequencing will only be 20 minutes which will further shorten the time to answer. In a typical use case of a MinION DNA sequencing experiment, 90% of the reads that may be processed further can be classified inside the reference to distinguish if it is of interest or not. With this technology, the time to sequence biological tissue to answer a biological question could dramatically be reduced.

## References

[1] J. D. Watson and F. H. C. Crick, "Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid," *Nature*, vol. 171, no. 4356, pp. 737-738, Apr. 1953 [Online]. Available: http://go.nature.com/1l8VADg. [Accessed: 18-May-2016].

[2] "Start using MinION - Community - Oxford Nanopore Technologies." [Online]. Available: http://bit.ly/1qpoVi0. [Accessed: 18-May-2016].

[3] J. C. Venter et al., "The Sequence of the Human Genome," *Science*, vol. 291, no. 5507, pp. 1304-1351, Feb. 2001 [Online]. Available: http://bit.ly/2h92VVj. [Accessed: 18-May-2016].

[4] M. Loose et al., "Real time selective sequencing using nanopore technology.," *bioRxiv*, p. 38760, Feb. 2016 [Online]. Available: http://bit.ly/2h8LfLR. [Accessed: 04-Feb-2016].

[5] "Dynamic Time Warping," in *Information Retrieval for Music and Motion*, Springer Berlin Heidelberg, 2007, pp. 69-84 [Online]. Available: http://bit.Ly/2gdEl3y. [Accessed: 03-Mar-2016].

[6] "mattloose/RUscripts," *GitHub*. [Online]. Available: http://bit.ly/2geeahP. [Accessed: 20-May-2016].

[7] D. Sart et al., "Accelerating Dynamic Time Warping Subsequence Search with GPUs and FPGAs," in *2010 IEEE 10th International Conference on Data Mining (ICDM)*, 2010, pp. 1001–1006.

[8] N. J. Loman and A. R. Quinlan, "Poretools: a toolkit for analyzing nanopore sequence data," *Bioinformatics*, vol. 30, no. 23, pp. 3399-3401, Dec. 2014 [Online]. Available: http://bit.ly/2h8Chyj. [Accessed: 15-Oct-2015].