Wageningen University and Research

Modelling the relation between event duration and GC-content in MinION nanopore sequencing data

Finnegan, Thijs
2018

Modelling the relation between event duration and GC-content in MinION nanopore sequencing data

Abstract

The MinION is a nanopore polynucleotide sequencing device that produces long reads, yet currently lacks accuracy basecalling accuracy. As the MinION uses a helicase to control the sequencing rate and helicase processing rate is in nature linearly dependent on GC-content, it is hypothesised that this relation could be used to improve nanopore sequencing accuracy. Here we show that linear models and basic random forests are not able to accurately predict GC-content from the sequencing rate in nanopore data. We conclude that variation in sequencing rate either follows a more complex model then was previously thought or is more heavily influenced by something other than GC-content.

Introduction

The field of genomics has seen major advancements over the last decades. These have been dependent on the development of novel genome sequencing methods. One of the largest developments has been the introduction and improvement of second generation sequencing, allowing for quick and affordable sequencing of genomes. Second generation sequencing methods break DNA strands up into small fragments and amplify them by PCR. The fragments are read in parallel and the overlapping small reads are pieced together to get the sequence of the entire DNA strand [2]. Although this greatly increased throughput over the previously developed Sanger sequencing methods, assembly becomes more difficult as the size of the strand increases. Second generation sequencing also has problems with recurrent sequences; short reads from recurrent sequences look similar, causing these regions to collapse during assembly [3].

In recent years novel sequencing methods have been developed that deal with these problems by producing long reads at a single molecule level. These methods are collectively referred to as third generation sequencing methods. As the reads are longer there is more overlap between reads, increasing the certainty that the overlaps are correct. This leads to third generation sequencing requiring less coverage and therefore requiring less reads to reach the same assembly accuracy. With more overlap and fewer reads the assembly process is also significantly easier than it is for second generation sequencing. The longer reads also allow for better reconstruction of repetitive sequences [4]. Another advantage is that some third generation sequencing methods such as nanopore sequencing don’t require samples to be amplified with PCR. This makes preparation easier and also removes errors that can be caused by PCR [5]. Nanopore sequencing is a third generation sequencing method that uses a membrane over which an electrical current is applied. A protein pore through which electrical current can flow is inserted into the membrane. Single stranded DNA is pulled through by an electrical potential, causing changes in the current. As measured current levels are mostly specific to the bases moving through the pore, the current signal can be translated into a sequence [6, 7]. The signal can be divided into events that correspond to a set of bases. The duration of each event is ideally equal to the amount of time it takes for the strand to move by one base.

The MinION developed by Oxford Nanopore Technologies is a small and relatively inexpensive polynucleotide sequencing device that utilizes this method. Since it was first released it has seen

enormous advancement in both the physical components as well as the basecalling software. Accuracy of reads is however still an important point for improvement and currently sits at 87.6% [4, 8]. Deletions are the most common errors [9]. Homopolymeric regions are particularly likely to cause deletions and often get truncated [4], because the current levels produced for these regions are homogeneous and thus difficult to divide into events. Using specific properties of the signal directly could be a novel way to improve accuracy.

A potential solution can be found in the mechanism controlling DNA strand processing speed. The speed at which DNA moves through the nanopore is by itself too high for accurate measurements [10, 11]. Because of this a molecular motor is used to control the speed and move the strand through the nanopore one base at a time [12]. As a molecular motor the MinION uses a helicase, which has been heavily modified over time. Helicase takes double stranded DNA and splits it into two single strands. In nature the speed at which DNA moves through a helicase is dependent on the fraction of nucleotides in the strand that is either guanine (G) or cytosine (C). An increase in GC-content linearly slows down the rate at which helicase processes DNA [13]. This is due to stronger base pairing between guanine and cytosine and more stable base stacking interactions compared to adenine (A) and thymine (T), making it harder for helicase to split the DNA [14]. Despite this being true in nature it cannot be assumed to be the same for the MinION. The helicase used by the MinION is heavily modified and operates in a controlled environment specific to nanopore sequencing [15]. This may affect helicase functionality, as the stability of DNA is affected by properties of its environment (e.g. salt concentration [14]).

Before sequencing poly-T adapters are attached to the double-stranded DNA. Helicase binds to these adapters which block its function until it reaches the pore [8]. When the helicase reaches the pore it starts to split the DNA and feeds a single strand through at an average rate of 450 bases per second [4]. The rate at which DNA moves through the pore varies, which influences the duration of events [7]. A logical candidate for the cause of this variation in event duration is the helicase since it controls the speed of the DNA. It would therefore also be expected that the GC-content affects the event duration. If there is a strong relation between GC-content and event duration then a model could be used to predict GC-content. This could be a novel way to increase accuracy by checking the results of basecalling against the model.

In this research the relation between GC-content and event duration was modelled. Models were constructed and scored on all probable combinations of distance and *k*-mer size, which are parameters that need to be known to create an accurate model. The best performing models had expected values for distance and *k*-mer size but the correlation found was too weak to be used for basecalling correction.

Methods

Data

For this research two datasets were used consisting of basecalled nanopore sequencing data in fast5-format. The first dataset was sequenced from chromosome 19 of the na12878 human DNA. After removing the initial quality scans this set contains 3998 fast5 files and over 20M bases. The second dataset is amplicon data obtained generated from the shufflon region of 6 IncI1 plasmids in 2 *Escherichia coli* strains, which was sequenced by the bacteriology and Epidemiology group of Brouwer using the MinION Mk1.8 (Oxford Nanopore Technologies), using an R9.4 flowcell (FLO-MIN106) and a native barcoding kit (EXP-NBD103). The dataset consists of 674496 fast5 files, together containing over 460M bases. In order to increase the quality of the reads both datasets were resquiggled using the resquiggle algorithm of the Tombo toolkit [16], which compares the basecalls to a reference sequence and then reassigns bases to the raw signal, thus correcting basecalling errors and event durations.

Simulated data was used to test the models and graphs. Simulated data was made by picking random bases and generating corresponding event durations by multiplying GC-content with a constant and adding Gaussian noise. The data was centred at a distance of 3 bases and values with a distance under 3 were removed to make it more similar to the real data.

Parameters

From fast5 files the bases and event duration were read. Event durations are expressed in number of measurements made before finding a new *k*-mer. This version of the MinION has a sampling rate of 4000 Hz and a sequencing rate of 450 bases per second. Bases were encoded as two classes, GC and AT. For the model it is necessary to determine the distance in bases (D) between the active site of the helicase and the constriction of the pore, as well as the number of bases (K) influencing the helicase processing rate at once (Figure 1).

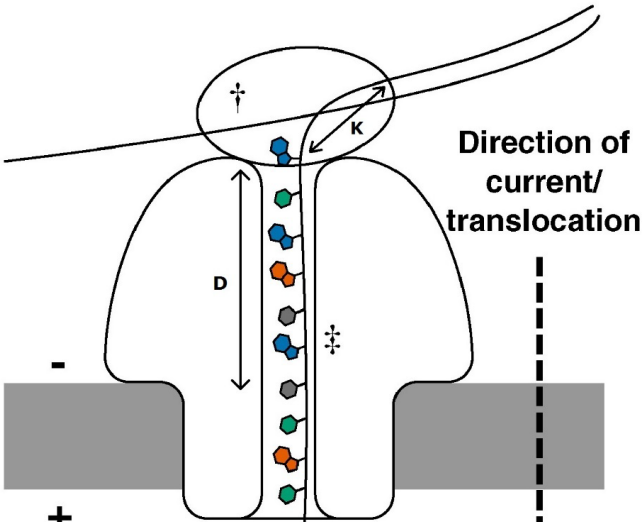

Figure 1: A helicase (top) and pore (bottom). Model parameters for k-mer size (K) and distance (D) to the pore constriction are marked. Modified image from [1]

It is necessary to know the distance between the position where bases are read, i.e. the constriction site of the pore, and the bases residing inside the helicase, where processing speed is expected to be controlled. If the distance is known then the event durations can be connected to the *k*-mers in the active site of the helicase during measurement. The size of *k* needs to be known to maximize model accuracy, since this should be the number of bases influencing each event

duration. To determine distance and *k*-mer size each likely combination of *k* and *d* was modelled and scored. With the number of bases influenced by a helicase being 8-10 according to literature [17, 18], *k*-mer sizes of 1 through 10 were tested. Because the length of the nanopore is 4 nm before the constriction where base reading takes place, and the number of bases in a nm of ssDNA is 1.4 [19, 20], distances of 4-40 bases were used.

Linear models

For the linear models two methods were used to interpret GC-content of the *k*-mers. First the event duration was used as a variable to predict the fraction GC-fraction of each *k*-mer, following the equation $Y = b_0 + b_1*x$. For the second method each base in the *k*-mer was categorized as either GC or AT. Each base in the *k*-mer was then used as a separate binary variable to predict event duration, with the formula $Y = b_0 + b_1*x_1 + b_2*x_2 ... + b_n*x_n$. Because of the large size of the amplicon dataset a random sample of 12000 reads was used while modelling with individual bases. This subset contains over 8 million bases.

Outliers

Extremely long event durations are likely caused by obstructions of the pore, thus outliers in event duration were removed to increase the quality of the data. Different cut-off values for event duration were tested. The one that gave the biggest increase in R^2 was an event duration of 40, which was therefore chosen as the cut-off value. Removing low values did not improve the model and also removed most of the data.

Modelling

The data was modelled with linear models and random forests. Models were constructed in Python 3.5 (the Python Foundation: www.python.org) using the Scikit-learn package (version 0.19.2) [21]. Linear models were trained and scored using cross-validation. The simpler models build on only GC-fraction were assessed with 5 fold cross-validation. The more complex models using individual bases as binary variables were assessed with 3 fold cross-validation to reduce computing time. Of each model an R^2, mean squared error and explained variance was determined. R^2 was used to choose the best models. Besides linear models the amplicon data was also modelled with random forests in case of a non-linear relationship.

Results

Resquiggle

When applying the resquiggle algorithm of the Tombo toolkit it is normal for part of the data to fail. This can for example be because of poor raw to expected signal matching. Of the human DNA dataset 551 out of 3997 files were not resquiggled, which is 13.8%. For the amplicon dataset this was 94824 out of 674496 files, or 14.1%. The failure rate for both is within the realm of expectation. The files that failed were not used during this research.

Data distribution

This research focuses on event duration and GC-content. Here we visualized these properties of the data set. Event duration describes how much time was between measurements of bases and is expressed in the number of measurements before a new base was found. GC-content describes the amount of guanine (G) or cytosine (C) in a group of bases, called a *k*-mer. Neither dataset contains event durations under 3. An event duration of 3 is the most common, with larger values decreasing in frequency. Events with a duration of 4 do not occur often enough to follow this trend (Figure 2a). Figure 3b shows GC-content for k-mers of size 3, which almost follows a random distribution, but with fewer *k*-mers containing only G and C bases. The overall GC-content of the human dataset is 47.8% and that of the amplicon dataset is 46.0%. This matches the literature, with chromosome 19 having a reported GC-content of 48% [22] and the amplicon shufflon having a GC-content of 46%.

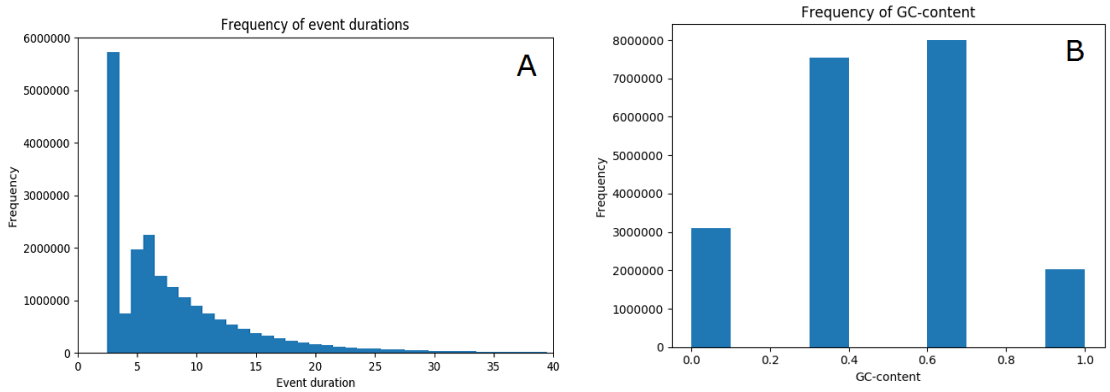

Figure 2: Distribution of event duration in bases (a) and GC-content for k-mer size 3 (b) in the human DNA dataset.

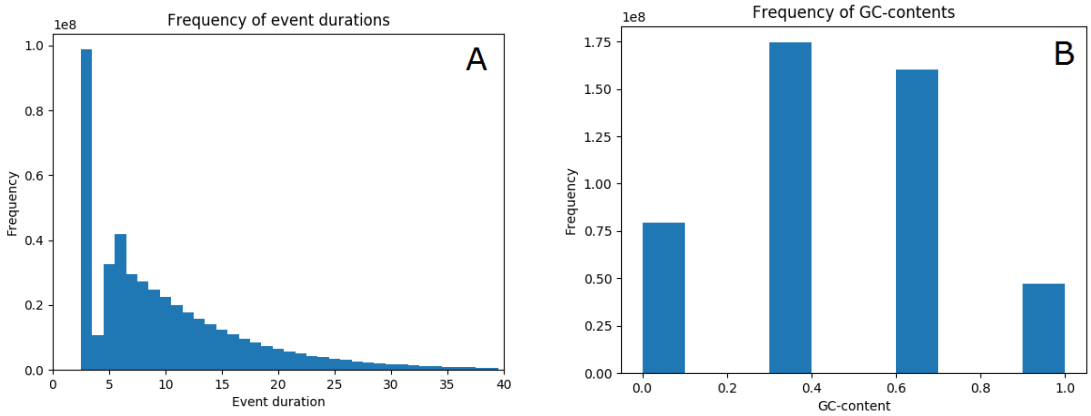

Figure 3: Distribution of event duration in bases (a) and GC-content for k-mer size 3 (b) in the Amplicon dataset.

The effect of GC-content on event duration was visualised by creating boxplots of the datasets. Shown below are the plots at distance 12 and *k*-mer size 3. The human DNA dataset shows a slight increase in event duration at higher GC-content, although the median stays the same until 100% GC-content (Figure 4). This shows that the relation between event duration and GC-content is probably very small.

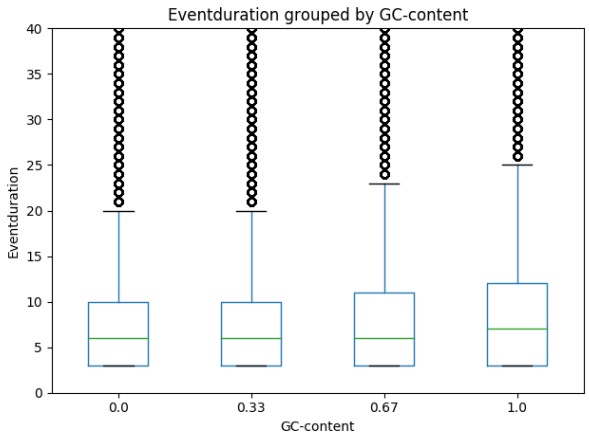

Figure 4: Effect of GC-content on event duration for the human DNA dataset.

The amplicon dataset shows a similar situation to the human dataset, however with even less variation (Figure 5). In this case there is almost no change in the distribution of event durations when increasing the GC-content. The exception is *k*-mers with 100% GC-content, which do have higher event durations on average. The amplicon dataset in general has higher event durations then the human dataset.

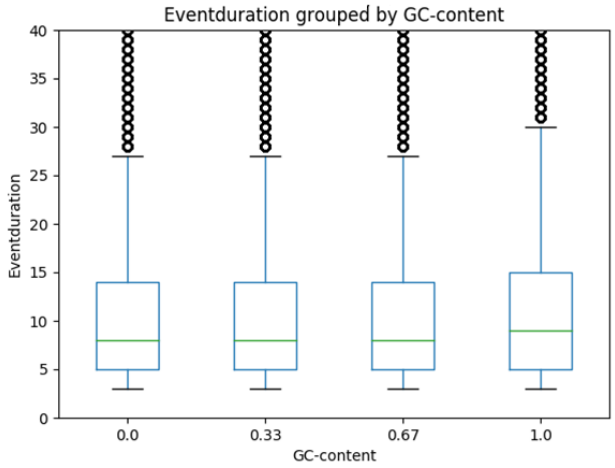

Figure 5: Effect of GC-content on event duration for the amplicon dataset

Simple linear models on human data

Simple linear models were made of percentage GC-content, with each combination of distance in bases and *k*-mer size modelled separately. 5-fold cross-validation was used to evaluate the models. The human DNA dataset gave the highest R^2 at a distance of 13 bases with a *k*-mer size of 3 bases. The R^2 had a value of 2.24e-3 with $p < 5e\text{-}3$ and an MSE of 0.0821. In this case R^2 is the portion of the GC-content variation that can be explained using event duration. An R^2 value of

2.24e-3 is very low, indicating that there is little to no predictive power. Figure 6 shows that distances with higher values decrease the R^2 even further. This suggests that the correct distance is around 11-13. Changing the *k*-mer size lowers the R^2, which supports the hypothesis that there is an optimum number of bases involved at a time. *K*-mer sizes of 1 and 5 got similar R^2 scores, at 2.02e-3 and 2.04e-3 respectively. *K*-mers with an even size and *k*-mers with larger sizes got lower R^2 scores.

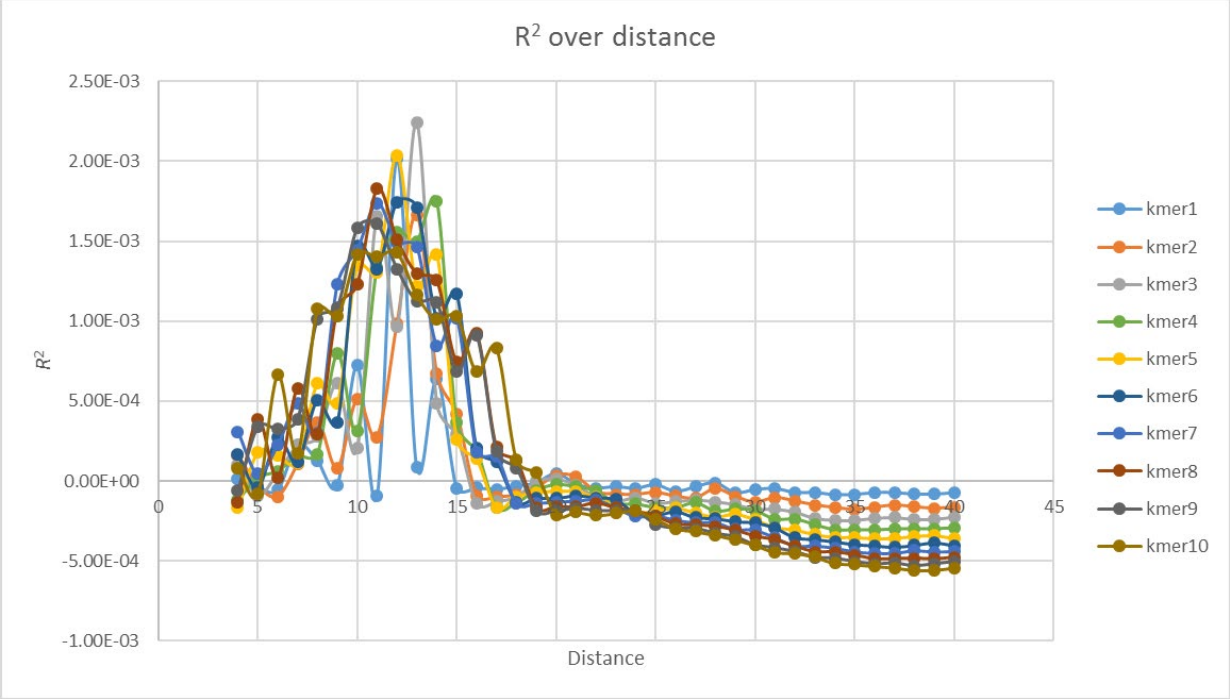

Figure 6: R^2 from different models of the human DNA dataset

A simple linear model on percentage GC-content mostly predicted a GC-content of 46%. No lower values were predicted, while large event durations caused predictions of well over 100% GC-content (Figure 7).

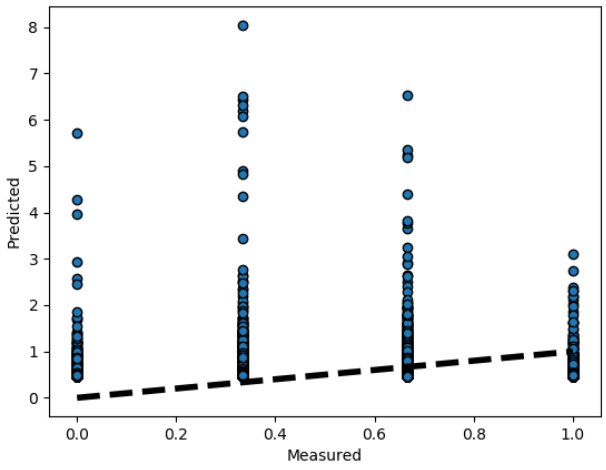

Figure 7: Predicted vs measured for the human DNA dataset

Simple linear models on amplicon data

The amplicon dataset gave similar results to the human dataset, with the best model having a distance of 12 (Figure 8). This time a *k*-mer size of 1 had a slightly higher R^2 than a *k*-mer size of 3. The corresponding R^2 were 3.34e-4 and 3.21e-4 respectively. The models with distance 12 and *k*-mer sizes of 6 or more had very low R^2 values. Except for distance 12, most distances had an R^2 value below 1e-5, but unlike the human dataset they were not negative. The exceptions were distances of 6 or lower, of which some had higher R^2 values with the highest being 2.69e-4 at a *k*-mer size of 4. These higher R^2 values had different distances for each *k*-mer size and were not found in the human dataset.

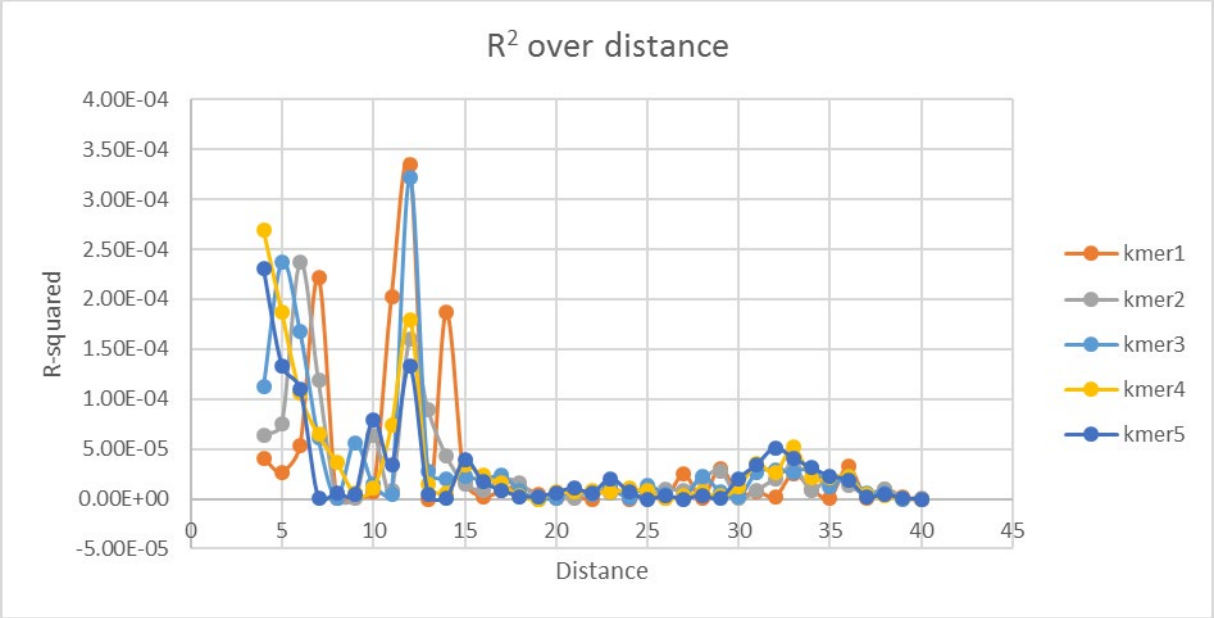

Figure 8: R^2 for different models of the amplicon dataset

Modelling individual bases

The datasets were also modelled using each base in the *k*-mers as a different variable. The models using individual bases have higher R^2 scores with increasing *k*-mer size. This is also true for adjusted R^2, as can clearly be seen in figure 10. The R^2 is still highest around distance 12, which is the most obvious in figure 9 but also true for the *E. coli* results.

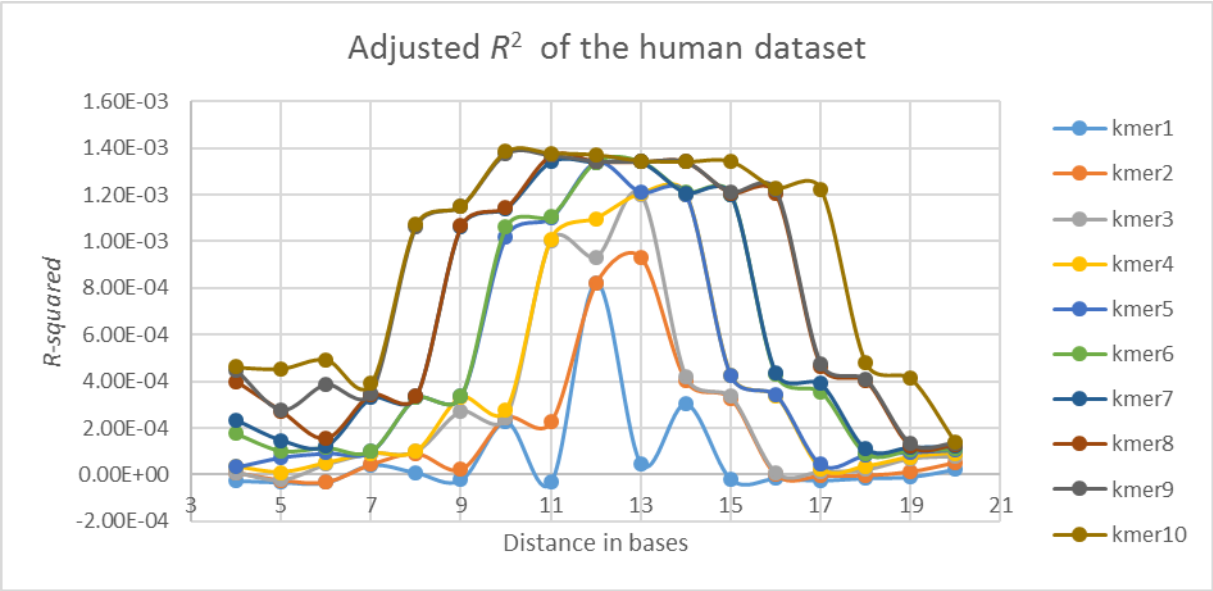

Figure 9: Adjusted R^2 for different models of the human dataset

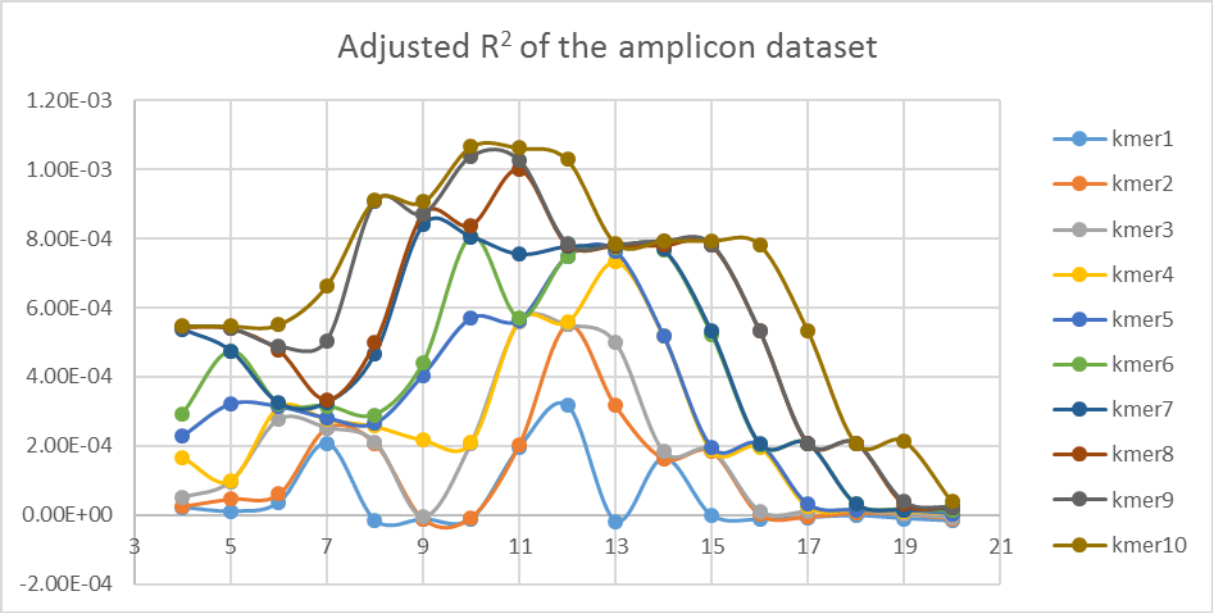

Figure 10: Adjusted R^2 for different models of the amplicon dataset

Random forests

Random forests were made of the amplicon dataset to capture any non-linear relations. A model was made for each combination of *k*-mer size and distance, using individual bases. Of each model a confusion matrix was made. The accuracy for each model was close to 54%. This shows that the random forests have very little predictive power, similar to the linear models. The confusion matrices show that the random forests always predict A or T, never G or C. The datasets contain only 46% GC-content, which makes always predicting A or T a winning strategy when there is no clear relation between GC-content and event duration. Clearly the random forest has no predictive power in this situation.

	Predicted AT	Predicted GC
Measured AT	1744385	0
Measured GC	1483658	0

Example confusion matrix at k-mer 1 and distance 4.

Discussion

Nanopore sequencing is a novel nucleic acid analysis method which produces long reads but currently lacks accuracy [8]. New methods need to be developed to improve this. One possible venue for this is to utilize the variation found in DNA processing rate. In the MinION, a nanopore sequencing device, a helicase is used to control the rate at which DNA moves through the pore. In nature the rate of helicase is linearly dependent on the amount of guanine (G) and cytosine (C) in the strand [13]. Explicitly making use of this by predicting GC-content from event duration could be a novel way to improve accuracy.

In order to model the relation between GC-content and event duration the distance between the helicase, where the processing rate is presumably determined, and the pore, where the actual signal is read out, needs to be determined, as well as the number of bases influencing event duration at once (the *k*-mer). Linear models were made either on percentage GC-content of the *k*-mer or on individual bases, where each base was grouped as either Adenine/Thymine or Guanine/Cytosine . The methods used were shown to be able to identify linear correlations in simulated data. When using real data from *E. coli* and human DNA sequencing the optimal distances found were 12 and 13 bases respectively, which is approximately what was expected considering the size of the pore and the helicase, the optimal *k*-mer sizes were 1 and 3 bases. These gave low R^2 values of 2.24e-3 on human data and 3.34e-4 for *E. coli* data with $p < 0.01$. Cut-off values were chosen to optimise the accuracy, so an even lower relation should be expected. Low p-values were probably because of the large amount of data. Random forests that were made always predicted a base to be either Adenine or Thymine, leading to a GC-content of 0 regardless of event duration, which shows that they have no predictive power. It is possible that the helicase is modified to limit variation in processing rate, which would explain the low accuracy of the models. We do not know for certain since such information is not published, but a constant processing rate is assumed when correcting for homopolymer stretches.

Future research could focus on more elaborate non-linear models, which may be able to predict the relation between GC-content and event duration, such as models that deal well with noise. It is likely that variation in event duration is partially caused by natural stochasticity of biological systems, which needs to be separated from meaningful variation. Research could also be done into other causes of the variation in event duration, for example by using it as a predictor for the location of deletions.

In conclusion, we found that linear models as well as random forests are not able to accurately predict GC-content from event duration. However given the weak but sensible correlation found, more elaborate models may be able to better capture the relation between GC-content and event duration, which can then be used in a novel way to improve nanopore sequencing accuracy.

References

1. Heather, J.M. and B. Chain, *The sequence of sequencers: The history of sequencing DNA.* Genomics, 2016. **107**(1): p. 1-8.
2. Goodwin, S., J.D. McPherson, and W.R. McCombie, *Coming of age: ten years of next-generation sequencing technologies.* Nature Reviews Genetics, 2016. **17**: p. 333.
3. Treangen, T.J. and S.L. Salzberg, *Repetitive DNA and next-generation sequencing: computational challenges and solutions.* Nature Reviews Genetics, 2011. **13**: p. 36.
4. Jain, M., et al., *Nanopore sequencing and assembly of a human genome with ultra-long reads.* Nature Biotechnology, 2018. **36**: p. 338.
5. Simpson, J.T., et al., *Detecting DNA cytosine methylation using nanopore sequencing.* Nature Methods, 2017. **14**: p. 407.
6. David, M., et al., *Nanocall: an open source basecaller for Oxford Nanopore sequencing data.* Bioinformatics, 2017. **33**(1): p. 49-55.
7. Magi, A., et al., *Nanopore sequencing data analysis: state of the art, applications and challenges.* Briefings in Bioinformatics, 2017: p. bbx062-bbx062.
8. de Lannoy, C., D. de Ridder, and J. Risse, *The long reads ahead: de novo genome assembly using the MinION.* F1000Research, 2017. **6**.
9. Magi, A., B. Giusti, and L. Tattini, *Characterization of MinION nanopore data for resequencing analyses.* Briefings in Bioinformatics, 2017. **18**(6): p. 940-953.
10. Manrao, E.A., et al., *Reading DNA at single-nucleotide resolution with a mutant MspA nanopore and phi29 DNA polymerase.* Nature biotechnology, 2012. **30**(4): p. 349-353.
11. Cherf, G.M., et al., *Automated Forward and Reverse Ratcheting of DNA in a Nanopore at Five Angstrom Precision().* Nature biotechnology, 2012. **30**(4): p. 344-348.
12. Bowen, R.V., et al., *Method for controlling the movement of a polynucleotide through a transmembrane pore*. 2017, Google Patents.
13. Manosas, M., et al., *Active and passive mechanisms of helicases.* Nucleic acids research, 2010. **38**(16): p. 5518-5526.
14. Yakovchuk, P., E. Protozanova, and M.D. Frank-Kamenetskii, *Base-stacking and base-pairing contributions into thermal stability of the DNA double helix.* Nucleic Acids Research, 2006. **34**(2): p. 564-574.
15. Feng, Y., et al., *Nanopore-based Fourth-generation DNA Sequencing Technology.* Genomics, Proteomics & Bioinformatics, 2015. **13**(1): p. 4-16.
16. Stoiber, M.H., et al., *De novo Identification of DNA Modifications Enabled by Genome-Guided Nanopore Signal Processing.* bioRxiv, 2017.
17. Lohman, T.M. and N.T. Fazio, *How Does a Helicase Unwind DNA? Insights from RecBCD Helicase.* BioEssays, 2018.
18. Velankar, S.S., et al., *Crystal structures of complexes of PcrA DNA helicase with a DNA substrate indicate an inchworm mechanism.* Cell, 1999. **97**(1): p. 75-84.
19. Goyal, P., et al., *Structural and mechanistic insights into the bacterial amyloid secretion channel CsgG.* Nature, 2014. **516**(7530): p. 250-253.
20. Yan, J. and J.F. Marko, *Localized single-stranded bubble mechanism for cyclization of short double helix DNA.* Physical review letters, 2004. **93**(10): p. 108108.
21. Pedregosa, F., et al., *Scikit-learn: Machine learning in Python.* Journal of machine learning research, 2011. **12**(Oct): p. 2825-2830.
22. Grimwood, J., et al., *The DNA sequence and biology of human chromosome 19.* Nature, 2004. **428**: p. 529.