MinCall — MinION end2end deep learning basecaller

Neven Miculinić, Marko Ratković, and Mile Sikić
{neven.miculinic, marko.ratkovic, mile.sikic}@fer.hr

Faculty of Electrical Engineering and Computing (FER), Zagreb, Croatia

Abstract. The Oxford Nanopore Technologies's MinION is the first portable DNA sequencing device. It's capable of producing long reads, over 100 kBp were reported, however, is has significantly higher error rate than other methods.
In this study, we created MinCall, an end2end basecaller model for the MinION. The model is based on deep learning and uses convolutional neural networks (CNN) in its implementation. For extra performances is uses cutting edge deep learning techniques and architectures, batch normalization and Connectionist Temporal Classification (CTC) loss. The best performing 270 layers deep model achieves state-of-the-art 90.5% match rate on E.Coli dataset using R9 pore chemistry and 1D reads.

Keywords: Basecaller, MinION, R9, CNN, CTC, Next generation sequecing

1 Introduction

In recent years, deep learning methods significantly improved the state-of-the-art in multiple domains such as computer vision, speech recognition, and natural language processing LeCun and Bengio (1998)Krizhevsky et al (2012). In this paper, we present application of deep learning in the field of Bioinformatics for DNA basecalling problem.

Oxford Nanopore Technology's MinION nanopore sequencing platform Mikheyev and Tin (2014) is the first portable DNA sequencing device. It's small weight, of only 90 grams, low capital cost, and long read length combined with high-throughput, real-time data analysis, and decent accuracy yield promising results in various applications. From clinical application such as monitoring infectious disease outbreaks Judge et al (2015)Quick et al (2016), characterizing structural variants in cancerNorris et al (2016) and even full human genome assembly Jain et al (2017).

Although MinION is able to produce long reads Loman (????a,b), they have a high sequencing error rate. This has been somewhat alleviated with new R9 pore model, replacing older R7 ones. In this paper, we show that this error rate can be reduced by replacing the default base caller provided by the manufacturer with a properly trained neural network model. In the future new R9.5 chemistry and 1D$\hat{2}$ reads should supersede current models.

In the MinION device, single-stranded DNA fragments move through nanopores, which causes drops in the electric current. The electric current is measured at each pore several thousand times per second, 4000 times exactly in our dataset. The electric current depends mostly on the context of several DNA bases passing through the pore at the time of measurement. As the DNA moves through the pore, the context shifts, and the electric current changes.

The MinION device typically yields reads several thousands bases long, even couple hundred thousand bases long reads were repoted Loman (????a,b). However the cost in on accuracy, signicifantly lower than older, more reliable and expensive sequencing methods.

The exact error rate metric is unreliable since multiple pipeline tools could be the issue. First the sample is prepared, hopefully, uncontaminated and matching reference genome as close as

possible, then sequenced using the MinION device obtaining raw data. Next, our model (or other groups ones) come along, basecall the sequence. To evaluate error rate metric basecalled read is aligned to the reference genome using aligners with their own errors/biases, mostly commonly used BWA-MEM Li (2013) and Graphmap Sović et al (2016).

2 Background

3 Sequencing overview

Conceptually, the MinION sequencer works as follows. First, DNA is sheared into smaller DNA fragments and adapters are ligated to either end of the fragments. The resulting DNA fragments pass through a protein embedded in a membrane via a nanometre-sized channel, a nanopore. A single DNA strand passes through the pore. Optionally, hairpin protein adapter can merge two DNA strands, allowing both template and complement read passing through the nanopore sequentially for more accurate reads. This technique is referred as 2D reads, while we focus on 1D reads containing only template DNA and no hairpin adapter.

Electrical current runs through the nanopore and exact nucleotides context within influences the nanopore's resistance. This resistive effect is our sensor data, that is the current fluctuations as DNA passes though the pore. The nanopore is 6 nucleotides wide, and many models use this information in their approaches, while we're opted out of this technicality.

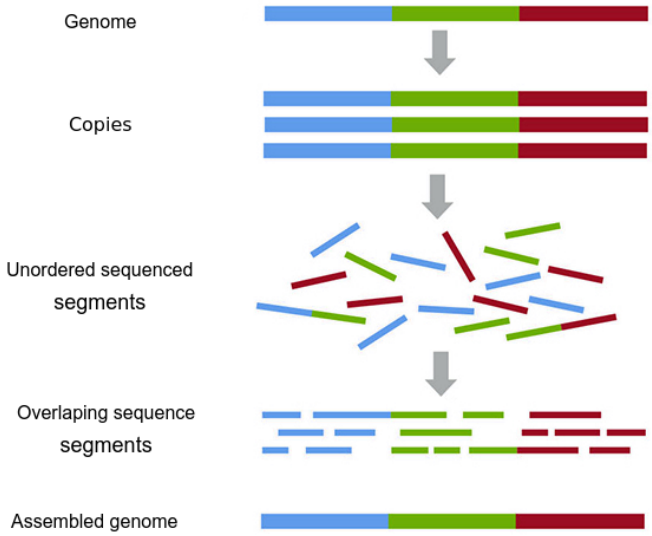


Fig. 1. Depiction of shotgun sequencing

4 Basecalling

The core of the decoding process is the basecalling step. Nowdays there's multiple basecalling options, what official and unoffical ones.

[1] Figure adapted from https://nanoporetech.com/how-it-works

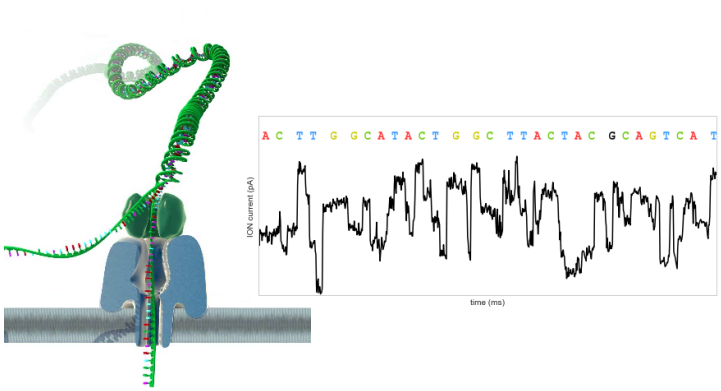


Fig. 2. DNA strain being pulled through a nanopore [1]

Earlier models were Hidden Markov model(HMM)-based where hidden state modeled DNA sequence of length 6 (6-mer) in the nanopore. Pore models were used in computing emission probabilities. Loman et al (2015); Schreiber and Karplus (2015); Szalay and Golovchenko (2015); Timp et al (2012) and the recent open source HMM-based basecaller Nanocall David et al (2016a). Modern basecallers use RNN base models, and we opted out using CNN instead with beam search.

We compared our model on R9 chemistry with Metrichorn (HMM based approach) and Albacore (RNN based approach). For detailed basecaller overview see the appendix A.

5 Method

Instead of opting for the traditional path using HMM or newly adopted RNN we tried using CNN (Convolutional neural networks) LeCun et al (1998), that is their residual version He et al (2016). For loss, we used CTC (Connectionist temporal classification) Graves et al (2006) between basecalled and the target sequence. The implementation used is open source warp-ctc Amodei et al (2015). Main computation framework is tensorflow Abadi et al (2015).

6 Data

Dataset used were E.Coli K-12 strands from Loman (????a) and Lamba basecalling[2]. In this research, models were trained on the E. Coli K-12 strand. (Should I cite someone for strand). Data was split into train and test subset, such that aligned reads map to different reference genome parts. For initial model data bootstrapping metrichorn was used, version (Ask Someone?).

The fast5 input training files were further split into smaller training blocks, consisting of fixed block size on raw signal. For each block target sequence, basecalled data is used in following way.

We're using basecalled knowledge which tells us on each raw read part which 6-mer were currently in the nanopore. Using this data, we get the intermediate target sequence for each block. To correct for model errors, we use aligned information, that is reference genome and alignment data (cigar string) to correct that information, yielding finalized target sequence. For further performance, we skip first and last training block, since most error aggregate on edges.

[2] Acquired from doc. dr. sc. Petra Korać i dr.sc. Paula Dobrinić

Adjacent matching bases are separated with surrogate one, for example, AAA -¿ AA'A for reasons described in the following section.

7 Results

7.1 Final model

Best performing model used has 270 total layers, divided into 3 90-layer blocks. Between each 90 layers blocks, MaxPool layer is inserted with the receptive width of 2 and stride 2, to ease computation effort and add precision.

Each block consists of 30 gated residual layers, each residual layer composed of 3 sequential Relu-Batch Normalization-Conv1D layers. Each convolutional layer uses the receptive width of 3 with 64 channels as output throughout the model.

7.2 Performance tables

As described in the previous section, the model was tested on E.Coli test set and compared to open source Nanonet and Albacore. Albacore doesn't have specific R9 chemistry mode, thus R9.4 was used instead which explains its lower performance on this task. Mean CIGAR operation are in table1 and KDE Match rate plot is in figure 7.2.

	Deletion rate	Error rate	Insertion rate	Match rate	Mismatch rate	Read length
albacore	0.060	0.194	0.070	0.867	0.063	9843
nanonet	0.088	0.190	0.040	0.897	0.062	5029
mincall_m270	0.077	0.172	0.040	0.905	0.056	9378

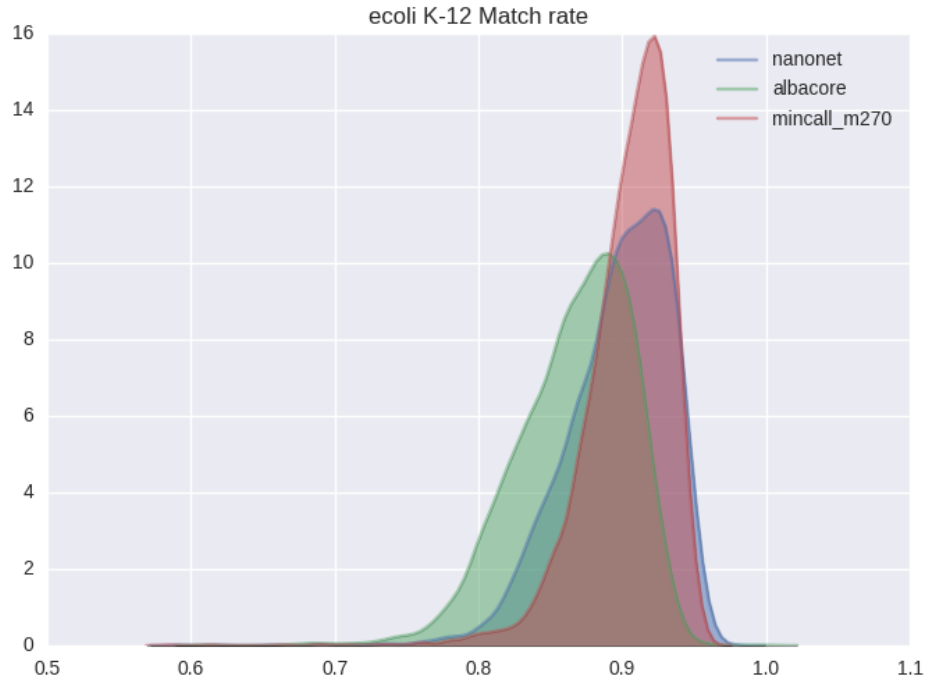
Table 1. Mean performance metrics on E.Coli dataset, 5k sample

	Base pairs/second
albacore	38000
nanonet	Slow
mincall_m270	3774

Table 2. Speed

8 Conclusion and further work

This model used advance state-of-the-art gated residual convolutional neural network, with top models having 270 layers and over 3M parameters, yet improvements over Metrichorn baseline are marginal. As the conclusion, it might be that we've reached Bayesian error rate for R9 chemistry. Furthermore, R9.5 and 1D$\hat{2}$ reads are under development which shall yield this paper's result obsolete quite soon, yet underlying code developed could easily be adjusted and trained on new data.



Unlike Nanonet which uses custom OpenCL kernels or Albacore — a novel ONT basecaller as of May 2017 lacking GPU support, this work used world-class computational framework tensorflow with highly optimized kernels and large development community. Therefore resulting paper's effect is showcasing Residual CNN approach or pure CNN approach with CTC loss is marginally better than already established basecaller and providing code in the contemporary framework.

9 Acknowledgments

First and foremost I'd like to thank my mentor izv. prof. dr. sc. Mile Sikić for setting up the problem, guiding us through its beginnings and providing helpful advice for its completion. Next Marko Ratkovic for his not only insightful conversations, but also meaningful code contributions.

Finally, I'm thanking various other people whose code, tools and advice I've used in completing this paper: Fran Jurišić, Ana Marija Selak, Ivan Sović and Martin Šošić.

Also some data were obtained in cooperation with Biologists doc. dr. sc. Petra Korać and dr.sc. Paula Dobrinić. Other were procured publicly shared from Loman labs Loman (????a)

Bibliography

Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M, Ghemawat S, Goodfellow I, Harp A, Irving G, Isard M, Jia Y, Jozefowicz R, Kaiser L, Kudlur M, Levenberg J, Mané D, Monga R, Moore S, Murray D, Olah C, Schuster M, Shlens J, Steiner B, Sutskever I, Talwar K, Tucker P, Vanhoucke V, Vasudevan V, Viégas F, Vinyals O, Warden P, Wattenberg M, Wicke M, Yu Y, Zheng X (2015) TensorFlow: Large-scale machine learning on heterogeneous systems. URL `http://tensorflow.org/`, software available from tensorflow.org

Amodei D, Anubhai R, Battenberg E, Case C, Casper J, Catanzaro B, Chen J, Chrzanowski M, Coates A, Diamos G, Elsen E, Engel J, Fan L, Fougner C, Han T, Hannun AY, Jun B, LeGresley P, Lin L, Narang S, Ng AY, Ozair S, Prenger R, Raiman J, Satheesh S, Seetapun D, Sengupta S, Wang Y, Wang Z, Wang C, Xiao B, Yogatama D, Zhan J, Zhu Z (2015) Deep speech 2: End-to-end speech recognition in english and mandarin. CoRR abs/1512.02595, URL `http://arxiv.org/abs/1512.02595`

Boža V, Brejová B, Vinař T (2017) DeepNano: Deep recurrent neural networks for base calling in MinION nanopore reads. PLOS ONE 12(6):e0178,751, DOI 10.1371/journal.pone.0178751, URL `https://doi.org/10.1371/journal.pone.0178751`

Community N (????) Basecalling overview. URL `https://community.nanoporetech.com/technical_documents/data-analysis/v/datd_5000_v1_reve_22aug2016/basecalling-overvi`, [Accessed; 12-July-2017]

David M, Dursi LJ, Yao D, Boutros PC, Simpson JT (2016a) Nanocall: an open source basecaller for oxford nanopore sequencing data. Bioinformatics p btw569

David M, Dursi LJ, Yao D, Boutros PC, Simpson JT (2016b) Nanocall: An open source basecaller for oxford nanopore sequencing data. bioRxiv DOI 10.1101/046086, URL `http://biorxiv.org/content/early/2016/03/28/046086`, `http://biorxiv.org/content/early/2016/03/28/046086.full.pdf`

Graves A, Fernández S, Gomez F, Schmidhuber J (2006) Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In: Proceedings of the 23rd international conference on Machine learning, ACM, pp 369–376

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 770–778

Jain M, Koren S, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Dilthey AT, Fiddes IT, Malla S, et al (2017) Nanopore sequencing and assembly of a human genome with ultra-long reads. bioRxiv p 128835

Judge K, Harris SR, Reuter S, Parkhill J, Peacock SJ (2015) Early insights into the potential of the oxford nanopore minion for the detection of antimicrobial resistance genes. Journal of Antimicrobial Chemotherapy 70(10):2775–2778

Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ (eds) Advances in Neural Information Processing Systems 25, Curran Associates, Inc., pp 1097–1105, URL `https://goo.gl/UpFBv8`

LeCun Y, Bengio Y (1998) The handbook of brain theory and neural networks. MIT Press, Cambridge, MA, USA, chap Convolutional Networks for Images, Speech, and Time Series, pp 255–258, URL `http://dl.acm.org/citation.cfm?id=303568.303704`

LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proceedings of the IEEE 86(11):2278–2324

Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. arXiv preprint arXiv:13033997

Loman N (????a) Nanopore R9 rapid run data release. URL `http://lab.loman.net/2016/07/30/nanopore-r9-data-release/`, [Online; posted 30-July-2016]

Loman N (????b) Thar she blows! Ultra long read method for nanopore sequencing. URL `http://lab.loman.net/2017/03/09/ultrareads-for-nanopore/`, [Online; posted 9-March-2017]

Loman NJ, Quick J, Simpson JT (2015) A complete bacterial genome assembled de novo using only nanopore sequencing data. Nature methods 12(8):733–735

Mikheyev AS, Tin MM (2014) A first look at the oxford nanopore minion sequencer. Molecular ecology resources 14(6):1097–1102

Norris AL, Workman RE, Fan Y, Eshleman JR, Timp W (2016) Nanopore sequencing detects structural variants in cancer. Cancer biology & therapy 17(3):246–253

Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L, Bore JA, Koundouno R, Dudas G, Mikhail A, et al (2016) Real-time, portable genome sequencing for ebola surveillance. Nature 530(7589):228–232

Schreiber J, Karplus K (2015) Analysis of nanopore data using hidden markov models. Bioinformatics p btv046

Sović I, Šikić M, Wilm A, Fenlon SN, Chen S, Nagarajan N (2016) Fast and sensitive mapping of nanopore sequencing reads with graphmap. Nature communications 7

Szalay T, Golovchenko JA (2015) De novo sequencing and variant calling with nanopores using poreseq. Nature biotechnology 33(10):1087–1091

Timp W, Comer J, Aksimentiev A (2012) Dna base-calling from a nanopore using a viterbi algorithm. Biophysical journal 102(10):L37–L39

A Basecallers

Here is currated basecaller list:

A.1 Official

Metrichor is an Oxford Nanopore company that offers cloud-based platform *EPI2ME* for analysis of nanopore data. Initially, base calling was only available by uploading data to the platform - that being the reason why this basecaller is often called Metrichor even though it is a name of the company.

With the release of R9 chemistry, this model was replaced by a more accurate recurrent neural network (RNN) implementation. Currently, Oxford Nanopore offers several RNN-based local basecaller versions under different names: Albacore, Nanonet and basecaller integrated into MinKNOW Community (????).

Albacore is basecaller by Oxford Nanopore Technologies ready for production and actively supported. It is available to the Nanopore Community served as a binary. The source code of Albacore was not provided and is only available through the ONT Developer Channel. Tool supports only R9.4 and future R9.5 version of the chemistry. For R9 tests in this paper we used R9.4 chemisty setting as instructed on ONT forums.

Nanonet[3] uses the same neural network that is used in Albacore but it is continually under development and does contain features such as error handling or logging needed for production use. It uses *CURRENNT* library for running neural networks. It supportes basecalling of both

[3] `https://github.com/nanoporetech/nanonet/`

R9 and R9.4 chemistry versions. However in our experiments it was painfully slow, which was as expected due to it's classification as not production ready.

Scrappie[4] is another basecaller by Oxford Nanopore Technologies. Similar to Nanonet, it is the platform for ongoing development. Scrappie is reported to be the first basecaller that specifically address homopolymer base calling. It became publicly available just recently in June, 2017 and supports R9.4 and future R9.5 data.

A.2 Third-party basecallers

Nanocall David et al (2016b) was the first third-party open source basecaller for nanopore data. It uses HMM approach like the original R7 Metrichor. Nanocall does not support newer chemistries after R7.3.

DeepNano Boža et al (2017) was the first open-source basecaller based on neural networks. It uses bidirectional recurrent neural networks implemented in Python, using the Theano library. When released, originally only supported R7 chemistry, but support for R9 and R9.4 was added recently.

[4] `https://github.com/nanoporetech/scrappie`