

SHORT REPORT

Open Access



Pair consensus decoding improves accuracy of neural network basecallers for nanopore sequencing

Jordi Silvestre-Ryan* and Ian Holmes*

*Correspondence:
jordisr@berkeley.edu;
ihh@berkeley.edu
Department of Bioengineering,
University of California, 94720
Berkeley, USA

Abstract

We develop a general computational approach for improving the accuracy of basecalling with Oxford Nanopore's 1D² and related sequencing protocols. Our software PoreOver (<https://github.com/jordisr/poreover>) finds the consensus of two neural networks by aligning their probability profiles, and is compatible with multiple nanopore basecallers. When applied to the recently-released Bonito basecaller, our method reduces the median sequencing error by more than half.

Main text

Nanopore sequencers, such as the MinION and related devices from Oxford Nanopore Technologies (ONT), allow for direct readout of individual DNA molecules [1]. However, the higher error rate of nanopore sequencing compared to other methods has limited its application in situations where deep coverage is unavailable, such as detection of rare variants or characterization of highly polymorphic samples. In principle, 2X coverage is available even for single duplexes, using ONT's 1D² protocol or related methods which sequence both strands of the duplex consecutively. In the 1D² protocol, special DNA adapters are used such that after the template DNA strand passes through the pore, its complementary strand very often follows. Combining the readout of both strands should improve accuracy; however, most neural network basecaller architectures are designed to operate on single strands. Here we present a general method for adapting existing basecallers to take advantage of the extra information in paired 1D² reads.

Nanopore sequencing works by threading a single strand of DNA through a protein nanopore embedded in a synthetic membrane. The DNA bases block the pore, perturbing the ionic current flowing through. The current can be measured, and the original sequence of nucleotides recovered computationally. This latter *basecalling* step makes heavy use of machine learning techniques and, increasingly, of neural networks.



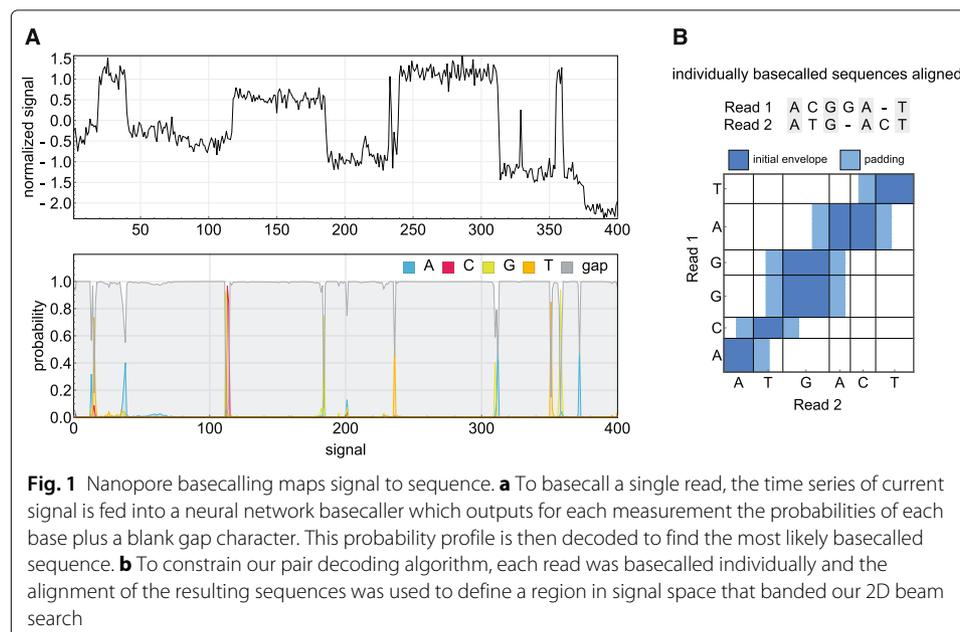
© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Early neural network basecallers (such as DeepNano [2], BasecRAWler [3], and certain ONT-developed basecallers) relied on a preprocessing step that segmented the current measurements into discrete events, corresponding to individual nucleotides passing through the pore. This aspect of basecalling shares similarities with speech recognition, where an audio time series must be segmented and then labeled with phonemes. Inspired by this similarity, later basecallers used Connectionist Temporal Classification (CTC), a method developed for speech recognition, which trains neural networks to do segmenting and classification simultaneously [4]. The community basecaller Chiron [5] successfully applied CTC to nanopore basecalling [6], while ONT incorporated CTC-style models into both production and research basecallers.

A CTC-trained neural network outputs a probability profile (Fig. 1a) defining a distribution $P(\ell|y)$ over possible basecalled sequences ℓ given the read y . By analogy to hidden Markov models, the task of finding the modal sequence of this distribution is termed “decoding”. While perfectly optimal decoding requires an intractably exhaustive search over sequences, heuristic algorithms (such as beam search or Viterbi search) can in practice be used to find reasonably good solutions.

The related task of “consensus decoding” arises when multiple reads $\{y_n\}$ are derived from the same underlying sequence ℓ , as is the case for 1D². Basecalling then yields multiple profiles $P(\ell|y_n)$. Our task is to find the single sequence that maximizes $P(\ell|\{y_n\})$; under a flat prior $P(\ell)$ and the assumption that the reads are independent, this will be the sequence that maximizes the product $\prod_n P(\ell|y_n)$, motivating the reframing of this problem as an exercise in profile-profile alignment [7].

To this end we have developed a beam search decoding algorithm for the pair decoding of two reads, making use of a constrained dynamic programming heuristic to speed calculations by focusing on areas of each read which are likely to represent the same sequence (full details provided in Additional file 1). We introduce our basecalling software PoreOver, which implements these decoding algorithms and



includes a basic recurrent neural network basecaller (PoreOverNet) for demonstration purposes.

DNA flows through the pore at an average of 450 bases/second; the electrical signal is recorded at 4000 Hz, yielding 9 measurements/base on average. Thus, if a read represents T bases, aligning two basecalled reads will take $\sim T^2$ steps, but aligning the raw signal measurements will take $\sim (9T)^2$ steps—an 81-fold increase compared to aligning basecalled sequences. To accelerate calculations we constrain our heuristic search to an “alignment envelope” containing the timepoints where the reads are most likely to align [8].

This envelope is estimated by doing a preliminary Viterbi decoding step on each read individually, then aligning the two sequences so obtained. This is faster than beam search, with similar performance (see Additional file 1), and explicitly maps each nucleotide to some range of timepoints. The two decoded sequences are then aligned globally, generating a nucleotide-level mapping between the reads, and (by extension) between the underlying time series. With some additional padding, this guide alignment defines the envelope for our banded 2D beam search (Fig. 1b).

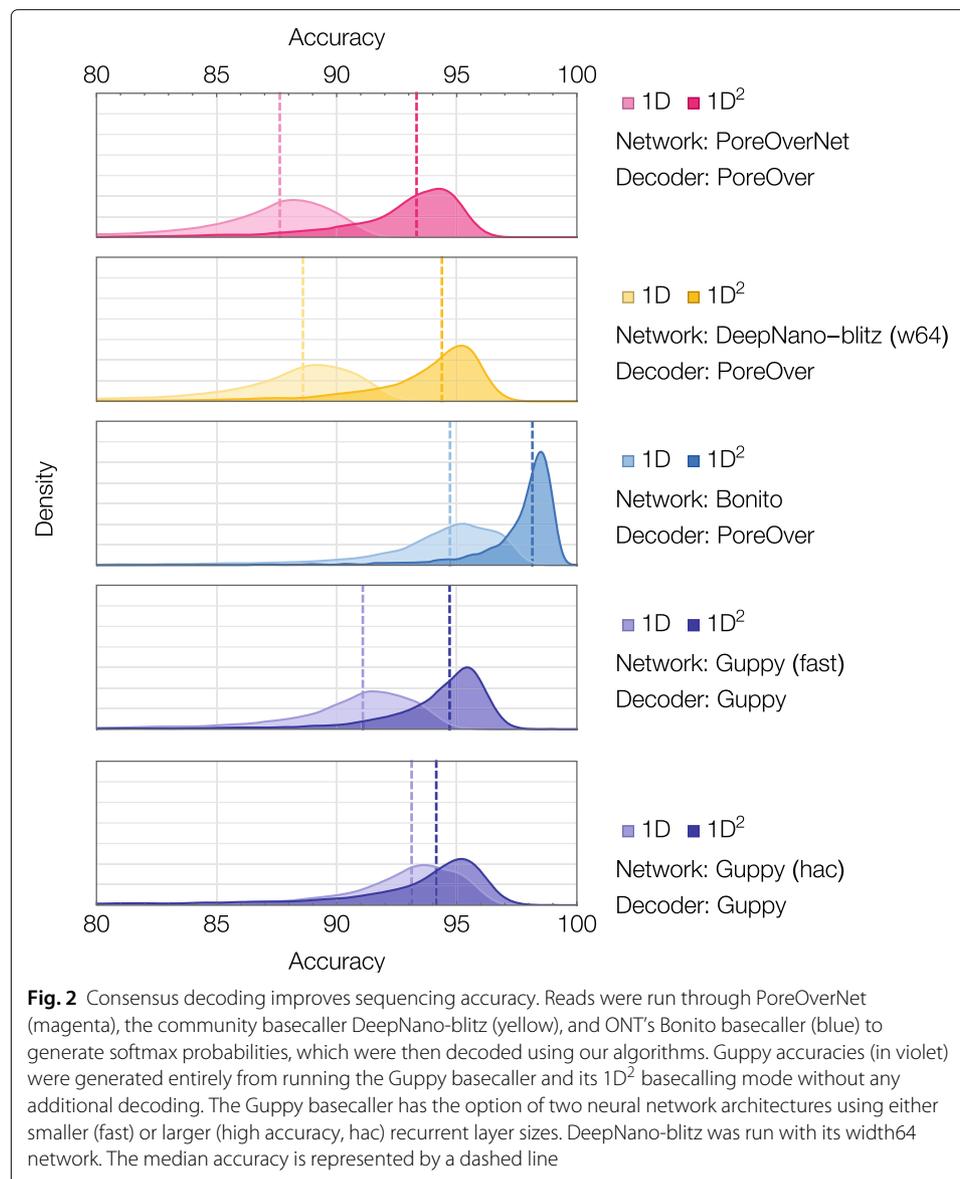
As nanopore reads can vary in length over orders of magnitude, a naive Needleman-Wunsch alignment may involve creating infeasibly large dynamic programming matrices. As a workaround, we use a modified Needleman-Wunsch with a fixed diagonal band. This appears to be sufficient for subsequent pair decoding, though exploiting recent advances in efficient pairwise alignment algorithms (such as [9]), may yield further improvements in accuracy and speed.

We tested our pair decoding algorithm on a sample of 5,000 R9.4 *E. coli* 1D² read pairs (Oxford Nanopore Technologies, personal communication), comprising 10,000 reads in total. Reads were run through a forward pass of our PoreOverNet basecaller to generate softmax probabilities, which were used for subsequent pair decoding.

After pair decoding, reads were aligned to the reference *E. coli* genome with Minimap [10] and the read accuracy calculated as (number of matches)/(length of alignment). We find that our banded 2D beam search improves the median accuracy from 87.6% for single reads to 93.2% for 1D² read pairs (Fig. 2), nearly halving the error rate of our PoreOverNet basecaller.

Our software can readily be adapted to work with the output of other neural network basecallers. Application to the recent DeepNano-blitz [11], showed a similar gain in accuracy from consensus decoding. We also applied our algorithm to the ONT basecaller Bonito [12], a research basecaller inspired by recent successes of purely convolutional neural networks in speech recognition, and compared results with Guppy, an earlier ONT basecaller which can make use of 1D². Our consensus method lifts Bonito’s median accuracy from 94.7% to 98.1%, better than halving the median error rate for single read basecalling and surpassing the consensus accuracy of Guppy’s 1D² method (Fig. 2). Unlike Guppy, our code is open source; further, it is modular in design, making it straightforwardly modifiable and re-usable for other basecallers. We thus envision the PoreOver as a consensus decoding tool to be used in concert with a state-of-the-art CTC basecaller such as Bonito. Since initial submission of this paper, the Bonito basecaller now includes an implementation of our pair decoding algorithm (as of version 0.2.0, [12]).

Generalizing beyond a pair of reads, consensus approaches are relevant to *polishing*, the task of refining a draft genome assembly by realigning reads to the draft. There are several



approaches to polishing via multi-read consensus: some analyze the raw current signal using a hidden Markov Model [13] or dynamic time warping [14], while others analyze the basecalled sequence using neural networks [15, 16]. To our knowledge none of the neural network methods explicitly use the intermediate basecaller probabilities (instead relying on previously basecalled sequence), while the methods that do use the raw signal do not use neural networks. The pairwise dynamic programming approach we describe could be extended to multiple reads, although the curse of dimensionality (a full dynamic programming alignment of N reads takes $\mathcal{O}(T^N)$ steps) would necessitate additional heuristics to narrow down the search space. These could include generalizing alignment envelopes to multiple sequences, or performing a stochastic search. With such heuristics, it should be possible to implement an algorithm to exploit the basecaller probabilities for general, multi-read consensus [7].

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13059-020-02255-1>.

Additional file 1: Supplementary text and figures.

Additional file 2: Review history.

Peer review information

Andrew Cosgrove was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Acknowledgements

We thank Tim Massingham and Marcus Stoiber (Oxford Nanopore Technologies) for helpful discussion, and the anonymous reviewers for their feedback and suggestions. This work used the computational cluster provided by the Berkeley Research Computing program.

Review history

The review history is available as Additional file 2.

Authors' contributions

JSR developed the software and conducted the benchmark analysis. JSR and IHH wrote the manuscript. Both authors read and approved the final manuscript.

Funding

The authors were supported by NIH/NCI grant CA220441, NIH/NHGRI training grant T32 HG000047, and by a research gift from Oxford Nanopore Technologies.

Availability of data and materials

Our software PoreOver [17] is available at <https://github.com/jordisr/poreover> under an MIT license. The *E. coli* 1D² reads used to test our pair decoding algorithm were generated by Oxford Nanopore Technologies and are available at [18].

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors received research funding (IHH) and travel reimbursement (JSR) from Oxford Nanopore Technologies.

Received: 15 April 2020 Accepted: 20 December 2020

Published online: 19 January 2021

References

1. Deamer D, Akeson M, Branton D. Three decades of nanopore sequencing. *Nat Biotechnol*. 2016;34(5):518.
2. Boža V, Břejová B, Vinař T. DeepNano: Deep recurrent neural networks for base calling in MinION nanopore reads. *PLoS ONE*. 2017;12(6):1–13. <https://doi.org/10.1371/journal.pone.0178751>.
3. Stoiber M, Brown J. BasecRAWler: streaming nanopore basecalling directly from raw signal. *bioRxiv*. 2017;133058. <https://doi.org/10.1101/133058>.
4. Graves A, Fernández S, Gomez F, Schmidhuber J. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In: Proceedings of the 23rd International Conference on Machine Learning, ICML '06. New York, NY, USA: ACM; 2006. p. 369–76. <https://doi.org/10.1145/1143844.1143891>.
5. Teng H, Cao MD, Hall MB, Duarte T, Wang S, Coin LJM. Chiron: translating nanopore raw signal directly into nucleotide sequence using deep learning. *GigaScience*. 2018;7(5):037. <https://doi.org/10.1093/gigascience/giy037>.
6. Wick RR, Judd LM, Holt KE. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biol*. 2019;20(1):129. <https://doi.org/10.1186/s13059-019-1727-y>.
7. Silvestre-Ryan J, Holmes I. Consensus decoding of recurrent neural network basecallers. In: Jansson J, Martín-Vide C, Vega-Rodríguez MA, editors. Algorithms for Computational Biology. Cham: Springer; 2018. p. 128–39.
8. Holmes I, Durbin R. Dynamic programming alignment accuracy. *J Comput Biol*. 1998;5(3):493–504.
9. Marco-Sola S, Moure JC, Moreto M, Espinosa A. Fast gap-affine pairwise alignment using the wavefront algorithm. *Bioinformatics*. 2020;1–8. <https://doi.org/10.1093/bioinformatics/btaa777>.
10. Li H. Minimap and miniasm: Fast mapping and de novo assembly for noisy long sequences. *Bioinformatics*. 2016;32(14):2103–10. <http://arxiv.org/abs/1512.01801>.
11. Boža V, Perešini P, Břejová B, Vinař T. DeepNano-blitz: a fast base caller for MinION nanopore sequencers. *Bioinformatics* (Oxford, England). 2020;36(14):4191–2. <https://doi.org/10.1093/bioinformatics/btaa297>.
12. Oxford Nanopore Technologies. Bonito. <https://github.com/nanoporetech/bonito>. Accessed Sept 2020.
13. Loman NJ, Quick J, Simpson JT. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat Methods*. 2015;12(8):733.
14. Chan RSL, Gordon P, Smith MR. Evaluation of dynamic time warp barycenter averaging (DBA) for its potential in generating a consensus nanopore signal for genetic and epigenetic sequences, vol. 2018-July. In: Proceedings of

the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS. New York: IEEE; 2018. p. 2821–4. <https://doi.org/10.1109/EMBC.2018.8512873>.

15. Shafin K, Pesout T, Lorig-Roach R, et al. Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nat Biotechnol.* 2020;38:1044–53.
16. Oxford Nanopore Technologies. Medaka. <https://github.com/nanoporetech/medaka>. Accessed Sept 2020.
17. Silvestre-Ryan J. PoreOver v1.0.0. 2020. <https://doi.org/10.6084/m9.figshare.13431101.v1>. Accessed Dec 2020.
18. Silvestre-Ryan J. E. coli 1D2 nanopore sequencing reads. 2020. <https://doi.org/10.6084/m9.figshare.13415867.v1>. Accessed Dec 2020.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

