Quantitative Biology 2018, 6(4): 359–368
https://doi.org/10.1007/s40484-018-0155-4

METHODOLOGY ARTICLE

WaveNano: a signal-level nanopore base-caller via simultaneous prediction of nucleotide labels and move labels through bi-directional WaveNets

Sheng Wang[1,†,*], Zhen Li[2,3,†], Yizhou Yu[2] and Xin Gao[1,*]

[1] Computational Bioscience Research Center (CBRC), King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Kingdom of Saudi Arabia
[2] Department of Computer Science, University of Hong Kong, Hong Kong SAR 999077, China
[3] School of Science and Engineering, Shenzhen Research Institute of Big Data, The Chinese University of Hong Kong, Shenzhen 518172, China
* Correspondence: realbigws@gmail.com, xin.gao@kaust.edu.sa

Received March 2, 2018; Revised March 28, 2018; Accepted March 28, 2018

***Background*: The Oxford MinION nanopore sequencer is the recently appealing third-generation genome sequencing device that is portable and no larger than a cellphone. Despite the benefits of MinION to sequence ultra-long reads in real-time, the high error rate of the existing base-calling methods, especially indels (insertions and deletions), prevents its use in a variety of applications.**

***Methods*: In this paper, we show that such indel errors are largely due to the segmentation process on the input electrical current signal from MinION. All existing methods conduct segmentation and nucleotide label prediction in a sequential manner, in which the errors accumulated in the first step will irreversibly influence the final base-calling. We further show that the indel issue can be significantly reduced via accurate labeling of nucleotide and move labels directly from the raw signal, which can then be efficiently learned by a bi-directional WaveNet model simultaneously through feature sharing. Our bi-directional WaveNet model with residual blocks and skip connections is able to capture the extremely long dependency in the raw signal. Taking the predicted move as the segmentation guidance, we employ the Viterbi decoding to obtain the final base-calling results from the smoothed nucleotide probability matrix.**

***Results*: Our proposed base-caller, WaveNano, achieves good performance on real MinION sequencing data from Lambda phage.**

***Conclusions*: The signal-level nanopore base-caller WaveNano can obtain higher base-calling accuracy, and generate fewer insertions/deletions in the base-called sequences.**

Keywords: nanopore sequencing; bi-directional WaveNets; base-calling; third generation sequencing; deep learning

Author summary: Oxford nanopore sequencing is a rapidly developed sequencing technology in recent years. Despite the benefits of this technique to sequence ultra-long reads in real-time, the high error rate of the existing base-calling methods, especially indels (insertions and deletions), prevents its use in many applications. Here we show that such indel errors are largely due to the segmentation process on the input electrical current signals, and propose a new deep learning model bi-directional WaveNet to perform the base-calling directly on the signal level. The experimental result suggests that our method achieves good performance on real nanopore sequencing data from Lambda phage.

[†] These authors contributed equally to this work.

INTRODUCTION

Over the last decade, high-throughput second-generation sequencing technology has revolutionized genomic research with the ability to sequence the whole genome of a variety of organisms on earth [1]. In 2014, Oxford Nanopore Technologies (ONT) released a third-generation sequencing platform, MinION, which is a portable, single-molecule genomic sequencing device no larger than an iPhone [2] (see Figure 1A). There are two key features of this device, long-reads and point-of-care [3].

MinION directly senses native, individual DNA single-strand without the need for polymerase chain reaction (PCR) amplification, which enables the device to sequence extremely long reads (typically from 12k to 120k bp, or even longer) of DNA without a reduction in the sequence quality [4]. This allows users to generate reads spanning most repetitive sequences, which most second-generation sequencing technologies based on short reads (typically from 75 to 150 bp) cannot unambiguously resolve [1].

MinION can be used for sequencing immediately at anywhere in real-time, as it is portable and does not require any special setup or calibration procedures [5]. It was reported that MinION has been used for diagnostic investigation during the Ebola outbreak in Guinea, west Africa [6]. Another report indicated that MinION has been tested on the International Space Station (ISS) for real-time sequencing of a Lambda phage and the results showed no difference in performance on the ISS and on earth [7].

The key innovation of the MinION sequencer is the direct measurement of the changes in the electrical current signal (denoted as raw signal) when a single-strand DNA passes through the nanopore [8]. In MinION, a few hundred of nanopores are implanted in a voltage-biased membrane. Single-strand DNA sequences pass through these pores. At each time point, there are five consecutive nucleotides in a pore (denoted as a 5-mer). The electrical current signal is measured for each time point of the pore. The underlying assumption is that with different 5-mers in the pore, the electrical current will be different. The goal is to decode the time-course electrical current signals into the sequence of nucleotides. This procedure is referred to as base-calling (Figure 1B). However, the frequency of the electrical current measurements and the speed of the DNA sequence passing through the pore are not coordinated, which causes the main technical difficulty for base-calling. In general, the frequency of the electrical current measurements is 7–9 times higher than the passing speed of the DNA sequence. That is, each 5-mer is on average measured by 8–10 times, yet the variance of measured times per 5-mer is very high as the passing speed is inconsistent. Here, a pore model is defined as the correspondence between the expected current signal and the 5-mer inside the pore at the same time point. Thus, given a pore model, we may annotate 5-mer label and move label upon the raw signal. Specifically, a 5-mer label indicates that a certain time point of the raw signal belongs to which 5-mer, whereas a move label indicates whether the 5-mer *stay*s or *move*s for the next time point of the signal (Figure 1B).

The most problematic issue in nanopore sequencing is the high sequencing error rate, especially in indels (insertions and deletions) [9]. To solve this issue, a variety of approaches have been proposed and they can be

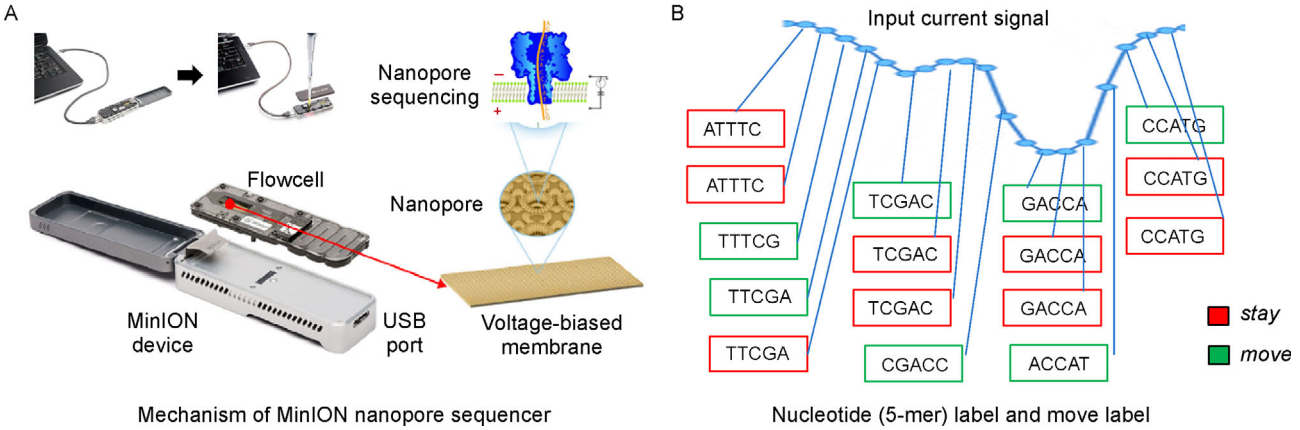

Figure 1. Mechanism of MinION nanopore sequencer and the two signal-level labels. (A) The MinION nanopore-based sequencing device. Sequencing is performed by adding the DNA sample to the flowcell. (B) When DNA molecules pass through the nanopore, each five-tuple DNA (denoted as 5-mer, *e.g.,* "ATTTC") inside the pore will cause a change in the magnitude of the current. Such correspondence between 5-mers and the expected current signal is denoted as the pore model. For consecutive time points, the 5-mer either *stay*s in the pore (shown in red) or *move*s by one nucleotide (shown in green). Such stay/move label can serve as the segmentation guidance to convert the current signal to a 5-mer event sequence.

roughly categorized into two groups: (i) machine learning based and (ii) consensus based. The former group relies on a machine learning model that learns the mapping between the input nanopore current signals (with length L_1) and the corresponding DNA sequence (with length L_3), *i.e.*, base-calling. The latter group not only relies on the base-called reads (here, a read is a decoded DNA sequence), but also requires additional resources, such as read mapping on a given reference genome [4], error correction using short reads from second-generation sequencing [10], or the overlapping base-called reads from the same nanopore experiments (such as Nanopolish [2,11] and PoreSeq [12]). In either case, a more accurate base-calling method is needed to facilitate the power and advantages of the nanopore sequencing. Here we focus on solving two key issues of the existing machine learning-based base-calling approaches: (i) serial base-calling, and (ii) model architecture.

Serial base-calling. In almost all existing methods [13,14], the base-calling process is divided into two serial steps: segmentation and decoding (see Figure 2A). The segmentation step takes the discrete time-course electrical current signals as input and outputs a segmented stepwise curve. This step can be considered as a clustering step, which tries to group current signals for the same 5-mers together. The decoding step then takes the segmented curve as input and decodes it to the nucleotide sequence. However, since the segmentation step is based on local signal information only and does not take global context into account, insertion (*i.e.*, one 5-mer is divided into multiple segments) and deletion (multiple 5-mers are merged to the same segment) issues commonly exist due to the noisy nature of the current signal measurement [13]. Such indel issues not only affect the event-level features, but also irreversibly harm the final base-calling performance.

Model architecture. Another key issue is that the machine learning model architectures in existing methods are not suitable for the ultra-long sequence of the nanopore current signals [15]. In brief, the first generation of base-calling methods is based on hidden Markov models (HMMs) [13]. The limitation of HMMs lies in the short-range dependency. To overcome this issue, the second generation of base-callers is established on recurrent neural network (RNN) architectures, such as long short-term memory (LSTM) [16] or gated recurrent unit (GRU) [17]. Thanks to a large hidden state space, the RNN base-callers can potentially capture long-distance dependencies in the input signal [14]. However, since the length of the nanopore signal is extremely long (about 100k bp on average), traditional RNN-based architectures still cannot handle these long sequences properly [15].

To further improve supervised learning methods for base-calling, it is possible for us to borrow ideas from very recent technology breakthrough in speech recognition via deep learning [15]. Deep learning [18] is a powerful machine learning technique that has revolutionized image classification [19], natural language processing [20], and bioinformatics [21]. Recently, a deep learning architecture WaveNets [15] demonstrated superior performance in speech generation. If we consider the nanopore signal as a speech signal, then base-calling is kind of similar to speech recognition. Thus, the recent developed WaveNet architecture might also work for base-calling.

In this paper, we propose a novel method, WaveNano, which jumps over the segmentation step and directly conducts 5-mer label and move label prediction simultaneously from the electrical current signals. The main contributions of this paper are as follows:

- We develop a method to accurately label the ground-truth 5-mer label and move label for each time point of the raw electrical current signal using dynamic time warping [8], which provides supervised training data for our base-calling method.
- We propose a novel model to simultaneously predict the 5-mer label and the move label for each time point via bi-directional WaveNets [15] with stacked residual blocks of convolutional neural network (shown in Figure 3).
- We show that using predicted move labels as the segmentation guidance can help solve the indel issue substantially. Experimental results on the Lambda phage demonstrate that our proposed method outperforms the existing methods and achieves good accuracy.

RESULTS

Lambda phage data preparation

We sequenced the genome of a Lambda phage, which is provided by ONT for calibration of the nanopore sequencers, based on the one-dimensional (1D) protocol. The reference DNA sequence of the Lambda phage is made available by ONT. Before sequencing, the Lambda phage DNA library preparation was performed using the genomic DNA sequencing kit (Oxford Nanopore). According to the manufacturers' instructions, we used the SQK-MAP-003 sequencing kit for R9.4 MinION flow cells. A new MinION flow cell was used for each sequencing run. The library was loaded onto the MinION flow cell and the genomic DNA 48-hour sequencing protocol was initiated using the MinKNOW software. We obtained around 24,000 reads, with an average length of 63,000 bp for electrical current signals.

Evaluation metrics

To evaluate the performance of our base-caller Wave-

Nano, we adopted a measurement similar to the one used in BLAST [22] to compare the similarity between the reference DNA sequence S_{ref} with length L_{ref} and the base-called sequence S_{call} with length L_{call}. The scoring function is 1 for the same nucleotide, -2 for mismatched nucleotides, -2 for gap open, and -1 for gap extension. Based on such a scoring function, the optimal alignment path could be obtained by dynamic programming [23]. Along the path, we can count the number of exact matches (denoted as M), the number of gaps (denoted as G), and the length of alignment (denoted as L_{ali}).

We used four evaluation metrics below to assess the quality of the base-called sequence: (i) sequence identity with respect to S_{ref} (defined as $\text{SeqID}_1 = M/L_{\text{ref}}$), (ii) sequence identity with respect to S_{call} (defined as $\text{SeqID}_2 = M/L_{\text{call}}$), (iii) sequence identity (defined as $\text{SeqID} = M/L_{\text{ali}}$), and (iv) the gap ratio (defined as $G_{\text{rate}} = G/L_{\text{ali}}$). Generally speaking, a higher value of the sequence identity and a lower value of the gap ratio indicate the high similarity between the base-called sequence and the reference sequence.

In addition, as shown in Equation (4), besides the optimization for the loss function of move label, we also try to optimize the AUC score of move label directly in the training process. Although we acknowledge that area under the precision-recall curve (AUPRC) is a better score to evaluate classification performance for imbalanced problem, currently it is very challenging to develop a simple approach to directly optimize AUPRC for a linear-structured data, such as sequence labeling problem here [24]. In this work, we will compare the AUC and AUPRC for move label prediction.

Compared methods

We compared WaveNano with the state-of-the-art and official base-calling tool Metrichor (https://metrichor.com/). Metrichor was initially trained on a hidden Markov model and the average base-calling accuracy in terms of SeqID is around 70% [13]. With the latest release of ONT R9, it was reported that Metrichor has evolved the base-calling algorithm to a bi-directional LSTM model [25] and the base-calling accuracy is boosted to close to 90% [9]. However, it should be noted that Metrichor is a cloud-based platform and the source code is not open to public. Recently, ONT released Albacore, which is a binary-only tool for offline base-calling. It was reported that these official tools have similar base-calling accuracy [14].

Comparison results

We trained WaveNano and conducted the evaluation on Lambda phage. The performance was obtained through five-fold cross validation on about 24,000 reads of Lambda phage. Experimental results show that WaveNano achieves better performance than Metrichor and Albacore, under all four measurements, SeqID_1, SeqID_2, SeqID, G_{rate} (Table 1). Specifically, WaveNano achieves 0.632 and 0.947 accuracy for the 5-mer label and move label prediction, respectively. Through Viterbi decoding with the segmentation guidance, the final base-calling obtains 0.956, 0.932, and 0.923 sequence identity, and 0.056 gap ratio, which implies base-calling from WaveNano not only predicts more accurate DNA sequences, but also solves the indel issue significantly better than the official base-callers, Metrichor and Albacore.

Table 1 also shows that the accuracy of Metrichor is similar to that of Albacore, especially in terms of SeqID and G_{rate}. In addition, Albabore has a higher SeqID_1 and a lower SeqID_2 than Metrichor, indicating that the DNA sequence predicted by Albacore has more matches to the ground-truth sequence than that predicted by Metrichor, but the sequence length predicted by Metrichor is shorter than that predicted by Albacore.

We further studied the importance of the move label prediction as the segmentation guidance and the bi-directional WaveNets, by removing each of them from WaveNano and evaluated the performance. Experimental results show that both the move guidance and the bi-directional WaveNets are important components in the success of WaveNano as removing each of them results in a significantly dropped 5-mer label accuracy to 0.564 and 0.597, respectively. Furthermore, for the final base-calling results, the gap ratio increases to 12.8% and 7.1%, respectively. Among the two components, the move label

Table 1 Base-calling performance on the Lambda phage genome (48.5 Kb)

	5-mer prediction	Move prediction	SeqID_1	SeqID_2	SeqID	G_{rate}
Metrichor	/	/	86.1	91.6	85.2	8.8
Albacore	/	/	87.8	88.4	85.9	8.2
WaveNano	**63.2**	**94.7**	**95.6**	**93.2**	**92.3**	**5.6**
w/o bi-WaveNet	59.7	92.2	93.7	91.3	89.4	7.1
w/o move guidance	56.4	/	91.1	84.6	83.3	12.8

"5-mer prediction" refers to the 5-mer label prediction accuracy, which is a 1024-class classification problem. "Move prediction" refers to the move label prediction accuracy, which is a binary classification problem. SeqID_1, SeqID_2, SeqID, G_{rate} are defined in the Section of Evaluation Metrics and shown in percentage.

as the segmentation guidance is more important than the bi-directional WaveNets.

The AUC (AUPRC) of move label prediction for WaveNano and WaveNano without bi-WaveNet is 0.872 (0.595) and 0.863 (0.581), respectively. If we remove the AUC loss term in Equation (4), these prediction results would become 0.867 (0.586) and 0.856 (0.569), respectively.

Runtime

It is difficult to directly compare the running time of Metrichor to that of WaveNano because Metrichor is a cloud-based tool and we did not know the exact parameter setting of Metrichor. However, we could compare the running time of WaveNano with the official offline base-caller, Albacore. It took WaveNano 0.0000416 s to base-call one time point of the signal or 0.5 s for a signal sequence with 12,000 time points, whereas it costs Albacore 2 s for base-calling the same signal sequence.

DISCUSSION AND CONCLUSIONS

In this paper we proposed a novel base-caller, WaveNano, for the third-generation nanopore sequencing. We showed that our method is better than the state-of-the-arts on processing the nanopore current signal data, obtaining higher base-calling accuracy, and generating fewer insertions/deletions on the Lamba page dataset. Consequently, base-called DNA sequences by WaveNano are more accurate and contain fewer gaps than those by Metrichor and Albacore, the current cutting-edge official base-callers.

The superiority of our method is rooted largely in the machine learning model bi-direction WaveNets (Bi-WaveNets), which is a deep learning model based on the dilated residual CNN. Such architecture could be regarded as an alternative model of RNN with gates (*e.g.*, LSTM, GRU). According to our knowledge, the official base-calling algorithm from ONT (*i.e.*, Metrichor and Albacore) is based on bi-directional LSTM (Bi-LSTM). It is worth mentioning that WaveNet is more suitable for capturing ultra-long dependency due to the following two reasons: (i) WaveNets are auto-regressive and consist of stacked causal filters with dilated convolutions to allow their receptive fields to grow exponentially with respect to the depth, which is essential to capture the ultra-long range temporal dependencies in the input data; (ii) the layers of the dilated convolutions make WaveNets a much faster model than using RNN with gate units. Consequently, comparing to RNN, WaveNets can exploit ultra long-range temporal dependencies in the signal sequence in an efficient and effective way. Moreover, our proposed model Bi-WaveNets can capture both upstream and downstream information, whereas the traditional WaveNets can only capture the upstream information.

Finally, it should be noted that a bunch of new base-callers recently appeared could directly work on the raw electrical current signals. Base-calling from raw signal (without segmenting the signal into events) first appears in Albacore v2.0 at September 2017. Albacore is the official off-line basecaller whose source code is not opened to public. Users can download Albacore from the Nanopore community website, but they need an account to log in. Later on, Scrappie, also from ONT research group and the source code is open to public, supports the base-calling from raw signal. Chiron [26] is the first-appeared third-party base-caller to perform raw signal base-calling, which is also based on deep learning. From this trend, we may foresee that the nanopore signal-level processing aided by deep learning should be the future.

METHODS

Here we propose WaveNano, a novel offline base-caller for third-generation nanopore sequencing. WaveNano simultaneously infers the 5-mer label and the move label of each time point of the input electrical current signal (see Figure 2B), by using bi-directional WaveNets with residual blocks and gated activation units. Exploiting the predicted move labels as the segmentation guidance, we employ Viterbi decoding with the predicted 5-mer label probability matrix to obtain the final DNA sequence.

Problem formulation

The base-calling process can be formulated as follows. Given an input electrical current signal sequence $X = x_1, x_2, ..., x_{L_1}$ with L_1 time points, we need to decode the final DNA sequence $B = b_1, b_2, ..., b_{L_3}$ with L_3 nucleotides, where x_i is the electrical current measurement of a 5-mer (*e.g.*, "ACGTT") at time point i, and b_j is a nucleotide that can take one of the four values from $\{A,T,C,G\}$. Note that the frequency of the electrical current measurements is about 7–9 times faster than the speed in which the single-strand DNA passes through the nanopore. For consecutive time measurements x_i and x_{i+1}, the 5-mer either *stay*s in the pore or *move*s by one nucleotide. We denote this annotation as the move label sequence X_m with length L_1. Such stay/move labels can later serve as the segmentation guidance to convert the electrical current signals to a 5-mer event sequence.

Previous methods, such as [13,14], divide the base-calling process into two serial steps: segmentation and decoding (see Figure 2A). In particular, the current signal sequence is firstly fragmented to an event sequence $X' = x'_1, x'_2, ..., x'_{L_2}$ with length $L_2 (L_3 < L_2 < L_1)$ through seg-

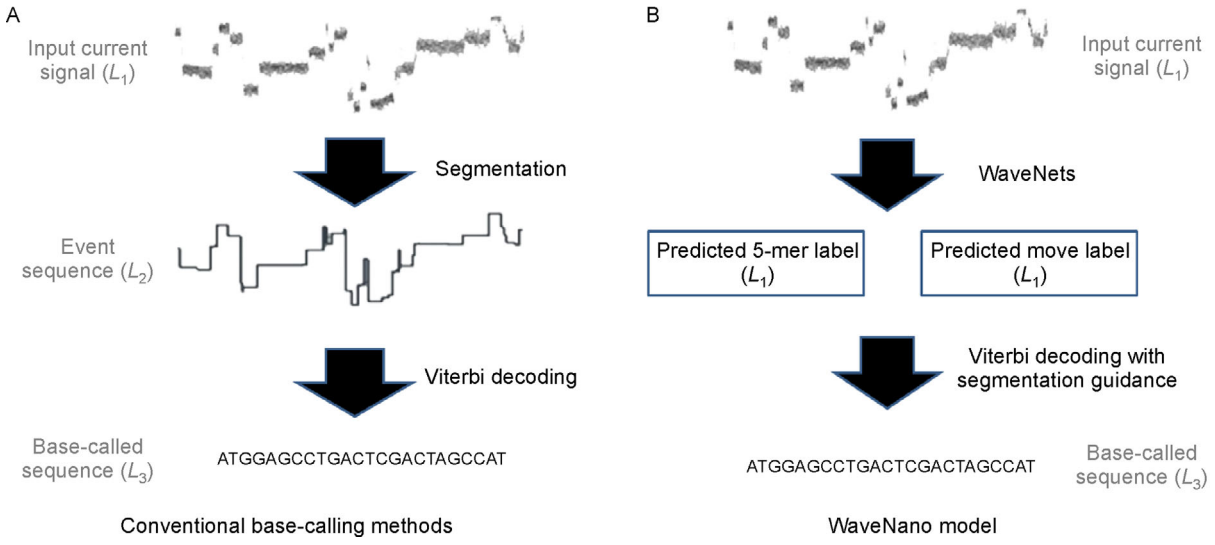

Figure 2. Comparison of the conventional base-calling methods and the WaveNano model. (A) Conventional base-calling methods conduct segmentation and decoding in a sequential manner. (B) WaveNano predicts 5-mer labels and move labels simultaneously, and then performs segmentation guided decoding.

mentation on X. Due to the noisy nature of the segmentation process, the indel issues in these methods irreversibly harm the final base-calling performance. Thus, we propose a novel approach, WaveNano, to simultaneously predict the 5-mer label y_1 (*i.e.*, which 5-mer among all possible $4^5 = 1024$ 5-mers is in the pore) and the move label y_2 (*i.e.*, whether the 5-mer moves by a nucleotide at the next time point) for each time point of the input current signal sequence X.

Signal-level ground-truth labeling

In order to train our deep learning model, we need to prepare the supervised training data first. For our method, training data refer to the ground-truth 5-mer label and move label annotation for each time point of the training electrical current signals, which are not directly available given the raw signals. We thus need to do signal-level labeling to assign the 5-mer label y_1 and move label y_2 to each time point of the signal, where the size of the label space for y_1 and y_2 is 1024 and 2, respectively.

Our original training data set contains the raw time-course electrical current signals X of length L_1 and their corresponding DNA sequence B of length L_3. However, it does not contain the corresponding 5-mer label and move label sequences, each of which should be of length L_1 as well. We first try to find an alignment between X and B.

Although it seems intractable to perform this alignment directly, we could use the ONT official pore model describing the electrical current signal that are expected to be observed for each 5-mer [13]. Given the DNA sequence B with length L_3, we can use 5-mer sliding window to generate all the L_3 5-mers (the last 4 5-mers contain less than 5 nucleotides, but are still in the pore). Each 5-mer is then converted to its expected electrical current signal value according to the ONT official pore parameters. After transforming B into the expected signal sequence S of length L_3, an optimal alignment path between the two signal sequences, X and S, could be obtained using dynamic time warping (DTW) [8].

To determine the optimal path via DTW, we recursively compute an $L_1 \times L_3$ matrix D, where the matrix entry $D(n,m)$ is the total cost of an optimal path between $X(x_1, ..., x_n)$ and $S(s_1, ..., s_m)$. Here $D(n,m) = \min\{D(n-1, m-1),\ D(n, m-1),\ D(n-1, m)\} + c(n, m)$ where c is a $L_1 \times L_3$ matrix containing distances between elements x_n in the sequence X and s_m in the sequence S. Here we use the Z-score difference to calculate $c(n,m)$. Note that recently there is a speed up approach for calculating DTW in $O(N)$ time complexity [27].

After aligning the reference DNA sequence B to the electrical current signal sequence X, it is straightforward to assign the 5-mer label $y_1(t)$ for each time point t. To assign the move label $y_2(t)$ at time point t, we just need to check whether the consecutive 5-mers at time points t and $t+1$ are the same or not. If same, then we assign *stay* at time point t, otherwise *move*.

Overall pipeline of WaveNano

The overall base-calling pipeline of WaveNano is presented in Figure 3, which takes the current signal with length L_1 as the input. Considering the large variance of the electrical current values at each time point, it is

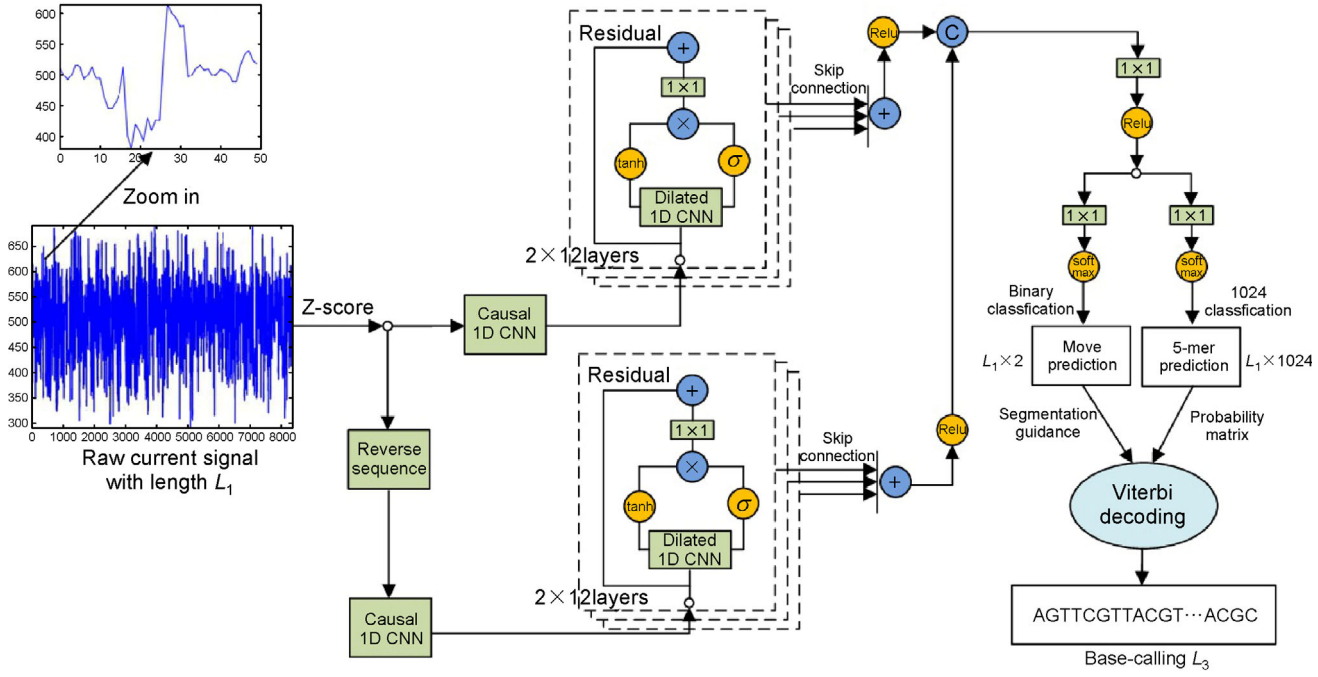

Figure 3. The overall pipeline of WaveNano. WaveNano takes the current signal X with length L_1 as the input. The calculated Z-score of the current signal feeds into the causal one-dimensional (1D) convolutional neural network (CNN) and stacked residual blocks with skip connections. Different from the original WaveNet which was proposed as a generative model, WavaNano employs bi-directional residual blocks as we can conduct base-calling with the dependency of the entire current signal. On top of the concatenated feature maps from the bi-directional blocks, 1D CNN with a 1×1 kernel is used for the 5-mer label (1024-class classification) and move label (binary classification) prediction. Using the move label prediction as the segmentation guidance, Viterbi decoding is leveraged with the smoothed 5-mer label probability matrix to obtain the final base-calling with length L_3. "C" donates concatenation.

necessary to calculate the Z-score normalization of the current signal [8]. In contrast to the existing machine learning models [13,14] which depend on the segmentation step by MinKNOW [13], WaveNano conducts training on the Z-score normalization of the original current signal with the ground-truth 5-mer label and move label obtained as described in the Section of Signal-level Ground-truth Labeling. Thus, WaveNano can essentially overcome the indel (insertions/deletions) issues that commonly exist in existing methods which are due to the segmentation errors. Besides, as WaveNano can conduct base-calling on the entire signal sequence, it leverages bi-directional WaveNets consisting of residual blocks with skip connections. Then the feature maps of the bi-directional WaveNets are concatenated together to capture the ultra long-dependency within the signal sequence. On top of the concatenated feature maps, we predict move labels (binary classification) and 5-mer labels (1024-class classification) simultaneously through one-dimensional convolutional neural network (1D CNN) with a 1×1 kernel. Using the predicted move labels as the segmentation guidance, the predicted 5-mer probability matrix (of size $L_1 \times 1024$) is segmented accordingly and then fed into the Viterbi decoding block to obtain the final base-calling with length L_3. It is worth mentioning that WaveNano does not rely on any other segmentation/decoding tools, and thus is a self-contained offline framework.

Bi-directional WaveNets for joint learning of 5-mer labels and move labels

Our proposed WaveNano method directly operates on the waveform of the normalized Z-score of the input electrical current signal. Given a waveform $x = x_1, ..., x_T$, different from the generative model WaveNet [15] (shown in Figure 4), the joint probability of the 5-mer label $p_1(\mathrm{x})$ and the move label $p_2(\mathrm{x})$ are factorized as follows:

$$p_1(\mathrm{x}) = \Pi_{t=1}^{T} p_1(x_t | x_1, ..., x_{t-1}, x_{t+1}, ..., x_T), \quad (1)$$

$$p_2(\mathrm{x}) = \Pi_{t=1}^{T} p_2(x_t | x_1, ..., x_{t-1}, x_{t+1}, ..., x_T). \quad (2)$$

That is, the 5-mer label and the move label at each time point are conditioned on all other time points.

Since the stacked dilated causal convolutions can have

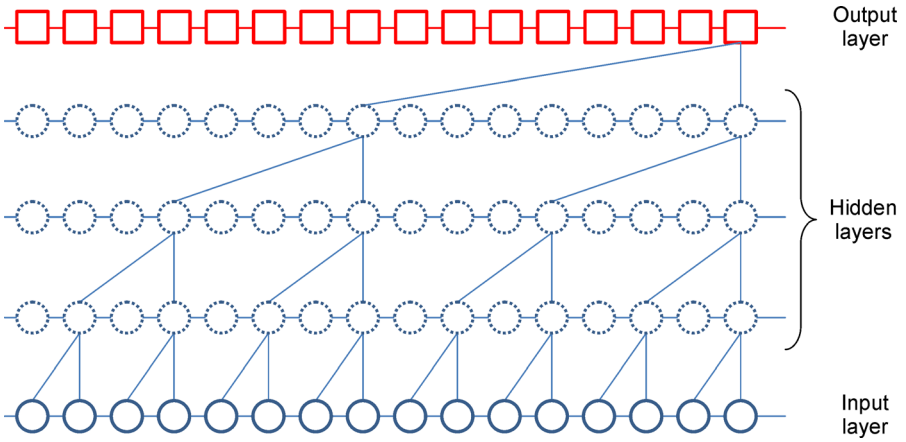

Figure 4. Basic architecture of WaveNets model. WaveNets [15] are autoregressive and consist of stacked causal filters with dilated convolutions for the purpose of growing their receptive fields exponentially with depth, which is critical to capture the ultra-long range temporal dependencies in the input nanopore electrical signals.

a very large receptive field [15], we exploit two parallel WaveNets with stacked dilated 1D CNNs by taking the forward current signal and the reversed current signal as inputs respectively. Similar to the configuration in which dilation is doubled for each layer up to a limit [15], we also use two repeat blocks in WaveNano, *i.e.*, $1, 2, 4, ..., 4096$. Furthermore, the gated activation units $z = \tanh(W_{f,k} * \mathrm{x}) \odot \sigma(W_{g,k} * \mathrm{x})$ [28], residual and parameterized skip connections [19] are also used in WaveNano to speed up convergence and enable a deeper model [15]. Specifically, residual connections are presented in the dotted box and all residual blocks are summed up through skip connections (Figure 3). In order to capture the long-range temporal dependency of all other time points, feature maps of bi-directional WaveNets are combined through the concatenation operation.

On top of the concatenated feature maps, WaveNano conducts the 5-mer label and move label prediction simultaneously though 1D CNN that is activated by softamx with a 1×1 kernel by minimizing the combined loss. It is worth mentioning that the move prediction is a class imbalanced problem (about 8 times more *stay* than *move*). Thus, we add the approximated AUC loss [29] for solving the class imbalanced problem of move prediction.

Specifically, we denote f as the move prediction layer with softmax activation illustrated in Figure 3. According to Ref. [30], the AUC of a predictor f is defined as $\mathrm{AUC}(f) = P(f(t_0) < f(t_1) | t_0 \in D^0, t_1 \in D^1)$, where D^0, D^1 are the samples with ground-truth labels *stay* and *move*, respectively. Its unbiased estimator, *i.e.*, Wilcoxon-Man-Whitney statistics, is $\mathrm{AUC}(f) = \frac{1}{n_0 n_1} \Sigma_{t_0 \in D^0, t_1 \in D^1} I(f(t_0) < f(t_1))$, where $n_0 = |D^0|, n_1 = |D^1|$, and $I(\ )$ is the indicator function. In order to add the noncontinuous AUC loss to the continuous cross-entropy loss and optimize the combined loss through gradient decent, we consider an approximation of the AUC loss by a polynomial approximation of indicator function $I(\ )$ with degree d [30], *i.e.*,

$$\mathrm{AUC}_{\mathrm{move}} = \frac{1}{n_0 n_1} \sum_{\substack{t_0 \in D^0 \\ t_1 \in D^1}} \sum_{k=0}^{d} \sum_{l=0}^{k} \alpha_{kl} f(t_1)^l f(t_0)^{k-l}, \quad (3)$$

where $\alpha_{kl} = c_k C_k^l (-1)^{k-l}$ is a constant.

Thus, the combined cross-entropy loss is:

$$\mathrm{loss} = \mathrm{loss}_{5\text{-mer}} + \lambda_1 \mathrm{loss}_{\mathrm{move}} + \lambda_2 \mathrm{AUC}_{\mathrm{move}}, \quad (4)$$

where $\mathrm{loss}_{5\text{-mer}} = -\frac{1}{T} \Sigma_i m_i^* \log(m_i)$, $\mathrm{loss}_{\mathrm{move}} = -\frac{1}{T} \Sigma_i s_i^* \log(s_i)$, and T is the length of the input signal. Besides, m_i, s_i are the predicted probabilities of the 5-mer label and the move label, m_i^*, s_i^* are ground-truth 5-mer label and move label respectively, and λ_1, λ_2 are the trade-off parameters.

Viterbi decoding with 5-mer labels and move labels as segmentation guidance

Given the predicted probabilities of the 5-mer labels p_1 and move labels p_2 for the electrical current signal sequence X, WaveNano first segments the 5-mer label sequence with the guidance of the predicted move labels. Specifically, for a certain time range t_1 to t_2 from the original signal sequence, if all their predicted *stay* labels are above a given threshold θ, then the predicted 5-mer label (*e.g.*, for a certain label l) for each time point t' in this segment is calculated by $p_1'(t', l) = \frac{1}{t_2 - t_1} \Sigma_{i=t_1}^{t_2} p_1(i, l)$. This produces a segmented event sequence X' with the

probabilities of the 5-mer labels p'_1, which can be interpreted as a smoothed predicted 5-mer label probability matrix. Finally, WaveNano runs the Viterbi decoding algorithm [22] on this smoothed probability matrix to compute the most likely 5-mer sequence S', which can then be transformed to the DNA sequence B directly.

It should be noted that the segmentation process in WaveNano is completely different from that in previous methods, such as Metrichor and Albacore, in the following two aspects: (i) our procedure employs a supervised learning model, WaveNet, which can capture ultra long-range temporal dependencies, whereas existing methods only exploit local information for segmentation; (ii) our procedure is more flexible than previous methods with the help of a tunable parameter θ. A larger θ will produce a longer base-called sequence, whereas a smaller θ will produce a shorter sequence. We found that setting θ to 0.9 as the default value leads to an appropriate base-called sequence with a comparable length to the reference DNA sequence.

Neural network architecture

In WaveNano, we exploit bi-directional Wavenets each with two repeat residual blocks, which consists of causal 1D CNN layers with kernel size 3 and dilated 1D CNN with dilation $1, 2, ..., 2^{12}$. The obtained 12 feature maps, each with 128 channels, are summed through skip connections. And the feature maps of bi-directional Wavenets are concatenated together as the contextual feature vector. The output from the contextual feature vector is regularized with dropout ($=0.5$) to avoid overfitting and fed to two 1D CNN layers with a 1×1 kernel. The output units for the 5-mer label and move label prediction are 1024 and 2 respectively with a softmax activation function. We set $\lambda_1 = 0.1$ and $\lambda_2 = 0.15$ for balancing the two jointly learned tasks.

Our code is implemented in Tensorflow, a publicly available deep learning software. Weights in our neural networks are initialized using the default setting in Tensorflow. We train all the layers in our deep network simultaneously using the Adam optimizer. The batch size is set to 10 and length is fixed to 12K through padding. The entire deep network is trained on a single NVIDIA GeForce GTX TITAN X GPU with 12GB memory. It takes about one to two weeks to train our deep network. In the testing stage, the 5-mer label and move label prediction of a read with length 12,000 takes 0.5 s on average.

ACKNOWLEDGEMENTS

We thank Minh Duc Cao and Lachlan J. M. Coin for providing the nanopore sequencing data for the Lambda phage sample. We thank Haotian Teng for providing helpful discussions. This work was supported by the Kind Abdullah Unviersity of Science and Technology (KAUST) Office of Sponsored Research (OSR) under Awards Nos. FCC/1/1976-04, URF/1/2601-01, URF/1/3007-01, URF/1/3412-01 and URF/1/3450-01.

COMPLIANCE WITH ETHICS GUIDELINES

The authors Sheng Wang, Zhen Li, Yizhou Yu and Xin Gao declare that they have no conflict of interests.

This article does not contain any studies with human or animal subjects performed by any of the authors.

REFERENCES

1. Cao, M. D., Nguyen, S. H., Ganesamoorthy, D., Elliott, A. G., Cooper, M. A. and Coin, L. J. (2017) Scaffolding and completing genome assemblies in real-time with nanopore sequencing. Nat. Commun., 8, 14515
2. Loman, N. J., Quick, J. and Simpson, J. T. (2015) A complete bacterial genome assembled *de novo* using only nanopore sequencing data. Nat. Methods, 12, 733–735
3. Li, Y., Han, R., Bi, C., Li, M., Wang, S. and Gao, X. (2018) DeepSimulator: a deep simulator for nanopore sequencing. Bioinformatics, 34, 2899–2908
4. Jain, M., Fiddes, I. T., Miga, K. H., Olsen, H. E., Paten, B. and Akeson, M. (2015) Improved data analysis for the MinION nanopore sequencer. Nat. Methods, 12, 351–356
5. Lu, H., Giordano, F. and Ning, Z. (2016) Oxford Nanopore MinION sequencing and genome assembly. Genom. Proteom. Bioinf., 14, 265–279
6. Quick, J., Loman, N. J., Duraffour, S., Simpson, J. T., Severi, E., Cowley, L., Bore, J. A., Koundouno, R., Dudas, G., Mikhail, A., *et al.* (2016) Real-time, portable genome sequencing for Ebola surveillance. Nature, 530, 228–232
7. Castro-Wallace, S. L., Chiu, C. Y., John, K. K., Stahl, S. E., Rubins, K. H., McIntyre, A. B. R., Dworkin, J. P., Lupisella, M. L., Smith, D. J., Botkin, D. J., *et al.* (2017) Nanopore DNA sequencing and genome assembly on the International Space Station. Sci. Rep., 7, 18022
8. Loose, M., Malla, S. and Stout, M. (2016) Real-time selective sequencing using nanopore technology. Nat. Methods, 13, 751–754
9. Jain, M., Olsen, H. E., Paten, B. and Akeson, M. (2016) The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biol., 17, 239
10. Goodwin, S.,Gurtowski, J., Ethe-Sayers, S., Deshpande, P., Schatz, M. C. and McCombie, W. R. (2015) Oxford Nanopore sequencing, hybrid error correction, and *de novo* assembly of a eukaryotic genome. Genome Res., 25, 1750–1756
11. Sovic, I., Šikić, M., Wilm, A., Fenlon, S. N., Chen, S. and Nagarajan, N. (2016) Fast and sensitive mapping of error-prone nanopore sequencing reads with GraphMap. Nat Commun., 7, 11307
12. Szalay, T. and Golovchenko, J. A. (2015) *De novo* sequencing and variant calling with nanopores using PoreSeq. Nat. Biotechnol., 33, 1087–1091

13. David, M., Dursi, L. J., Yao, D., Boutros, P. C. and Simpson, J. T. (2017) Nanocall: an open source basecaller for Oxford Nanopore sequencing data. Bioinformatics, 33, 49–55
14. Boža, V., Brejová, B. and Vinař, T. (2017) DeepNano: deep recurrent neural networks for base calling in MinION nanopore reads. PLoS One, 12, e0178751
15. Van Den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu K. (2016) Wavenet: A generative model for raw audio. ArXiv, 1609.03499
16. Hochreiter, S. and Schmidhuber, J. (1997) Long short-term memory. Neural Comput., 9, 1735–1780
17. Chung, J., Gulcehre, C., Cho, K. H. and Bengio, Y. (2014) Empirical evaluation of gated recurrent neural networks on sequence modeling. ArXiv, 1412.3555
18. LeCun, Y., Bengio, Y. and Hinton, G. (2015) Deep learning. Nature, 521, 436–444
19. He, K., Zhang, X., Ren, S., and Sun, J. (2016) Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas
20. Hirschberg, J. and Manning, C. D. (2015) Advances in natural language processing. Science, 349, 261–266
21. Wang, S., Sun, S., Li, Z., Zhang, R. and Xu, J. (2017) Accurate *de novo* prediction of protein contact map by ultra-deep learning model. PLoS Comput. Biol., 13, e1005324
22. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990) Basic local alignment search tool. J. Mol. Biol., 215, 403–410
23. Pearson, W. R. and Miller, W. (1992) Dynamic programming algorithms for biological sequence comparison. In Methods in Enzymology. pp. 575–601, Elsevier
24. Wang, S., Ma, J. and Xu, J. (2016) AUCpreD: proteome-level protein disorder prediction by AUC-maximized deep convolutional neural fields. Bioinformatics, 32, i672–i679
25. McIntyre, A. B., Rizzardi, L., Yu, A. M., Alexander, N., Rosen, G. L., Botkin, D. J., Stahl, S. E., John, K. K., Castro-Wallace, S. L., McGrath, K., *et al.* (2016) Nanopore sequencing in microgravity. npj Microgravity, 2, 16035
26. Teng, H., Cao, M. D., Hall, M. B., Duarte, T., Wang, S. and Coin, L. J. M. (2018) Chiron: translating nanopore raw signal directly into nucleotide sequence using deep learning. Gigascience, 7, giy037
27. Han, R., Li, Y., Wang, S. and Gao, X. (2017) An accurate and rapid continuous wavelet dynamic time warping algorithm for unbalanced global mapping in nanopore sequencing. bioRxiv, 238857
28. van den Oord, A., Kalchbrenner, N., Vinyals, O., Espeholt, L., Graves, A., and Kavukcuoglu, K. (2016) Conditional image generation with pixelcnn decoders. In Advances in Neural Information Processing Systems
29. Wang S., Sun S., and Xu J. (2016) AUC-maximized deep convolutional neural fields for protein sequence labeling. In Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2016. Lecture Notes in Computer Science, Frasconi P., Landwehr N., Manco G., Vreeken J. (eds) vol 9852. Springer, Cham
30. Calders T., and Jaroszewicz S. (2007) Efficient AUC optimization for classification. In Knowledge Discovery in Databases: PKDD 2007. Lecture Notes in Computer Science, Kok J. N., Koronacki J., Lopez de Mantaras R., Matwin S., Mladenič D., Skowron A. (eds), vol 4702. Springer, Berlin, Heidelberg

