深度学习在生物学中的应用：
对MinION测序结果中base calling的计算

作者：writer 来源：科学网

本文原地址：https://www.iikx.com/news/progress/3531.html

本文仅供学习交流之用，版权归原作者所有，请勿用于商业用途！

深度学习在生物学中的应用： 对MinION测序结果中base calling的计算。记得去年阿尔法狗(Alpha Go)的新闻出来后，有些生物学领域的研究人员可能会考虑是否需要跨专业学习一下Artificial Intelligence
(AI)，看看它能否在生物领域也掀起热浪。结果当他们刚刚了解到阿尔法狗的命脉乃来自Deep Learning (深度学习)真传时，它的亲兄弟AlphaFold 就以迅雷不及掩耳之势(2018年12月新闻发布会)在蛋白质折叠预测领域独领风骚。有生物学背景的人都知道，虽然科学家们破译了基因组，但从DNA到蛋白质翻译过程受各种基因和/或蛋白质的调控、修饰，并且蛋白质从翻译产生到能发挥功能的这一过程也是在细胞内经历了多次折叠。但人家AlphaFold则不畏这些千难万苦，硬生生的通过氨基酸序列直接预测蛋白质的3D结构(AlphaFold 的新闻发布链接：https://deepmind.com/blog/alphafold/)。所以当AlphaFold一出世，大家都惊呼它是能把诺贝尔奖抱回家的人选之一。

既然AlphaFold和AlphaGo是亲兄弟，那它们两个到底有什么相同点呢?其实这个相同点就是Deep Learing算法。而不同点则在于该算法分别应用在了围棋领域和蛋白质研究领域。所以，无论是生物，物理，还是化学领域的科研人员是时候学习一下Deep
Learing算法，说不定你就是下一个诺贝尔获奖者呢!原来从事Deep
Learing研究的专家和学者已经进军到生物领域的各个方向了。

Quantitative Biology (QB)期刊向来提倡交叉，尤其是计算、数学、物理前沿领域与生命科学的交叉，并且期刊也一直在跟踪刊登这些交叉领域的前沿热点文章。在Deep Learning这个火热阶段，QB编辑部邀请到了该领域的曾坚阳教授和裴剑锋教授作为QB期刊2018年第四期的Guest Editors，为QB刊组织了一场关于Deep Learning或Neural Network在第三代测序分析碱基识别、线粒体形态定量分析和药用蛋白预测中应用的盛宴(感兴趣的研究人员可以登陆期刊网站先睹为快：http://journal.hep.com.cn/qb/EN/2095-4689/current.shtml)。

今天重点推送的是Deep Learning在MinION测序仪base-calling中的应用【1】(WaveNano:a signal-level nanopore base-caller via simultaneous prediction of nucleotide labels and move labels through bi-directional WaveNets， http://journal.hep.com.cn/qb/EN/10.1007/s40484-018-0155-4)。

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.**

QB期刊

文章简介：

这篇文章是来自沙特阿卜杜拉国王科技大学(KAUST)的Xin Gao教授团队与香港中文大学(深圳)的Zhen Li 博士合作完成。 Xin Gao 教授团队目前已经完成了一系列与纳米孔测序相关的工作。该团队关于Deep Learning在MinION测序仪数据模拟器中的工作(DeepSimulator: a deep simulator for Nanopore sequencing)于今年9月份发表在了生物信息学领域老牌期刊Bioinformatics上的哦【2】。此外，该团队还完成了一款全新的cwDTW算法，可以高效的联配超长的纳米孔信号，并以此为基础进行信号标注(signal labeling)从而能够检测单核苷酸多态性(SNP)【3】。该工作在国际顶级生物信息学会议ECCB

2018上口头展示，同时亦发表于Bioinformatics。在这里向Xin Gao教授及其团队表示祝贺!

英国生物技术公司Oxford Nanopore自2014年推出MinION测序仪后，由于其小巧的身材(iphone大小)，要求不高的运行环境，较长的reads读取(超过15kb)，较快的测序速度，实时的测序数据监测等特点，一经问世就受到广泛关注。该测序仪于2016年登上了国际太空站(ISS)，完成了第一次太空测序，并证实了对Lambda phage的测序结果在ISS和地球上并无差别【4】。 MinION测序仪的基本工作原理是基于纳米孔测序技术，通过检测单链DNA分子通过纳米孔时引起电流变化的不同，用于碱基的识别(见Figure 1)。由于电流检测的频率通常是DNA序列通过纳米孔速度的7-9倍，因此这对base-calling造成巨大的技术挑战。此外，较高的测序错误率，尤其是对indels(插入和缺失)的测序，是纳米孔测序仪面临的一个主要问题。

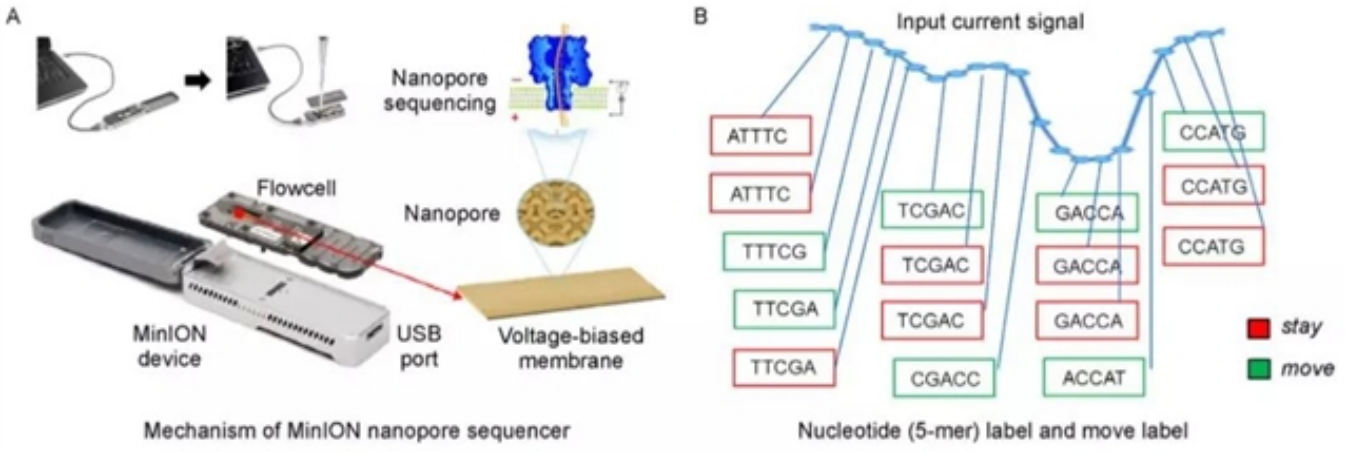

Figure 1. Mechanism of MinION nanopore sequencer and the two signal-level labels. (A) The MinION nanopore-based sequencing device. Sequencing is performed by adding the DNA sample to the flowcell. (B) When DNA molecules pass through the nanopore, each five-tuple DNA (denoted as 5-mer, *e.g.*, "ATTTC") 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 *stays* in the pore (shown in red) or *moves* by one nucleotide (shown in green). Such stay/move labels serve as the segmentation guidance to convert the current signal to a 5-mer event sequence.

为了解决纳米孔测序仪较高错误率的问题，目前已经出现了多种算法。这些算法可以简单地归为两类，即基于机器学习(Machine Learning)的算法和基于共有序列(Consensus)的算法。这两种算法的基本原理都是通过serial base-calling过程(见Figure 2A)进行碱基识别，而这一过程势必会增加错误率。此外，这两种算法所用到的机器学习中的模型建构(Model Architecture)更适用于短片段的计算。为了解决上述问题，本文作者们采用了Google DeepMind团队在语音合成和语音识别方面新开发的具有完美表现的WaveNets深度学习方法【5】将纳米孔中的信号当作语音信号，而base-calling则类似于语音识别过程，开发了一种基于机器学习的新算法-WaveNano(见Figure 2B和Figure 3)。这种算法不依赖任何segmentations/decoding工具，而完全是一种self-contained 的线下工具。

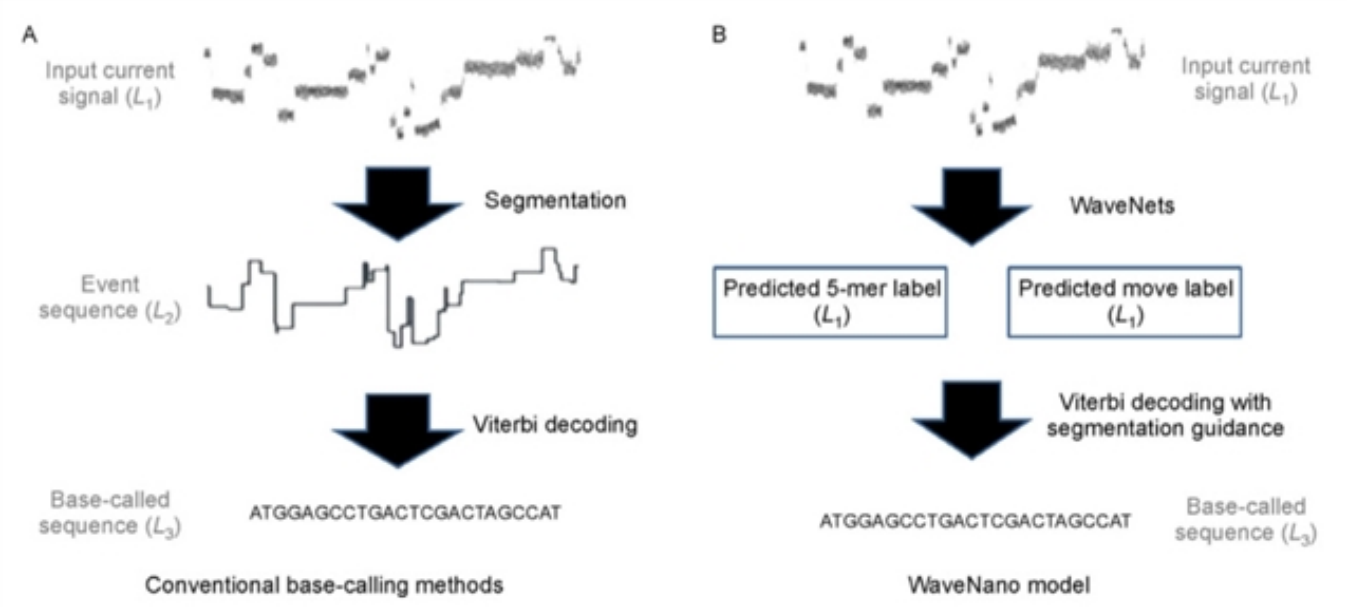

Figure 2. Comparison of the conventional base-calling methods and the WaveNano model. (A) Conventional base-calling methods conduct segmentation and decoding in a sequential manner. (B) WaveNano predicts 5-mer labels and move labels simultaneously, and then performs segmentation guided decoding.

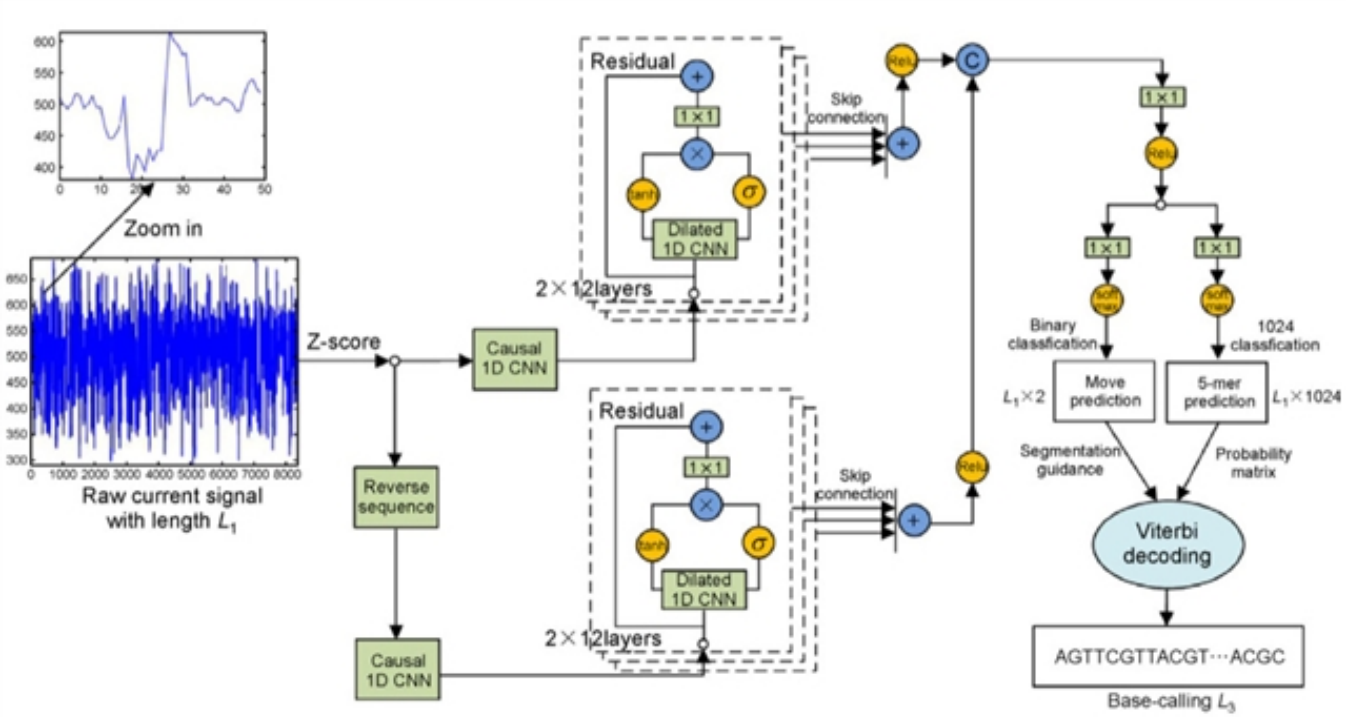

Figure 3. The overall pipeline of WaveNano. WaveNano takes the current signal X with length L_1 as the input. The calculated Z-score of the current signal feeds into the causal one-dimensional (1D) convolutional neural network (CNN) and stacked residual blocks with skip connections. Different from the 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.

通过该算法，文章作者对Lambda phage的基因组用MinION进行了测序，结果得到了大约24,000个

reads，电流信号平均为63,000bp。同时，作者还将WaveNano与官方的Metrichor算法以及Albacore算法进行了结果比较(如Table 1)，结果表明WaveNano不仅能预测比较准确的DNA序列，同时该算法对indel的处理结果明显优于Metrichor和Albacore。此外，WaveNano的运行时间约为1个信号序列为0.5s，而Albacore的运行时间则为2s。

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 [illegible] and shown in percentage.

由此可见，WaveNano算法对于分析MinION产生的Lambda phage测序结果具有良好的表现，尤其对于indel序列的分析，其结果要比目前商用的Metrichor和Albacore具有更高的准确度。

Reference

1. Sheng Wang, Zhen Li, Yizhou Yu and Xin Gao. (2018) WaveNano:a signal-level nanopore base-caller via simultaneous prediction of nucleotide labels and move labels through bi-directional WaveNets. Quant. Biol., 6 (4): 359-368.

2. Yu Li, Renmin Han, Chongwei Bi, Mo Li, Sheng Wang, Xin Gao. (2018) DeepSimulator: a deep simulator for Nanopore sequencing. Bioinformatics, 34 (17), 2899-2908

3. Renmin Han, Yu Li, Xin Gao, Sheng Wang. (2018) An accurate and rapid continuous wavelet dynamic time warping algorithm for end-to-end mapping in ultra-long nanopore sequencing. Bioinformatics, 34 (17), i722-i731

4. 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

5. 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

关于期刊：

Quantitative Biology(QB)期刊于2013年创刊，由高等教育出版社、清华大学和北京大学共同主办。由教育部主管，高等教育出版社出版、德国施普林格(Springer)出版公司负责海外发行。

QB期刊主要报道生命科学的定量研究和系统整合的前沿热点，包括系统生物学、合成生物学、生物信息学以及计算生物学领域的最新研究成果和前沿进展，旨在推动定量生物学领域科研水平

提升，同时为生命科学与计算机、数学、物理等交叉研究领域打造一个学术水平高、可读性强、具有全球影响力的交叉学科期刊品牌。

目前，QB以网络版和印刷版两种形式出版，全球免费。同时，被QB收录的文章全部免费出版，不收取任何费用。此外，对于特别优秀文章开通绿色通道，从投稿到第一个决定只需1周。

更多的文章请点击：

http://journal.hep.com.cn/qb

https://link.springer.com/journal/40484

欢迎投稿，投稿直通：

https://mc.manuscriptcentral.com/qb

主编阵容：

期刊目前由清华大学/美国德克萨斯大学达拉斯分校张奇伟教授和北京大学定量生物学中心汤超教授担任共同主编。清华大学张学工教授担任执行主编。此外，期刊还邀请了40余名国内外知名领域专家、学者组成编委会，保证刊物的高学术水平。

共同主编：张奇伟(Michael Q. Zhang)教授

共同主编：汤超 教授

执行主编： 张学工教授

(来源：科学网)

更多 科学进展 请访问 https://www.iikx.com/news/progress/

本文版权归原作者所有，请勿用于商业用途，爱科学iikx.com转发