SIMON – A Novel Diffusion Machine Learning Model that Accurately Simulates ONT Signal Data

Raehash Shah[1], Chris Remillard[2], Jared Richardson[2], David VanHoute[2], Andreas Toba[2]
[1]Carnegie Mellon University, [2]QC Virology – IOPS Regeneron

Abstract

In the past decade, nanopore sequencing has become the preferred method for generating accurate long reads in real time. As the DNA molecule passes through the biological nanopore of an Oxford Nanopore Technologies (ONT) sequencer, the sequencer reads the electrical current signal of the nanopore and translates the signal data into the sequence of the DNA molecule. As more applications rely on having large amounts of sequencing data, a tool is proposed that can accurately simulate signal data. SIMON (SIgnal siMulated by diffusiON) is a novel method that uses a conditional diffusion model and structured state space model to generate signal data. Once SIMON generated signal data, signal and sequence evaluation metrics were used to evaluate SIMON. From these evaluation metrics, although SIMON did well in generating signaling data, SIMON can be improved in accuracy and be used in multiple applications like a spiked cell line and comparison analysis.

Motivation

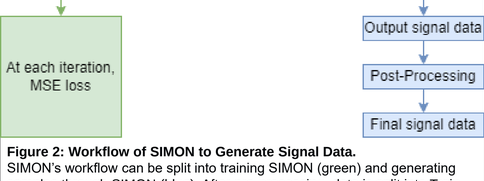

Figure 1: Conditional Diffusion-based Structured State Space Models Generated ECG Data Indistinguishable from Real ECG Data.
While looking for different approaches to simulate ONT Nanopore signal data, J. Alcaraz and N. Strodthoff published a paper in 2023 demonstrating a conditional diffusion approach to simulate ECG data which outperformed previous approaches for generating waveform data. Therefore, this project adapts the architecture of this model to fit the properties of Nanopore signal data. Figure taken from (Kong et al., 2021)

Model Architecture

Figure 2: Workflow of SIMON to Generate Signal Data.
SIMON's workflow can be split into training SIMON (green) and generating samples through SIMON (blue). After pre-processing, data is split into Train and Sample and a database unique to the samples worked with is created mapping the k-mer to signal values. Once signal data trains SIMON, a ground truth signal is passed into the trained model to generate signal data which undergoes post-processing of adding noise, dynamic time warping and trimming to generate the final signal data.

Introduction

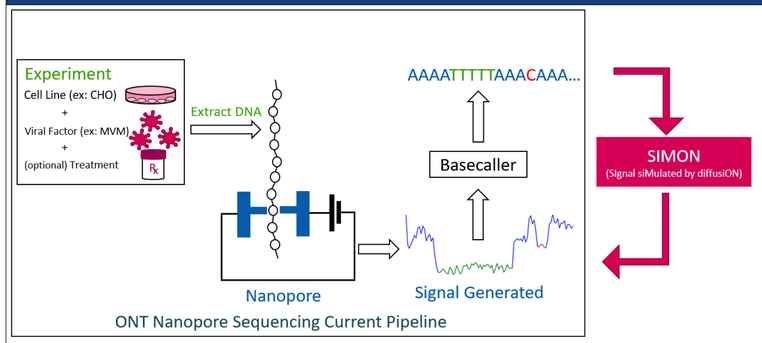

Figure 3: SIMON's integration into QC Virology's current platform.
Currently, as a part of upstream testing, QC Virology focuses on identifying viral species present in the sample when mammalian cells or CHO cells that can produce the drug are cultured in the bio-reactor. The method of identifying viral samples is through a form of 3rd generation sequencing, ONT Nanopore Sequencing. During sequencing the DNA is passed through the nanopore of the device which corresponds to a gradient charge change which is measured as a signal value. This signal value is then passed into a RNN (Recurrent Neural Network) model called the basecaller which matches the changes observed in the gradient to a corresponding nucleotide base. SIMON, as a tool, would act as the reverse of the Basecaller where SIMON would translate sequencing data back into signaling data. This would be a valuable tool for QC Virology as this tool would assist in validation of computational tools and make it possible to work with sequencing data of viruses outside of the scope of the QC Virology's lab capabilities

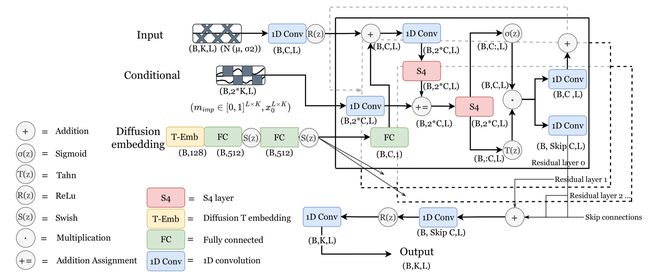

Figure 4: SIMON's computational architecture (A Structured State-Space Model called SSSDS4).
Adopted from J. Alcaraz and N. Strodthoff, this is the proposed architecture for the SSSDS4 which is the backbone computation for SIMON. A structured state-space model allows for the ability to capture long-term dependencies in time series data. SSSDS4, a type of structured state space model incorporates conditional diffusion which has been the foundational concept for other signal generation specifically speech synthesis in DiffWave. However, the important difference is the S4 layer which is a diffusion layer within each residual block instead of a bidirectional dilated convolution which would allow for more diffusion-based embedding. In addition, a secondary S4 layer was added after joining it with the conditional information so that the model has added flexibility when incorporating conditional information. This model architecture had strong success with ECG data, so was used for ONT nanopore signal data. (Figure taken from Alcaraz et al., 2023)

Results

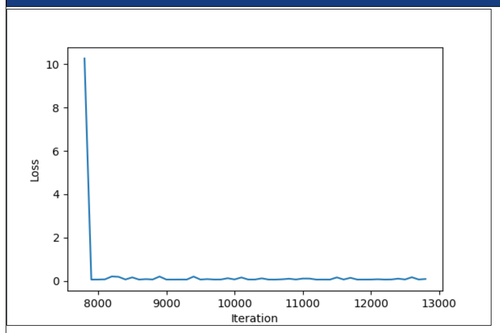

Figure 5: SIMON Training Loss Function.
At each iteration of training, a mean-squared loss value was computed comparing a randomly sampled signal to the final signal that has undergone diffusion adding noise to each iteration. As we can see the final loss value is 0.018 which corresponds to an accuracy of 98.19%

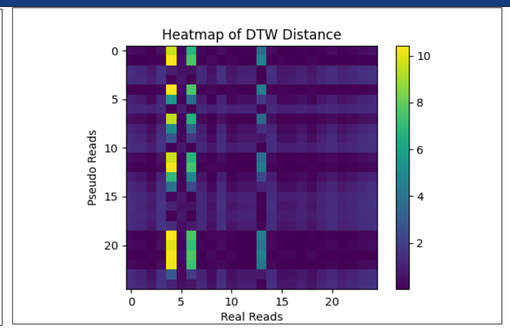

Figure 6: DTW Distance Heatmap of Real and Simulated Signal Data.
25 random samples (both real reads and simulated reads) were used to compute the dynamic time warping distance between each of the reads. Notice that although there are distance values that are relatively large, most of the dynamic time warping distance is small suggesting strong pattern matching.

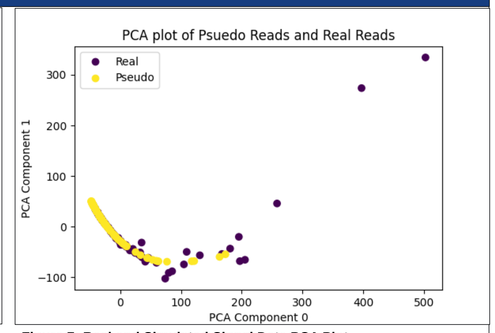

Figure 7: Real and Simulated Signal Data PCA Plot.
The plot of generated reads and the real reads were plotted on a PCA plot, to demonstrate how indistinguishable the generated reads are from the real reads generated. As we can see some of the pattern of real reads match the pattern of pseudo reads. The accuracy of binary classifier is a 46%.

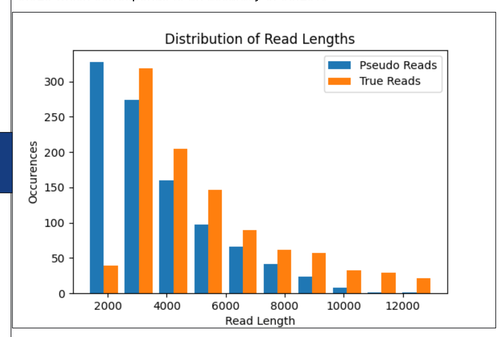

Figure 8: Distribution of Simulated Signal Data compared to Real Signal Data.
Read length distributions of simulated signal data and true signal data. Notice that there is differences between the distribution of signal data compared to real signal data which is where SIMON can be improved especially for smaller reads because the assumption has been made of the read length distribution. This affects the number of reads generated being less than the true reads.

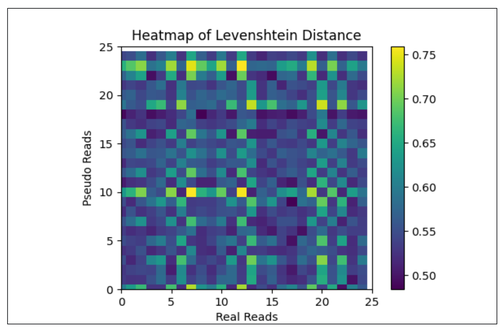

Figure 9: Levenshtein Distance Heatmap of Real and Simulated Signal Data.
A levenshtein distance was computed between 25 random samples of real and simulated data. On average, the plot suggests that 2 operations of mimatches or indels is necessary for every 3 base pairs. Therefore, although this is high, we can use this distance metric as a supplement to other evaluation metirics.

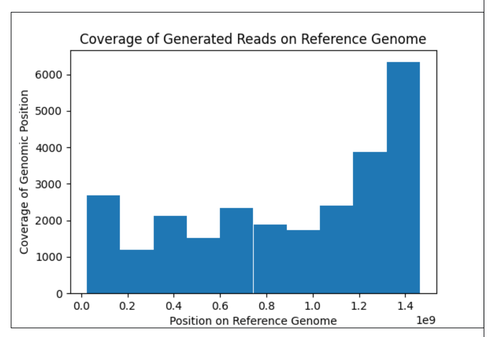

Figure 10: Coverage of Basecalled Simulated Signal Data on the Reference Genome Signal Data.
Coverage of the reads are shown meaning the number of times a possible read is at each genomic position. Notice that this indicates a large coverage value which suggests the generated reads are quite representative of the genome. There is also higher coverage on positions later in the genome which might be a room for exploration for read generation.

Evaluation Metrics

Signal Data Performance	Basecalled Sequence Data Performance
Dynamic Time Warping (DTW) Distance Heatmap	Levenshtein Distance Heatmap
Binary Classifier Accuracy	Alignment of Simulated Reads to Reference Genome

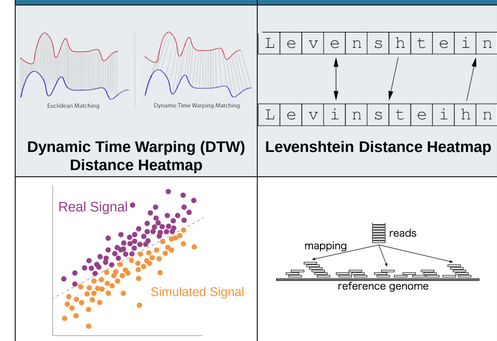

Figure 11: Evaluation Metrics of Signaling and Sequencing Data.
To evaluate how well SIMON could simulate data, the signaling data generated by SIMON was compared to real signaling data based on the Dynamic Time Warping Distance which measures how much the simulated data varies over time compared to real data. In addition, to assess how indistinguishable the simulated data is, a binary machine learning Logistic Regression Classifier trained on the properties of a simulated signal data and real signal data attempted to classify the differences between the two signaling data. In addition, the simulated signaling data was basecalled using the standard basecaller Guppy to generate sequencing data. Using this sequencing data, the Levenshtein Distance between the true reads and the generated reads was computed to see what deviations (mismatches, indels) took place in the basecalled reads. As well, the basecalled simulated reads were aligned to the reference genome of the species to see where they align to the genome and how well the alignment is to the genome.

Conclusion

	Sequencing Run Signal Data	SIMON Simulated Signal Data (in theory)
	Slow (days)	Fast (hours)
	Accurate	Accurate
	Expensive	Cheaper

Future Directions

- Improve accuracy of SIMON by providing SIMON with more training data
- Capture more features of signaling data by making SIMON's architecture larger with more nodes
- Compare SIMON performance compared to other signal generation models
- Generate signaling data for a sample containing both viral and mammalian factors

References

1. Alcaraz, J. M. L., & Strodthoff, N. (2023). Diffusion-based Time Series Imputation and Forecasting with Structured State Space Models. *Transactions on Machine Learning Research*. https://openreview.net/forum?id=hHiIbk7ApW
2. Alcaraz, J. M., & Strodthoff, N. (2023). Diffusion-based conditional ECG generation with structured state space models. *Computers in Biology and Medicine*, *163*, 107115. https://doi.org/10.1016/j.compbiomed.2023.107115
3. Berger, B., Waterman, M. S., & Yu, Y. W. (2021). Levenshtein Distance, Sequence Comparison and Biological Database Search. *IEEE transactions on information theory*, *67*(6), 3287–3294. https://doi.org/10.1109/tit.2020.2996543
4. Han, R., Li, Y., Gao X., & Wang, S., (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. https://doi.org/10.1093/bioinformatics/bty555
5. Joshi. D, Mao, S., Kannan, S., & Diggavi S. (2020). QAlign: aligning nanopore reads accurately using current-level modeling. *Bioinformatics* 37 (5) 625-633. https://doi.org/10.1093/bioinformatics/btaa875
6. Kong, Z., Ping, W., Huang, J., Zhao, K., & Catanzaro, B. (2021). DIFFWAVE: A VERSATILE DIFFUSION MODEL FOR AUDIO SYNTHESIS. *International Conference on Learning Representations*. https://openreview.net/forum?id=a-xFK8Ymz5J
7. Pagès-Gallego, M., de Ridder, J. Comprehensive benchmark and architectural analysis of deep learning models for nanopore sequencing basecalling. *Genome Biol* **24**, 71 (2023). https://doi.org/10.1186/s13059-023-02903-2

REGENERON