Research Statement: Abstracts

Minindu Weerakoon[1*]

[1] Auburn University, Auburn, AL 36849
* Correspondence: wmw0016@auburn.edu;

In the Computational Biology Lab under the mentorship of Professor Haynes Heaton, I am dedicated to advancing the field of sequence analysis. My primary objective is to critically examine the existing challenges and limitations within the state-of-the-art methodologies in sequence analysis. By identifying these core issues, I aim to not only propose innovative solutions but also to implement them effectively, thereby setting new benchmarks in accuracy, efficiency, and applicability of sequence analysis in biological research. This endeavor involves a comprehensive approach, incorporating the latest in computational strategies, algorithm development, and data analysis techniques, to usher in a new era of advancements in computational biology.

Abstract topoqual [1]

Pacific Biosciences (PacBio) circular consensus sequencing (CCS) aka high fidelity (HiFi) technology has revolutionized modern genomics by producing long (10+kb) and highly accurate reads by sequencing circularized DNA molecules multiple times and combining them into a consensus sequence. Currently the accuracy and quality value estimation is more than sufficient for genome assembly and germline variant calling, but the estimated quality scores are not accurate enough for confident somatic variant calling on single reads. Here we introduce TopoQual, a tool utilizing partial order alignments (POA), topologically parallel bases, and deep learning to polish consensus sequences and more accurately predict base qualities. We correct ~31.9% of errors in PacBio consensus sequences and validate base qualities up to q59 which is one error in 0.9 million bases enabling accurate somatic variant calling with HiFi data.

Abstract CarM [2]

Continual Learning (CL) is an emerging machine learning paradigm in mobile or IoT devices that learns from a continuous stream of tasks. To avoid forgetting of knowledge of the previous tasks, episodic memory (EM) methods exploit a subset of the past samples while learning from new data. Despite the promising results, prior studies are mostly simulation-based and unfortunately do not promise to meet an insatiable demand for both EM capacity and system efficiency in practical system setups. We propose CarM, the first CL framework that meets the demand by a novel hierarchical EM management strategy. CarM has EM on high-speed RAMs for system efficiency and exploits the abundant storage to preserve past experiences and alleviate the forgetting by allowing CL to efficiently migrate samples between memory and storage. Extensive evaluations show that our method significantly outperforms popular CL methods while providing high training efficiency.

Abstract Carousel memory [3]

In the burgeoning field of machine learning, Continual Learning (CL) stands out for its ability to adaptively learn from an unending sequence of tasks, ensuring that knowledge from past tasks is not forgotten. This capability is crucial, especially when considering the deployment of CL systems on devices with limited memory, such as mobile and IoT devices. Traditional approaches to mitigate the challenge of memory-related performance degradation involve the use of episodic memory (EM), which selectively stores samples from previously encountered data. However, this strategy hits a snag due to the limited memory capacity of typical hardware on such devices, which constrains the EM size and, by extension, the system's ability to maintain high accuracy levels for practical applications.

One critical limitation of existing CL methodologies is their inability to recover samples that exceed the episodic memory's storage capacity, leading to a permanent loss of potentially valuable information. This not only results in a

loss of data but also exacerbates the problem of catastrophic forgetting, where the system loses the ability to recall previously learned information.

To tackle these challenges, our research introduces a groundbreaking approach to episodic memory management termed Carousel Memory (CarM). CarM leverages a novel hierarchical storage system that utilizes not only the high-speed RAM typically used for EM but also the larger-capacity internal storage devices found in mobile and IoT devices. This method significantly expands the available memory resource without sacrificing access speed, thanks to an efficient management strategy that seamlessly migrates data between the fast-access memory and the larger internal storage.

Our innovative approach ensures that CarM can be integrated with existing CL frameworks to enhance their performance. Through rigorous testing with seven established CL methodologies, CarM has demonstrated remarkable improvements in final average accuracy rates—by as much as 28.4%—without compromising on training efficiency. This represents a substantial advancement in the field, suggesting that Carousel Memory could play a pivotal role in the future development of Continual Learning applications, particularly those intended for memory-constrained environments.

Abstract BSP++ [unpublished]

This study introduces an innovative adaptation of the Longest Common Subsequence with k mismatches (LCSk++) algorithm, applied to Partial Order Alignment (POA) graphs, aiming to improve sequence alignment within complex biological datasets. By extending LCSk++ to graph structures, specifically POA graphs that represent multiple sequence alignments, we enhance its utility in bioinformatics. Our method leverages dynamic programming and graph traversal to identify conserved regions across sequences (backbone), which allows us to perform local and semi global alignment on the banded poa graph, thereby improving the accuracy and speed of sequence alignment and consensus sequence construction.

References

1. Weerakoon, M., Heaton, H., Lee, S., & Mitchell, E. (2024). TopoQual polishes circular consensus sequencing data and accurately predicts quality scores. bioRxiv, 2024-02.
2. Lee, S., Weerakoon, M., Choi, J., Zhang, M., Wang, D., & Jeon, M. (2022, July). CarM: Hierarchical episodic memory for continual learning. In Proceedings of the 59th ACM/IEEE Design Automation Conference (pp. 1147-1152).
3. Lee, S., Weerakoon, M., Choi, J., Zhang, M., Wang, D., & Jeon, M. (2021). Carousel Memory: Rethinking the Design of Episodic Memory for Continual Learning. arXiv preprint arXiv:2110.07276.