WORKFORCE DEVELOPMENT & EDUCATION

Improving Metagenome Assembly Scaffolding With Machine Learning

U.S. DEPARTMENT OF ENERGY
Office of Science

Emmanuel Adara[1], Mentor (Primary Investigator) Steven Hofmeyr[2]
[1]University of Alabama, [2]Lawrence Berkeley National Laboratory

Abstract

The project is set to determine machine learning approaches that can outperform the current heuristics in MetaHipMer (MHM), the learning approach can in future work, be integrated into the MHM pipeline to offer improved assembly quality to all MHM users, with beneficial results for microbiome work depending on metagenome assembly.

Background info

The ExaBiome project has developed MetaHipMer (MHM), the only metagenome assembler to scale to thousands of compute nodes in modern supercomputers. Consequently, it is the only assembler able to coassemble terabase sized metagenomes, leading to opportunities for new scientific discoveries. MHM is a complex software pipeline that uses heuristics extensively. In one particular stage, called scaffolding, heuristics are used to determine the best path through the *contig graph*, where the vertices are contiguous sequences of DNA (contigs), and the edges are possible links between those contigs. The best path is the most contiguous one with the fewest errors. Currently, this path algorithm relies on heuristics that are derived from domain specific knowledge of genomes. As such, we believe there is room for improvement, specifically through the application of machine learning.

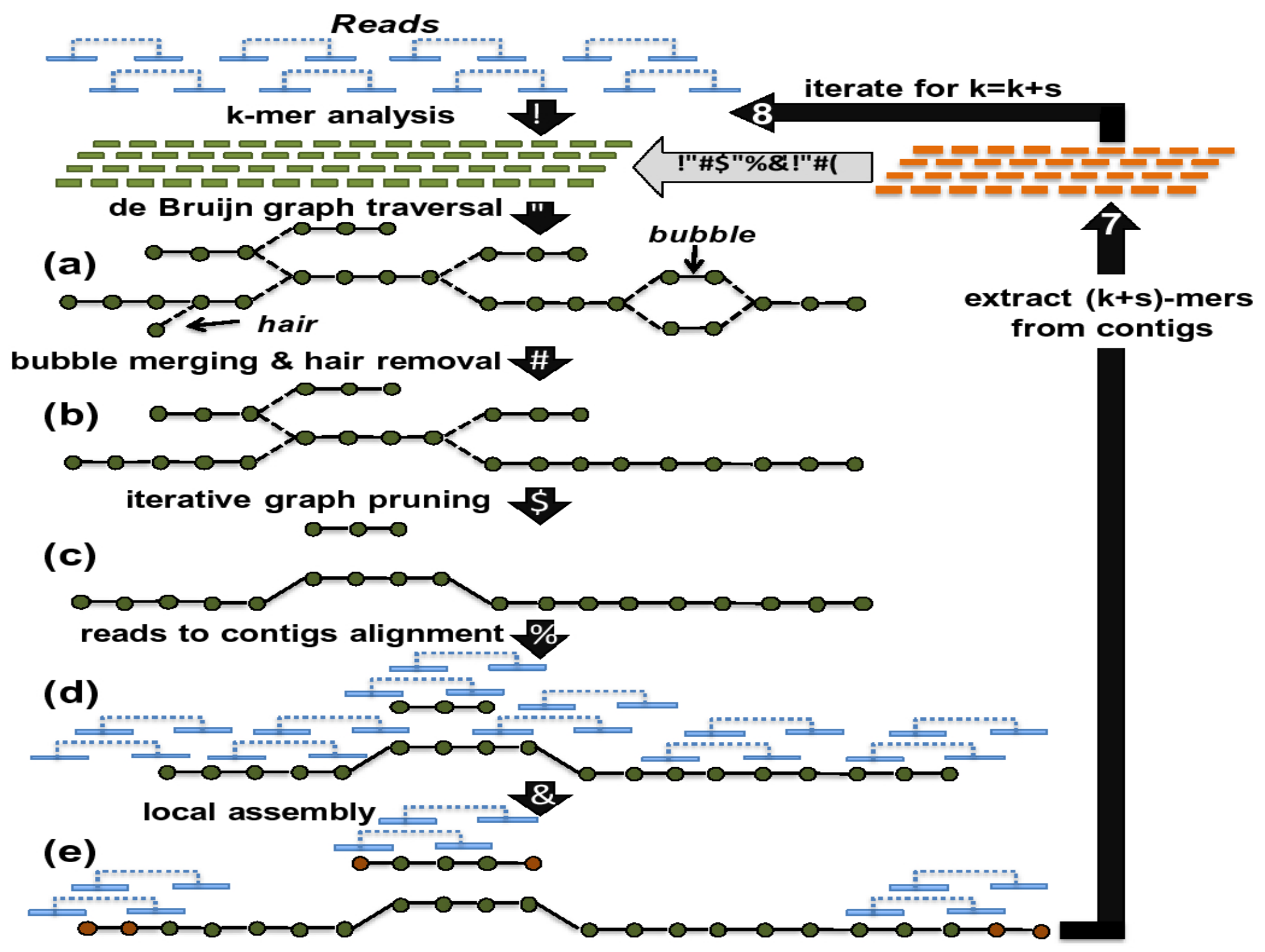


Figure 1: [1]Iterative contig generation workflow in MetaHipMer method.

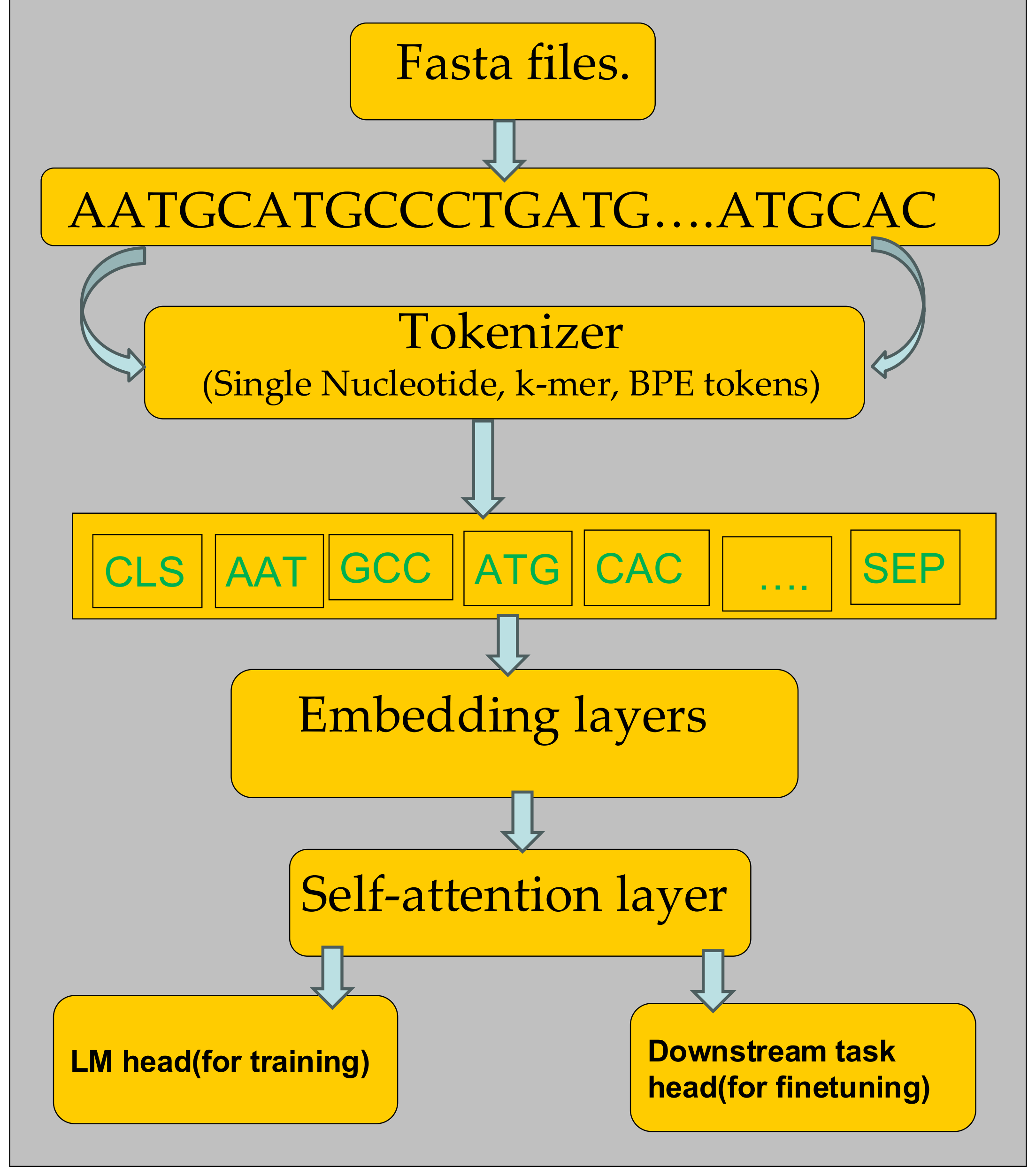
Research Question

This project aims to test the hypothesis that machine learning can improve scaffolding over the current heuristic approach and hence improve the quality of assemblies produced by MHM.

Method

- •determine a ground truth for whether a given edge in the contig graph is actually correct.
- • generate large labeled datasets that can be used to train machine learning systems to differentiate correct from incorrect edges.
- •experimenting with various machine learning approaches in the problem context
- •fine-tuning the most promising ones to maximize the accurate detection of correct edges in the contig graph.

LLMs Overview



Fasta files.

AATGCATGCCCTGATG….ATGCAC

Tokenizer
(Single Nucleotide, k-mer, BPE tokens)

CLS | AAT | GCC | ATG | CAC | …. | SEP

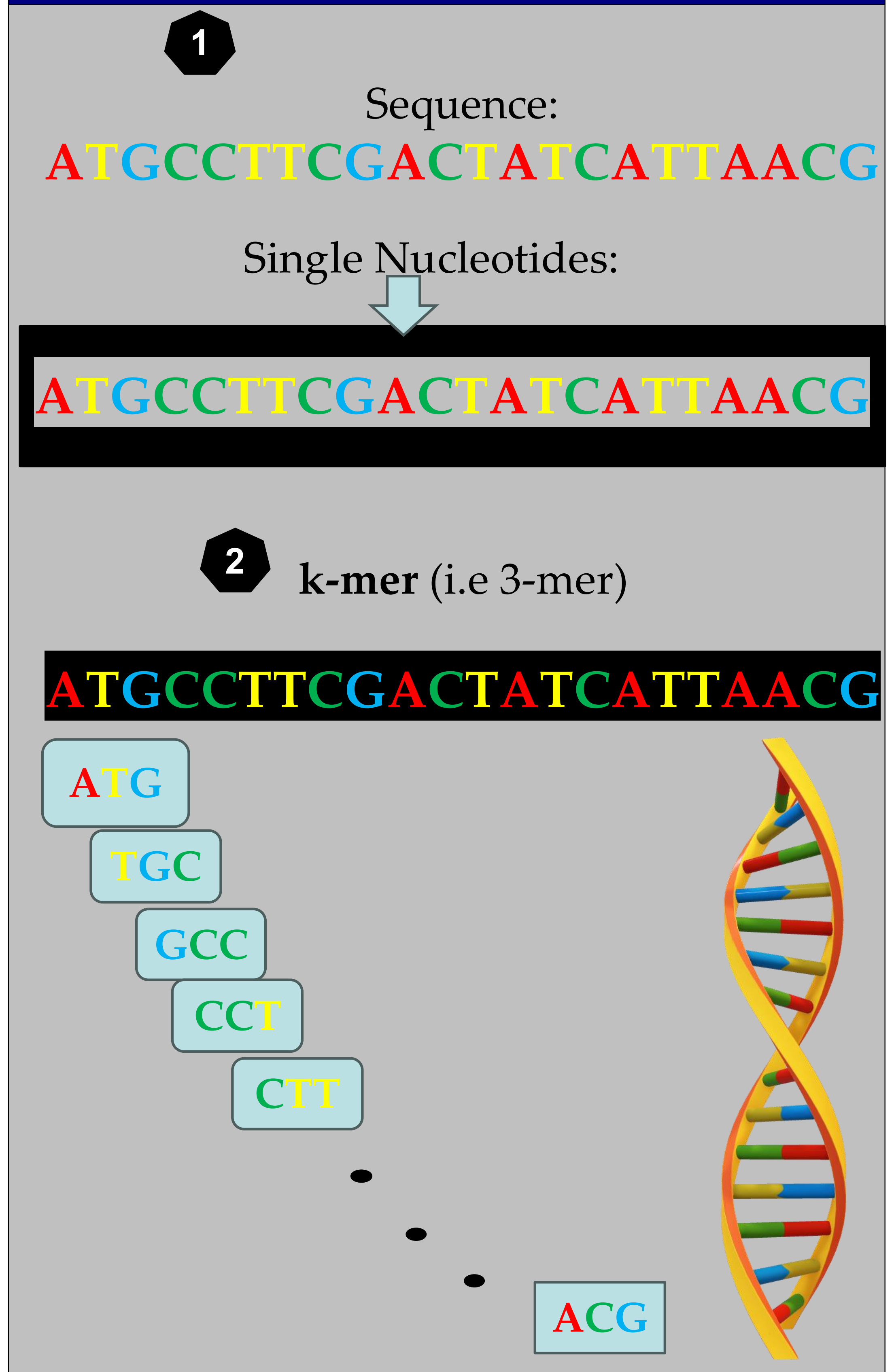
Embedding layers

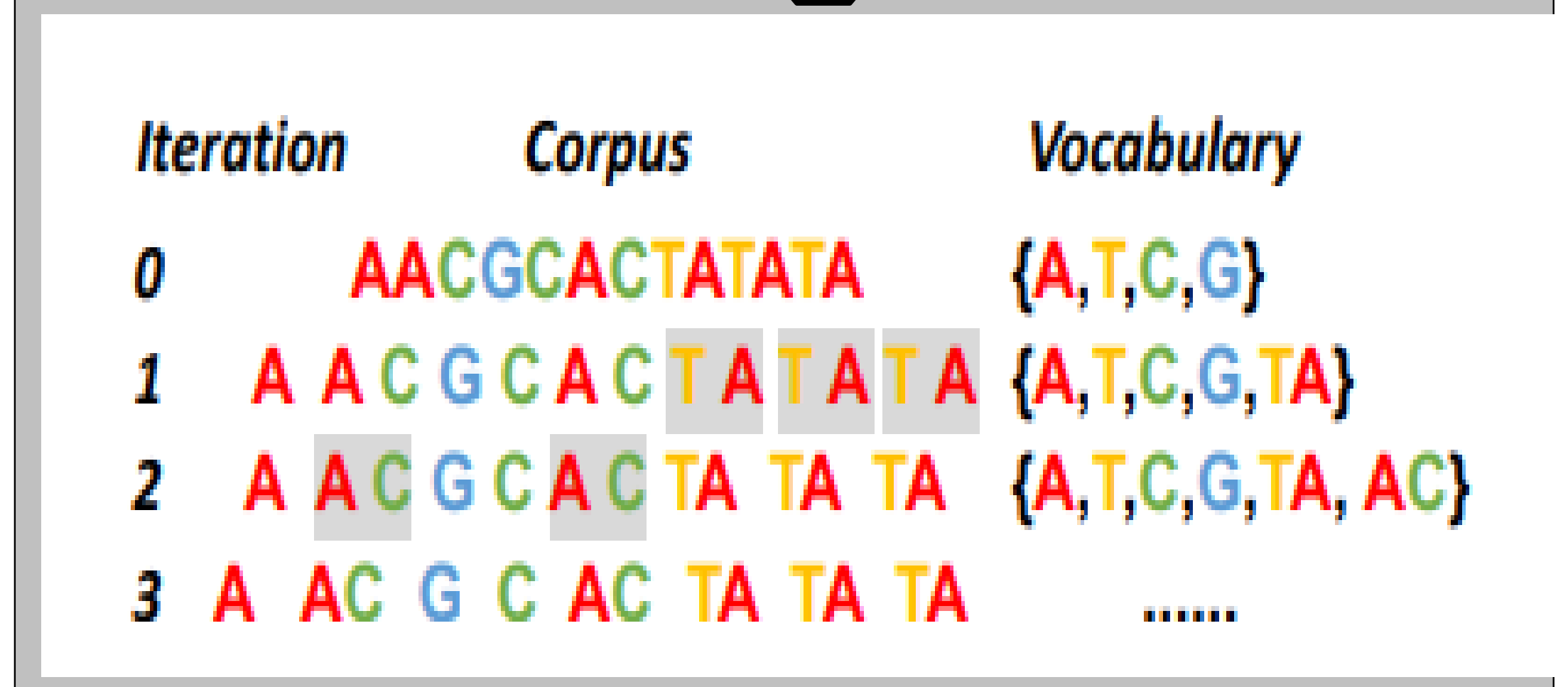
Self-attention layer

LM head(for training)

Downstream task head(for finetuning)

Tokenizers





Iteration	Corpus	Vocabulary
0	AACGCACTATATA	{A,T,C,G}
1	A A C G C A C T A T A T A	{A,T,C,G,TA}
2	A A C G C A C TA TA TA	{A,T,C,G,TA, AC}
3	A AC G C AC TA TA TA	……

Figure 2: [2]BPE tokenizer vocabulary construction

Workforce Development & Education has made arrangements with the Creative Services Office (CSO) to print posters for all Workforce Development & Education Office programs.

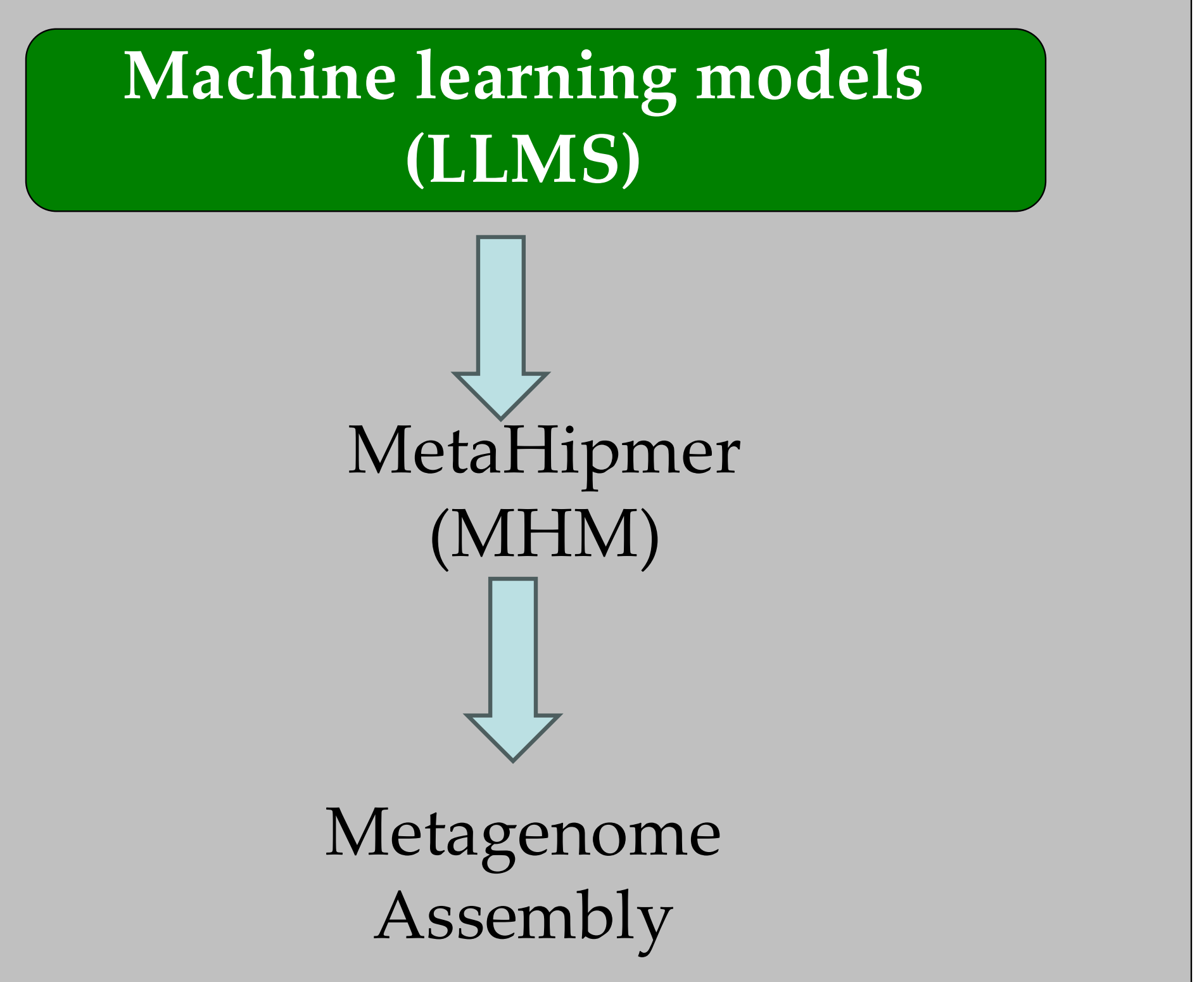
& Education Poster Due <Date> in subject line of email

- • Workforce Development & Education

Progress so far…

- • Review of different machine learning approaches
- •Comparison of different Large language models(LLM) and their applications to genome with different tokenization.
- •Application of LLMs to perform downstream task like classification regression task(prediction) on genomic datasets.

What Next?



Machine learning models (LLMS)

MetaHipmer (MHM)

Metagenome Assembly

- • Test, evaluate and incorporate ML scaffolding method to MetaHipmer

Acknowledgement

- Steven Hofmeyr, LBNL
- Sustainable Research Pathways (SRP)
- Berkeley Lab
- Computer Language & System Software (CLaSS) group
- Andrew Tritt, Performance and Algorithm Group, LBNL

[1] Georganas, E. et al. Extreme scale de novo metagenome assembly. in SC18: International Conference for High Performance Computing, Networking, Storage and Analysis, 122–134 (IEEE, 2018).

[2] DNABERT-2: Efficient Foundation Model and Benchmark for Multi-Species Genome, https://doi.org/10.1093/bioinformatics/btab083