Design of a new read simulator to improve taxonomic profiling of metagenomic sequences using deep learning

Student: Eric Rangel, URI
ejrangel2@alaska.edu
Mentor: Cecile Cres, URI
cecile_cres@uri.edu
PI: Ying Zhang, URI
yingzhang@uri.edu

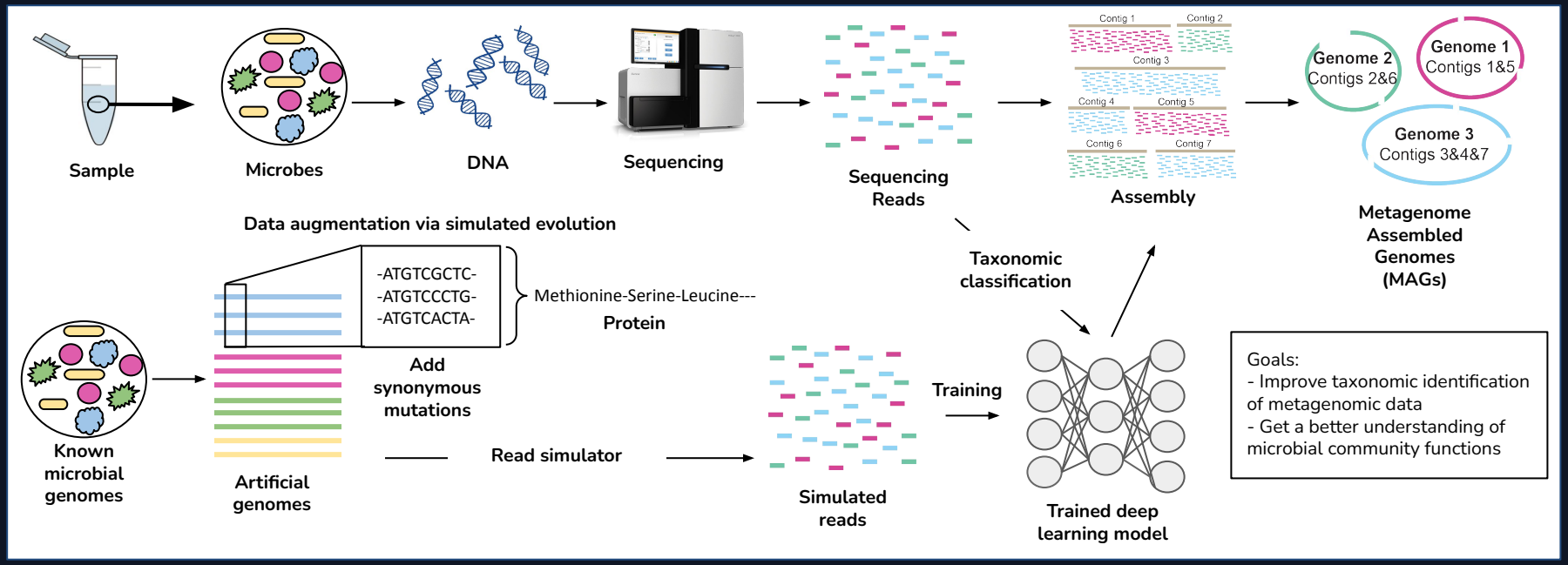
Date: January 12th, 2022

CAREERS CYBERTEAM

Design of a new read simulator to improve taxonomic profiling of metagenomic sequences using deep learning

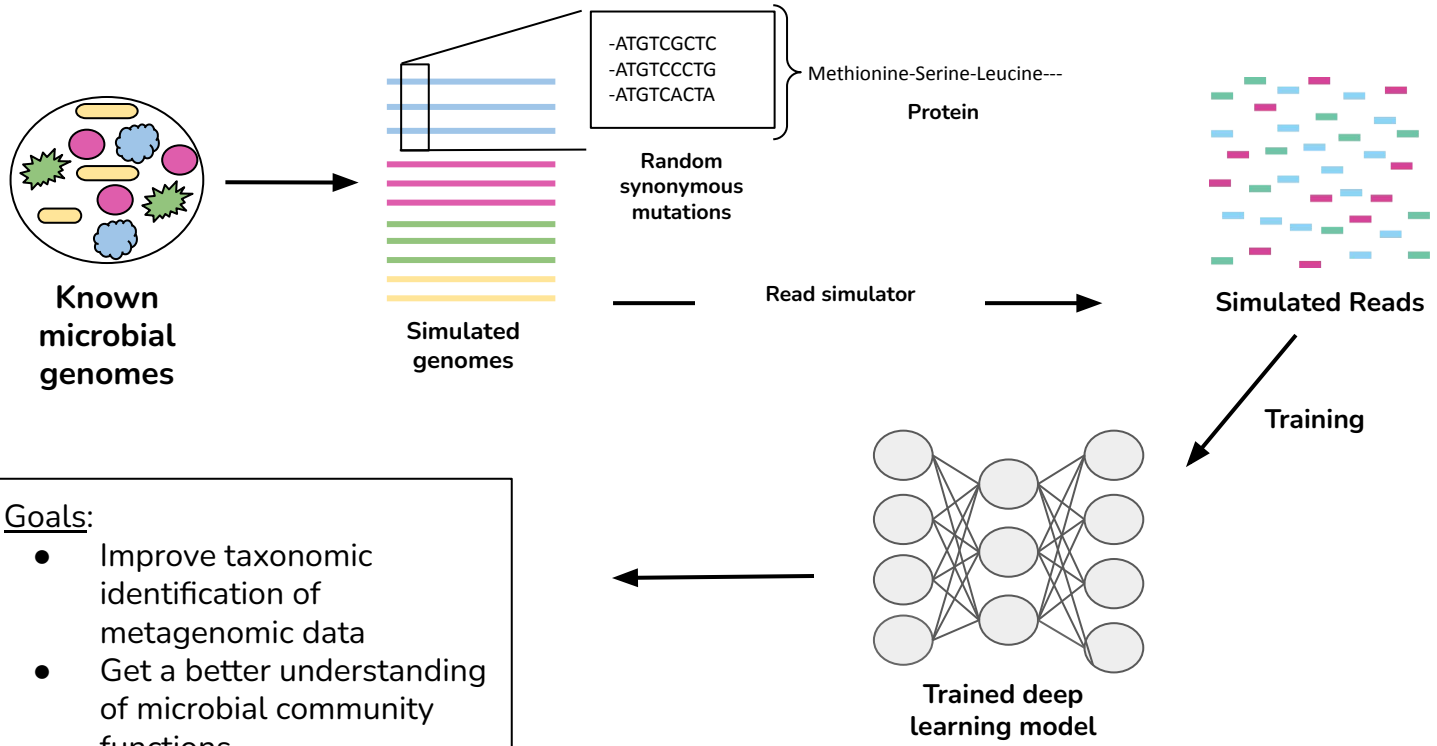
- Timeframe
  - June 14th, 2021
  - December 31st, 2021

Design of a new read simulator to improve taxonomic profiling of metagenomic sequences using deep learning



CAREERS CYBERTEAM

Design of a new read simulator to improve taxonomic profiling of metagenomic sequences using deep learning



Goals:

- Improve taxonomic identification of metagenomic data
- Get a better understanding of microbial community functions

CAREERS CYBERTEAM

Design of a new read simulator to improve taxonomic profiling of metagenomic sequences using deep learning

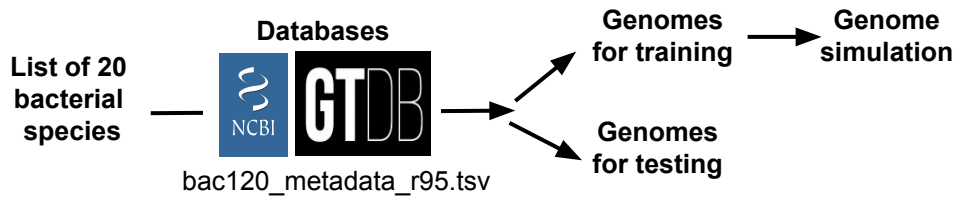
- Goals:

1. Use CCI (AiMOS, NPL clusters) for computational resources to implement a read simulator able to perform data augmentation via simulated evolution and train deep learning models
2. Evaluate the ability of the read simulator in generating datasets that can improve the performance of deep learning models in identifying unknown microbial genomes
3. Compare with deep learning models trained with datasets built using state-of-the-art read simulators

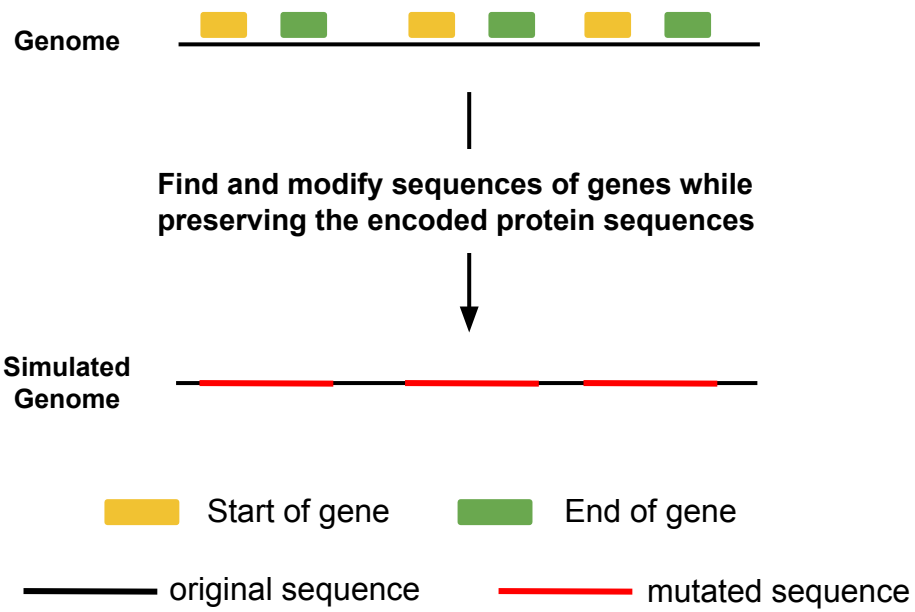
CAREERS CYBERTEAM

Goal 1: Read Simulator

1. Select bacterial genomes

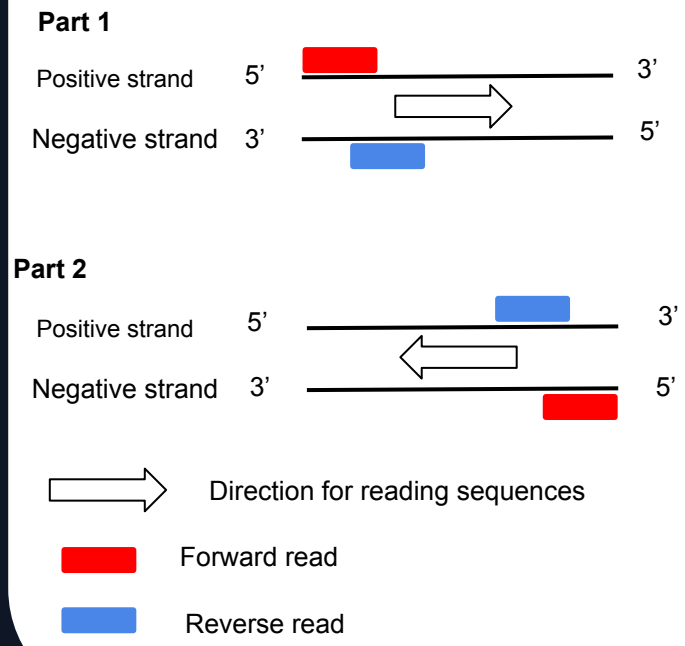


2. Find genes & Mutate Genomes



3. Simulate reads

Read length : 250 bp
insert size: 400 bp



CAREERS CYBERTEAM

Goal 2: Can our model accurately identify reads from other genomes that we didn't use for training?

Reads simulated with training genomes

Validation set

Number of reads: 5,327,060

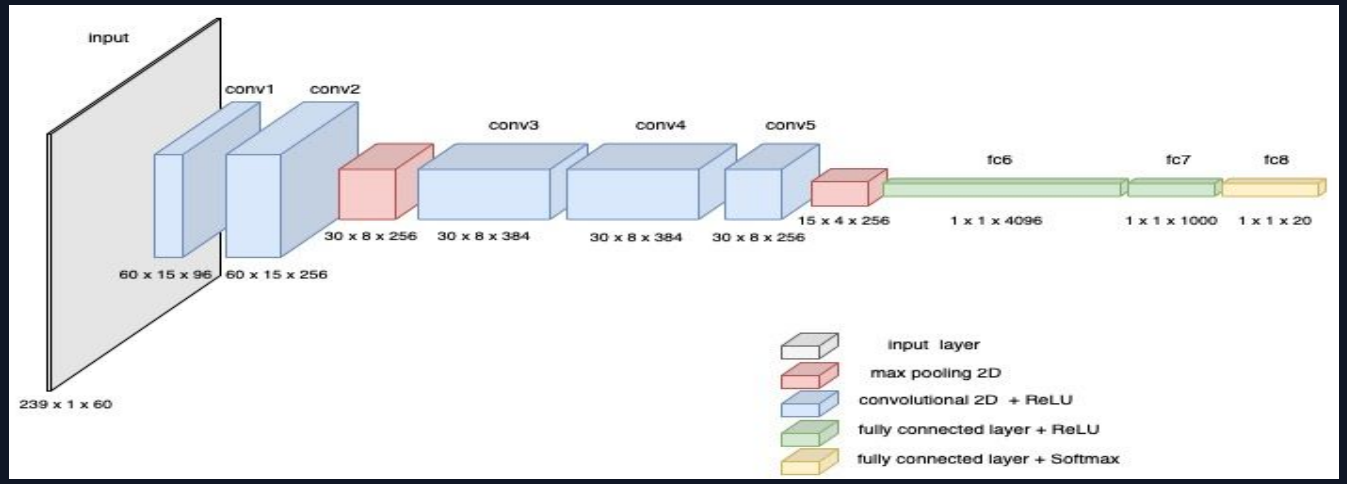
Evaluate model during training

Deep learning model

Training set

Number of reads: 12,430,940

Build deep learning model



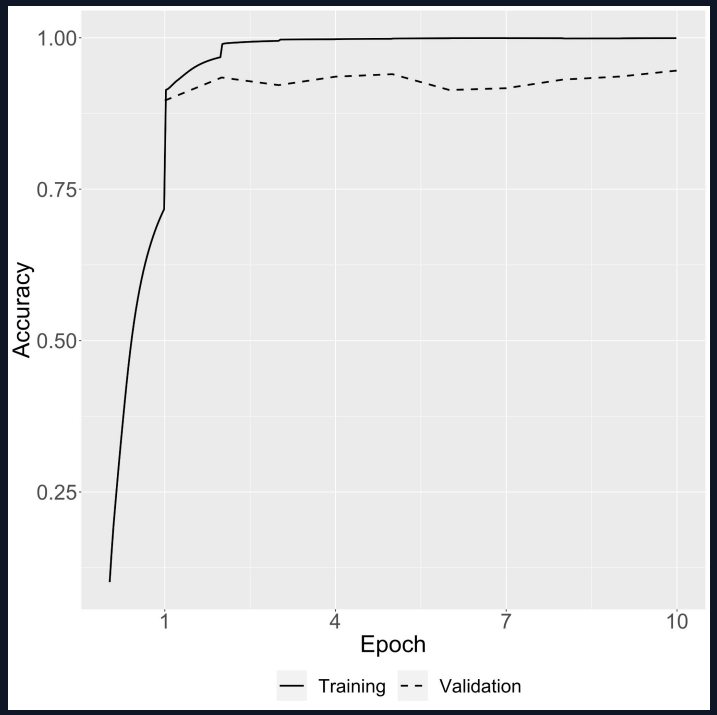
Deep Learning model architecture

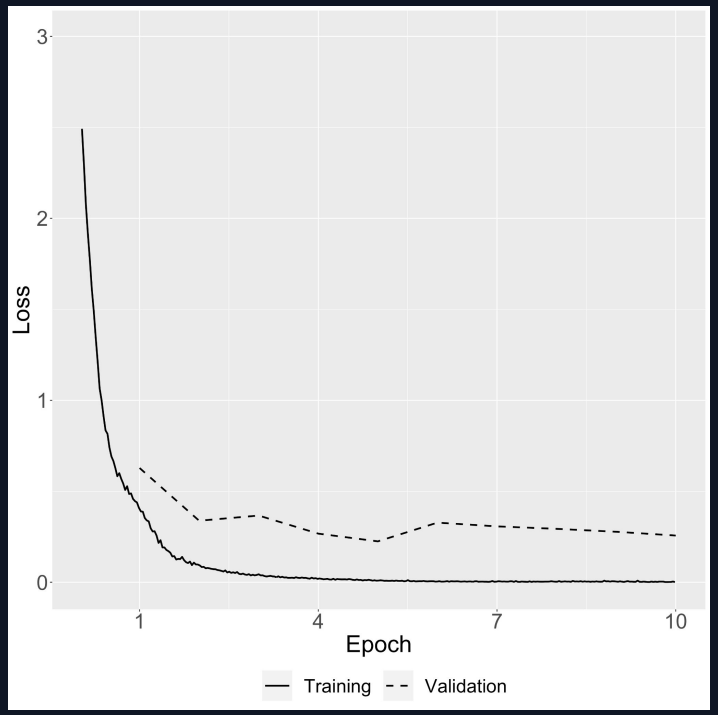
CAREERS CYBERTEAM

Goal 2: Can our model accurately identify reads from other genomes that we didn't use for training?

Training results

Learning curves





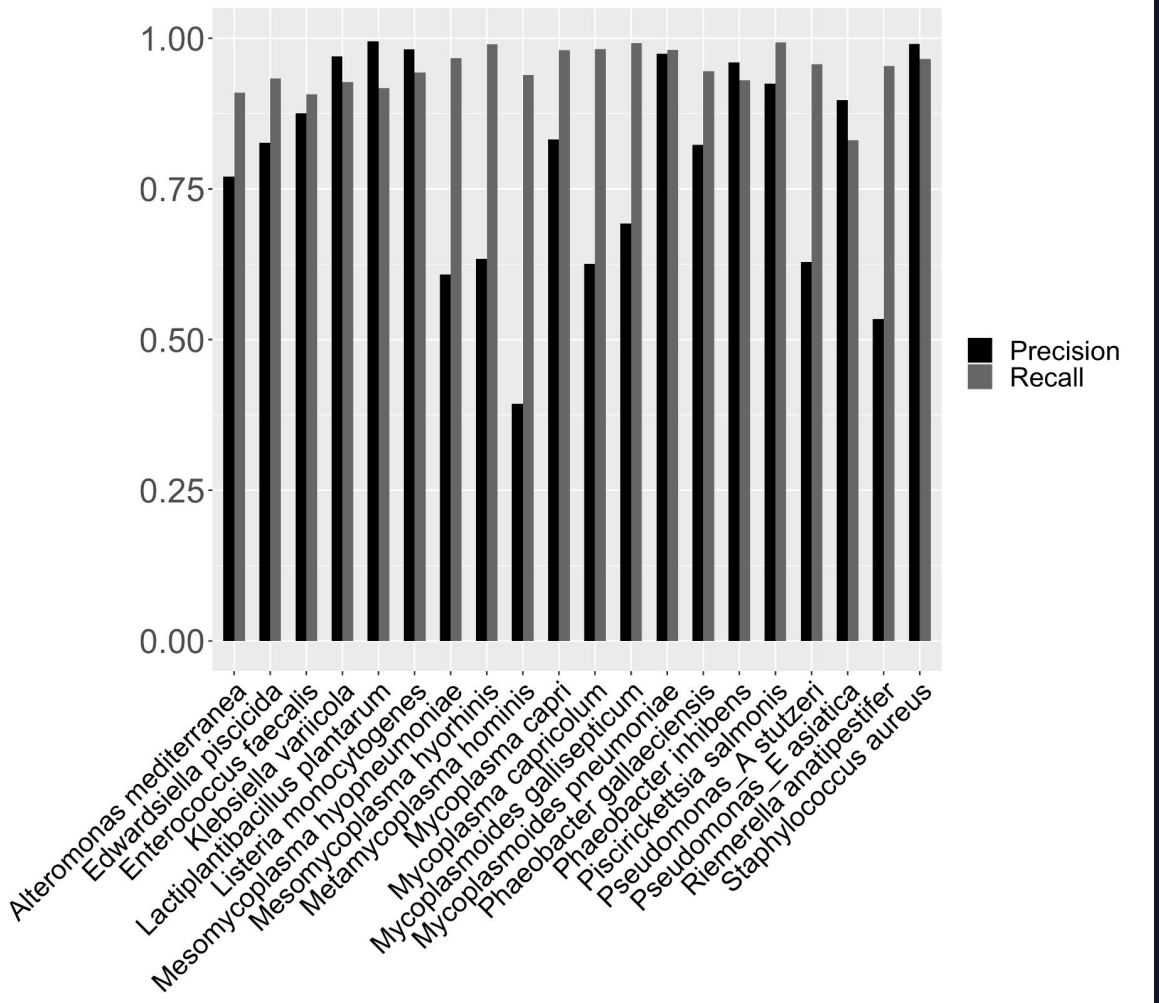
Cluster: Andromeda (URI)
Run time: 6h27
GPUs: 1
Training accuracy at epoch 10: 99.94%
Validation accuracy at epoch 10: 94.59%

CAREERS CYBERTEAM

Goal 2: Can our model accurately identify reads from other genomes that we didn't use for training?

Testing results

Number of reads in Testing set: 36,871,152



Cluster: Andromeda (URI)
Run time: 27 minutes
GPUs: 1
Testing accuracy: 94.93%

Precision = # reads of species *s* correctly classified / # reads of species *s*

Recall = # reads of species *s* correctly classified / (# reads of species *s* + # reads of species *s* incorrectly classified to other species)

CAREERS CYBERTEAM

What I learned?

- Python
  - Matplotlib
  - Biopython
  - Pandas
- Slurm Jobs
  - AIMOS/NPL (CCI) + Andromeda (URI)
  - Batch Scripts

- Machine Learning Theory
  - GANs
  - Tensorflow

CAREERS CYBERTEAM

- What went well?
  - Communication
  - Support Service for Systems at CCI
  - Organizations working together

- what could have gone better?
  - Documentation on Systems at CCI would be helpful so users will not rely heavily on Support Services

Publications/Contributions

- Github:

https://github.com/zhanglab/ReadsClassification/tree/Read_Simulator/Read_Simulator