Machine Learning and Population Genetics

Population genetics seeks to quantify the evolutionary events that led to the diversity we observe in present-day genomes (see these Population Genetics course notes for background information). This group of projects aims to use Machine Learning (ML) to infer these evolutionary events, from natural selection to population size changes, to migration, mutation, and recombination. One idea is to treat genetic data from many individuals as an "image" that can be fed into a convolutional neural network (CNN). See these CNN course notes for background information, and Flagel et al 2018 and Chan et al 2018 for population genetics applications. Recently my work has been leaning toward using generative adversarial networks (GANs) to simulate realistic data for understudied populations (see Wang et al 2021 and Riley et al 2024 - the first paper grew out of a senior thesis project!)

- Transformer architectures for genetic data. Transformers have been successful in many natural language applications (i.e. ChatGPT), but are in the very early stages of application to population genetic data. Transformers are naturally permutation-invariant, allowing them to be applied to an unordered collection of individuals. One idea is to create a novel transformer architecture for genetic inference. Potential applications include local ancestry inference (LAI), which is important for applying disease-association results to admixed populations. Other applications include natural selection and recombination rate variation.

- Domain-adaptation. Often in machine learning for population genetics, we train on simulations and test on real data. This presents a problem when the simulations are not very realistic. There are several ways to overcome this problem, including networks that seek to learn the difference between real and simulated data, then filter it out from inference results. GANs and weighted training examples are other options. This project aims to interpret and compare different approaches to simulation mis-specification.

- Interpretation. We have made progress interpreting CNN methods in terms of summary statistics (which are usually based on biologically relevant information). However, we don't yet understand exactly what CNNs are learning on their own. The filters and hidden layers of CNN methods for images often have a very nice, intuitive interpretation. A very interesting project would be to create a similarly intuitive framework for DNA data, possibly by feeding in the final hidden layers into an existing interpretation algorithm or using a "model-of-the-model" approach such as decision trees on top of a neural network architecture.

- Tree-based interpretability. Local tree inference is a huge area in population genetics, and results from tree-based inference are frequently fed into neural network methods. One project involves 1) understanding if CNN-based NNs are learning local tree summary stats and 2) assessing the impact of trees as a feature vs. raw genetic data.

- Privacy GAN. Generating synthetic data is very important in clinical settings where the real data cannot be released due to privacy concerns (i.e. exposing the disease status of specific individuals). This project seeks to create realistic simulated data that preserves privacy (note: this project builds upon previous work in the lab).

- GAN latent space. In my previous work on GANs, we find that some evolutionary parameters are more difficult to infer than others, and also that training sometimes gets "stuck" in regions of the parameter space that are not very realistic. A very interesting project could explore the latent space of evolutionary models through mathematical and/or visualization techniques, with the end goal of quantifying uncertainty and making GAN training more robust.

- Data representation. How to best represent genomic data being fed into a neural network is an open question. There are techniques for representing variation and sorting individuals that may improve and/or bias learning results. The goal of this project is to understand the relationship between data preprocessing, architecture choices, and inference results.

- Mosquito genetics. A more biological project involves applying GAN methods to mosquito data, to better understand mosquito demographic patterns and the effect on insecticide resistance. We now have mosquito data from many populations and climates in Africa, and the goal is to understand the timeline and scope of migration and population structure.

- Pangenome. Finally, if you're more interested in biology (and microbiology in particular), there is an option to investigate the "pangenome" (consisting of all genes within a species) by modeling core and accessory genes in a computational setting (co-advised by Professor Eric Miller).