Demographic Model Selection with Deep Learning

Ariella L. Gladstein[1*] and Daniel R. Schrider[1]

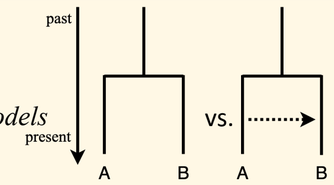
1) Department of Genetics, University of North Carolina, Chapel Hill, NC
aglad@med.unc.edu, https://github.com/agladstein

Abstract: One way to go about obtaining a better understanding of a population's demographic history is to design several competing models and see which best fits the data. We developed a method using a convolutional neural network (CNN) to choose the best demographic model using genome sequence data. We approach demographic model choice as a classification problem, where we train a CNN on simulated genome sequence data from a variety of models and parameter space to learn to recognize the genomic patterns resulting from each type of model. This allows us to input genome sequence from an unknown demographic history, and output posterior probabilities for each model the neural network was trained on. We also designed a feature vector of 30 genomic summary statistics to use in a random forest or approximate Bayesian computation. Overall, the random forest performs the best, while the CNN approaches the accuracy of the random forest with more training data. ABC, on the other hand, was unable to complete all the given classification problems due to linear dependency among the summary statistics.

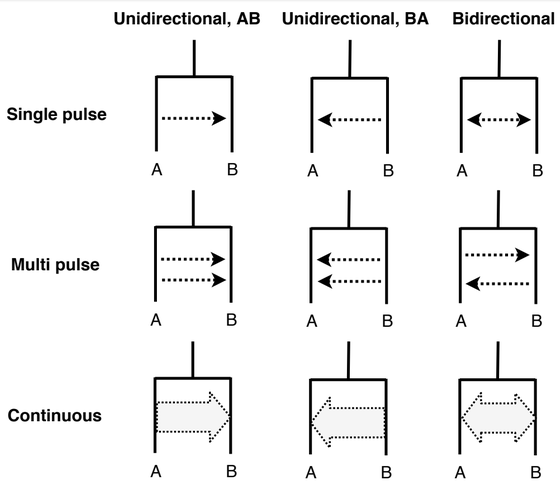
Motivation

- An important task of demographic inference is model selection.
- Model selection can be used to test evolutionary hypotheses regarding populations' histories.
- Current methods to choose the best model can be grouped into two general categories: likelihood-based and Approximate Bayesian Computation (ABC).
- Likelihood and ABC approaches vary in accuracy, ABC is more flexible than likelihood approaches, and both are computationally burdensome.

Goal:
Accurately classify demographic models, specifically geneflow models of varying complexity.



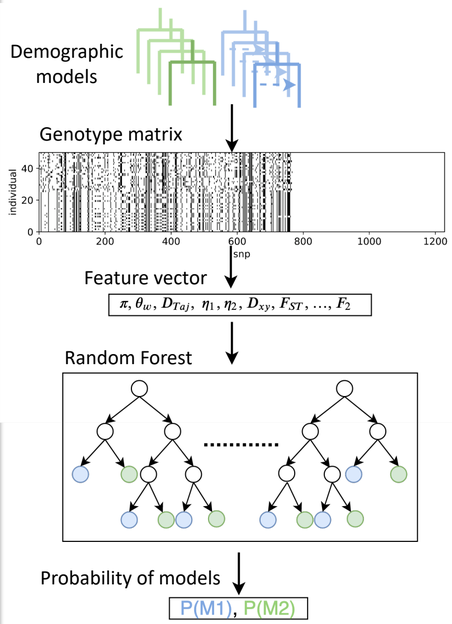
Geneflow Models



- Simulate models with human-like parameters:
  - Mutation rate = 1.2e-8, recombination rate = 1e-8
  - N_e parameters ~U(1000, 40000)
  - Time parameters ~U(1, 4000) generations
  - Migration rate ~U(0.1, 0.9)

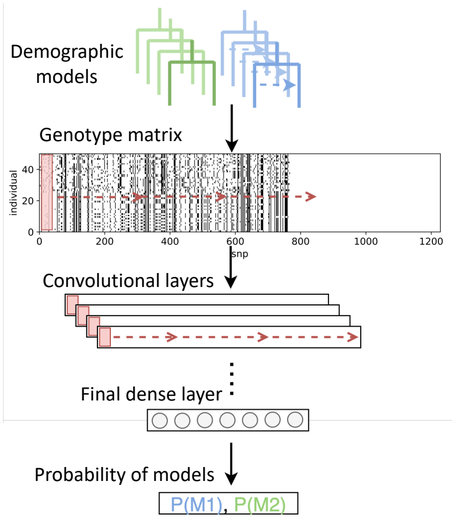
Methods

Random Forest



1. Simulate demographic models[1].
2. Calculate feature vector of summary statistics from genotype data. Feature vector includes 30 stats[2,3].
3. Train RF on feature vector from *n* simulations[4].
4. Predict the demographic model.

Convolutional Neural Network

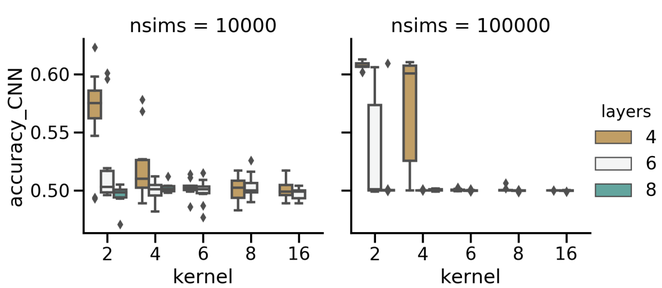


1. Simulate demographic models[1].
2. Train CNN on genotype matrix[4,5].
3. Predict the demographic model.

CNN architecture exploration

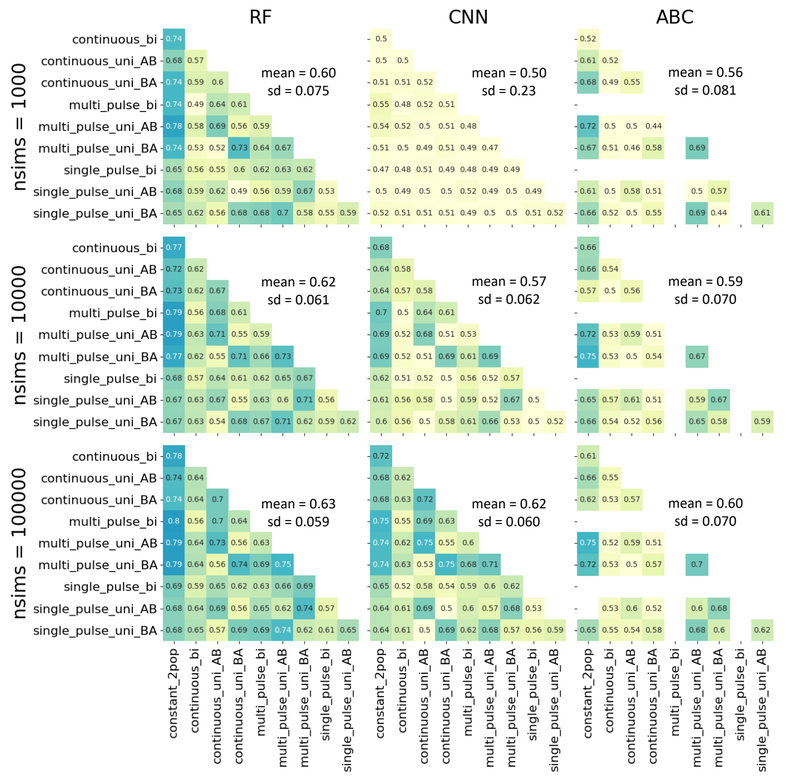
We explored a variety of architectures based on LeNet[7], VGG16[8], ReLERNN[9], including:

- batch size
- convolutional kernel size
- number of layers
- dropout rate
- pooling size
- dilation rate



Results

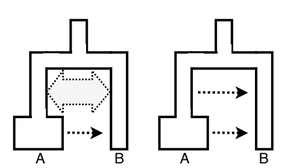
Accuracy of binary classification



- In general, RF has the best accuracy, even with few simulations.
- The CNN approaches the RF's accuracy with more training data.
- ABC has the worst accuracy and is sometimes unable to preform the regression step, due to linear dependence among the stats.

Future Directions

- Multiple loci
- More complex models
- Multi-way classification



References

1. Kelleher et al. (2016), Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes, PLoS Comput Biol 12(5): e1004842
2. Miles et al. (2019, April 26). cggh/scikit-allel: v1.2.0 (Version v1.2.0). Zenodo.
3. Schrider et al. (2018) Supervised machine learning reveals introgressed loci in the genomes of *Drosophila simulans* and *D. sechellia*. PLOS Genetics 14(4): e1007341
4. Pedregosa et al. (2011) Scikit-learn: Machine Learning in Python. JMLR 12, pp. 2825-2830.
5. Flagel et al. (2019). The Unreasonable Effectiveness of Convolutional Neural Networks in Population Genetic Inference. *MBE*, *36*(2), 220–238.
6. Chollet et al. (2015) Keras. https://keras.io.
7. Lecun et al. (1998). Gradient-Based Learning Applied to Document Recognition. Proceedings of the IEEE.
8. Simonyan & Zisserman (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. Arxiv, 1–14.
9. Adrion et al. (2019) Inferring the landscape of recombination using recurrent neural networks. bioRxiv 662247.