Deep learning for genetic ancestry inference
Jan Sokol [jsokol], Matthew Aguirre [magu] (at) stanford.edu

CS230, WINTER 2020, STANFORD UNIVERSITY

Abstract
In genomics, local ancestry inference (LAI) is used to estimate the ancestral composition of a genomic sequence at high resolution. Here, we describe an approach to LAI which leverages deep learning techniques developed for image segmentation. We consider two formulations of the ancestry inference problem — namely, local and global inference — and benchmark our algorithms using real and simulated genotype data from the 1000 Genomes Project.

Problem Formulation
As genomic samples have become more diverse, LAI has emerged as a key processing step in ancestry and disease association studies. The input is a set of genetic variants and the output is an ancestry assignment, either for the entire sample (global ancestry) or as a segmented mask over individual base pairs (LAI; Figure 1). In this project we use global ancestry prediction as a stepping stone to LAI.

Dataset
We use a sample of n=5,008 haplotypes from 2,504 individuals in the 1000 Genomes Project (1KG) [2], which collected whole genome sequences of individuals across 26 world populations (Table 1). We use a sample of n=5,008 haplotypes from 2,504 individuals in the 1000 Genomes Project (1KG) [2], which collected whole genome sequences of individuals across 26 world populations (Table 1). We formulate the global ancestry problem as

Experiments
• As initial validation of our approach, we trained a model with three fully connected layers to predict global continental ancestry from a subsample of 500 genetic variants (approximately the window size of RFMix). Internal layers were of size 500 and 30 with ReLU activation, followed by an output layer of size 5 (softmax activation). This model interpolated the training set of 4,000 randomly chosen haplotypes and generalized well to the remaining 1,008 test samples (99.7% and 82% accuracy), and was quite robust to the sizes and activations of the internal layers.

• We then built a CNN model of global ancestry which has an input conv layer with filters of size/stride 512 with same padding, followed by a smoothing convolutional layer with size 64, a stride of 4, and valid padding; followed by two fully connected layers to predict the output. We trained this model using four loss functions: (1) categorical cross entropy loss; (2) entropy loss weighted by the distance between true and predicted output or (3) inverse distance; and (4) Haversine distance, where we predict coordinates of origin for each sample.

• Finally, we implemented a model with U-Net architecture for LAI based on a publicly available github repository [3]. This choice was informed by the high performance of the CNN model, and because U-Nets have been shown to work well for segmentation tasks such as this one [4]. Though this model fit the training set well, it failed to generalize to a holdout test set (data not shown); additional work to extend this architecture to a valid LAI model is required.

Summary
• Here, we demonstrate high fidelity global ancestry prediction from the equivalent of one chromosome of array genotypes.
• We formulate the global ancestry problem as multi-class labeling and as coordinate regression.
• Multi-class classification is accurate up to country/region (e.g. CHS/CHB are both Chinese; ITU/STU are from southern Europe). Weighting loss by inverse distance improves accuracy between nearby populations likely to share similar genetic signatures.
• While less accurate, coordinate predictions recapitulate the human migratory history of admixed groups: CEU individuals in the test set are scattered across the Atlantic between mainland USA and northern Europe; likewise for other American groups (like PUR) and western Africa.

Future Directions
• Work remains to make a U-Net model for LAI viable; options include:
  1. Additional data augmentation to reduce overfitting
  2. Using a wider set of genetic data (e.g. all chromosome 1)
  3. Alter model hyperparameters, potentially including dropout at higher layers in the U-Net
• Use of our coordinate-based predictions in population-genetic studies may be of particular interest to domain experts.

References and Acknowledgements
We would like to thank Alex Ioannidis [ioannidis] and Daniel Mas Montserrat for their contributions to the ideation of this project. We have submitted all of our code on gradescope.