Deep learning for genetic ancestry inference

Jan Sokol [jsokol], Matthew Aguirre [magu](at)stanford.edu

Link to introductory video: https://youtu.be/t-yDAD6sPCU

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.

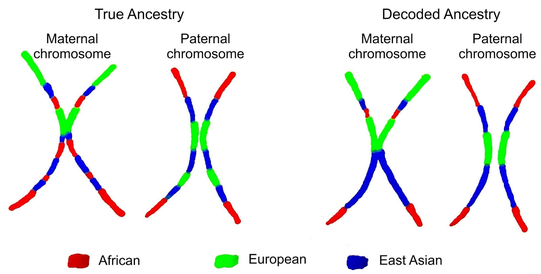

Figure 1: Pictorial representation of LAI. Ground truth (left) and decoded ancestry (right) across segments of a pair of chromosomes.

The current gold standard for LAI, RFMix [1], uses random forests to estimate parameters of a conditional random field model of ancestry in genomic windows of size 400kb (400,000 base pairs).

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).

Population	Code	Color in fig 3
Sri Lankan Tamil in the UK	STU	
Toscani in Italy	TSI	
Punjabi in Lahore, Pakistan	PJL	
Japanese in Tokyo, Japan	JPT	
Chinese Dai in Xishuangbanna, China	CDX	
Utah residents (CEPH) with Northern and Western European ancestry	CEU	
Han Chinese in Beijing, China	CHB	
Gujarati Indians in Houston, TX	GIH	
African Ancestry in Southwest US	ASW	
Gambian in Western Division, The Gambia - Mandinka	GWD	
Luhya in Webuye, Kenya	LWK	
Iberian populations in Spain	IBS	
Colombian in Medellin, Colombia	CLM	
Finnish in Finland	FIN	
Puerto Rican in Puerto Rico	PUR	
Mende in Sierra Leone	MSL	
Bengali in Bangladesh	BEB	
Esan in Nigeria	ESN	
Mexican Ancestry in Los Angeles, California	MXL	
Kinh in Ho Chi Minh City, Vietnam	KHV	
African Caribbean in Barbados	ACB	
Peruvian in Lima, Peru	PEL	
Han Chinese South	CHS	
Yoruba in Ibadan, Nigeria	YRI	
Indian Telugu in the UK	ITU	
British in England and Scotland	GBR	

Table 1: Sample locations and and three-letter codes of 1KG populations, with our own color labels for each group.

We here consider a subset of p=57,876 variants on Chromosome 1 (downsampled for computational tractibility) for ancestry inference.

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.

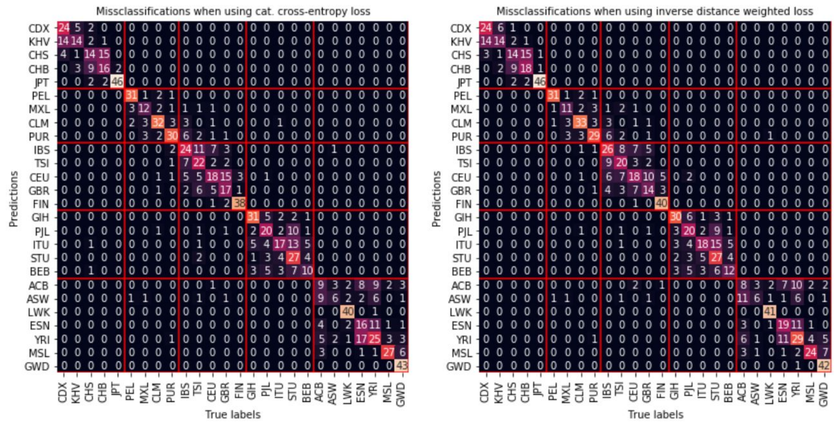

Figure 2: Confusion matrix for CNN with categorical cross-entropy loss (L) and inverse distance weighted cross-entropy loss (R) .

- 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; cross-entropy loss weighted by (2) the distance between true and predicted output or (3) inverse distance; and (4) Haversine distance, where we predict coordinates of origin for each sample.

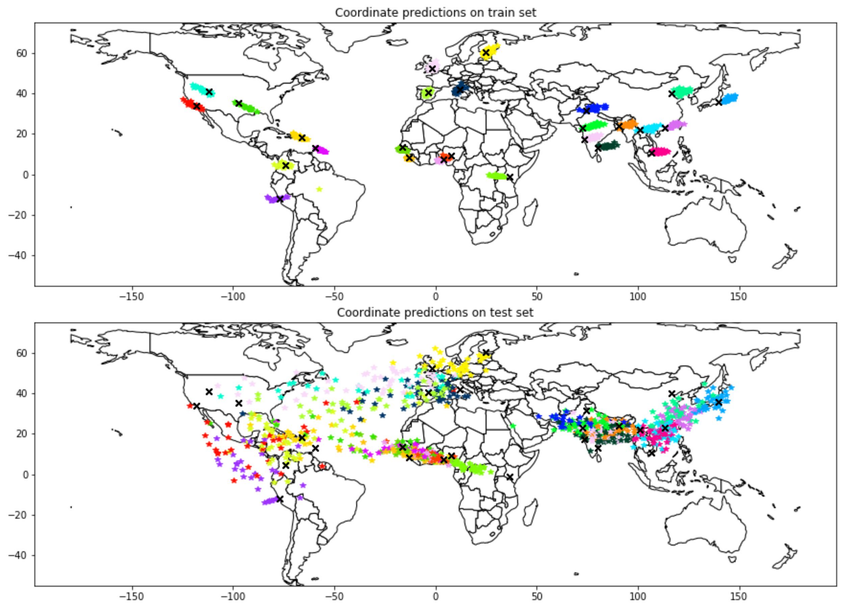

Figure 3: Geographic predictions for training (top) and test (bottom) sets from CNN model of global ancestry, formulated as coordinate regression.

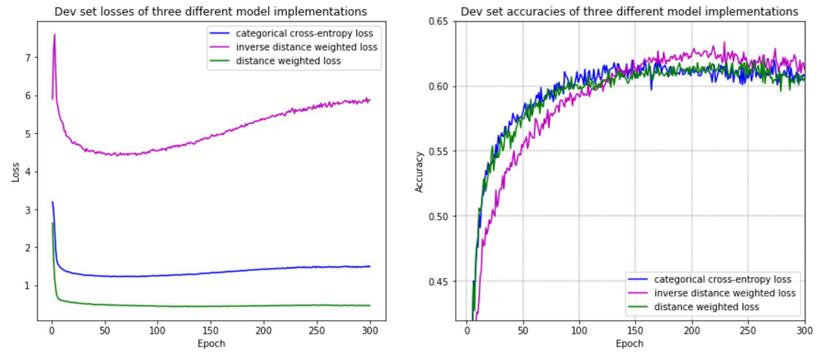

Figure 4: Training and test loss (L) and accuracy (R) for global ancestry model with unweighted, distance, and inverse-distance weighted loss.

- 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 doman experts.

References and Acknowledgements

[1] Maples B., *et. al.*, *Am J Hum Genet.*, 2013. (PMC3738819).
[2] The 1000 Genomes Project Consortium, *Nature*, 2015. (PMC4750478).
[3] https://github.com/zhixuhao/unet
[4] Ronneberger, *et. al.*, *MICCAI*, 2015. (arXiv:1505.04597).

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.