CS230

Deep Learning for Local Ancestry Inference

Jan Sokol
Biomedical Informatics Training Program
Stanford University
jsokol@stanford.edu

Matthew Aguirre
Biomedical Informatics Training Program
Stanford University
magu@stanford.edu

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.

1 Introduction

Local ancestry inference (LAI), also known as ancestry deconvolution, is used to estimate the ancestral composition of genomic sequences at the resolution of individual base pairs. As human genetic studies have grown in size and scope to accommodate increasingly diverse samples, LAI has emerged as a critical step for analyses ranging from genome-wide association studies to the inference of human population history.

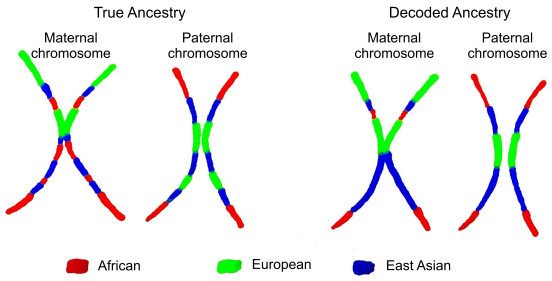

Figure 1: Pictoral representation of ancestry deconvolution. The left chromosome pair shows the ground truth ancestry of each genetic segment. The right chromosome pair represents hypothetical inferred ancestries of each genetic segment.

The input for LAI is a human genome sequence, and the output is a masked annotation of each position of the sequence which indicates its population of origin (e.g. African, European). At various stages of this project, we also consider the "global ancestry" problem, which has the same input as LAI (i.e. a genome) but only seeks to label the population of origin for the entire sample. The key difference here is the admixture assumption: in the case of global ancestry, individuals are assumed to belong to one population, whereas LAI explicitly considers admixed individuals whose ancestry comprises multiple populations. This results in a significant dimensionality difference for these two models: global ancestry is a single-class output, but LAI is a mask of size p (genome length).

CS230: Deep Learning, Winter 2018, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

In general, analyses of genomic data are subject to the classic $p \gg n$ problem, as the human genome is 3 billion base pairs in length. However, not all genomic loci vary in humans, with $\sim 99\%$ of sequence shared between individuals. Variants near one another on chromosomes can also be highly correlated, sometimes in population-dependent ways. In the case of LAI, the modeling task is further complicated by variation which is ubiquitous in humans but does not encode information useful to infer ancestry (i.e. the junk feature problem).

2 Related work

Prior work on ancestry inference has relied heavily on Hidden Markov Models (HMMs), though these models have evolved in complexity as genotyping technologies have matured. The earliest tool for global ancestry, STRUCTURE [1], used a model-comparison approach to assess the likelihoods of samples originating from the one or several populations based on a set of unlinked genotypes; a similar Bayesian method was also considered [2].

This tool was later extended to account for correlation across genetic variants by adding a "linkage" model [3], which allowed for local ancestry estimation. Other similar HMM-based models were also developed, with specific considerations for trans-ethnic mapping of disease genes [4], variable admixture times [5], varied geographic population distributions [6], or which permit the use of fine-scale reference panels of genetic variation [7], or allow much faster computation (ADMIXTURE) [8]. These approaches to the LAI problem have been extensively reviewed [9].

The current gold standard tool for LAI in research settings, RFMix [10], uses independently trained random forest models to predict ancestry within genomic windows of size $\sim 400kb$ (400,000 base pairs). For computational tractability, these random forests estimate parameters of a conditional random field model of ancestry within each window, rather than predicting ancestry directly.

3 Dataset and Features

In this work we make use of a reference dataset of genetic variation called the 1000 Genomes Project (1KG) [11]. This dataset contains the whole-genome sequences of $n = 2,504$ individuals in 29 distinct world population groups (e.g. "Northern Europeans in Utah", or "Mende in Sierra Leone"). For our analysis, we have an effective $n = 5,008$ phased haploytpes in 1KG, as every individual has two copies of each chromosome.

As genomic data are very wide ($p = 81,271,745$ over the entire 1KG cohort), we work with two subsets of the human genome to speed up computation. The first is a subset of $p_1 = 57,876$ variants on chromosome 1 which are present on the genotyping array used in a large population cohort study in the UK [12]. Microarrays are an affordable genotyping technology which assay $\sim$1 million genomic variants; they are commonly used in large-scale genetic studies and by direct-to-consumer genetic testing companies [13] [14]. The second subset is the entirety of chromosome 21, which contains $p_2 = 1,105,538$ genetic variants. Results in this document are from the first subset only.

We also collected augmented data resulting from simulating admixture between individuals. Each simulated genome is an approximation to that of an individual with diverse ancestral background (e.g. father from Europe, mother from Africa). These genomes are created by one of two approaches: (1) naively stitching together genotypes of non-admixed individuals, with the number of stitch points sampled as a Poisson variable with rate proportional to the number of generations of mixing [15]; or (2) with the `msprime` [16] software, which simulates genotypes from genealogical trees sampled according to a coalescent model due to Hudson [17].

4 Methods

4.1 Predicting global continental ancestry with a small FCNN

To ensure that our subsamples of genomic sequence contain sufficient information to predict ancestry, we first implemented a small neural network consisting of three fully-connected layers. Our "small-net" took as input a contiguous block of 500 genomic variants. The alleles of each variant (e.g. A vs. T) were one-hot encoded, and passed to a fully connected layer of 500 nodes (followed by ReLU

activation); a second fully connected layer of 30 nodes (ReLU activation); and a final, fully connected output layer of five nodes (softmax activation).

Our small net has five output nodes since there are five continental ancestry groups in our data: AFR (Africa), AMR (Americas), EAS (East Asia), EUR (Europe), and SAS (South Asia). This network was trained on a random subsample of 4,000 haplotypes with categorical cross-entropy loss, and evaluated on the remaining 1,008 in a train-test holdout design. Code and results for this model are in `model1.ipynb`, which is implemented primarily in `tensorflow` [18], with some tools for model assessment borrowed from `scikit-learn` [19].

4.2 Global ancestry prediction with a CNN

We have also implemented a convolutional neural network (CNN) for global ancestry prediction in Keras [20]. This network consists of two convolutional layers, followed by a fully-connected layer, then an output layer to predict either continental (5 labels) or population-level (26 labels) ancestry.

In light of the large window size used by RFMix, we chose an initial filter size of 512 with a (very large) stride of 256. Given the relative sparsity of genetic variation, we found 64 filters to be a sufficient cover of the likely landscape of variation in a window of this size. Likewise, for the second convolutional layer we chose 32 filters of size 64 with a stride of 4, to account for possible longer-range correlation across the chromosome. As is standard in CNNs for imaging-type tasks, we follow the convolutional layer with a fully-connected layer prior to the output layer. Given the size of the input from the convolutional layers, we decided to have 64 nodes in this layer.

Through ad-hoc experimentation we found that global ancestry prediction is quite robust to these parameter choices, including removal of the fully-connected layer prior to output. This is not too surprising since genetic variants differ in frequency across populations to the point where simple linear models, such as those in early models of global ancestry [1] [3], can perform quite well.

4.3 Generalized loss functions for Global ancestry prediction

We also experimented with creating a loss function that penalizes misclassifications based on the extent to which the prediction and the ground truth are related. For example, predicting that a sample from Finland is from China would be more wrong than predicting that it is from somewhere else in Europe. We therefore implemented a customized loss function that penalized misclassifications proportionally to the great-circle distance between the ground truth label and the prediction.

To obtain a second measure of the extent to populations in our dataset are related, we used the total genetic variance contained in each population relative to the total genetic variance between any two given populations, denoted F_{ST}, as computed by the 1000 Genomes Project Consortium [21]. We concluded that F_{ST} may serve as a good proxy for the genetic relatedness of any two populations in our dataset. Consequently, we implemented a second model that penalized misclassifications using the F_{ST} score (see `CNN_Global.ipynb`).

We further experimented with formulating global ancestry as regression, using the geographic coordinates of origin for each of the 1KG populations. For this task, we treat latitude and longitude as separate output parameters in the interval $[0, 1]$ and then scale the resulting output to $[-\pi/2, \pi/2]$ for latitude and $[-\pi, \pi]$ for longitude. Since the distance between two coordinates on the Earth is an arc on a great circle, we use the Haversine distance as our loss function for a single sample:

$$\ell(y, \hat{y}) = arcsin\left(\sqrt{sin\left(\frac{\phi_i - \hat{\phi}_i}{2}\right)^2 + cos(\phi_i)cos(\hat{\phi}_i)sin\left(\frac{\theta_i - \hat{\theta}_i}{2}\right)^2}\right)$$

where $y = (\phi, \theta)_i$ are the latitude/longitude for sample i. The architecture of the CNN otherwise remained the same as above (see `CNN_Global_Haversine_v2.ipynb`).

4.4 Local ancestry inference with a U-Net

To generalize our approach to the LAI task, we re-implemented a publically available U-Net architecture (`https://github.com/zhixuhao/unet`), which has been shown to perform well at

segmentation tasks [22]. This model consists of five "downward" convolutional layers with max-pooling, and five "upward" convolutional layers with up-sampling and feed-forward links from earlier layers in the U-shape. We naively kept the hyperparameters as-is in the public implementation (up to changes necessary to accommodate a one-dimensional input), and refer the interested reader to the GitHub and reference linked above for more information on this architecture (see `CNN_LAI.ipynb`).

5 Experiments/Results/Discussion

5.1 Toy example: Predicting continental ancestry from a window of genomic sequence

We first decided to check whether our subsampled data contained sufficient information to predict ancestry. To accomplish this, we built a global ancestry model using a small window consisting of the first 500 variants on chromosome 1, which roughly corresponds to the window size used by RFMix at the density of our subsample ($\sim$1 variant per $1.2kb$). We trained a fully connected neural network (FCNN; see Methods) on a random subsample of 4,000 haplotypes from 1KG and tested on the remaining 1,008. As this is a proof of concept experiment, we decided against using an additional holdout validation set. We found that the FCNN was able to interpolate the training set within a few dozen epochs (99.75% training accuracy), and that its predictions generalized reasonably well to the test set (82.7% accuracy). This suggested that genomic windows of approximately 500 variants contain sufficient information to predict the local ancestry of individuals across an entire chromosome, at or near this resolution.

5.2 A next step: Predicting global ancestry with an entire chromosome

Given the success of our small FCNN at predicting global ancestry in a small window, we decided to use a convolutional neural network (CNN) architecture for the global/local ancestry problem. We use the same evaluation framework for this model as for the FCNN, training on a random sample of 4000 1KG haplotypes and testing on the remainder. In the case where we predict continental ancestry, the CNN is also able to rapidly interpolate the training set and achieve excellent performance on the test haplotypes ($\sim$98.5% classification accuracy), with most errors due to mislabeling American samples (AMR) as African (AFR) or European (EUR), or vice versa.

In the case where we predict population of origin (26 output classes rather than 5), we see significantly reduced test set accuracy ($\sim$60%). However, misclassifications rarely occur outside continental ancestry blocks (see figure 3), with many errors owing to non-identifiability of nearby populations (e.g. Yoruba [YRI] and Esan [ESN], both from Nigeria).

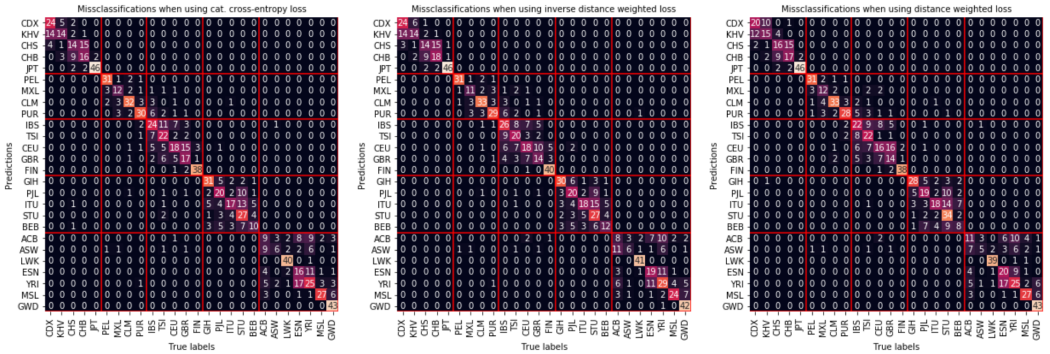

Figure 2: Confusion matrices for population-level classification task with categorical cross-entropy loss (Left), inverse distance weighted loss (Center), and distance loss (Right). Populations are grouped by continent, and separated by red bars.

To try to improve the performance of our model, we implemented a custom loss function that penalizes misclassifications based on the distance between the prediction and the ground truth (measured either by distance or genetic relatedness; see Methods). When we used the great-circle distance as a proxy for the genetic relatedness, our model's accuracy decreased. However when we used the relative

genetic variation between any two populations (F_{ST}), our global ancestry predictions improved slightly (from about 61% to about 62%).

We implemented one model in which misclassifications of closely related populations are highly penalized (i.e. by inverse distance), and another in which misclassifications of closely related population are penalized less (i.e. proportional to distance. Interestingly, accuracy only increased for the first model. As expected, our model that penalized misclassifications of closely related populations more was slightly better at distinguishing between closely related populations (Figure 2).

5.3 Generalizations: Coordinate loss and a U-Net for LAI

Given the geographic diversity of the 1KG populations and the relative similarity of neighboring populations, we decided to consider ancestry prediction as a regression problem by having the model output the latitude and longitude coordinates of each sample. Unfortunately, we found that Haversine loss performs worse compared to mean squared error over the coordinates (data in Jupyter notebook). However, we are able to predict ancestries quite well overall (Figure 3); though there are noticeable differences in training and test set performance, this is likely an accurate reflection of reality as some populations (e.g. Europeans in North America – teal in Figure 3) should be predicted as the midpoint of their geographic source and ancestral geography. Viewed through this interpretation, the apparent poor performance of our model on the test set actually reflects the geographic migratory history of each of these populations (e.g. European migration and African slave trade to the Americas).

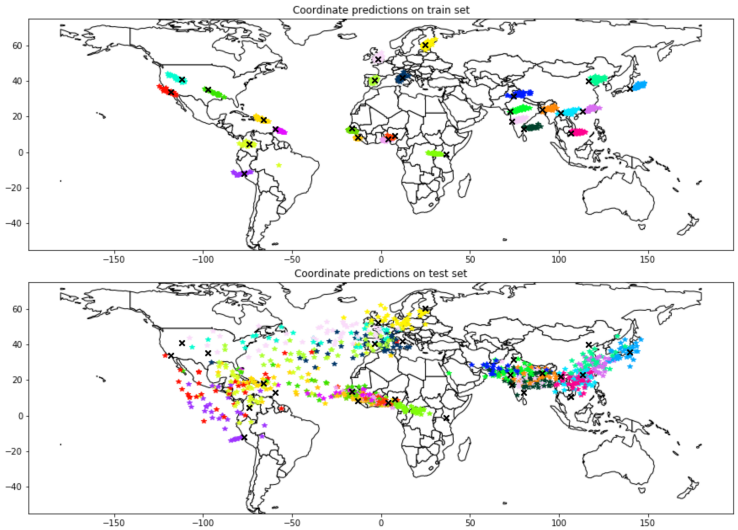

Figure 3: Coordinate predictions for genetic samples using the Haversine loss model.

To extend our CNN architecture to the local ancestry task, we also implemented a U-Net [22], which performs well for image segmentation tasks. Since LAI is essentially a 1D segmentation problem, we hoped this model would perform well here; however, we found that this model failed to learn, and instead converged on predicting one ancestry at all sites (data in Jupyter notebook).

6 Conclusion/Future Work

Here, we present an application of deep learning genetic ancestry inference. Our CNN model discriminates global ancestry at regional resolution from the equivalent of one chromosone of array genotyped genetic data. When formulated as coordinate regression, our model remains predictive and and recapitulates the migratory history of admixed populations in the Americas.

However, significant work remains to translate these successes into a viable model for LAI. Future directions for U-Net development include (1) hyperparameter tuning to avoid local minima; (2) using a wider set of data (e.g. all chromosome 1) to model genetic data at finer resolution; (3) further data augmentation to reduce overfitting. We suspect that our findings will be of interest to the population genetics community, and we will pursue further development of our model.

7 Contributions

J.S. implemented the FCNN and extended the CNN loss function to include misclassification weights. M.A. implemented the CNN and U-Net. J.S. and M.A. jointly developed the Haversine loss function, performed testing of all models, and co-wrote the manuscript. J.S. and M.A. would like to acknowledge Alexander Ioannidis (`ioannidis@stanford.edu`) and Daniel Mas Montserrat for assistance with the data collection and ideation for this project.

All of our code is submitted on gradescope.

8 Appendix

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	

Figure 4: For brevity, we refer to the 1000 Genomes populations by their canonical three letter codes in the main text of this paper – we here include their full names and countries of origin as reference.

References

[1] Jonathan K Pritchard, Matthew Stephens, and Peter Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.

[2] Hua Tang, Jie Peng, Pei Wang, and Neil J Risch. Estimation of individual admixture: analytical and study design considerations. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 28(4):289–301, 2005.

[3] Daniel Falush, Matthew Stephens, and Jonathan K Pritchard. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164(4):1567–1587, 2003.

[4] Nick Patterson, Neil Hattangadi, Barton Lane, Kirk E Lohmueller, David A Hafler, Jorge R Oksenberg, Stephen L Hauser, Michael W Smith, Stephen J O'Brien, David Altshuler, et al. Methods for high-density admixture mapping of disease genes. *The American Journal of Human Genetics*, 74(5):979–1000, 2004.

[5] Hua Tang, Marc Coram, Pei Wang, Xiaofeng Zhu, and Neil Risch. Reconstructing genetic ancestry blocks in admixed individuals. *The American Journal of Human Genetics*, 79(1):1–12, 2006.

[6] Sriram Sankararaman, Srinath Sridhar, Gad Kimmel, and Eran Halperin. Estimating local ancestry in admixed populations. *The American Journal of Human Genetics*, 82(2):290–303, 2008.

[7] Alkes L Price, Arti Tandon, Nick Patterson, Kathleen C Barnes, Nicholas Rafaels, Ingo Ruczinski, Terri H Beaty, Rasika Mathias, David Reich, and Simon Myers. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS genetics*, 5(6):e1000519, 2009.

[8] David H Alexander, John Novembre, and Kenneth Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome research*, 19(9):1655–1664, 2009.

[9] Yushi Liu, Toru Nyunoya, Shuguang Leng, Steven A Belinsky, Yohannes Tesfaigzi, and Shannon Bruse. Softwares and methods for estimating genetic ancestry in human populations. *Human genomics*, 7(1):1, 2013.

[10] Brian K Maples, Simon Gravel, Eimear E Kenny, and Carlos D Bustamante. Rfmix: a discriminative modeling approach for rapid and robust local-ancestry inference. *The American Journal of Human Genetics*, 93(2):278–288, 2013.

[11] 1000 Genomes Project Consortium et al. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061, 2010.

[12] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, et al. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018.

[13] Catherine A. Ball, Jake K. Byrnes, Daniel Garrigan, Eurie Hong, Keith Noto, Josh Schraiber, Alisa Sedghifar, Shiya Song, Barry Starr, David Turissini, and Yong Wang. Ethnicity estimate white paper. Technical report, AncestryDNA, 2013.

[14] Eric Y. Durand, Chuong B. Do, Joanna L. Mountain, and J. Michael Macpherson. Ancestry composition: A novel, efficient pipeline for ancestry deconvolution. Technical report, 23andme, 2014.

[15] Simon Gravel. Population genetics models of local ancestry. *Genetics*, 191(2):607–619, 2012.

[16] Jerome Kelleher, Alison M Etheridge, and Gilean McVean. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS computational biology*, 12(5), 2016.

[17] Richard R Hudson. Properties of a neutral allele model with intragenic recombination. *Theoretical population biology*, 23(2):183–201, 1983.

[18] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[20] François Chollet et al. Keras. `https://keras.io`, 2015.

[21] 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526:68–74, 2015.

[22] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.