CS230

CS230 Final Project

Hodan Farah and Zouberou Sayibou*
Department of Computer Science
Stanford University
hfarah@stanford.edu, zouberou@stanford.edu

Abstract

As climate change continues to worsen at quickening rates, we are interested in how plants and crops such as corn and cassava which humans rely heavily on may adapt. It known that genomic variation leads some individuals to be more fit than others. These variations are often referred to as alleles or Single Nucleotide Polymorphisms (SNPs). This paper outlines a new deep learning method called Locator 2.0, which sets to correlate genetic variation, specifically SNPs, to climate. We anticipate that our model will be able to find SNPs which are predictive of temperature and precipitation.

1 Introduction

The planet is warming and climate change is affecting the globe in drastic ways. As weather and ecosystems change considerably, scientists have not yet answered the question of whether the plants, our producers, can survive and adapt to such change. Studies have shown that many plants which have lost their race against extinction were influenced by humans directly/indirectly, and/or by atmospheric and oceanic changes. To determine whether plants can survive in the face of a rapidly changing climate, we must understand how plants adapt and populations evolve. Speciation theory, a concept from ecology, explains how species evolve over time (Weissing et al, 2011). It is also known that genomic variation leads some individuals to be more fit than others. These variations are often referred to as alleles or Single Nucleotide Polymorphisms (SNPs).
However, we still don't know what SNPs will allow our staple food crops like corn, rice, cassava, etc. to survive in a rapidly warming climate. To do so, we need to understand how and what genetic variation allows plants to adapt to changing climate conditions. However, this is tricky because genomes are complex. Polygenic traits, epistasis, and a variety of other non-linear interactions make prediction difficult. The plant *Arabidopsis thaliana* has made a great contribution to key concepts in molecular understanding in biology. Therefore, we aim to use this model organism to predict how the SNP profile of an individual changes under different climates using deep learning. Specifically, we aim to train a neural network that can successfully predict climate based on whole-genome SNPs to identify SNPs that are predictive of suitable climates for the plant.
There are 19 bioclim variables, and we will use these as hyperparameters and find correlation values between the variables and our SNPs to train our model on relevant SNPs. Predicting climate based on whole-genome SNPs to identify SNPs that are predictive of climate will give us great insight into the genetic variation which may allow *Arabidopsis thaliana* to adapt and evolve in the face of a changing climate and warming temperatures, which ultimately would give us insight into how key plants and crops heavily relied on by the human species may adapt as well.

*Stanford University c/o 22 and 23

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

2 Related work

It can be seen in the literature available that researchers have approached linking climate to genetic variation in many different ways. One example is an outlier detection method which was used to determine SNPs associated with climate in black spruce (Prunier et al, 2011). Another paper used a Genome Wide Association Study algorithm coupled with a Genome-environment association (two methods in molecular biology which link genetic variation to a trait or disease) to identify genetic variation linked to heat tolerance for the common bean (Lopez et al, 2019). More closely, another piece of literature also attempted to identify genetic variation and pathways which may be climate adaptive in *Arabidopsis thaliana* and found that amino acid–changing variants were "significantly enriched among the loci strongly correlated with climate, suggesting that our scan effectively detects adaptive alleles. Moreover, from our results, we successfully predicted relative fitness among a set of geographically diverse."(Hancock et al, 2011). Most important to our study, a classic piece of literature described a deep learning method authors coined "Locator" which uses genetic variation to predict where an organism came from geographically (Battey et al, 2019). It does this using control organisms, in this case meaning samples which have both known locations of origin and mapped genomes. Locator operates upon the assumption that there is some function which relates a likelihood of observing specific genetic variation in relation to location and uses a "deep, fully connected neural network to approximate this mapping for a set of genotyped individuals with known locations." In this paper we use a variation of this method which instead uses genetic variation, specifically SNPs, to be predictive of temperature for *Arabidopsis thaliana*, which we call Locator 2.0.

3 Dataset and Features

As implied by the name, our Locator 2.0 model heavily utilizes the ideas and tools described within the Locator deep learning method and model. As done in Locator, we have implemented a fully connected deep neural network, however our model is implemented using pytorch rather than tensorflow. Like Locator, we utilized the early stopping method for 100 epochs in our neural network. Also as done in Locator, we used the ADAM optimization algorithm and euclidean distance for our loss function. We used the same architecture, but tailored the output layer to our purposes to match the number of bioclim variables. We tried to implement a scheduler as done in the Locator model, however this proved to be ineffective at having any positive effect on our model and thus were left out of our Locator 2.0 neural network.
Our data set consists of 2029 *Arabidopsis thaliana* samples from 1001 genomes project and the rest were from the Moi Laboratory at Stanford University (Kawakatsu et al, 2016). Each sample has roughly 10 millions SNPs that were either present in the individual's DNA samples or imputed from the rest of the population using standard SNP imputation protocol. Given that we are predicting the climate of these samples, Moi Lab collected BioClimatic variables of each sample given their location from WorldClim2.[2] Bioclimatic variables are Derived from the monthly temperature and rainfall data and are often used in species distribution modeling and related ecological modeling techniques. They "represent annual trends (e.g., mean annual temperature, annual precipitation) seasonality (e.g., annual range in temperature and precipitation), and extreme or limiting environmental factors (e.g., the temperature of the coldest and warmest month, and precipitation of the wet and dry quarters)."[3]

4 Methods

4.1 Pre-processing samples

We considered using the same preprocessing approach used by the original Locator. As per the original Locator, we removed any samples where there wasn't a location or climate data. We split the remaining samples as 0.9/0.1 for train/test sets respectively which resulted in 1556/173 samples respectively in the train and test set.

4.2 SNPs Approach

As described in Locator, we trained our model using four different number of SNPs, ten thousand SNPs, 100 thousand, 500 thousand, and 1 million SNPs. Like Locator, we considered randomly

[2]WorldClim. (n.d.). Retrieved December 9, 2022, from https://www.worldclim.org/data/bioclim.html
[3]WorldClim. (n.d.). Retrieved December 9, 2022, from https://www.worldclim.org/data/bioclim.html

chosen SNPs for each of the categories (10000, 100000, 500000, and 1 Million) respectively.
In our own approach, we also considered choosing the mostly highly correlated SNPs in relation to each set of Temperature and Precipitation variables. We took the training set and calculated correlation values using Spearman correlation for each of the 12 Temperature and 8 Precipitation variables per SNP. We then took the mean correlation across the Temperature and Precipitation BioClim variables for each SNP. This process took upwards of 24 hours to run, and thus we used parallel computing to speed up this process to allow us to debug and run our code efficiently, this time choosing the SNPs with the highest correlation values and ran our model on each of these sets.

4.3 Baseline

In the baseline, we considered running linear regression on our 4 categorical sets of SNPs. For each of these sets of SNPs, we compared the randomly chosen SNPs and highest correlated SNPs approaches. This resulted in 8 linear models. The highest accuracies are the random choice of 500 thousand SNPs which has an average accuracy of 0.68 in Temperature while random choice of 100 thousand SNPs which has an average accuracy 0.49 accuracy in Precipitation.

4.4 Neural Network Approach

Since we cannot fit the entire set of 10 million SNPs on the GPU, we first split the genome randomly into two groups of roughly 5 million SNPs each. We then trained a model on each of these groups. The test losses were extremely high, and the average accuracy was below 0.05 across all the bioclim variables. There also was a high amount of noise in the data set, likely due to irrelevant SNPs to our model, giving us a low accuracy. Thus, we looked to the literature available to inform how we should proceed. As done in the Locator paper cited, we instead randomly picked 1 million SNPs and performed on mini-batch gradient descent with 32 samples. We considered using batch gradient descent on 10 thousand, 100 thousand, and 1 million randomly chosen SNPs. Figure 2A, 2B, and 2C shows much improvement in our model. It noticed that after some certain epochs the loss plateaus.

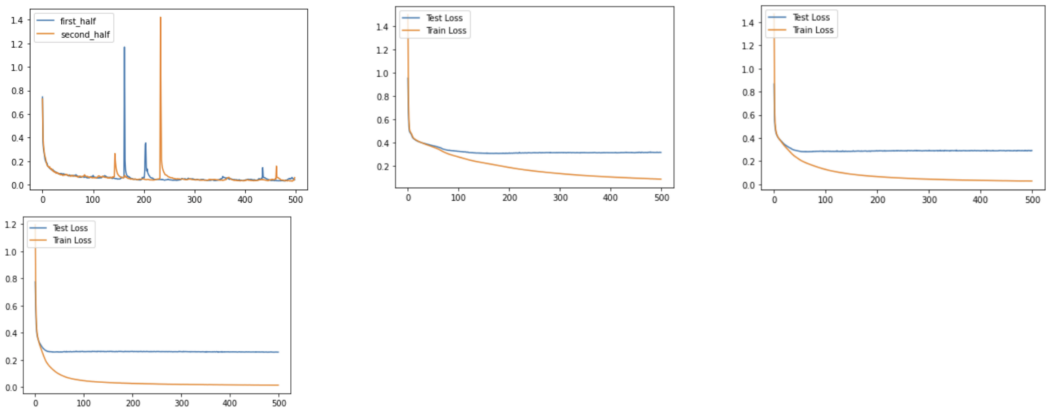


Figure 2: train and test losses of various Locator 2.0 models. A. [Batch Gradient Descent on 10 thousand randomly chosen SNPs]. B. [Batch Gradient Descent on 100 thousand randomly chosen SNPs] C. [Batch Gradient Descent on 1 Million randomly chosen SNPs]

To resolve that, we introduced the early stopping strategy when the test loss doesn't decrease at approximately 0.01 after 100 epochs, the training stops. This has also helped with faster training of our models.With the early stopping approach, we train a model on each sets of SNPs (10 thousand SNPs on randomly chosen, 10 thousand SNPs on high correlation,...etc) for Temperature variables and Precipitation variables respectively. These resulted in 16 unique models, displayed in Figure 3.

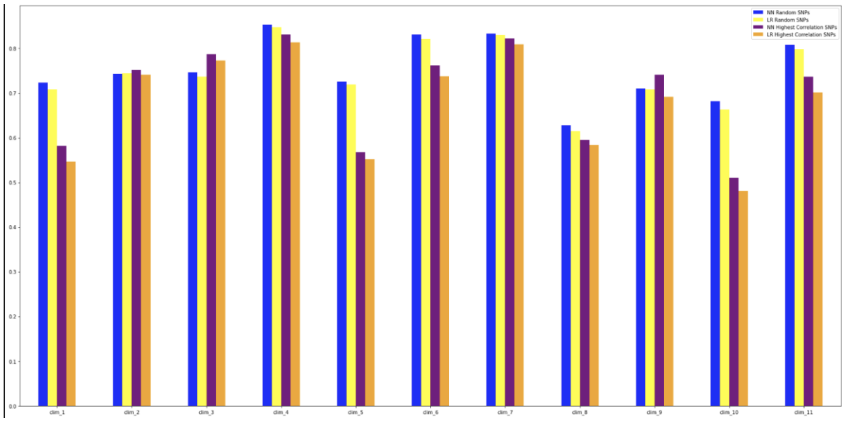
Temperature	Highest Correlated SNPs		Random SNPs	
SNPs	LR	NN	LR	NN
10k	0.379±0.006	**0.68±0.010**	0.38±0.005	**0.65±0.005**
100k	0.594±0.014	**0.65±0.014**	0.60±0.014	**0.78±0.007**
500k	0.632±0.017	**0.68±0.013**	0.63±0.0168	**0.78±0.010**
1M	0.675±0.012	**0.70±0.012**	0.68±0.012	**0.76±0.**

Precipitation	Highest Correlated SNPs		Random SNPs	
SNPs	LR	NN	LR	NN
10k	0.137±0.004	**0.45±0.018**	0.30 0.006	**0.45±0.003**
100k	0.364±0.009	**0.46±0.014**	0.49 0.012	**0.50±0.019**
500k	0.439±0.009	**0.48±0.012**	0.41 0.022	**0.42±0.021**
1M	0.443±0.010All	**0.46± 0.013**	0.47 0.011	**0.49±0.008**

Figure 3: A [All Temperature models], B [All Precipitation models]

5 Experiments/Results/Discussion

Initially, we used the aggregated set of 19 bioclim variables to find correlation values between the variables and our SNPs to help us reduce the number of SNPs we trained our model on to ones that are correlated to climate generally. Of these models, those generated via random sampling and via top correlated values are performing in relatively similar fashions. However, bioclim variables 1 to 12 are related specifically to temperature, and variables 13 to 19 are related to precipitation. Although temperature and precipitation are related and affect one another, aggregating the data in this fashion likely led to poor results as ecologically these are two distinct variables which should be considered separately. Thus, we disaggregated our set of 19 into the two sets for temperature and precipitation as described. We tweaked our model accordingly in both scenarios and re-ran our model in the same fashion as explained above with randomly selected and highest correlated SNPs. Our model improved when tailored to the 11 temperature parameters/bioclim variables. To quantify this, we ran a linear regression as a baseline on our model, on both the set of randomly chosen SNPs and the highest correlated SNPs. Our model performed the best on the set of 500 thousand randomly chosen and highest correlated SNPs, which is displayed in our accuracy results and seen in Figure 3. Most importantly, as seen in Figure 3, our Locator 2 model was the best performing and had highest accuracies and performing better than baseline.



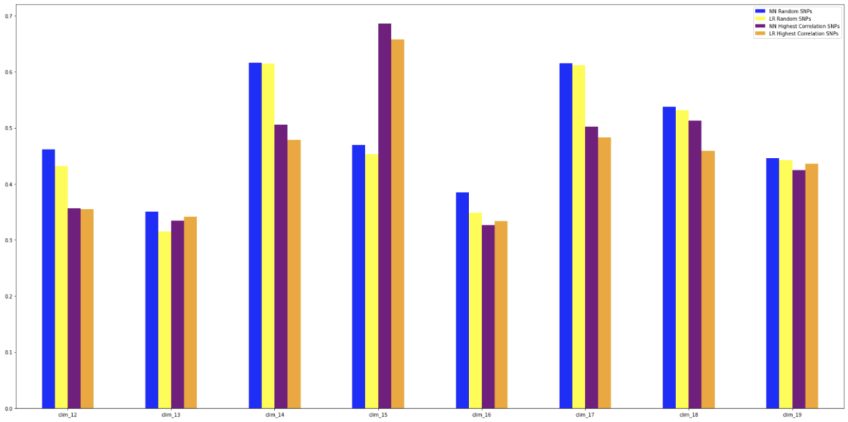


Figure 4: A [Best performing temperature model, 1mil SNPs], B [Best performing precipitation model, 1mil SNPs]

6 Conclusion/Future Work

To improve our model, we have many ideas for next steps. First, we are going to try to tweak our hyperparameters to find the best ones for our network as a starting point in fixing our model. Additionally, we are also working on coding a Genome-Wide Association Study (GWAS) to help us with the issue of finding relevant SNPs to act as a baseline, in a similar fashion that Locator was able to use samples with known geographic origins as a baseline. GWAS is a technique in bioinformatics that is used to identify genomic variants, or SNPs in this case, that are statistically associated with a risk for a disease or a particular trait. In our case, we are looking for SNPs that are correlated to and predictive of temperature. This GWAS study would behave in a similar fashion to our linear regression, but it would allow us to control for things that are important in the realm of genetics, such as population structure and family history of samples. Once a model is trained, we will use a variety of neural network interpretability techniques like DeepLIFT, "a method for decomposing the output prediction of a neural network on a specific input by backpropagating the contributions of all neurons in the network to every feature of the input" (Shrikumar, et al. 2017), to try and extract what important genomic features lead to increased frequency in different climates.
Our ultimate goal is to build a comprehensive map from variation in the genome to organism fitness (the ability of an organism to pass on genetic material to offspring). In future work, we hope to predict a proxy variable, allele frequency, which we have a lot of data on and correlates highly with fitness. Specifically, we aim to train a deep neural network to predict the change in allele frequency from the first generation to the second from both the DNA sequence and the climate at a given location. We will try training different neural architectures, potentially including convolutional neural networks, Long Short-Term Memories (LSTMs), and/or Transformers. The goal is to train a neural network that can successfully predict the change in allele frequency for a given SNP in a given environment. More specifically, we will train the network to predict the allele frequency change per SNP of a given year and predict the allele frequency change in future years, and check to see if the neural network works on them.

7 Contributions

As mentioned in our dataset section, the Moi lab both provided data for our project and helped guide our methodology and contributed to our work. In particular, Lauren Gillespie and Dr. Moises Exposito-Alonso were key contributors to this project. Additionally, special thanks to Surag Nair who was our TA this quarter and provided valuable contributions and feedback which helped guide the progress of our model.

References

[1] Battey, C. J., Ralph, P. L., Kern, A. D. (2019). Predicting geographic location from genetic variation with Deep Neural Networks. https://doi.org/10.1101/2019.12.11.872051

[2] Hancock, A. M., Brachi, B., Faure, N., Horton, M. W., Jarymowycz, L. B., Sperone, F. G., Toomajian, C., Roux, F., Bergelson, J. (2011). Adaptation to climate across the arabidopsis thaliana genome. Science, 334(6052), 83–86. https://doi.org/10.1126/science.1209244

[3] Kawakatsu, T. (n.d.). 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. Shibboleth authentication request. Retrieved December 9, 2022, from https://www-sciencedirect-com.stanford.idm.oclc.org/science/article/pii/S0092867416306675?via

[4] López-Hernández, F., amp; Cortés, A. J. (2019). Last-generation genome–environment associations reveal the genetic basis of heat tolerance in common bean (phaseolus vulgaris L.). Frontiers in Genetics, 10. https://doi.org/10.3389/fgene.2019.00954

[5] López-Hernández, F., amp; Cortés, A. J. (2019). Last-generation genome–environment associations reveal the genetic basis of heat tolerance in common bean (phaseolus vulgaris L.). Frontiers in Genetics, 10. https://doi.org/10.3389/fgene.2019.00954

[6] PRUNIER, J. U. L. I. E. N., LAROCHE, J. É. R. Ô. M. E., BEAULIEU, J. E. A. N., amp; BOUSQUET, J. E. A. N. (2011). Scanning the genome for gene snps related to climate adaptation and estimating selection at the molecular level in boreal black spruce. Molecular Ecology, 20(8), 1702–1716. https://doi.org/10.1111/j.1365-294x.2011.05045.x

[7] Weissing, F. J., Edelaar, P., amp; van Doorn, G. S. (2011, March). Adaptive speciation theory: A conceptual review. Behavioral ecology and sociobiology. Retrieved December 9, 2022, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3038232/

[8] Weissing, F. J., Edelaar, P., amp; van Doorn, G. S. (2011, March). Adaptive speciation theory: A conceptual review. Behavioral ecology and sociobiology. Retrieved December 9, 2022, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3038232/

[9] WorldClim. (n.d.). Retrieved December 9, 2022, from https://www.worldclim.org/data/bioclim.html