PREDICTING PLANT SENSITIVITY BASED ON ROOT SYSTEM RESPONSES TO PHYTOHORMONES IN ARABIDOPSIS USING TRANSFER LEARNING

Rohith Ravindranath
rr3415@columbia.edu

Columbia University
Computer Science Department

Jordynn Lurie
jml2347@columbia.edu

Columbia University
Biomedical Engineering Department

ABSTRACT

Perturbation is widely used in agriculture as a way to grow biologically fit plants. However, plants of the same species have genetic variations based on their global location, which gives rise to different sensitivity levels to perturbation. We propose a new architecture based on transfer learning and SNP data to describe the sensitivity of arabidopsis accessions in response to phytohormone perturbation. We propose a two-step method. The first is to pre-train sub-networks that are able to predict phenotype values that describe the accession's root system. The second is to add dense layers on top, fine tune the full model, and predict the accessions sensitivity to perturbation. Our model is able to converge on the small dataset that is available to us and provides a general framework that can be used in the agriculture industry.

1. INTRODUCTION

Currently, there is an abundance of data related to how a plant's root system architecture (RSA) responds to various phytohormones. RSAs are both easy to perturb with molecules of interest and of high adaptive relevance as the RSA provides the framework for plant growth and productivity. Phytohormonal pathways are the driving factor for RSA adaptation, allowing the plant to optimize its stability and uptake of vital nutrients. In a paper that analyzed 192 Arabidopsis accessions, [1], it was observed that all Arabidopsis RSAs responded to phytohormones in the same way, however the degree to which each accession responded was dependent on its individual genotype. Understanding the magnitude of a given plant's RSA response to phytohormone perturbation gives us an understanding of how sensitive that plant is to its own phytohormonal pathway, and thus its ability to optimize its RSA for its environment. The RSA is complex and can be described by 10 measurable phenotypes including root length, root branching, root mass density etc.

While there are various machine learning methods that have shown efficacy in accurately predicting each of these phenotypes, we wish to take a more holistic approach that incorporates all the measured phenotypes in predicting an accession's sensitivity. We aim to use SNP data from each accession to predict the various root phenotypes and finally compute an overall fitness score. Formally, the biological question we aim to answer goes as follows - Can we predict the overall evolutionary sensitivity of an accession based on root system responses to phytohormones given the respective SNPs?

With respect to our methodology, we plan to develop a supervised model that uses a transfer learning approach. We are interested in building a model that uses pre-trained sub-networks to (i.) predict the accession's RSA based on the SNP and (ii.) then use those intermediary values to predict the overall sensitivity (singular output). In our proposed architecture, the intermediary outputs are created independently from one another.

2. DATA

We will use two separate data sources/sets from the 1001 Genomes Database. First, we will use the Arabidopsis Single Nucleotide Polymorphism (SNP) data[2]. This data contains meta-data about each accession which we will use to coordinate the phenotype data. The format of the data is in a binary matrix.

Our second data source is from the AraPheno database which contains curated public phenotypes for Arabidopsis accessions. We will have to download all the data (from multiple studies[1]) that phenotype various traits of the root systems in response to Abscisic Acid, Cytokinin, and Auxin. Specifically the data contains the id of the accession, phenotype after exposure to drug, and specific drug used.

Finally, we used data derived in a paper[1] that analyzed the same aforementioned Arabidopsis accessions. In this paper, mean values for the 10 root traits described in Table 2 across 192 Arabidopsis accessions and four conditions [auxin (IAA), cytokinin (CK), abscisic acid (ABA), and no hormone (C)] were used to perform a principal component analysis (PCA). They then calculated the Euclidian distance (Ed) in the RSA space defined by PC1 and PC2 for each accession from each hormone treated to the control treated sample. This

Root Architecture Mapping 1	
Phenotype Name	**Trait Ontology (TO)**
Mean(TRL)	Lateral Root Length
Mean(TLRL)	Lateral Root Length
Mean(R)	Root Branching
Mean(P2)	Root Length
Mean(LRR)	Root Mass Density
Mean(LRL)	Lateral Root Length
Mean(LRD R)	Root Mass Density
Mean(LRD P)	Root Mass Density
Mean(LR.no)	Lateral Root Number

Table 1. Mapping of Phenotype Name to Ontology

Root Architecture Mapping 2	
Phenotype Name	**Trait Name**
Mean(TRL)	Total Root Length
Mean(TLRL)	Total Lateral Root Length
Mean(R)	Length of Branching Zone
Mean(P2)	Primary Root Length
Mean(LRR)	Length Ratio
Mean(LRL)	Average Lateral Root Length
Mean(LRD R)	Density in R
Mean(LRD P)	Density in P
Mean(LR.no)	Lateral Root Number

Table 2. Mapping of Phenotype Name to Trait Name

distance indicates the effect of the treatment in relation to its RSA under control conditions and thereby is corrected for developmental differences of the accessions. For each accession, the average Ed was calculated for all conditions which represents a measure of how profoundly a genotype differs from the norm in altering its RSA in response to all hormonal perturbations.

Since our data is in various data files, preprocessing was needed to get the required features into a singular data file. Each accession has a unique identifier that is present in all data sources, hence we used that id to merge all the phenotype data together. With respect to the SNP data, we re-ordered the matrix to ensure the order of the accessions lined up correctly with the order in our final phenotype dataset.

3. METHODS

Our proposed method takes a transfer learning approach. We first create several small models, one for each phenotype trait of the root system. We train these individual models independently hoping that this will allow the models to better capture how the phenotype trait is influenced by both the genetic variation in each accession and one of three drugs (auxin, cytokinin, and abscisic acid). After training these models, we freeze their weights and add two dense layers on top of these

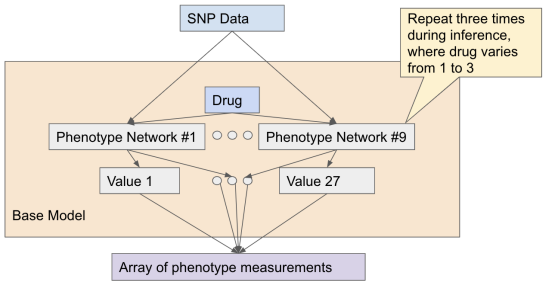

Fig. 1. Architecture of base model, illustrating the sub-networks with their inputs and outputs.

models. We then fine-tune the model with a different set of inputs and outputs. Our final model will take input the SNP data of an accession and the output will be the sensitivity that specific accession has on perturbation. The following subsections go into detail regarding base models and the full model.

3.1. Base Models

Our base model, contains multiple smaller networks, each unique to a specific root phenotype trait. In our dataset, we have 9 phenotype traits and their exposure to perturbation. Hence our base model consists of 9 smaller networks. The input to each of these inner networks is the accession SNP data and the corresponding drug (numeric value from one to three). Equation 1 illustrates the tuples for our inner models. These inputs are concatenated together and fed into the networks. The output is the predicted measurement of a specific phenotype trait. The input (X) to all nine inner models is the same, while the output (Y) is the specific phenotype measurement that the network is being trained for.

$$(X_{1,1}, X_{2,1}, Y_1) \ldots (X_{1,n}, X_{2,n}, Y_n) \quad (1)$$

Each of the nine models have the same layers. Each network starts off with an embedding layer, followed by two fully connected layers, and finally an output layer with a sigmoid activation function. The two fully connected layer had 1028 and 128 units respectively.

During inference time, these nine models are run 3 times each for each drug in our data set. The output of our base model is 9*3 = 27 scalars that describe the root architecture after exposure to each drug. These values are then fed into dense layers which output a singular value that describes the accession's overall sensitive to perturbation. Figure 2 breaks down our base model and shows how the sub-models are combined.

With respect to training these sub-networks, they are trained independently with the same input but different corresponding output. These sub-networks are trained over 10 epochs each, using the Adam optimizer with a learning rate or 0.01. Our loss function is mean absolute error. We also

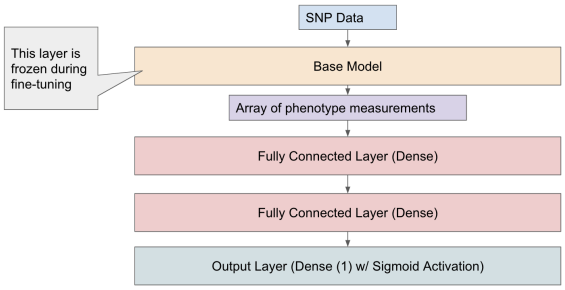

Fig. 2. Architecture of full model.

have a 80-20 validation split when training about model. Our intuition is that training these sub-network independently will allow the sub-networks to better capture how the genetic variation in each accession combine with a specific perturbation affects a specific phenotype trait.

3.2. Full Model

Given our base model, we take a transfer learning approach and add full connected layers on top of our base model. We finally add an output layer with a sigmoid activation to output a singular value describing the sensitivity of the specific accession to perturbation. The two fully connected layer had 1028 and 128 units respectively.

When training our full model, the base model is set to non-trainable, hence the weights cannot be altered when training out full model. Only the newly added layers are trained. This method is called fine-tuning. Using the knowledge already gained from the sub-networks in our base model, our top layers are trained to use that knowledge to make an even better prediction. This is also the premise of transfer learning, where knowledge learned from the source dataset to the target dataset.

$$(X_1, Y_1) \dots (X_n, Y_n) \tag{2}$$

Our input to the full model is just the accessions SNP data and the output is average sensitivity to perturbation. Equation 2 shows the input space for our full model. Notice that this input is different from the sub-model. With our sub-model, we have an added input of the specific drug. This input (X_2) is concatenated with the SNP data within the base model.

The full model is trained over 10 epochs each, using the Adam optimizer with a learning rate or 0.01. Our loss function is mean absolute error. We also have a 80-20 validation split when training about model. Initially, we had the learning rate set to 0.1, however we notice our loss after each epoch had high fluctuation. Hence we lowered the rate to 0.01.

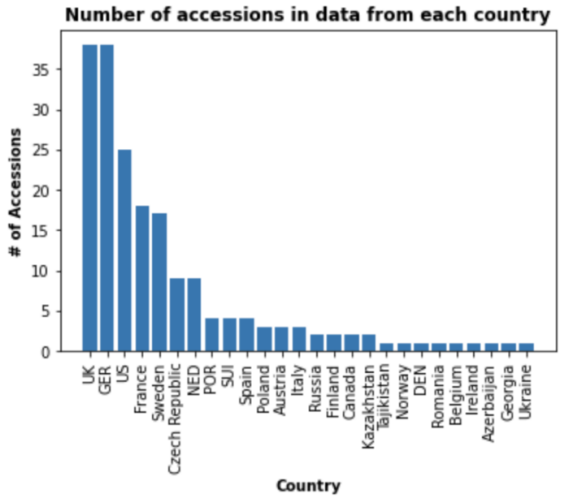

Fig. 3. Number of Accessions from Each Country.

4. RESULTS

We show results of both the sub-networks and the full model after transfer learning and fine tunning. We also show results of our exploratory analysis in order to give us better insight on how our model is performing.

4.1. Exploratory Data Analysis

We performed exploratory data analysis in order to understand the distribution of our accessions across country of origin, to see if there is any correlation between the 10 measured phenotypes, and to visualize the overall change of each phenotype in response to each phytohormone compared to control.

The bar plot displaying the number of accessions from each country in Figure 3 shows that the majority of our accession data comes from the UK, Germany, or the US, with the rest of the data makeup shown in the graph. The percentages of data represented by each country were also calculated. These percentages allowed us to ensure that each country is equally represented in our training data.

A sample heatmap in Figure 4 shows the correlations of phenotypic traits between each other and latitude, with the lighter squares representing a higher correlation. The traits that are shown to be highly correlated in this figure (i.e. root number ($LR.No$), lateral root density (LRD_P), total lateral root length ($TLRL$), and length ratio (LRR)) contribute to the same principal components as calculated in [1]. Demonstrating the reputability of the paper's findings.

The percent change of each trait per country after treatment with each of the 3 phytohormones are displayed in the bar plots above. Overall percent change for each trait after treatments are shown in Figure 5. Although treatment with each phytohormone did cause different degrees and directions

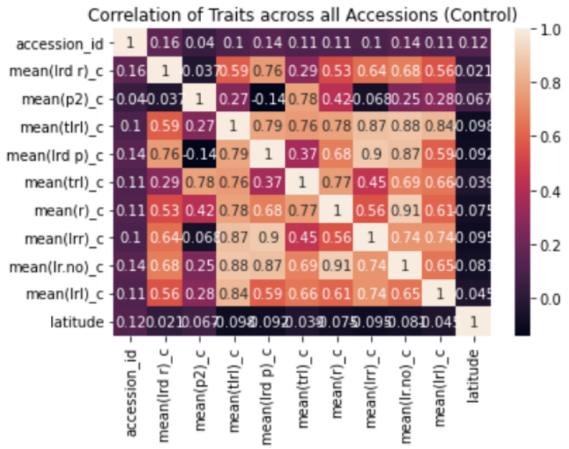

Fig. 4. Correlation of Traits across all Accessions (control).

of change in each trait, Auxin treatment most notably caused the most robust root system architecture response.

4.2. Sub-Networks

As mentioned before, there are 9 sub-networks. Each for a specific phenotype trait that describes the root architecture. All 9 models have the same network architecture, but we train these models individually to achieve the lowest loss possible for each trait. We observe in the Figure 6 and Figure 7 that they all nine models converge together. The train and validation loss converge together. This is a good indication that there is no over-fitting or under-fitting occurring. This is mainly due to the Dropout layer we added after the first Fully Connected Layer. Previously, we did not add this layer and we saw the model over-fitting to the training data. However, that is not the case after including the Dropout layer. We also notice that there is a differences on the final loss value for each of the sub-networks. This can be described by the variance in each of the phenotype traits and also the range of values for each of the phenotype traits.

4.3. Full-Model

When training the full model, we create a custom layer. This custom layer takes in the full model input, the SNP data of the accessions, appends a scalar indicating which drug the sub-networks should predict upon. We then run through all nine models three times each outputting a total of 27 scalar values. These values are then passed to dense layers. After training the model via transfer learning and fine-tuning, we observe that the model converges in Figure 8, however we hoped that the final loss would be lower. We also predict on a test set and notice we get around the same loss, indicating a good fit and no over-fitting in our model. All of our sub-networks and our model were assessed using Mean Absolute Error (MAE),

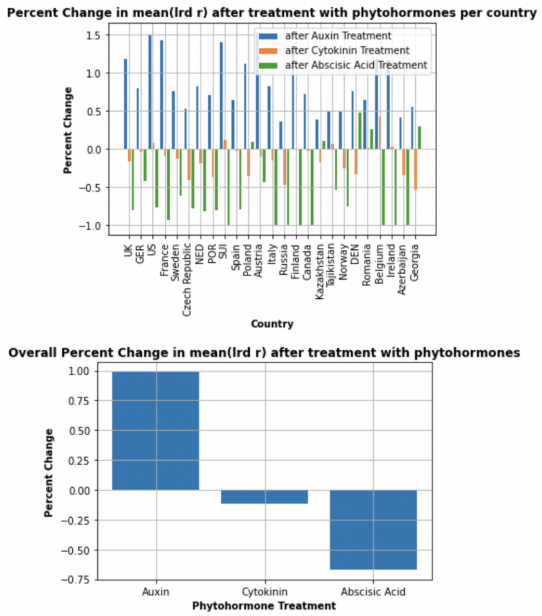

Fig. 5. Country-wise and Overall percent change in mean(lrd r) after treatment with phytohormones.

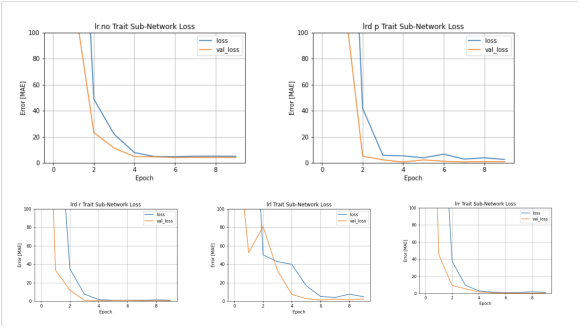

Fig. 6. Loss of sub networks for each phenotype trait.

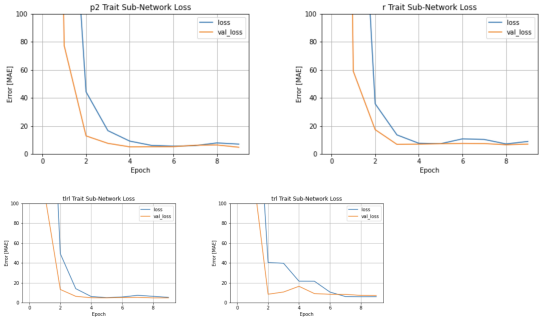

Fig. 7. Loss of sub networks for each phenotype trait.

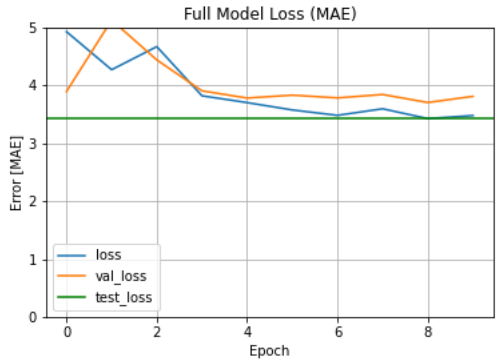

Fig. 8. Loss of full model after transfer learning.

since our output is continuous.

5. DISCUSSION

5.1. Model

With respect to our sub-networks, we trained them independently rather than a single model that predicts for all phenotypes. Our intuition was that each sub-network would capture the specific genetic variation in the SNP for the respective trait. We see this to be true in our figures, which show all nine models converge with no overfitting. This was the first sign that we were on the right track with our proposed architecture. Furthermore, we observe that we didn't have to train our models for for 10 epochs. Rather, in most cases 5 epochs seemed to be sufficient based on the plots. We also see this present in our full model with all the sub-networks integrated as a custom layer.

One challenge we did face early in our project was deciding on the architecture of our sub-networks. Originally we had two dense layers with relu activation. The models were able to converge, but was misleading. When predicting with our original version of the model, the sub-networks would predict the same value for all values of X. After further investigation, we realized that it was because we were using a relu activation function instead of a sigmoid function. We think that the relu activation function performed poorly since our SNP data is binary. Furthermore, we also added in a dropout layer (p = 0.3) after our first dense layer. We hoped that adding the dropout layer would ensure that our model would not overfit. After seeing the success with these two changes in our sub-network, we then implemented these changes in the top layers of our full model.

Another challenge we faced regarded our SNP data. Our SNP data seemed to be too long for our model. For each accession, the SNP data was 214,553 values long. At first, the model would not compile since it would have to create too many parameters for a single layer. We then shorten that number by half. Our model did compile and run, however the loss was really high (MAE = 123.0). One reason was that the length of the SNP was still too long and the model couldn't effectively identify patterns in such a long string of binary values. After trail and error, we came a sweet spot of 50,000 base pairs for our SNP data for each accession. This length seemed to provided the lowest loss for our model.

One limitation of our data was the number of examples we had. In total, we only had 192 accessions and their respective data. It is known that for deep learning networks to perform, an abundance of data is needed. Although our models did converge during training, we expect better performance if we had access to more training examples. In the future, we would also try to normalize the phenotype measurements since we have a small dataset, normalizing our data would make sure our model does not get impacted by outliers and would have a lower variance.

5.2. Application in Real World

We believe that our model that predicts Arabidopsis sensitivity to key phytohormones can have applications in counteracting the negative effects climate change has on agriculture. The effects of climate change have been significantly accelerating since the last century. It is expected that, by 2050, the global mean temperature will increase by 1.5–2°C, causing weather extremes that will negatively affect agricultural production[3]. The global population is also expected to reach nine billion by 2050[3]. Projections show that feeding world's population would require raising the overall food production by around 70% by 2050, however, the current trajectory shows that the rates of global production in key crops would increase far below what is needed to produce enough food to meet the raising population demands[3]. A higher temperature causes changes in water viscosity of soil and root hydraulic conductance that damage roots. However, this increased temperature also triggers alteration of key phytohormone hormone levels that trigger signal transduction pathways preparing plants to overcome the stress of situation[3]. If farmers can plant crops that they know are more reactive to these pathways, they can increase the likelihood of plant survival amid increasing temperatures.

5.3. Next Steps

To take this work further, there are several areas of improvement that can be done to the model and the data preprocessing. First, we are currently using the first 50,000 values of the SNP data. In future work, it would be interesting to see if we selected specific base pairs in the SNP data that are known to affect the root system or sensitivity and see how that affects our model. We think selecting specific sections of the SNP data using a priori knowledge might result in a better overall loss. It would also be good to have more data available (i.e. more training examples). With respect to our model, we believe there could be some exploration done

to improve the architecture of the sub-networks. It would be interesting to see how embeddings in our sub-network would affect the overall model. Currently, we are feeding the intermediary outputs from our sub-networks directly into the top layers. It would be interesting, as a next step, to see if any data transformation to those 27 scalar values would affect the model performance as well.

5.4. Code

You can view our code and dataset here.

6. REFERENCES

[1] Kristina Metesch Wolfgang Busch Daniela Ristova, Marco Giovannetti, "Natural genetic variation shapes root system responses to phytohormones in arabidopsis," *the plant journal*, vol. 96, pp. 468–481, October 2018.

[2] Florian Jupe Eriko Sasaki Robert J. Schmitz Mark A. Urich Taiji Kawakatsu, Shao-shan Carol Huang and Rosa Castanon, "Epigenomic diversity in a global collection of arabidopsis thaliana accessions," *Cell*, vol. 166, pp. 492–505, July 2016.

[3] L. Oñate-Sánchez M. Pernas J. Calleja-Cabrera, M. Boter, "Root growth adaptation to climate change in crops," *Frontiers in Plant Science*, May 2020.