Top of Form

|  |  |  |
| --- | --- | --- |
| Client: | Ying Li | File: 25-045 |
| Dept: | Horticulture & Landscape Architecture | Faculty:  Student: |
| Date: | 3/26/25 | Initial Meeting:  Follow-up: |
| Consultant and Attendees: Dr. Ying Li, Sumeeth Guda, Dr. Chong Gu | | |
| Statement of Problem:  Does incorporating chromatin data into a random forest model enhance the accuracy of the predictions of transcription factors for legumes. | | |
| Goal of this Project: Journal Publication / Grant Application | | |
| Background:  The client is a faculty member in the Horticulture department who is investigating if gene regulatory networks are important for integrating CO2 availability and nitrogen supply in legumes. The background in the experiment is that increasing atmospheric CO2 concentrations has the potential to improve agricultural output. However, the photosynthetic gains from rising atmospheric CO2 concentrations are generally lower than the maximum predicted gains. This is caused by a down-regulation of photosynthesis that is connected to plant carbon-nitrogen imbalance when grown at elevated CO2. While this phenomenon occurs in legumes, it is not as severe due to their ability to exchange carbon for nitrogen with symbiotic soil bacteria. The client is interested in seeing how legumes sense, integrate, and respond to nitrogen and carbon/energy status through dynamic control of gene expression. And see how this enables the development of strategies for achieving maximum crop yield and quality in future climates. The client will build a gene regulatory network which combines gene expression data with epigenomic data measured across nitrogen and CO2 treatments. Their hypothesis for this experiment is that incorporating chromatin data will enhance the accuracy of the predictions of the transcription factors.  They came to the SCS to get consultation before for a grant proposal that is due on 3/28/25 as well as to gain the perspective of how a statistician would adjust the model parameters to account for the external Chromatin data. | | |
| Progress of project at this time: Design (No data collected) | | |
| Relevant information presented at the meeting:  Variables + Proposed Design:   * The study design has 3 different nitrogen supplies (5, 10, 20) each with 2 different CO2 concentrations (420 ppm, 880 ppm), for two different varieties of beans (Snap093, Huntington). * Transcriptome and chromatin data will be collected via chlorophyll fluorescence imaging.   Meeting Notes:  From the get-go, Dr. Li made her intentions clear of why she came to the SCS, what she is looking for is consultation on her design and suggestion on how to tweak her random forest model. Dr. Li's explained that her lab studies gene expressions in plant genomes, focusing on how different environmental conditions influence gene activity. When a gene is active, it transcribes RNA copies, which can be measured to determine gene activity. This research studies how proteins and other biological components are produced in the plant genome. The experiment involves varying CO₂ concentrations (420, 880 ppm), N₂ concentrations (5, 10, 20), and two bean types (Snap093, Huntington) to observe how different conditions impact gene expression and the relationships between genes. The primary goal is to analyze gene expression responses and differentiation under these conditions.  A key regulatory factor in this process is transcription factors (TFs), which bind to gene proteins and regulate RNA production. There are 2,000 TF genes among the 30,000 total genes within this study. The main challenge is understanding how all TFs control all genes within a directed gene network, where the connections may not be strongly linked. The lab uses GENIE3, a random forest-based model, to infer gene regulatory networks by treating each gene as a potential target and predicting its expression based on other genes. However, GENIE3 does not distinguish TF genes from target genes and focuses on the aggregated predictive power of the model rather than individual local predictions. The original GENIE3 treats each of the p genes as target gene one at a time and it uses the other p-1 genes as predictors. The model will then aggregate the impact of each gene in the p-1 models in which it serves as predictors.  Dr. Li seeks to modify GENIE3 by incorporating chromatin data, which includes information on chromatin openness (measured via ATAC-seq) and histone marks, both of which influence TF activation. The hypothesis is that open chromatin allows TF genes to be active, while closed chromatin and histone modifications can suppress TF expression. However, GENIE3 does not natively support integrating external weights or additional datasets beyond the gene expressions. The client is fixing a target gene and is focusing on the prediction model itself using the TF genes as predictors. One suggestion Dr. Gu had was to introduce chromatin openness as a sample-specific weight, which could modify predictor variables in the model.  There are technical challenges in implementing this modification. GENIE3 optimizes tree splits using an OLS estimate to minimize squared error, and it is unclear how to directly incorporate weights into this process. Regression-based approaches could conceptually estimate gene influence by attaching weights, but they may be computationally slow and difficult to fit due to data complexity. A possible approach is to analyze a single target gene and a single histone modification at a time, using openness measures across samples to weight predictor genes. The openness of TF genes could act as a modifier for predictor variables. However, regression might not be feasible due to scalability issues, and GENIE3 remains the standard method for Dr. Li’s field.  Concerns:  There were some concerns regarding the project.   * Firstly, it was identified that GENIE3 had some technical limitations. The random forest model does not distinguish between TF and target genes and cannot naturally integrate external datasets like chromatin openness or histone marks as weights. This is a problem, because it does not allow adding weights or modifying tree splits based on chromatin openness. * Regarding the analysis of the results, while regression could theoretically introduce weighting, it may be computationally slow and difficult to fit due to the complexity of gene expression data. | | |
| Recommendations for Design and/or Analysis:  As discussed in the meeting, there is not a direct way to incorporate chromatin data into GENIE3’s tree-based model unless weights can be incorporated model. However, Dr. Gu recommended a potential workaround by fixing a single target gene and using different samples of that gene to create sample-specific weights. The openness of the TF gene could be used to modify the impact of the data, and if implemented as a multiplicative adjustment, this would mathematically be equivalent to modifying the predictor variable. In essence, chromatin data could be used to weigh the samples, transforming the approach into weighted least squares (WLS) regression. If GENIE3 allows for weight insertion, this method could be directly applied; otherwise, different versions of the weights could be created using histone data as an alternative strategy. | | |
| Who will carry out these actions?  Client:   * Look into the GENIE3 documentation and see if it supports weight insertion in the model initialization. * Let the consultant know if there are any issues or if she has any questions about the analysis process.   Consultant:   * Answer any question the client has regarding the analysis of the model results and direct them to the correct resources to accomplish their task. | | |
| Status: Follow up required. | | |

Bottom of Form