**Client 25-045**

**3/26/25 11:30 AM**

**Goal: Journal Article / Grant Proposal**

**Phase: Design (No data collected)**

**Deadline: 3/28/25**

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.
* They came to the SCS to get consultation before for a grant proposal that is due on 3/28/25.
* Grant Form: <https://portal.nifa.usda.gov/web/crisprojectpages/1032112-optimizing-nitrogen-photosynthesis-relationships-for-future-climates-improving-predictive-accuracy-of-gene-networks-with-epigenetic-states.html>
* 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 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 of 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.

Hypothesis: Incorporating chromatin data will enhance the accuracy of the predictions of the transcription factors.

Variables + Proposed Design:

* The study design has 3 different nitrogen supplies each with 2 different CO2 concentrations, for two different varieties of beans (3x2x2 design).
* Transcriptome and chromatin data will be collected via chlorophyll fluorescence imaging

Research Questions:

* Transcriptome data will be used to infer a gene regulatory network using a random forest-based approach called GENIE 3.

Statistical Issues:

* The client has experience with GENIE3, however, they want to adapt GENIE3 to include the chromatin data and would like a statistician’s expertise's opinion on how to incorporate it (in the sense that which layers/parameter to change in the random forest).

Questions to ask client:

1. What is the target variable in the dataset? What is the proposed test-train-validation split?
2. How many trees are being boosted during their simulations? How many layers are in the trees?
3. This project seems more like a machine learning project compared to an applied statistics-based project.