MC REU Proposal – Summer 2022

**ALEX LI**

zql5144@psu.edu

Undergraduate Student

Computer Science & Math Program, Penn State Harrisburg

Project Title

Differential analysis for gene expression and cannabinoid pathways in Hemp across different experimental conditions

Supervisor

Dr. Hien Nguyen

Computer Science & Math Program, Penn State Harrisburg

Secondary supervisor

Dr. Shobha Potlakayala

Biology & Science Program, Penn State Harrisburg

**Objectives of the Proposed Project**

The objective of this project is to develop statistical models to analyze the dynamics of the genes and biosynthesis pathways of cannabinoids in Hemp under different suppressing conditions.

**Introduction**

In modern times the world is facing a variety of problems, ranging from global warming to fulfilling the growing resource demand that is a direct result of the vast growing human population. As result of global warming and increasing food demands, it is becoming increasingly more important to cultivate crops that are resistant under drought stress. It is also important to understand how plants react under drought stress, in terms of productivity and the effects on biological composition (Li et al. 2021). A promising drought resistant crop that provides food, fiber, and medical material that will become increasingly important in coming years is *Cannabis sativa L*. *Cannabis sativa L*, which is also known as industrial hemp, is a close relative to marijuana, that possesses high amounts of cannabinoids (CBDs). In total there are over 500 metabolic compounds found in the plant, some of which possess the qualities necessary to be useful in a variety of applications including the medical field. Some of the ailments and diseases that can be treated by the compounds are diabetes, stress induced ailments, as well as a multitude of pain management applications among a variety of other things. There has also been a shift in research projects favoring industrial hemp over marijuana, due to the high concentration of CBDs in hemp and the social stigma surrounding marijuana (Lata et al. 2009). One of the largest setbacks in the research of industrial hemp is the lack of standardized tissue culture methods that can generate homogenous hemp lines. If a standardized tissue culture method is developed, it could be used to overexpress the CBD’s to be used in a variety of beneficial applications.

In the past droughts have caused large amounts of crop loss and famine across the world with varying degrees of intensity. Recently droughts have become increasingly common and will most likely become more common as the food demand continually increases, due to pressure from a rapidly growing population, and climate change (Muhammed et al. 2022). In the past other varieties of plants have, such as sugarcane, have been studied to determine the effect of drought and high salinity on plant productivity and genetics. To determine how to properly cultivate and treat plants that are under varying levels of drought and saline stress (Li et al. 2022). All of which shows the need to determine the effect of drought on industrial hemp, as a means of determining the proper watering regimen to maximize plant production and the validity of cultivating industrial hemp in an extremely dry environment. This allows greater resource conservation and management, as well as greater efficiency and production per plant capita.

Currently there is minimal research on the effect of drought on industrial hemp and identification of genes involved in response to drought stress. This project will study different lengths of drought stress on industrial hemp and their effects on the relative changes in the gene expression profiles of the flower tissues collected at three different time points after the drought treatment. Parallelly, flower tissues will also be sent for complete metabolite profiling to understand the changes in the in the production of metabolites. This project would allow us to demonstrate the ability of industrial hemp to grow on marginal lands and demonstrate the potential to expand the agricultural lands in Pennsylvania. The drought experiments involving industrial hemp plants will be conducted at the NIH Level II greenhouse facilities of the Central Pennsylvania Research and Teaching Laboratory for Biofuels (CPRTLB) at Penn State Harrisburg. Transcriptome and metabolite analyses will be performed by statistical models.

**Project goals**

1. Learn how to interpret biological data, especially the transcriptome data.
2. Learn how to use bioinformatics tools to organize and extract biological data.
3. Learn how to develop statistical models for hypothesis testing
4. Determine the most differentially expressed genes in the drought conditions
5. Study how the cannabinoid biosynthesis pathways rewire in different conditions.
6. Present the findings at a conference.

This MC REU program would provide me with the valuable research experience that would shape my academic pathways and my future career.

**My Research Background**

While studying computer science at Penn State Harrisburg, I am fortunate to meet many brilliant professors. They rigorously trained me in mathematics and further developed my interest in applied statistics.

Having taken two advanced math classes (STAT 318 and MATH 448), I am equipped with data analytics skills. I have worked with statistical models and their real-world applications. For example, I used regression models to predict the annual stock returns based on the Dow Jones index since 1896.

Besides, as a computer science major, I am skilled in programming. I know R and Python, two powerful programming languages for statistical purposes. I have actively participated in a machine learning research group and successfully implemented several key algorithms and evaluated their effectiveness. I have also assisted graduate students in their research projects.

**Personal Statement**

I have always had a strong interest in utilizing my knowledge to better this world. Ever since I was a child, I love using technology to make positive changes around. During high school, I was fascinated in robot development and wanted to invent a robot that could free people from their house chores. I spent countless hours in the robotic club. Eventually, I did not invent that robot (yet), but won a national robotic championship and went to Malaysia to compete internationally.

Later down on the road, I realized that, at the core of any scientific project, it is crucial to effectively analyze the data and accurately make predictions and decisions. Therefore, I am pursuing a computer science degree with a minor in mathematics. I want to learn more and solve real-world problems I encounter.

**Materials and Methods**

The gene sequencing step will be performed by our collaborators in the Department of Biology, Penn State Harrisburg. We will use Kallisto - a bioinformatics tool (Bray et al. 2016) – to quantify the abundance of transcripts. The results of this step will be the expression of genes in different conditions. Data normality will be tested Shapiro-Wilk test. If the normality is satisfied, we will apply Analysis of variance (ANOVA) to detect the differentially expresses genes. Otherwise, we will apply non-parametric tests such as Kruskal-Wallis test.

To infer the relationship between genes, we use Fisher’s Exact Test, Pearson’s Chi-square, or correlation coefficient. Given the genes in a cannabinoid biosynthesis pathway, we will be able to detect the changes in the gene-gene relationship and understand why the amount of a certain cannabinoid goes up or down under the experimental conditions.

**Tentative Schedule**

|  |  |
| --- | --- |
| Week 1 | Extract reference gene sequences |
| Week 2 | Run Kallisto to get transcript abundance |
| Week 3 | Run normality tests, then perform differential analysis with ANOVA or Kruskal-Wallis as baseline models. |
| Week 4 | Run with other statistical models, such as Nested ANOVA |
| Week 5 | Extract cannabinoid biosynthesis pathways |
| Week 6 | Run Pearson’s chi-square test to detect gene-gene relationship as baseline models |
| Week 7 | Run with other statistical models, such as Fisher’s Exact Test, or directional chi-square. |
| Week 8 | Revise models |
| Week 9 | Work on the manuscript |
| Week 10 | Revise the manuscript |

**Results**

We aim to develop a statistical framework to analyze the differential gene expression and differential biosynthesis pathways for industrial hemp. The experimental condition in this study is drought stress, which is an important challenge and affect agricultural productivity. Moreover, our statistical framework can be extended to study other experimental conditions, and other plants.

**References**

Muhammad Aslam, Mehtab, et al. “Mechanisms of Abscisic Acid-Mediated Drought Stress Responses in Plants.” *International Journal of Molecular Sciences,* vol. 23, no. 3, Jan. 2022, p. 1084. DOI.org (*Crossref*), https://doi.org/10.3390.ijms23031084.

Li P., Lin P., Zhao Z., Li Z., Liu Y., Huang C., Huang G., Xu L., Deng, Z., Zhang, Y., et al. Gene Co-Expression Analysis Reveals Transcriptome Divergence between Wild and Cultivated Sugarcane under Drought Stress. *International Journal of Molecular Sciences*, 2022, 23, 569. https://doi.org/10.3390/ ijms23010569

Li, H., Li, X., Sun, M. *et al.* Molecular characterization and gene expression analysis of tomato WOX transcription factor family under abiotic stress and phytohormone treatment. *J. Plant Biochem. Biotechnol.* **30,**973–986 (2021). <https://doi.org/10.1007/s13562-021-00723-8>

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol, 34(5), 525–527.