BIOSYSTEMS CHARACTERIZATION, MODELING AND OPTIMIZATION

Yazeli E. Cruz Rivera

A thesis submitted in partial fulfillment of the requirements for the degree of

MASTERS OF SCIENCE IN

INDUSTRIAL ENGINEERING

UNIVERSITY OF PUERTO RICO

MAYAGÜEZ CAMPUS

2018

Approved by:

_____	_____
Mauricio Cabrera-Ríos, PhD President, Graduate Committee	Date
_____	_____
Mayra Mendez, PhD Member, Graduate Committee	Date
_____	_____
Jaime Seguel, PhD Member, Graduate Committee	Date
_____	_____
Graduate Studies Representative	Date
_____	_____
Viviana Cesaní, PhD Chairperson of the Department	Date

Abstract

There is a need for interdisciplinary work to better understand, represent, and make competitive decisions in regards of biosystems. In this work biosystems understanding is sought through statistical characterization, representation through statistical modeling, and decision-making through mathematical optimization, all within a biological framework. This thesis applies these concepts in an orderly fashion through three biosystems in particular: biofuel derived from algae, hairy vetch as a cover crop for corn, and the identification of important genes in Alzheimer's Disease. The first case consists on supporting decision-making on lipid extraction to obtain biofuel from microalgae. The second case consists on defining how hairy vetch management and latitude affect the economics of hairy vetch. The last case consists of characterizing Alzheimer's disease through differentially expressed genes using microarray experiments. After finding these genes an interaction is established within each other. These cases were identified independently but can be organized in the proposed framework.

Resumen

El trabajo interdisciplinario es necesario para entender, representar y tomar decisiones competitivas al respecto de biosistemas. Este trabajo se basa en utilizar métodos de caracterización, modelaje estadístico y toma de decisiones por medio de optimización matemática dentro del marco biológico. Este trabajo aplica estos conceptos de una forma ordenada en tres biosistemas: biocombustible obtenidos de algas, *hairy vetch* como cocultivo del maíz y la identificación de genes importantes en la enfermedad del Alzheimer. El primer caso consiste en apoyar la toma de decisiones en la extracción de lípidos para obtener biocombustible de algas. El segundo caso consiste en definir cómo el mantenimiento de *hairy vetch* y la latitud afecta su economía. El último caso consiste en caracterizar el Alzheimer por medio de genes que se diferencian en su expresión usando microarreglos e interacciones entre sí. Los casos han sido identificados independientemente, pero son organizados de acuerdo con el marco propuesto.

Acknowledgements

This work was supported by the USDA-NIFA Award 2015-38422-24064 sub award 1000000920 granted to Dr. Krystel Castillo, Dr. Michael Persans, Dr. Hudson Devoe and Dr. Mauricio Cabrera-Rios. This award also provided me the opportunity to spend a summer at the Beltsville Agricultural Center (BARC) where I was able to collaborate with Dr. Steven Mirsky and DR. John Spargo. This work was also supported by the NIH MARC Assisting Bioinformatics Efforts at Minority Schools project 2T36GM008789.

I want to thank my advisors Dr. Mauricio Cabrera-Rios and Dr. Clara Isaza for providing me with this opportunity, guidance and support since my undergraduate years. I am grateful to have been part of the Applied Optimization Group since 2012. I would also like to thank my committee members Dr. Mayra Mendez and Dr. Jaime Seguel for the advice and input into this work.

Copyright ©

By

Yazeli E. Cruz Rivera

Table of Content

Contents

List of Figures

List of tables

Chapter 1. Introduction

1.1 Introduction

A biosystem is defined by the National Center for Biotechnology Information (NCBI) [1] as: "A group of molecules that interact in a biological system". Since we are surrounded by Biosystems and are ourselves Biosystems, there has been a large effort to better understand the behavior and find solutions to problems that may arise in their study. That is why the unifying themes in this thesis are characterization, modeling and optimization in the experiments of three different biosystems, as shown in Figure 1 and explained later.

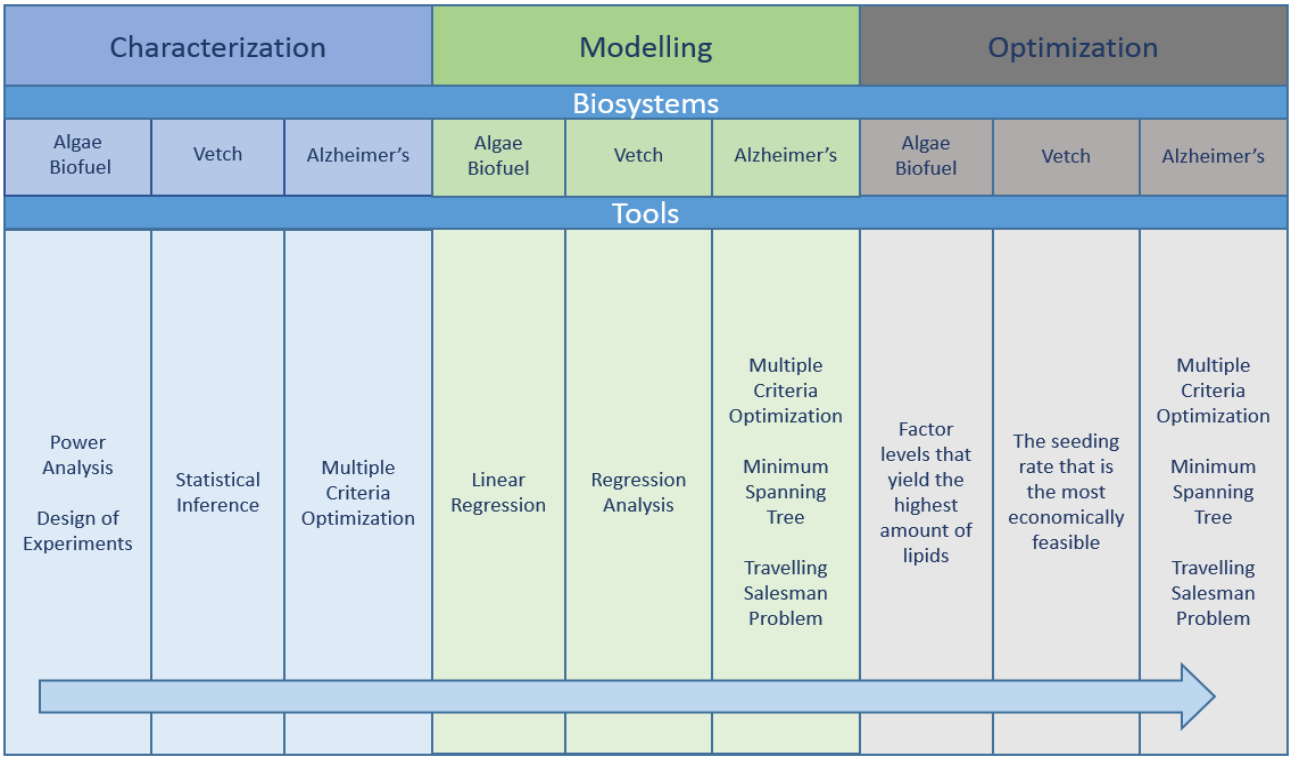


Figure 1: Thesis outline

Characterization refers to determining what are the key variables and interactions that define the biosystem's responses under study. Modeling refers to mimicking in a mathematical sense how the key variables and interactions relate to explain the responses. Optimization refers to the use of the models to arrive to values in the important variables that provide the best possible values for the biosystem's responses. These pieces are bonded in this work by an interdisciplinary effort of knowledge on biology and industrial engineering to support each study. Industrial Engineering tools (Figure 1) such as statistical inference, design of experiments and mathematical optimization are used here to approach three biosystems and their objectives: (1) Maximizing lipid production in algae through power analysis (2) Finding a Hairy Vetch seeding rate that maximizes corn yield and nitrogen production in the east of the United States through statistical analysis and (3) finding potentially expressed genes for Alzheimer's through mathematical optimization.

The first case in this thesis, the maximization of lipid production in algae is a collaborative effort with Dr. Krystel Castillo of the University of Texas at San Antonio, Dr Michael Persons and Dr Hudson Devoe from the University of Texas at Rio Grande Valley and Dr Clara Isaza from the Ponce Health Sciences University that comes as a response to the worldwide demand to move to clean and renewable biofuel. Algae was chosen due to its ability to grow in harsh conditions such as arid territories and wastewater [2] and has shown to have high biomass production in comparison to other energy products [3]. Nannochloris represents the algae strand from salt water and Ooscystis from fresh water. Lipids are derived from the biomass obtained from algae cells. Lipid extraction in this project consists of two experimental phases: algae cell growth and lipid extraction. In the first phase, it is necessary to maximize cell growth as well as the number of viable cells. The variables chosen for the initial experimentation are light, sucrose and salinity levels because they are important for cultivating algae. A design of experiments was proposed to

find the levels for these variables that could provide the highest lipid yield. An initial power and sample size assessment was done with the data collected by our collaborators in Texas before moving on to the execution of the experiment. In the second stage, the amount of lipid extracted per cell will be maximized as a proxy to efficiency. These three aspects –time to cell growth, number of viable cells, and efficiency- will be mapped to the experiment variables on each stage including formulation, setup and processing conditions.

The second biosystem is the subject of continuation of a study presented in Mirsky et al. [4] on how planting date and termination date influence the biomass of hairy vetch across seeding rate and latitudinal gradient across the eastern United States. This case is also the result of collaboration, this time with Dr. Steven Mirsky and Dr. John Spargo from the USDA Beltsville Agricultural Research Center. Hairy vetch is a legume cover crop [5]. As defined by the Sustainable Agriculture Research & Education cover crops are used to "Fix atmospheric nitrogen (N) for use by subsequent crops, reduce or prevent erosion, produce biomass and add organic matter to the soil". The use of cover crops has incremented greatly over the last couple of years due to the growing resistance that weeds are obtaining after year of chemical fertilizer treatments in the US and the shift towards organic agriculture practices.

Due to the gain in use of cover crops there is a need to establish a threshold where it is economically viable for the farmer to use the cover crop method compared to traditional types of mineral fertilizer. A very important component that comes into the equation is how much Nitrogen does the hairy vetch supply. The analysis was conducted through an initial design of experiments where there are different treatment levels of management (Seeding rates, Planting dates and Termination dates) and biomass from hairy vetch and the cash crop, in this case corn, was obtained from the different sites through 2011 to 2014. The data obtained from this experiment was analyzed

through statistics and mathematical optimization to answer the question: “How does hairy vetch management and latitude affect the economics of hairy vetch?”.

The third biosystem is Alzheimer’s disease, in particular the identification of important genetic components –akin to biomarkers- that are still elusive in this disease. The importance of an Alzheimer’s disease biomarker stems from its role in early detection, diagnosis, and prognosis of Alzheimer’s Disease. A biomarker is defined as a substance that can be biologically measured and it is related with an increased risk of a disease”. A good biomarker candidate is a molecule or cellular event that can be measured and is characterized by its distinct behavior in different states [6]. A step towards a biomarker is to find a differentially expressed gene. The search for biomarkers can be taken to a genetic level of relative expression in the presence and absence of a health condition. One of the options for this purpose is the use of microarray experiments, which can measure the relative gene expression of thousands of genes simultaneously. High throughput biological experiments like microarrays have been used to detect potential Alzheimer’s disease biomarkers maintaining the issue of selecting genes for normalization and parameters for the analysis as reviewed in Cooper-Knock et al 2012 [7].

This work relates to an ongoing effort within our research group at UPRM to find Alzheimer’s differentially expressed genes that are potential biomarkers and elicit a possible signaling path among them using publicly available microarray data, in particular for this work, first reported by Dunkley et al 2006 [8]. The list of biomarkers is found through the application of multiple criteria optimization (MCO) following the methods that our group described in Watts-Oquendo et al 2012 [6], Sánchez-Peña et al 2013 [9], Lorenzo et al 2015 [10] and Camacho-Caceres [11] . The characterization of the signaling path in Alzheimer’s is carried out using the well-known Travelling Salesman Problem (TSP) combinatorial optimization formulation [12] as was used in

the paper by Lorenzo et al 2015 [10] and Isaza et al 2017 [13]. An important and distinctive characteristic of the work in Alzheimer's presented in this thesis is the evaluation of alternative schemes to build the TSP problem using probabilities instead of correlation values.

These three cases in the study of biosystems, all with different purposes, are here unified by the need of statistical characterization, empirical modeling and mathematical optimization, which will be approached with tools from the Industrial Engineering realm. The applicability of the analysis framework is strongly supported by the fact that the three cases were identified independently in three research centers and were approached similarly.

1.2 Objective

The objective of this thesis is to provide an analysis structure based on statistical characterization, empirical modeling and mathematical optimization to support the study of biosystems. To this end, three studies are conducted: a design of experiments biofuel production from algae, characterization and viability of hairy vetch, and determination of important genetic components of Alzheimer`s disease based on probability to later validate through a biological structure.

1.3 Motivation

The growing interdisciplinary efforts in biology and engineering call into action the need for reproducibility. Reproducibility is defined by the National Science Foundation as [14]: “the ability of a researcher to duplicate the results of a prior study using the same materials and procedures as were used by the original investigator. So in an attempt to reproduce a published statistical analysis, a second researcher might use the same raw data to build the same analysis files and implement the same statistical analysis to determine whether they yield the same results.”.Reproducibility is needed to find efficient and accurate solutions. The need of reproducibility in science has been a popular topic in the past years due to recent studies that have shown alarmingly low reproducibility percentages in science studies. Some of the most important studies that go in depth in this topic are Begley et. al [15] and the open science collaboration work on estimating the reproducibility of psychological science [16]. Begley et. al [15] estimated that there is approximately 85% of biomedical research that is not reproducible. By having additional experimental runs, the Open Science Collaboration [16] was able to find that only 36% of the results were significant instead of the 97% originally published by the preceding studies of

psychological science. Companies that include industrial laboratories such as Amgen and Bayer have shown reproducibility rates of 25% and 11% from studies in related areas to oncology [17]. Reproducibility is possible by formally combining the series of steps advocated in this work: understanding the system (Characterization), recreating the system's behavior (Modeling) and obtaining the best response for the system (Optimization). Each step includes existing analysis tools such as statistical inference, analysis and design of experiments and mathematical optimization that can be used to create the capability of better decision making in planning for Biosystems framework. Characterization, modelling and optimization provide a better understanding of a noisy system. The Biosystems framework is to be referenced for future use. An outline of the motivation is shown in figure 2.

Figure 2: Motivation outline

1.4 Scope

The scope of this thesis includes three cases of different Biosystems studies, each of them identified independently in three research groups. The first one aims to maximize the amount of lipid yield from the Nannochloris and Ooscystis algae through design of experiment and mathematical optimization. The second case aims to find the hairy vetch seeding rate that provides the most of corn yield through the eastern United States under the lowest cost scheme for farmers through statistical inference. Lastly, the third case consists on identifying the highest differentially expressed genes for Alzheimer's Disease through network optimization models. This work with AD proposes and evaluates using probabilities in lieu of linear correlations as heretofore explored in our research group.

1.5 Thesis Organization

This thesis is organized as follows: The second chapter is divided in two parts, first a compendium and overview of relevant literature concerning issues on Biofuel derived from algae, hairy vetch as a cover crop seed and Alzheimer's disease. This also includes general background information of the proposed methodologies of this thesis. The third chapter is dedicated to the biofuel production study from its characterization of the Biosystem through statistical inference, modeling through regression analysis to its optimization. Chapter 4 follows a similar organization but instead it is focused on the USDA hairy vetch/corn study. Chapter 5 is focused on the Alzheimer's case study and its proposed methodologies of the Traveling Salesperson Problem(TSP) from genes found through Multiple Criteria Optimization(MCO). The TSP, as an original idea in this work is explored using probabilities as opposed to linear statistical correlations among pairs of genes. For completeness, a comparison of the proposed methodology to GeneMANIA, a program that is focused on constructing gene networks. The sixth and last chapter constructs the general conclusions of this thesis and routes future work. A timeline is provided in figure 3.

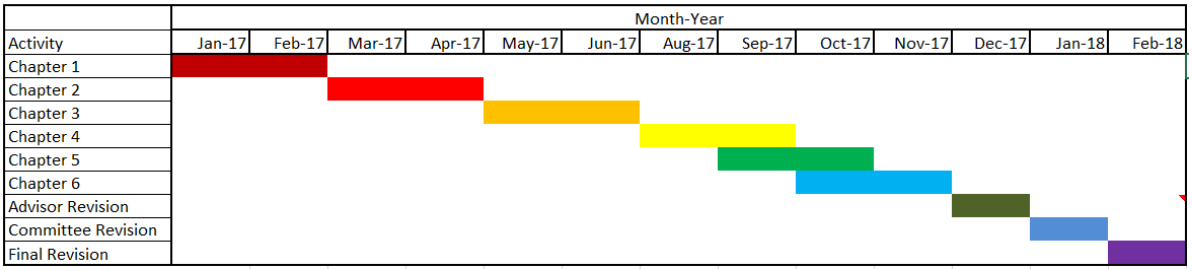


Figure 3: Gantt chart for thesis

Chapter 2. Literature Review

This chapter includes the individual literature review for each of three cases. Each major section has a series of sub sections that better details the background for each case.

2.1 Biofuel

According to the Biofuel Organization, biofuel is defined as any fuel whose energy is obtained through a process of biological carbon fixation. Biofuel -also known as agro-fuel- is derivate from biomass or bio waste (any matter derivate from plants or animals). Those are divided in two generations. Those in the first generation are made from sugar, vegetable oil and starch. Those in the second generation ones are greener fuels made from sustainable feedstock [18].

First and second generation fuels are used for many purposes, but the main use is for transportation. In 2012, biofuels accounted for roughly 7.1 percent of the total transport fuel consumption, or 13.8 billion gallons, unchanged from the previous year [19]. Based on projections, in 2017 the amount of water withdrawn for biofuel production would increase by 74% if agricultural practices remain the same [20]. Biofuel remains to this day the only widely available source of clean, renewable transportation energy [21].

Even though biofuel is the only "clean energy" available it is still a pollutant. Biofuel's combustion produces CO_2 will happen to the atmosphere, it is assumed to be the same as plants or algae which comes absorbed during their growth, so they both released into the atmosphere practically same amount of CO_2 that was consumed [22], [23]. It is biodegradable, non-toxic, and typically produces about 60 % less carbon dioxide emissions [2]. Smog emissions are 65% lower than diesel derived from petroleum. The upcoming sub-section will go more into detail intro second generation biofuel production, specifically biofuel derived from microalgae.

2.1.1 Microalgae for Biofuel

Microalgae are a diverse group of single-celled organisms that have the potential to offer a variety of solutions for our liquid transportation fuel requirements through several avenues. Algae can produce biomass very quickly and have the capacity to produce energy- rich oils [24], [25]. Most of them have been found that accumulate high oils levels in the total dry biomass [25]. Groups considered in microalgae are: diatoms, green algae, golden brown, prymnesiophytes, eustigmatophytes [24]. Since they are single-celled organisms that duplicate by division, high-throughput technologies can be used to rapidly evolved strains. By using these organisms, time is reduced in processes that take years in crop plants, to a few months in algae. Algae have a reduced impact on the environment compared with terrestrial sources of biomass used for biofuels [26].

Biofuel produced from algae not only minimizes land it can also remediate waste streams. Potential waste streams include municipal wastewater to remove nitrates and phosphates before discharge, and fuel gas of coal or other combustible-based power plants to capture sulfates and CO_2 [22], [23]. Algae production strains also have the potential to be bioengineered, allowing improvement of specific traits [27], [28] and production of valuable co-products, which may allow algae biofuels to compete economically with petroleum. These characteristics make algae a platform with a high potential to produce cost-competitive biofuels [29]. Thus, it is necessary to identify the factor levels for light, salinity and sucrose in the chosen algae strains that will ensure the highest amount of lipid yield through reproducibility to remain cost effective.

2.1.2 Biofuel production

Biofuel production and consumption is dominated by the United States and Brazil. In 2011, the two represented 70% of global biofuel consumption and 74% of global production. Both fuel

types have been growing for use and consumption during the past decade [24]. The process to obtain biofuel from algae is divided in five stages: microalgae cultivation, photobioreactors, harvesting, biomass processing and biofuel production.

Microalgae cultivation can be done in open or closed ponds. Open ponds can be naturals waters, tanks or containers. Closed ponds systems cost more than open ponds and they allow more species for cultivation. In closed ponds systems microalgae can increase the amount and concentration of carbon dioxide which increases the rate of growth of microalgae [30].

The next stage is the use of photobioreactors. They are closed and expensive systems of cultivation with a high biomass productivity. They provide uniform and efficient lighting, plus a cleaning system. Principal factors are incorporated such as algae species, temperature, nutrients, water, pH, CO_2, among others, to control the system.

There are several types of harvesting methods including centrifugation, filtration, ultrafiltration, sedimentation, chemical flocculation, and flotation. The concentration of the micro-algae from the various harvesting methods can vary from 0.5 to 27% dry weight. Further dewatering or drying may be required prior to energy extraction from the microalgae. The centrifugation method is the most effective method in terms of dry solid output concentration with a range of 10-22%. Other advantages are the low cost, reduce energy input, and cost of subsequent stages [31].

The process that transforms the lipids extracted from the algae to biofuel is called transesterification. Bumara & Varma [32] state that : “Oil extracted from the algae is mixed with alcohol and an acid or a base to produce the fatty acid methylesters that makes up the biodiesel”.

2.1.3 Design of Experiments for Biofuel

Design of experiments (DOE) is defined as a systematic and efficient method to determine the relationship and the variations between factors affecting a process and the output of that process [33]. It is a very powerful tool for the optimization of biotechnological process and mathematical modeling of the statistical relevant terms at a given probability level. DOE provides the experimentation of different levels of the chosen variables to achieve the main goal to extract the largest amount of lipids per algae cell [33].

Few studies in the past have included DOE when it comes to biofuel production. In Hallenbeck et. al [34] Nannochloropsis gaditana algae was under study using a DOE-Response Surface Method (RSM) in order to maximize biomass production, lipid content and total lipid production with light intensity (250, 325 and 400 I E/m^2), inoculum size (50000, 225000 and 400000 counts/1 L) and CO_2 concentration (0.04%, 4.52% and 9%) as factors. A 3^k factorial Box-Behken design was generated for the cultures growth experiment with a total of 15 runs and center point that was replicated 5 times. The response variables were cell counts, chlorophyll and lipid fluorescence. The response surfaces shown in this paper were incomplete. This is because only one of the three variables to one level were fixed, instead of showing the entire surface area. There was no initial power analysis for the number of replicates needed and they used a Box-Behken design to estimate second degree quadratic polynomials. A power analysis is needed to find the minimum sample size to replicate the experiment in other to explain random deviation.

In Wei et. al [35] a L_9 test was performed with the Nannochloropsis Oculata algae for ten days. The factors included limited nitrogen supplementation (0,0.22 and 0.44 mmol N L^{-1}), culture temperature (10, 20 and 30°C) and high iron concentration (1.2 x 10-2, 1.2 x 10^{-1} and 1.2 mmol Fe L^{-1}). An ANOVA was used to identify differences among groups. When differences among

groups were identified, the means were compared through Duncan's multiple-range test. Three outputs were measured: total lipid yield (dry weight of lipid cells in harvest), neutral lipid/total lipid ratio and lipid yield. Their highest amount of total lipid content of 60.44 +/ 0.68% dry weight was with treatment 6 (0.22 mmol N L^{-1}, 1.2 mmol Fe L^{-1} and 10°C), the highest neutral lipid/total lipid ratio with treatment 3 (0 mmol N L^{-1}, Iron 1.2 mmol Fe L^{-1} and 30°C) and the highest lipid yield. There were only nine (taken from Taguchi) experimental points runs and one replicate. A power analysis was not pursued in this experiment either.

Lastly, In Bohnen and Bruck [36] the authors applied normality tests and Analysis of Variance within a p-value of 0.05 for the null hypothesis on data from Massart et. al [37]. These data were obtained from a face centered response surface area from three factors at three levels: light intensity (100, 200 and 300 μmol m^{-2} s^{-1}), potassium nitrate (1, 2 and 3 g l^{-1}) and sodium chloride (10, 30 and 50 g l^{-1}) for biomass productivity (mg l^{-1} day^{-1}) and lipid content (wt% of dry mass). By using the Shapiro-Wilk test for normality, biomass productivity was found to follow a normal distribution, but lipid content did not. Bohnen and Bruck [36] expanded Massart's et. al [37] raw data equation modelling on biomass productivity by re-analyzing the data thus reducing the original Biomass productivity = $k_0 + k_1[KNO_3] + k_2[NaCl] + k_3[Light] + k_{11}[KNO_3]^2 + k_{22}[NaCL]^2 + k_{33}[light]^2 + k_{12}[KNO_3][NaCL] + k_{13}[KNO_3][Light] + k_{23}[NaCL][Light]$ to an equation with a smaller mean square error Biomass productivity = $k_0 + k_2[NaCL] + k_3[Light] + k_{33}[Light]^2$ and stated that nitrogen source is statistically insignificant with a α=0.05. Since lipid content raw data was found to not follow a normal distribution, a Box-Cox transformation was done. The result was the final model Ln(Lipid Content) = $k_0 + k_2[NaCL] + k_{22}[Light]^2 + k_{33}[Light]^2$ that obtained a better R^2 than the raw data models. Bohnen and Bruck [36] is the only paper where design of

experiments pre-requisites such as a Box-Cox transformation for initially non-normal data are applied even though it is an overview of such.

The one key difference of this project to the ones mentioned before is planning. An operation characteristic curve (OC curve) must be done before a DOE to find the minimum sample size to estimate the minimum difference with a power of at least 80% ($\beta = 0.80$). In the initial phase, three factors will be tested (light, salinity and sucrose). In this case, a central composite design was chosen due to the information that can be provided by the star points. After executing the experiment, an ANOVA is done to obtain the coefficients to represent the functions for absorbance, absorbance slope and the lipid fluorescence/ cells ratio graphically.

2.2 Hairy vetch

This work is a continuation of Mirsky et. al [4] of how hairy vetch biomass is affected by seeding rate, latitude, seeding date and termination date on Eastern U.S. (Massachusetts, New York, Pennsylvania, Maryland and North Carolina). Mirsky et. al focused on what is the seeding rate that provided the highest hairy vetch biomass by a given seeding date and termination date. The next step is to find how much is the dollar per pound of plant available Nitrogen (PAN) provided by hairy vetch across a latitudinal gradient and to optimize cost-benefit of hairy vetch based on corn performance across a latitudinal gradient. The focus of this literature review will be on past studies of PAN estimations. Another way to estimate PAN is through fertilizer equivalent/nitrogen fertilizer replacement value (NFRV). NFRV is the equivalent to the amount of nitrogen provided by the cover crop compared to common nitrogen fertilizers such as Urea, Ammonium sulfate, Ammonium nitrate and Anhydrous ammonia.

This literature review includes an introduction of legume cover crop, nitrogen derived from the cover crop, literature estimates of PAN, basic costs and how we are going to use the empirical data provided to establish a generic estimation of how much Nitrogen (N) content is provided by the biomass that is passed on to corn.

2.2.1 Legume Cover Crop

Cover crops are defined by the Minnesota Department of Agriculture[38] as : “grasses, legumes or forbs planted to provide seasonal soil cover on cropland when the soil would otherwise be bare—i.e., before the crop emerges in spring or after fall harvest.”. In this project, the focus is the winter annual legume cover crop, hairy vetch. Hairy vetch is planted mostly in the fall to

provide soil protection during the winter. Sustainable Agriculture Research & Education [39] listed many of the benefits that legume cover crops provide are:

- Fix atmospheric nitrogen (N) for use by subsequent crops
- Reduce or prevent erosion
- Produce biomass and add organic matter to the soil
- Attract beneficial insects

Specifically, hairy vetch has many added benefits such as growing well in areas with hard freezing. It also produces a large amounts of vegetation and up to one hundred pounds of nitrogen per acre for the upcoming cash crop [39]. The complete amount of nitrogen produced by hairy vetch it is not completely passed on to the next crop, which in this case is corn. This is due to the nitrogen cycle as shown in Figure 4.

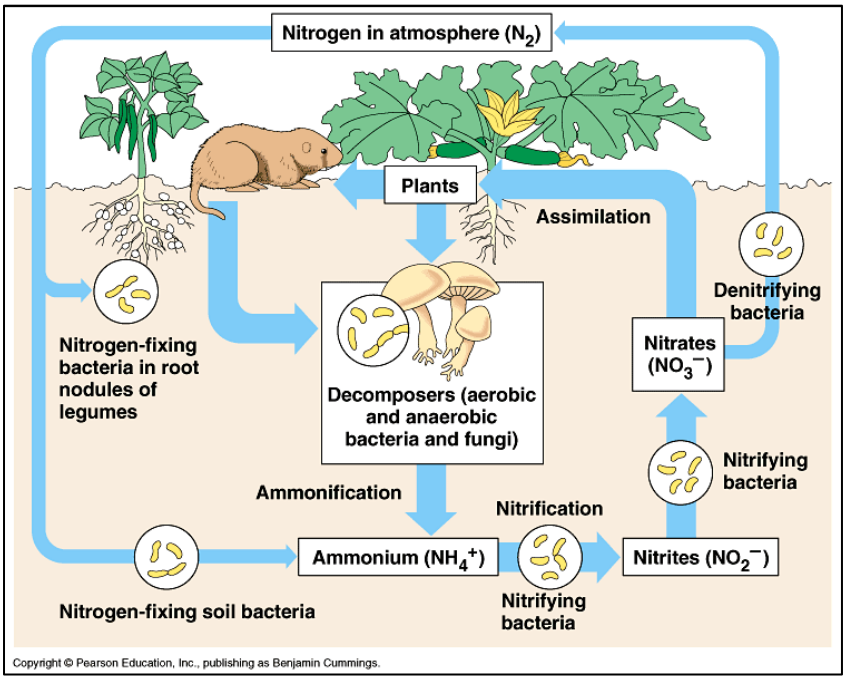


Figure 4: Nitrogen Cycle [40]

Nitrogen can be lost due to different causes in the soil system. The following list of nitrogen loss causes and definitions are provided by the University of Minnesota Extension [41] :

- Leaching: The loss of soluble NO3--N as it moves with soil water, generally excess water, below the root zone.
- Denitrification: The process by which bacteria convert NO_3--N to N gases that are lost to the atmosphere.
- Volatilization: N is lost as ammonia (NH_3) gas.
- Crop removal: The removal of crops is the largest cause of nitrogen loss in soil.
- Soil erosion and runoff: This can be prevented with fertilizer and conservation tillage.

2.2.2 Hairy vetch on Corn yield plant available Nitrogen and fertilizer equivalent

Crude PAN

Crude PAN in the Beltsville Agricultural Center (BARC) is estimated by the following formula:

$$\text{N uptake in corn from vetch} = \frac{N\ obtained\ by\ vetch - N\ obtained\ in\ control\ group}{Total\ N\ in\ vetch\ tissue} \quad (2.1)$$

Decker et al. evaluated winter cover crops the Coastal Plains and Piedmont for no-tillage corn from 1986 to 1988 in Maryland [42]. The seeding for the Coastal Plain was done during late September/early October and terminated during late April/early May. Furthermore, Piedmont's seeding was during mid/late September and termination during early/mid May. With formula (1) it was estimated that the Coastal Plain contained a crude PAN percentage of 3.25% in 1986,

36.61% in 1987 and 19.57% in 1998 with an average of 25.73%. Holderbaym et. al [43] study was also done in the Poplar Hill and Forage farm in the Beltsville Agricultural Center (BARC) in Maryland with a seeding rate of 28 kg ha^{-1}. The crude PAN estimation was 26.50% for Poplar Hill in 1984. In 1985, Poplar Hill obtained 42.71% and Forage Farm 7.37%. In both papers the soil N uptake values where not reported, thus yielding lower crude PAN percentages than average.

From Kuok & Jellum's study in Pennsylvania [44] using (1) the crude PAN estimate for 1991 was 44.39%, for 1992 was 51.27%, for 1993 was 80.29%, for 1994 was 33.41%, for 1995 was 19.82% and 41.71% for 1996 with a seeding rate of 36 kg ha-1. Clark et. al [45] hairy vetch had a seeding rate of 28 kg ha^{-1} . With formula (1) the Coastal Plain of Maryland obtained an estimated 1.20% to 4.48% of crude PAN from 1990 to 1991. The piedmont obtained 11.51% to 88.14% from 1990 to 1991 being the early May + Mid-May termination date the one that contained the highest discrepancies in the amount of percentage.

Fertilizer equivalent

One of the other methods used to estimate N derived from legumes. Larue & Patterson [46] define this as: "The N added to the soil by the legume crop gives a yield in the successive crop that is compared to the response to different levels of added N fertilizer.".

Grain yield regression analysis was also done in Decker et. al to find a fertilizer replacement value for hairy vetch using a quadratic plus-plateu to find economic optimum yields [42]. Since this study was published in 1994 the values used for fertilizer nitrogen (FN) cost were \$0.55 kg^{-1} (0.25 lb^{-1}) and corn at \$98.20 Mg^{-1} (\$2.50 bu^{-1}). The FN rate for economic optimum yield in the Coastal Plain for hairy vetch was 65 and 125 kg ha^{-1} for Piedmont. Legume cover crops had the

lowest FN rate for economic optimum yield out of all the cover crops and they also maximized corn yield.

Kettering et. al [47] summarized the works of Samson et. al [48], Dou & Fox [49], Sarrantonio & Scott [50] and Stute & Posner [51] in northeastern states such as New York and Pennsylvania. The Nitrogen fertilizer replacement values and the corresponding planting date for hairy vetch are presented in Table 1:

Location	Cover Crop		Tillage	Seeding Time	NFRV (kg ha^{-1})	Reference
Ontario	Hairy vetch		CT	Existing winter wheat fields"	~112	Samson et al. [48]
NY	Hairy vetch		CT	Late Aug.	52	Sarrantonio and Scott [50]
NY	Hairy vetch		NT	Late Aug.	17	Sarrantonio and Scott [50]
PA	Hairy vetch	yr1	CT	Early Aug. after wheat harvest	103	Dou and Fox [52]
PA	Hairy vetch	yr2	CT	Early Aug. after wheat harvest	30	Dou and Fox [52]
PA	Hairy vetch	yr1	NT	Early Aug. after wheat harvest	149	Dou and Fox [52]
PA	Hairy vetch	yr2	NT	Early Aug. after wheat harvest	15	Dou and Fox [52]
PA	Hairy vetch	yr1	NT	Early Aug. after wheat harvest	48	Dou and Fox [52]
PA	Hairy vetch	yr2	NT	Early Aug. after wheat harvest	57	Dou and Fox [52]
WI	Hairy vetch		CT	Mid-late Apr. with oat	>108	Stute and Posner [51]
WI	Hairy vetch		CT	Late Jul.-early Aug. after oat	73+	Stute and Posner [51]

Table 1: Kettering et. al [47] *summarized NFRV for northeastern states*

Clark et. al [53] evaluated spring management for hairy vetch during 1990 to 1991 to estimate fertilizer equivalents. Figure 5 shows that FE for hairy vetch is about 100 to 140 kg N ha^{-1}. The percentage of the fertilizer nitrogen replacement value (2.2) is 44% for 1990 and 50% for 1991.

$$\% \text{ of nitrogen fertilization value} = \frac{\text{Fertilizer N upake}}{\text{Corn N uptake}} \times 100 \quad (2.2)$$

In Spargo et. al [54] study of legume cover crops and organic amendments to meet the nitrogen (N) needs in corn at Beltsville Agricultural Research Center (BARC). The percentage of the fertilizer nitrogen replacement value was 81.81% in 2009 and 92.73% in 2010. In Wagger's study

of cover crop management in before planting corn in North Carolina there is an estimated fertilizer percentage equivalent to 35% in 1984 and 24% in 1985 for hairy vetch [55].

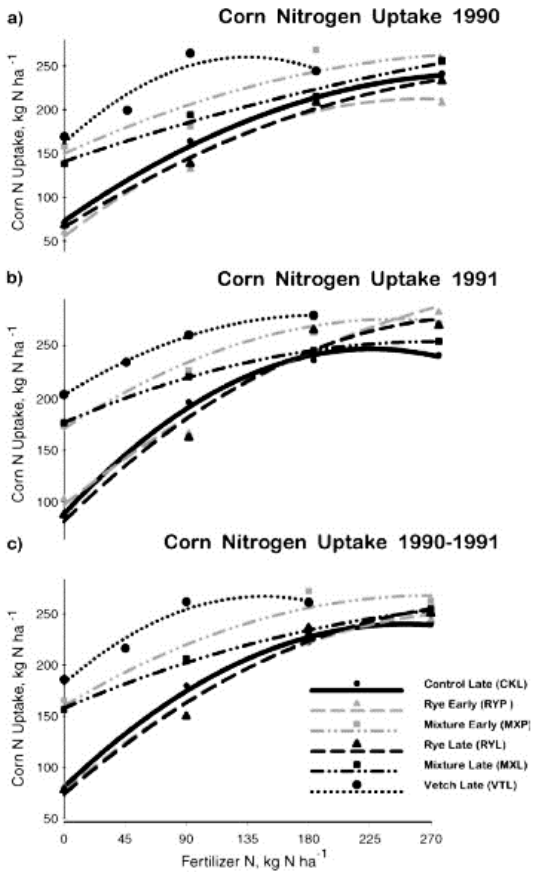


Figure 5: Corn N uptake (kg N ha^{-1}) from Clark et. al [53]*, originally figure 2c of the paper.*

Cornell University's agronomy fact sheet literature review of the nitrogen benefits of winter cover crops found that hairy vetch obtained a NFRV over 70 lbs N/acre in 50% of studies and a NFRV of 50 lbs N/acre in 80% of the studies in the northeast of the United States [56]. Blevins et. al [57] study of cover crops in Kentucky yielded a NFRV of 65 to 135 kg ha^{-1}. Ebelhar's [58] studies of hairy vetch as a cover crops obtained NFRV from 90 to 100 kg ha^{-1}. By finding the most accurate NFRV and crude PAN values, an equitable economical comparison between hairy vetch

and nitrogen fertilizer will be provided. This comparison ties directly to choosing the best seeding rate by the biomass provided by the seeding rate and the nitrogen provided by that biomass that is later passed on to corn.

Lastly, the NFRV that are used for decision making in chapter 4 are obtained from Spargo's [54] study of how corn grain yield is affected by hairy vetch biomass. The NFRV was found by first finding the Nitrogen uptake. This is done by taking the with vetch by using the equations in figure 6. The with vetch equation value of 249 is equaled to the without vetch equation ($249=141+1.03x-0.00163x^2$) and solve for x (Nitrogen uptake). The value of x is 132.74 kg/ha^{-1}. To find the nitrogen replacement value formula 2.3 was used. The N fertilizer value for 2010 provided in the paper is 152 kg/ha^{-1}. The resulting NFRV is 0.88.

$$NFRV=\frac{Nitrogen\ Uptake}{N\ fertilizer} \quad (2.3)$$

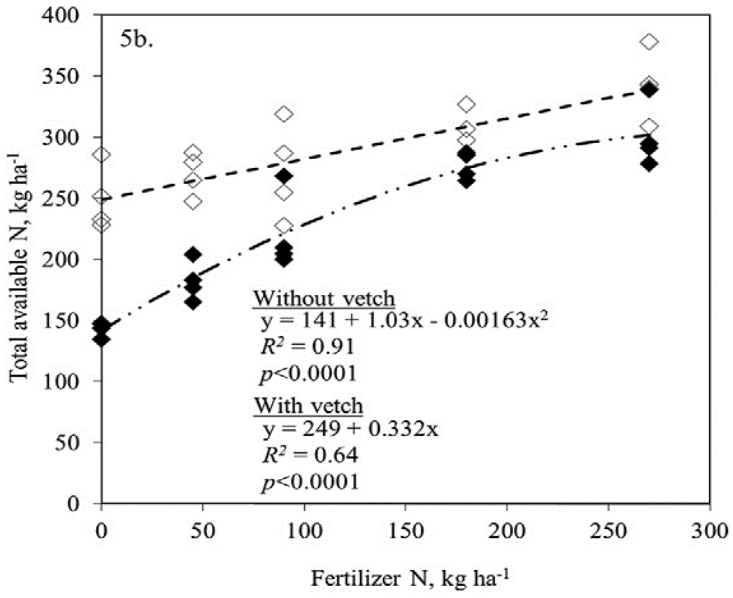
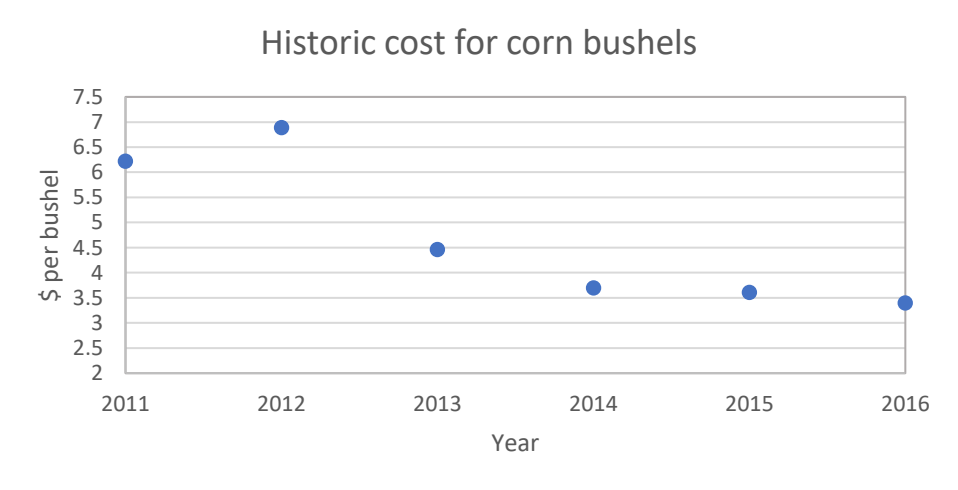


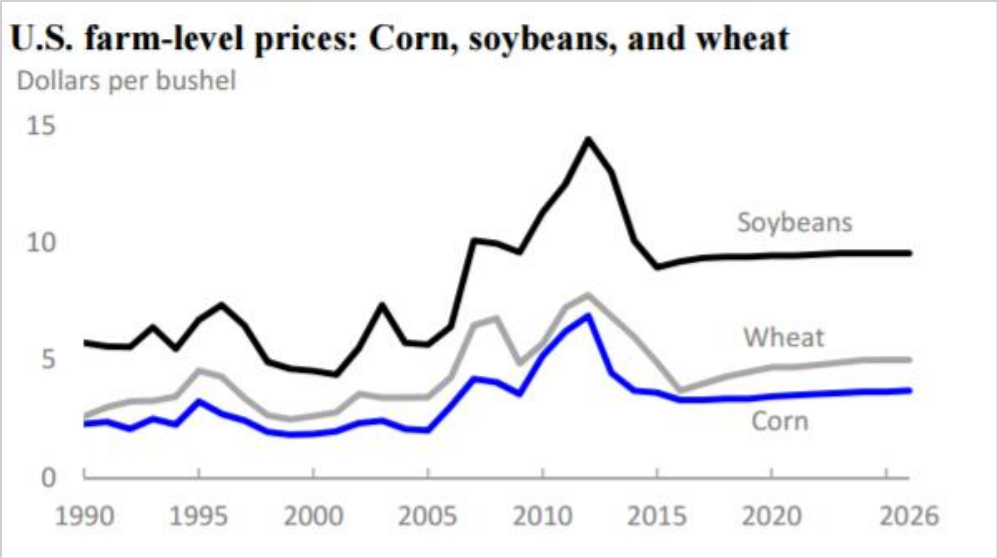
Figure 6: Total N uptake in response to fertilizer N and without a hairy vetch cover crop in 2010 [54]

2.2.3 Costs

An important part of this case study is the costs that come with planting hairy vetch. These costs will help find the hairy vetch seeding rate that is the most cost efficient for the farmers. Costs that are included are corn bushel, nitrogen fertilizer and hairy vetch seed. Graph 1 shows the historic costs per bushel of corn provided by the USDA [59]. The prices for a bushel of corn have been steadily declining since 2012. Graph 2 shows the forecast for the cost per bushel for corn until 2016 [60]. The forecast shows that the cost per bushel of corn will increment slowly for the next ten years.

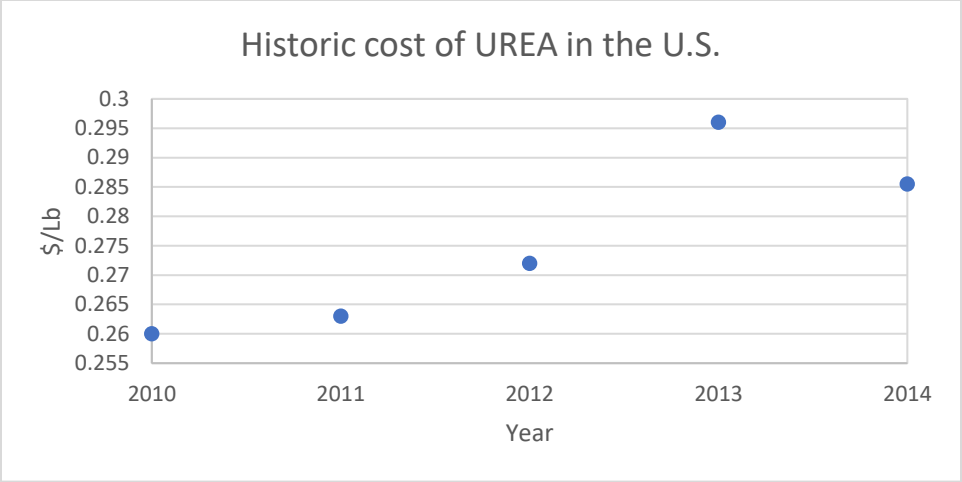


Graph 1: Historic costs of corn bushels in the U.S. [59]



Graph 2: Forecast per corn bushel until 2026 [60]

There is no historic cost reports or forecasts for the cost of hairy vetch seeds available in the USDA site. Spargo et. al [54] concluded that PAN from hairy vetch had a cost of \$1.33 kg^{-1} based on cost \$4.21 kg^{-1} for vetch seed and a seeding ratio 34 kg ha^{-1}. Current cost for hairy vetch seed is 2.05 \$/lb [61] . Incidentally, Graph 3 shows the historic cost (\$/Lb) of urea in the U.S. provided by the USDA [62]. The reports that include the cost for 2015 have not been released yet. Also, there is not readily available forecasts of the cost of urea in the future. The current cost of urea in the U.S. is 0.2683 \$/lb [62].



Graph 3: Historic costs of hairy vetch in the U.S. [62]

2.3 Alzheimer's Disease

Alzheimer's disease is defined by the National Institute of Health (NIH) as: "an irreversible, progressive brain disorder that slowly destroys memory and thinking skills, and eventually the ability to carry out the simplest tasks." [63]. The following is a list that the NIH compiled of effects that Alzheimer's have on patients [63]:

- Amyloid plaques (Abnormal clumps)
- Tangle bundles of fibers
- Loss of connections between nerve cells

And it slowly destroys:

- Memory
- Thinking skills
- The ability to carry out the simplest tasks

As also stated by the NIH, the first symptoms tend to appear approximately when the patient is in their mid-60s. There is as estimate of 5 million patients in the United States and 44 million patients worldwide [64]. It is the sixth leading cause of death in the U.S. and it is projected that by 2050 there will more than 16 million patients in the U.S. only.

There are three major states of Alzheimer's disease: Mild, Moderate and Severe. In the mild state people undergo large memory loss. In the moderate stage, the damage grows deeper into parts of the brain that command reasoning, language, conscious though and sensory processing [63]. The last stage, severe, the brain tissue has grown notably smaller and plaques and tangles are noticeable. The patient at this point is mostly bed ridden until the body stops working.

Understanding how Alzheimer's occurs and why it occurs is, indeed, a cross disciplinary endeavor undertaken by our research group at UPRM. An analysis pipeline described in [11] and [10] has been adopted here to approach the identification of genes that are deemed differentially expressed in the presence of Alzheimer's disease through Multiple Criteria Optimization (MCO) [11] and structured in a mathematical graph through the Traveling Salesman Problem integer optimization formulation [10]. With the support of Prof. Clara Isaza (Ponce Health Sciences University), a biological interpretation of the results has also been included for completeness as well as a previously established comparison of our analysis pipeline with a somewhat similar code available online. Besides the natural question of representation of a biological signaling path as a cycle -which is an ongoing effort to date in our group-, an important one is the possibility of using information beyond the usual statistical linear correlation among pairs of genes. To this end, an original contribution of this work includes the construction of a model for Alzheimer´s using probabilities to relate pairs of genes as an alternative configuration. This variation is here presented and assessed.

2.4 Review

In summary, the literature review for the first case of deriving biofuel for microalgae is the comparison of factors, strains and yield obtained from past studies. For the second case, hairy vetch as a cover crop, the values for NFRV and the costs for nitrogen fertilizer, corn bushels and the cost for the hairy vetch seed are to be used for a comparison of the between the nitrogen that hairy vetch provides and the nitrogen that is provided by fertilizer. Lastly, Alzheimer's, is an overview of the gravity of the illness and this works' contribution in the analysis pipeline that the Applied Optimization Group has been working on the past years.

Chapter 3. USDA biofuel study

This chapter focuses on biofuel production biosystem characterization, modelling and optimization. The main goal in this in this study was to maximize the extraction of lipids in algae to produce more biofuel. Lipid extraction in this project is seen as a two-phase experimental process:

1. **Algal cell growth:** Necessary to minimize the cell growth time while maximizing the number of viable cells.
2. **Lipid extraction:** The amount of lipid extracted per cell will be maximized as a proxy to efficiency.

3.1 Characterization

An initial experiment was carried out to observe cell and lipid growth for Nannochloris and Ooscystis in summer 2015 and 2016 by Michael Persan's research group at University of Texas, Brownsville. This data was used for the creation for initial ideas and characterization. The unit of time used is days.

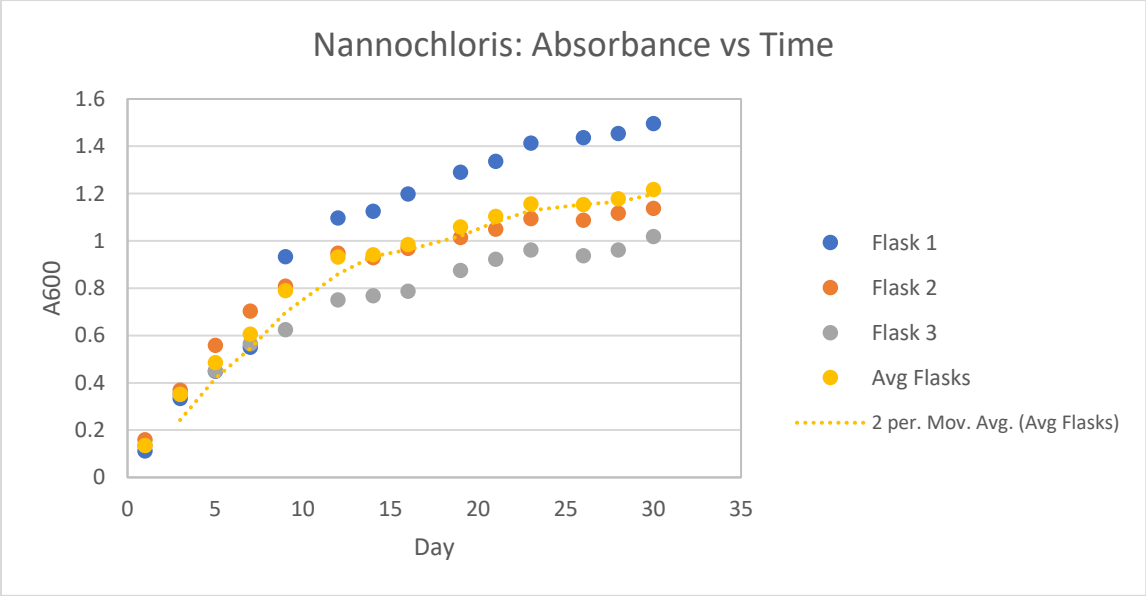
3.1.1 Characterization – Absorbance vs Time

Absorbance is the measurement for the number of cells. The actual data for Nannochloris is represented in graph 4 and 5 in F/2 + NH4CL + K2HPO4 and Ooscystis in 0.5% sucrose medium graph 5 and 6. There were three samples for each type of algae. Figure 8 represents the ideal behavior of this graph. The two aspects observed in this type of graph was:

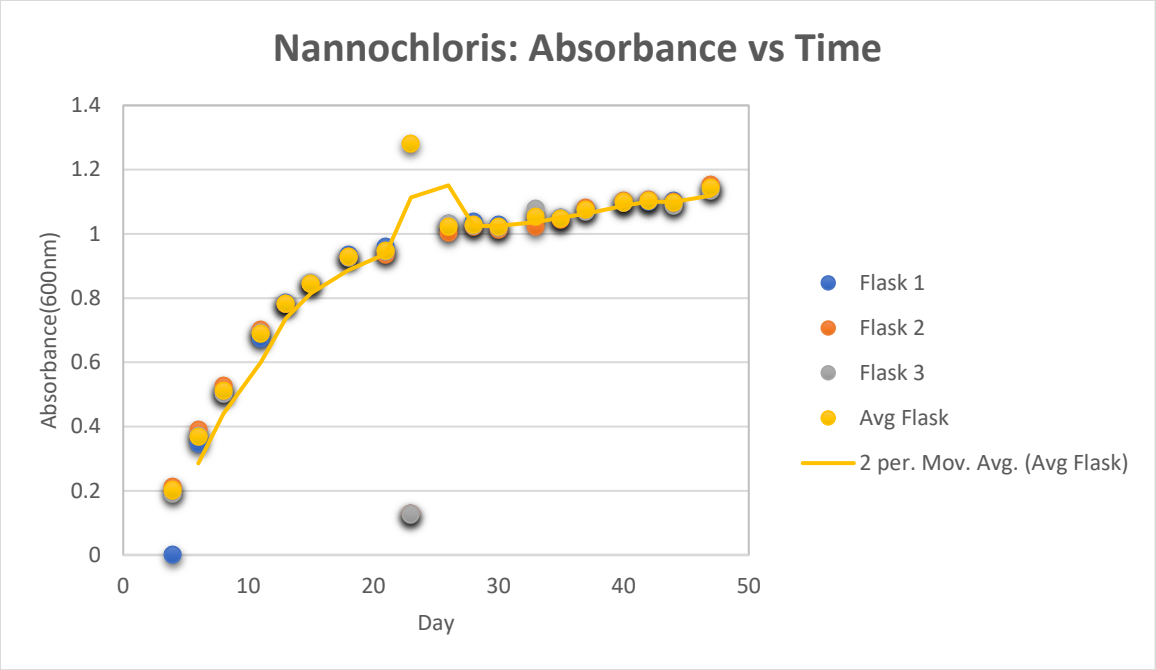
F_1= Max absorbance

F_2 = Absorbance Slope

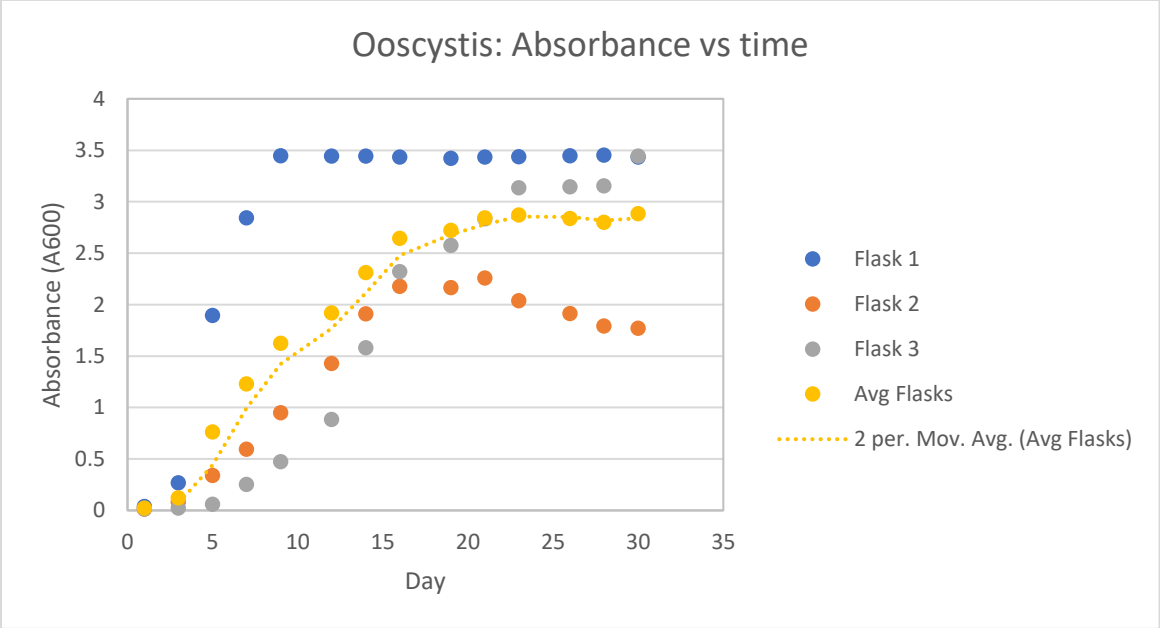
F_1 is about growing the most number of cells and F_2 is the slope that represents growing the most number of cells in the shorted time as possible. As seen in the graph 4 and graph 5, Nannochloris does have a similar behavior to figure 7 (Except for an outliner that was due to an incorrect reading of the data that day, as verified by telephone by Dr. Persan) in all of the runs. As shown in graph 6 and graph 7 Ooscystis does not have a similar behavior to figure 7. Graph 7 shows that the Ooscystis example never reached a maximum plateau during the length of the experiment. That is why it is important that an experiment is reproducible, so one can better understand the behavior and trace back to why irregularities happen.



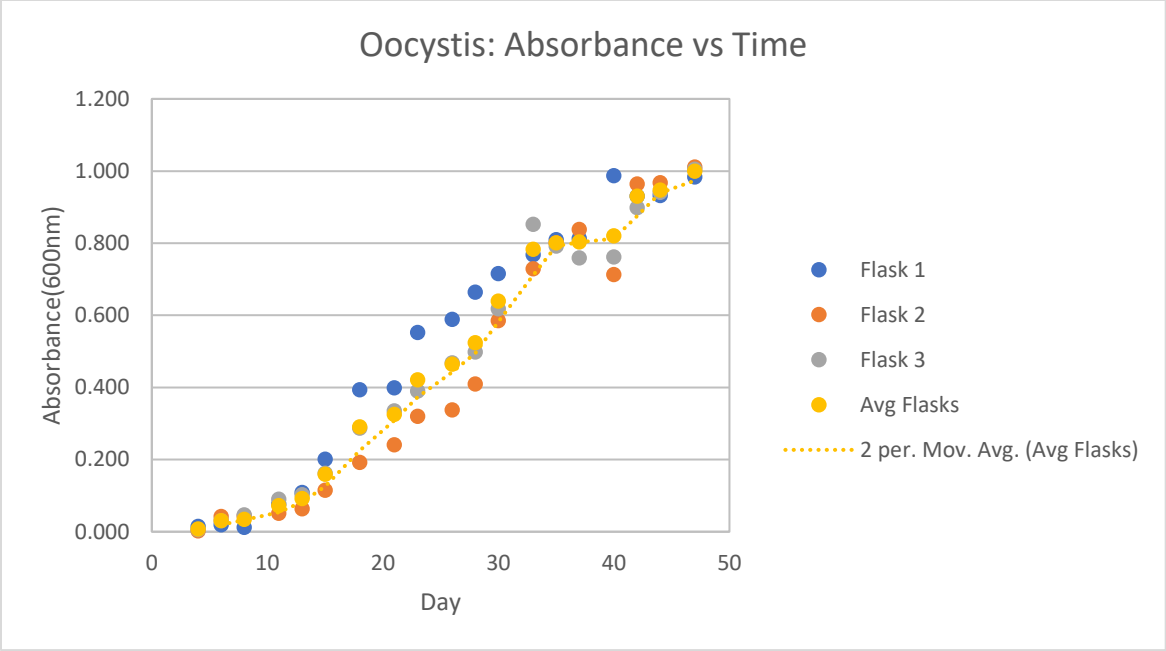
Graph 4: Absorbance data from initial run of Nannochloris for 2015



Graph 5: Absorbance data from initial run of Nannochloris for 2016



Graph 6: Absorbance data from initial run of Oocystis for 2015



Graph 7: Absorbance data from initial run of Oocystis for 2016

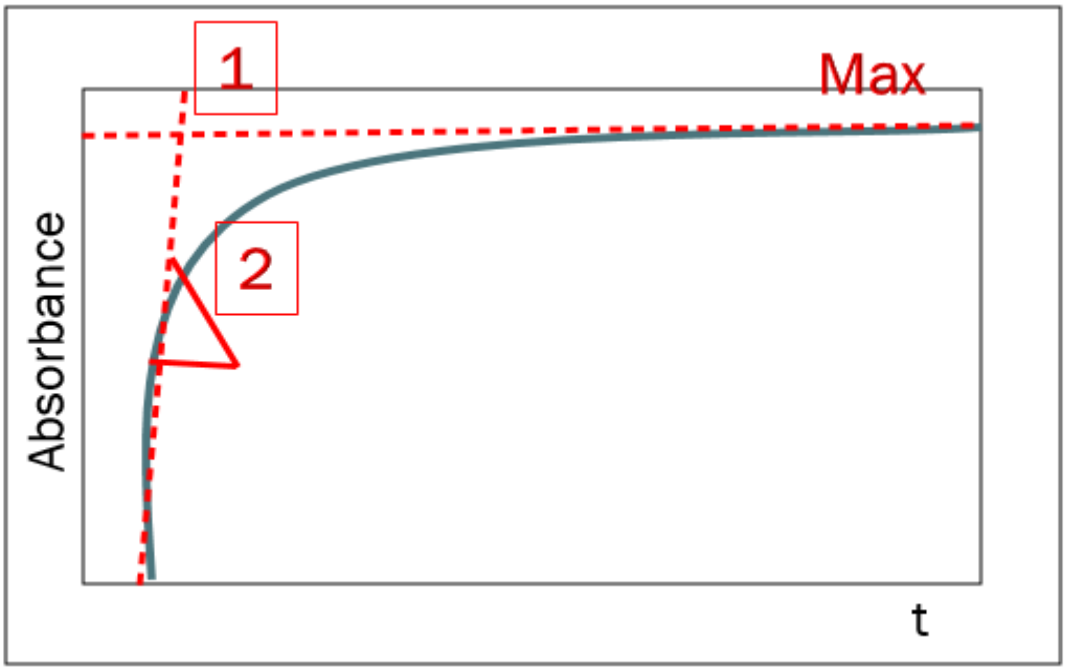
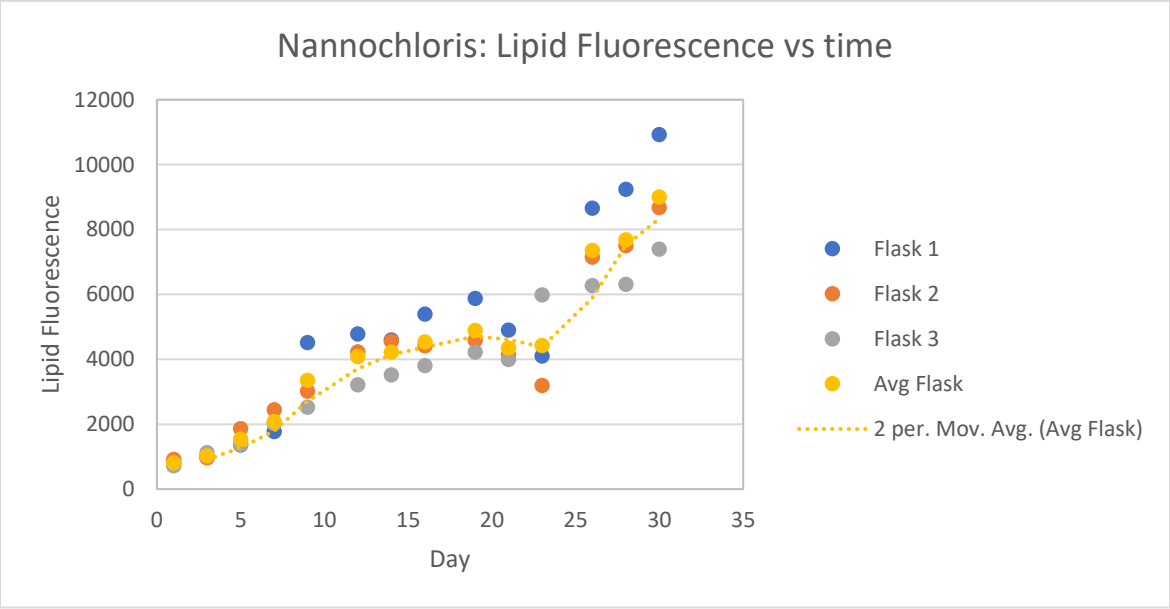


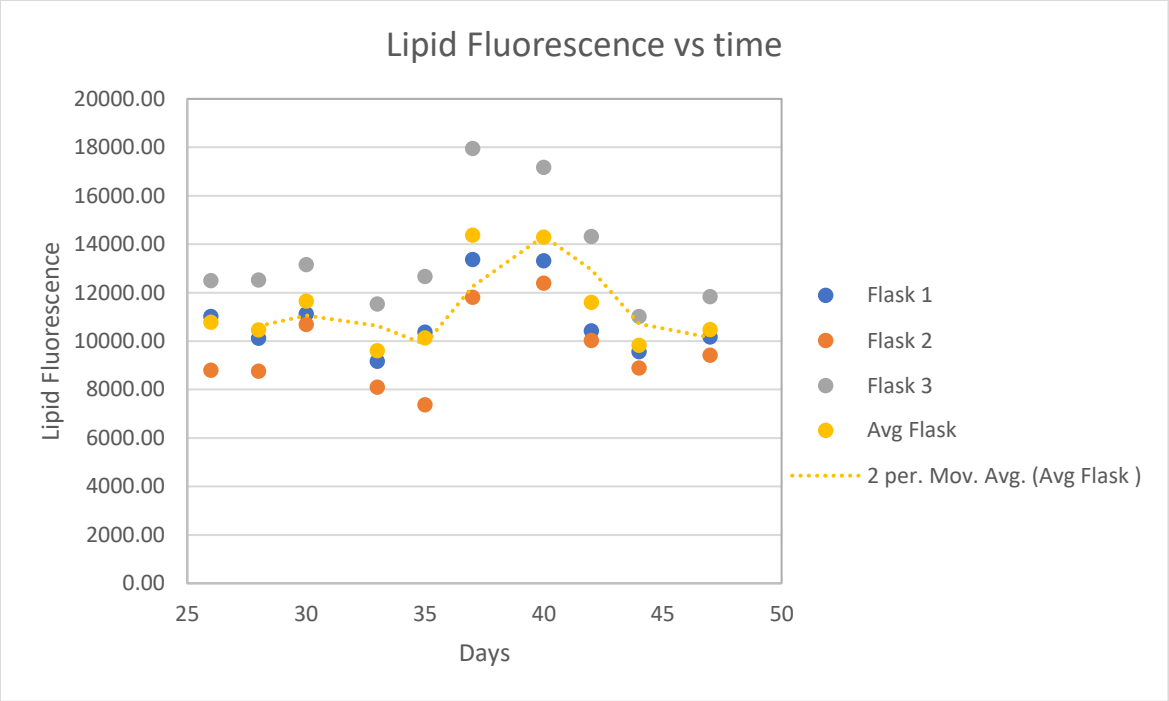
Figure 7: Ideal absorbance graph behavior [65]

3.1.2 Characterization – Lipid Fluorescence vs Time

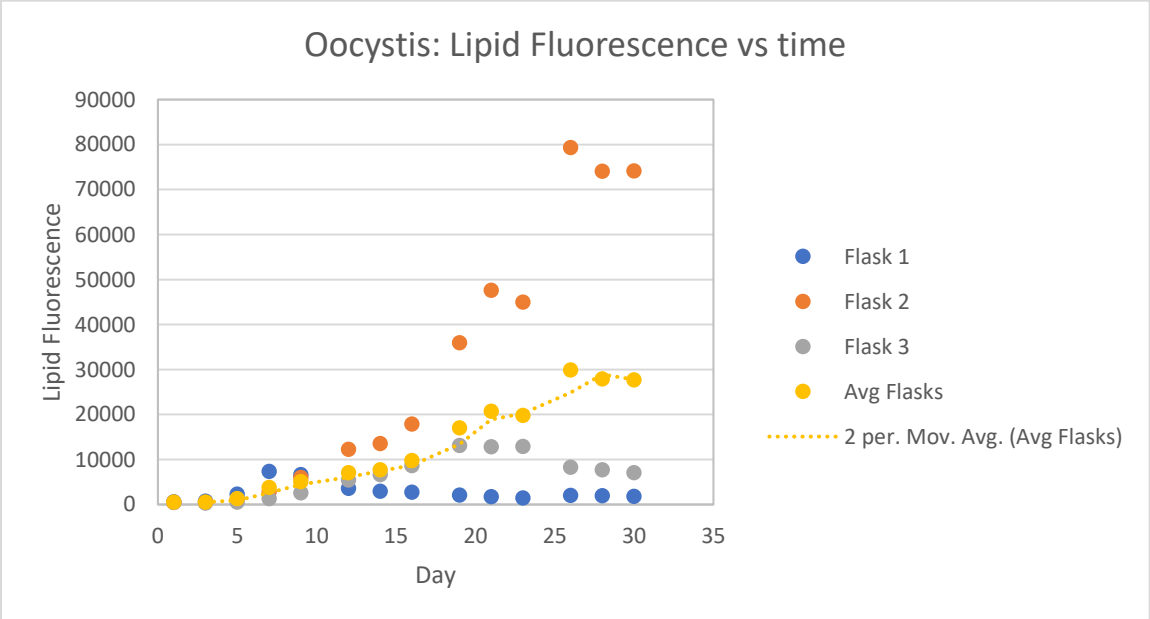
To achieve the maximum amount of lipids in a cell the amount of lipids was measured. The graph for lipid fluorescence vs time was done as the first step towards the $\frac{Lipid\ Fluorescence}{cells}$ vs time. Lipid fluorescence is the direct measurement of the amount of lipids. The goal is to obtain the most amount of lipid. Graph 8 and 9 show the initial data runs for Nannochloris. Graph 10 and 11 show the initial data runs for Ooscystis. The data for the 2016 run (Graph 9 and Graph 11) of both algae was taken from day 25 to the end of the experiment thus not allowing an accurate comparison to the proposed behavior in figure 8. In the 2015 runs (Graph 8 and graph 10) there is no linear behavior. The amount of lipids was supposed to increase in time. Nannochloris had some dips in growth but this is due to not standardizing procedures. Ooscystis in the other had had two samples that did not grow at all.



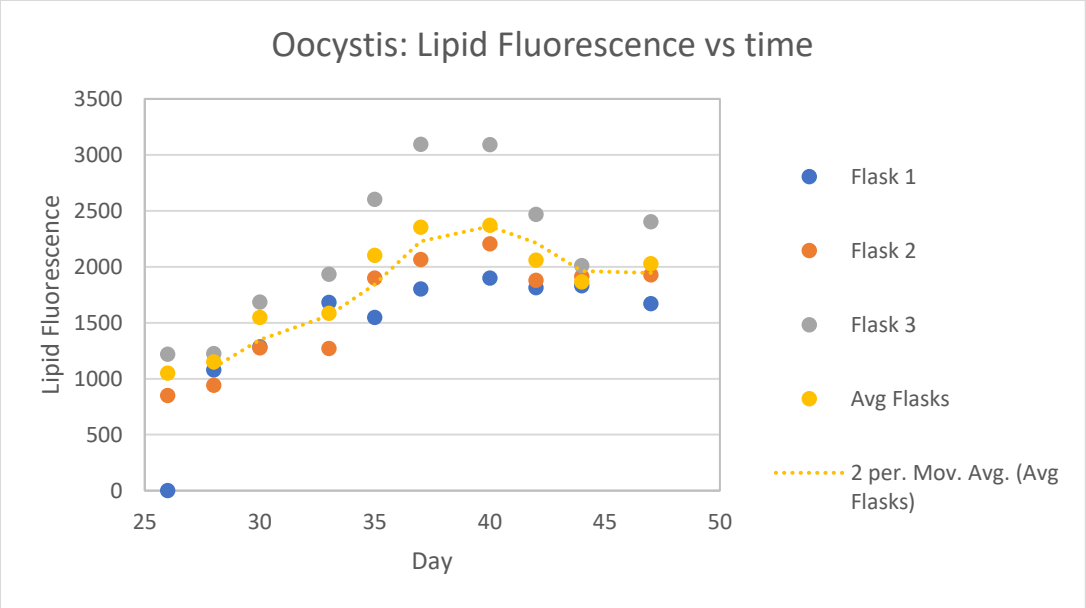
Graph 8: Lipid fluorescence data from initial run of Nannochloris for 2015



Graph 9: Lipid fluorescence data from initial run of Nannochloris for 2016



Graph 10: Lipid fluorescence data from initial run of Ooscystis for 2015



Graph 11: Lipid fluorescence data from initial run of Ooscystis for 2016

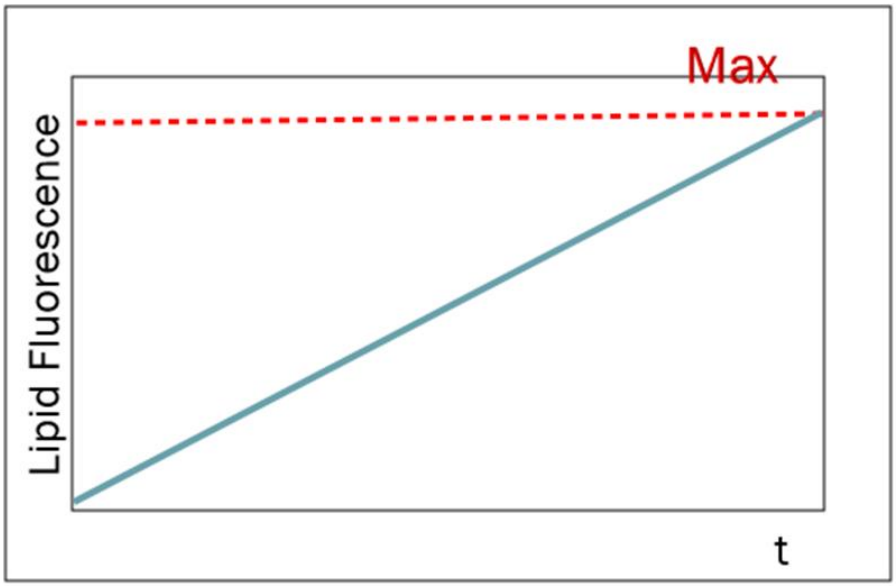


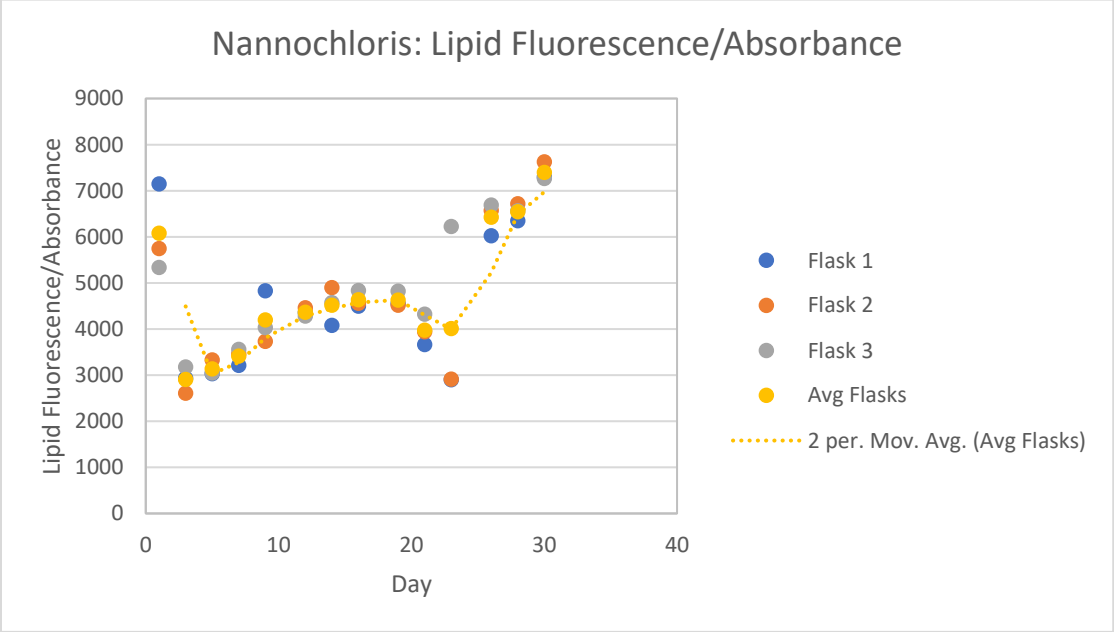
Figure 8: Proposed graph behavior for lipid fluorescence [65]

3.1.3 Characterization – Lipid Fluorescence/cells vs Time

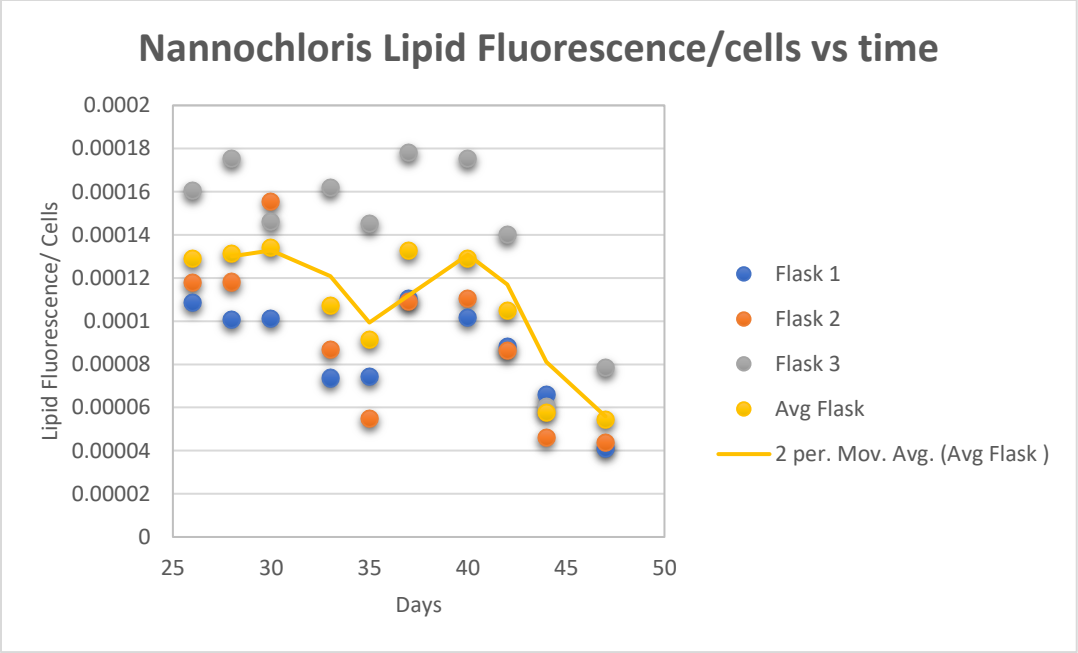
The rate of Lipid fluorescence/ cells is the amount of lipids in each cell, which is the main objective of this study and the third function that was maximized.

$$F_3 = \frac{Lipid\ Fluorescence}{cells}$$

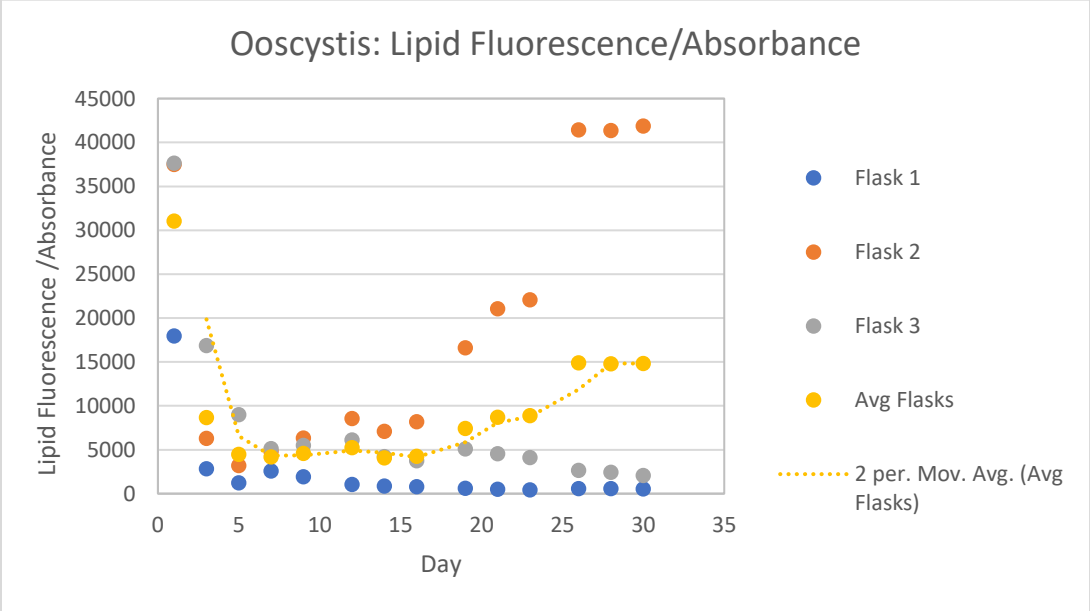
Graphs 12 and 13 are the initial data runs for Nannochloris. Graphs 14 and 15 are the initial data runs for Ooscystis. The data from Graphs 13 and 14 were taken after the 25th day of the experiment of the experiment thus not allowing an accurate comparison to the proposed behavior. In graph 12 the Nannochloris data does not behave at all like the proposed behavior in figure 9. There are a few dips instead of growing into a plateau. The data for Ooscystis in graph 14 also shows behavior that is different from the proposed behavior in graph



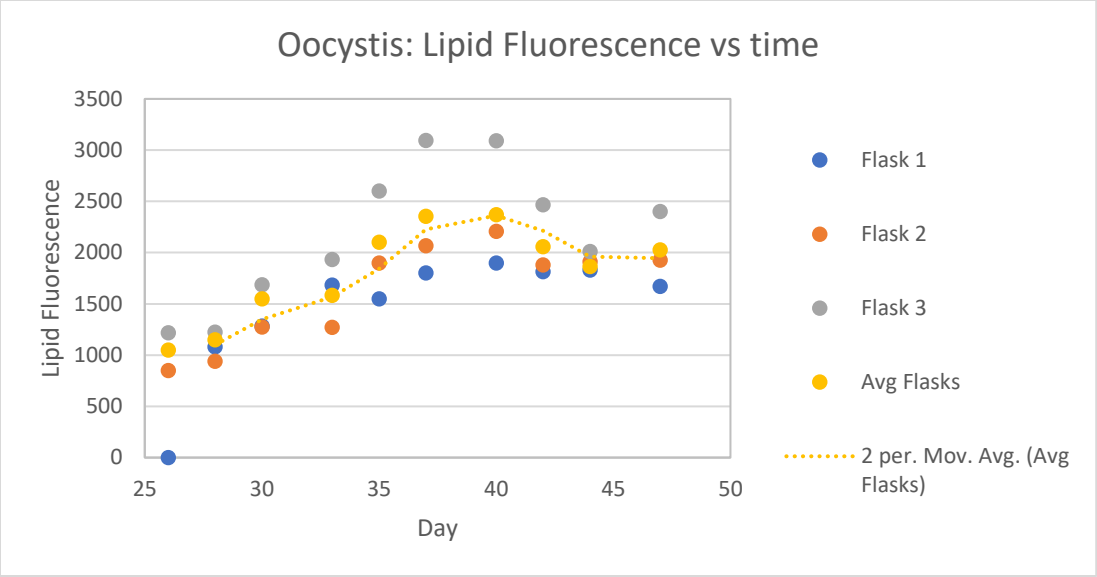
Graph 12: Nannochloris Lipid Fluorescence/cells data from initial run for 2015



Graph 13: Nannochloris Lipid Fluorescence/cells data from initial run for 2016



Graph 14: Ooscystis lipid Fluorescence/cells data from initial run for 2015



Graph 15: Ooscystis lipid Fluorescence/cells data from initial run for 2015

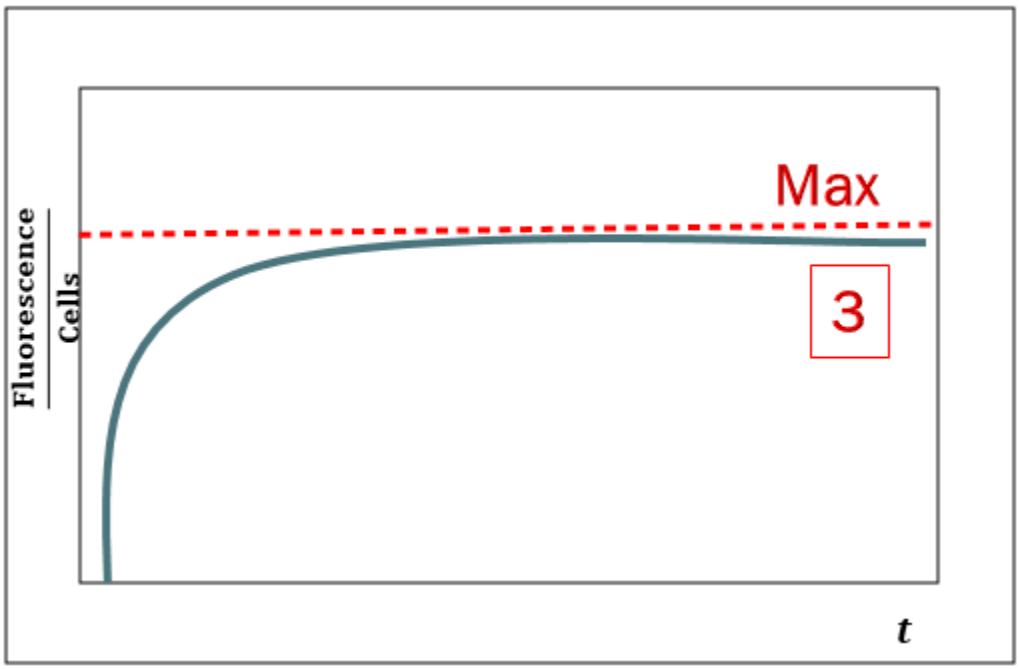


Figure 9: Proposed graph behavior for lipid fluorescence/cells [65]

3.1.4 Characterization – Power Analysis

An initial power analysis was done to find the minimum sample size to replicate the design of experiments and explain random deviation in the study. With the sample size, a design of experiment was created for both Nannochloris and Ooscystis algae.

3.1.5 Power Analysis

For both algae the data that was used for this analysis is the cell count column in the Cell Spec tab due to this data being a direct input from reading the experiment. Cells/ mL was not chosen due to being subjective to the number of grids and dilution. This study is necessary to find the minimum sample size to replicate the design of experiment and explain the random deviation in the study. The random deviation is the deviation in the same point in time of the runs. This report is based on the data taken in summer 2016 from Nannochloris and Ooscystis algae. These data has a total of 3 runs for each algae, each flask is a run. Paired t-tests were performed. In the paired t-test the difference is expressed as the difference between the population paired means that you would like to be able to detect. The analysis is based on what is the minimum difference needed to achieve a power (1-β) of 80%, where power is the probability of rejecting the null hypothesis correctly [33]. In this case since this is a retrospective study, the minimum number of cells between the existing runs is used as the difference. Miu (μ) is estimated by $\bar{x}$, which is the average amount of cells. Sigma (σ) is estimated by the standard deviation (S) of the data and it is a function of $\mu_1 - \mu_0$. The S chosen for both algae was the average of the standard deviations. The null hypothesis is that the replication obtains the same average amount of cells as the original run and the alternate hypothesis that it does not obtain the same average amount of cells in the replication.

Nannochloris

The data used (Table 2) for these algae was from day 6 to 47 due to not having information available from flask 1 in day 4. The standard deviation used for this analysis is 1862.03 cells. The power analysis is show in figure 10 and table 3. Table 3 shows that an experiment with a sample size of 3 needs a minimum difference of 6077.7 cells to achieve a power of 80%. A sample size of 5 is recommended due to the sample sizes larger than it are less sensible than those under 5 (Refer to table 3 and graph 16). A smaller sample size is not recommended due to every 2 consecutive points have a difference less than 6077.7 cells (Refer to graph 17).

Day	Cell count flask 1	Cell count flask 2	Cell count flask 3	Average Cell count in the same point of time	Standard deviation in same point of time
6	1701	1911	1639	1750.33	142.55
8	5939	3469	1988	3798.61	1996.15
11	3890	2900	2843	3210.83	588.88
13	6543	5128	5580	5750.00	722.66
15	6808	4629	3763	5066.67	1569.34
18	11364	5700	8932	8665.33	2841.40
21	8167	5263	8346	7258.33	1730.76
23	12264	10507	7860	10210.22	2216.92
26	10155	7475	7795	8475.00	1463.69
28	10035	7410	7155	8200.00	1594.26
30	10975	6880	9010	8955.00	2048.05
33	12470	9340	7135	9648.33	2680.83
35	13935	13445	8725	12035.00	2876.99
37	12095	10840	10070	11001.67	1022.13
40	13095	11245	9805	11381.67	1649.25
42	11825	11610	10220	11218.33	871.24
44	14480	19375	18190	17348.33	2553.73
47	24790	21705	15105	20533.33	4947.67

Table 2: Cell count data for Nannochloris

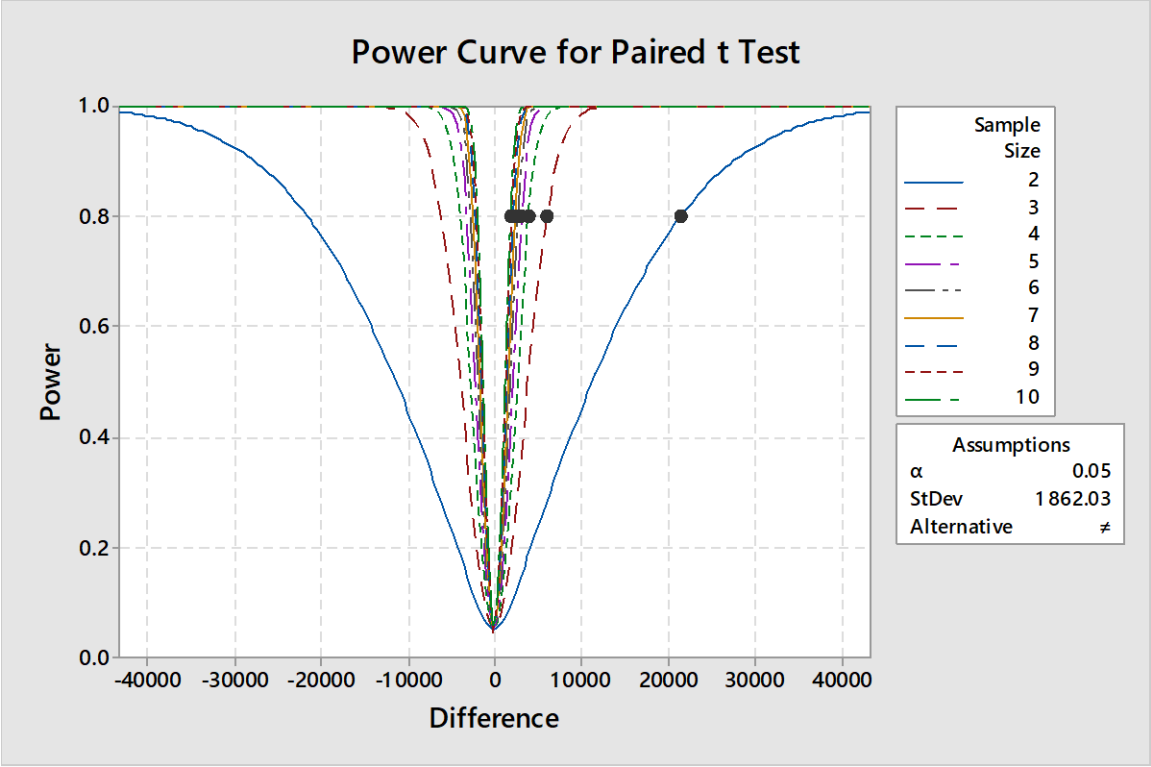
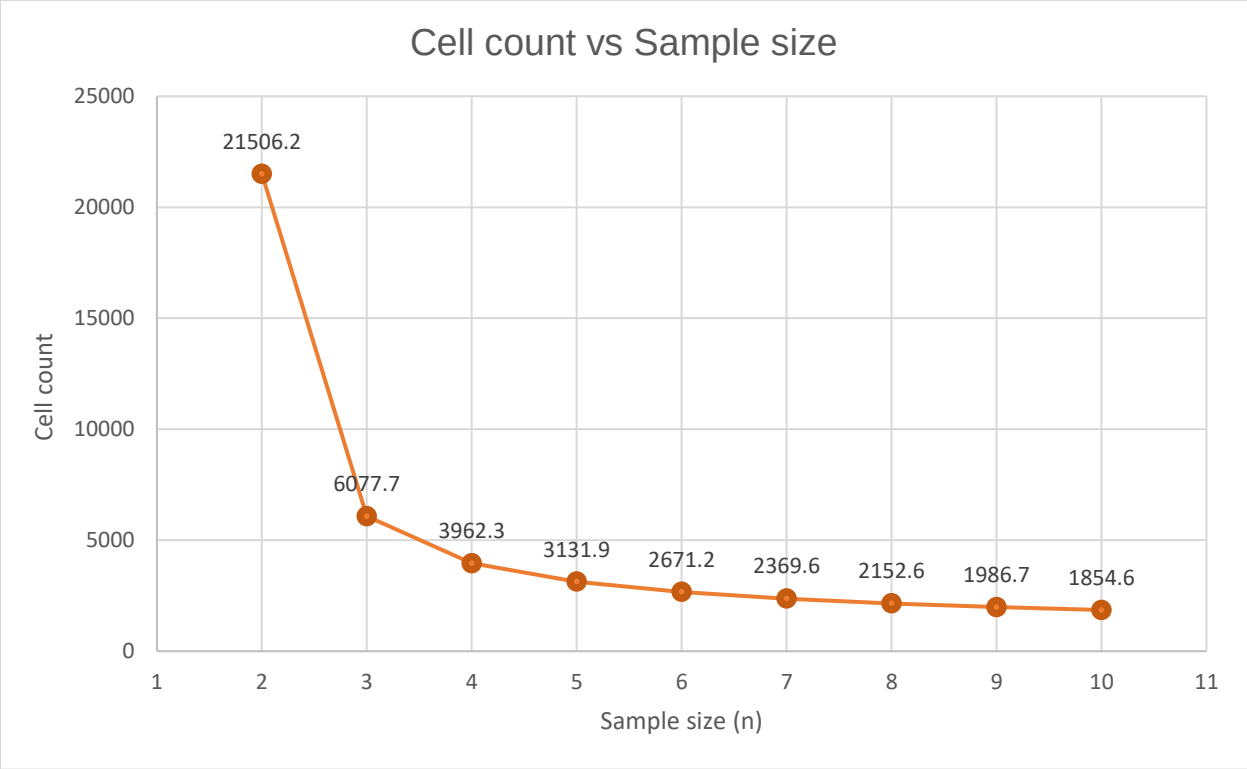


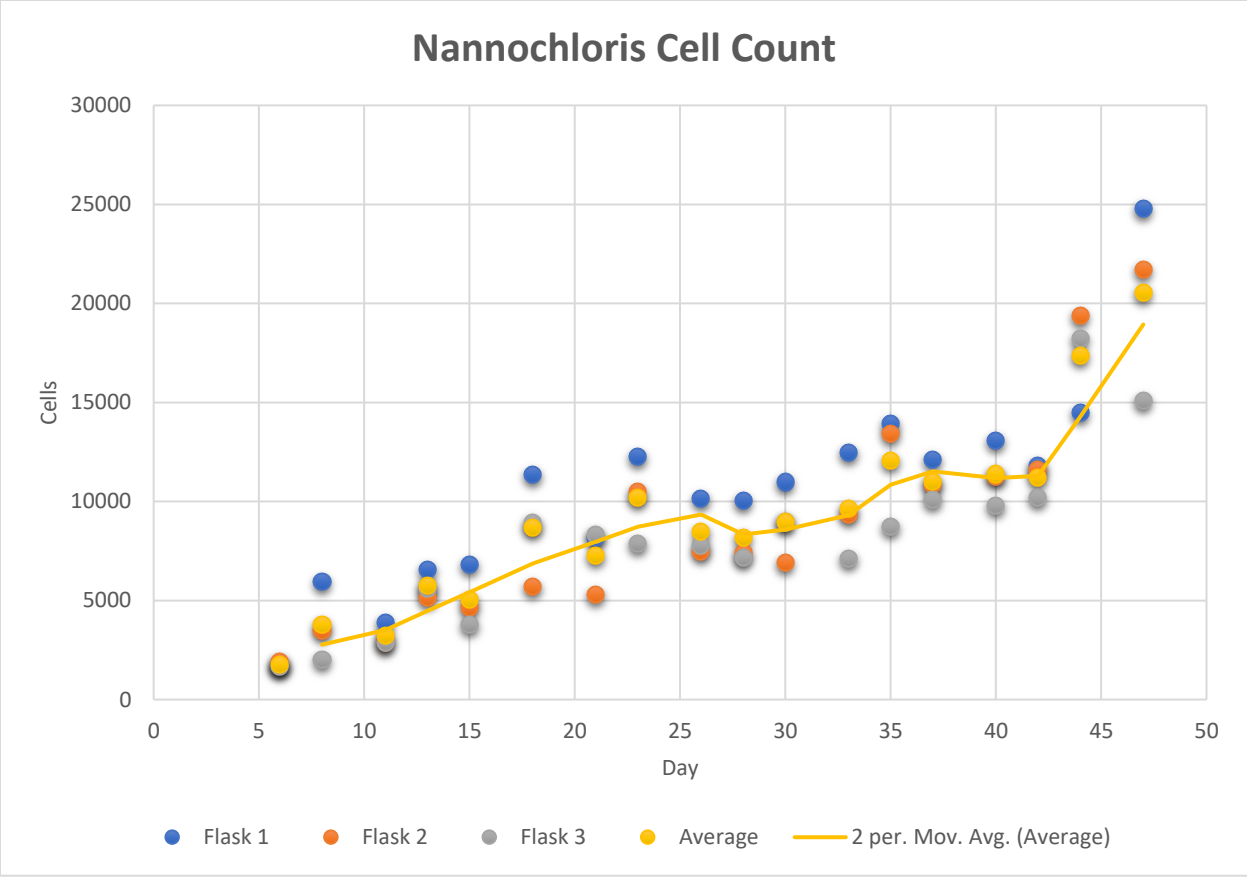
Figure 10: Nannochloris power curve for paired t test at 1-β =0.80 and s=1862.03

Sample size	Power	Difference ($\mu_1 - \mu_0$)
2	0.8	21506.2
3	0.8	6077.7
4	0.8	3962.3
5	0.8	3131.9
6	0.8	2671.2
7	0.8	2369.6
8	0.8	2152.6
9	0.8	1986.7
10	0.8	1854.6

Table 3: Difference needed to achieve 1-β =0.80



Graph 16: Nannochloris cell count vs sample size



Graph 17: Nannochloris cell count per flask

Ooscystis

The data used (Table 4) for these algae was from day 4 to 47. The standard deviation used for this analysis is 174.7 cells. The power analysis is show in figure 11 and table 5. Table 4 shows that an experiment with a sample size of 3 needs a minimum difference of 570.23 cells to achieve a power of 80%. A sample size of 5 is recommended due to the sample sizes larger than it are less sensible than those under 5 (Refer to table 5 and graph 18). A sample size is not recommended due to every 2 consecutive points have a difference less than 570.23 cells except day 47 (Refer to graph 19).

Day	Cell count flask 1	Cell count flask 2	Cell count flask 3	Average Cell count in the same point of time	Standard deviation in same point of time
4	2.40	6.40	1.20	3.33	2.72
6	6.80	7.60	4.60	6.33	1.55
8	14.60	21.00	25.80	20.47	5.62
11	48.20	12.00	33.20	31.13	18.19
13	151.00	48.20	97.60	98.93	51.41
15	77.80	60.20	142.00	93.33	43.06
18	285.40	149.00	215.20	216.53	68.21
21	404.20	172.00	271.00	282.40	116.52
23	464.00	225.60	353.20	347.60	119.30
26	548.80	215.60	314.80	359.73	171.08
28	682.80	366.80	473.20	507.60	160.78
30	677.00	589.00	518.00	594.67	79.65
33	820.00	844.00	1015.00	893.00	106.33
35	868.00	940.00	917.00	908.33	36.77
37	1558.00	804.00	1324.00	1228.67	385.93
40	1280.00	700.00	576.00	852.00	375.81
42	1660.00	984.00	1400.00	1348.00	340.99
44	2218.00	1234.00	1752.00	1734.67	492.23
47	2444.00	958.00	1672.00	1691.33	743.19

Table 4: Cell count data for Ooscystis

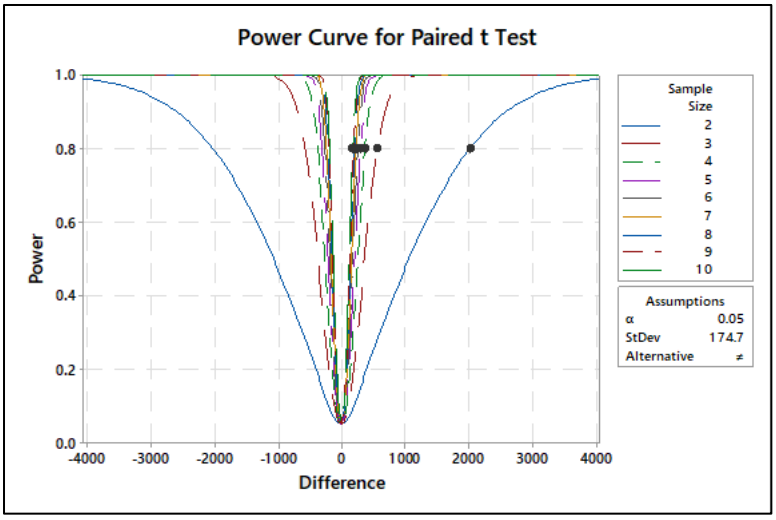
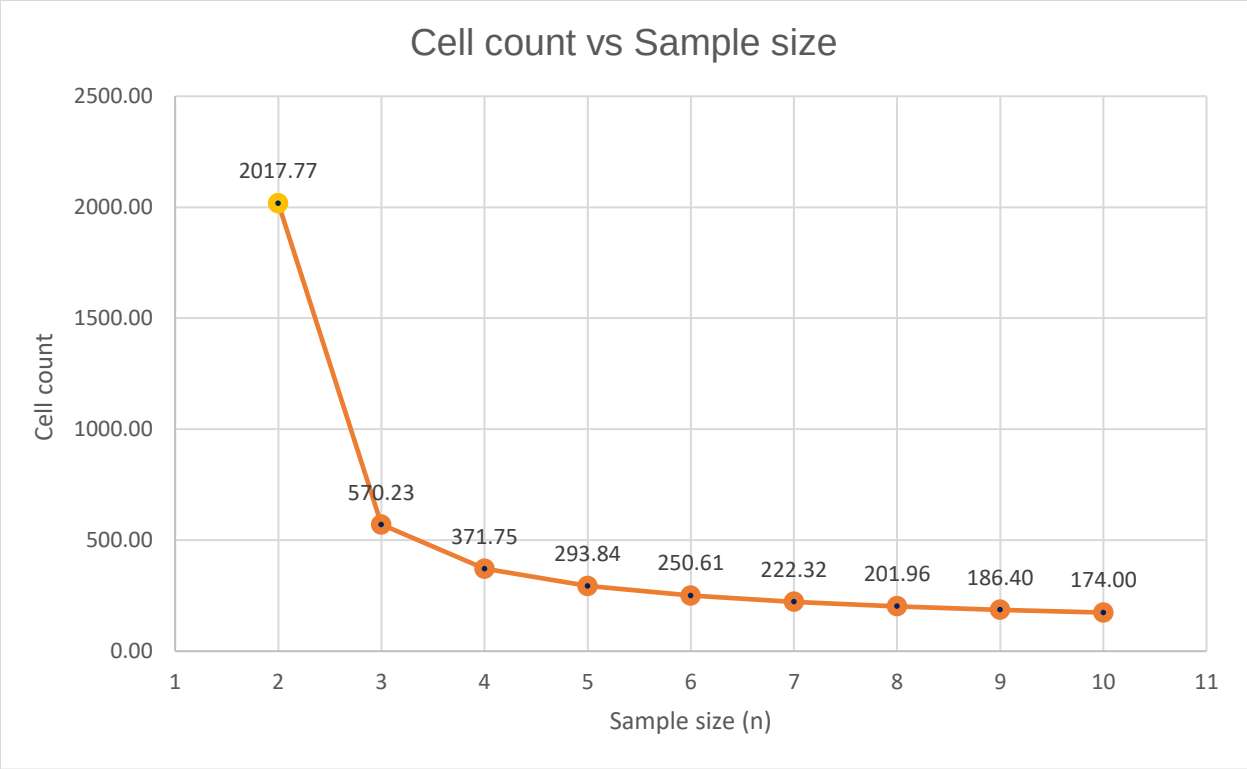


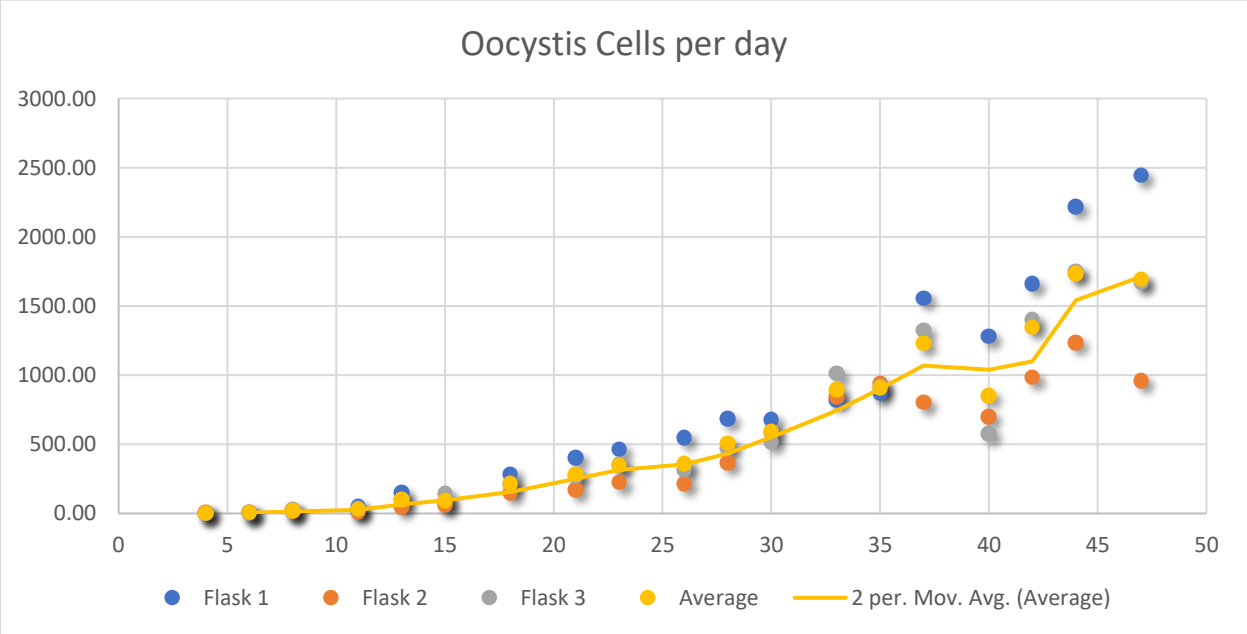
Figure 11: Ooscystis power curve for Paired t test at 1-β =0.80

Sample Size	Power	Difference ($\mu_1 - \mu_0$)
2	0.8	2017.77
3	0.8	570.23
4	0.8	371.75
5	0.8	293.84
6	0.8	250.61
7	0.8	222.32
8	0.8	201.96
9	0.8	186.40
10	0.8	174.00

Table 5: Difference needed to achieve 1-β =0.80



Graph 18: Ooscystis cell count vs sample size



Graph 19: Ooscystis cell count

3.1.6 Characterization - Design of Experiments

The next step in the characterization process is to find which variables are important for our key performance measurements. As part of our interdisciplinary effort, dynamic spreadsheet example was created to represent the experimental design proposed with light, sucrose and salinity as the factors. In the first modelling example central composite design was proposed due to its capability to support the building of second-order response surface models [33]. The beta coefficients found in this experiment will be used on the second part of the modelling example. The functions are represented as a mathematical optimization model as shown in equation 1.

Maximize F_1 = absorbance slope= f_1 (x_1=light, x_2=sucrose, x_3=salinity)

F_2 = absorbance= f_2 (x_1=light, x_2=sucrose, x_3=salinity)

$$F_3 = \frac{Lipid\ Fluorescence}{Cells} = f_3\ (x_1=\text{light}, x_2=\text{sucrose}, x_3=\text{salinity})$$

Subject to $\bar{x}_{\min} \leq x \leq \bar{x}_{\max}$

Equation 1: Mathematical model for yield

The proposed experimental design contains: 8 corner points, 6 start points and 5 replications for the center point (Refer to figure 12). The 5 replications points where chosen through the power analysis done in the previous section. This sums to a total of 19 runs. The center point is described as the current operation conditions [33]. The proposed model is represented in table 6. The coded levels are provided due to it being applied to both Nanncochloris and Ooscystis. The real values of the factors for each algae are given in table 7 and 8. The main difference between Nannochloris and Ooscystis are the salinity levels due to Nannochloris being a salt water algae and Ooscystis not.

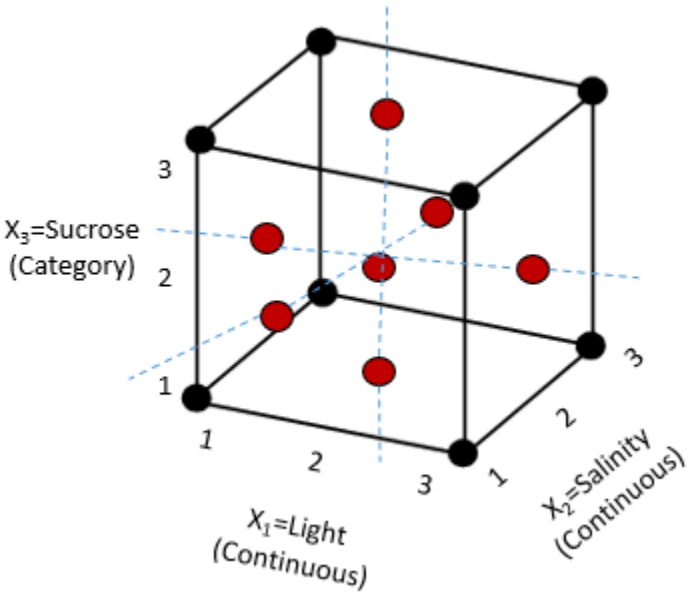


Figure 12: Experimental region

	x_1= light	x_2=sucrose	x_3=salinity
1	1	1	1
1	1	1	3
1	1	2	2
1	1	3	1
1	1	3	3
1	2	1	2
1	2	2	1
1	2	2	3
1	2	2	2
1	2	2	2
1	2	2	2
1	2	2	2
1	2	2	2
1	2	3	2
1	3	1	1
1	3	1	3
1	3	2	2
1	3	3	1
1	3	3	3

Table 6: Coded central composite for algae

Levels	x_1=Light (Hours)	x_2= Salinity (ppt)	x_3=Sucrose (%)
1	8	15	0
2	16	35	0.025
3	24	50	0.5

Table 7: Experimental info for Nannochloris

Levels	x_1=Light (Hours)	x_2= Salinity (ppt)	x_3=Sucrose (%)
1	8	0	0
2	16	15	0.025
3	24	35	0.5

Table 8: Experimental info for Oocystis

In continuation, a step by step example of the proposed design of experiment is explained with dummy data and later the modelling per se of the three main linear functions as explained in this chapter.

β_0	1
β_1	1
β_2	1
β_3	1

	X Nannochloris			
Run	Intercept	x_1 =Light (Hr)	x_2=Salinity (ppt)	x_3=Sucrose (%)
1	1	8	15	0
2	1	8	15	0.05
3	1	8	35	0.025
4	1	8	50	0
5	1	8	50	0.05
6	1	16	15	0.025
7	1	16	35	0
8	1	16	35	0.025
9	1	16	35	0.025
10	1	16	35	0.025
11	1	16	35	0.025
12	1	16	35	0.025
13	1	16	35	0.05
14	1	16	50	0.025
15	1	24	15	0
16	1	24	15	0.05
17	1	24	35	0.025
18	1	24	50	0
19	1	24	50	0.05

Y Nannochloris		
Y deterministic	Y stochastic	Y response
24	0.32748896	24.33
24.05	0.71604382	24.77
44.025	0.97448636	45.00
59	0.99903946	60.00
59.05	0.21172563	59.26
32.025	0.562986	32.59
52	0.77016774	52.77
52.025	0.60268957	52.63
52.025	0.53969143	52.56
52.025	0.93862634	52.96
52.025	0.81186273	52.84
52.025	0.22145608	52.25
52.05	0.24188413	52.29
67.025	0.60497511	67.63
40	0.61658211	40.62
40.05	0.23228527	40.28
60.025	0.48938217	60.51
75	0.46030686	75.46
75.05	0.48164132	75.53

X^TX

19	304	640	0.475
304	5504	10240	7.6
640	10240	24650	16
0.475	7.6	16	0.018125

$(X^TX)^{-1}$

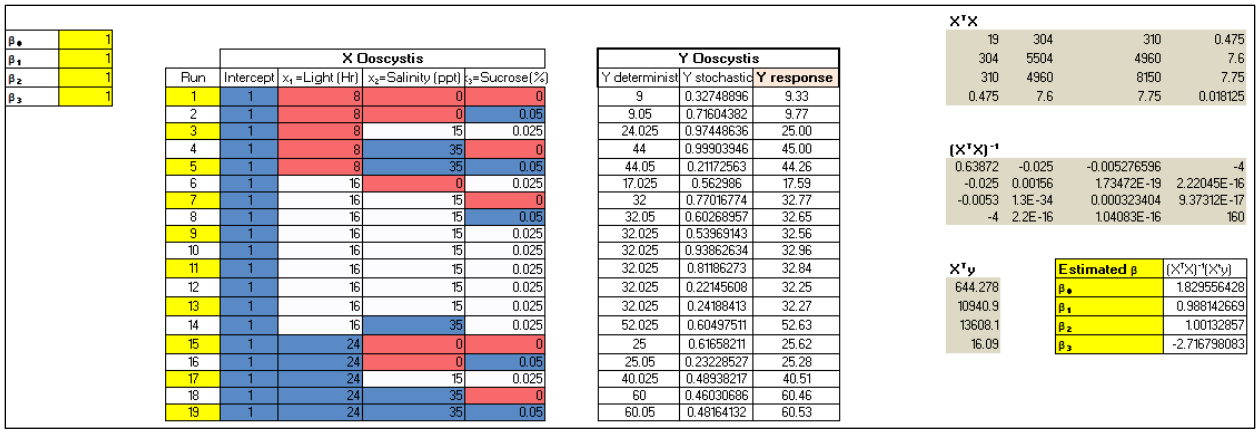
0.91957	-0.025	-0.010893617	-4
-0.025	0.00156	2.40741E-35	-2.22045E-16
-0.0109	0	0.000323404	0
-4	-2E-16	-1.73472E-17	160

X^Ty

974.278
16220.9
35916.4
24.331

Estimated β	$(X^TX)^{-1}(X^Ty)$
β_0	1.816804723
β_1	0.988142669
β_2	1.002093234
β_3	-4.160019848

Figure 13: Design of experiment for Nannochloris example with dummy data



β_0	1
β_1	1
β_2	1
β_3	1

	X Ooscystis			
Run	Intercept	x_1 =Light (Hr)	x_2=Salinity (ppt)	x_3=Sucrose(%)
1	1	8	0	0
2	1	8	0	0.05
3	1	8	15	0.025
4	1	8	35	0
5	1	8	35	0.05
6	1	16	0	0.025
7	1	16	15	0
8	1	16	15	0.05
9	1	16	15	0.025
10	1	16	15	0.025
11	1	16	15	0.025
12	1	16	15	0.025
13	1	16	15	0.025
14	1	16	35	0.025
15	1	24	0	0
16	1	24	0	0.05
17	1	24	15	0.025
18	1	24	35	0
19	1	24	35	0.05

Y Ooscystis		
Y determinist	Y stochastic	**Y response**
9	0.32748896	9.33
9.05	0.71604382	9.77
24.025	0.97448636	25.00
44	0.99903946	45.00
44.05	0.21172563	44.26
17.025	0.562986	17.59
32	0.77016774	32.77
32.05	0.60268957	32.65
32.025	0.53969143	32.56
32.025	0.93862634	32.96
32.025	0.81186273	32.84
32.025	0.22145608	32.25
32.025	0.24188413	32.27
52.025	0.60497511	52.63
25	0.61658211	25.62
25.05	0.23228527	25.28
40.025	0.48938217	40.51
60	0.46030686	60.46
60.05	0.48164132	60.53

X^TX

19	304	310	0.475
304	5504	4960	7.6
310	4960	8150	7.75
0.475	7.6	7.75	0.018125

$(X^TX)^{-1}$

0.63872	-0.025	-0.005276596	-4
-0.025	0.00156	1.73472E-19	2.22045E-16
-0.0053	1.3E-34	0.000323404	9.37312E-17
-4	2.2E-16	1.04083E-16	160

X^Ty

644.278
10940.9
13608.1
16.09

Estimated β	$(X^TX)^{-1}(X^Ty)$
β_0	1.829556428
β_1	0.988142669
β_2	1.00132857
β_3	-2.716798083

Figure 14: Design of experiment for Ooscystis example with dummy data

Figure 13 and 14 is a screenshot taken from the excel worksheet created to model with dummy data the proposed design of experiment. This exercise is done through matrix notation $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ because it is the predicted value**.** This spreadsheet is dynamic, the values of the betas is interchangeable. The y stochastic represents the natural variation that an experiment may have. The estimated betas are found through least squares estimation and used in the second part of the modelling example.

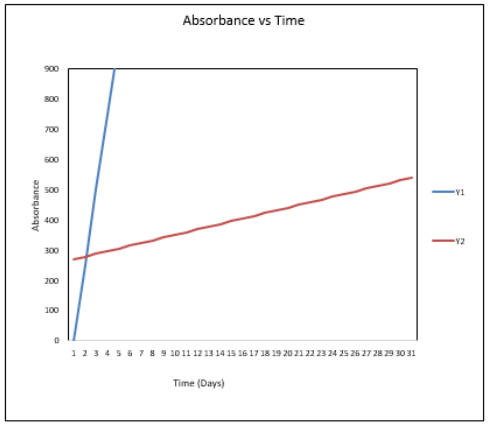
3.2 Modeling

Lastly, another dynamic spreadsheet was created to better explain the proposed modeling process to our collaborators. Figure 15 represents the three functions show in equation 1 The spreadsheet is completely dynamic to the decision variables and the input of coefficients that are derived from the design of experiments. Function 1 just has one component (slope) whereas function 2 and 3 are divided into two components (Intercept and slope). The data graphed is obtained from the time table where the functions growth is time dependent. Depending on how the inputs are changed the graphs to the right are going to change. This part of the example gives more visibility into finding which decision variable levels will help maximize all functions and what could be a trade-off.

Decision Variables		
X_1= Light	X_2= Salinity	X_3= Sucrose Source
2	2	2

Coefficients	Y_1	Y_2		Y_3	
	Absorbance Slope	Max Absorbance Intercept	Max Absorbance Slope	Intercept Lipid Fluorescence/Cells	Slope Lipid Fluorescence/Cells
β_0	10	10	1	200	0
β_1	10	20	1	10	1
β_2	100	10	1	10	1
β_3	10	100	2	10	1
Responses	**250**	**270**	**9**	**260**	**6**

Time(Days)	Y_1 = Absorbance Growth Rate	Y_2 = Absorbance Growth	Y_3 = Lipid Fluorescence/Cells Growth
0	0	270	260
1	250	279	266
2	500	288	272
3	750	297	278
4	1000	306	284
5	1250	315	290
6	1500	324	296
7	1750	333	302
8	2000	342	308
9	2250	351	314
10	2500	360	320
11	2750	369	326
12	3000	378	332
13	3250	387	338
14	3500	396	344
15	3750	405	350
16	4000	414	356
17	4250	423	362
18	4500	432	368
19	4750	441	374
20	5000	450	380
21	5250	459	386
22	5500	468	392
23	5750	477	398
24	6000	486	404
25	6250	495	410
26	6500	504	416
27	6750	513	422
28	7000	522	428
29	7250	531	434
30	7500	540	440



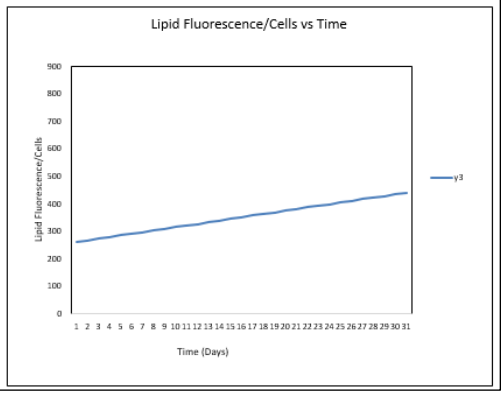


Figure 15: Modelling example for objectives spreadsheet

3.4 Summary

The interdisciplinary work in this chapter came into conclusion that the main goal in this biosystem is to maximize the amount of lipids obtained per algae cell in the least amount of time using light, sucrose and salinity as factors. This goal was separated into three functions: (1) Absorbance, (2) Absorbance Slope and (3) the ratio of lipids in each cell. The initial characterization of Nannochloris and Ooscystis consisted on comparing the initial experiments done in the University of Texas, Brownsville with the information available in the Biotek website [65]. These initial experiments were not consistent with the proposed behavior. The next step was to use this data to find a minimum sample size through a power analysis. This analysis showed that a sample size of 5 for both algae would be enough to help explain random deviation in the study. The sample size was the base of the design of experiment. The experimental design proposed was a central composite design with the sample size of 5 applied as replications in the center point. This design can be applied to both Nannochloris and Ooscystis. The only difference is the level of salinity between them since Nannchloris is a salt water algae and Ooscystis is a fresh water algae. This experimental design will find the value for each beta coefficient meaning that it will better explain the effect of each factor in the growth of each algae. The beta coefficient found will be plugged in the dynamic spreadsheet created. This spreadsheet was designed to model the future results using the three functions discussed before. This spreadsheet will also help visualize the best level for each factor. These models are stepping stones for the future optimization work.

Chapter 4. USDA hairy vetch/corn study

This chapter focuses on finding an economic threshold for farmers that use hairy vetch as a cover crop. This is done by finding the best hairy vetch seeding rate for hairy vetch and corn/silage. The costs of usage of hairy vetch is later compared to the costs of nitrogen fertilizer. Since the main driver for decision making is corn/silage yield, the best hairy vetch seeding rate found for corn/silage, this will be the seeding rate used for comparisons between nitrogen fertilizer and the trade-off between using the best for hairy vetch vs corn/silage.

4.1 Characterization

Statistical inference was used to characterize hairy vetch biomass from Massachusetts (MA), New York (NY), Pennsylvania (PA), Maryland (MA) and North Carolina (NC) from 2011-2014.

4.1.1 Characterization: Hairy Vetch

The data characterized for hairy vetch was obtained from MA, NY, PA, MD and NC (Kin, Sali and Gold) for 2011-2012, 2012-2013 and 2013-2014. A series of ANOVAs were done to find the hairy vetch biomass yield's relationship with planting date (PD), hairy vetch seeding rate (SR) and harvesting date (HD) across all states at $\alpha=0.05$ taking all years as replicates. These ANOVAs were used to create regressions to predict hairy vetch biomass based on the unitary changes found from the data of the actual factors. The null hypothesis is that the factors (PD, SR and HD) do not affect hairy vetch biomass and the alternate hypothesis is that it does affect hairy vetch biomass. Also residual normality (Kolmogorov-Smirnov test), residual independence (Runs test) and residual constant variance (Lavene's test) and fits (lack-of-fit test) were evaluated.

Table 9 has a visual help were those p-values that are significant are highlighted.

Tests		All years						
	Terms	MA	NY	PA	MD	NC Kin	NC Sali	NC Gold
Significance of factors	Regression	0	0	0	0	0.001	0.139	0
	PD	0	0	0.069	0	0.113	0.103	0
	SR	0	0	0	0.006	0.153	0.145	0.009
	HD	0	0.866	0	0.231	0.983	-	0.008
	PD_day*SR	0	0.001	0.189	0.666	0.171	0.614	0.392
	PD*HD	0.119	0.182	0	0.012	0.724	-	0
	SR*HD	0.002	0.041	0.491	0.689	0.129	-	0.582
	PD^2	0.086	0.008	0.005	0	0.718	0.055	-
	SR^2	0	0.026	0.167	0.638	0.543	0.8	0.665
	HD^2	0	0.993	0	0	-	-	-
K-S	Residual Normality	0.005	0.01	0.005	0.005	0.375	0.291	0.708
Runs test	Residual Independence	0.233	0	0	0	0.582	0.373	0.079
Lavene	Residual Constant Variance	0.014	0	0	0.405	0.022	0.224	0.957
Model fit	Lack of fit	0	0	0	0	0.851	0.268	0.667
	R-Sq	61.68%	46.30%	35.53%	55.44%	0.26%	13.27%	54.79%
	R-Sq(Adj)	60.34%	44.49%	33.84%	54.52%	0.19%	5.67%	50.63%

Table 9: P-values, test results and model fit across Massachusetts (MA), New York (NY), Pennsylvania (PA), Maryland (MD) and North Carolina (NC Kin, NC Sali and NC Gold)

In table 9 all the leading factors (PD, SR and HD) have significance either individually, by interaction or by second order in MA, NY, PA, MD and NC Gold. Lack of significance of such factors in NC Kin's and NC Sali's can be due to the lack of data available compared to the other sites. The regression fit for NY and PA comply with all the significance tests. MA only failed in residual independence and MD only failed in constant variance. All states except for the NC sites are significant in the lack-of-fit test. The state with the highest fit (R-Sq) was MA later followed by MD, NY, NC Gold and PA. NC Kin and NC Sali had low fits. The second order regression had an overall good representation for MA, NY, PA and MD which are the states are that going to be used to compare the best SR for hairy vetch vs corn. These findings are important because it was statistically proven that: 1) hairy vetch depends on PD, SR and HD and 2) it leads the way to ensure proper modeling. Table 10 shows the coefficients found in the regression adjusted that will be used in section 4.2.

Terms		MA	NY	PA	MD	NC Kin	NC Sali	NC Gold
β_0	Constant	1218.9	960	2034	-47	7563	5647	8191
β_1	PD	-615.1	-1242	-194	-2318	-476	266	-3091
β_2	SR	474.9	522	772	500	348	320	743
β_3	HD	580.8	-28	878	303	6	-	1761
β_{12}	PD*SR	-389.7	-535	-145	-96	424	-115	382
β_{13}	PD*HD	193	278	-578	-757	-243	-	-2778
β_{23}	SR*HD	376	313	-118	-60	415	-	-236
β_{11}	PD^2	-173	512	-594	2514	-250	1206	-
β_{22}	SR^2	-529	-370	-250	98	-229	-90	-171
β_{33}	HD^2	714	2	806	1004	-	-	-

Table 10 Coefficients for hairy vetch biomass modeling across Massachusetts (MA), New York (NY), Pennsylvania (PA), Maryland (MD) and North Carolina (NC Kin, NC Sali and NC Gold)

4.1.2 Characterization: Corn

The corn/silage yield data for this study was obtained for MA and MD for 2011-2012,2012-2013 and 2013-2014. The PA and NY data is from 2012-2013 and 2013-2014. The different sites at NC did not participate in this part of the study. This section consists of the study of how seeding rate from hairy vetch as a factor affects corn/silage yield across these states at $\alpha=0.05$. Also, what is the seeding rate that provides the best corn yield through the family error rate. The following tests were used:

Test name	Feature	Null hypothesis	Alternative hypothesis
Hsu MCB	The difference between each group and the best of the other groups.	All means are equal.	At least one mean is different.
Dunnett	The difference between each treatment group and the control group.	A treatment's mean is equal to the control's mean.	A treatments mean is different to the control's mean.

Table 11: Summary of tests [66]

In Hsu's test highest is best, in this case the highest corn yield. When the confidence interval contains zero, there is no difference between groups and the best of the groups. If the confidence interval is entirely above zero it means that the group is significantly better. If the confidence interval is entirely below zero, the group is significantly worse. In Dunnet's when the confidence interval contains zero, there is no difference to the control group. If the confidence interval is entirely above zero, the group is significantly better than the control group.

A one-way Analysis of Variance (ANOVA) and a second order regression were also done to find if hairy vetch seeding rate is a statistically significant for corn yield. The null hypothesis is that hairy vetch seeding rate does not affect corn yield and the alternate hypothesis is that it does affect corn yield.

Massachusetts – All years as replicates

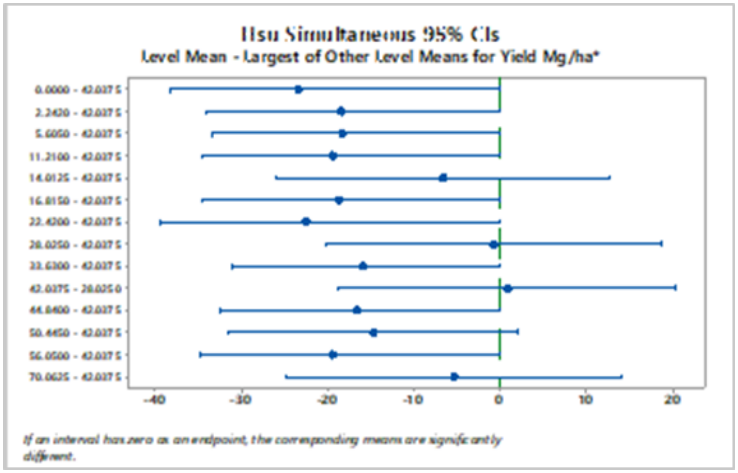


Figure 16: Hsu for Massachusetts using all years as replicates

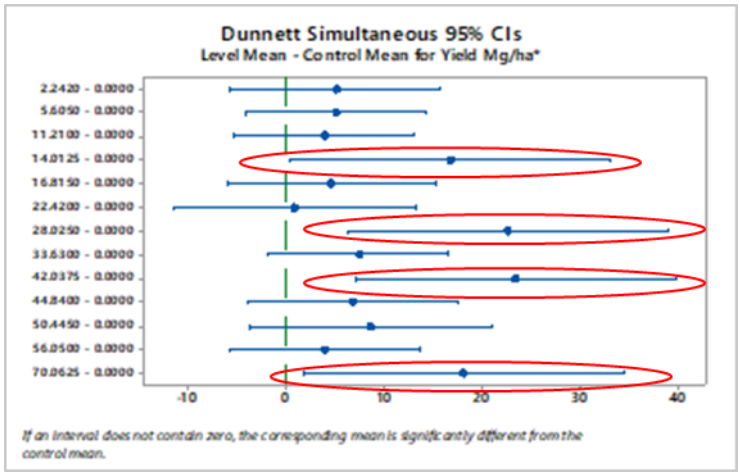


Figure 17: Dunnett's for Massachusetts using all years as replicates

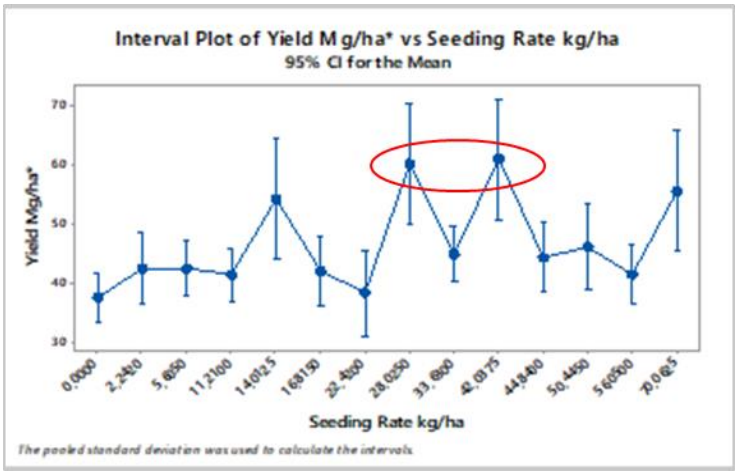


Figure 18: Interval plot for Massachusetts for all years as replicates

Figure 16 shows that most of the intervals perform significantly worse compared to 14.013-42.038,8.025-42.038, 42.038-28.025 and 70.063-42.048, which are not significantly different between each other. In Figure 17 you can see that the seeding rates of 14.02,28.025,42.038 and 70.063 kg/ha^{-1} perform significantly better than the control group of 0 kg/ha. Figure 18 further states that 14.015,28.025,42.038 and 70.063 kg/ha obtained higher yields than the rest of the seeding rates. In an economic point of view a seeding rate of 14.015 or 28.025 kg/ha^{-1} are recommended. The seeding rate of hairy vetch was determined to be statistically significant with a p-value of practically 0

New York – All years as replicates

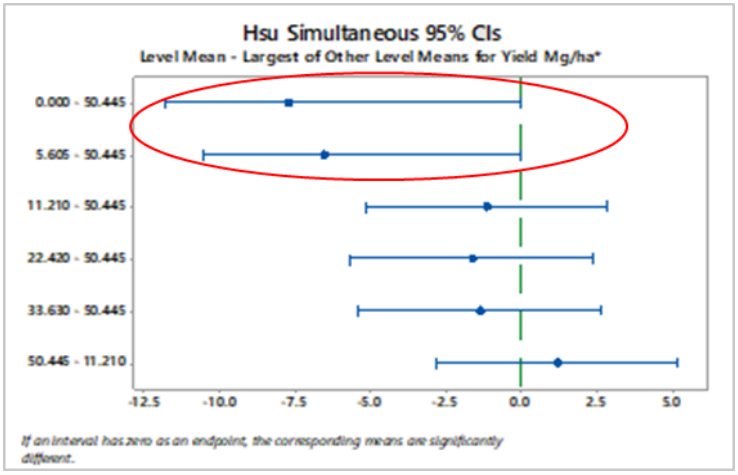


Figure 19: Hsu for New York using all years as replicates data

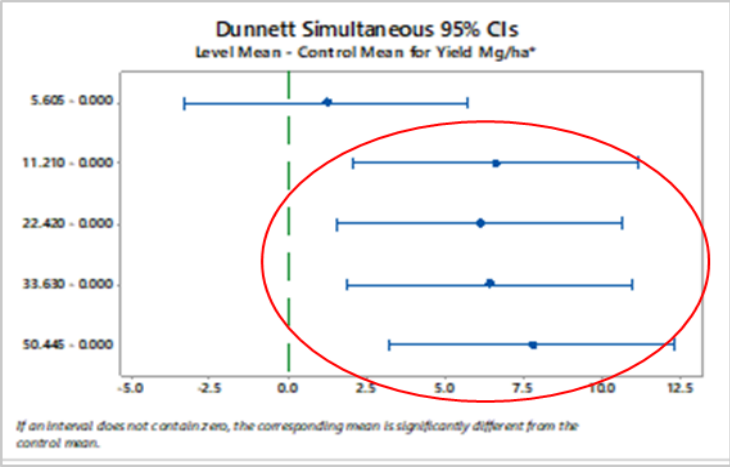


Figure 20: Dunnett's for New York using all years as replicates data

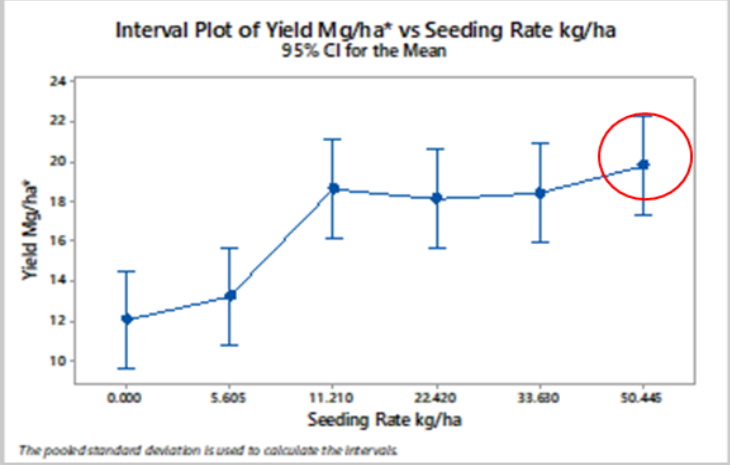


Figure 21: Interval plot for New York using all years as replicates data

Figure 19 shows that the intervals 0-50.445 and 5.605-50.445 are significantly worse. Figure 20 shows that 11.210, 22.420, 33.630 and 50.445 kg/ha^{-1} perform significantly better than the control group. Figure 21 shows that 50.445 kg/ha^{-1} provided the largest amount of corn yield. The seeding rate of hairy vetch was determined to be statistically significant with a p-value of 0.

Pennsylvania- All years as replicates

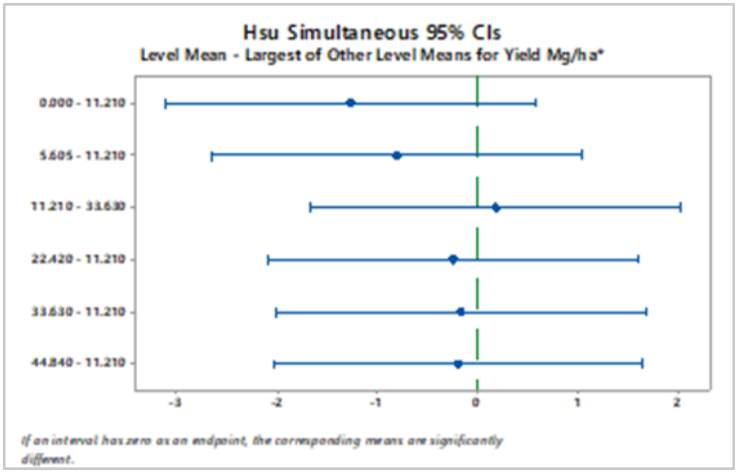


Figure 22: Hsu for Pennsylvania using all years as replicate

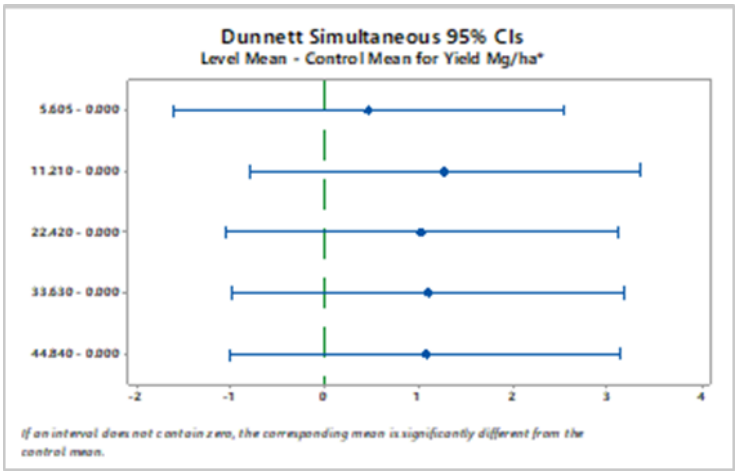


Figure 23: Dunnett's for Pennsylvania using all years as replicate

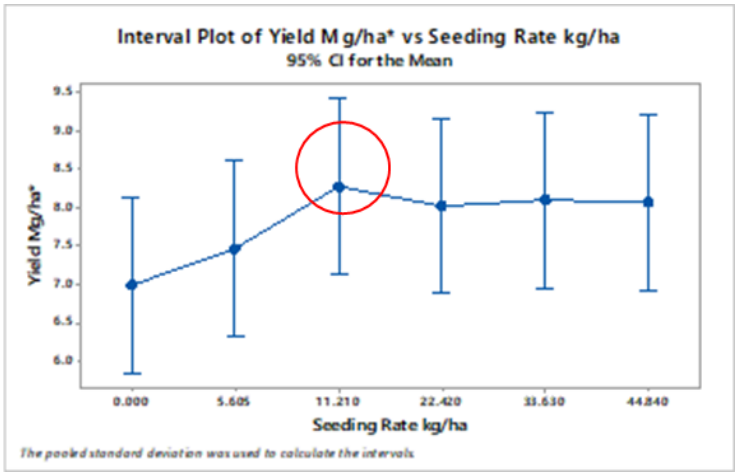


Figure 24: Interval plot for Pennsylvania using all years as replicate

Figure 22 shows that none of the confidence intervals are significantly different from each other. Figure 23 shows that none of the seeding rates are significantly different than the control group. Figure 24 shows that the seeding ate of 11.210 kg/ha^{-1} obtained the highest corn yield. The seeding rate of hairy vetch was determined to not be statistically significant with a p-value of 0.611.

Maryland – All years as replicates

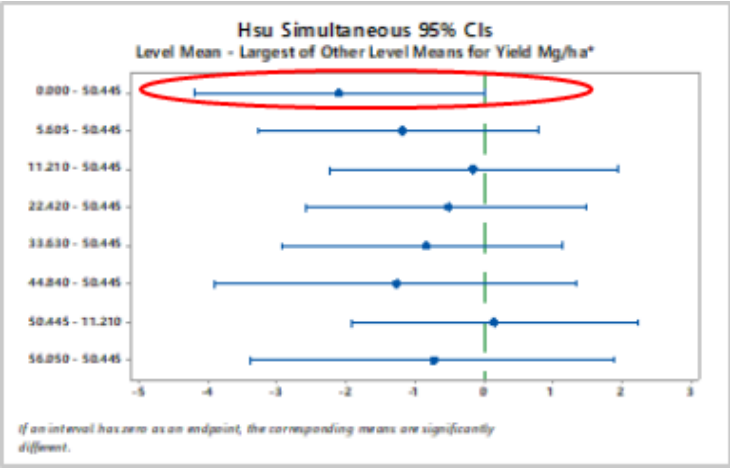


Figure 25: Hsu for Maryland using all years as replicates

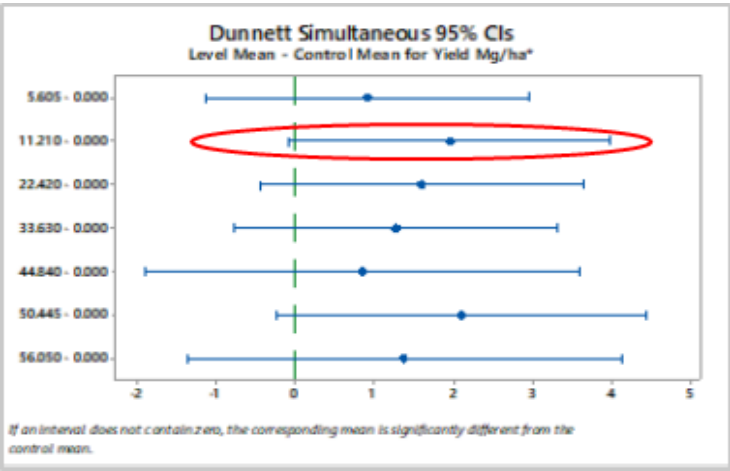


Figure 26: Dunnett's for Maryland using all years as replicates

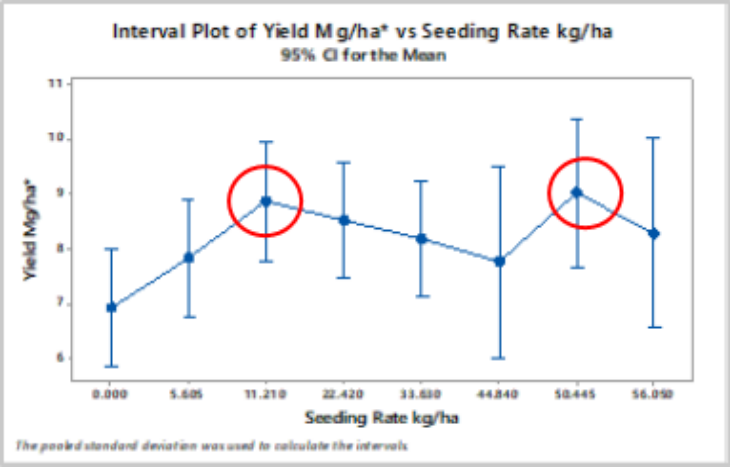


Figure 27: Interval plot for Maryland using all years as replicates

Figure 25 shows that most of the intervals have no significant difference. Figure 26 shows that none of the of the seeding rates have significantly different yields than the control group. Figure 27 shows that the seeding rate of 50.445 kg/ha^{-1} obtained the highest corn yield. The seeding rate of hairy vetch was determined to not be statistically significant with a p-value of 0.219.

Summary

Table 12 shows that hairy vetch seeding rate is only significant two of the four participating states. These states also happen to be the coldest states (MA and NY). In table 13 seeding rate was only significant in NY, this can be caused due to the lack of other possible factors used to cultivate corn. Table 14 has the summary of the coefficient found by adjusting a second order regression.

Info	All years			
	MA	NY	PA	MD
SR p-value	0	0.003	0.611	0.219
R-Sq	22.03%	13.65%	3.84%	7.13%
R-sq(adj)	15.45%	10.11%	0.00%	1.97%
Best SR through interval plot (kg/ha^{-1})	42.0375	50.445	11.21	50.445

Table 12: One-way ANOVA corn yield summary for all years as replicates

Info	All years			
	MA	NY	PA	NY
Regression p-value	0.016	0	0.555	0.12
SR p-value	0.114	0	0.418	0.228
SR^2 p-value	0.309	0.028	0.543	0.173
Lack-of-fit	0.001	0.1	0.877	0.381
R-Sq	4.90%	15.47%	1.16%	3.19%
R-sq(adj)	3.75%	14.27%	0.00%	1.71%

Table 13: Second order regression for all years as replicates

Coefficients		All years			
		MA	NY	PA	MD
β_0	Intercept	45867	18854	8123	8632
β_2	SR	2932	3420	254	406
β_{22}	SR^2	-3098	-2920	-350	-837

Table 14: Coefficients from the second order regression

4.2 Modeling: Corn Yield and Hairy Vetch Biomass vs Hairy Vetch Seeding rate

The coefficients obtained from the second-order ANOVAs in section 4.1.1. The regressions were graphed to give a visual aid to find the hairy vetch seeding rate that provides the largest amount of hairy vetch biomass. The hairy vetch biomass graph consists on graphing the different SR with same PD and HD obtained the highest hairy vetch biomass with different SR. Since no data corn yield data for the sites in NC was recollected, only the optimal hairy vetch seeding rate will be provided.

The latest HD from hairy vetch biomass is considered the PD for corn/silage yield. There is no information of the HD for corn/silage so the second-order ANOVA was done for corn/silage yield only using hairy vetch seeding rate. There is only corn yield information for MA, NY, PA and MD. This case consists on comparing hairy vetch biomass and corn/silage yield obtained from second order regressions to find the best hairy vetch seeding rate for both. The graphs help visualize which hairy vetch seeding rate provides the highest hairy vetch biomass and corn yield due to its convex shape. By deriving the equation given by the second order regression and setting it to zero, the highest point the convex shape will be the best seeding rate. Let that x_1 = PD, x_2= SR and x_3=HD. The equations were derived and solved for x_2. The main driver for this case is the best hairy vetch seeding rate for corn/silage yield. The hairy vetch biomass regression is represented in equation 1 and for corn/silage yield equation 2:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_{12} x_1 x_2 + \hat{\beta}_{13} x_1 x_3 + \hat{\beta}_{23} x_2 x_3 + \hat{\beta}_{11} x_1^2 + \hat{\beta}_{22} x_2^2 + \hat{\beta}_{33} x_3^2$$

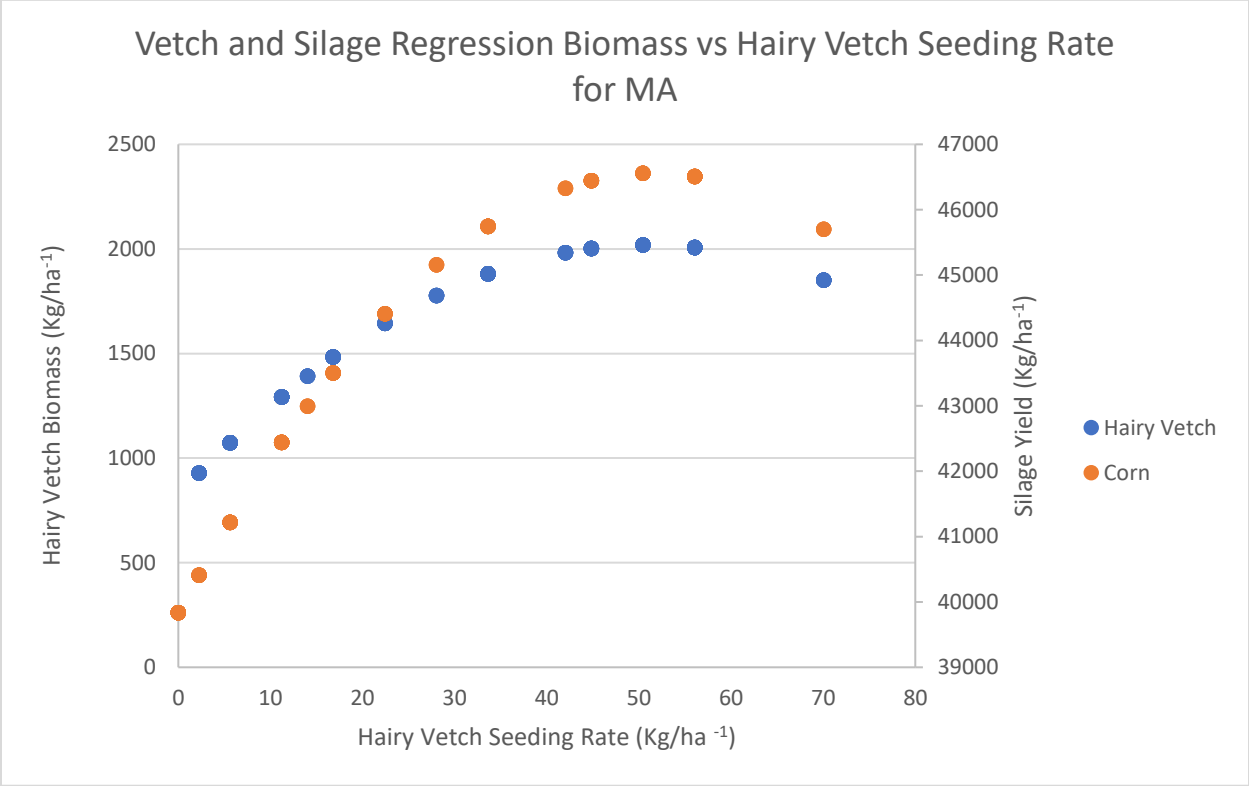
Equation 2 Hairy vetch biomass regression equation

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_2 x_2 + \hat{\beta}_{22} x_2^2$$

Equation 3 corn/silage yield regression equation

All years as replicates

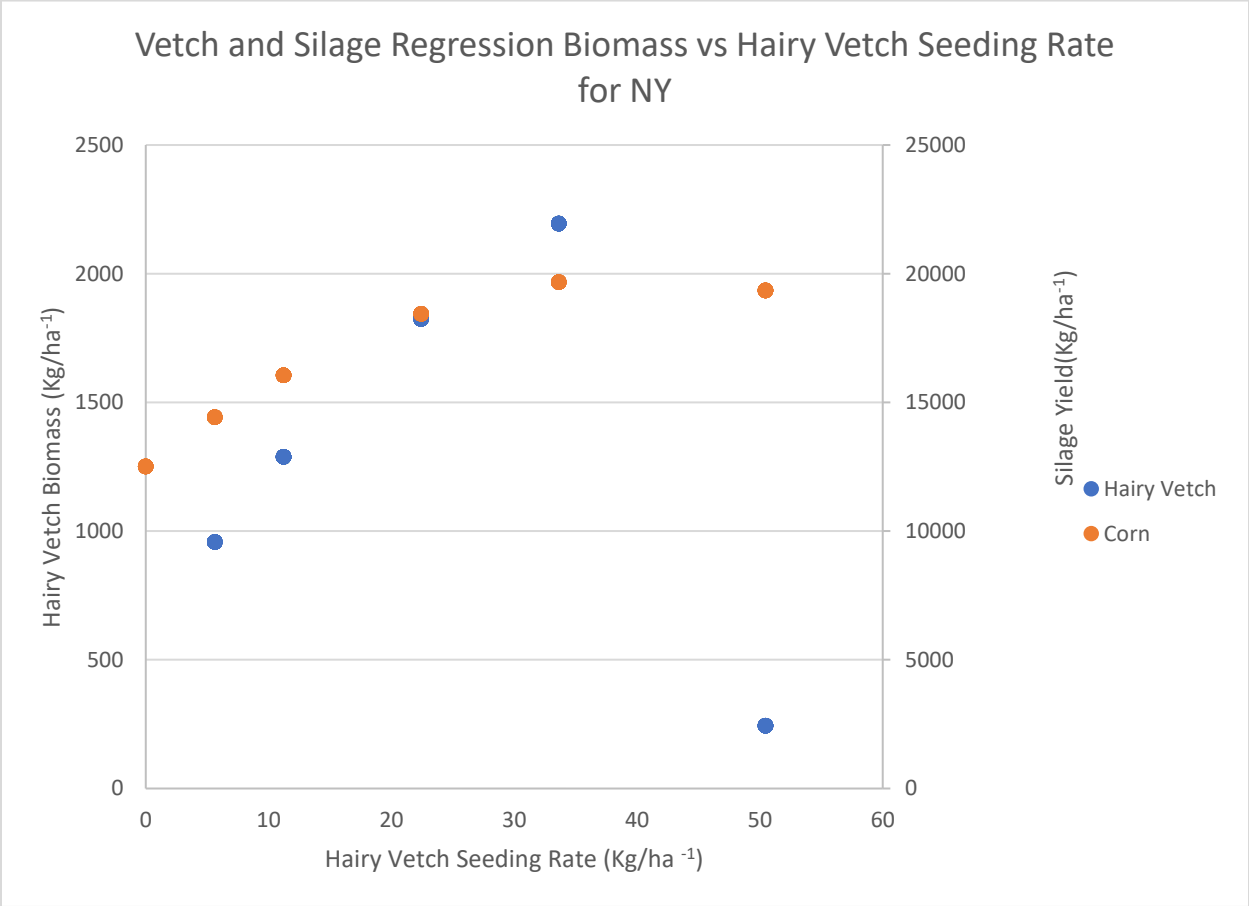
Massachusetts



Graph 20: Hairy Vetch Seeding Rate vs Regression Hairy Vetch and Corn Biomass for MA all years as replicates

Graph 20 shows that by deriving and solving for x_2 the best hairy vetch seeding rate is for hairy vetch biomass 37.80 kg/ha^{-1}. After deriving, setting to 0 and solving for x_2 the best hairy vetch seeding rate is 51.61 kg/ha^{-1} for silage yield. This is the only state where the best seeding rate for silage yield was higher than the one for hairy vetch biomass.

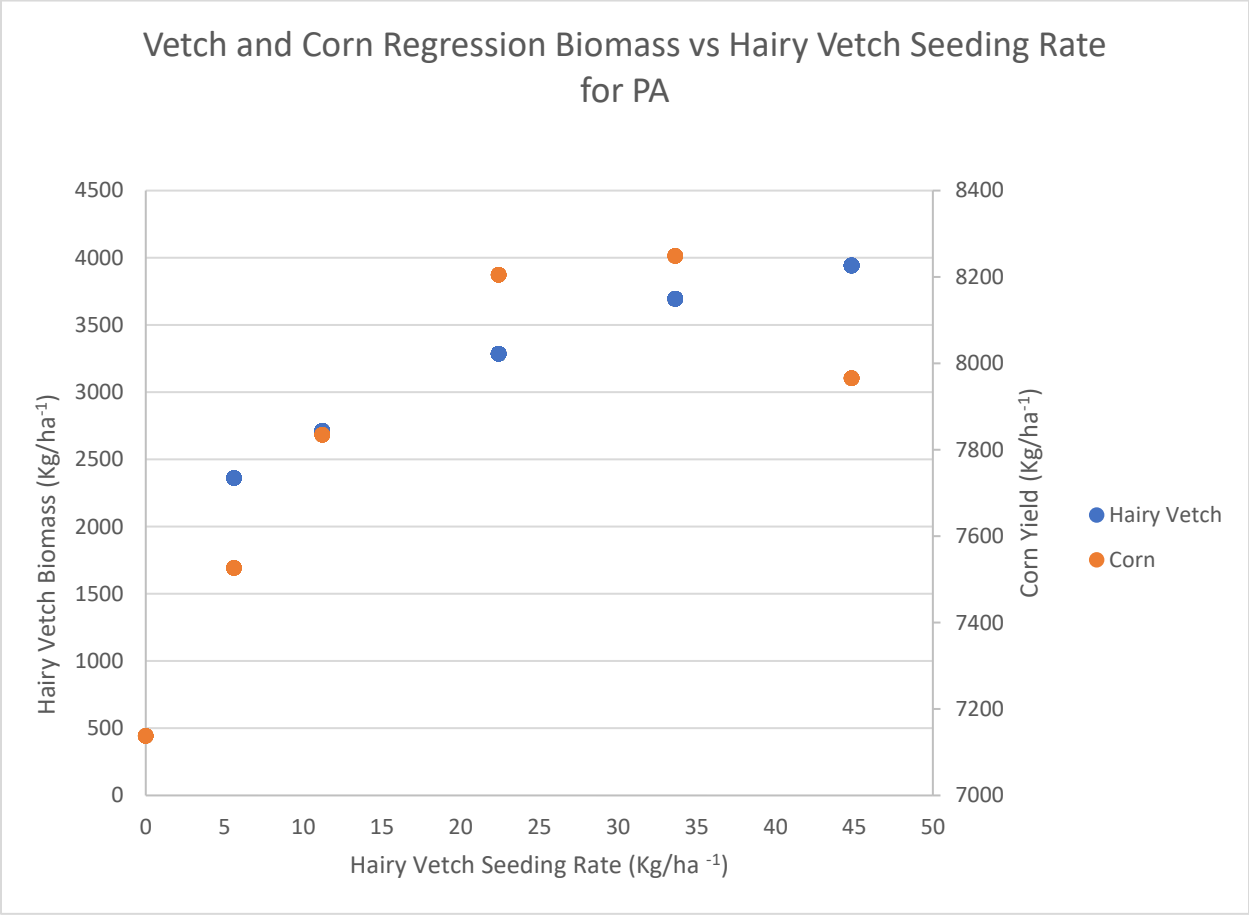
New York



Graph 21:Hairy Vetch Seeding Rate vs Regression Hairy Vetch and Corn biomass for NY all years as replicates

Graph 21 shows that when deriving and solving for x_2 the best hairy vetch seeding rate is 53.04 kg/ha^{-1} for hairy vetch biomass. After deriving, setting to 0 and solving for x_2 the best hairy vetch seeding rate is 39.99 kg/ha^{-1} for silage yield. Due to hairy vetch's high seeding rate needed to obtain the most biomass, it can be stated that for New York it is not necessary to use more than is 39.99 kg/ha^{-1} of hairy vetch seeds because it gives the highest yield of silage.

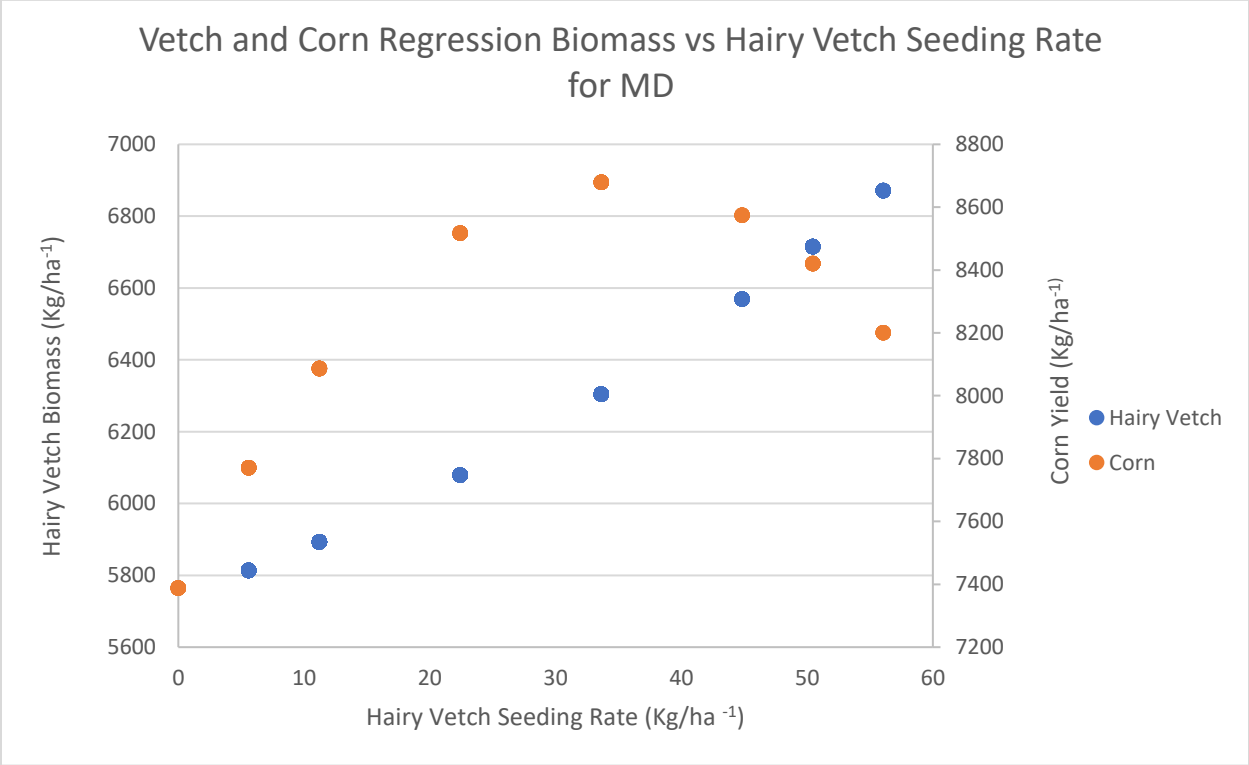
Pennsylvania



Graph 22: Hairy Vetch Seeding Rate vs Regression Hairy Vetch and Corn biomass for PA all years as replicates

When deriving, setting to 0 and solving for x_2 the best hairy vetch seeding rate is 56.23 kg/ha^{-1} for hairy vetch biomass. As can be seen in graph 22 hairy vetch biomass does not show a convex form. Thus, giving a best seeding rate value higher than what was used. After deriving and solving for x_2 the best hairy vetch seeding rate is 29.53 kg/ha^{-1} for corn yield. Due to hairy vetch's high seeding rate needed to obtain the most biomass, it can be stated that for Pennsylvania it is not necessary to use more than is 29.53 kg/ha^{-1} of hairy vetch seeds because it gives the highest yield of corn.

Maryland



Graph 23: Hairy Vetch Seeding Rate vs Regression Hairy Vetch and Corn biomass for MD all years as replicates

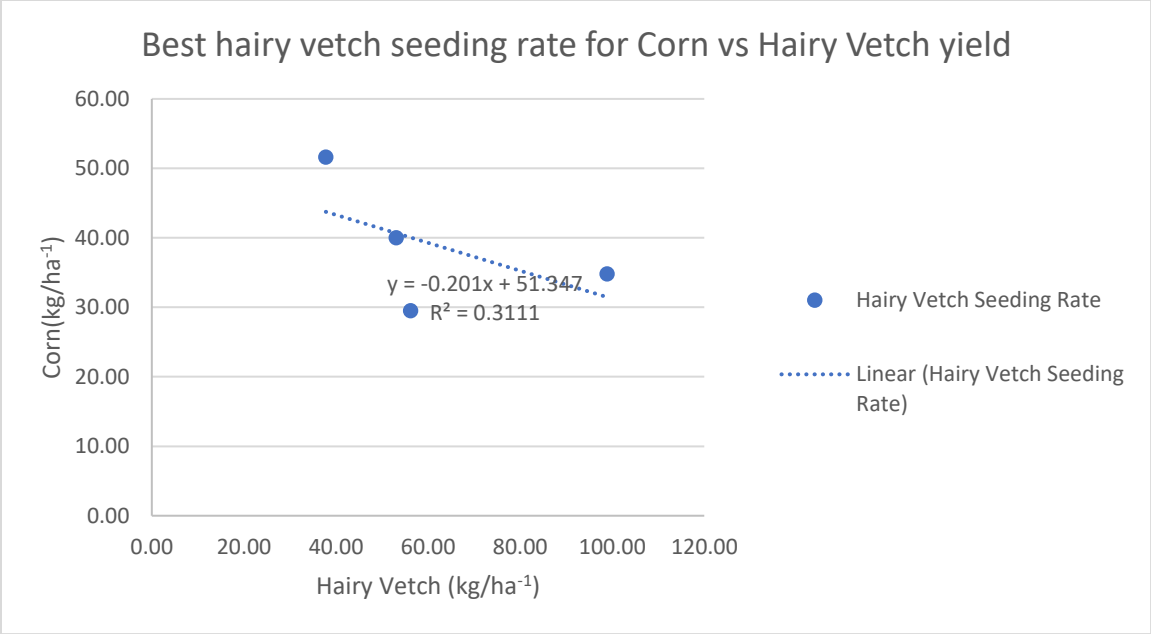
When deriving, setting to 0 and solving for x_2 the best hairy vetch seeding rate is 98.86 kg/ha^{-1} for hairy vetch biomass. As can be seen in graph 23hairy vetch biomass does not show a convex form. Thus, giving a best seeding rate value higher than what was used. After deriving and solving for x_2 the best hairy vetch seeding rate is 34.82 kg/ha^{-1} for corn yield. Due to hairy vetch's high seeding rate needed to obtain the most biomass, in can be clearly stated that for Maryland it is not necessary to use more than 34.82 kg/ha^{-1} of hairy vetch seeds because it gives the highest yield of corn.

Comparison

The following comparison analysis was done to create a basic rule of thumb for farmers to work with in a simple manner. Table 15 shows that for most states except for Massachusetts's that the hairy vetch seeding rate needed to obtain the most corn yield is less than the hairy vetch seeding need to obtain the best hairy vetch biomass. In graph #24 the best seeding rates per plant per state for corn and hairy vetch are plotted. The graph shows that if states were to be grouped, it is not necessary to spend more on buying hair vetch seeds to obtain the most biomass due to being able to obtain a high corn yield with a smaller seeding rate. This due to corn yield being the main revenue of the operation. As a rule of thumb, a seeding rate of 38.99 kg/ha^{-1} (Average from the seeding rates of the four states) can be used to obtain the highest corn yield.

Best Hairy Vetch seeding rate per state (kg/ha^{-1})					
Plant	MA	NY	PA	MD	Average
Hairy Vetch	37.80	53.04	56.23	98.86	61.48
Corn	51.61	39.99	29.53	34.82	38.99

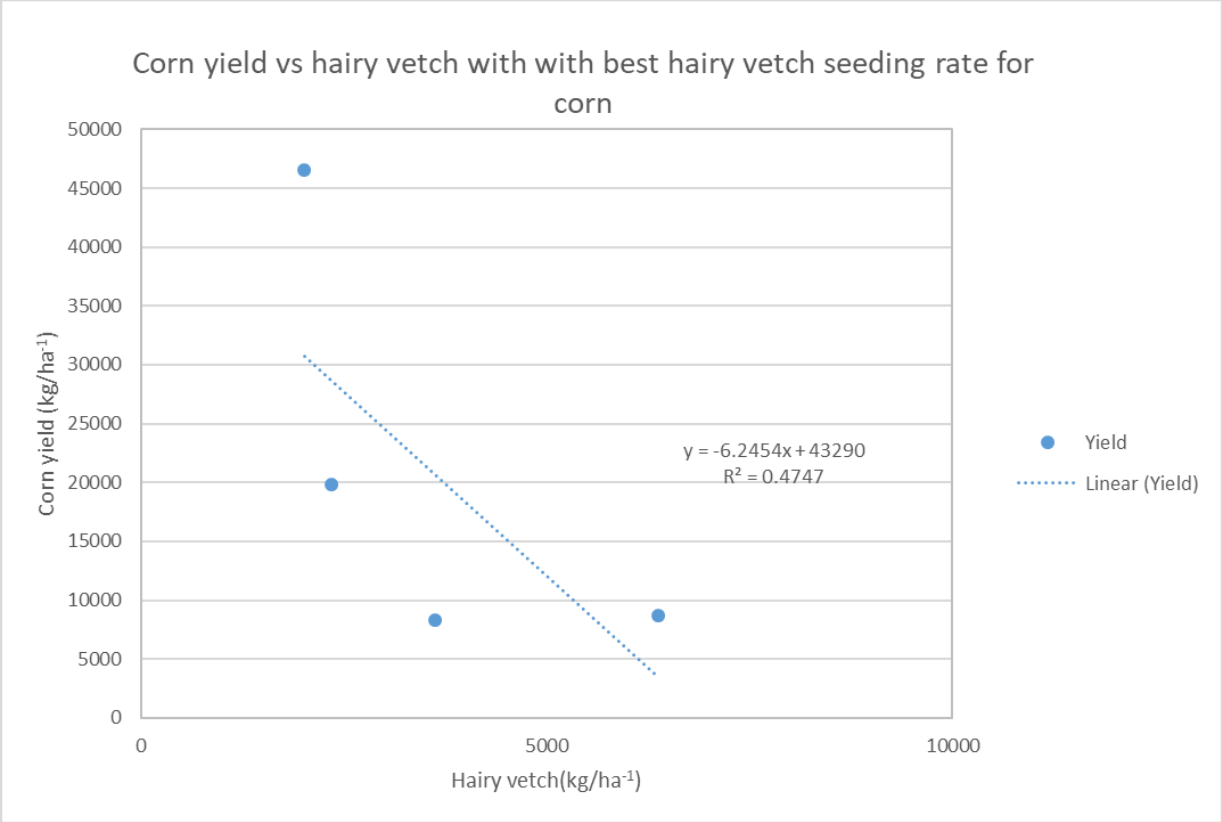
Table 15: Best Hairy Vetch seeding rate per state



Graph 24: Graph of Best SR for Corn vs Best SR for Hairy Vetch

Yield (With Best SR for corn kg/ha^{-1})				
Plant	MA	NY	PA	MD
HV	2018.39	2351.08	3628.87	6379.08
Corn	46560.72	19855.40	8270.62	8681.23
%HV/Corn	4.33%	11.84%	43.88%	73.48%

Table 16: Corn yield and hairy vetch biomass with the best SR from Corn



Graph 25: Best corn yield and hairy vetch biomass using the best seeding rate for corn yield

The next analysis consists on substituting the values of the best hairy vetch seeding rate for corn for the original second order equations for corn and hairy vetch. The original data shows that the amount of corn yield is much higher than the biomass from hairy vetch. It can be can be assumed that a percentage of nitrogen of the hairy vetch biomass will be passed on to the corn crop.

4.3 Optimization

This section uses the values obtained from the second order regressions in 4.2 to create an optimal economic analysis that will compare the usage of hairy vetch vs nitrogen (N) fertilizer. Another comparison will be between using the best hairy vetch seeding rate of corn vs hairy vetch. The yield is the main driver for decision making. First, we want to maximize corn yield and second we want to minimize costs. The following formulas were used to find the values needed for the analysis.

$$\text{Cost for N fertilizer } (\$/ha^{-1}) = \text{N replacement value} * \text{HV biomass} * \text{Fertilizer cost} \quad (4)$$

$$\text{Cost of best SR for corn/silage } (\$/ha^{-1}) = \text{Best SR for Corn/silage} * \text{ HV seed cost} \quad (5)$$

$$\text{Cost of best SR for HV } (\$/ha^{-1}) = \text{Best SR for HV} * \text{ HV seed cost} \quad (6)$$

The assumptions for this analysis are:

- HV biomass yield was obtained by substituting the best HVSR for corn/silage yield in the original regression equations for hairy vetch biomass and corn/silage yield
- The N replacement value of 0.88 averaging values from "Organic and Supplemental Nitrogen Sources for Field Corn production" [54] and assumed for all states.
- The cost of corn/silage used was $5.96/kg and was obtained from averaging costs in section 2.2.3.
- The cost of urea used was $0.62/kg and was obtained from averaging costs in section 2.2.3.

State	Hairy Vetch Seeding Rates (kg/ha^{-1})	Hairy Vetch Biomass from Regression (kg/ha^{-1})	Nitrogen replacementent value (kg/ha^{-1})	Best hairy vetch Seeding Rate for Hairy Vetch (kg/ha^{-1})	Best Hairy Vetch seeding rate for corn (kg/ha^{-1})	Cost with best hairy vetch seeding rate for corn ($/ha^{-1})	Cost equivalent for hydrogen fertilizer with ($/ha^{-1})	Cost with levels hairy vetch seeding rate for Hairy Vetch ($/ha^{-1})	Cost savings when using cover crop vs nitrogen fertilizer (%)	Costs savings by using best hairy vetch seeding rate for corn instead of hairy vetch (%)
MA	0.00	825.64	726.57	37.80	51.61	307.44	447.10	0.00	31.24%	-
	2.24	928.40	816.99	37.80	51.61	307.44	502.75	13.36	38.85%	-2201.96%
	5.61	1073.86	945.00	37.80	51.61	307.44	581.52	33.39	47.13%	-820.79%
	11.21	1293.17	1137.99	37.80	51.61	307.44	700.28	66.78	56.10%	-360.39%
	14.01	1391.99	1224.95	37.80	51.61	307.44	753.79	83.47	59.21%	-268.31%
	16.82	1483.58	1305.55	37.80	51.61	307.44	803.39	100.17	61.73%	-206.93%
	22.42	1645.08	1447.67	37.80	51.61	307.44	890.85	133.55	65.49%	-130.20%
	28.03	1777.68	1564.36	37.80	51.61	307.44	962.65	166.94	68.06%	-84.16%
	33.63	1881.37	1655.60	37.80	51.61	307.44	1018.80	200.33	69.82%	-53.46%
	37.80	**1939.76**	**1776.18**	**37.80**	**51.61**	**307.44**	**1093.00**	**225.17**	**71.87%**	**-36.53%**
	42.04	1982.71	1744.78	37.80	51.61	307.44	1073.68	250.41	71.37%	-22.77%
	44.84	2002.04	1761.79	37.80	51.61	307.44	1084.15	267.11	71.64%	-15.10%
	50.45	2019.01	1776.73	37.80	51.61	307.44	1093.34	300.50	71.88%	-2.31%
	51.61	**2018.80**	**1776.54**	**37.80**	**51.61**	**307.44**	**1093.22**	**307.44**	**71.88%**	**0.00%**
	56.05	2007.08	1766.23	37.80	51.61	307.44	1086.88	333.88	71.71%	7.92%
	70.06	1850.80	1628.70	37.80	51.61	307.44	1002.25	417.35	69.33%	26.34%
NY	0.00	586.00	515.68	53.04	39.99	238.22	317.33	0.00	24.93%	-
	5.61	958.00	843.04	53.04	39.99	238.22	518.78	33.39	54.08%	-613.47%
	11.21	1288.50	1133.88	53.04	39.99	238.22	697.75	66.78	65.86%	-256.74%
	22.42	1825.00	1606.00	53.04	39.99	238.22	988.28	133.55	75.90%	-78.37%
	33.63	2195.50	1932.04	53.04	39.99	238.22	1188.91	200.33	79.96%	-18.91%
	39.99	**2331.90**	**2052.07**	**53.04**	**39.99**	**238.22**	**1262.77**	**238.24**	**81.14%**	**0.01%**
	50.45	2440.00	2147.20	53.04	39.99	238.22	1321.31	300.50	81.97%	20.73%
	53.04	**2444.46**	**2151.13**	**53.04**	**39.99**	**238.22**	**1323.73**	**315.98**	**82.00%**	**24.61%**
PA	0.00	1973.97	1737.10	56.23	29.53	175.91	1068.95	0.00	83.54%	-
	5.61	2363.06	2079.49	56.23	29.53	175.91	1279.65	33.39	86.25%	-426.85%
	11.21	2711.33	2385.97	56.23	29.53	175.91	1468.25	66.78	88.02%	-163.43%
	22.42	3285.43	2891.18	56.23	29.53	175.91	1779.13	133.55	90.11%	-31.71%
	29.53	**3564.83**	**3137.05**	**56.23**	**29.53**	**175.91**	**1930.44**	**175.89**	**90.89%**	**-0.01%**
	33.63	3696.25	3252.70	56.23	29.53	175.91	2001.60	200.33	91.21%	12.19%
	44.84	3943.82	3470.56	56.23	29.53	175.91	2135.66	267.11	91.76%	34.14%
	56.23	**4028.14**	**3544.76**	**56.23**	**29.53**	**175.91**	**2181.32**	**334.97**	**91.94%**	**47.49%**
MD	0.00	5745.34	5055.90	98.86	38.99	232.26	3111.23	0.00	92.53%	-
	5.61	5814.43	5116.70	98.86	38.99	232.26	3148.64	33.39	92.62%	-595.63%
	11.21	5893.19	5186.01	98.86	38.99	232.26	3191.29	66.78	92.72%	-247.81%
	22.42	6079.76	5350.19	98.86	38.99	232.26	3292.33	133.55	92.95%	-73.91%
	33.63	6305.05	5548.44	98.86	38.99	232.26	3414.32	200.33	93.20%	-15.94%
	38.99	**6426.40**	**5655.23**	**98.86**	**38.99**	**232.26**	**3480.03**	**232.25**	**93.33%**	**-0.01%**
	44.84	6569.05	5780.77	98.86	38.99	232.26	3557.29	267.11	93.47%	13.05%
	50.45	6715.57	5909.70	98.86	38.99	232.26	3636.63	300.50	93.61%	22.71%
	56.05	6871.77	6047.16	98.86	38.99	232.26	3721.21	333.88	93.76%	30.44%
	98.86	**8384.07**	**7377.98**	**98.86**	**38.99**	**232.26**	**4540.16**	**588.90**	**94.88%**	**60.56%**

Table 17: Economic analysis of all hairy vetch seeding rates across Massachusetts (MA), New York (NY), Pennsylvania (PA) and Maryland (MD)

Table 17 is the summary of the overall economic analysis that compares nitrogen fertilizer vs hairy vetch as a cover crop. The last two columns have the concluding results with the cost savings using the best HVSR for corn vs N fertilizer and savings between using the best HVSR for corn instead of the best one for HV. To calculate these results formulas 7 and 8 were used.

$$\text{Cost savings using HVSR corn vs N fertilizer (\%)} = \frac{Cost\ equivalent\ for\ N\ fertilizer - Cost\ with\ best\ HVSR\ for\ Corn/Silage}{Cost\ equivalent\ for\ N\ fertilizer} \quad (7)$$

$$\text{Costs savings by using best for corn/silage vs HV (\%)} = \frac{Cost\ with\ best\ HVSR\ for\ Hairy\ Vetch - Cost\ with\ best\ HVSR\ for\ Corn/Silage}{Cost\ with\ best\ HVSR\ for\ Hairy\ Vetch} \quad (8)$$

State	HV Biomass from Regression (kg/ha^{-1})	N replacementent value(kg/ha^{-1})	Cost with best HVSR for corn/silage ($/ha^{-1})	Cost equivalent for N fertilizer ($/ha^{-1})	Cost with best SR for Hairy Vetch ($/ha^{-1})	Cost savings by using best HVSR for corn/silage vs N fertilizer (%)	Costs savings by using best for corn/silage vs HV (%)
MA	2018.39	1776.18	307.44	1093.00	225.17	71.87%	-36.53%
NY	2351.08	2068.95	238.22	1273.16	315.95	81.29%	24.60%
PA	3628.87	3193.41	175.91	1965.11	334.96	91.05%	47.48%
MD	6379.08	5613.59	207.42	3454.41	588.90	94.00%	64.78%

Table 18: Economic analysis summary for best in each state

Table 18 is a summary of table 17. As it can be seen, the use of hairy vetch as a cover crop was overall more favorable across all states. The HV Biomass from regression is obtained by plugging the values of the best hairy vetch seeding rate for corn/silage yield in equation 2. The N replacement value column is obtained by multiplying the NFRV of 0.88 with the value found in the column before in the HV Biomass from regression column. The following columns use equations 4, 5 and 6. In both cost savings columns there is a trend that the more south the state is, the more beneficiary it is to use a cover crop to obtain nitrogen. The only state that did not have savings from using the best HVSR for corn/silage instead the on for hairy vetch was Massachusetts. The recommended HVSR for MA is the best HVSR for vetch which is 37.80 kg/ha^{-1}. For NY, PA and MD the HVSR recommended is the best one for corn/silage being 39.99, 29.53 and 34.82 kg/ha^{-1}.

Chapter 5. Alzheimer's Disease Study

This work discusses the finding of Alzheimer's potential biomarkers using publicly available microarray data first reported by Dunckley et al 2006 [8] and proposes a possible signaling path among them. The identification of potential AD biomarkers from microarray data is casted as a multiple criteria optimization (MCO) problem. Our group first explored MCO in the identification of potential cancer biomarkers in Sánchez-Peña et al 2013 [9] . The results of the method are successfully validated as related to the illness through comparison with available scientific literature in many cases. However, there were genes identified through our methods that turned out to have a considerable amount of evidence to also be related to cancer, which have not been formally associated to it. These offer an important opportunity for future work. The aim of a MCO problem is to find the best compromises between two or more conflicting criteria. Formally, these best compromises are located in the Pareto efficient frontier of the set of candidates evaluated in all criteria of interest (also called performance measures in this work). We propose that the genes in the efficient frontier in the present analyses, built with performance measures related to changes in gene expression, are potential AD biomarkers. It is our premise that the changes in expression of these important genes are correlated among them, that is, that there is a signal among them. We have proposed that this signal can be modeled as a cyclical correlation path, the elicitation of which constitutes a highly combinatorial optimization problem. In Mathematics, and particularly in the field of Operations Research, this problem is called the Traveling Salesman Problem (TSP). This study attempts the characterization of the signaling path in Alzheimer's using the well-known TSP combinatorial optimization formulation [9] as was first used in our paper, Lorenzo et al 2015 [10] in the context of cervix cancer.

The proposed method as an MCO problem and the characterization via the TSP formulation is tested on an AD microarray database reported by Dunckley et al 2006 [8]. The results from the MCO problem are validated through scientific literature and are the input for the TSP. The TSP path provides correlations that have not been reported yet but that are biological plausible thereby offering new research opportunities.

5.1 Characterization, Modeling and Optimization

The procedure to select the genes of interest through MCO is explained in detail Sánchez-Peña et al 2013 [6], [9], [10]. In brief, MCO will select a family of solutions (genes in this case) that have the best compromises between the performances measures considered in the analysis. The solution to the MCO problem is called the Pareto-efficient frontier, which in turn contains *efficient solutions*, as depicted in Figure 28.

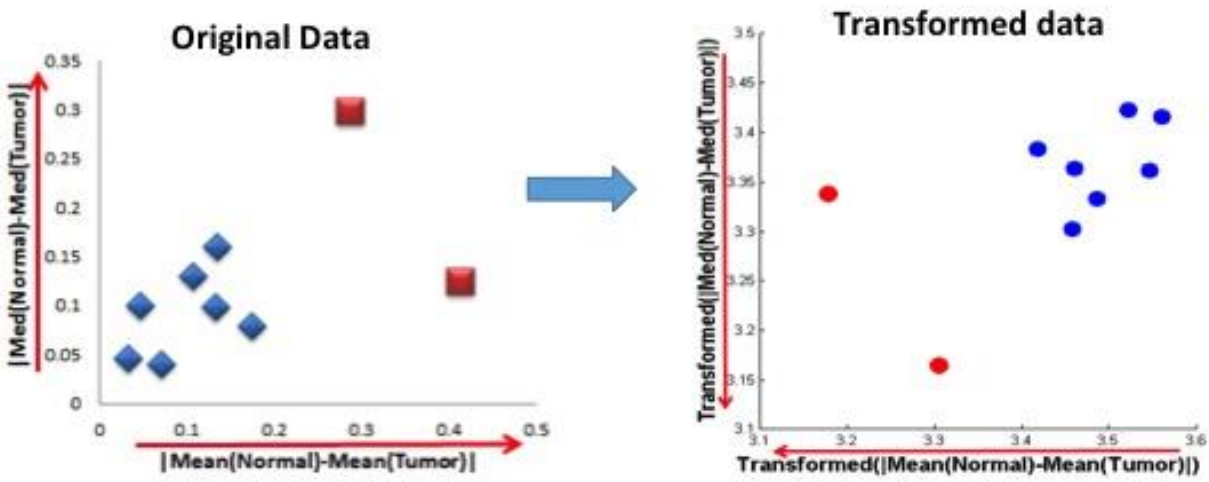


Figure 28: General representation of multiple criteria optimization problem considering two performance measures to be maximized. The solutions represented as big squares are deemed Pareto-efficient.

In order to find how the gene expression changes were related in the genes identified through MCO, the relationship between each pair of genes was modeled as linear statistical correlation. The basic correlation formula denoted as ρ_{XY} and between random variables X and Y is [67]:

$$\rho_{XY} = \frac{cov(X,Y)}{\sqrt{V(X)V(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

The correlation between X and Y could be zero, positive or negative and is bounded as follows:

$$-1 \leq \rho_{XY} \leq +1$$

Because correlation values range from -1 to 1, when their absolute values are closer to 1, these indicate strong correlations. Simply put, two genes will be strongly correlated if the absolute value of their correlation is close to 1. If each gene is represented through a node in a graph, then the undirected arc joining a pair of genes can contain their absolute correlation value. The MCO procedure identified 10 genes shown in Table 18, and the most correlated cyclical path between the changes in their expression leads naturally to the Travelling Sales Problem (TSP) formulation. The TSP formulation as described in Orlin, Ahuja & Magnati et. al 1993 [68] considers c_{ij} representing the cost of travelling from city *i* to city *j*. A binary variable y_{ij}, indicates whether or not the salesman travels from city *i* to city *j*. Additionally let us define flow variables x_{ij} on each arc (*i,j*) and assume that the salesman has *n*-1 units available at node 1, which is arbitrarily selected as a "source node", and he must deliver 1 unit to each of the other nodes [9]. The optimization model is as follows:

$$Minimize \sum_{(i,j)\in A} c_{ij} y_{ij} \qquad (P1a)$$

$$\sum_{1\le j\le n} y_{ij} = 1 \quad \forall i = 1,2,\dots,n \qquad (P1b)$$

$$\sum_{1\le i\le n} y_{ij} = 1 \quad \forall\, j = 1,2,\dots,n \qquad (P1c)$$

$$Nx = b \qquad (P1d)$$

$$x_{ij} \le (n-1) y_{ij} \quad \forall (i,j) \in A \qquad (P1e)$$

$$x_{ij} \ge 0 \quad \forall (i,j) \in A \qquad (P1f)$$

$$y_{ij} = 0 \; or \; 1 \quad \forall (i,j) \in A \qquad (P1g)$$

Following the description in [7], let $A' = \{(i,j): y_{ij} = 1\}$ and let $A'' = \{(i,j): x_{ij} > 0\}$. The constraints (P1b) and (P1c) imply that exactly one arc of A' leaves and enters any node i; therefore, A' is the union of node disjoint cycles containing all of the nodes of N. In general, any integer solution satisfying (P1b) and (P1c) will be union of disjoint cycles; if any such solution contains more than once cycle; they are referred to as subtours, since they pass through only a subset of nodes.

In constraint (P1d) N is an *nxm* matrix, called the *node-arc incidence matrix* of the minimum cost flow problem. Each column N_{ij} in the matrix corresponds to the variable x_{ij}. The column N_{ij}

has a +1 in the *i*th row, and -1 in the *j*th row; the rest of its entries are zero. Constraint (P1d) ensures that *A"* is connected since we need to send 1 unit of flow from node 1 to every other node via arcs in *A"*. The forcing constraints (P1e) imply that *A"* is a subset *A'*. These conditions imply that the arc set *A'* is connected and thus cannot contain subtours [10], [68].

An illustration of how the resulting graph would look like is shown in Figure 29. In our method, we solve the TSP to optimality capitalizing in the shortlist provided by the first part of the analysis, the MCO procedure. A Matlab code aided by the branch-and-bound method was used to this end [11], [69].

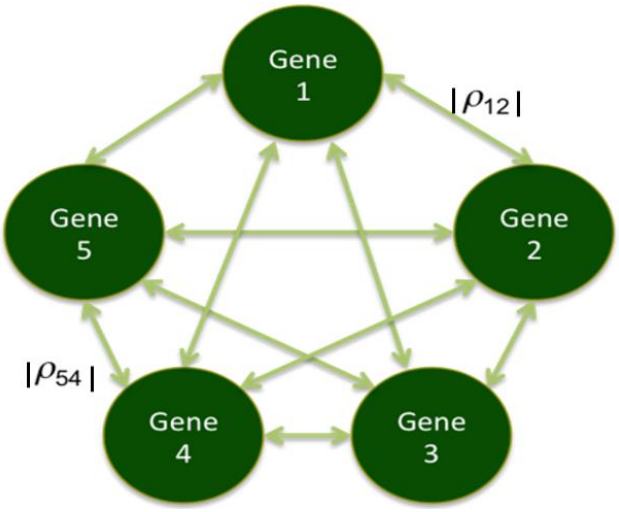


Figure 29: The Travelling Sales Problem (TSP) representation. A solution must visit each node once and return to its initial node, thereby creating a cyclic path

In summary, TSP consists of finding the most correlated cyclical tour, in mathematical terms. This means that the tour to be selected is the one that produces the biggest sum of correlations at the end, visiting each gene exactly once. Coordinated pairwise behavior can be measured as a statistical correlation. The statistical correlation was computed as linear following the method presented by Lorenzo et al 2015 [10], [11].

5.2 Results

The first results of the proposed method include the analysis of the microarray database GSD2795 used by Dunckely et. al. [8] related to Alzheimer's disease focused on neurofibrillary tangles (NFTs). The database consists of 19 cases and 14 control tissues supplied from the brain banks of the Alzheimer's Disease Center (ADC) program. The database has a total of 54,675 genes. Using the MCO procedure, the first 3 frontiers were identified, and they contained 10 potential biomarkers that are listed in Table 18. Reducing the 54,675 genes to only 10 of them evidences the screening power of the MCO method.

Accession Number	Identifier
1553551_s_at	ND2
1553538_s_at	COX1,MT-CO1
224373_s_at	ND4, Hnrnpm,DCAF6
1555653_at	HNRNPA3
1553588_at	ND3,SH3KBP1
201492_s_at	RPL41
212788_x_at	FTL
203540_at	GFAP
200095_x_at	RPS10
229353_s_at	NUCKS1

Table 19: List of 10 potential biomarkers identified in the first 3 frontiers through the MCO problem

The computations of the correlations were carried out in a pairwise manner and the results are presented in the correlation matrix in Figure 31. Their absolute values allow to assess how strong these correlations are in a scale from 0 (not correlated) to 1 (perfectly linearly correlated). Two genes will be strongly correlated if the absolute value of their correlation is close to 1 and their correlation decreases, as the coefficient gets closer to 0. Out of the 10 potential biomarkers, the two most correlated were COX1 and ND2. COX1 correlates with four biomarkers while ND2 correlates with three biomarkers. Figure 31 shows the result of modeling the expression changes of the selected genes as a TSP.

	COX1	ND2	ND3	HNRNPA3	RPS10	RPL41	GFAP	FTL	ND4	NUCKS1
COX1	0	0.989	0.959	0.971	0.690	0.529	0.167	0.618	0.996	0.745
ND2	0.989	0	0.942	0.972	0.704	0.558	0.200	0.626	0.992	0.774
ND3	0.959	0.942	0	0.951	0.654	0.601	0.117	0.651	0.958	0.660
HNRNPA3	0.971	0.972	0.951	0	0.680	0.548	0.131	0.603	0.977	0.715
RPS10	0.690	0.704	0.654	0.680	0	0.859	0.670	0.917	0.706	0.888
RPL41	0.529	0.558	0.601	0.548	0.859	0	0.565	0.952	0.551	0.655
GFAP	0.167	0.200	0.117	0.131	0.670	0.565	0	0.639	0.184	0.579
FTL	0.618	0.626	0.651	0.603	0.917	0.952	0.639	0	0.628	0.713
ND4	0.996	0.992	0.958	0.977	0.706	0.551	0.184	0.628	0	0.756
NUCKS1	0.745	0.774	0.660	0.715	0.888	0.655	0.579	0.713	0.756	0

Figure 30: Correlation matrix indicating how strong the correlations between the 10 potential biomarkers are, values close to 1 indicates strong correlations

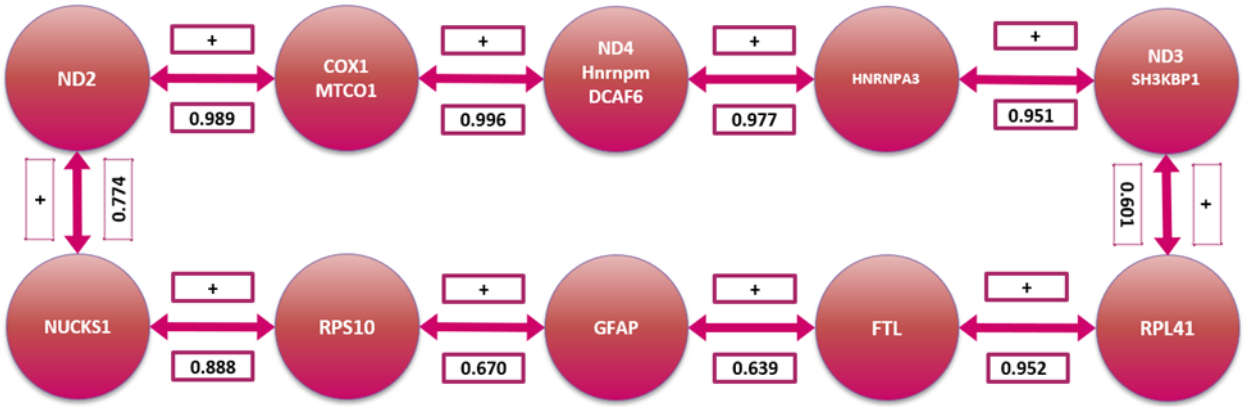


Figure 31: Gene coordinated behavior pathway determined by the Travelling Sales Problem solution

5.3 Discussion

Starting with the first node in Figure 31, this corresponds to ND2 that codes for the mitochondrial encoded NADH Dehydrogenase 2. Out of the 10 biomarkers identified using the proposed method, 4 are related to the mitochondria. Mitochondria play a role in neuronal cell survival due to their role as regulator of energy metabolism and cell death pathways. MT-ND2 gene codes for the mitochondrial encoded NADH Dehydrogenase 2. There was a report that mutations in this gene have been observed in AD brains [70], but a later report concluded that the mutation was not specifically associated with AD [71]. Drosophila ND2 mutants show progressive neurodegeneration [72]. Next gene in the TSP path corresponds to COX1 which is a mitochondrial gene. COX1 can also be identified as MT-CO1. It encodes a protein that forms part of the cytochorome C Oxidase enzyme complex. In an experiment that was testing the protective effect of melatonin, COX1 was found to have an increment in its expression[73], [74]. In that study it was also found that the expression of ND1 and ND4 were increased as a result of the melatonin treatment. The next gene in the path codes for MT-ND4, or mitochondrial encoded NADH Dehydrogenase 4, that forms part of the core subunit of the mitochondrial NADH dehydrogenase (Complex I). The mRNA expression for this gene has been reported to decrease in the hippocampus and inferior parietal lobule of Alzheimer's Disease patients [75]. Probe 224373_s_at recognizes ND4, Hnrnpm and DCAF6 [76]. *DCAF6* is related to androgen receptors which are involved with the development of Alzheimer's [75], [77], as well as in the temporal cortex of AD patients[78]; this supports the usefulness of the method that can move past the intermediary genes . Mutations of the *DCAF6* gene have been also linked to maternally inherited schizophrenia [79]. A study for the expression of mitochondrial ND2 and ND4 genes in amyotrophic lateral sclerosis (ALS), found that the anterior neurons in the cervical spinal cord had reduced mtDNA gene levels and an increment in the amount of mtDNA deletions [80].

Following the correlation path, the next probe (HNRNPA3) corresponds to control for the microarray. In literature some of the probes recognize more than one gene for example probe 1553588_at recognizes ND3 and SH3KBP1[81]. The next gene, MT-ND3, codes for another member of mitochondrial Complex I. The product of this gene was shown to bind to a peptide corresponding to 25 amino acids of the C-terminal of amyloid-beta [82]. The authors of the report proposed that the ND3 Amyloid- beta interaction could explain in part the lower activity of Complex I in astrocytes and neurons.

Following the TSP path, the next gene is *RPL41*, which codes for the Ribosomal Protein L41. It has been suggested that the *RPL41* product plays different roles in cell proliferation and differentiation during neurogenesis [83].This protein was found also to help with virus replication in some avian viruses: infectious bursal disease virus [84] and Sindbis virus[85]. It also promotes the expression of the c-myc proto-oncogene[86]. *RPL41, it is also associated with ATF4 degradation* [87] that in turn mediates neurodegeneration in Alzheimer's Disease and transmission of a neurodegenerative signal through some brain regions [88], [89]

FTL gene codes for the ferritin light polypeptide protein, which is the next node in Figure 32. Ferritin is the main intracellular iron storage protein. It has been reported that levels of ferritin are lower in the peripheral blood mononuclear cells from AD patients, and it has been proposed that this change is one of the factors responsible for the dysregulation of iron found in AD patients [90].FTL is also involved in the proteolytic cleavage of the β-amyloid precursor protein by its interaction with PEN-2 promoting γ-secretase activity and consequently the production of β-amyloid [91]. *FTL* product is associated with neurodegenerative disorder related with iron accumulation in the brain, primarily in the basal ganglia [92] and enhances oxidative damage [93]–[95].

Other selected gene is *GFAP*; this gene is particularly interesting because it is mostly expressed in the brain. *GFAP* overexpression is a characteristic of astrocyte reactivity [96]. Mutations in this gene are responsible for Alexander disease (a rare disorder of the central nervous system*)*, leukodystrophy, and Alzheimer's disease. In the case of Alexander disease, myelin is destroyed[97].

RPS10, next gene in the TSP pathway, codes for one of the proteins of the 40S ribosomal subunits. The expression of this gene was found to be lower in Schizophrenia patients than in controls [98]. Changes in expression of this gene has been observed in colorectal cancer [99]. The RPS10 protein interacts with the HIV-1 Nef protein [100]. Mutations in the *RPS10* are linked to diamond-blackfan anemia [101]. *RPS10* is part of the ribosomal protein family that has been reported to change its expression in neurodegenerative diseases such as Alzheimer's [102].

The next gene in Figure 32, corresponds to *NUCKS1* that codes for nuclear casein kinase and cyclin-dependent kinase substrate 1. Its product has been shown recently to participate in homologous recombination DNA repair [103]. Is interesting to note that the *NUCKS1* product is also used by the HIV-1 for the viral transcription of its genetic material [104] and has been reported as a biomarker for some cancers [105], [106].The expression change of the *NUCKS1* has been linked to mood disorders and to Parkinson's disease [107], [108]

In summary, grouping genes that code for proteins of a bigger complex one finds: *ND4*, *ND2*, and *ND3*, mitochondrial encoded NADH dehydrogenase (Complex I) genes. Another mitochondrial-encoded gene is *COX1* and its product also forms part of one of the electron transport complexes. The correlations between the expression changes for these genes are high, more than 0.9. These results coincide with the reports of mitochondrial genes expression change in Alzheimer [109], [110] and other neurodegenerative diseases [111].There are two genes that

code for different ribosomal proteins: *RPS10* and *RPL4* in the TSP separated by *FTL* and *GFAP*. Some of the selected genes have been reported to change their expression in different cancers and some are known to help in viral infections.

5.4 Comparisons

5.4.1 Correlation VS Complement of p-value: TSP

Another statistical performance measurement is the p-value. In order to use the p-value in the in this code, it was subtracted by one. This was done because p-values show more significance when the value is lower, in contrast to correlation. This value is referred to as the complement of p-value. This measurement provided a different optimal path than correlation. Figures 31 and 33 correspond to the route found initially with correlation values. Figures 32 and 34 correspond to the route found with the complement of p-values. Figure 31 and 32 the optimal routes found with both performance measurements. Both measurements provided two different routes. Figure 31 shows higher correlations between the genes than in Figure 34 except for the relationship between NUCKS1 and ND2. Figure 32 and 33 shows that there are the same values of the complement of p-value. Graph 26 compares the values of each of the routes. This graph shows that the routes found with the correlation values initially had the best values in correlation and complement of p-value.

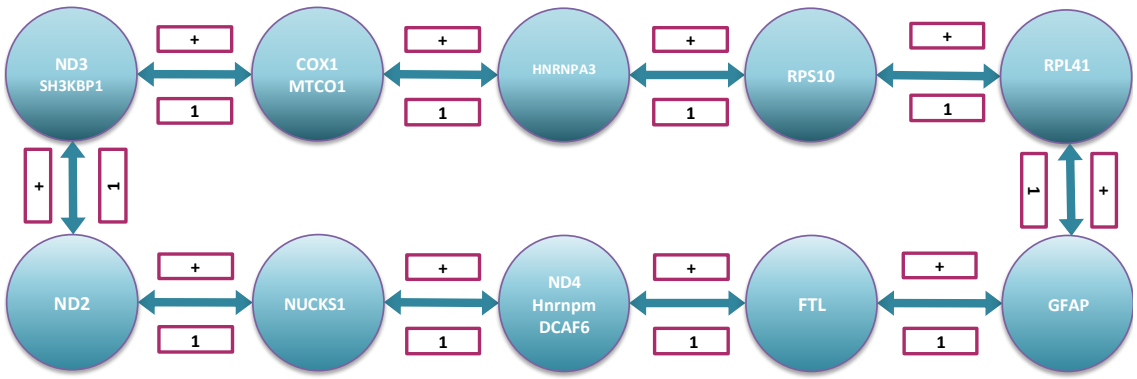


Figure 32: TSP by complement of p-value

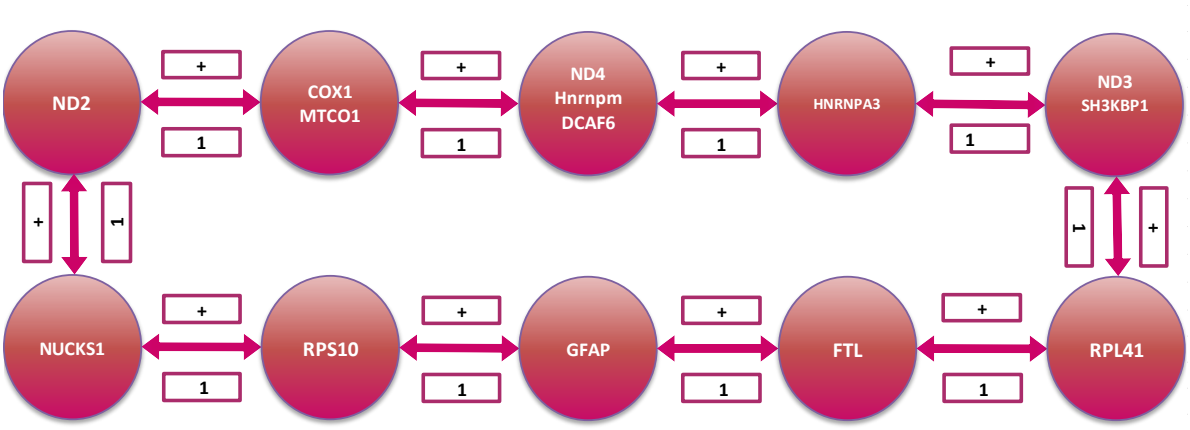


Figure 33: TSP of the correlation pathway with corresponding complement of p-value

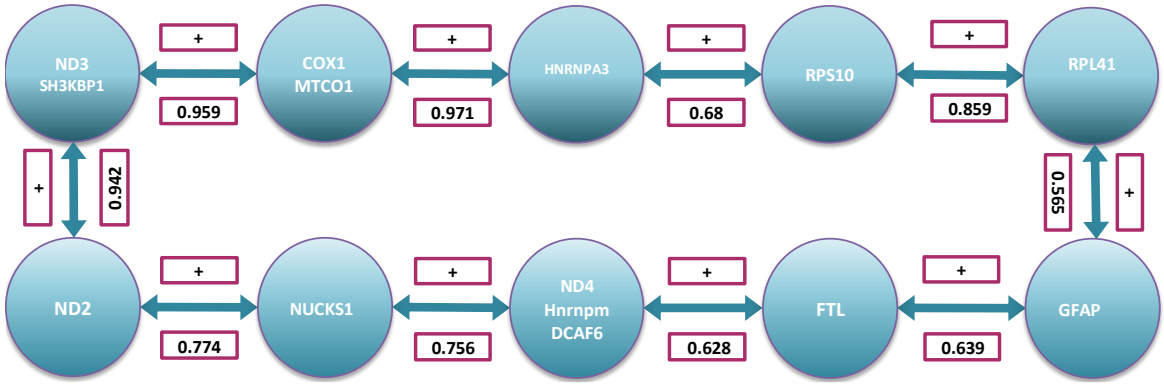
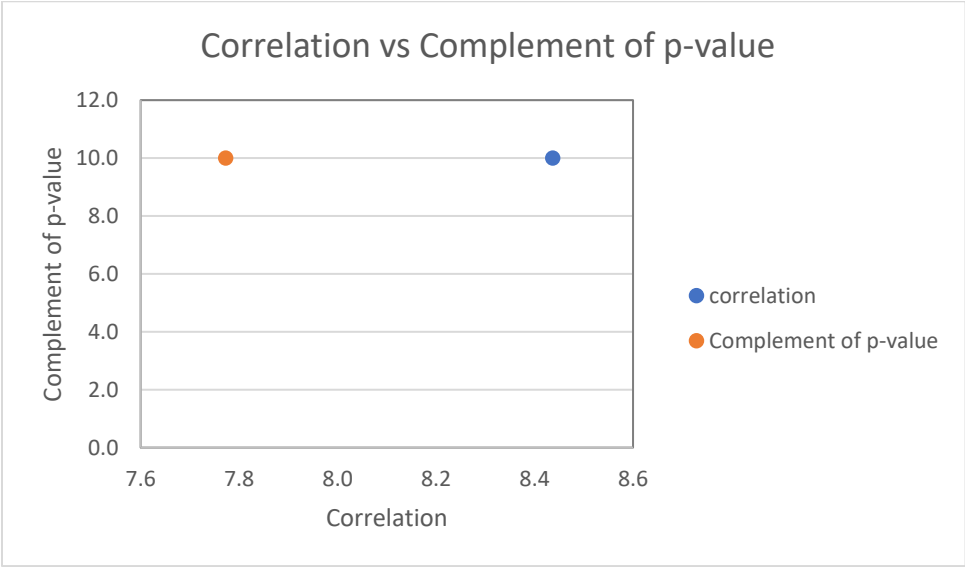


Figure 34: TSP of the complement of p-value pathway with corresponding correlation



Graph 26: Correlation between correlation and complement of p-value

5.4.2 Correlation VS Complement of P-value: Minimum Spanning Tree

Another method used for finding an optimal signaling paths between genes in our research group is the Minimum Spanning Tree (MST) [13]. MST provides "tree that maximizes the linear correlations"[13]. Figure 35 shows the MST obtained with the correlation matrix. This tree used all the genes. The similarities with figure 31 are the linear relationship between ND2 with COX1 and RPL41 with ND3. Figure 36 shows the MST obtained from the complement of p-values. In this case, the tree did not include all the genes. The similarities with figure 32 is the direct relationship between FTL and RPL4.

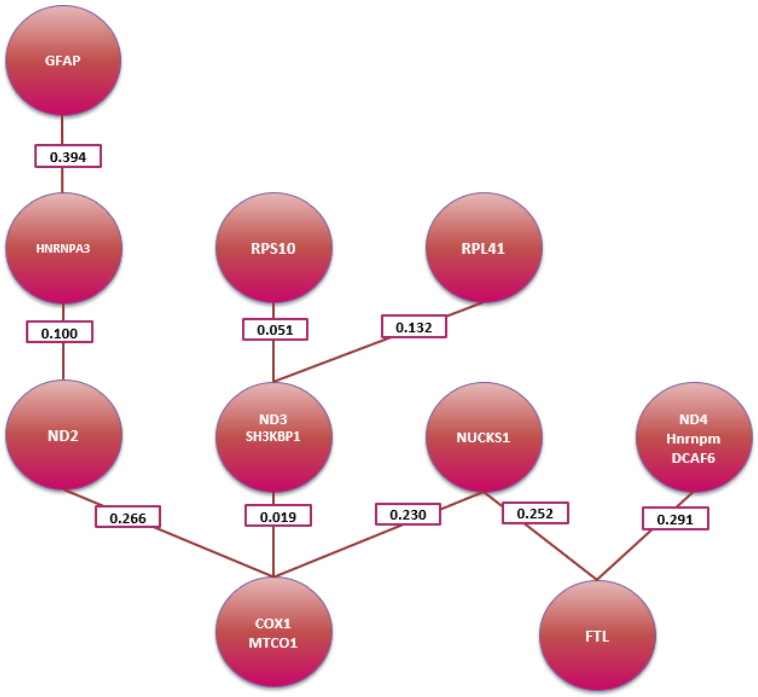


Figure 35: MST with correlation values

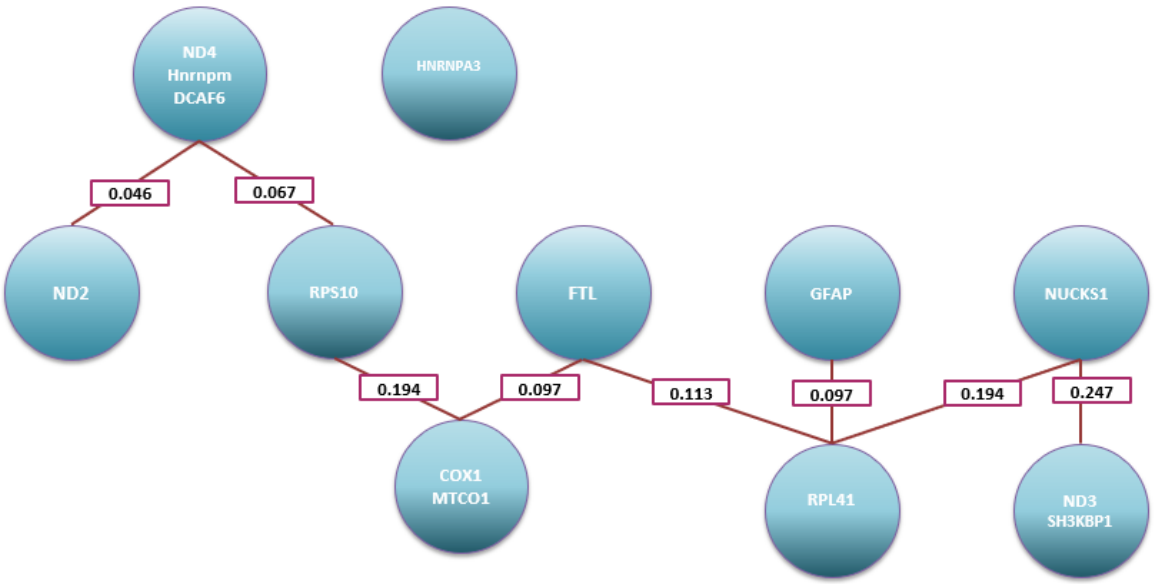


Figure 36: MST with complement of p-values

5.4.3 TSP & MST signaling paths vs GeneMANIA

In this section we compare TSP and MST with the GeneMania interface. GeneMANIA is described as: “A flexible user-friendly web interface for generating hypotheses about gene function, analyzing gene lists and prioritizing genes for functional assays.” [112]. This interface is characterized by obtaining data from databases such as BioGRID, GEO, IRefIndex and I2D when given a gene list and widens the amount of genes with similar functionality of properties [13] . In this case the gene list given is table 18 that was also used for TSP and MST methods.

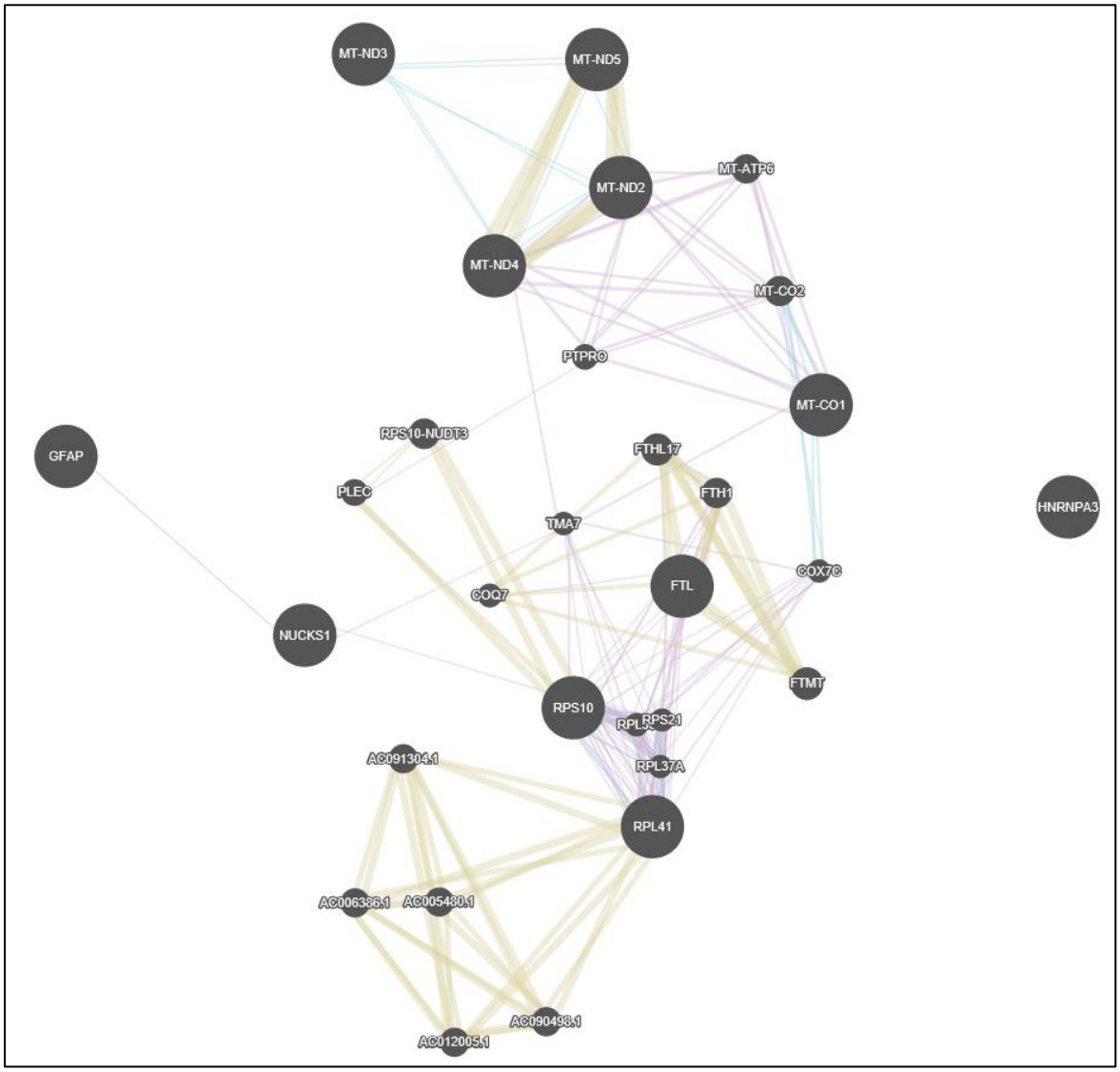


Figure 37: Results in Genemania

Figure 37 shows the overlapping networks found in GeneMANIA. These networks are called co-expression, shared proteins domains, pathways and co-localization. To construct these networks, geneMANIA includes additional genes from the original list. The default amount of additional genes is 20 but can be determined by the user in a range from 0 to 100 [13]. A total of 19 connections were found in the geneMANIA network. As can be seen most of the genes are connected overall except for HRNPA3. This a similarity found with the MST in figure 36.

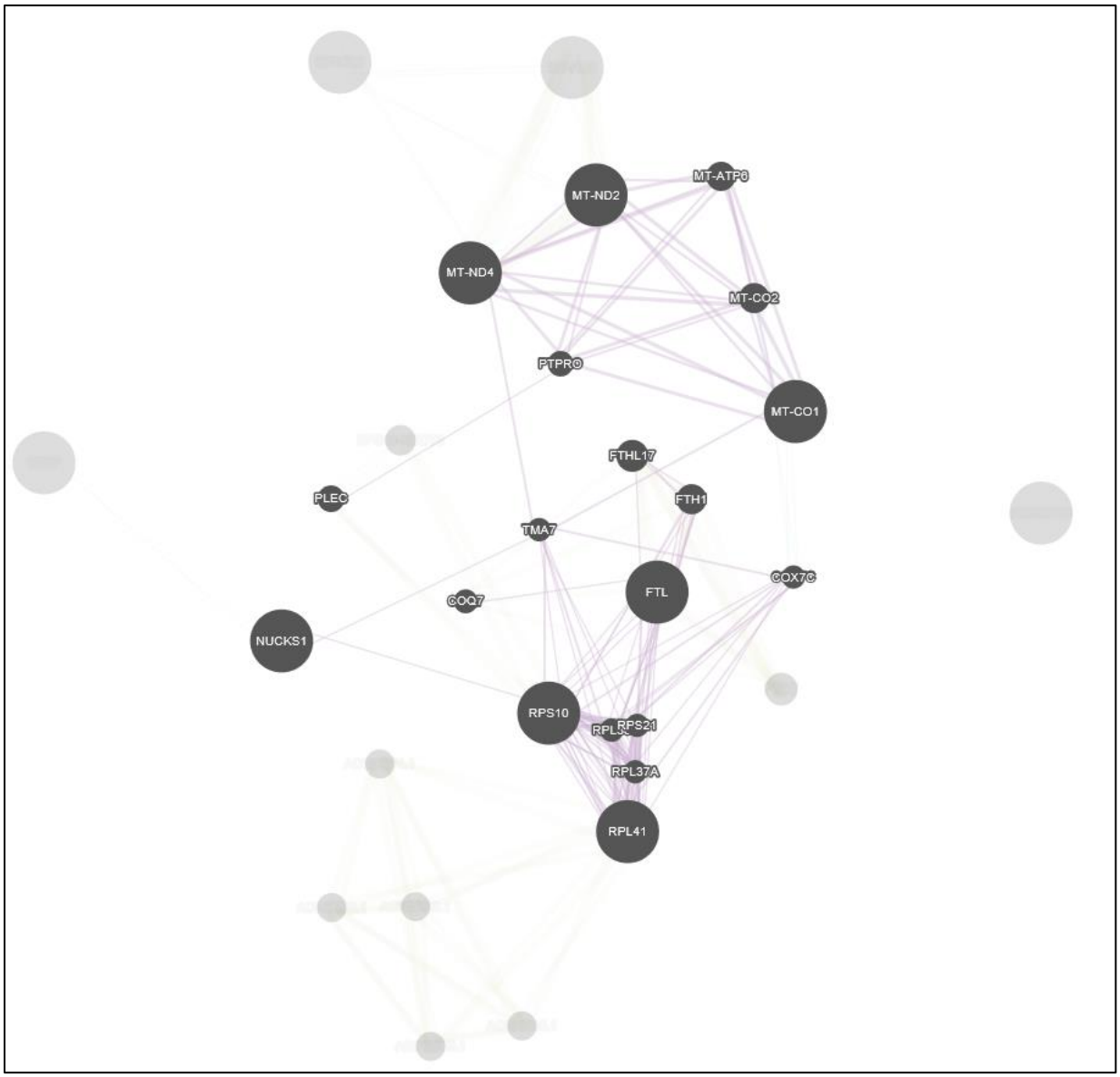


Figure 38: Co-expression results

The largest network is the co-expression shown in figure 38 with 13 connections. A total of 7 genes from the original list of 10 are connected in this network. This network identifies studies that link genes through co-expression from Gene Expression Omnibus (GEO) data series with GSE identifiers [112]. The connection is created if two genes have similar expression levels across conditions in publications of gene expression studies [113].

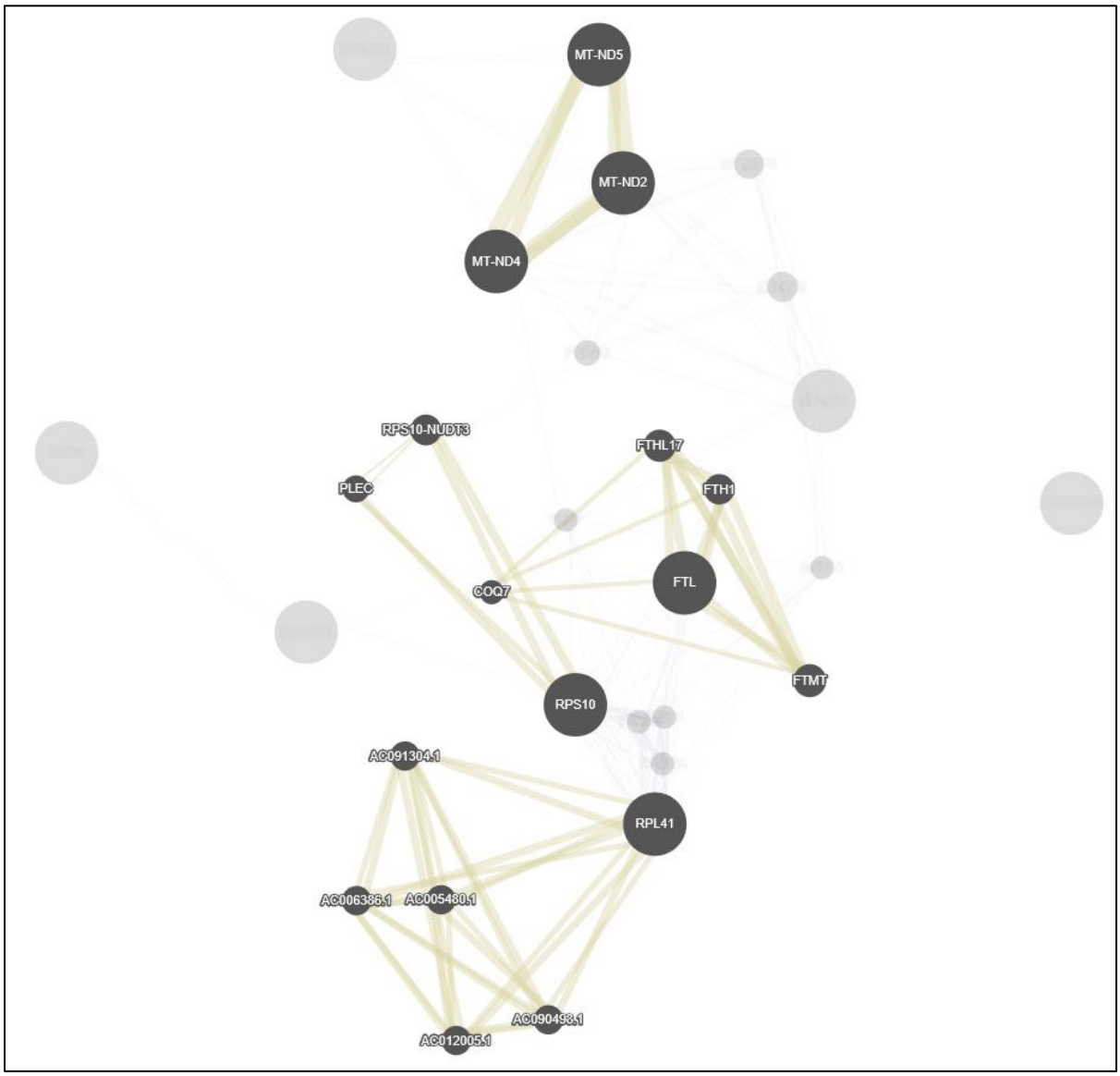


Figure 39: Shared protein domains

Figure 39 is the shared protein domains network. This network only contains two connections. Also, it only included 5 out of the 10 original genes from the query list. This network identifies studies that link genes through protein domains profiles that are transformed to shared protein domains from the InterPro, SMART and Pfam databases [112][113].

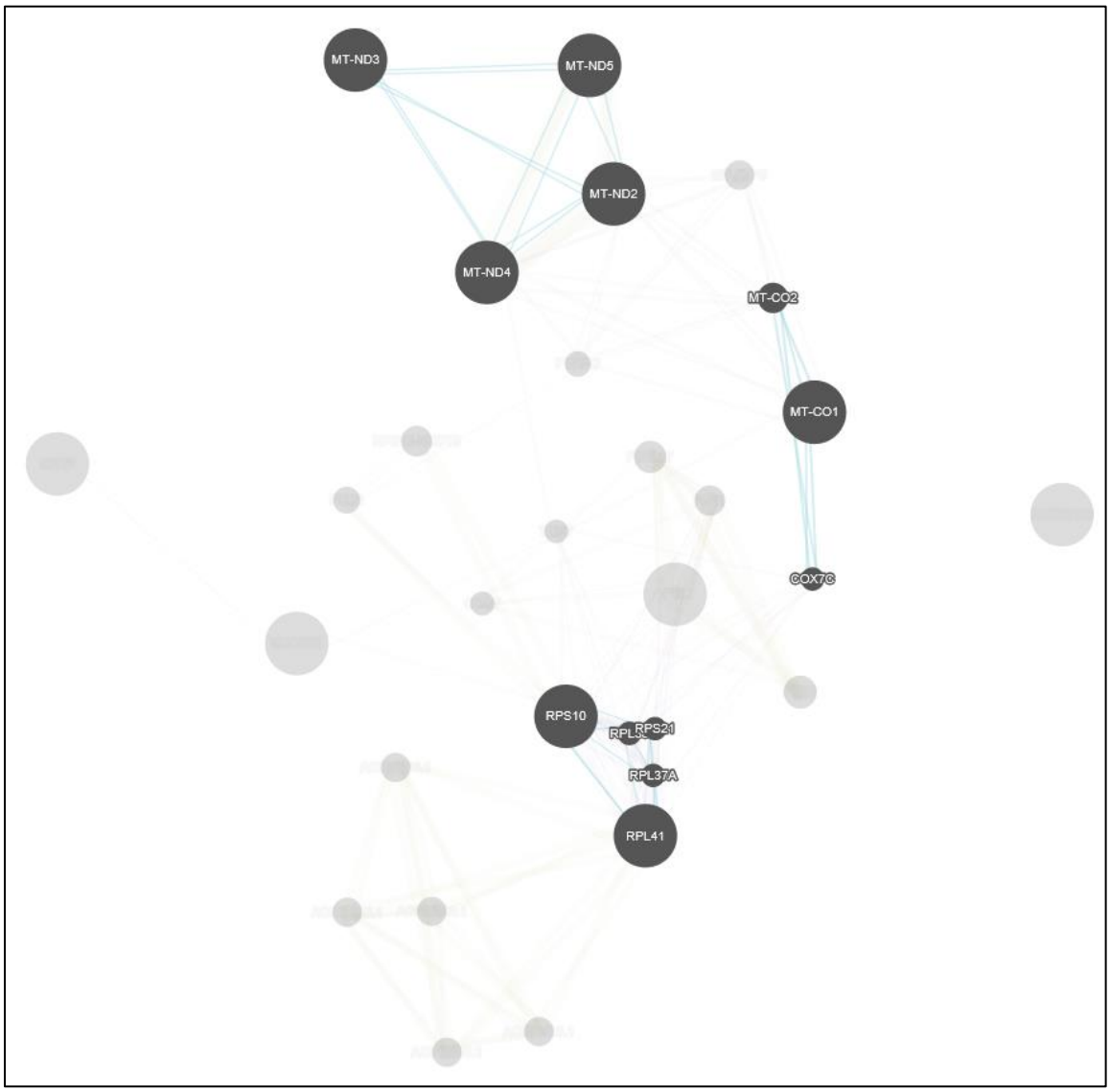


Figure 40: Pathway results

Figure 40 is the pathway domains network. This network only contains two connections. It includes 6 of the original genes from the query list. This network is one the newest ones in the

GeneMANIA interface. A connection is created when two genes respond the same part inside a pathway [113]. The databases used to create this links are BioCyc and Reactome [113].

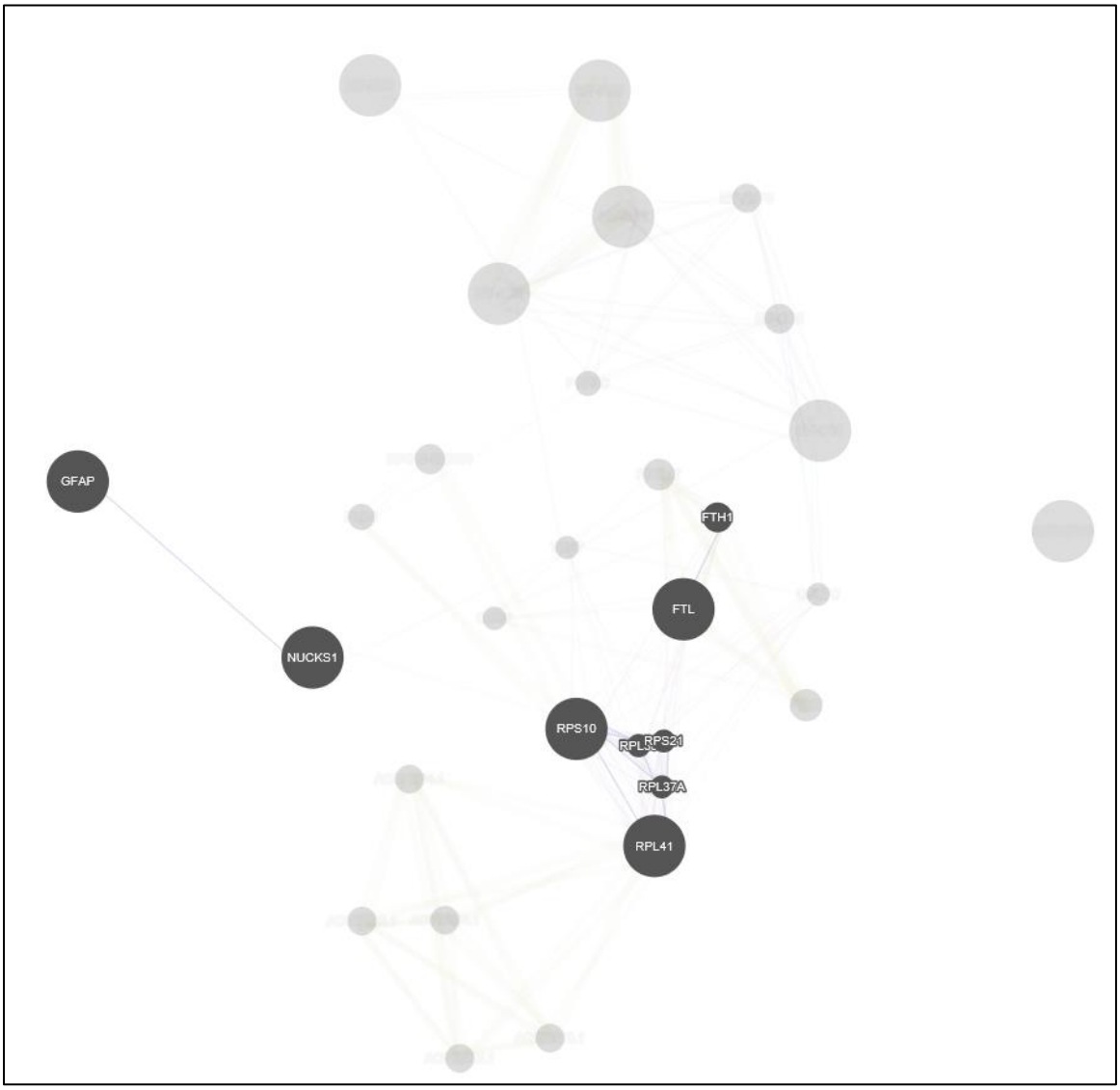


Figure 41: Co-localization results

The final network shown in figure 41 is co-localization. These results only have one connection. This network only included 5 out of the original 10 genes provided in the query list. A connection in this network is created when genes are: "both expressed in the same tissue or if their gene products are both identified in the same cellular location." [113].

All the networks provided by GeneMANIA were incapable in connecting all the original 10 genes from the query list and it is dependable of publications. The benefit of TSP and MST is that they work on obtaining a global solution. TSP for example was able to provide an optimal path with all the genes provided in both correlation and complement of p-value performance measurements. MST was also able to provide an optimal path with the provided genes with the correlation values.

5.5 Summary

The analysis pipeline designed by the AOG can be and has been applied to public databases of different illnesses [6], [9], [10]. This detail shows the versatility of the pipeline. In its interdisciplinary inception, it joins the biology and the industrial engineering fields through characterizing, modelling and optimizing the AD biosystem. These 3 steps are largely intertwined in this specific biosystem case. The effectiveness of the pipeline is shown from its initial characterization by applying MCO to the AD data to find 10 potential biomarker genes from over 50,000 genes. These 10 potential biomarker genes are used to further characterize, model and optimize through TSP. Some of these genes had not been linked yet to AD but have been to other illnesses. The optimal models found by TSP and MST using correlation and the complement of p-values are global solutions. These models are discussed and validated through biological evidence. The AOG pipeline is much stronger compared to other gene networking methods due to discovering new potential gene biomarkers and providing unbiased solutions.

Chapter 6. Conclusions and future work

The framework created in this work prioritizes the importance and the steps that need to be done for proper replication in researching different biosystems. These biosystems are connected through characterization, modelling and optimization as steps. The characterization of each biosystem is the most crucial step due to it providing us the guidance for appropriate modelling and optimization of each case. The initial understanding of the data will lead in the end to better decision making. A variety of mathematical tools such as graphical visualization, statistical inference and multiple criteria optimization were adjusted to each case.

For the first case, the USDA biofuel study, there was data from the initial experiment runs. These experiments were done only using the second level of each factor. The reference from Biotek [65] of the behavior for lipid fluorescence (amount of lipids) and absorbance (number of cells) was used to compare how the results must behave. By using graphical visualization, the discrepancies from the initial experiments are shown. This gives reason to ensure a proper design of experiments to ensure reproducibility. For the second case, USDA hairy vetch/corn study, a proper experiment was designed across five states in the east of the United States for various years. Since there were various years of data available these were taken as replications. Second order regressions with ANOVA with factors PD, SR and HD were tested for significance across states. A one-way ANOVA and a second order regression with ANOVA were done for corn/silage yield and hairy vetch seeding rate to test for significance. For the third case, Alzheimer's Disease, characterization, modelling and optimization are finely intertwined. The initial characterization is finding the genes that are differentially expressed. These genes give more information about the illness. Multiple criteria optimization was used to find those genes. The two measurements were mean and the

median. These measurements were chosen to their aim of centrality without bias. All these tools used how different biosystems can be initially understood, paving way for better decision making.

After characterizing each case, modeling was ensured. For the USDA biofuel study, a power analysis was done as the first step. The power analysis has the benefit that it can help save time and money by providing the minimum sample size to replicate the design of experiment. Later, a central composite design was elaborated including with a total of 19 runs. The sample size found in the power analysis was added to the central point of the design. A design of experiment also helps with the traceability of the results. Another aspect included in the modeling of the USDA biofuel case is that of graphical visualization, but in the linear aspect. These graphs were created to better illustrate the goals of this case to our collaborators. Two individual graphs were created. The first graph represents the absorbance vs time curved derived into two of the functions that are to be maximized: absorbance and absorbance slope. The second graph is the derivative of the lipid fluorescence/cells curve, which is the third function that is to be maximized. The modeling for the second case, USDA hairy vetch/corn study, was done using the coefficient values found in the characterization of the data provided and it is heavily intertwined with the optimization. Only the years and states that provided information for both hairy vetch and corn yield were used to compare. Using the coefficients, a graphical visualization was constructed. The use of second order regressions for this case brings the benefit of having functions with curvature. This comes in handy due to their convex shape. These functions are derived and solved for hairy vetch seeding rate, this provides the highest point of the graph that is in the hairy vetch seeding rate that provides the largest amount of hairy vetch biomass and corn/silage yield. Lastly, the Alzheimer's disease case, the modeling also goes in hand with the optimization. The travelling salesman problem was used

to find the path of maximum correlation between the genes found in the characterization of the disease.

The modeling step helps to naturally move to the optimization framework for each case. The optimization for the USDA biofuel case will be left as future work. This includes solver methods to find the design point of the experiment that provides the largest amount of yield and if that point is inside the proposed design. The best hairy vetch seeding rate found are the optimized values. The best hvsr for corn/silage and vetch were used to create an economic analysis where a comparison with the cost with nitrogen fertilizer was done. Since the main driver is to obtain the most amount of corn/silage yield at the lowest cost the economic analysis showed that there were large costs savings in using hairy vetch as a cover crop instead of a nitrogen fertilizer. In the future work for this chapter, this framework can be applied to other cover and cash crops. For the third case, a comparison of different performance measures, optimal signaling paths and another interface such as GeneMANIA. The dominant performance measure was the correlation path. As concluded in chapter 5, the framework developed in our research group is consistent in creating optimal signaling paths with no bias. The proposed method identifies genes already reported as relevant to inflammation and neurodegenerative diseases. Results also suggest that infections could be related to AD development as other reports have proposed [114]–[117]. Genes without previous report relevance in AD can be proposed for further biological validation as well as the gene expression connections that have not been explored yet.

List of papers, presentations, conferences and awards

PAPERS

- *"Characterization of Alzheimer's disease: An Operations Research Approach".* Yazeli E. Cruz-Rivera, Jaileene Perez-Morales.Yaritza M. Santiago, Valerie M. Gonzalez, Clara E. Isaza, Mauricio Cabrera-Rios. Journal paper for **Journal of Alzheimer's**, 2018. In revision.
- *"Characterization of Alzheimer's disease: An Operations Research Approach".* Yazeli E. Cruz-Rivera, Yaritza M. Santiago, Valerie M. Gonzalez, Clara E. Isaza, Mauricio Cabrera-Rios. Refereed proceeding for **IEOM 2015.**

REFEREED CONFERENCES AND PRESENTATIONS

- *"Biosystems Characterization, Modelling and Optimization".*Yazeli E. Cruz-Rivera, Nilvia Cuevas Feliciano, Adriana Cardona, Clara E. Isaza and Mauricio Cabrera-Ríos. **AAAS 2017**
- *"Biosystems Characterization, Modelling and Optimization".*Yazeli E. Cruz-Rivera, Nilvia Cuevas Feliciano, Adriana Cardona, Clara E. Isaza and Mauricio Cabrera-Ríos. **SACNAS 2016**
- *"On effective biofuel production from algae: Initial Ideas".*Yazeli E. Cruz-Rivera, Nilvia Cuevas Feliciano, Adriana Cardona, Clara E. Isaza and Mauricio Cabrera-Ríos. **Sigma Xi 2016**
- *"Characterization of Alzheimer's disease: An Operations Research Approach".* Yazeli E. Cruz-Rivera, Yaritza M. Santiago, Valerie M. Gonzalez, Clara E. Isaza, Mauricio Cabrera-Rios. **ERN 2016.**
- *"Characterization of Alzheimer's disease: An Operations Research Approach".* Yazeli E. Cruz-Rivera, Yaritza M. Santiago, Valerie M. Gonzalez, Clara E. Isaza, Mauricio Cabrera-Rios. **IEOM 2015.**
- *"Characterization of Alzheimer's disease: An Operations Research Approach".* Yazeli E. Cruz-Rivera, Yaritza M. Santiago, Valerie M. Gonzalez, Clara E. Isaza, Mauricio Cabrera-Rios. **ABRCMS 2014.**
- *"Graph-Models to Lead Genetic Signaling Path Discovery: Preliminary Ideas and Results*". Yazeli Cruz-Rivera, Enery Lorenzo, Nicole Ortiz, Clara Isaza, Mauricio Cabrera-Ríos. **BMES 2013.**
- *"Coordinated changes on relative genetic expression of potential lung cancer biomarkers".* Nicole Ortiz, Yazeli Cruz, Enery Lorenzo, Jesus Rodriguez, Clara Isaza, Mauricio Cabrera. **ISERC 2013.**

AWARDS

- AAAS 2017 Joshua Neimark Travel award recipient.
- SACNAS 2016 travel award.
- 2nd place in the undergraduate paper student competition in IEOM 2015 conference.
- FASEB MARC BMES 2013 travel award recipient.

References

[1] "NCBI BioSystems Database Overview." [Online]. Available: https://www.ncbi.nlm.nih.gov/Structure/biosystems/docs/biosystems_about.html. [Accessed: 20-Mar-2017].

[2] T. M. Mata, A. A. Martins, and N. S. Caetano, "Microalgae for biodiesel production and other applications: A review," *Renew. Sustain. Energy Rev.*, vol. 14, no. 1, pp. 217–232, 2010.

[3] X. Miao and Q. Wu, "Biodiesel production from heterotrophic microalgal oil," *Bioresour. Technol.*, vol. 97, no. 6, pp. 841–846, 2006.

[4] and J. T. S. S.B. Mirsky*1, V.J. Ackroyd1, S. Cordeau2, 3, W.S. Curran4, M. Hashemi5, S.C. Reberg-Horton6, M. Ryan3, "Effect of Seeding Rate and Date on Hairy Vetch (Vicia villosa Roth) Biomass Production across the Eastern United States." Agronomy Journal, 2017.

[5] "Cover Crop Guide." [Online]. Available: http://covercrops.cals.cornell.edu/index.php. [Accessed: 03-Feb-2017].

[6] E. Watts-Oquendo, M. Sánchez-Peña, C. E. Isaza, and M. Cabrera-Ríos, "Potential colon cancer biomarker search using more than two performance measures in a multiple criteria optimization approach.," *P. R. Health Sci. J.*, vol. 31, no. 2, pp. 59–63, Jun. 2012.

[7] J. Cooper-Knock, J. Kirby, L. Ferraiuolo, P. R. Heath, M. Rattray, and P. J. Shaw, "Gene expression profiling in human neurodegenerative disease.," *Nat. Rev. Neurol.*, vol. 8, no. 9, pp. 518–30, Sep. 2012.

[8] T. Dunckley, T. G. Beach, K. E. Ramsey, A. Grover, D. Mastroeni, D. G. Walker, B. J. LaFleur, K. D. Coon, K. M. Brown, R. Caselli, W. Kukull, R. Higdon, D. McKeel, J. C. Morris, C. Hulette, D. Schmechel, E. M. Reiman, J. Rogers, and D. A. Stephan, "Gene expression correlates of neurofibrillary tangles in Alzheimer's disease.," *Neurobiol. Aging*, vol. 27, no. 10, pp. 1359–71, Oct. 2006.

[9] M. L. Sánchez-Peña, C. E. Isaza, J. Pérez-Morales, C. Rodríguez-Padilla, J. M. Castro, and M. Cabrera-Ríos, "Identification of potential biomarkers from microarray experiments using multiple criteria optimization.," *Cancer Med.*, vol. 2, no. 2, pp. 253–65, Apr. 2013.

[10] E. Lorenzo, K. Camacho-Caceres, A. J. Ropelewski, J. Rosas, M. Ortiz-Mojer, L. Perez-Marty, J. Irizarry, V. Gonzalez, J. A. Rodríguez, M. Cabrera-Rios, and C. Isaza, "An Optimization-Driven Analysis Pipeline to Uncover Biomarkers and Signaling Paths: Cervix Cancer.," *Microarrays (Basel, Switzerland)*, vol. 4, no. 2, pp. 287–310, Jun. 2015.

[11] K. I. Camacho-Cáceres, J. C. Acevedo-Díaz, L. M. Pérez-Marty, M. Ortiz, J. Irizarry, M. Cabrera-Ríos, and C. E. Isaza, "Multiple criteria optimization joint analyses of microarray experiments in lung cancer: from existing microarray data to new knowledge.," *Cancer Med.*, vol. 4, no. 12, pp. 1884–900, Dec. 2015.

[12] G. Reinelt, "TSPLIB—A Traveling Salesman Problem Library," *ORSA J. Comput.*, vol. 3, no. 4, pp. 376–384, Nov. 1991.

[13] C. Isaza, J. Rosas, E. Lorenzo, A. Marrero, M. Ortiz-Mojer, L. Perez-Marty, and M. Cabrera-Ríos, "Biological Signaling Pathways and Potential Mathematical Network Representations: Biological Discovery through Optimization."

[14] K. Bollen, J. Cacioppo, R. Kaplan, J. A. Krosnick, and J. L. Olds, "Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science," *Rep. Subcomm. Replicability Sci. Advis. Comm. to Natl. Sci. Found. Dir. Soc. Behav. Econ. Sci.*, pp. 1–29, 2015.

[15] C. G. Begley and J. P. A. Ioannidis, "Reproducibility in science: Improving the standard for basic and preclinical research," *Circ. Res.*, vol. 116, no. 1, pp. 116–126, 2015.

[16] Open Science Collaboration, "Estimating the reproducibility of psychological science," *Science (80-.).*, vol. 349, no. 6251, p. aac4716-aac4716, 2015.

[17] T. M. Errington, E. Iorns, W. Gunn, F. E. Tan, J. Lomax, B. A. Nosek, K. Gilbert, J. Moore, S. Renaut, D. Rennison, D. Laitin, T. Madon, L. Nelson, B. Nosek, M. Petersen, R. Sedlmayr, J. Simmons, U. Simonsohn, M. Van der Laan, J. Huguenard, K. Kelner, W. Koroshetz, D. Krainc, S. Lazic, M. Levine, M. Macleod, J. McCall, R. Moxley, K.

Narasimhan, L. Nobel, S. Perrin, J. Porter, O. Steward, E. Unger, U. Utz, and S. Silberberg, "An open investigation of the reproducibility of cancer biology research," *Elife*, vol. 3, pp. 726–728, Dec. 2014.

[18] "Biofuels - Biofuel Information - Guide to Biofuels." [Online]. Available: http://biofuel.org.uk/. [Accessed: 15-Mar-2017].

[19] "USDA ERS - U.S. Bioenergy Statistics." [Online]. Available: https://www.ers.usda.gov/data-products/us-bioenergy-statistics.aspx. [Accessed: 15-Mar-2017].

[20] J. Hoogeveen, J.-M. Faurès, and N. van de Giessen, "Increased biofuel production in the coming decade: to what extent will it affect global freshwater resources?," *Irrig. Drain.*, vol. 58, no. S1, pp. S148–S160, Feb. 2009.

[21] "Bioenergy (Biofuels and Biomass) | EESI." [Online]. Available: http://www.eesi.org/topics/bioenergy-biofuels-biomass/description. [Accessed: 15-Mar-2017].

[22] P. He, S. Xu, H. Zhang, S. Wen, Y. Dai, S. Lin, and C. Yarish, "Bioremediation efficiency in the removal of dissolved inorganic nutrients by the red seaweed, Porphyra yezoensis, cultivated in the open sea," *Water Res.*, vol. 42, no. 4–5, pp. 1281–1289, 2008.

[23] I. Douskova, J. Doucha, K. Livansky, J. MacHat, P. Novak, D. Umysova, V. Zachleder, and M. Vitova, "Simultaneous flue gas bioremediation and reduction of microalgal biomass production costs," *Appl. Microbiol. Biotechnol.*, vol. 82, no. 1, pp. 179–185, 2009.

[24] Nrel, "A look back at the U. S. Department of Energy's aquatic species program: biodiesel from algae," *Report*, vol. 328, p. 291 p, 1998.

[25] L. Rodolfi, G. C. Zittelli, N. Bassi, G. Padovani, N. Biondi, G. Bonini, and M. R. Tredici, "Microalgae for oil: Strain selection, induction of lipid synthesis and outdoor mass cultivation in a low-cost photobioreactor," *Biotechnol. Bioeng.*, vol. 102, no. 1, pp. 100–112, 2009.

[26] G. C. Dismukes, D. Carrieri, N. Bennette, G. M. Ananyev, and M. C. Posewitz, "Aquatic phototrophs: efficient alternatives to land-based crops for biofuels," *Curr. Opin.*

Biotechnol., vol. 19, no. 3, pp. 235–240, 2008.

[27] J. N. Rosenberg, G. A. Oyler, L. Wilkinson, and M. J. Betenbaugh, "A green light for engineered algae: redirecting metabolism to fuel a biotechnology revolution," *Curr. Opin. Biotechnol.*, vol. 19, no. 5, pp. 430–436, 2008.

[28] L. A. Zaslavskaia, J. C. Lippmeier, C. Shih, D. Ehrhardt, A. R. Grossman, and K. E. Apt, "Trophic Conversion of an Obligate Photoautotrophic Organism Through Metabolic Engineering," *Science (80-.).*, vol. 292, no. 5524, pp. 2073–2075, 2001.

[29] M. Hannon, J. Gimpel, M. Tran, B. Rasala, and S. Mayfield, "Biofuels from algae: challenges and potential.," *Biofuels*, vol. 1, no. 5, pp. 763–784, 2010.

[30] "Diseño de Foto-Bioreactores para el Cultivo Micro Algas Oleaginosas Parte 2. Bioproceso y Especifidades « Biotecnología Práctica." [Online]. Available: https://bioreactorcrc.wordpress.com/2011/05/21/diseo-de-foto-bioreactores-para-el-cultivo-micro-algas-oleaginosas-parte-2-bioproceso-y-especifidades/. [Accessed: 16-Mar-2017].

[31] J. J. Milledge and S. Heaven, "A review of the harvesting of micro-algae for biofuel production," *Rev. Environ. Sci. Biotechnol.*, vol. 12, no. 2, pp. 165–178, 2013.

[32] B. Kumara Behera and A. Varma, "Microbial Resources for Sustainable Energy," no. June, pp. 1–27, 2016.

[33] "Design and Analysis of Experiments - Douglas C. Montgomery - Google Books." [Online]. Available: https://books.google.com.pr/books/about/Design_and_Analysis_of_Experiments.html?id=kMMJAm5bD34C&redir_esc=y. [Accessed: 17-Mar-2017].

[34] P. C. Hallenbeck, M. Grogger, M. Mraz, and D. Veverka, "Bioresource Technology The use of Design of Experiments and Response Surface Methodology to optimize biomass and lipid production by the oleaginous marine green alga , Nannochloropsis gaditana in response to light intensity , inoculum size and CO 2," *Bioresour. Technol.*, vol. 184, pp. 161–168, 2015.

[35] L. Wei, X. Huang, Z. Huang, and Z. Zhou, "Orthogonal test design for optimization of lipid

accumulation and lipid property in Nannochloropsis oculata for biodiesel production," *Bioresour. Technol.*, vol. 147, pp. 534–538, 2013.

[36] B. Fm and B. Tb, "A Critical Review of Algae Process Optimization Using Design of Experiment Methodologies," vol. 1, pp. 1–8, 2013.

[37] A. Massart, É. Aubry, and A. L. Hantson, "Optimization of the medium composition of the microalga 'Dunaliella Tertiolecta Butcher' in order to combine high cell density and accumulation of lipids for biodiesel production," *Biotechnol. Agron. Soc. Environ.*, vol. 14, no. SPEC. ISSUE 2, pp. 567–572, 2010.

[38] "Cover Crops." [Online]. Available: http://www.mda.state.mn.us/protecting/conservation/practices/covercrops.aspx. [Accessed: 26-Feb-2017].

[39] "Legume Cover Crops." [Online]. Available: http://www.sare.org/Learning-Center/Books/Managing-Cover-Crops-Profitably-3rd-Edition/Text-Version/Legume-Cover-Crops. [Accessed: 26-Feb-2017].

[40] "nitrogencycle5 (752×600)." [Online]. Available: https://www.quia.com/files/quia/users/repasy_p/nutrientcycles/nitrogencycle5. [Accessed: 28-Mar-2017].

[41] "Understanding nitrogen in soils : Nitrogen : Nutrient Management : Agriculture : University of Minnesota Extension." [Online]. Available: http://www.extension.umn.edu/agriculture/nutrient-management/nitrogen/understanding-nitrogen-in-soils/. [Accessed: 26-Feb-2017].

[42] A. M. Decker, A. J. Clark, J. J. Meisinger, F. R. Mulford, and M. S. Mclntosh, "Legume Cover Crop Contributions to No-Tillage Corn Production," 1990.

[43] J. F. Holderbaum, A. M. Decker, J. J. Meisinger, F. R. Mulford, and L. R. Vough, "(1990) Fall-Seeded Legume Cover Crops for No-Tillage Corn in the Humid East (AJ)," no. i, 1982.

[44] S. K. E. J. Jellum, "Long-term winter cover cropping effects on corn (Zea mays L .) production and soil nitrogen availability," no. 9904, pp. 470–477, 2000.

[45] A. J. Clark, A. M. Decker, J. J. Meisinger, and M. S. Mcintosh, "(1997) Kill Date of Vetch, Rye, and a Vetch-Rye Mixture: I. Cover Crop and Corn Nitrogen (AJ)," 1995.

[46] T. A. Larue and T. G. Patterson, "How Much Nitrogen do Legumes Fix?," *Adv. Agron.*, vol. 34, no. C, pp. 15–38, 1981.

[47] Q. M. Ketterings, S. N. Swink, S. W. Duiker, K. J. Czymmek, D. B. Beegle, and W. J. Cox, "Integrating Cover Crops for Nitrogen Management in Corn Systems on Northeastern U . S . Dairies," pp. 1365–1376, 2015.

[48] J. W. Doran and M. S. Smith, "Role of cover crops in nitrogen cycling," *Cover Crop. clean water Proc. an Int. Conf. West Tennessee Exp. Station. April 9-11, 1991, Jackson, Tennessee.*, no. c, pp. 85–90, 1991.

[49] Z. Dou, "The contribution of nitrogen from legume cover crops double-cropped with winter wheat to tilled and non-tilled . maize," *Eur. J. Agron.*, vol. 3, no. 2, pp. 93–100, 1994.

[50] M. Sarrantonio and T. W. Scott, "Tillage Effects on Availability of Nitrogen to Corn Following a Winter Green Manure Crop," vol. 1668, pp. 1661–1668, 1988.

[51] J. K. Stute and J. L. Posner, "Legume cover crops as a nitrogen source for corn in an oat-corn rotation.pdf." pp. 385–390, 1995.

[52] Z. Dou, R. H. Fox, and J. D. Toth, "Seasonal Soil Nitrate Dynamics in Corn as Affected by Tillage and Nitrogen Source," vol. 864, pp. 858–864, 1995.

[53] A. J. Clark, J. J. Meisinger, A. M. Decker, and F. R. Mulford, "Effects of a Grass-Selective Herbicide in a Vetch–Rye Cover Crop System on Nitrogen Management," pp. 36–42, 2007.

[54] J. T. Spargo, M. A. Cavigelli, S. B. Mirsky, J. J. Meisinger, and V. J. Ackroyd, "Organic supplemental nitrogen sources for field corn production after a hairy vetch cover crop," *Agron. J.*, vol. 108, no. 5, pp. 1992–2002, 2016.

[55] M. G. Wagger, "(1989) Cover Crop Management and Nitrogen Rate in Relation to Growth and Yield of No-till Corn (AJ)," vol. 538, no. 1, pp. 533–538, 1989.

[56] Q. M. Ketterings, S. N. Swink, S. W. Duiker, K. J. Czymmek, D. B. Beegle, and B. Cox, "Nitrogen Benefits of Winter Cover Crops," 2008.

[57] R. L. Blevins, J. H. Herbek, and W. W. Frye, “Legume cover crops as a nitrogen source for no-till corn and grain sorghum.pdf.” 1990.

[58] S. A. Ebelhar, W. W. Frye, and R. L. Blevins, “Nitrogen from legume cover crops for no-tillage corn.pdf.” 1984.

[59] “Feed Grains Custom Query.”

[60] “USDA ERS - USDA Agricultural Projections to 2026.” [Online]. Available: https://www.ers.usda.gov/publications/pub-details/?pubid=82538. [Accessed: 08-Apr-2017].

[61] “Hairy Vetch | SmartStore.” [Online]. Available: https://smartstore.greencoverseed.com/productdisplay/hairy-vetch. [Accessed: 14-Mar-2017].

[62] “USDA/NASS QuickStats Ad-hoc Query Tool.” [Online]. Available: https://quickstats.nass.usda.gov/results/61A6DC9A-75AD- 3680-9365- 170A6BDFF471. [Accessed: 14-Mar-2017].

[63] “Alzheimer’s Disease Fact Sheet | National Institute on Aging.” [Online]. Available: https://www.nia.nih.gov/alzheimers/publication/alzheimers-disease-fact-sheet. [Accessed: 17-Mar-2017].

[64] “Alzheimer’s Statistics.” [Online]. Available: http://www.alzheimers.net/resources/alzheimers-statistics/. [Accessed: 17-Mar-2017].

[65] “Application Note: Monitoring of Algal Growth Using Their Intrinsic Properties.” [Online]. Available: https://www.biotek.com/resources/application-notes/monitoring-of-algal-growth-using-their-intrinsic-properties/. [Accessed: 28-Jun-2017].

[66] “Minitab.” [Online]. Available: https://www.minitab.com/es-mx/. [Accessed: 14-Mar-2017].

[67] G. C. R. Douglas C. Montgomery, *Applied Statistics and Probability for Engineers*. 2010.

[68] J. B. O. R. K. Ahuja, T. L. Magnanti, *Network Flows: Theory, Algorithms, and Applications*. 1993.

[69] “MATLAB and Statistics Toolbox Release.” The MathWorks, Inc., Natick, Massachusetts, United States.

[70] F. H. Lin, R. Lin, H. M. Wisniewski, Y. W. Hwang, I. Grundke-Iqbal, G. Healy-Louie, and K. Iqbal, “Detection of point mutations in codon 331 of mitochondrial NADH dehydrogenase subunit 2 in Alzheimer’s brains.,” *Biochem. Biophys. Res. Commun.*, vol. 182, no. 1, pp. 238–46, Jan. 1992.

[71] V. Petruzzella, X. Chen, and E. A. Schon, “Is a point mutation in the mitochondrial ND2 gene associated with Alzheimer’s disease.,” *Biochem. Biophys. Res. Commun.*, vol. 186, no. 1, pp. 491–7, Jul. 1992.

[72] J. L. Burman, L. S. Itsara, E.-B. Kayser, W. Suthammarak, A. M. Wang, M. Kaeberlein, M. M. Sedensky, P. G. Morgan, and L. J. Pallanck, “A Drosophila model of mitochondrial disease caused by a complex I mutation that uncouples proton pumping from electron transfer.,” *Dis. Model. Mech.*, vol. 7, no. 10, pp. 1165–74, Oct. 2014.

[73] R. H. Swerdlow, “Mitochondria and cell bioenergetics: increasingly recognized components and a possible etiologic cause of Alzheimer’s disease.,” *Antioxid. Redox Signal.*, vol. 16, no. 12, pp. 1434–55, Jun. 2012.

[74] V. N. Anisimov, I. G. Popovich, M. A. Zabezhinski, S. V Anisimov, G. M. Vesnushkin, and I. A. Vinogradova, “Melatonin as antioxidant, geroprotector and anticarcinogen.,” *Biochim. Biophys. Acta*, vol. 1757, no. 5–6, pp. 573–89, Jan. .

[75] M. Y. Aksenov, H. M. Tucker, P. Nair, M. V Aksenova, D. A. Butterfield, S. Estus, and W. R. Markesbery, “The expression of several mitochondrial and nuclear genes encoding the subunits of electron transport chain enzyme complexes, cytochrome c oxidase, and NADH dehydrogenase, in different brain regions in Alzheimer’s disease.,” *Neurochem. Res.*, vol. 24, no. 6, pp. 767–74, Jun. 1999.

[76] “MT-ND4 Gene - GeneCards | NU4M Protein | NU4M Antibody.” [Online]. Available: http://www.genecards.org/cgi-bin/carddisp.pl?gene=MT-ND4&keywords=nd4. [Accessed: 26-Jan-2016].

[77] P.-H. Chen, Y.-P. Tsao, C.-C. Wang, and S.-L. Chen, “Nuclear receptor interaction protein,

a coactivator of androgen receptors (AR), is regulated by AR and Sp1 to feed forward and activate its own gene expression through AR protein stability.," *Nucleic Acids Res.*, vol. 36, no. 1, pp. 51–66, Jan. 2008.

[78] R. Fukuyama, K. Hatanpää, S. I. Rapoport, and K. Chandrasekaran, "Gene expression of ND4, a subunit of complex I of oxidative phosphorylation in mitochondria, is decreased in temporal cortex of brains of Alzheimer's disease patients.," *Brain Res.*, vol. 713, no. 1–2, pp. 290–3, Mar. 1996.

[79] L. Martorell, T. Segués, G. Folch, J. Valero, J. Joven, A. Labad, and E. Vilella, "New variants in the mitochondrial genomes of schizophrenic patients," *Eur. J. Hum. Genet.*, vol. 14, no. 5, pp. 520–528, Mar. 2006.

[80] P. M. Keeney and J. P. Bennett, "ALS spinal neurons show varied and reduced mtDNA gene copy numbers and increased mtDNA gene deletions.," *Mol. Neurodegener.*, vol. 5, p. 21, Jan. 2010.

[81] "MT-ND3 Gene - GeneCards | NU3M Protein | NU3M Antibody." [Online]. Available: http://www.genecards.org/cgi-bin/carddisp.pl?gene=MT-ND3&keywords=nd3. [Accessed: 26-Jan-2016].

[82] M. E. Munguia, T. Govezensky, R. Martinez, K. Manoutcharian, and G. Gevorkian, "Identification of amyloid-beta 1-42 binding protein fragments by screening of a human brain cDNA library.," *Neurosci. Lett.*, vol. 397, no. 1–2, pp. 79–82, Jan. .

[83] M. Ueno, H. Nakayama, S. Kajikawa, K. Katayama, K. Suzuki, and K. Doi, "Expression of ribosomal protein L4 (rpL4) during neurogenesis and 5-azacytidine (5AzC)-induced apoptotic process in the rat.," *Histol. Histopathol.*, vol. 17, no. 3, pp. 789–98, Jan. 2002.

[84] Y. Chen, Z. Lu, L. Zhang, L. Gao, N. Wang, X. Gao, Y. Wang, K. Li, Y. Gao, H. Cui, H. Gao, C. Liu, Y. Zhang, X. Qi, and X. Wang, "Ribosomal protein L4 interacts with viral protein VP3 and regulates the replication of infectious bursal disease virus.," *Virus Res.*, vol. 211, pp. 73–8, Jan. 2016.

[85] L. Green, B. Houck-Loomis, A. Yueh, and S. P. Goff, "Large ribosomal protein 4 increases efficiency of viral recoding sequences.," *J. Virol.*, vol. 86, no. 17, pp. 8949–58, Sep. 2012.

[86] A. Egoh, S. Nosuke Kanesashi, C. Kanei-Ishii, T. Nomura, and S. Ishii, “Ribosomal protein L4 positively regulates activity of a c-myb proto-oncogene product.,” *Genes Cells*, vol. 15, no. 8, pp. 829–41, Aug. 2010.

[87] A. Wang, S. Xu, X. Zhang, J. He, D. Yan, Z. Yang, and S. Xiao, “Ribosomal protein RPL41 induces rapid degradation of ATF4, a transcription factor critical for tumour cell survival in stress.,” *J. Pathol.*, vol. 225, no. 2, pp. 285–92, Oct. 2011.

[88] S. M. Fayaz and G. K. Rajanikant, “ATF4: the perpetrator in axonal-mediated neurodegeneration in Alzheimer’s disease.,” *CNS Neurol. Disord. Drug Targets*, vol. 13, no. 9, pp. 1483–4, Jan. 2014.

[89] J. Baleriola, C. A. Walker, Y. Y. Jean, J. F. Crary, C. M. Troy, P. L. Nagy, and U. Hengst, “Axonally synthesized ATF4 transmits a neurodegenerative signal across brain regions.,” *Cell*, vol. 158, no. 5, pp. 1159–72, Aug. 2014.

[90] Â. C. Crespo, B. Silva, L. Marques, E. Marcelino, C. Maruta, S. Costa, A. Timóteo, A. Vilares, F. S. Couto, P. Faustino, A. P. Correia, A. Verdelho, G. Porto, M. Guerreiro, A. Herrero, C. Costa, A. de Mendonça, L. Costa, and M. Martins, “Genetic and biochemical markers in patients with Alzheimer’s disease support a concerted systemic iron homeostasis dysregulation.,” *Neurobiol. Aging*, vol. 35, no. 4, pp. 777–85, Apr. 2014.

[91] X. Li, Y. Liu, Q. Zheng, G. Yao, P. Cheng, G. Bu, H. Xu, and Y. Zhang, “Ferritin light chain interacts with PEN-2 and affects γ-secretase activity.,” *Neurosci. Lett.*, vol. 548, pp. 90–4, Aug. 2013.

[92] P. Maciel, V. T. Cruz, M. Constante, I. Iniesta, M. C. Costa, S. Gallati, N. Sousa, J. Sequeiros, P. Coutinho, and M. M. Santos, “Neuroferritinopathy: missense mutation in FTL causing early-onset bilateral pallidal involvement.,” *Neurology*, vol. 65, no. 4, pp. 603–5, Aug. 2005.

[93] A. G. Barbeito, H. J. Garringer, M. A. Baraibar, X. Gao, M. Arredondo, M. T. Núñez, M. A. Smith, B. Ghetti, and R. Vidal, “Abnormal iron metabolism and oxidative stress in mice expressing a mutant form of the ferritin light polypeptide gene.,” *J. Neurochem.*, vol. 109, no. 4, pp. 1067–78, May 2009.

[94] R. Vidal, L. Miravalle, X. Gao, A. G. Barbeito, M. A. Baraibar, S. K. Hekmatyar, M. Widel, N. Bansal, M. B. Delisle, and B. Ghetti, "Expression of a mutant form of the ferritin light chain gene induces neurodegeneration and iron overload in transgenic mice.," *J. Neurosci.*, vol. 28, no. 1, pp. 60–7, Jan. 2008.

[95] M. A. Baraibar, A. G. Barbeito, B. B. Muhoberac, and R. Vidal, "A mutant light-chain ferritin that causes neurodegeneration has enhanced propensity toward oxidative damage.," *Free Radic. Biol. Med.*, vol. 52, no. 9, pp. 1692–7, May 2012.

[96] L. Ben Haim, M.-A. Carrillo-de Sauvage, K. Ceyzériat, and C. Escartin, "Elusive roles for reactive astrocytes in neurodegenerative diseases.," *Front. Cell. Neurosci.*, vol. 9, p. 278, Jan. 2015.

[97] Y.-S. Chen, S.-C. Lim, M.-H. Chen, R. A. Quinlan, and M.-D. Perng, "Alexander disease causing mutations in the C-terminal domain of GFAP are deleterious both to assembly and network formation with the potential to both activate caspase 3 and decrease cell viability.," *Exp. Cell Res.*, vol. 317, no. 16, pp. 2252–66, Oct. 2011.

[98] D. Martins-de-Souza, W. F. Gattaz, A. Schmitt, C. Rewerts, S. Marangoni, J. C. Novello, G. Maccarrone, C. W. Turck, and E. Dias-Neto, "Alterations in oligodendrocyte proteins, calcium homeostasis and new potential markers in schizophrenia anterior temporal lobe are revealed by shotgun proteome analysis.," *J. Neural Transm.*, vol. 116, no. 3, pp. 275–89, Mar. 2009.

[99] J. M. Frigerio, J. C. Dagorn, and J. L. Iovanna, "Cloning, sequencing and expression of the L5, L21, L27a, L28, S5, S9, S10 and S29 human ribosomal protein mRNAs.," *Biochim. Biophys. Acta*, vol. 1262, no. 1, pp. 64–8, May 1995.

[100] W. Abbas, I. Dichamp, and G. Herbein, "The HIV-1 Nef protein interacts with two components of the 40S small ribosomal subunit, the RPS10 protein and the 18S rRNA.," *Virol. J.*, vol. 9, p. 103, Jan. 2012.

[101] L. Doherty, M. R. Sheen, A. Vlachos, V. Choesmel, M.-F. O'Donohue, C. Clinton, H. E. Schneider, C. A. Sieff, P. E. Newburger, S. E. Ball, E. Niewiadomska, M. Matysiak, B. Glader, R. J. Arceci, J. E. Farrar, E. Atsidaftos, J. M. Lipton, P.-E. Gleizes, and H. T. Gazda,

"Ribosomal protein genes RPS10 and RPS26 are commonly mutated in Diamond-Blackfan anemia.," *Am. J. Hum. Genet.*, vol. 86, no. 2, pp. 222–8, Feb. 2010.

[102] M. Fittschen, I. Lastres-Becker, M. V Halbach, E. Damrath, S. Gispert, M. Azizov, M. Walter, S. Müller, and G. Auburger, "Genetic ablation of ataxin-2 increases several global translation factors in their transcript abundance but decreases translation rate.," *Neurogenetics*, vol. 16, no. 3, pp. 181–92, Jul. 2015.

[103] A. C. Parplys, W. Zhao, N. Sharma, T. Groesser, F. Liang, D. G. Maranon, S. G. Leung, K. Grundt, E. Dray, R. Idate, A. C. Østvold, D. Schild, P. Sung, and C. Wiese, "NUCKS1 is a novel RAD51AP1 paralog important for homologous recombination and genome stability.," *Nucleic Acids Res.*, vol. 43, no. 20, pp. 9817–34, Nov. 2015.

[104] H.-Y. Kim, B.-S. Choi, S. S. Kim, T.-Y. Roh, J. Park, and C.-H. Yoon, "NUCKS1, a novel Tat coactivator, plays a crucial role in HIV-1 replication by increasing Tat-mediated viral transcription on the HIV-1 LTR promoter.," *Retrovirology*, vol. 11, p. 67, Jan. 2014.

[105] L. Gu, B. Xia, L. Zhong, Y. Ma, L. Liu, L. Yang, and G. Lou, "NUCKS1 overexpression is a novel biomarker for recurrence-free survival in cervical squamous cell carcinoma.," *Tumour Biol.*, vol. 35, no. 8, pp. 7831–6, Aug. 2014.

[106] A. Kikuchi, T. Ishikawa, K. Mogushi, M. Ishiguro, S. Iida, H. Mizushima, H. Uetake, H. Tanaka, and K. Sugihara, "Identification of NUCKS1 as a colorectal cancer prognostic marker through integrated expression and copy number analysis.," *Int. J. Cancer*, vol. 132, no. 10, pp. 2295–302, May 2013.

[107] X. Liu, R. Cheng, M. Verbitsky, S. Kisselev, A. Browne, H. Mejia-Sanatana, E. D. Louis, L. J. Cote, H. Andrews, C. Waters, B. Ford, S. Frucht, S. Fahn, K. Marder, L. N. Clark, and J. H. Lee, "Genome-wide association study identifies candidate genes for Parkinson's disease in an Ashkenazi Jewish population.," *BMC Med. Genet.*, vol. 12, p. 104, Jan. 2011.

[108] J. Savitz, M. B. Frank, T. Victor, M. Bebak, J. H. Marino, P. S. F. Bellgowan, B. A. McKinney, J. Bodurka, T. Kent Teague, and W. C. Drevets, "Inflammation and neurological disease-related genes are differentially expressed in depressed patients with mood disorders and correlate with morphometric and functional imaging abnormalities.,"

Brain. Behav. Immun., vol. 31, pp. 161–71, Jul. 2013.

[109] B. M. Francis, J. Yang, B. J. Song, S. Gupta, M. Maj, R. P. Bazinet, B. Robinson, and H. T. J. Mount, "Reduced levels of mitochondrial complex I subunit NDUFB8 and linked complex I + III oxidoreductase activity in the TgCRND8 mouse model of Alzheimer's disease.," *J. Alzheimers. Dis.*, vol. 39, no. 2, pp. 347–55, Jan. 2014.

[110] A. H. Bhat, K. B. Dar, S. Anees, M. A. Zargar, A. Masood, M. A. Sofi, and S. A. Ganie, "Oxidative stress, mitochondrial dysfunction and neurodegenerative diseases; a mechanistic insight.," *Biomed. Pharmacother. = Biomédecine pharmacothérapie*, vol. 74, pp. 101–10, Aug. 2015.

[111] A. Brockington, P. R. Heath, H. Holden, P. Kasher, F. L. P. Bender, F. Claes, D. Lambrechts, M. Sendtner, P. Carmeliet, and P. J. Shaw, "Downregulation of genes with a function in axon outgrowth and synapse formation in motor neurones of the VEGFdelta/delta mouse model of amyotrophic lateral sclerosis.," *BMC Genomics*, vol. 11, p. 203, Jan. 2010.

[112] K. Zuberi, M. Franz, H. Rodriguez, J. Montojo, C. T. Lopes, G. D. Bader, and Q. Morris, "GeneMANIA Prediction Server 2013 Update," *Nucleic Acids Res.*, vol. 41, no. W1, pp. W115–W122, Jul. 2013.

[113] "Help · GeneMANIA." [Online]. Available: http://pages.genemania.org/help/. [Accessed: 19-Dec-2017].

[114] F. Bibi, M. Yasir, S. S. Sohrab, E. I. Azhar, M. H. Al-Qahtani, A. M. Abuzenadah, M. A. Kamal, and M. I. Naseer, "Link Between Chronic Bacterial Inflammation and Alzheimer Disease," *CNS Neurol. Disord. -Drug Targets*, vol. 13, pp. 1140–1147, 2014.

[115] D. K. V. Kumar, S. H. Choi, K. J. Washicosky, W. A. Eimer, S. Tucker, J. Ghofrani, A. Lefkowitz, G. McColl, L. E. Goldstein, R. E. Tanzi, and R. D. Moir, "Amyloid-β peptide protects against microbial infection in mouse and worm models of Alzheimer's disease.," *Sci. Transl. Med.*, vol. 8, no. 340, p. 340ra72, 2016.

[116] S. A. Harris and E. A. Harris, "Herpes Simplex Virus Type 1 and Other Pathogens are Key Causative Factors in Sporadic Alzheimer's Disease," *J. Alzheimer's Dis.*, vol. 48, no. 2, pp.

319–353, Sep. 2015.

[117] N. S. Lurain, B. A. Hanson, J. Martinson, S. E. Leurgans, A. L. Landay, D. A. Bennett, and J. A. Schneider, “Virological and Immunological Characteristics of Human Cytomegalovirus Infection Associated With Alzheimer Disease,” *J. Infect. Dis.*, vol. 208, no. 4, pp. 564–572, Aug. 2013.