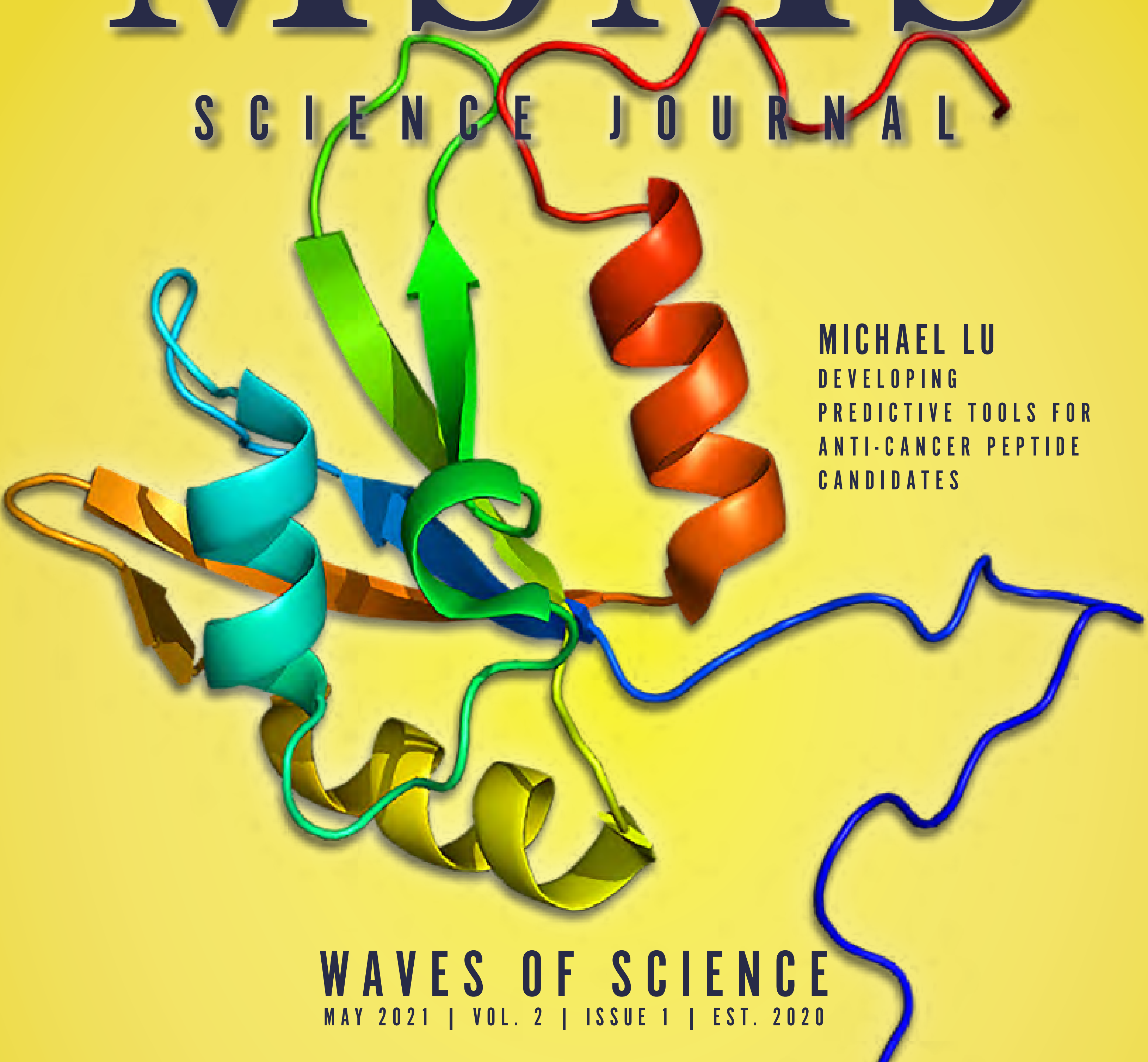
ISSN 2692-191X

MSMS

SCIENCE JOURNAL

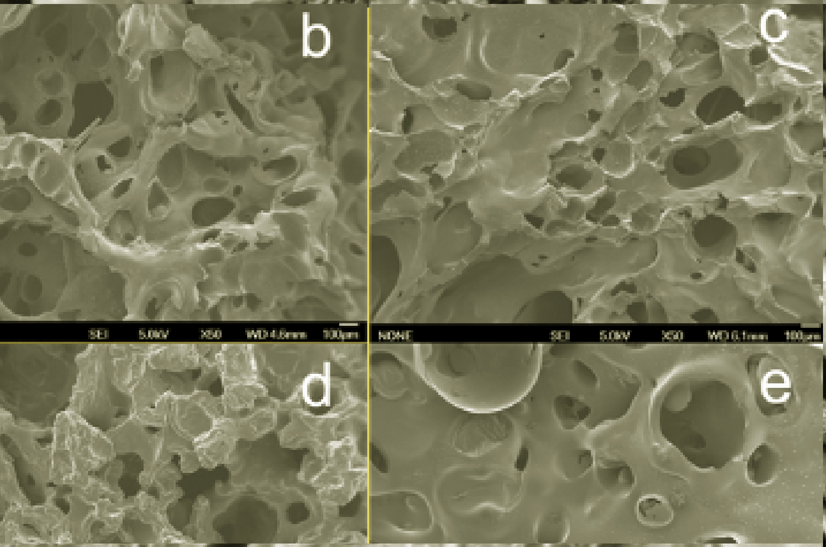


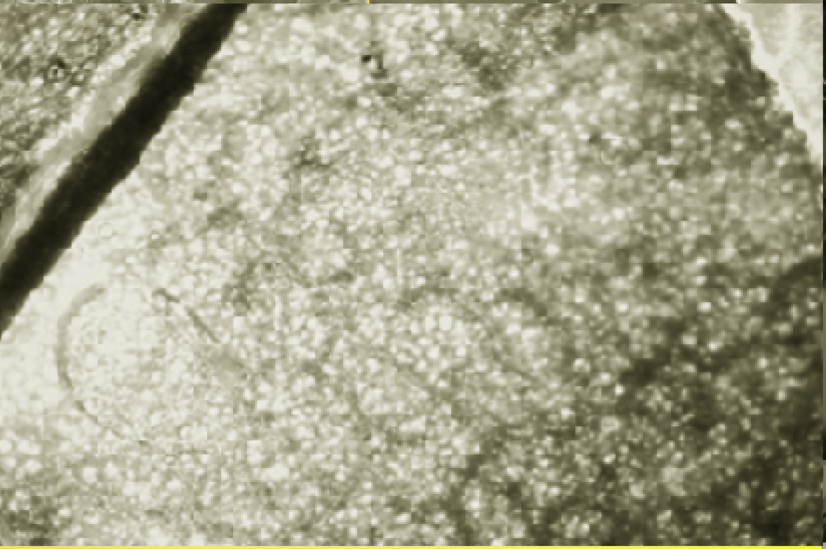
MICHAEL LU

DEVELOPING PREDICTIVE TOOLS FOR ANTI-CANCER PEPTIDE CANDIDATES

WAVES OF SCIENCE

MAY 2021 | VOL. 2 | ISSUE 1 | EST. 2020


b
c
SEI 5.0kV X50 WD 4.6mm 100µm
NONE SEI 5.0kV X50 WD 6.1mm 100µm
d
e



Cover Diagram: Michael Lu (2021)

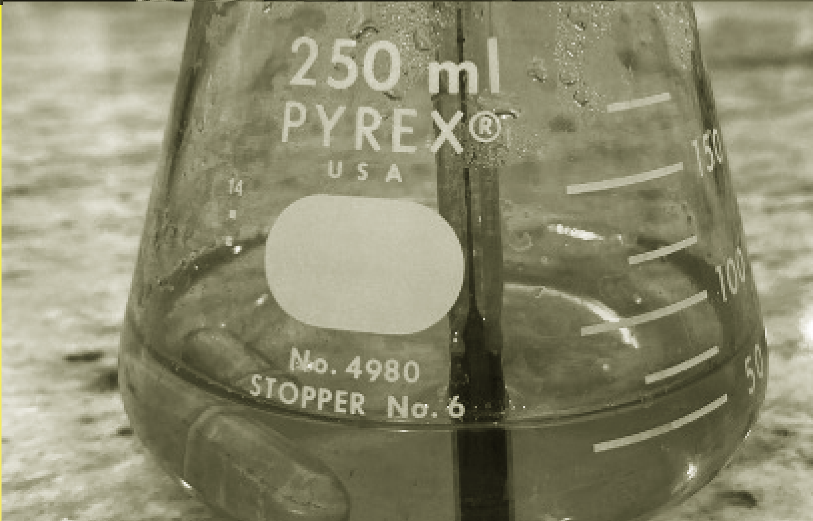
Faculty Advisor: Dr. Tina Gibson

Cover Design: Skylar Nguyen (2021)

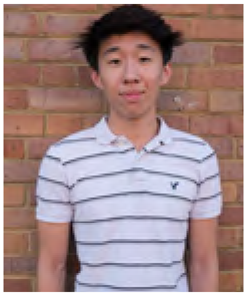
Photography: Caleb Jenkins (2022)

Editor: Skylar Nguyen

Editor elect: Jessica Yan (2022)



CONTENTS



A Novel Evolution-Based Technique for Generating Anticancer Peptides

Michael Lu

Abstract

Cancer threatens millions of individuals every year and remains a central issue in the scientific community. It's important to search for new alternative treatments, and anticancer peptides (ACP's) are a potential key solution. Characterized by peptides between the length of 10-25 amino acids, ACP's can target a variety of cancers while avoiding common side effects caused by conventional drugs. The current in vitro method of discovery of ACP's is both time-consuming and expensive, so computational tools have the potential to expedite the discovery process. The goal of this study is to build a machine learning-based genetic algorithm that can produce new anticancer peptide candidates through generations of training. First, a random forest and a support vector machine model were trained with a 91.2% testing accuracy to determine if a randomly given peptide carries anticancer properties. Then, these classifiers were used to construct a genetic algorithm. A population of 100 random peptides were generated and evaluated using the classifiers trained before, and the top five performing peptides were chosen to repopulate a new generation of 100 peptides with random mutations. After training for 25 generations, the population's average chance of exhibiting anticancer properties grew to 81%, and the top five peptides in the final generation averaged a fitness score of 0.9. These results reflect the potential of evolutionary algorithms in developing new ACP-based treatments to cancer.

Introduction

Because cancer affects millions around the world every year, it is important to investigate new alternative treatments to cancer. Over the last several years, scientists have started exploring the use of anticancer peptides (ACP's) as a new type of therapeutic treatment. Unfortunately, the use of conventional drugs against cancer is becoming increasingly ineffective due to resistance and harmful side effects. However, scientists have found that ACP's can avoid these harmful side effects. In fact, they are safer and have many comparative advantages such as high levels of activity, specificity, affinity, and they are less immunogenic and have better delivery control (Gholibeikian et al., 2019). ACPs are peptide chains that are usually 10-25 residues in length. The three main types of ACPs are pore-forming peptides, cell penetration peptides, and tumor-targeting (Marqus et al., 2017). Currently, there are only 10 anticancer peptides that are currently being developed as drugs due to the difficulty in discovery and development (Shoombuatong et al., 2018). To accelerate the slow pace of discovery, machine learning can be used. More specifically, machine learning algorithms are capable of recognizing patterns in large datasets. Classifier machine learning algorithms are able to classify between different classes. In the realm of anticancer peptides, classifier algorithms can train on datasets of hundreds of ACP's to predict whether random peptides carry anticancer properties. Evolutionary computational methods can also be used to generate new examples based off of a dataset. A genetic algorithm is a type of computational evolutionary algorithm that models natural selection and evolution (Whitley, 1994). Given an objective function and a randomly generated population, then this population can "evolve." The engineering goal of this research is to use both machine learning classifiers and genetic algorithms to generate new predicted candidates.

Datasets: Datasets of verified anticancer peptides were collected from three sources: the Data Repository of Antimicrobial Peptides (DRAMP) database (Fan et al., 2016; Kang et al., 2019; Liu et al., 2018; Liu et al., 2017), the Antimicrobial Peptide Database (APD) (Wang et al., 2016; Wang et al., 2009; Wang et al.) database, and the Anticancer Peptide and Protein Database (CancerPPD). Sequences shorter than 12 residues were removed to allow for future physicochemical calculators, and the datasets were combined for a total of 584 verified ACP's. For the non-ACP class, 584 unique anticancer peptide sequences were sampled randomly from the Swiss Protein database (UniProt Consortium, 2019) which contains 561,568 annotated and reviewed peptide sequences. These naturally occurring sequences were between the lengths of 10-25 residues.

Data Processing

Physicochemical features were extracted from each peptide sequence with the PyBioMed library for Python. The features contain 10,049 total features (e.g. amino acid composition, charge, hydrophobicity, etc.), and datasets were normalized in the range from 0 to 1 for each feature using the following formula:

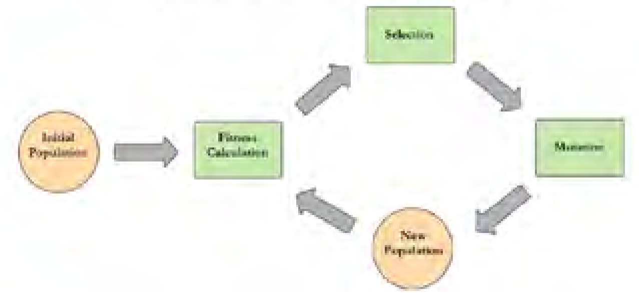
$$X_{i,k} = \frac{X_{i,k} - min\ (X_{i,k})}{max(X_k) - min\ (X_k)}$$

where X_(i,k) represents a feature for ith sample and kth feature. Lastly, random forest feature importance was used to isolate the top 1000 features. These features were deemed the most influential features for anticancer peptide prediction. The dataset was reduced to these features. These processed datasets were then divided into train-test splits and fed into machine learning models.

Machine Learning Models

There were two different types of machine learning models tested for the classification model. The first is the Support Vector Machine, which is a type of model that finds the hyperplane that maximizes the margins between the two classes (Suykens et al, 1999). The second model is Random Forest (RF), which is an ensemble learning method that utilizes a group of weak predictors known as decision trees (Liaw et al, 2002). To design the evaluation function, a mathematical expression of both confidence predictions from the random forest and support vector machine are used. These machine learning models can be used in conjunction with a genetic algorithm to generate new instances of ACP's. A genetic algorithm is an evolution-based computational algorithm that finetunes a random starting population into a desired population through multiple generations of natural selection. The cycle is shown below:

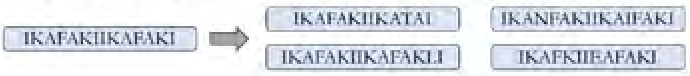
Figure 1: The events in a generation cycle.



In each generation, an objective/fitness function is used to evaluate and label every sequence in the population with a fitness score. In this study, the fitness score represented the machine learning classifier's confidence level in a peptide carrying anticancer properties (0.0-1.0). To generate new ACP candidates, accurate machine learning models are needed to act as a "natural selective force." The complete list of steps in a generation is shown:

1. An initial population of 100 random peptides was generated by the modlamp python package. This step isn't repeated.
2. The 100 peptides were converted into 10,049 physicochemical features and cut down to 1,000 features using random forest feature selection
3. The 100 peptides are evaluated through the fitness function, multiplying the two outputs of the machine learning models
4. The top 5 peptides are selected
5. Mutations are applied onto the peptides with a 10% chance of a deletion, insertion, or substitution at each letter

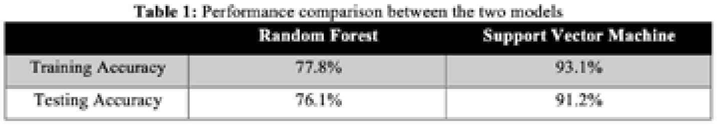
Figure 2: Example of generation mutation in the 1st generation of training



Results

The machine learning models were trained on 80% of the data and tested on 20% of the data. After 10 iterations of training and testing, the random forest algorithm achieved a 77.8% training accuracy and a 76.1% testing accuracy, and the support vector machine achieved a training accuracy of 93.1% and a testing accuracy of 91.2%.

Table 1: Performance comparison between the two models



	Random Forest	Support Vector Machine
Training Accuracy	77.8%	93.1%
Testing Accuracy	76.1%	91.2%

The fitness function was defined as the product of the confidence scores between the random forest and support vector machine models. This fitness score is used to evaluate the population in every generation. After 20 generations of the generation cycle with a mutation rate of 10%, the following figures display the population's progression in fitness:

Figure 3: Average Fitness over 25 Generations

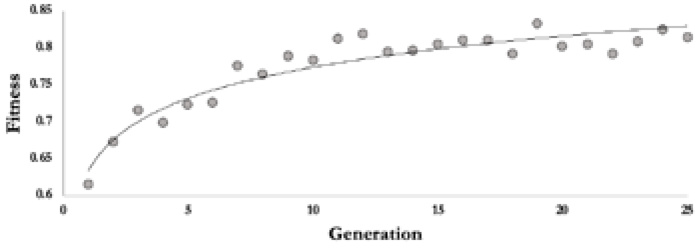
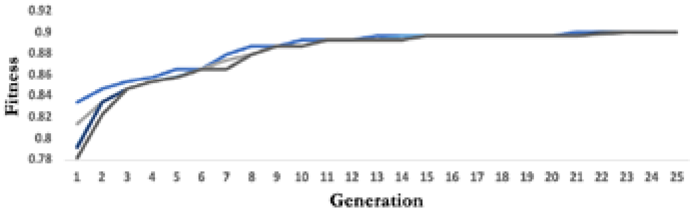


Figure 4: Fitness of the Top 5 Peptides



The population's average fitness increases from 0.6 to 0.8, and the scores of the top 5 peptides in each generation increase from 0.8 to 0.9. After 25 generations of training, the fitness scores level off and increases in fitness scores were negligible.

Discussion

After 25 generations, the machine learning models finetuned a population of 100 random peptides into anticancer peptide candidates with an 80% confidence level for each peptide. In other words, these peptides exhibit crucial physicochemical properties that allow for cell penetration, membrane penetration, or tumor targeting. The machine learning models used achieved relatively high accuracies between 70-90%. This was found by testing the model on the test peptide datasets. One weakness in the generation cycle is the potential for the finetuning of the population to be towards antimicrobial peptides in general instead of cancer-specific peptides. The random forest model achieved an accuracy between 70-80% to determine whether a general antimicrobial peptide is specific to targeting cancer. This means that around 20-30% of the final population likely includes peptides that are antimicrobial but not anticancer. However, the general peptide in the final population should still be specific to cancer.

Future Works

This research can be expanded in a few ways. The top peptide candidates among the final generation can be synthesized and tested in a wet lab experiment against cancer cell lines to find the actual efficacy of the peptides. This experiment would also help verify the specific types of cancer that each peptide could target. In addition, on the backend, the machine learning models can likely also be advanced. Machine learning capabilities are improved every year and using deep learning models can potentially improve classification accuracy and provide better fitness evaluations.

References

Chiangjong, W., Chutipongtanate, S., & Hongeng, S. (2020). Anticancer peptide: Physicochemical property, functional aspect and trend in clinical application. International Journal of Oncology, 57(3), 678-696.

Fan L.; Sun J.; Zhou M.; Zhou J.; Lao X.; Zheng H.; Xu H. DRAMP: a comprehensive data repository of antimicrobial peptides. Sci Rep. 2016 Apr 14;6:24482. PMID: 27075512

Gholibeikian, M., Bamoniri, A., HoushdarTehrani, M. H., Mirjalili, B. B. F., & Bijanzadeh, H. R. (2019). Structure-activity relationship studies of Longicalcynin A analogues, as anticancer cyclopeptides. Chemico-biological interactions, 108902.

Liaw, A., & Wiener, M. (2002). Classification and regression by random forest. R news, 2(3), 18-22.

Marqus, S., Pirogova, E., & Piva, T. J. (2017). Evaluation of the use of therapeutic peptides for cancer treatment. Journal of biomedical science, 24(1), 21.

Shoombuatong, W., Schaduangrat, N., & Nantasenamat, C. (2018). Unraveling the bioactivity of anticancer peptides as deduced from machine learning. EXCLI journal, 17, 734.

Suykens, J. A., & Vandewalle, J. (1999). Least squares support vector machine classifiers. Neural processing letters, 9(3), 293-300.

UniProt Consortium. (2019). UniProt: a worldwide hub of protein knowledge. Nucleic acids research, 47(D1), D506-D515.

Wang, G., Li, X. and Wang, Z. (2016) APD3: the antimicrobial peptide database as a tool for research and education. Nucleic Acids Research 44, D1087-D1093. Paper PDF

Whitley, D. (1994). A genetic algorithm tutorial. Statistics and computing, 4(2), 65-85.

Reinforced Lignin Foams with Higher Adsorption Capability

Jessica Yan

Abstract

In a previous study, open-cell lignin foams were prepared from Kraft lignin through a simple baking process. Lignin foams demonstrated good biosorbent capabilities for removing heavy metals and spilled oil from water. However, the foams need to be reinforced in strength since raw lignin foams are brittle and exhibit poor mechanical properties, and the adsorption capability needs to be further improved. To prepare lignin foams with reinforced mechanical strength and improved adsorption capabilities, two hypotheses were made in this project: the first hypothesis was that lignin foam strength can be improved by adding plastic polymers from recycled plastics, and the second hypothesis was that the adsorption capacity of the foam can be improved by adding wood waste-derived activated carbon (AC) which has a higher surface area and adsorption capacity to heavy metals, spilled oil and other pollutants. The mechanical performance of the foam reinforced with waste plastic was tested using a mechanical compression machine; the strength of the foam was significantly improved (more than 12 times) by adding 10wt% polyethylene. To enhance the adsorption capacity of lignin foams, different amounts of AC (0.5, 1, 2 and 3wt%) were coated onto the surface of the foam. The adsorption capabilities of lignin foam to copper and oil in water were improved by at least 10% when coated with 1wt% activated carbon. According to the obtained data, two research hypotheses were proven valid through this project.

Introduction

Toxic heavy metals are harmful to aquatic life and cause a variety of diseases and disorders as they accumulate in the human body via the food chain (Tchounwou, et al, 2012). Currently, most conventional methods that are used to remove heavy metals are uneconomical and have disposal issues (Gunatilake, 2015). Spilled oil, another source of pollution to water bodies, severely affects the marine environment, causing a decline in aquatic life (Saadoun, 2005). Various methods have also been developed to recover organic solvents and oil from water (Doshi, et al, 2018). Among these methods, adsorption is regarded as an effective and economical technology due to its high efficiency, simple design, and smooth operation (Crini, et al, 2018).

This study demonstrates a novel method for resolving oil spills using Kraft lignin. Lignin, which is usually in a fine powder from pulp mills, has been studied extensively due to its unique physicochemical properties for effectively adsorbing organic and inorganic pollutants from water (Ge & Li, 2018). However, lignin is difficult to recover and recycle in its fine powder form.

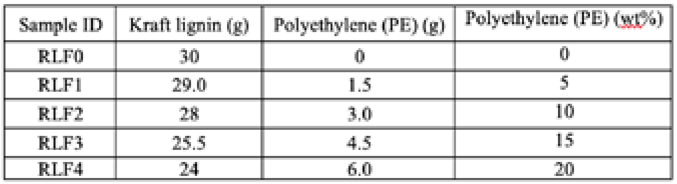
Lignin-derived foam is a promising adsorbent for pollution control as it is a lightweight, three-dimensional porous structure with a high surface area. Previously, lignin foams have been prepared from lignin powder through a simple baking process and tested as efficient adsorbents. However, the foams need to be reinforced in strength since they exhibit poor mechanical properties, and the adsorption capability can be further improved. Thus, the purpose of this project is to prepare lignin foams with reinforced mechanical strengths and improved adsorption capabilities.

Methods

Preparation of reinforced lignin foams (RLFs) with recycled polyethylene (PE)

To fabricate reinforced lignin foams, precursors of the reinforced lignin foams consisted of Kraft lignin powder and recycled shredded polyethylene (PE) in different weight ratios were blended. The mixtures were transferred into a cylinder mold that was heated in an oven to a temperature of 450°F (~230°C) and held for 60 minutes. The oven was cooled to room temperature, and the mold was taken out. This procedure resulted in an open-celled, self-expanded lignin foam. The precursor compositions of the reinforced lignin foams are listed in Table 1. The foam samples with different PE contents are labeled as RLF0-RLF4.

Table 1. The formulas of reinforced lignin foams (RLFs)

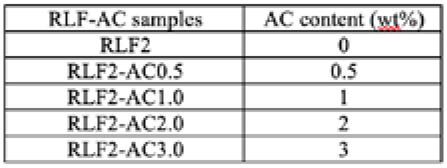


Sample ID	Kraft lignin (g)	Polyethylene (PE) (g)	Polyethylene (PE) (wt%)
RLF0	30	0	0
RLF1	29.0	1.5	5
RLF2	28	3.0	10
RLF3	25.5	4.5	15
RLF4	24	6.0	20

Impregnation of reinforced lignin foams (RLFs) with activated carbon (AC)

To improve the adsorption capabilities of RLFs, RLF2 foam samples were created and coated with activated carbon at varying weight percentages. First, ethanol was poured into a glass beaker, and activated carbon was added to the beaker and sonicated. The RLF blocks were added and stirred to let the AC particles penetrate to the foam pores. The RLF blocks were taken out and dried in an oven. The foam blocks were washed in ethanol again to remove the AC particles that were trapped in the pores without sticking with the surface. The RLF2 samples coated with varying AC contents are labeled as RLF2-ACs in Table 2.

Table 2. The reinforced lignin foams (RLFs) coated with different AC contents



RLF-AC samples	AC content (wt%)
RLF2	0
RLF2-AC0.5	0.5
RLF2-AC1.0	1
RLF2-AC2.0	2
RLF2-AC3.0	3

Evaluation of lignin foam adsorption capability to copper ion in water

A stock solution of copper was prepared by dissolving copper sulphate in distilled water. The effect of adsorption mass on the equilibrium adsorption of copper was investigated with lignin foams in copper solution (Figure 1).

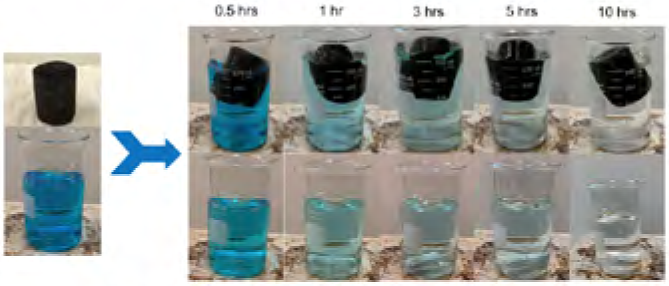


Figure 1. Copper removal by lignin foam over time

Evaluation of lignin foam adsorption capability to cooking oil in water

For adsorption tests, cooking oil was transferred into a beaker with de-ionized water (Figure 2). The effect of removal time and mass of lignin foam on the effectiveness of oil was studied.

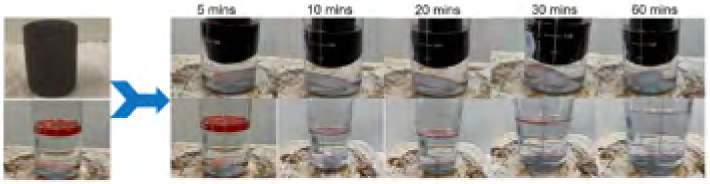


Figure 2. Oil removal by lignin foam over time

The adsorption capacity (Q) was calculated using the weight of the lignin foam before and after adsorption as following Equation 7:

Results

Effect of polyethylene on the compressive strength of lignin foams

Figure 3 shows the effect of PE content on the compressive strength of lignin foams. Pure lignin foam (RLF0) presented a very poor mechanical property in terms of compressive strength; the compressive strength of the RLF0 sample was the lowest, as it was 0.48± 0.19 MPa. When the PE content was 10 wt%, the compressive strength of RLF increased to 5.92 ± 0.75 MPa and was about 12 times higher than the RLF0 sample.

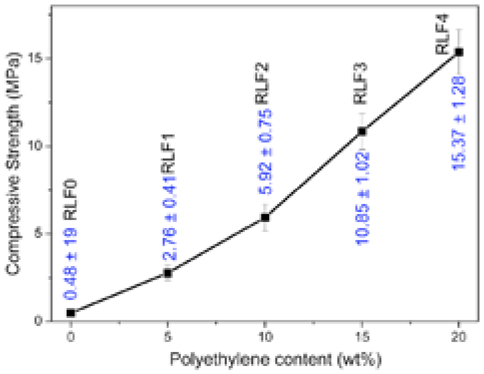
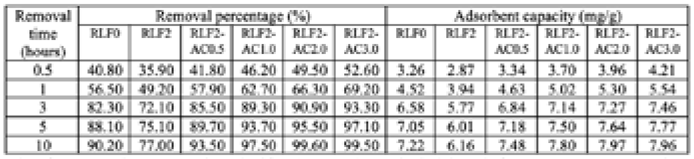


Figure 3. Effect of polyethylene content on the compressive strength of reinforced lignin foams

Effect of contact time of lignin foams in adsorption of heavy metal ions in water

The effect of contact time on the removal efficiency of copper ions from aqueous solutions using lignin foam was studied. The results are shown in Table 3. Overall, the experimental results show the copper removal rates increased as the coated AC content increased, and the adsorption capability of RLF-AC samples to copper ions increased as coated AC content increased.

Table 3 Relationships between contact time and the removal percentage and adsorption capacity for copper ions by reinforced lignin foams.



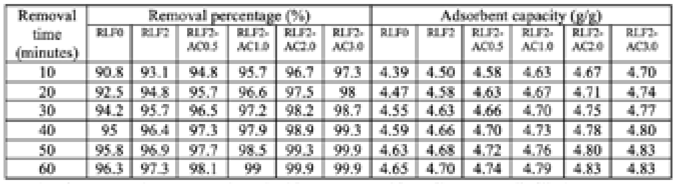
Removal time (hours)	Removal percentage (%)						Adsorbent capacity (mg/g)					
	RLF0	RLF2	RLF2-AC0.5	RLF2-AC1.0	RLF2-AC2.0	RLF2-AC3.0	RLF0	RLF2	RLF2-AC0.5	RLF2-AC1.0	RLF2-AC2.0	RLF2-AC3.0
0.5	40.80	35.90	41.80	46.20	49.50	52.60	3.26	2.87	3.34	3.70	3.96	4.21
1	56.50	49.20	57.90	62.70	66.30	69.20	4.52	3.94	4.63	5.02	5.30	5.54
3	82.30	72.10	85.50	89.30	90.90	93.30	6.58	5.77	6.84	7.14	7.27	7.46
5	88.10	75.10	89.70	93.70	95.50	97.10	7.05	6.01	7.18	7.50	7.64	7.77
10	90.20	77.00	93.50	97.50	99.60	99.50	7.22	6.16	7.48	7.80	7.97	7.96

(5 g foam samples were each soaked in 200 mL 200 mg/L $CuSO_4$ solution at room temperature).

Effect of removal time of lignin foams in adsorption of oil in water

The effect of contact time on the removal efficiency of oils from oil-water samples using lignin foam was studied. The removal percentage and capacity of adsorbent obtained during the experiments were presented in Table 4. From the table, it is apparent that the percentage of removal increased with the increase of sorption time. Also, the removal percentage and the corresponding adsorption capability increased when PE is added to lignin foam, and the removal rates and the adsorption capabilities to oil are further improved for the RLF-AC samples

Table 4. Relationships between contact time and the removal percentage and adsorption capacity for cooking oil by lignin foam.



Removal time (minutes)	Removal percentage (%)						Adsorbent capacity (g/g)					
	RLF0	RLF2	RLF2-AC0.5	RLF2-AC1.0	RLF2-AC2.0	RLF2-AC3.0	RLF0	RLF2	RLF2-AC0.5	RLF2-AC1.0	RLF2-AC2.0	RLF2-AC3.0
10	90.8	93.1	94.8	95.7	96.7	97.3	4.39	4.50	4.58	4.63	4.67	4.70
20	92.5	94.8	95.7	96.6	97.5	98	4.47	4.58	4.63	4.67	4.71	4.74
30	94.2	95.7	96.5	97.2	98.2	98.7	4.55	4.63	4.66	4.70	4.75	4.77
40	95	96.4	97.3	97.9	98.9	99.3	4.59	4.66	4.70	4.73	4.78	4.80
50	95.8	96.9	97.7	98.5	99.3	99.9	4.63	4.68	4.72	4.76	4.80	4.83
60	96.3	97.3	98.1	99	99.9	99.9	4.65	4.70	4.74	4.79	4.83	4.83

(3g foam samples were each soaked in 15 mL cooking oil-105 mL distilled water at room temperature.)

Discussion

The results show that the mechanical strength of the lignin foam was successfully reinforced by adding waste plastic to the lignin powder. The compressive strength of the lignin foam was significantly increased by approximately 12 times from 0.48 MPa to 5.92 MPa when adding 10 wt% PE. In addition, the adsorption capabilities of lignin foam to copper and oil in water were improved by at least 10% when coating 1wt% activated carbon to the surface of the foam. Through this project, two research hypotheses were proven valid. By integrating plastic polymers with lignin, the mechanical strength greatly improved, and by coating lignin foams' surfaces with activated carbon, the adsorption capabilities were improved as carbon-based adsorbents possess high surface area, excellent mechanical properties, and high adsorption capacities.

Implications and Recommendations

Globally, around 2 billion people are affected by contaminated water, and approximately 100 million animals die each year due to water pollution. In addition, over a million gallons of oil are spilled into ocean waters annually. Lignin foams are estimated to save millions of animals through rapid adsorption, and foams can significantly reduce the harmful impacts of heavy metals and oils on the marine environment. Overall, reinforced lignin foams can become a sustainable, cost-effective alternative to tackling global water problems. By utilizing reinforced lignin foams as a solution alongside other options, such as bioremediation, oil booms usage, precipitation, and ion exchange, to resolve oil spills and water crises, countries around the world will have sufficient access to clean water and sanitation sources for billions of people, and environments will be restored and protected.

References

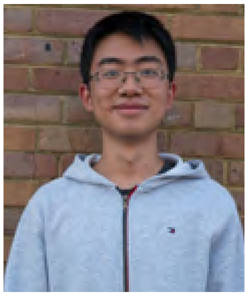
Crini, G., et al. (2018). Adsorption-oriented processes using conventional and non-conventional adsorbents for wastewater treatment, G. Crini, E. Lichtfouse (Eds.), Green Adsorbents for Pollutant Removal: Fundamentals and Design, Springer International Publishing, Cham, p. 23.

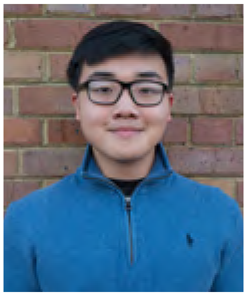
Doshi, B., Sillanpaa, M., Kalliola, S. (2018). A review of bio-based materials for oil spill treatment, Water Res., 135, pp. 262-277.

Ge, Y., Li, Z. (2018) Application of lignin and its derivatives in adsorption of heavy metal ions in water: a review. ACS Sustain. Chem. Eng., 6 (5), pp. 7181-7192.

Gunatilake, S. (2015). Methods of Removing Heavy Metals from Industrial Wastewater, Methods.

Tchounwou PB., et al. (2012). Heavy metal toxicity and the environment. Mol Clin Environ Toxicol Exp Supplementum. 101:133–64.





Assessing the Efficacy of COVID-19 Policies Using Machine Learning

Andrew Yu

Aaron Wan

Abstract

To limit the spread of COVID-19, government policies are necessary. Current research focuses primarily on documenting proposed policies rather than assessing the efficacy of said proposals. Thus, the goal of this research is to analyze COVID-19 policies using machine learning methods. Specifically, the aims of this project are to develop an accurate machine learning model to predict the effectiveness of COVID-19 policies and use that model to find the most effective COVID-19 policies. Data from the CoronaNet Research Project and Our World in Data was used to provide a list of over 50,000 government policies in 195 countries as well as the corresponding COVID-19 data. A random forest regression algorithm achieved an R2 value of 0.88 on training data and 0.63 on test data. These high values affirmed the strength of our model. Using the feature importances attribute of the model, the top 15 most important policy features were identified, and these results were affirmed by a feature selection algorithm. The model is novel because it conducts a holistic analysis of COVID-19 policies, meaning it can used by policymakers to compare the effectiveness of a wide range of policies. Additionally, unlike others, the model can predict the impact of proposed policies. Based on these findings, it is recommended that governments increase policy adherence through methods such as publicizing information concerning the pandemic. It is also recommended for governments to prioritize policies that restrict external borders and regulate businesses. Policies should target all residents and be mandated through consequences like jail time and fines.

Introduction

On December 26, 2019, the first reported case of an unknown disease was reported in Wuhan, China. Further inspection by the Chinese Center for Disease Control and Prevention determined that this disease was the first strain of the novel coronavirus (2019-nCoV) on January 7, 2020. Since then, COVID-19 has spread quickly, creating an international pandemic that has taken hundreds of thousands of lives (Jung et al., 2020). Effective government intervention policies are needed to combat the spread of this deadly virus and reduce mortality in those already infected. Because of the novelty of COVID-19, current research is mainly focused on documenting the policies governments have proposed rather than assessing the efficacy of said proposals (Cheng et al., 2020). Unfortunately, the little research that has been conducted is far too specific in its policy considerations (Kai et al., 2020) or the scope of implementation (Jiwei et al., 2020). As such, the goal of this paper is to provide a comprehensive analysis of the various COVID-19 policies that have been proposed across the globe in order to determine what kind of policies affect COVID-19 the most. To accomplish this, multinational data describing various governmental policies was combined with corresponding data for the cases and deaths due to COVID-19 during the timeframe of the policies. This data was then incorporated into a random forest regression model, which was analyzed to produce a set of rankings for optimal policies that were able to control COVID-19 cases and deaths.

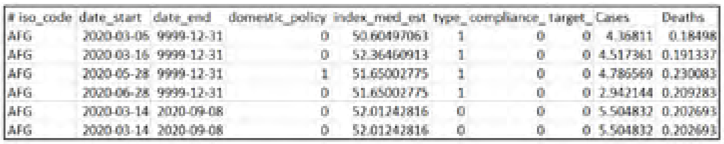
Data Selection

Two databases were utilized: the CoronaNet Research Project and Our World in Data's COVID-19 data. The first dataset documented over 50,000 policy responses to COVID-19 from 195 countries. From this dataset, factors such as whether the policy was domestic and the index_med_est (the anticipated level of response to the policy, from 0-100, as calculated by the CoronaNet Research Project) were used. Other categorical policy factors included the type of policy (social distancing, border restrictions, health monitoring, etc.), the compliance mechanism (mandatory, voluntary, etc.), and the policy's target population, were selected. To mathematically represent these categorical variables, factors were one-hot encoded into multiple columns with binary data representing whether a specific policy fell under that category. The latter dataset from Our World in Data contained daily country-specific COVID-19 data. From this dataset, the number of daily new cases per million individuals and daily new deaths per million individuals were taken.

In the case of null data values, the means of comparable data were taken to replace the missing datapoint. For instance, in cases where index_med_est¬ was unavailable for a certain policy, it was assumed that the anticipated level of response to the policy would be similar to that of other policies proposed by the same government, and the average index_med_est¬ of other policies from the same country was used.

To merge the two datasets, each policy was matched with the corresponding average COVID-19 data for the timeframe of the policy. Additionally, excess data for countries in one dataset and not the other was removed.

The final data was arranged as shown below, although additional columns representing each of the possible types, compliance levels, and target populations of the policy are hidden.



# iso_code	date_start	date_end	domestic_policy	index_med_est	type_	compliance_	target_	Cases	Deaths
AFG	2020-03-06	9999-12-31	0	50.60497063	1	0	0	4.36811	0.18498
AFG	2020-03-16	9999-12-31	0	52.36460913	1	0	0	4.517361	0.191337
AFG	2020-05-28	9999-12-31	1	51.65002775	1	0	0	4.786569	0.230083
AFG	2020-06-28	9999-12-31	0	51.65002775	1	0	0	2.942144	0.209283
AFG	2020-03-14	2020-09-08	0	52.01242816	0	0	0	5.504832	0.202693
AFG	2020-03-14	2020-09-08	0	52.01242816	0	0	0	5.504832	0.202693

The model was trained to take inputs of whether the policy was domestic, the expected level of response by people within the nation, the type of policy, method of compliance to enforce the policy, and the targeted demographic of the policy. In total, 40 input variables were incorporated in the model. The output of the model was trained to be the average number of new cases and deaths, per million, that occurred during the timeframe the policy was implemented.

Model Development

After testing multiple machine learning models, a random forest regressor was determined to be the best fit for the data due to its accuracy and ability to control for over-fitting. Using the Python package scikit-learn, the model was trained on 80% of the dataset. The remaining 20% of the dataset was used as test data to affirm the model's effectiveness.

After training the model, the model's feature importance's attribute was analyzed. This attribute was calculated based on the proportion of nodes within the decision trees that utilized each input variable. More influential input variables were used within the model more times and thus had a higher feature importance.

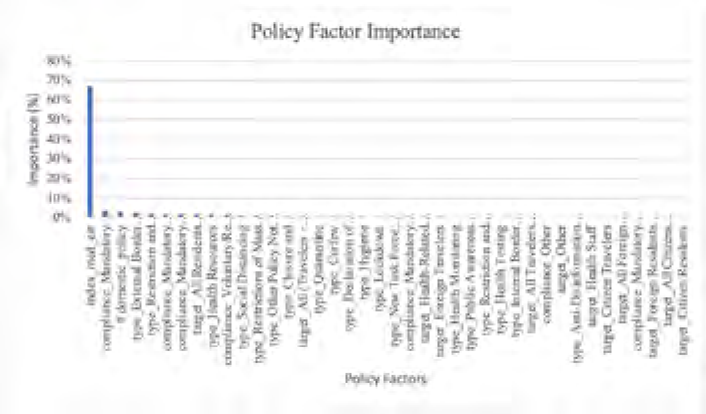
To confirm model results, a recursive feature elimination selector, or RFE selector, was utilized. The goal of this algorithm was to recursively eliminate subsets of data and train the model on these subsets. If the algorithm was able to train a model to achieve relatively high accuracy without certain input variables, those variables were deemed as less influential. The selector eventually reached a final subset of input variables which were determined to be the most important.

Results

The final model achieved an R^2 value of 0.878 on training data and an R^2 of 0.626 on test data. The R^2 value of 0.626 on test data indicates that, given the features of any policy, the model's predictions explain 62.6% of the variance in new cases and deaths during that policy's implementation.

Because predicting the outcome of COVID-19 involves biological and behavioral factors and the fields of biology and social sciences consider an R^2 of 0.5 as high, this R^2 value indicates that the random forest regression model is a good fit for the data.

The final importance levels of all 40 policy factors, as a percentage of time the factor was used in the model, are displayed below:



The most significant factor was revealed to be index_med_est, which represents how likely people will respond to the policy. Regarding policy type, policies that restrict external borders and regulate businesses were determined to affect COVID the most. The most effective compliance mechanism was mandating with legal penalties such as jail time, and the demographic that should be targeted was determined to be all residents.

After ranking each of the policy factors, the RFE selector was used to confirm the findings. The RFE selector ended up selecting half of the 40 features as significant. These 20 policy factors are displayed below:

Feature Selector Results			
domestic_policy	index_med_est	type_Closure and Regulation of Schools	type_Curfew
type_Declaration of Emergency	type_External Border Restrictions	type_Health Resources	type_Lockdown
type_Other Policy Not Listed Above	type_Quarantine	type_Restriction and Regulation of Businesses	type_Restriction and Regulation of Government Services
type_Restrictions of Mass Gatherings	type_Social Distancing	compliance_Mandatory (Unspecified/Implied)	compliance_Mandatory with Fines
compliance_Mandatory with Legal Penalties (Jail Time)	compliance_Voluntary/Recommended but No Penalties	target_All (Travelers + Residents)	target_All Residents (Citizen Residents + Foreign Residents)

After eliminating all other features from the model and only using these factors to train the model, the random forest regressor was able to achieve an R^2 value of 0.870943. This value indicates that only these policy factors are necessary to create accurate predictions.

Discussion

This research is one of the first to conduct a holistic analysis of COVID-19 policies. The results from the random forest regressor provide a ranking for the importance of policy factors in controlling both COVID-19 cases and deaths. These rankings can be used by other researchers and policymakers to compare the efficacy of a wide range of policies based on their features in order to determine which policies should be prioritized. Additionally, the model also provides predictions for the impacts of future policies.

Recommendations

It is recommended for governments to take action based on the results of this research. The most important factor that was identified by the model was the country policy activity score. This indicates that it is essential that governments take action to increase policy adherence through methods such as publicizing information concerning the pandemic. (Al-Hasan, 2020) Even if a policy is perfect, it will not matter if citizens do not adhere to it.

It is also recommended for governments to prioritize policies that restrict external borders and regulate businesses. These policies would limit the spread of COVID-19 from external sources while also restricting domestic businesses, which are likely the largest source of gatherings as they provide essential goods and services.

Policies should target all residents of the country. Although travelers are a source for COVID-19, residents remain in the country long-term, meaning they are the ones most likely to spread the disease.

In order to increase adherence, policies should be mandated. The most effective consequences are legal penalties such as jail time and fines.

Policymakers could use the model to predict the impact of proposed policies. However, further improvements may be necessary to increase the model's accuracy and utility in the field.

References

Al-Hasan, A. (2020, November 11). Citizens' Adherence to COVID-19 Mitigation Recommendations by the Government: A 3-Country Comparative Evaluation Using Web-Based Cross-Sectional Survey Data. Journal of Medical Internet Research, 22(8). https://www.jmir.org/2020/8/e20634/

Cheng, C., Barceló, J., Hartnett, A. S., Kubinec, R., & Messerschmidt, L. (2020, June 23). COVID-19 Government Response Event Dataset (CoronaNet v.1.0). https://www.nature.com/articles/s41562-020-0909-7

Kai, D., Goldstein, G.-P., Morgunov, A., Nangalia, V., & Rotkirch, A. (2020, April 22). Universal Masking is Urgent in the COVID-19 Pandemic: SEIR and Agent Based Models, Empirical Validation, Policy Recommendations. https://arxiv.org/pdf/2004.13553.pdf

Roser, M., Ritchie, H., Ortiz-Ospina, E., & Hasell, J. (2020, March 04). Coronavirus Pandemic (COVID-19) - Statistics and Research. Retrieved January 17, 2021, from https://ourworldindata.org/coronavirus

Development of Filter embedded with Silver Nanoparticles for Water Sanitation

Skylar Nguyen

Abstract

In the 21st century, 780 million people are without clean water, and 2.5 billion people do not have access to water with improved sanitation. Dirty drinking water can cause dysentery, infertility, and waterborne illness. Developing filters embedded with silver nanoparticles aims to create an inexpensive, convenient solution to cleaning contaminated water. First, water will be collected from different regions along Mississippi and Alabama, and these water samples will be put through a bacteria testing kit. After the kit confirms that the water samples have bacteria, those samples will be disposed of, and new water samples from the same sources will be put through the silver nanoparticle filter. The silver nanoparticles will be created from a plant extract, Ocimum basilicum. The plant extract will be used as a reducing and capping agent; it will be suspended over the silver nanoparticle mixture and added until the silver nanoparticle mixture turns yellow. Lower possible environmental bacteria results due to the bioactive properties of Ocimum basilicum and the antimicrobial properties of silver nanoparticles. The null hypothesis is rejected. The filter embedded with silver nanoparticles did decrease possible environmental bacteria in water sourced from all three water sources. Not only do the filters aim to sanitize water, the silver nanoparticles could also have health implications in water gels or bandages to disinfect wounds. The application of this project has the ability to lower people's chances of contracting a waterborne illness and allows for reduced gender inequalities through lessening the time women clean water.

Background

In the 21st century, there are 780 million people without clean water, and 2.5 billion people without access to water with improved sanitation (more than 35% of the population does not have access to cleaner drinking water). The World Health Organization (WHO) and UNICEF have conducted research showing that rural areas within sub-Saharan Africa, Southern Asia, and Eastern Asia have the highest concentration of people without clean water (Gleick, 2002). Dirty drinking water results in water-related diseases. Diseases that arise from contaminated water include but are not limited to cholera, typhoid, dysentery, malaria, and yellow fever.

Through creating more sustainable, low-cost options to sanitize water, this phenomenon can end. Silver nanoparticles have shown high antimicrobial rates when infused in water filtration systems, wound wrappings, and water gels. Dankovich and Gray (2015) impregnated silver nanoparticles into paper pages made of cellulose, filtered contaminated water through the paper, and reduced 200,000 colony forming units (CFU) of Escherichia coli per 100mL to less than ten CFU. Similarly, Che et al. (2019) and Krishnaraj et al. (2010) both synthesized silver nanoparticles and used its antibacterial properties to fight against water borne pathogens. Pankongadisak et al. (2014) used the antibacterial properties of silver nanoparticles to disinfect wounds through inserting calcium alginate beads impregnated with silver nanoparticles into hydrogel under wound dressings.

The intent of this engineering goal is to embed silver nanoparticles into a filter. The filter will be tested on water collected from different sources. A bacteria test before and after using the filter will be conducted to show the effectiveness of bacteria removal. Accomplishing the goal of removing bacteria from the water samples with the filter embedded with silver nanoparticles will add to water sanitation research through creating an inexpensive and convenient way to clean water.

Engineering Goal

Currently, there are filters embedded with silver nanoparticles, but many of them are not cost effective or widely used. The objectives of this engineering design are to embed silver nanoparticles into a commercially made paper filter and to use the filter to lower the possible environmental bacteria in water from the Black Prairie region of Mississippi, Gulf Coast region of Mississippi, and Southeastern Plains region of Alabama. The null hypothesis is that the commercially made paper filter embedded with silver nanoparticles will not lower the possible environmental bacteria in 100mL water samples. The alternate hypothesis is that the commercially made paper filter embedded with silver nanoparticles will lower the concentration of possible environmental bacteria in 100mL water samples.

Methodology

Water samples from different regions along the Mississippi and Alabama coasts were put through an at home bacteria testing kit. After 24 hours, the kit showed if there was bacteria or no bacteria in the water samples. The water samples were disposed of. After this test, new water samples were taken from the same sources, and these new samples were put through the silver nanoparticle filter.

Fresh Ocimum basilicum was collected. The Ocimum basilicum was rinsed and cleaned thoroughly with distilled water. After cleaning the Ocimum basilicum, it was cut into small pieces. Then 25g of Ocimum basilicum was measured and put into 100mL of distilled water. The plant and water mixture was heated to 75 degrees Celsius until the plant extract was created. The plant extract was created when the solution in the mixture turned green.

To create the silver nitrate mixture, 10mL of silver nitrate solution mixed with 50mL of distilled water. The silver nitrate solution is heated to 90 degrees Celsius while the plant extract is suspended over silver nitrate mixture and added drop by drop until the silver nitrate mixture turns yellow. The plant extract is used as the reducing and capping agent for the silver nanoparticles. In the yellow solution, silver nanoparticles are observed.

Image 1: Plant Extract

Image 2: Silver nanoparticles immediately after using the plant extract as a suspending and capping agent

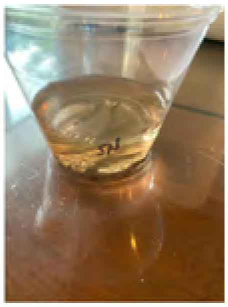


Image 3: Formation of silver nanoparticles before embedded in filter

The filters were placed in the silver nanoparticle solution. After saturating the filter in the solution for ten minutes, the filter was hung until dry. This procedure was repeated on each filter three times. Water was drained through the filter coated in silver nanoparticles. A direct microscopic count was performed to count possible environmental bacteria before and after the water samples were run through the filter.

Image 4: Water from Gulf Coast region before filtration

Image 5: Filter embedded with silver nanoparticles

Image 6: Water from Gulf Coast region after filtration

Results

The water put through the commercially made paper filter embedded with silver nanoparticles lowered possible environmental bacteria in water samples from all three regions. The concentration of possible environmental bacteria from ten samples from each region was averaged. Water from the Gulf Coast region decreased from 1562.67 to 55.33 possible bacteria cell count in 100mL samples. Water from the Black Prairie region decreased from 1420.67 to 52.67 possible bacteria cell count in 100mL samples. Water from the Southeastern Plains region decreased from 1658.67 to 57.33 possible bacteria cell count in 100mL samples.

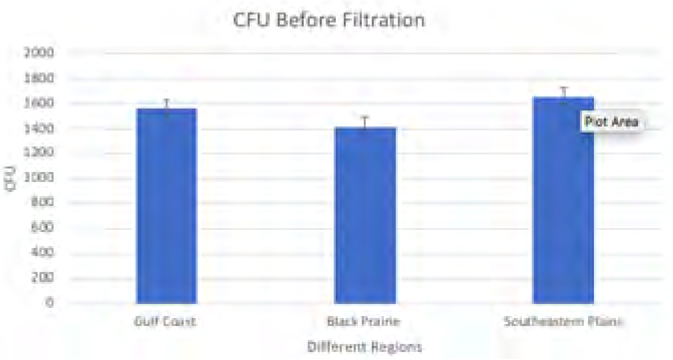


Diagram 1

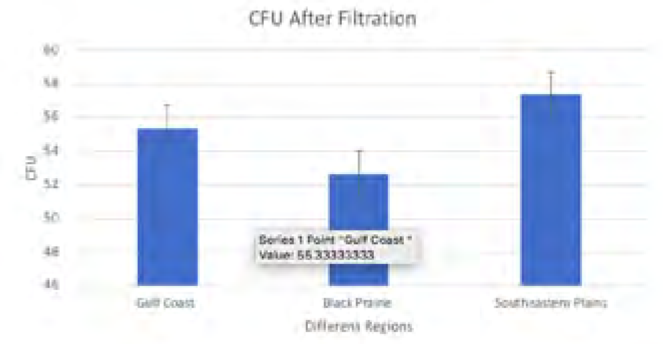


Diagram 2

Conclusions

The null hypothesis is rejected. The null hypothesis is rejected because the filter embedded with silver nanoparticles did decrease possible environmental bacteria in water sourced from all three regions. This can be seen in the decrease in possible environmental bacteria within all the 100mL samples from all three regions. Additionally, it can also be seen that all the error bars overlap into the average bars, showing that there was a consistent decrease in possible environmental bacteria in 100mL samples from the same areas. This means that the filter consistently killed the density of bacteria in the water. This result could be due to the bioactive properties within Ocimum basilicum and the antibacterial properties within silver nanoparticles being able to lower the concentration of the specific bacteria found within the water samples.

Implications

This project contributes to Sustainable Development Goal 3: good health and well-being. Lowering the bacteria concentration of a dirty water source will lower the chance of contracting a waterborne illness. This project contributes to Sustainable Development Goal 6: clean water and sanitation. Filters embedded with silver nanoparticles will clean contaminated water through using the bioactive properties within Ocimum basilicum & the antibacterial properties within silver nanoparticles to lower the bacteria concentration in water. This project contributes to Sustainable Development Goal 10: reduced inequalities. Creating a cheap and accessible way to clean water will reduce inequalities through allowing more women and girls to go to school and receive an education instead of spending hours of their day gathering and cleaning dirty water.

Recommendations

The filters could be commercialized as an inexpensive and convenient way to sanitize small amounts of contaminated water. Embedding the silver nanoparticles in bandages has the ability to disinfect wounds.

Future Research

For future research, more testing, a new design, and a different procedure for creating nanoparticles could be used. Through more testing, more water sources could be tested with higher volume and concentration of bacteria. The effectiveness of the filter on high concentrations of pathogenic bacteria, and the maximum amount of water the filter could withstand before breaking could be calculated. A new design could allow for a filter that is more cost effective. A new design could filter larger amounts of water more quickly and kill a higher percentage of the bacteria. This could be done through using a different capping and reducing agent for the nanoparticles and developing a filter from different materials. Using nanoparticles created through a different procedure or a different capping and reducing agent could decrease more types of bacteria after filtration.

References

Che, W., Xiao, Z., Wang, Z., Li, J., Wang, H., Wang, Y., & Xie, Y. (2019). Wood-based Mesoporous Filter Decorated with Silver Nanoparticles for Water Purification. ACS Sustainable Chemistry Engineering. 7(50), 5134-5141

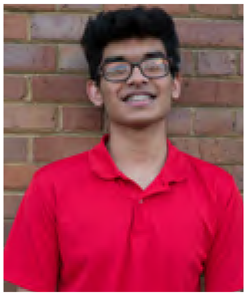
Dankovich, T. & Gray, D. (2015). Bactericidal paper impregnated with silver nanoparticles for point-of-use water treatment. American Chemical Society. 45(5), 1992-1998

Glelick, P. (2002). Dirty Water: Estimated Deaths from Water-Related Diseases 2000-2020. Pacific Institute Research Report. 1(1), 1-2

Krishnaraj, C., Jagan, E. G., Rajaekar, S., Selvakumer, P., Kalaichelvan, P. T., & Mohan, N. (2010) Synthesis of silver nanoparticles using Acalypha indica leaf extracts and its antibacterial activity against water borne pathogens. Colloids and Surfaces B: Biointerfaces, 76(1), 50-56

Pankongadisak, P., Ruktanonchai, U. R., Supaphol, P., & Suwantong, O. (2014). Development of silver nanoparticles-loaded calcium alginate beads embedded in gelatin scaffolds for use as wound dressings. Society of Chemical Industry. 1(1), 1-2

All photos, images, graphs, and charts created by finalist unless otherwise noted.



Shanay Desai

Single-Cell Genomic Profiling of the Adult Human Heart and Transcriptomic Analysis of Differentially Expressed Genes (DEGs) in Myocardial Pathophysiology

Abstract

High mortality rates of cardiovascular diseases (CVD) in humans are contributed to the lack of detailed characterization of the human heart. Largely due to limited access in sampling and overwhelming extracellular diversity in myocytes, a clear transcriptomic profile of the human heart is difficult. Here, Next Generation Sequencing (NGS) techniques were utilized to create a genomic profile from an initial 34,543 young and adult Mus musculus mice cell dataset. With 2554 cells upregulated in the young phenotype (48.5% correlation area) and 2700 cells upregulated in the old phenotype (51.5% correlation area), a total of 255 significantly enriched genes ($p<0.05$) were tested for downstream analysis. Gene set enrichment analysis (GSEA) showed that histone H3 deacetylation, thyroid hormone receptor binding, regulation of circadian sleep/wake cycle, and target of rapamycin (TOR) signaling were the most significantly enriched biological agents and cell-to-cell signaling targeted pathways within the aged heart. Hdac8, Rps6kb1, Nr0b2, Thrap3, and Per3 were also indicated as significant gene groups after performing comparative leading-edge analyses and enrichment map visualizations. As cardiac biologists begin to examine the heart at transcriptomic levels, they gain insights into cellular functionality of gene groups and potential therapeutic strategies useful to cardiovascular disease development, progression, and prevention. This study supplies the foundational set of gene groups and targeted pathways needed for further wet-lab testing.

Keywords: *cardiac aging, transcriptomics, cardiovascular disease, gene set enrichment analysis, false discovery rate, Rps6kb1*

Introduction

Cardiovascular disease (CVD) represents the number one cause of morbidity and mortality in both young and old adults worldwide. Approximately 655,000 humans in the United States die of heart disease each year. One person dies approximately every 36 seconds from a CVD. With the increasing life expectancy patterns, it is estimated that humans born after 2000 will live up to their 100th birthday (Christensen et al., 2009). Since CVD's still represent the leading cause of death nationally, aging poses a significant risk factor for CVD development. In addition, the human genome consists of 2 – 3 % genes that have protein coding functions while the remaining 97 – 98 % are not transcribed into proteins, called noncoding RNAs (ncRNAs). These ncRNAs are divided into two distinct groups: small-ncRNAs and long-ncRNAs. Small-noncoding RNAs (sncRNAs) have transcripts with less than 200 nucleotides. Long-noncoding RNAs (lncRNAs) are categorized as transcripts that have 200+ nucleotides. SncRNAs play important roles in protein scaffolding, epigenetic patterns, and RNA post-translation processing (Bar et al., 2016). However, the long-term functions and characteristics of lncRNAs have yet to be extensively investigated.

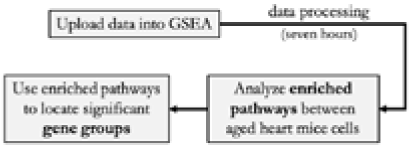
In addition to the human genome, single-cell RNA sequencing (scRNA-seq) has emerged as a major genomic (gene) and transcriptomic (protein) profiling technique. scRNA-seq allows scientists to study ncRNAs with unprecedented detail without knowledge of genetic lineage traces. Additionally, scRNA-seq allows scientists to examine individual cells, as opposed to its predecessor bulk RNA sequencing, which examines specialized cells in groups. scRNA-seq is crucial for targeted therapeutic options and provides a better understanding for cells in their biological context. Evidence suggests that lncRNAs play a major role in the development of aging related disease (Boon et al., 2016). However, few studies have been carried out to explain the role of lncRNAs in relationship to cardiac aging related diseases and locate specific gene clusters. The goal of this study is to identify those uninvestigated cardiac gene groups in myocytes (heart cells) of adolescent and adult Mus musculus using scRNA-seq techniques. This will be useful to determine cell-to-cell signaling pathways and potential biomarkers in the heart, which may explain the importance of aging in CVDs.

Methods

The enrichment analysis model was used to accurately and efficiently characterize, filter, and sort gene groups based on cell-to-cell functionality, normalized expression coefficients, and false discovery rates. To accomplish this, three computational "steps" were carried out:

1. Quality Control (QC): Before analyzing gene expressions in datasets, the "cell barcodes" must correspond to a testable cell within RStudio. If the cell isn't testable and is included in the downstream analysis, the results may be inaccurate. QC was done to ensure that the data is of proper quality to be tested for downstream analysis. The open-source database can be found on National Center for Biotechnology Information (NCBI) Gene Expression #GSE14814. Computer software packages downloaded in RStudio include cowplot, dplyplot, Seurat, and BioConductor.
2. Normalization: Normalization restructures the gene dataset to minimize redundancy. A gene capture in the dataset may contain redundant numbers (cells) due to invariability. Normalization takes these "unformatted" datasets and scales them to obtain correct gene expressions between cells types that is similar across the entire dataset.
3. Gene Set Enrichment Analysis (GSEA): GSEA is a visual analytical tool that is applied to interpreting biological data. The main goal is to understand the shared biological functions that exist, including discovery of the same biological pathway, shared genes and regulators, common cellular compartmentalization, and even association with diseases.
4. Analysis: Using the previous steps and Microsoft software, the significantly expressed genes and important biological pathways can be visualized and analyzed for similarities and differences among cluster groups.

GSEA Data Processing Guide: GSEA was utilized to collect data on cardiomyocytes harvested from three (3) 8-week-old mice and three (3) 18-month-old mice. The following figure shows the data processing outline..



Upload data into GSEA

data processing

(seven hours)

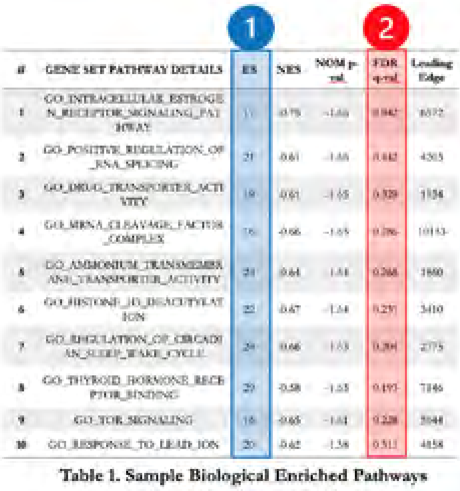
Analyze **enriched pathways** between aged heart mice cells

Use enriched pathways to locate significant **gene groups**

Results

Results of the data processing are shown in Table 1.

v



#	GENE SET PATHWAY DETAILS	ES	NES	NOM p-val	FDR q-val	Leading Edge
1	GO_INTRACELLULAR_ESTROGEN_RECEPTOR_SIGNALING_PATHWAY	11	-0.75	-1.66	0.042	6572
2	GO_POSITIVE_REGULATION_OF_RNA_SPLICING	21	-0.61	-1.66	0.442	4303
3	GO_DRUG_TRANSPORTER_ACTIVITY	19	-0.61	-1.65	0.329	1134
4	GO_MRNA_CLEAVAGE_FACTOR_COMPLEX	16	-0.66	-1.65	0.296	10153
5	GO_AMMONIUM_TRANSMEMBRANE_TRANSPORTER_ACTIVITY	21	0.64	1.61	0.268	1900
6	GO_HISTONE_H3_DEACETYLATION	22	-0.67	-1.64	0.251	3410
7	GO_REGULATION_OF_CIRCADIAN_SLEEP_WAKE_CYCLE	24	0.66	1.63	0.304	2775
8	GO_THYROID_HORMONE_RECEPTOR_BINDING	20	-0.58	-1.55	0.193	7846
9	GO_TOR_SIGNALING	16	-0.65	-1.61	0.228	3044
10	GO_RESPONSE_TO_LEAD_ION	20	-0.62	-1.56	0.511	4858

Table 1. Sample Biological Enriched Pathways (young cardiomyocytes)

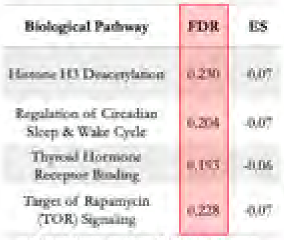
In this research study, two characteristics of this table are important:

1. **Enrichment Score (ES):** The level of degree to which the gene is overrepresented (top or bottom) in the ranked set of genes in the dataset.

2. **False Discovery Rate (FDR):** The estimated probability that the ES represents a false positive finding. In other words, a lower FDR correlates to a higher degree of certainty.

Dataset Output: From the dataset, a total of 34,543 initial features (genes) were analyzed according to their ES and FDR. In young cardiomyocytes, 2,554 / 5,254 gene sets were upregulated in the young phenotype. Four gene sets were significantly enriched with an FDR < 25%. In total, 176 gene sets were significantly enriched at a nominal p-value < 1%. In adult cardiomyocytes, 2,700 / 5,254 gene sets were upregulated in the adult phenotype. Zero gene sets were significantly enriched with an FDR < 25%. In total, 79 gene sets were significantly enriched at a nominal p-value < 1%.

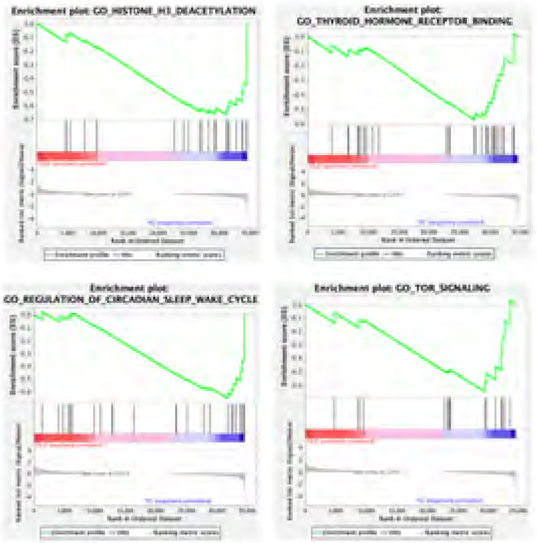
Interpretation of Data: Within the 8-week-old mice, four biological pathways (with an FDR < 25%) were important. These four pathways are shown in Table 2. Each pathway is responsible for a correlated cause to cardiac aging. Furthermore, each pathway contains gene cluster groups that express the specific pathway within the heart.



Biological Pathway	FDR	ES
Histone H3 Deacetylation	0.230	-0.07
Regulation of Circadian Sleep & Wake Cycle	0.204	-0.07
Thyroid Hormone Receptor Binding	0.193	-0.06
Target of Rapamycin (TOR) Signaling	0.228	-0.07

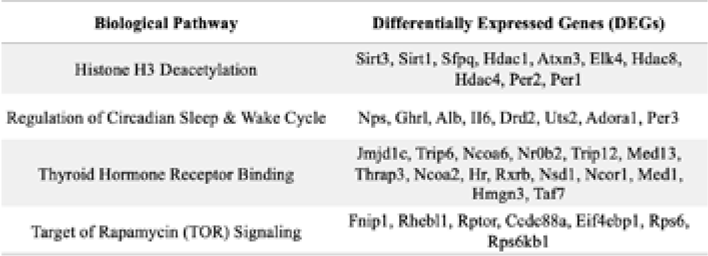
Table 2. Four Significant Cellular Pathways

Enrichment Plot Analysis: Each pathway expressed above can be visualized with an enrichment plot – a unique dataset visualization technique that clearly shows which genes are overexpressed or under expressed in a cluster. Each of the four enrichment plots can be found below.



The following pathways (and subsets of genes) within each enrichment plot were examined using a ranking metric score system. The lowest "point" on the green line correlates to the location of significant gene groups. In all four enrichment plots, the genes are negatively correlated in relationship to their respective enrichment score (ES).

Significant Gene Clusters: Within each enrichment plot, differentially expressed genes (DEGs) are shown. These genes, shown under the green line, have been identified as novel biomarkers for their respective biological pathway. The gene clusters can be found in Table 3.



Biological Pathway	Differentially Expressed Genes (DEGs)
Histone H3 Deacetylation	Sirt3, Sirt1, Sfpq, Hdac1, Atxn3, Elk4, Hdac8, Hdac4, Per2, Per1
Regulation of Circadian Sleep & Wake Cycle	Nps, Ghrl, Alb, Il6, Drd2, Uts2, Adora1, Per3
Thyroid Hormone Receptor Binding	Jmjd1c, Trip6, Ncoa6, Nr0b2, Trip12, Med13, Thrap3, Ncoa2, Hr, Rxrb, Nsd1, Ncor1, Med1, Hmgn3, Taf7
Target of Rapamycin (TOR) Signaling	Fnip1, Rhebl1, Rptor, Ccdc88a, Eif4ebp1, Rps6, Rps6kb1

Table 3. Differentially Expressed Genes (DEGs)

Finally, all significant genes identified here can be confirmed for future wet lab testing and confirmation of existence. Because these gene sets have been previously neglected, future bioinformatics tests are needed to effectively elucidate the role of every gene in comparison to its biological makeup.

Discussion

Several questions were created to efficiently create this data science model:

- Can single-cell techniques be utilized to elucidate the role that DEGs play in the progression of cardiac aging development?
- What cellular pathways contribute the most in cardiovascular disease and cardiac aging?
- What biomarkers and therapeutics can be identified from these shared genes and biological pathways with regards to cardiac aging?

Hdac8, Rps6kb1, Nr0b2, Thrap3, and Per3 were indicated as a few significant gene groups (likely causes of enriched pathways) after performing comparative leading-edge analyses. Beyond useful as a valuable resource tool, this study supplies insights into cellular functionality and therapeutic strategies useful to cardiovascular disease development, progression, and prevention. After analysis, this model was successful in (a) mapping biological pathways, (b) identifying biomarkers, and (c) creating a genomic profile to compare young cardiomyocytes to old cardiomyocytes. The integrity of biological databases creates errors not shown within the data. Sequencing frameshifts, high levels of redundancy, and unstructured data cells are potential sources of error that contribute to invalid results. Data storage, standardization, interoperability, and retrieval are dependent upon more bioinformatics models to reflect accurate findings.

Implications and Recommendations

Although little evidence exists to explain biological compositions and networks within the human heart, our understanding of cardiovascular aging and its homeostasis mechanisms is limited. Cellular heterogeneity has been proven as crucial in pathophysiological processes of other organs; however, this still remains unclear due to many technical and procedural malfunctions. Different organs of the human body have been shown to assist in the human heart function. The need to use single-cell technology on the cellular and biological level is evident in explaining cardiac aging as a whole because it provides a comprehensive understanding of the cellular heterogeneity within the adult human heart.

As evidenced by the United Nations Sustainable Development Goal 3.4, the target goal is intended to "reduce mortality of non-communicable diseases by one third through better treatments" (Gordon, 2019). This goal aligns with the overall Sustainable Development Goal 3, which is intended to "ensure healthy lives and promote well-being for all at all ages." Better scRNA-seq methods combined with new knowledge of biologically important cells that contribute to these non- communicable diseases will be the next step in reaching Goal 3.4 specifically and Goal 3.

Future Research

The clear resolution imaging and rich data provided from scRNA-seq continues to attract cardiac biologists to understand the intracellular networks, epigenetic regulators, and potential heart defects that cause these disease to develop (Gladka et al., 2018). Using these biologically important understandings, novel targets for future therapeutic opportunities would be possible. As we progress in the twenty-first century, scRNA-seq will become more mainstream as costs decrease and widely distributable kits are available for the public. Personalized medical applications and further specific research with scRNA-seq technology will provide biologists with a closer end to these detrimental diseases, including cardiovascular diseases.

References

Bär, C., Jesus, B. B., Serrano, R., Tejera, A., Ayuso, E., Jimenez, V., . . . Blasco, M. A. (2014). Telomerase expression confers cardioprotection in the adult mouse heart after acute myocardial infarction. Nature Communications, 5(1). doi:10.1038/ncomms6863

Boon, R. A., Hofmann, P., Michalik, K. M., Lozano-Vidal, N., Berghäuser, D., Fischer, A., Knau, A., Jaé, N., Schürmann, C., & Dimmeler, S. (2016). Long Noncoding RNA Meg3 Controls Endothelial Cell Aging and Function: Implications for Regenerative Angiogenesis. Journal of the American College of Cardiology, 68(23), 2589–2591. https://doi.org/10.1016/j.jacc.2016.09.949

Christensen, K., Doblhammer, G., Rau, R., & Vaupel, J. W. (2009). Ageing populations: the challenges ahead. Lancet (London, England), 374(9696), 1196–1208. https://doi.org/10.1016/S0140-6736(09)61460-4

Gladka, M. M., Molenaar, B., de Ruiter, H., van der Elst, S., Tsui, H., Versteeg, D., Lacraz, G., Huibers, M., van Oudenaarden, A., & van Rooij, E. (2018). Single-Cell Sequencing of the Healthy and Diseased Heart Reveals Cytoskeleton-Associated Protein 4 as a New Modulator of Fibroblasts Activation. Circulation, 138(2), 166–180. https://doi.org/10.1161/CIRCULATIONAHA.117.030742

Gordon, L. (2019). Prevention and Treatment of Non-Communicable Diseases to Achieve a Reduction of Premature Mortality (SDG 3.4): Meta-Narrative Review of Global Strategies. SSRN Electronic Journal. doi:10.2139/ssrn.3294760

Raeed Kabir

Vidhi Patel

Analysis of the Differential Impacts of Material and Social Stressors on Mental Health during the COVID-19 Pandemic

Abstract

Depression cripples a large percentage of the population, and the novel coronavirus pandemic has acted as an additional stressor, further exacerbating a decline in mental health. To effectively combat this, more information surrounding the virus' impact on mental health is vital. The virus has introduced two substantial issues, affecting even those who have not contracted the virus: a disruption in the economy that results in large scale unemployment and loss in income, as well as laws that have limited the sociability of individuals. This project acknowledges the presence of COVID-19 and examines financial, isolative, and emotional factors as potential culprits for the rise in depression. Our project follows a "horse-race" model, where financial stressors are tested relative to limitations in sociability to see which dimension of life is more predictive of having depressive symptoms.

Introduction

The first case of 2019-nCOV (COVID-19) in the United States was reported in early January of 2019 (Holshue et al., 2020). This relatively novel virus has taken a toll on the mental health of many Americans, tripling the prevalence of depression (Ettman et al., 2020). However, a limited number of studies have been done around the causes for this. These studies focus on financial wellbeing as the primary stressor in peoples' lives, but this study expands on this literature by using more explanatory variables. COVID-19 brought social distancing laws (Hwang et al., 2020) and began to deteriorate the economy (Ibn-Mohammed et al., 2020), so financial wellbeing and social isolation were the two primary variables tested. The research question of this study then becomes the following: which of these factors can best explain the rise in depression across the nation. Following a "horse-race" research method (Davidson & MacKinnon, 1981), this study compares two major variables for their predictive power for depression: finance and isolation.

Methodology

Survey Instrument:

Fifty-two original questions were sent in a survey through a variety of mediums, including Facebook, a satisfactory method of data collection (Boas et al., 2020). Additional platforms include LinkedIn, Twitter, and Email. There were questions from three main dimensions: economic, emotional, and isolative stressors. This study controls for various factors that may otherwise conflate the data such as gender, education, and ethnicity. Because of the timeline of the project, participants were asked if they believe they suffer from seasonal affective disorder ("Seasonal Affective Disorder," 2020), otherwise known as seasonal depression, to prevent conflation in findings. The survey measured mental health with the Patient Health Questionnaire (PHQ-8), an instrument used to gauge depressive symptoms (Kroenke et al., 2009).

PHQ-8 Sensitivity Analysis:

Two surveys were sent out through various forms of media: one survey with five entertaining graphical images placed every ten questions and one without. There was no correlation between the presence of entertaining images and the PHQ-8 score, as the regression coefficient was -0.2410 with a large p-value of 0.785. A p-value < 0.1 is to be significant for the remainder of this study.

Summary of Descriptive Statistics:

There were 143 participants who took the survey from mid-December to early February. All analyses use this sample size. The target population is adults with financial responsibilities, so the average age of the sample is 44.18 years.

Demographics: A majority of participants' race was Caucasian or White (63.4%) and African American or Black (26.76%). Participants were mostly female (78.3%), with males being (21.7%). Most resided in Mississippi (79.9%) and a small amount in Kentucky (4.20%). The education levels were as follows: High School Diploma or Equivalent, 8.4%; Bachelor's Degree, 34.2%; Master's Degree, 35.7%; JD/MD/Ph.D., 18.2%; Other Terminal Degree, 3.5%.

Fiscal Statistics: The average income per year for 2019 (pre-tax) for our sample population was $83,356. For 2020 (pre-tax), the average was $85,944. 26.6% of the population was low income (< $40,000), while 25.2% was high income (> $100,000). Isolation Statistics: The average person has 3.84 conversations in person per day in the last week. 53.15% of individuals believe they have more virtual interactions now than they did before the pandemic, and 67.83% of the population says that they had more in person conversations before COVID-19.

Mental Health Statistics: The depression rates of the population were gauged using the PHQ-8, where scores could range from 0 to 24. 78.32% of participants were not depressed (score < 10); 20.43% participants scored in the range of major depression (≥ 10); and 0.70% of participants scored severe major depression (≥ 20). Additionally, 26.57% of individuals report that they suffer from SAD, and another 12.59% of individuals are not sure if they do.

Contemporaneous Study:

The first set of analyses used more objective measures of depression (PHQ-8), finance, and interactions. These measures tested current levels of depression, 10 months after the pandemic began.

Index variables are comprised of several other variables and are used in this study to summarize the predictive power of both dimensions: financial status and isolation. Increasing the value of an index maps with higher stress in said dimension. The constituent parts to each index largely dictates the results of the regression, so there are two models, where the indices are built differently.

Multivariate Regression Model:

In the first regression, some variables are chosen to become a part of the indices based on how representative they are. Gender and seasonal depression were thought of as meaningful controls.

$$Y_i = \alpha + \beta_1 Finance_i + \beta_2 Isolation_i + \beta_3 SAD_i + \beta_4 Gender_i + \varepsilon_i$$

$$Finance_i = Finance\ Index\ Variable$$

$$Isolation_i = Isolation\ Index\ Variable$$

$$Gender_i = Gender\ Binary$$

$$SAD_i = Seasonal\ Depression\ Binary$$

X_i	β	p-Value
Finance Index Variable	+10.90***	0.00
Isolation Index Variable	+1.57	0.371
Female Binary	+1.01	0.250
Seasonal Depression Binary	+2.17***	0.008

* = p < 10%; ** =p< 5%; *** =p< 1%

Machine Learning Model:

To eliminate subjectivity, the LASSO algorithm is a unique machine learning model in that it allows unrelated variables to be zeroed out completely. New indices were built without bias and new controls were used.

$$\frac{1}{2N}(y - X\beta^{*})'(y - X\beta^{*}) + \lambda\sum_{j=1}^{p}|\beta_j|$$

$$Y_i = \alpha + \beta_1 Finance_i + \beta_2 Isolation_i + \beta_3 SAD_i + \beta_4 Education_i + \varepsilon_i$$

$$Finance_i = Finance\ Index\ Variable$$

$$Isolation_i = Isolation\ Index\ Variable$$

$$Education_i = Level\ of\ Education$$

$$SAD_i = Seasonal\ Depression\ Binary$$

X_i	β	p-Value
Finance Index Variable	+4.36***	0.000
Isolation Index Variable	+1.79**	0.049
Level of Education	-0.49*	0.090
Seasonal Depression Binary	+2.20***	0.005

* = p < 10%; ** =p< 5%; *** =p< 1%

Key Findings: In Model 1, finance has a large coefficient of regression with a significant p-value. The finance index is 7 times more predictive than isolation. In Model 2, finance is 3 times more predictive, supporting the first model.

Event Study:

This begins an analysis on past cases of depression and their relationship to COVID-19, in contrast with the contemporaneous study. For each of the three dimensions – finance, emotion, and isolation – participants were asked to rank the worst five months in the year 2020. The frequency of Mississippi residents ranking any given month as the worst was then graphed on the vertical axis.

The second graph superimposes the COVID-19 death numbers for Mississippi in 2020 on top of Graph 1. All data was gleaned from the Mississippi State Department of Health ("Provisional Death Count," 2021).

Key Findings: For Graph 1, the third month shows a sudden shock that stressed all dimensions of wellbeing. March was also the first month COVID-19 became an issue in Mississippi. There is little variation in finance, but emotional wellbeing seems to track closely with isolation. For Graph 2, people become desensitized to the stressors of the pandemic after some time. Sociability begins to peak as deaths peak in August.

An additional exploration is done to analyze the change in income across years. Graph 3 is a kernel density plot that shows these findings.

Key Findings: The average income from 2020 was about $2,500 dollars higher, proposing the average individual is doing better financially, but few increased and their jumps were so significant that this offset the declining wellbeing of the remainder of the population. This study found two types of displacement: upwards displacement was accompanied by an average of 0 weeks of unemployment, and downwards unemployment was accompanied by an average of 5.9 weeks of unemployment.

The high-income group (> $100,00) averaged 1.78 weeks unemployed. The low-income group (< $40,000) averaged 3.63 weeks unemployed. However, the average participant is 2.74 weeks unemployed.

Time Series to Isolate an Exogenous Shock:

Next, the study used time series analysis to visualize relationship between wellbeing and each month as an independent variable.

The results are in the table below, where the likelihood of January being ranked the worst month became the baseline, and any change in this likelihood of being ranked the worst is given in this table. For example, July was 11.9% more likely than January to be ranked the worst month for emotion.

Key Findings: February was statistically the same month as January. March introduced a stress that had a causal effect on emotional and isolative health. The presence of COVID-19 could be said to have caused a decline in mental health. Emotions covary more greatly with isolation in the past. The population does begin to become desensitized to the pandemic.

Multivariate Time Series Regression:

In order to draw a perfect parallel between this study and the contemporaneous study, the study runs a similar regression at the individual level, i, but at discrete steps in time, months. This will absorb the time-variant relationship of emotions over the course of the year.

Key Findings: Isolation is 3 times more likely than finance to predict deteriorated mental health in the early stages of the pandemic (opposite of the contemporaneous study).

Conclusion:

Both financial and isolative stressors are valuable predictors for depression. The differences in these findings are not a result of bad data, but they highlight the differences between the studies.

The contemporaneous study finds financial stress the greatest predictor of depression, and the past studies find isolative stress as the greatest predictor earlier on. It is understood that COVID-19 does something in the world that directly hurts the emotional health of the people, but it is not the virus itself, as the number of deaths does not track with emotion. Additionally, the potency of the exogenous shock wears off and sociability become restored as people become desensitized to the pandemic. The reasons for being depressed likely shift. After this shift, it is entirely possible that silent changes in financial wellbeing have accumulated across the year and have become more predictive of being depressed 10 months later.

Thus, there is not one "horse-race" winner. Because the underlying mechanisms of depression are more nuanced than expected, there are two "horse-race" winners at different moments in the pandemic.

This study further elucidates univariate relationships between lifestyle and depression. It is found that virtual interactions like video call and text message have no effect on wellbeing, but interactions in person are quite significant. It is also notable that those that are on Food Stamps and unemployment benefits are likely to have a 5-7 point increase in PHQ-8 scores because of their financial stress. Further resources must be available for at-risk groups.

Limitations:

1) The PHQ-8 test is not a test that diagnoses depression, but it is a strong indicator for depressive symptoms. 2) The time series is reserved for Mississippi-only residents, and the contemporaneous study contains 20% of non-Mississippi residents. Still, the LASSO algorithm verifies that state residence is not a strong predictor of depression. 3) Because of the size of the study (143 participants), coefficients of regression and p-values may be lower or higher than they should be. 4) The sample is widely Caucasian, high-income individuals. There are no accurate analyses on ethnic impacts on depression because of this. 5) The studies based on ranking questions are subject to memory loss and perception. Although this is not objective, peoples' perception of their own wellbeing is not trivial. Understanding when people believed they suffered is quite important.

Implications:

Using regressions, this study highlights key aspects of life related to depression, so institutions can reallocate resources with policy to quell the surge of depression during current or future pandemics.

Graph 6

Being financially vulnerable and socially isolated had enormous effects on wellbeing. Assuming PHQ-8 scores did not change from December 2020 to when most participants took the survey (January/February 2021), a proportion was used to calculate what the average PHQ-8 score likely would have been during the peak month for depression, April 2020.

The average individual in the population would have roughly had a PHQ-8 score of 14.53. This is halfway between major depression and severely major depression for the average individual. People in the U.S. suffered severely during the COVID-19 pandemic, and aid should be administered to at-risk individuals, regardless of where this stress came from.

References:

Davidson, R. and MacKinnon, J. (1981). "Several Tests for Model Specification in the Presence of Alternative Hypotheses" Econometrica, 49(3), 781-793. https://doi.org/10.2307/1911522

Ettman, C., Abdalla, S., Cohen, G., Sampson, L., Vivier, P., Galea, S. (2020). Prevalence of Depression Symptoms in US Adults Before and During the COVID-19 Pandemic. JAMA Network Open, 3(9), Article 2019686. doi:10.1001/jamanetworkopen.2020.19686

Holshue, M., DeBolt, C., Lindquist, S., Lofy, K., Wiesman, J., Bruce, H., Spitters, C., Ericson, K., Wilkerson, S., Tural, A., Diaz, G., Cohn, A., Fox, L., Patel, A., Gerber, S., Kim, L., Tong, S., Lu X., Lindstrom, S., … Pillai, S. (2020). First case of 2019 novel coronavirus in the United States. The New England Journal of Medicine, 382(10), 929-36. https://www.nejm.org/doi/full/10.1056/NEJMoa2001191

Hwang, T., Rabheru, K., Peisah, C., Reichman, W., Ikeda, M. (2020). Loneliness and social isolation during the COVID-19 pandemic. International Psychogeriatrics, 32(10), 1217-1220. https://doi.org/10.1017/S1041610220000988

Ibn-Mohammed, T., Mustapha, K., Godsell, J., Adamu, Z., Babatunde, K., Akintade, D.D., Acquaye, A., Fujii, H., Ndiaye, M.M., Yamoah, F., Koh, S.C.L. (2021). A critical analysis of the impacts of COVID-19 on the global economy and ecosystems and opportunities for circular economy strategies. Resources, Conservation, and Recycling, 164, Article 105169. doi: 10.1016/j.resconrec.2020.105169

Kroenke, K., Strine, T., Spritzer, R., Williams, J., Berry, J., Mokdad, A. (2009). The PHQ-8 as a measure of current depression in the general population. Journal of Affected Disorders, 114(1-3),163-173. https://doi.org/10.1016/j.jad.2008.06.026

Seasonal Affective Disorder (SAD). (2020). American Psychiatric Association. https://www.psychiatry.org/patients-families/depression/seasonal-affective-disorder

Provisional Death Count, 2020-2021 (2021). Mississippi Department of Health. https://msdh.ms.gov/msdhsite/_static/resources/10682.pdf

Acknowledgements:

Dr. Reshmaan Hussam is thanked for her passionate guidance and invaluable enthusiasm as the mentor of this project.

Hayden Anderson

Transplantation of TNF/FasL Ligands and TNFR/FAS Receptors: A Study of the Effects of Agrobacterium Tumefaciens on Wisconsin Fast Plants

Abstract

"How can we cure cancer" has been a central question for many medical professionals for decades. However, besides radiation treatment and chemotherapy, these professionals have not been able to find an easy cure to cancer. That is why, for this project, the question was changed a more specific alternative, "How can we detect cancer successfully?" Wisconsin Fast Plants, a type of plant that grows exponentially compared to the regular plants, were grown as a control group and a test group was also grown. The test group was then inoculated with Agrobacterium tumefaciens, a plant alternative to cancer that quickly grows tumors in the stem which dehydrates the leaves and ultimately causes death. The plants were then inoculated and observed to see the effects of the bacterium on the plant. After they went through their life cycle, the leaves of the plant were observed under a light microscope. The control group had the normal cells that any other plant would; however, the plants with the cancer seemed to have signs of plant apoptosis. It was then concluded that if scientists were able to use this apoptosis to kill the cancer cells, then they would be able to use this for treatment of cancer. Using scientists' knowledge of experimental mitochondrial transplantation, it can be concluded that in the near future, scientists will be able to transplant not only organelles from cells, but also the receptors and cells that cause apoptosis to induce apoptosis into cancer cells.

Introduction

In 2020, approximately 1.8 million people were diagnosed with cancer in the United States. With different cancers ranging from breast to colon cancer, this deadly disease kills millions each year according to the National Cancer Institute. Not only this, but there are common denominators of the side effects of treatment such as anemia, pain, and infection that sometimes make treatment worse than the actual disease (American Cancer Society, 2020). However, over the recent years, scientists, and doctors all over the world have been seeking for better treatments that will not only help the pain of cancer treatment, but also survival rates of cancer all together. One research topic that medical professionals have been studying is induction of apoptosis into cancer cells (Ucker & Levine, 2018). Apoptosis is the programmed death of a cell due to internal deoxyribose nucleic acid (DNA) damage inside of the cell. Cancer results by too much apoptosis or a lack thereof in said damaged cells. Thus, the cancer cells spread and grow tumors. However, apoptosis is actually very important in treatment of cancer. Due to how important (and deadly) apoptosis can be, it has been one of the most researched topics by cell biologists over the years. More specifically, cell biologists have studied three main biochemical changes in apoptosis: caspases, the breakdown of proteins and DNA, and changes in the membrane/recognition by phagocytic cells (nigms.nig.gov, 2020). These cell biologists and other scientists have also started experimenting with mitochondrial transplantation in which a mitochondria "floats" to where it's supposed to go in the cell (Elliott et al., 2012). With this, a plant bacterium that is relatively similar to cancer has been used to research these concepts such as apoptosis and cancer that is called Agrobacterium tumefaciens. As these concepts are being developed and studied, the purposes of this project are as follows: first, to better understand how apoptosis and cancer works in the human body, second, to see how cancer directly affects cells through inoculation of agrobacterium tumefaciens in the leaves of a Wisconsin Fast Plant, and third, to incorporate this project and past research to show that we could possibly transplant the apoptosis ligands (TNF and FasL) and its receptors (TNFR and Fas) in order to help find a better treatment for cancer. It was hypothesized through this experiment that the Wisconsin Fast Plants would have a higher death rate and would be affected negatively if inoculated with the agrobacterium tumefaciens.

Methods

Growing Plants: The plants were first planted in November and lasted throughout December. They were planted on a water reservoir (a bucket with water in it and a cloth mat) to simulate the water that is formulated in the ground. The plants were then planted using potted mix, quads, and other miscellaneous planting items. Three different quads were planted – two for control and one for inoculation. After around a week of them growing, they were pollinated the plants with dried honeybees. This was done using one bee per plant in order to decrease cross pollination. They then grew for another week until they were inoculated with the bacterium.

During and After Inoculation: The leaves on the plant were inoculated with the bacterium, Agrobacterium tumefaciens. The plant absorbed the bacterium via a painting method of inoculation to the bottom on selected leaves of the plant.

Diagram 1 and 2:

Plants Directly After Inoculation

Close Up of Leaves after Inoculation

After inoculation, plants were placed under a grow light to help regulate the temperature of the plants. This also would also give the plants light at all times. After around two (2) weeks, the plants showed major deterioration and they were put underneath a light microscope in which the effects of the bacterium on the cells of the plant were looked at it. This way, the effects of the plant could be looked at that most people would not see to the human eye.

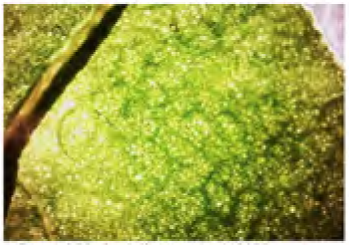
Results

Results Without a Microscope: Without a microscope, the control group looked like normal plants. They had regular veins, green leaves, and were healthy plants in general. However, there were many symptoms from the bacterium in the test group. First, the test group began to deteriorate faster than the control group. This was shown through the blackening of the leaves and flies that flew around the plant. Second, the bacterium caused many of the leaves to dehydrate due to the tumors that grew inside of the plant (shown through the droopy and yellow leaves). With some of the leaves dehydrating, other leaves absorbed more water than others and caused the veins to be enlarged in the plant.

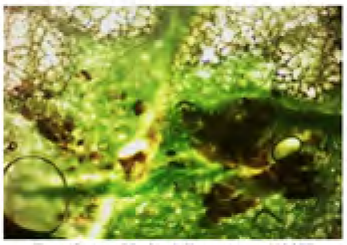
Leaves deteriorating from the test group

Enlarged veins from the test group

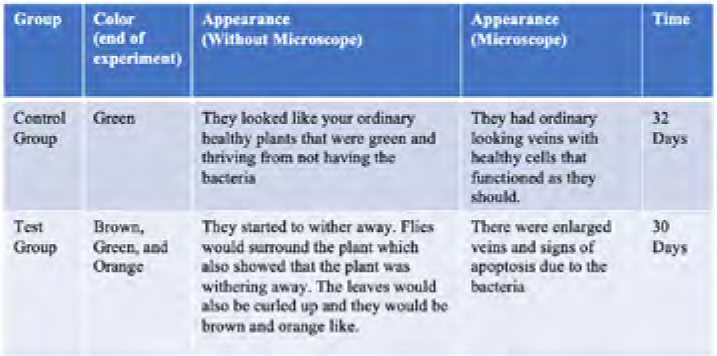
Results Underneath the Microscope: Underneath the microscope, the control groups looked like normal plant cells: alive and working cells, regular veins, and an overall healthy plant. In the test group, there were enlarged veins (as shown in the "results to the human eye" section) that were caused by overhydration from the bacterium. However, the most important factor in the test group was how the cells reacted to the bacterium. The cells inside of the plant went through programmed cell death, otherwise known as apoptosis. The apoptosis in these plants were shown through black and brown spots on the cells of the plant.



Control Under Microscope (400X)



Test Group Under Microscope (400X)



Group	Color (end of experiment)	Appearance (Without Microscope)	Appearance (Microscope)	Time
Control Group	Green	They looked like your ordinary healthy plants that were green and thriving from not having the bacteria	They had ordinary looking veins with healthy cells that functioned as they should.	32 Days
Test Group	Brown, Green, and Orange	They started to wither away. Flies would surround the plant which also showed that the plant was withering away. The leaves would also be curled up and they would be brown and orange like.	There were enlarged veins and signs of apoptosis due to the bacteria	30 Days

Discussion

The results of this experiment showed that there is a correlation between apoptosis and cancerous cells. Throughout this project, three major factors were observed: what the plants looked like, how fast the plants were dying, and what the cells looked like underneath the microscope.

As expected, the Agrobacterium tumefaciens caused negative effects on the test group; however, the results were not as expected. Apoptosis in cells was not expected since the mistake was not directly inside of the plants, but instead the bacterium was inoculated. However, with this, it is able to be concluded that whether or not cancer happens from an excessive amount of apoptosis or lack thereof, there is apoptosis inside of the cells that cancer happens most of the time. It cannot be concluded that apoptosis happens one hundred percent of the time; however, it can be hypothesized that it does happen most of the time-based studies and research. From this, it can be concluded that apoptosis is one of the major symptoms of cancer and scientists may be able to use this to their advantage using professional research.

Implications and Recommendations

Apoptosis is taught as programmed cell death any many biology classes; however, many do not talk about what happens behind the scenes. Apoptosis is the self-destruction of a cell that happens due to a mistake in the DNA. Due to the cell not wanting to replicate the incorrect DNA. However, sometimes cells make another mistake and do not go through apoptosis which leads to tumors. The symptoms make some of the major causes of cancer are broad; however, it can be concluded that is the mutations inside of the cells and the apoptosis that occurs because of these cells are very common causes of cancer. Before exploring the possibilities of using apoptosis in cancer cells, background information must be included. Apoptosis happens when two ligands named tumor necrosis factor (TNF) and FasL link to TNFR and FAS receptors. These receptors then release instructions into the cell that tell the cell to induce apoptosis. With this information, it can conclude that cancer happens in cells for many reasons, but apoptosis is one of the most notable features here. With this information, it can most notably be concluded that apoptosis happens in cells and it can be questioned if induction of apoptosis in cancer cells is possible. It is more noticeably possible due to the advancement of research that has been given on many different areas of the cell. Recently, medical professionals have been researching mitochondrial organelle transplantation in humans in order to help with mitochondrial failure, which results in problems with homeostasis. Using this information, it is totally possible to transplant many cellular items such as organelles, receptors, ligands, etc. If it were possible to use experimentation and research of transplantation of these cellular items, then it is also possible to cure cancer using induction of apoptosis in these damaged cells. How the cure could happen is through apoptosis into cancer cells is through transplantation of TNFR/FAS receptors into the cancerous cells (Wong, 2011). Through this, scientists and researchers would also have to make sure to have a way to include ligands inside of the body in order to provide a way for the receptors to send instructions inside of the cell to induce apoptosis. If transplanted receptors were injected into the cancerous cells, then it may help the well-being of the world since cancer is one of the leading causes of death in that nation.

Future Research

Through this research, the medical field could firstly, give a more successful treatment for cancer due to making treatment easier to live with than the alternative, that of radiation. Second, this information could be utilized in many other neurodegenerative diseases such as Alzheimer's and Parkinson's disease. Third, with the information of apoptosis that would be learned from this research, cell biologists would be able to learn more about how apoptosis occurs and how cells react to apoptosis.

References

Cancer Facts & Figures 2020, American Cancer Society (ACS), Atlanta, Georgia, 2020.

Elliott RL, Jiang XP, Head JF. Mitochondria organelle transplantation: introduction of normal epithelial mitochondria into human cancer cells inhibits proliferation and increases drug sensitivity. Breast Cancer Res Treat. 2012 Nov;136(2):347-54. doi: 10.1007/s10549-012-2283-2. Epub 2012 Oct 19. PMID: 23080556.

Studying cells. (n.d.). Retrieved April 01, 2021, from https://www.nigms.nih.gov/education/fact-sheets/Pages/studying-cells.aspx

Ucker DS, Levine JS. Exploitation of Apoptotic Regulation in Cancer. Front Immunol. 2018;9:241. Published 2018 Feb 27. doi:10.3389/fimmu.2018.00241

Wong, R.S. Apoptosis in cancer: from pathogenesis to treatment. J Exp Clin Cancer Res 30, 87 (2011). https://doi.org/10.1186/1756-9966-30-87

Effect of Caffeine on the Life Span and Reproduction Rate of *Drosophila melanogaster*

Victoria Callahan

Abstract

The purpose of this experiment was to test the effect of long-term caffeine usage on Drosophila melanogaster. The effect on their lifespan and reproduction rates were studied. Caffeine was administered to the F1 generation of flies in amounts of one and two drops through their food. The null hypothesis regarding lifespan was rejected using a chi square calculation. Data showed that Group A, the group with only one drop of caffeine, had a longer lifespan than either of the other two groups. Group B, two caffeine drops, also had a longer lifespan than the control. This data was analyzed by combining the data from different vials of flies given the same treatment and taking the average, standard deviation, range, and median of that data. The null hypothesis for reproduction rate was also rejected using a chi square calculation. Data showed that Group A had later transfer dates and Group B generally had earlier transfer dates than the control. This data was analyzed by averaging the days since the vial was created for each group (control, A, and B). This research helps to promote the Sustainable Development Goal 3: Good Health and Wellbeing for all People. Further research applying this knowledge to humans can also help to aide medical professionals in making recommendations regarding caffeine to patients.

Introduction

Caffeine is the most widely used of all psychoactive drugs and has become an integral part of cultures around the world (Wu et al., 2009). The average daily consumption of caffeine in the US and Canada is 210-238mg/day. Recently, caffeine has begun to be seen as "a model drug of abuse" and the use of caffeine has gained more attention. Caffeine is absorbed quickly when ingested and there is no blood brain barrier in either adults or fetuses (Fredholm et al.,1999). D. melanogaster's lifespan can be altered by various stressors, such as the quality of their living environment. When they are no longer able to respond to stressors, their physiological balance cannot be maintained, and they eventually die. Caffeine can fragment the sleep cycle of D. melanogaster and disrupt their central biological clock. Disruption of a D. melanogaster's circadian clock can increase the chances of obtaining diseases that are associated with ageing. (Wu et al., 2009). A fly's diet can also have a significant effect on its lifespan (Halim et al., 2020). Under optimal conditions, D. melanogaster may live as long as 56 days (Flagg, n.d.). A study about the effects of long-term caffeine usage showed dramatic adaptations that did not necessarily align with those found when caffeine was used in the short term. These changes were not all detrimental and were in some cases beneficial (Fredholm et al., 1999). Male and female fruit flies have specific identifying characteristics which were used throughout this study. Most notable was the male's underside, which is much darker than the females along with being more rounded (Flagg, n.d.). This study will look at the effect of caffeine on the lifespan and reproductive rates of Drosophila melanogaster.

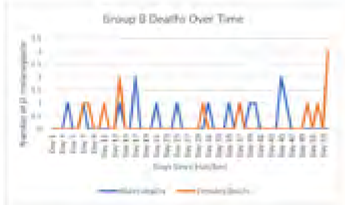
Methodology

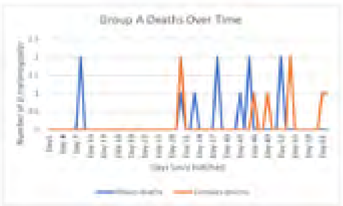
Procedures:

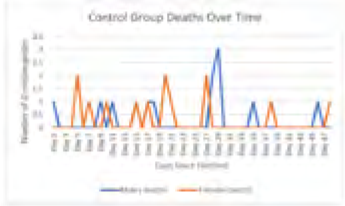
1. Obtain an F1 generation from a wild type parental generation
2. Check F1 vial every day for hatched larvae and if possible, create a group that includes roughly 10 flies and is a mix of male and female flies. Continue this until multiple vials have been created for the control group, one caffeine drop group(A), and 2 caffeine drop group(B)
3. Label the vial with the date of creation and group name(ex. Test Group 1A or control group 2)
4. To mix up the food in the tube add the proper amount (either 0,1, or 2 drops) of caffeine to 23.4ml of water stir mixture. Then add equal amounts of Instant Drosophila medium and caffeine and water mixture to tube. Sprinkle a little yeast on top.
5. Everyday record the number of living and dead flies in each group in lab notebook as well as their genders until most flies in each group have died. Remove dead flies from tube.
6. When larvae appear in the tube with one of the groups, transfer flies to a new tube with food of the same caffeine concentration as they were previously in. Record the date of transfer.

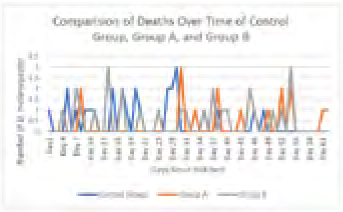
To analyze the data regarding lifespan, the data from each tube (under the same conditions) was combined to make one combined data set for each of the three conditions: control group, Group A, and Group B. The average lifespan, range, median, and standard deviation was calculated. Data regarding reproduction was averaged. The data for the day each transfer took place was also averaged with the day for the equivalent transfer in all the other tubes put under the same conditions. If a group did not have multiple tubes with data, no average was taken. Missing data was marked as an error and accounted for by standard deviation. Finally, a specific first transfer average and standard deviation was taken using the same procedure as stated above. The flies were considered to have reproduced when they were transferred due to the presence of larvae.

Data Analysis and Results







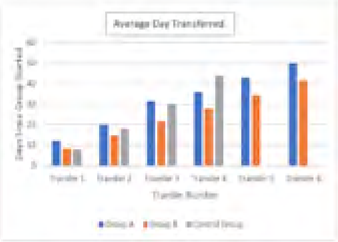


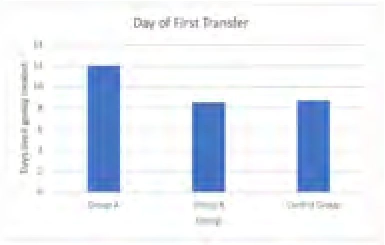
Interpretation of Results

It is observed within the results for the three test groups that both caffeine groups had longer average lifespans with Group A having a much longer lifespan then either Group B or the control group by 7.4 12.9 days for males and 16.8 and 28.3 for females respectively. Both caffeine groups had a longer range of lifespan than the control group. Group B death rates are evenly distributed over the timeline showing that the flies died consistently over time, while Group A death rates are skewed left showing that they had longer lifespans on average, and the control groups deaths are skewed right showing that they had shorter lifespans on average. Death occurs when a fly is no longer able to respond to stressors and their physiological balance can not be maintained (Wu et al., 2009). The lifespan averages show that the two groups with caffeine had extended lifespans and were thus able to respond to the different stressors better than the control group flies.

With the control group, the average male lifespan was greater than that of the females while the opposite holds true for both Group A and Group B. Group A female average lifespan was significantly higher than any other lifespan at 48.3±12.5 days.

Each group (control, A, and B) originally had three tubes consisting of 10 flies each and the data from these tubes would be combined to give data for all three groups. However, this did not go as planned due to flies escaping during tube transfers. These flies were noted and taken out of the data and calculations, as no death date could be determined for them. Group A ended up with only one tube of data due to the majority of flies in the third tube escaping.





Interpretation of results

There is little difference between the average day of the first transfer due to larvae of the control group and Group B. There is a difference between the control group, Group B, and Group A. Group A has a much later date at 12±4.2 days after the group was first created. Group A also has a much higher standard deviation than the control group and Group B.

Looking at the average day each group was transferred due to larvae, Group B consistently transferred earlier than both Group A and the control group. The exception to this is with the first transfer where the Group B average day is 8.5 which is 0.5 higher than the control group average. Group A is consistently transferred later than both Group B and the control except for transfer four.

Both group A and B have six transfers and produced larvae consistently with the transfer days, stretching out towards the end of the lifespan range for each group. Group B's last transfer occurs nine days before the maximum lifespan observed in this study. While another transfer would be expected within this time frame, the lack of one can be attributed to the tendency for Group B males to die faster than Group B females. This leaves only females in the test vial near the end. The control group only had four transfers and can be attributed to the shorter lifespans seen within those flies with no difference in reproductive habits.

There was missing data for the transfer days particularly within the Group A data set. These would be due to human error in recording the data and should be mitigated by standard deviation. There were also a few transfers made for reasons other than larvae appearing within the tube, such as a spider appearing within the vial. This error was accounted for in the data with standard deviation and can be prevented in the future by raising flies in a non-home environment.

Discussion

The null hypothesis regarding lifespan data will be rejected because the p value of 54.2 is greater than the chi square critical value of 9.488. This shows that the continuous consumption of caffeine over time does influence the lifespan of D. melanogaster. The fact that both amounts of caffeine extended lifespan would be supported by the research done by Fredholm et al.(1999) and his colleagues, showing that the effects of long-term caffeine usage were not always negative.

The null hypothesis for the caffeine's effect on reproduction rate is rejected because the p value of 26.3 is greater than the chi square critical value of 11.070. While the differences seen between the control group and the A and B group transfer rates were small, there was still a slight difference. Small alterations of D. melanogaster's genetic code can significantly alter their behaviors (Benzer & Konopka, 1971). It is possible that something similar to this was happening to a small extent. It is also possible that the caffeine was reaching the brain of the D. melanogaster causing this to occur. It is known that there is no blood brain barrier to caffeine (Fredholm et al.,1999).

Implications and Recommendations

This research can help promote the Sustainable Development Goal 3: Good Health and Wellbeing for all People. This research shows long-term caffeine usage can lengthen the lifespan of Drosophila melanogaster. This is important information for medical researchers going forward as patients are continuing to ingest high amounts of caffeine. D. melanogaster and humans have more than 50% of gene homologous (Halim et al., 2020). This implies that it is possible that the same could apply to humans. While this does not provide enough research for medical professionals to begin making recommendations regarding caffeine usages, it does provide evidence that they may need to worry less about daily caffeine consumption by patients. The research showed small amounts improved the lifespan length, although the decrease in lifespan length shown in Group B compared to Group A does suggest that there is a cutoff where the effect is no longer seen or could potentially become harmful. The information regarding reproduction is also important as it suggests a possible slowdown in reproduction for D. melanogaster. As stated above, this could be transferred to human studies as future researchers could include in an investigation of caffeine's effects on humans.

Future Research

For future research, an empirical study of caffeine users of different types could be conducted to see if this research transfers to humans in their everyday lives. It could work as a series of questionaries regarding frequency and amount of caffeine consumption and follow the participants for x number of years from the start of the study. More research can also be conducted regarding reproduction in D. melanogaster with a larger sample size and more data to ensure the small sample size did not significantly affect the results.

References

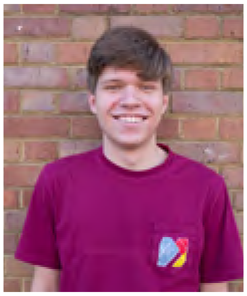
Benzer Seymour, Konopka Ronald J., R. J. (1971). Clock Mutants of Drosophila melanogaster. Proc. Nat. Acad. Sci. USA Vol. 68, No. 9, , 2112-2116.

Flagg, R. O. (n.d.). Carolina Drosophila Manual. Burlington, North Carolina: Carolina Biological Supply Company.

Fredholm, Bertil B, et al. "Actions of Caffeine in the Brain with Special Reference to Factors That Contribute to Its Widespread Use." Pharmacological Reviews, vol. 51, no. 1, 1 Mar. 1999, pp. 83–133., pharmrev.aspetjournals.org/content/51/1/83?ijkey=82bc2be89e4ba5c421952840b805153539420735&keytype2=tf_ipsecsha.

Halim MA, Tan FHP, Azlan A, Rasyid II, Rosli N, Shamsuddin S, Azzam G. Ageing, Drosophila melanogaster and epigenetics. Malays J Med Sci. 2020;27(3):7–19. https://doi.org/10.21315/mjms2020.27.3.2

Wu, M. N., Ho, K., Crocker, A., Yue, Z., Koh, K., & Sehgal, A. (2009). The effects of caffeine on sleep in Drosophila require PKA activity, but not the adenosine receptor. The Journal of neuroscience : the official journal of the Society for Neuroscience, 29(35), 11029– 11037. https://doi.org/10.1523/JNEUROSCI.1653-09.2009



Jackson Flowers

Using Geometric Algebra to Estimate the Sparsity of Data and the Size of the Minimal Overcomplete Basis for Data through Observations

Abstract

Dictionary Learning is a type of machine learning with extensive applications, including facial recognition and image denoising. Dictionary Learning works by taking some dataset and outputting an overcomplete basis that can be used to represent each datapoint as a linear combination of as few basis vectors as possible. However, implementations of Dictionary Learning require as an input the number of basis vectors to generate, and there is currently no general procedure for determining the value of this number. Here we provide a procedure using the geometric outer product to bound the number of necessary basis vectors in the case where the dimensions of the subspaces comprising the dataset are distinct, the case where the dimensions are all the same, and the general case. We also provide a system of equations that gives the dimensions of each of the subspaces as solutions, and we provide bounds on the number of necessary basis vectors given the dimensions of the underlying subspaces. This brings Dictionary Learning closer to being able to be used on any dataset without prior knowledge about the dataset's underlying structure.

Introduction

Dictionary Learning is a type of machine learning that is useful in applications where many other machine learning techniques fail because Dictionary Learning outputs a representation of data that is then interpreted for results, rather than directly output the results from the data. These applications include facial recognition [1], low-dose x-ray CT reconstruction [2], and image denoising [3]. In short, Dictionary Learning takes some dataset and outputs a (potentially overcomplete) basis that can be used to represent each datapoint as a linear combination of as few basis vectors as possible [4], which is more commonly known as a sparse representation of the data. However, implementations of Dictionary Learning, such as those in Scikit-learn [5], require as an input the number of basis vectors (often called atoms) to generate, and there is currently no general procedure for determining the optimal value for this number [6]. In practice, usually one guesses and checks the number of basis vectors until one gets a reasonable result, as the number need only be greater than the minimum number necessary to represent the data, so very little research has been done on methods to find the number of basis vectors needed; however, using too many basis vectors can lead to overfitting and higher computation times. However, as processes involving Dictionary Learning become more and more automated, it will become necessary for a computer to be able to find the number of basis vectors necessary to sparsely represent data, which we call the overcomplete dimension of a dataset.

We outline a procedure to bound this overcomplete dimension. This number is closely related to the dimensions of the subspaces that the dataset is composed of, which we derive algorithms for estimating in general. We also examine specific subcases of the problem, when these subspaces have either all the same dimension or all distinct dimensions and develop algorithms to estimate the dimensions in these subcases.

Background on Geometric Algebra

First, we define the outer product using standard definitions (also known as the geometric outer product, the wedge product, and the exterior product) on R^n and note some of its basic properties. The outer product of a set {x_1,…,x_r} of vectors from an inner product space is the oriented r-dimensional parallelepiped with edges x_1,…,x_r. We denote the outer product of \{x_1,…,x_r \} by x_1∧⋯∧x_r, and we denote the volume of this parallelepiped by |x_1∧⋯∧x_r |. If this volume is 0, we write x_1∧⋯∧x_r=0.

From this definition, we immediately observe that if the inner product space is R^n and r>n, then x_1∧⋯∧x_r=0. More generally, x_1∧⋯∧x_r=0 if and only if {x_1,…,x_r} is a linearly dependent set. This can be easily proven by induction on r. To compute the outer product in practice, more results are necessary, such as those outlined by Macdonald [7]. Additionally, a more rigorous definition of the outer product can be achieved through geometric algebra [7] or through category theory [8].

For our purposes, we usually wish to determine whether |x_1∧⋯∧x_r |≈0 (i.e. less than ϵ for some fixed ϵ). The fastest way to do this is through an LU decomposition. First, we let A=[x_1,⋯,x_r], then compute the LU factorization of A using standard techniques; thus, A=LU for square lower triangular L and upper triangular U. Because |x_1∧⋯∧x_r |=0 if and only if \{x_1,…,x_r \} is linearly dependent, it follows immediately from the definition of an LU factorization that |x_1∧⋯∧x_r |≈0 if and only if U_(i,i)≈0 for some i.

Estimating the Probability of Linear Dependence

Let Ω=U_(i=1)^M▒Ω_i be a dataset consisting of a union of M>1 nontrivial subspaces of R^n. Order these so that dim⁡Ω_1≤⋯≤"dim" Ω_M, and let "dim" Ω_i=σ_i. We also let p_i=P(x∈Ω" and " x∈Ω_i). If we randomly select x_1,…,x_r from Ω, the probability of linear dependence P_r=P(x_1∧⋯∧x_r=0) is as follows:

$$P_r = \sum_{k=1}^{M} (-1)^{k+1} \sum_{1\le i_1\le\cdots\le i_k\le M} \sum_{\sigma_{i_1}<j_1,\ldots,\sigma_{i_k}<j_k}^{\sum j_i\le r} \frac{r!}{j_1!\cdots j_k!\,(r-\sum j_i)!} p_{i_1}^{j_1}\cdots p_{i_k}^{j_k}\left(1-\sum p_{i_l}\right)^{(r-\sum j_l)}.$$

This equation is novel, but a full derivation is excluded for conciseness. In the case where each of the subspaces has the same dimension, say σ, we can reduce the above equation. By symmetry, each term in the second sum has the same value, and since there are (M¦k) terms, the sum reduces to the following:

$$P_r = \sum_{k=1}^{M} (-1)^{k+1} \binom{M}{k} \sum_{\sigma\le j_i,\sum j_i\le r} \frac{r!}{j_1!\cdots j_k!\,(r-\sum j_i)!} M^{-\sum j_i}\left(1-\frac{k}{M}\right)^{r-\sum j_i}.$$

Algorithms to Estimate the Sparsity

Next, we propose three different algorithms to estimate the sparsity of a dataset, defined as the largest highest dimension of the underlying subspaces (σ_M), each of which reward different use-cases.

Firstly, we propose an algorithm for estimating the sparsity in the case where each of the subspaces has the same dimension and proportion, i.e. σ_1=⋯=σ_M=σ and p_1=⋯=p_M=1/M. We estimate P_r using Monte-

Carlo simulations, calling each estimate Q_r. The algorithm is as follows:

With 1000 samples, find the smallest value of r such that Q_r≠0. Then, perform 100,000 samples to improve the estimate. We denote this value of r as r_0. Note that this is not necessarily σ+1, as while P_(σ+1)>0, P_(σ+1) may be arbitrarily small.

Next, with 100,000 samples each, estimate Q_(r_0+1) and Q_(r_0+2).

Through brute force, find the least squares solution for σ and M according to the equation for P_(r_0), P_(r_0+1), and P_(r_0+2).

Next, we motivate the prior algorithm. Initially, we seek to find a nonzero value of Q_r with minimal r, but this isn't necessarily the global nonzero minimum of P_r. With 20 subspaces of R^20 of dimension 10, for example, P_11≈10^(-13), which is not reasonably detectable through Monte-Carlo methods. Thus, we only use 1000 samples initially to strike a balance between computing time and accuracy, though this number can be varied. The general idea of the algorithm is that we calculate a few values of Q_r, then find the values of M and σ that give theoretical probabilities closes to these values of Q_r through brute force. Lower values of r_0 significantly decrease the time of the brute force, which is why we seek the smallest value of r_0 in the first step. In the listed algorithm, we only calculate 3 terms arbitrarily, and this number may be increased/decreased to increase/decrease accuracy respectively. Because necessarily 0<σ<r_0, there are a finite number of values of σ to test, but there are technically an infinite number of values of M to test. Thus, the only way for the algorithm to actually function is to assume some maximum value for M, which can be done by considering the specific context of the problem the algorithm is being applied to.

Next, we propose an algorithm for estimating the sparsity in the case where each subspace has a distinct dimension:

Calculate Q_r for 1≤r≤n, and σ_1 is one less than the first r such that Q_r≠0. Also, p_1=(Q_(σ_1+1))^(1/(σ_1+1)).

Suppose that we know the first m values of σ_i and p_i. Let ρ_m (r) denote the value of the equation for P_r when the first sum is evaluated from i=1 to m instead of from i=1 to M for a given r value. σ_(m+1) is one less than the first r value for which Q_r-ρ_m (r) is nonzero. In addition, p_(m+1)=(Q_(σ_(m+1))-ρ_m (r))^(1/(σ_(m+1)+1)). By the principle of mathematical induction, we have a formula for all σ_i and p_i.

The formulas used in the above algorithm stem from the fact that when r=σ_i+1, there is only one term in the formula for P_r involving σ_i and p_i. Thus, we may inductively calculate each σ_i to find the sparsity σ_M.

Finally, we propose an algorithm for the general case, i.e. when σ_i are arbitrary.

Using 1000 samples, find the first value r such that Q_r≠0. We call this value r_0. Then, calculate Q_r for all r_0≤r≤n-1 using 10,000 samples.

Through brute force, calculate the least squares solution for σ_1,...,σ_M and M, assuming some fixed distribution for p_1,...,p_M.

This is a direct generalization of the algorithm for when each subspace has the same dimension. In this, we typically assume that each subspace has approximately the same proportion in the dataset, but any distribution may be assumed; some distribution must be assumed to be able to calculate each P_r.

We implement each of these algorithms in the Julia programming language and have verified their efficacy and accuracy.

Estimating the Size of the Overcomplete Basis

A set of vectors D={D_1,...D_N} is an overcomplete basis for Ω if each Ω_i can be spanned by exactly σ_i elements of D. The size of an overcomplete basis D is |D|. An overcomplete basis D for Ω is minimal if every other overcomplete basis for Ω has size greater than or equal to the size of D. The overcomplete dimension for Ω is the size of any minimal overcomplete basis for Ω.

Henceforth, D refers to some minimal overcomplete basis for Ω with size N. In general, minimal overcomplete bases are not unique, but in select cases, minimal overcomplete bases are unique up to scalar multiplication. We seek to bound N given each σ_i.

Firstly, we make a simplifying observation. If Ω is some dataset with overcomplete dimension N and Λ is some dataset with overcomplete dimension η, then the overcomplete dimension of Ω∪Λ is less than or equal to N+η. This is due to the requirement that the overcomplete basis corresponding to the overcomplete dimension be minimal. Because the overcomplete dimension of some subspace Ω_i is just σ_i by definition, it follows that N≤∑σ_i.

If two subspaces have dimensions that sum to less than n, they have no guaranteed nontrivial intersection, as can be seen clearly through the example of subspaces slightly shifted from each other. Thus, the following upper bound for N holds:

$$N \leq \begin{cases} \displaystyle\sum_{i=1}^{M/2} \min(n, \sigma_{2i} + \sigma_{2i-1}) & \text{if } M \text{ is even} \\ \sigma_1 + \displaystyle\sum_{i=1}^{(M-1)/2} \min(n, \sigma_{2i} + \sigma_{2i+1}) & \text{if } M \text{ is odd.} \end{cases}$$

This upper bound essentially "pairs" subspaces together to see if they intersect, hence the reliance on the parity of M.

Implications

We have given a procedure that allows one to directly bound the number of basis vectors needed to sparsely represent data, as well as calculate the dimensions of each of the subspaces comprising the data. Specifically, we have developed three different algorithms to estimate the dimensions of the subspaces comprising a sparse dataset for three different use case: when each of the dimensions is the same, when each of the dimensions is distinct, and when no assumptions are made about the dimensions whatsoever. Each of these provides a method to estimate the sparsity, as well as provide inputs to our derived bounds for the overcomplete dimension. Moreover, these bounds are essentially the best possible bounds, except for slight possible improvements as σ_i→n. In the sparse case, however, these bounds cannot be improved, as can be shown by a construction of a limiting case. This helps bridge the gap between data collection and dictionary learning techniques.

Future Work

There is still potential work to be done to improve the three proposed algorithms. In particular, the number of samples taken and terms used is arbitrary, so more work could be done to find the minimal sample size. In addition, the algorithms themselves should be tested with real-world datasets, as all testing done for this project was with manufactured datasets. Additionally, more work can be done regarding the effect of noise and small datasets on the accuracy of the algorithms. Finally, it may be possible to gain information by considering Ω⊗G^n, where G^n is the geometric algebra in n dimensions, as this could help bring out more details about the subspaces.

References

[1] Q. Zhang and B. Li. Discriminative K-SVD for dictionary learning in face recognition. In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 2691-2698. IEEE, 2010.

[2] Q. Xu, H. Yu, X. Mou, L. Zhang, J. Hsich, and G. Wang. Low-dose X-ray CT reconstruction via dictionary learning. IEEE transactions on medical imaging, 31(9):1682-1697, 2012.

[3] W. Dong, X. Li, L. Zhang, and G. Shi. Sparsity-based image denoising via dictionary learning and structural clustering. In CVPR 2011, pages 457-464. IEEE, 2011.

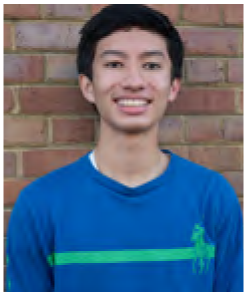
[4] L. Tosic and P. Frossard. Dictionary Learning. IEEE Signal Processing Magazine, 28(2):27-38, 2011.

[5] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12:2825–2830, 2011.

[6] T. Tong, J. Caballero, K. Bhatia, and D. Rueckert. Chapter 6 - dictionary learning for medical image denoising, reconstruction, and segmentation. In G. Wu, D. Shen, and M. R. Sabuncu, editors, Machine Learning and Medical Imaging, pages 153 – 181. Academic Press, 2016.

[7] A. Macdonald. Linear and geometric algebra. Alan Macdonald, 2010.

[8] S. Mac Lane and G. Birkhoff. Algebra. Chelsea Publishing Company, Incorporated, 3rd edition, 1999.



Analyzing Global Factors Affecting the HIV Epidemic Utilizing Statistical Tools and Machine Learning

Nicholas Djedjos

Abstract

Human Immunodeficiency Virus (HIV), the virus that causes Acquired Immunodeficiency Syndrome (AIDS), infected an estimated 1.7 million individuals globally in 2019. However, there are few global studies that analyze HIV at a global level. This study was undertaken to analyze the quantitative values of potential causative socioeconomic and health factors of HIV incidence rate worldwide. Drawing from multiple information sources, such as UNAIDS and WHO, possible country level indicators of HIV, such as adolescent fertility rate, contraceptive prevalence rate, mean years of education, medical doctors, and health expenditures were compiled from 185 countries. Bivariate and multivariate analysis was utilized to extract the strength of the possible factors. Packages such as sci-kit learn and NumPy in conjunction with Python were used for statistical analysis, and Microsoft Excel was used for data storage. Through univariate, bivariate, and machine learning analysis, the strongest positive correlation was found between HIV incidence rate and adolescent fertility rate. The strongest negative correlations were found between medical doctor density and mean years of education. The multiple linear regression model had a lower test accuracy and had a conflicting positive correlation with mean years of education compared to the other analyses, signalling that most of the data could be non-linear. Curbing adolescent fertility rate, increasing the number of medical doctors, and increasing mean years of education may decrease HIV infections worldwide. This can be implemented by local, state, and national governments.

Introduction

Human Immunodeficiency Virus (HIV) /Acquired Immune Deficiency Syndrome (AIDS) continues to affect millions of individuals and is an enormous health issue worldwide. Unlike its retrovirus cousins, HIV specifically targets CD4 lymphocytes, which are white blood cells that are designed to fight viral infection. This, combined with HIV's high mutation rate and ability to remain dormant, creates unique challenges for researchers (Retroviruses, 1997). Though antiretroviral therapy is instrumental for longer patient life spans and slowing the progression of the HIV pandemic, there is currently no cure for the retrovirus (WHO, 2013). Thus, it is instrumental that potential causes for disease transmission are examined.

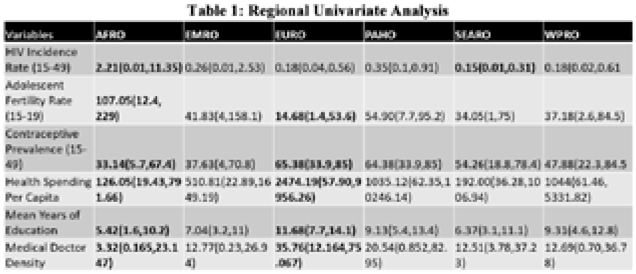
As HIV has touched nearly every country in the world, understanding why HIV incidence rates vary country to country may be useful to developing policies to combat the disease. This variance can be attributed to different influential and behavioral risk factors per population. Though there has been research involving specific countries and regions and HIV prevalence rate, few studies cover the broad impact of predictor variables on HIV incidence rate throughout the world.

The objective of this study is to examine the influence of social, health, and economic factors on global HIV incidence rates through the usage of statistical analyses. Specifically, a multiple linear regression analysis predicts the outcome (Y) based on a set of predictor variables (Xi) to assess the impacts of multiple variables in the same model (Alexopoulos, 2010) and a random forest feature importance model could evaluate the importance of features on a classification task (Sci-Kit Learn, 2011). Trained on current epidemiological data derived from various health databases, a dual machine learning model system could reveal to private, public, and academic sectors the policies necessary to prohibit further transmission of the disease.

Methodology:

Data: Five characteristics were analyzed due to their possible HIV Incidence correlation: adolescent fertility rate per 1000 women ages 15-19 and medical doctor density per 10,000 people, from the World Health Organization database, mean years of education of total schooling for ages 25+ from the our World In Data dataset, health spending per capita from the World Bank dataset, and contraceptive prevalence, both traditional and modern, from the Population Reference Bureau fact sheet. The HIV Incidence rate was retrieved from the UNAIDS fact sheet. There were 1386 data points procured for regional comparisons and of those 1386 data points, 819 were utilized for modelling. Datasets were sorted with a Python script to align with country specific data. For further HIV feature analysis, a "low" HIV incidence rate was assigned to a country wide incidence rate of less than 1.0, and a "high" incidence rate was assigned to a country wide incidence greater than 1.0. The 1.0 boundary was chosen due to the mean HIV incidence rates of the recorded countries being 1.01.

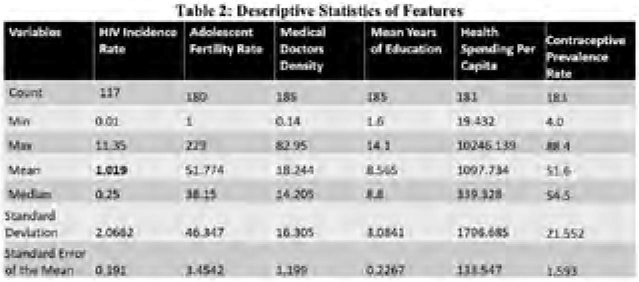
Table 1: Regional Univariate Analysis



Variables	AFRO	EMRO	EURO	PAHO	SEARO	WPRO
HIV Incidence Rate (15-49)	**2.21(0.01,11.35)**	0.26(0.01,2.53)	0.18(0.04,0.56)	0.35(0.1,0.91)	**0.15(0.01,0.31)**	0.18(0.02,0.61
Adolescent Fertility Rate (15-19)	**107.05(12.4, 229)**	41.83(4,158.1)	**14.68(1.4,53.6)**	54.90(7.7,95.2)	34.05(1,75)	37.18(2.6,84.5)
Contraceptive Prevalence (15-49)	**33.14(5.7,67.4)**	37.63(4,70.8)	**65.38(33.9,85)**	64.38(33.9,85)	54.26(18.8,78.4)	47.88(22.3,84.5
Health Spending Per Capita	**126.05(19.43,791.66)**	510.81(22.89,1649.15)	**2474.19(57.90,9956.26)**	1035.12(62.35,10246.14)	192.00(36.28,1006.94)	1044(61.46, 5331.82)
Mean Years of Education	**5.42(1.6,10.2)**	7.04(3.2,11)	**11.68(7.7,14.1)**	9.13(5.4,13.4)	6.37(3.1,11.1)	9.33(4.6,12.8)
Medical Doctor Density	**3.32(0.165,23.147)**	12.77(0.23,26.94)	**35.76(12.164,75.067)**	20.54(0.852,82.95)	12.51(3.78,37.23)	12.09(0.70,36.78)

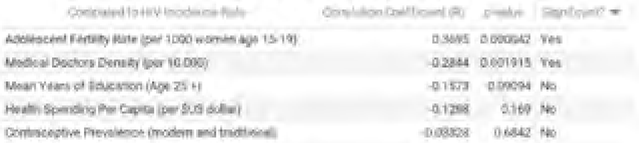
Each cell includes the mean, and within the parenthesis, is the minimum and maximum value of the feature.

Table 2: Descriptive Statistics of Features



Variables	HIV Incidence Rate	Adolescent Fertility Rate	Medical Doctors Density	Mean Years of Education	Health Spending Per Capita	Contraceptive Prevalence Rate
Count	117	180	185	185	181	181
Min	0.01	1	0.14	1.6	19.432	4.0
Max	11.35	229	82.95	14.1	10246.139	88.4
Mean	**1.019**	51.774	18.244	8.565	1097.734	51.6
Median	0.25	38.15	14.205	8.8	339.328	54.5
Standard Deviation	2.0682	46.347	16.305	3.0841	1796.685	21.552
Standard Error of the Mean	0.191	3.4542	1.199	0.2267	133.547	1.593

Table 3: Bivariate Analysis (p < 0.05)



Correlated to HIV Incidence Rate	Correlation Coefficient (R)	p-value	Significant?
Adolescent Fertility Rate (per 1000 women age 15-19)	0.3695	0.000042	Yes
Medical Doctors Density (per 10,000)	-0.2844	0.001915	Yes
Mean Years of Education (Age 25 +)	-0.1573	0.09094	No
Health Spending Per Capita (per $US dollar)	-0.1288	0.169	No
Contraceptive Prevalence (modern and traditional)	-0.03828	0.6842	No

For bivariate analysis, a Pearson coefficient test was utilized as well as a p-test to determine statistical significance. The most influential factor found was adolescent fertility rate, which had a statistically significant positive correlation of 0.3695 and the least influential factor was Contraceptive Prevalence which had a weak, negative correlation of -0.03828.

Machine Learning Models:

The data was imported into a Multiple Linear Regression model, where backwards tracking was utilized to find the most influential features and their correlation. A second machine learning method, Random Forest Feature importance, paired with a Random Forest Classifier, further analyzed the five features. The training data was chosen randomly, with 80 percent of the data points functioning as training data and 20 percent as test data. Each machine learning model was run ten times, and the average was determined to be the overall accuracy and feature importance.

Results-

Univariate Analysis

The univariate analysis indicates that AFRO WHO Region had the highest HIV Incidence Rate, highest adolescent fertility rate, lowest medical doctor density, lowest mean years of education, lowest health spending per capita, and lowest contraceptive prevalence rate. Meanwhile, the EUR WHO Region had nearly the opposite, with one of the lowest HIV Incidence Rates, lowest adolescent fertility rate, highest medical doctor density, highest mean years of education, highest health spending per capita, and highest contraceptive prevalence.

Bivariate Relationships:

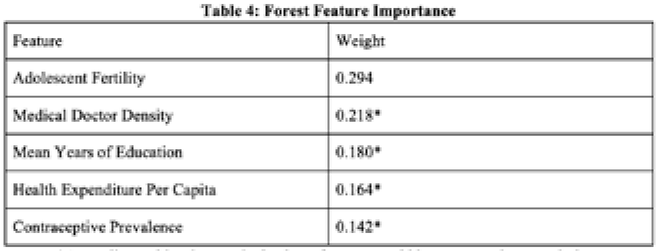
A Pearson correlation test showcased the correlation coefficient between each feature and HIV Incidence Rate. A statistically significant, positive correlation was found between Adolescent fertility rate and HIV incidence rate. A statistically significant, negative correlation was found between medical doctor density and HIV incidence rates.The other three variables had statistically insignificant, negative correlations with HIV incidence rate. The most influential factor found was adolescent fertility rate and the least influential factor was Contraceptive Prevalence. The bivariate analysis also aligns with the univariate analysis, as the region with the highest HIV incidence rate also had the highest adolescent fertility rate, as well as one of the lowest HIV incidence rates had the lowest adolescent fertility rate.

Modeling:

The multiple linear regression model with five features had an accuracy of 27.8% on training data and a 23% test accuracy rate with the mean years of education having the largest correlation weight. The least important feature, health spending per capita was subsequently dropped. The next multiple linear regression model had a 25.8% accuracy on training data and 27% test accuracy rate on testing data, and the mean years of education had the largest correlation weight. Dropping the next least important feature, contraceptive prevalence, decreased the accuracy of the model.

The random forest feature importance model designated the strongest feature importance to adolescent fertility, with medical doctor density second. The third important feature varied between mean years of education and health spending per capita. To test the accuracy of the features, a random forest classifier was utilized with the same randomized dataset the feature importance model used. The random forest classifier model had an accuracy of 94.72% on training data and 84% on test data.

Table 4: Forest Feature Importance



Feature	Weight
Adolescent Fertility	0.294
Medical Doctor Density	0.218*
Mean Years of Education	0.180*
Health Expenditure Per Capita	0.164*
Contraceptive Prevalence	0.142*

*According to bivariate analysis, these features could have a negative correlation.

Figure 1: Multiple Linear Regression Formula

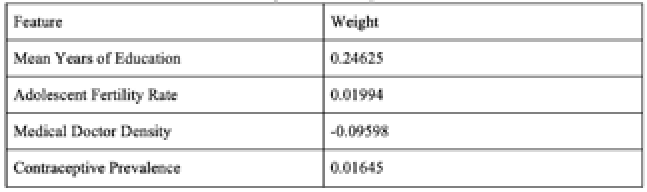
$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + b_5x_5$$

y = dependent variable

b_0 = y intercept

b_n = slope coefficient

Table 5: Multiple Linear Regression Trial 2



Feature	Weight
Mean Years of Education	0.24625
Adolescent Fertility Rate	0.01994
Medical Doctor Density	-0.09598
Contraceptive Prevalence	0.01645

Discussion:

Notably, the top three features in each model were the same: adolescent fertility rate, medical doctor density, and mean years of education. Although the multiple linear regression model had a 27% test accuracy, it gave considerable weight to mean years of education compared to the other features, more than eleven times the weight of the second important feature. Meanwhile, the forest feature importance model gave the most weight to adolescent fertility, which aligns with the bivariate statistical correlation. Utilizing the same random data as the feature model, a classifier model was able to distinguish between low and high HIV incidence rates at a country level 94.7% of the time with training data and 82% of the time with test data. A possible reason for the low accuracy and positive correlation of mean years of education with HIV incidence rate for the multiple linear regression, which conflicted with the bivariate analysis and machine learning model, is that most of the data is nonlinear. This idea aligns with the random feature importance and decision trees classifier having a higher accuracy.

Implications and Recommendations

It is crucial that a data driven approach is utilized to combat the transmission of HIV (Courtenay-Quirk et al., 2016). The top three recurring features, adolescent fertility rate, medical doctor density, and mean years of education should be examined further. Moreover, this research aligns with Sustainable Development Goal 3 by the United Nations, which is the ability to ensure healthy lives and wellbeing. Being knowledgeable on the most influential factors causing HIV in the world will allow broad efforts from multiple agencies to make specific adjustments to prevent HIV spread.

Adolescent Fertility Rate

A statistically significant, positive relationship was found between adolescent fertility rate and HIV incidence rate. Around the world, countries that suffer from disproportionately high adolescent fertility rates also have a greater percentage of non-marital sex, domestic abuse, and unwanted pregnancies, all of which could contribute to sexually transmitted diseases like HIV. Measures at both the government level and local level that could curb adolescent fertility rate include stronger social programs to improve socioeconomic status and improving gender inequality standards through national laws.

Medical Doctor Density

A statistically significant, negative relationship was found between medical doctor density and HIV incidence rate. One explanation is for patients to be properly diagnosed and treated for HIV, there must be medical doctors. Their roles as preliminary screeners, monitors of infection, and appropriate managers of antiretroviral therapy greatly affect the spread of the virus. Because of this, increasing medical doctor density through financial incentives or government involvement could decrease HIV incidence rate especially in countries with low densities of doctors.

Mean Years of Education

Although mean years of education was statistically insignificant, its heavy weight in the multiple linear regression model could provide insight on the HIV epidemic. Educated people have more financial stability and as a result, do not have to turn to risky behaviors such as prostitution for an adequate income. Further, health education helps individuals be more aware about the risks of HIV, and how to protect themselves. Increasing mean years of education at the community, state, and national level may reduce HIV incidence rate according to the bivariate analysis.

Future Research

The data utilized in this model may be used in more machine learning models such as a support vector machine, neural network, or K-Nearest Neighbors classifier to further analyze the data. In the future, including more features and data points to fully encapsulate the HIV pandemic could yield stronger statistical relationships. In addition to more data points, a future model could track yearly HIV cases instead of HIV incidence rate.

References

Alexopoulos, E. C. (2010). Introduction to Multivariate Regression Analysis. Hippokratia, 14(Suppl 1), 23–28. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3049417/

Cari Courtenay-Quirk, Hilary Spindler, Aimee Leidich, and Pam Bachanas (2016). Building Capacity for Data-Driven Decision Making in African HIV Testing Programs: Field Perspectives on Data Use Workshops. AIDS Education and Prevention: Vol. 28, No. 6, pp. 472-484. https://doi.org/10.1521/aeap.2016.28.6.472

Current health expenditure per capita (current US$) | Data. (2017). [Dataset]. The World Bank. https://data.worldbank.org/indicator/SH.XPD.CHEX.PC.CD

Our World in Data (2017). Mean years of schooling 1870-2017[Dataset] https://ourworldindata.org/grapher/mean-years-of-schooling-1?tab=chart

Population Reference Bureau (2019). Family Planning Data Sheet. https://www.prb.org/2019-world-population-data-sheet/

Retroviruses. (1997). In J. M. Coffin, S. H. Hughes, & H. E. Varmus (Eds.), PubMed. Cold Spring Harbor Laboratory Press. https://www.ncbi.nlm.nih.gov/books/NBK19376/

Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011

United Nations AIDS (2019). HIV Estimates with Uncertainty Bounds 1990-2019I,(latest available year) [Dataset]. https://www.unaids.org/en/resources/fact-sheet#:~:text=New%20HIV%20infections&text=In%202019%2C%20around%201.7%20million,3.7%20million%5D%20people%20in%201998.

The Effect of *Moringa oleifera* on Purification of Natural Bodies of Water

Haliee Sexton

Madeline Raynor

Abstract

Water contamination is currently a growing issue around the world, and the Moringa oleifera seed showed promise in successful water purification. Knowing this, it can be determined that the Moringa seed would be successful in filtering toxic agents from natural bodies of water as well. In this study, a sample of unfiltered, natural water was collected from a local retainment pond, and the contaminants in the water were tested for dissolved solids, along with other potentially toxic elements that could be present in the sample. Crushed Moringa oleifera seeds were incorporated into the water sample. The solution was left to sit for 24 hours before the testing process was repeated. There were substantial levels of potentially harmful elements in the sample, such as lead, nitrate, sulfate, and zinc. The presence of zinc nitrate can irritate the skin and eyes, and even cause nausea and vomiting. Lead is also extremely toxic when consumed. All values, excluding the zinc, were reduced to zero after filtration. Additionally, using a digital pH meter, the pH of the sample after filtration was determined to be 6.5, 0.5 higher than the original value of 6, which is closer to the pH of pure water. The study showed that the moringa seed has significant effects on the filtration of toxic agents out of bodies of water. The removal of multiple elements and increased pH demonstrates that the Moringa seed was successful in the purification of the water sample, and further experiments will clarify the limitations of its abilities in filtration.

Introduction

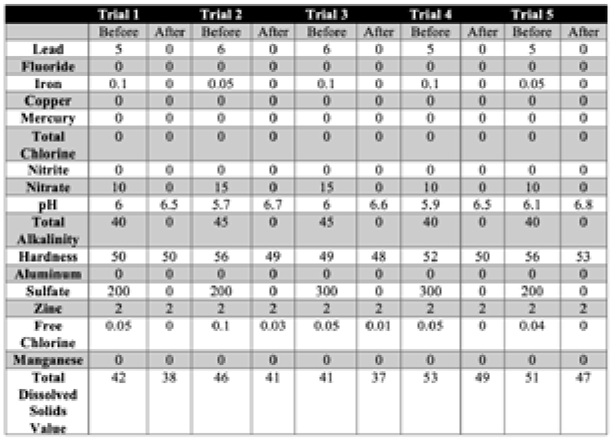
The Moringa oleifera plant is native to Africa, and it is used primarily for its medicinal properties. The plant, however, can be used as a natural filter for water. According to researchers Francis Kweku Amagloh and Amos Benang (2009), Moringa is a non-toxic, natural source that can be used in water purification. The chemicals typically used in this process, like chlorine, could cause extreme health hazards if something were to go wrong in the filtering process. Their research determined that the filtering effects of the plant's seed had extremely similar results as the previous filtering method used by professionals (Amagloh & Benang, 2009). Compared to aluminum sulfate and ferric sulfate, the percentages of E. coli and turbidity removed by the Moringa oleifera were very similar and only slightly lower. Researchers from Great Britain have also come to similar conclusions, going so far as to make the observation that older seeds have less effective filtering properties. These researchers made the powdered Moringa into a paste that could be further used for research (Pritchard, et al. 2010). Their research, however, did not concern the effects of the plant on naturally occurring water sources. Small bodies of water, such as holding ponds or small lakes, often run-off into rivers and streams. These streams lead into huge bodies of water, so if the ponds and lakes become polluted with toxins, the large bodies of water will as well. The powdered Moringa oleifera plant could potentially filter out the toxins on a large scale, depending on the concentration of the seeds

Figure 1: Moringa oleifera seeds in the Moringa oleifera plant. From "25 Major Health & Beauty Benefits of Moringa Seeds," by Bharat Sharma, 2016 (goodhealthall.com)

Methods

The first step was to collect a sample of unfiltered, natural water from a local retainment pond. The presence of contaminants in the water was investigated using water test strips. These strips show the presence of potentially toxic elements such as lead, fluoride, iron, copper, mercury, chlorine, nitrite, nitrate, sulfate, zinc, and manganese. The water sample was also tested for dissolved solids, and the pH level was determined using a digital pH tester. After recording the results, finely crushed Moringa seeds were incorporated into the water sample. The new solution was left to sit for 24 hours, and the new levels of the contaminants was retested. The testing process was repeated for five trials. The data was then inserted into a chart to run a statistical analysis (Table 1).

Table 1:

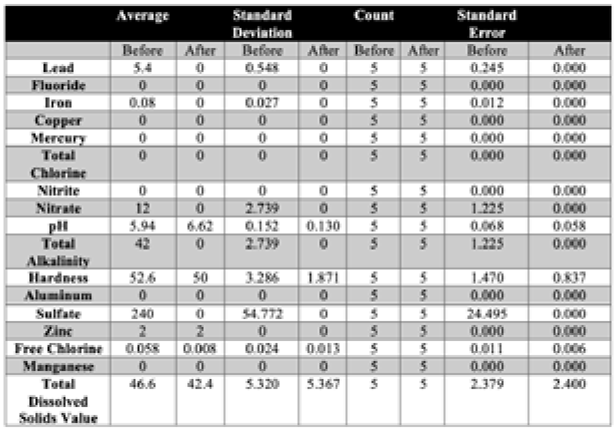


	Trial 1		Trial 2		Trial 3		Trial 4		Trial 5	
	Before	After	Before	After	Before	After	Before	After	Before	After
Lead	5	0	6	0	6	0	5	0	5	0
Fluoride	0	0	0	0	0	0	0	0	0	0
Iron	0.1	0	0.05	0	0.1	0	0.1	0	0.05	0
Copper	0	0	0	0	0	0	0	0	0	0
Mercury	0	0	0	0	0	0	0	0	0	0
Total Chlorine	0	0	0	0	0	0	0	0	0	0
Nitrite	0	0	0	0	0	0	0	0	0	0
Nitrate	10	0	15	0	15	0	10	0	10	0
pH	6	6.5	5.7	6.7	6	6.6	5.9	6.5	6.1	6.8
Total Alkalinity	40	0	45	0	45	0	40	0	40	0
Hardness	50	50	56	49	49	48	52	50	56	53
Aluminum	0	0	0	0	0	0	0	0	0	0
Sulfate	200	0	200	0	300	0	300	0	200	0
Zinc	2	2	2	2	2	2	2	2	2	2
Free Chlorine	0.05	0	0.1	0.03	0.05	0.01	0.05	0	0.04	0
Manganese	0	0	0	0	0	0	0	0	0	0
Total Dissolved Solids Value	42	38	46	41	41	37	53	49	51	47

Results: Each of the before and after results of the five trials is recorded in parts per million as well as pH, and each value is an estimation based off of the values retrieved from the water testing strips. One can observe a decrease in the values of lead, iron, nitrate, the total alkalinity, hardness, sulfate, free chlorine, and the total dissolved solids value. The pH also became less acidic with the introduction of the seed.

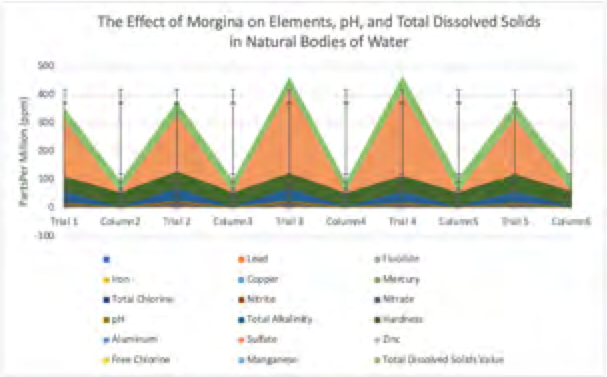
Statistical Analysis: To summarize the data found from the experimentation, the before and after values from each trial were averaged together. From this average, the standard deviation between each of the before and after data points were determined, as well as the standard error. On average, there is not much variation between each data point, and for data sets such as sulfate, the large standard deviation in explainable. The values on the water testing strips for sulfate are separated by intervals of 50, so the difference between each value determined from these strips will vary by about 50. The standard error values are small, so there is an unlikely chance that the data points are inaccurate.

Table 2:



	Average		Standard Deviation		Count		Standard Error	
	Before	After	Before	After	Before	After	Before	After
Lead	5.4	0	0.548	0	5	5	0.245	0.000
Fluoride	0	0	0	0	5	5	0.000	0.000
Iron	0.08	0	0.027	0	5	5	0.012	0.000
Copper	0	0	0	0	5	5	0.000	0.000
Mercury	0	0	0	0	5	5	0.000	0.000
Total Chlorine	0	0	0	0	5	5	0.000	0.000
Nitrite	0	0	0	0	5	5	0.000	0.000
Nitrate	12	0	2.739	0	5	5	1.225	0.000
pH	5.94	6.62	0.152	0.130	5	5	0.068	0.058
Total Alkalinity	42	0	2.739	0	5	5	1.225	0.000
Hardness	52.6	50	3.286	1.871	5	5	1.470	0.837
Aluminum	0	0	0	0	5	5	0.000	0.000
Sulfate	240	0	54.772	0	5	5	24.495	0.000
Zinc	2	2	0	0	5	5	0.000	0.000
Free Chlorine	0.058	0.008	0.024	0.013	5	5	0.011	0.006
Manganese	0	0	0	0	5	5	0.000	0.000
Total Dissolved Solids Value	46.6	42.4	5.320	5.367	5	5	2.379	2.400

Modeling: The line graph shows the evident change between each element tested before and after the introduction of the Moringa seed. Each value is recorded in parts per million, and each element is color coded to coincide with the diagram. The vertical lines indicate standard deviation as well as standard error.



Discussion

These results partially accepted our hypothesis, showing that the seed has potential filtration abilities that can be further developed and implemented into a more widespread process for water purification. The results showed that the seed is effective in making the water less acidic. The water treated with the moringa showed a pH higher and closer to 7, which is the pH of pure water. The moringa also showed successful infiltration of nitrogen-centered compounds like nitrate, as well as lead, iron, sulfate, and free chlorine. Many bodies of water are contaminated by fertilizers used in surrounding soil, which has large amounts of nitrogen (Fertilizer 101, 2014). Therefore, an interesting use for moringa could be decontamination in the water on land largely treated with fertilizers. Another important note is the successful filtration of lead, which is very toxic if consumed in large amounts, and people drinking impure water in developing countries could be at a high risk of lead consumption. Additionally, as a result of the moringa filtration, the amounts of alkalinity, hardness, and dissolved solids were lowered, further showing that the moringa can potentially be used as a filter for water high in these contaminants.

Implications and Recommendations

The research is applicable the Sustainable Development Goals numbers 6, 3, and 10: clean water and sanitation, good health and well-being, and reduced inequalities. Because the Moringa oleifera seed is effective in filtering toxins out of natural water sources, it can be used to improve water quality as well as water pollution. About forty percent of American households rely on groundwater as a source of drinking water, and most of this groundwater becomes polluted by pesticides and fertilizer (Denchak, 2018). The research can also be applied to sustainable development goal number 3, good health and well-being. Good health strongly depends on clean and safe water for everyone. This research also concerns sustainable development goal number 10, reduced inequalities. A large part of inequalities between first and third world countries is the ability to have clean water, and successful filtration with the Moringa would greatly benefit toward this goal. From the research, it was determined that the Moringa seed was effective in filtering out nitrogen-centered compounds such as nitrite and nitrate. Nitrogen is also one of the main contents in fertilizer, so conclusively, the seed would be effective in filtering out the toxins found in fertilizer as well as the excess nitrogen (Fertilizer 101, 2014). The filtering effects of the seed can also help to combat water pollution caused by run-off into natural water sources such as ponds and rivers.

Future Research

The moringa seed shows much potential in water filtration, but it was not completely successful in decontaminating all of the water. A process implementing the seed needs to be researched and developed further in order to have a higher success rate, instead of just the seed by itself. Research could be dedicated to find something else inexpensive and natural to accompany the seed to make it more efficient and productive in complete filtration, based on the results that were only successful in part.

References

Amagloh, F. B. (2009, February). Effectiveness of Moringa oleifera seed as a coagulant for water purification. African Journal of Agricultural Research, 4, 119-123.

Denchak, Melissa (2018, May). "Water Pollution: Everything You Need to Know." Retrieved from https://www.nrdc.org/stories/water-pollution-everything-you-need-know

Fertilizer 101: The Big 3 - Nitrogen, Phosphorus and Potassium. (2014, May). Retrieved from https://www.tfi.org/the-feed/fertilizer-101-big-3-nitrogen-phosphorus-and-potassium

Pritchard, M. C. (2010). A comparison between Moringa oleifera and chemical coagulants in the purification of drinking water - An alternative sustainable solution for developing countries. Physics and Chemistry of the Earth, 35(13-14), 798-805.

Pritchard, M. C. (2010, July). A study of the parameters affecting the effectiveness of Moringa oleifera in drinking water purification. Physics and Chemistry of the Earth, Parts A/B/C, 35(13-14), 791-797.

Sharma, B (2016). [Picture of the Moringa oleifera seed inside of the Moringa plant] [Photograph]. Good Health All. https://goodhealthall.com/moringa-seeds-health-beauty-benefits/

.

Investigating Growth Promotion and Soil Enrichment in *Allium sativum*

Maggie Buck

Abstract

In farming, Allium sativum (garlic) has been said to influence living things around it. This has been caused by various reasons such as the chemical composition of garlic and its effect on the soil. The effectiveness of garlic as an insecticide is well-known, but there is little known about garlic's long-term effects on the soil. This study examines the effects of garlic grown on its own in the soil checked every two weeks over twelve weeks. The effects of garlic on the soil over time is shown in this experiment by comparing the control with the later soil samples. A soil sample was taken every two weeks from the plant trays and tested using a Vernier pH meter and a Rapitest Soil Test Kit. The pH was tested by putting the meter in water that had the soil mixed into it and allowed to settle. The soil nutrients were tested by using Rapitest Kit. The results show that growing Allium sativum used potash (potassium). This shows Allium sativum can survive in poorer soil than is usually believed. This is especially significant with the low levels of nitrogen as it is usually thought that is nitrogen is a necessity to grow Allium sativum.

Alternate hypothesis: Allium sativum effects soil composition with its use of nutrients in the soil.

Null hypothesis: Allium sativum does not effect soil composition.

Introduction

In farming, Allium sativum (garlic) has been said to influence living things around it. This has been caused by various reasons such as the chemical composition of garlic and its effect on the soil. Garlic is an important plant in its use in intercropping as insect barriers such as in its use in between rows of cabbage which increased plant survival rate by 10.2% (Mischeck & Katsaruware, 2014). The effect that garlic has on the plants it is grown with is mostly unknown with a lack of research of garlic's effect on the soil itself. Garlic is important in agriculture as a pesticide as well with its use to control numerous insects such as Tetranychus urticae where 50% of insects were killed at a concentration of 7.49 mg/l of garlic distillate (Attia, et al., 2011; Attia, et al., 2011). The effectiveness of garlic as an insecticide is well-known, but there is little known about garlic's long-term effects on the soil and the plants it is grown with. The effects on the soil especially with enzymes need to be researched due to them playing key roles in nutrient cycling which can change the nutrients in the soil over time (Soil Enzymes, 2011). Garlic is known to have a strong need for nitrogen and phosphorus with its growth stunted without it and that some fertilization is needed to grow garlic (Adem & Tadesse, 2014)

Methods

Data: A soil sample was taken every two weeks from the plant trays and tested using a Vernier pH meter and a Rapitest Soil Test Kit. The pH was tested by putting the meter in water that had the soil mixed into it and allowed to settle. The soil nutrients were tested by mixing soil and distilled water, waiting 24 hours, filling the testing container with the water, adding the powder inside the testing capsules, waiting the required time, and comparing the color to the other side.

Results

Set-up: Before and After Growth Period

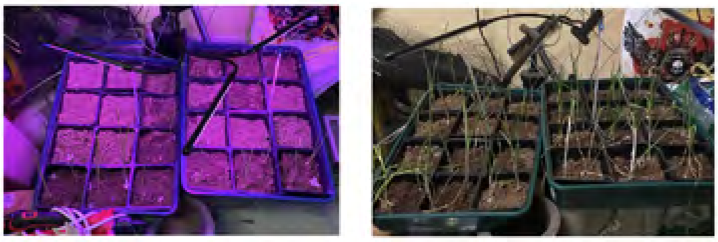


Table 1: Data that resulted from testing: Nitrogen, Phosphorus and Potash measurements on a 0-5 scale; pH scale 1 – 14.

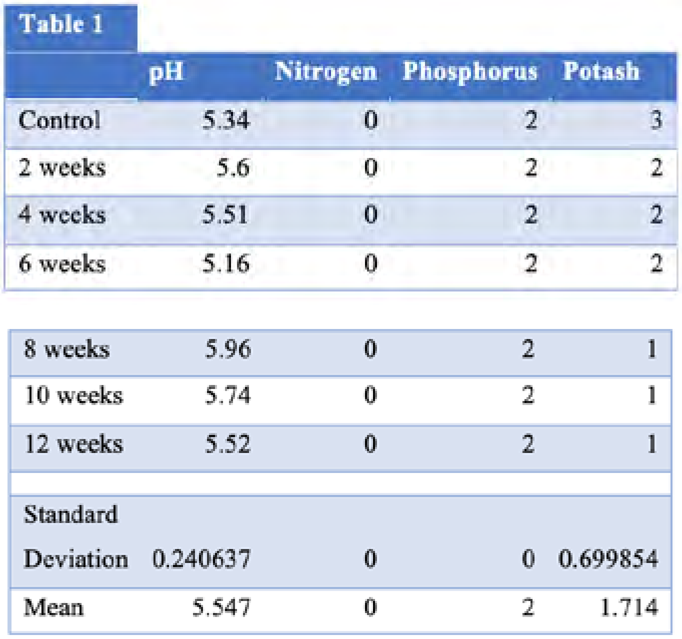
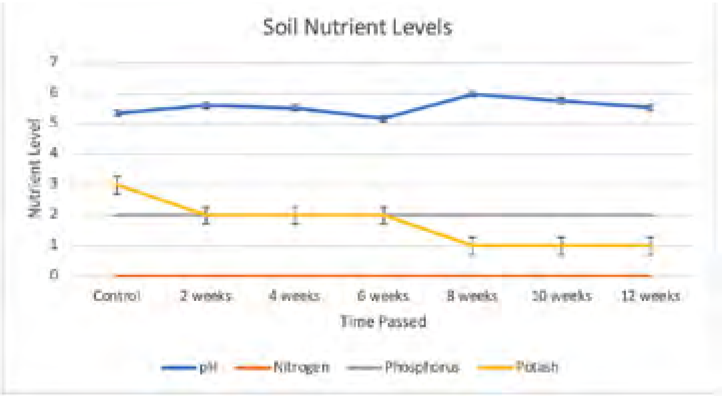


Table 1

	pH	Nitrogen	Phosphorus	Potash
Control	5.34	0	2	3
2 weeks	5.6	0	2	2
4 weeks	5.51	0	2	2
6 weeks	5.16	0	2	2
8 weeks	5.96	0	2	1
10 weeks	5.74	0	2	1
12 weeks	5.52	0	2	1
Standard Deviation	0.240637	0	0	0.699854
Mean	5.547	0	2	1.714



Discussion

The results show that growing Allium sativum used potash (potassium). This also shows that Allium sativum can survive with minimal nitrogen.

Implications and Recommendations

This shows Allium sativum can survive in poorer soil than is usually believed. This is especially significant with the low levels of nitrogen as it is usually thought that is nitrogen is a necessity to grow Allium sativum. This can also be used to further the effort in achieving SDG 15, Life on Land, by allowing for more efficient farming and fertilization of farmland and more efficient fertilization decreases fertilizer runoff which causes algae blooms and other events. It can also help achieve SDG 2, Zero Hunger, by allowing more efficient farming for Allium sativum.

Future Research

This research could be used in further research of how Allium sativum effects the soil it is grown in such as soil enzymes or an investigation of how Allium sativum effects the growth of other plants this way. There could also be soil research into the levels of sulfur in soil for garlic as a it is considered a key nutrient in growing it due to that being a key component in the smell of garlic.

References

Attia, S., Grissa, K., Lognay, G., Heuskin, S., Mailleux, A. C., & Hance, T. (2011). Chemical Composition and Acaricidal Properties of Deverra scoparia Essential Oil (Araliales: Apiaceae) and Blends of Its Major Constituents Against Tetranychus urticae (Acari: Tetranychidae). Journal of Economic Entomology.

Attia, S., Grissa, K., Lognay, G., Mailleux, A., Heuskin, S., Mayoufi, S., & Hance, T. (2011). Effective concentrations of garlic distillate (Allium sativum) for the control of Tetranychus urticae (Tetranychidae). Journal of Applied Entomology.

Adem, B. E., & Tadesse, S. T. (2014). Evaluating the Role of Nitrogen and Phosphorus on the Growth Performance of Garlic (Allium Sativum L.). Asian Journal of Agricultural Research, 221-217.

Mischeck, D., & Katsaruware, R. (2014). Onion (Allium cepa) and garlic (Allium sativum) as pest control intercrops in cabbage based intercrop systems in Zimbabwe. IOSR Journal of Agriculture and Veterinary Science.

Soil Enzymes. (2011, September 19). Retrieved from Soil Quality for Environmental Health: http://soilquality.org/indicators/soil_enzymes.html

.

NOTES

NOTES

NOTES

SCIENCE JOURNAL ENTRIES

Name/County	Competition and Awards
Hayden Anderson/Lowndes	Region V 1st Place Plant Sciences, 2nd Place State Science Fair, Alternate ISEF Recipient Special Award: Society for In Vitro Biology
Maggie Buck/Tate	Region VII 1st Place Plant Sciences, 4th Place State Science Fair
Victoria Callahan/Madison	Region II 3rd Place Animal Sciences, 1st Place State Science Fair
Shanay Desai/Madison	Region II 1st Place Biomedical/Health Science, 2nd Place State Science Fair, ISEF Recipient
Nicholas Djedjos/Madison	Region II 2nd Place Mathematics/System Software; 2nd Place State Science Fair Special Awards: United States Agency for International Development, Yale Science and Engineering Award, Mu Alpha Theta Award
Jackson Flowers/DeSoto	Region VII 1st Place Mathematics/System Software; 3rd Place State Science Fair
Raeed Kabir/Harrison	Region I 1st Place Behavioral/Social Sciences; ISEF Recipient Special Awards: Outstanding Research in Psychological Sciences, American Psychological Association Award, Mu Alpha Theta Award, Yale Science and Engineering Award Junior Science and Humanities Symposia: 1st Place
Michael Lu/Oktibbeha	Region V 1st Place Mathematics/Systems Software;1st Place State Science Fair, ISEF Recipient Special Award: Mu Alpha Theta Award Science Talent Search top 300 Finalist
Skylar Nguyen/Jackson	Region III 1st Place Earth/Environmental Science; 2nd Place State Science Fair, ISEF Recipient
Vidhi Patel/Pike	Region I 1st Place Behavioral/Social Sciences; ISEF Recipient Special Awards: Outstanding Research in Psychological Sciences, American Psychological Association Award, Mu Alpha Theta Award, Yale Science and Engineering Award Junior Science and Humanities Symposia: 1st Place
Maddie Raynor/Harrison	Region III 2nd Place Earth/Environmental Science; 4th Place State Science Fair
Hailee Sexton/Harrison	Region III 2nd Place Earth/Environmental Science; 4th Place State Science Fair
Aaron Wan/Oktibbeha	Region V 1st Place Biomedical/Health Science; 1st Place State Science Fair ISEF Recipient Special Award: US Agency for International Development Award
Jessica Yan/Oktibbeha	Region V 1st Place Earth/Environmental Science; ISEF Recipient Special Awards: American Meteorological Society, Association For Women Geoscientists, NASA Earth System Science Award, NOAA Award, RICOH Sustainable Development Award, US Stockholm Junior Water Prize, Yale Science/Engineering Association Award, Justice Manning Award for Environmental Excellence
Andrew Yu/Oktibbeha	Region V 1st Place Biomedical/Health Science; 1st Place State Science Fair ISEF Recipient Special Award: US Agency for International Development Award

MSMS Science Journal
Waves of Science

The MSMS Science Journal is a collection of research projects conducted and defended by the students at The Mississippi School for Mathematics and Science
1100 College Street, MUW-1627
Columbus, Mississippi 39701
The MSMS Science Journal is available to read on our website at www.themsms.org