Rebecca Vidal

Dr. Yoo

Probabilistic Graphical Models

4 November 2022

REVISED Class Project Proposal

Abstract:

The objective of this class project is to expand from my current ongoing Master's degree research project pertaining to different gene families and its potential association with Alpha-1 Antitrypsin Deficiency (AATD). For the purposes of this class, the intention is to ultimately be able to construct a Bayesian Gene Network. As I obtain more knowledge from this course, the exact details of how to approach said will evolve accordingly.

Introduction:

As briefly mentioned in the abstract, the research pertains to AATD. The research aims to analyze zinc transporter gene families of SLC30A and SLC39A, MTF1 gene (Metal Regulatory Transcription Factor 1), and MT (Metallothionein) gene family and its potential association with Alpha-1 Antitrypsin Deficiency (AATD). There are already research articles indicating the association between AATD and metal ion dyshomeostasis (e.g. zinc). Yet, further research needs to be conducted in order to determine if previously mentioned gene families have any association with AATD.

Background:

Alpha-1 Antitrypsin Deficiency (AATD) is a genetic disorder in which there are either low or no levels of the protein "Alpha-1 Antitrypsin" (AAT). AAT is produced within the liver and its purpose is to aid in protecting the lung (e.g. protect the lung from environmental factors such as air pollution). A deficiency of AAT can lead to multiple lung complications / diseases most notably COPD. Nevertheless, deficiency of said can also lead to liver complications such as cirrhosis. Hence, this research aims to further understand any potential correlation between above referenced zinc transporter families and AATD.

Research Questions:

1) From GEO Series, which genes have statistical significance?
2) Of said with statistical significance, which should be of focus for the research project?

3) What is the relationship between these genes and how can that be represented with a Bayesian Network?

Methods:

For the research project, the exact genes and GEO (Gene Expression Omnibus) series have already been established. The series under investigation are as follows: a) GSE 1122, b) GSE141593, c) GSE36478, and d) GSE 36478. Of the four series, the GSE 1122 is the only one from *Homo sapiens* whereas the others are from mouse models. The GSE 141593 series contains data of 5 biological replicates for each of Wild Type, PiZ, and PiZ/Chop-/- mouse livers at both ages 6 weeks and 36 weeks. The GSE 36478 series contains two platforms of data sets as follows: a) GPL 339 platform and GPL 340 platform. The GSE 36478 series consists of both Wild Type and PiZ mouse and distinguishes if female or male. The GSE 93115 series consists of 3 biological replicates for both the Wild Type and PiZ mouse. For each series, review of the following genes: SLC30A1, SLC30A2, SLC30A3, SLC30A4, SLC30A5, SLC30A6, SLC30A7, SLC30A8, SLC30A9, SLC39A1, SLC39A2, SLC39A3, SLC39A4, SLC39A5, SLC39A6, SLC39A7, SLC39A8, SLC39A9, SLC39A10, SLC39A11, SLC39A12, SLC39A13, SLC39A14, MT1, MT2, MT3, MT4, and MTF1.

In short, the research process as follows: a) determine which have any statistical significance from the GEO DataSets already provided on NCBI, b) narrow down from those with statistical significance, and c) test said with liver cells and conduct PCR tests. To determine statistical significance, there were four main GEO Series that were analyzed. Not every GEO Series had all of the genes under investigation. Regardless, for those in which data was available, through Microsoft EXCEL, data analysis of "t-Test: Paired Two Sample for Means" between the wild type mice and PiZ mice. To clarify, the "PiZ" mice are those already with the mutant mice (AAT variant). Some GEO Series had "doubles" of these genes with the same GEO Series. For those, means were calculated and then determined if statistically significant. Afterwards, for those with statistical significance, through Microsoft EXCEL, graphs will be made. Upon analysis of said, the "top" genes to investigate will be narrowed down. Further scientific literature review of said will be conducted, and depending on that, PCR tests will be conducted. However, the latter portion of my research will most likely occur throughout the Spring 2023 semester. There are different types of probabilistic graphical models; however, for the purposes of this course, I will be using Bayesian Networks (BN). BN consisting of both nodes and direct edges. In BN, it essentially shows the relationship between the nodes (i.e random variable) and the direct edges (i.e. conditional probability). Evidently, this can help indicate any potential relationship between each gene and symptomology / disease. I will have further details once I have a firmer understanding of how to construct said. Thus, for the time being, a portion of my "Timeline" indicates "TBD".

Timeline:

Week of October 24th: Finalize calculations of means of those genes “doubled” within a GEO Series and finalize the graphs of those with statistical significance

Week of October 31st: Weekly update meeting with Dr. Liuzzi. During said meeting, reviewed current data and discussed to finalize data analysis for GSE 141593 series. In the process of confirming meeting with Zhenghua for her help in extracting and reviewing the GEO DataSets.

Week of November 7th: Will have weekly meeting with Dr. Liuzzi and finalize which genes to focus on for the purposes of the research project. Upon decision of said, then I will begin working on constructing the Bayesian Gene Network. Furthermore, plan to meet with Zhenghua.

Week of November 14th: (TBD)

Week of November 21st: (TBD)

Week of November 28th: (TBD)

Week of December 5th: Finalizing and proof-reading any last details of the project

References

About alpha-1 antitrypsin deficiency. Genome.gov. (n.d.). Retrieved October 18, 2022, from https://www.genome.gov/Genetic-Disorders/Alpha-1-Antitrypsin-Deficiency

Bayesian network. Bayesian Network - an overview | ScienceDirect Topics. (n.d.). Retrieved November 1, 2022, from https://www.sciencedirect.com/topics/mathematics/bayesian-network

Emwas, A.-H., Alghrably, M., Dhahri, M., Sharfalddin, A., Alsiary, R., Jaremko, M., Faa, G., Campagna, M., Congiu, T., Piras, M., Piludu, M., Pichiri, G., Coni, P., & Lachowicz, J. I. (2021). Living with the enemy: From protein-misfolding pathologies we know, to those we want to know. *Ageing Research Reviews*, *70*, 101391. https://doi.org/10.1016/j.arr.2021.101391

Koller, D., & Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. Massachusetts Institute of Technology.

Wang, L., Audenaert, P., & Michoel, T. (2019). High-dimensional bayesian network inference from systems genetics data using genetic node ordering. *Frontiers in Genetics*, *10*. https://doi.org/10.3389/fgene.2019.01196

035411 - piz strain details. (n.d.). Retrieved October 21, 2022, from https://www.jax.org/strain/035411