ISBDS-CFDE Training Course, 2024

Offered by the UNM SoM DoIM Translational Informatics Division

Research Plan

Template #1

Student name:	
Student email:	
Student affiliation:	
Advisor name:	
Advisor email:	
Date:	

Students are welcome to use this document as a template for their own customized independent study research plan by changing some elements and/or by adding additional research objectives. All students must complete a preliminary research plan containing the sections included below and submit it for instructor approval.

Project Title

The title should be descriptive and comprehensible to the intended audience.

Exploring genes positively and negatively regulated by the asthma drug salmeterol using the Condensed Knowledge Graph (KG)

Problem statement

The problem statement provides a clear and concise description of the issue that the project will address.

Completion of this project will impart experience with Neo4j desktop, Cypher, the National Library of Medicine's NCBI Gene lookup page, and the CFDE Data Portal, which are powerful tools for studying associations between drugs, genes, and diseases. The deployment of a condensed KG within Neo4J will be used to explore associations between drugs, genes, and diseases. Specifically, genes that are positively and negatively regulated by the asthma drug salmeterol will be identified from the condensed KG and further explored using other resources.

Data sources

Source and dataset specification sufficient to facilitate verification of accessibility by the student for the project.

- Condensed Knowledge Graph (dumpfile is available to students)
- National Library of Medicine's NCBI Gene lookup page (https://www.ncbi.nlm.nih.gov/gene)
- CFDE Data Portal (https://data.cfde.cloud/)
- PubMed (https://pubmed.ncbi.nlm.nih.gov/)
- Google Scholar (https://scholar.google.com/)

Importance

Why is your project problem important to biomedical science, medicine, public health or others? Cite relevant references if possible.

Asthma is a chronic inflammatory airway disease and is the most common chronic lung disease worldwide[1]. Investigating the links between human genes and existing asthma drugs is an important pursuit that could motivate future research into more effective treatments for asthma.

This project is also important because it expands the skills of the student to pursue future biomedical research that involves accessing and exploring data from knowledge graphs combined with other resources.

1) National Asthma Education and Prevention Program. Expert Panel Report 3 (EPR-3): Guidelines for the Diagnosis and Management of Asthma-Summary Report 2007. J Allergy Clin Immunol. 2007 Nov;120(5 Suppl):S94-138. doi: 10.1016/j.jaci.2007.09.043. Erratum in: J Allergy Clin Immunol. 2008 Jun;121(6):1330. PMID: 17983880.

Objectives

Provide a list of the specific answers, results, datasets, tools, and/or other deliverables that will be generated by completing the project.

- Installation of Neo4J Desktop
- Learn to instantiate and query the condensed KG via Neo4J on a local machine
- Create a Cypher query to find disease nodes related to asthma, with a focus on "allergic asthma"
- Using Cypher queries on the condensed KG, identify about a dozen genes that are positively regulated by salmeterol, and about a dozen genes that are negatively regulated by salmeterol.
- For each of the genes identified as being positively and/or negatively regulated by salmeterol, tables will be created which list the name of each gene, any corresponding protein(s), protein function, top expression tissues, and a column for notes.
- Explore a number of the genes positively and negatively regulated by the asthma drug salmeterol using NIH resources. Identify and discuss any trends or patterns between the types of proteins the genes encode, expression profile sample types, or functional associations described in PubMed article content.

Methodology

Methodology defines the methods and logic steps that will be taken to solve the project problem and to achieve proposed objectives. Software should be specified, particularly any specialized scientific software tools.

The provided instructions for installing neo4j and accessing the condensed KG will be used to explore salmeterol as a treatment for the disease “allergic asthma.” Cypher queries will be developed to return genes that are regulated by a compound (salmeterol) linked to allergic asthma via drug indication.

For example, a cypher query exploring negative regulation of salmeterol takes advantage of the “negatively_regulated_by” relationship between a compound and a gene:

```
match (d:Disease)-[r1:indication]-(c:Compound)-[r2:negatively_regulated_by]-(p:Gene)
where d.node_label contains "Allergic asthma"
return * limit 15;
```

Next, we explore positive regulation using the same approach to generate a list of genes. The list of positively regulated genes will be compared to the list of negatively regulated genes.

To further study the specific genes identified via cypher queries, the CFDE Data Portal located at https://data.cfde.cloud/, PubChem and PubMed will be used. Additionally, the National Library of Medicine’s NCBI Gene lookup page (https://www.ncbi.nlm.nih.gov/gene) will also be used to investigate each of the genes identified from the condensed KG as either positively or negatively regulated by the asthma drug salmeterol. This resource can be used to learn more about the relevant gene products (proteins), access a tissue expression profile, reveal a list of relevant PubMed articles, and more.
For each of the 13 genes identified as being positively or negatively regulated by salmeterol, the previously-described resources will be used to construct a table listing the name of each gene, any corresponding protein(s) and their function, top expression tissues, and a column for notes. Based on these observations, any apparent trends will be identified and discussed.

The table will be used to address the following questions:

- Is there any apparent trend between the types of proteins the genes encode (i.e. transmembrane proteins, hydrophobic proteins, structural proteins, enzymes, etc) among the two classes of genes?
- Are the top expression profile sample types consistent among genes within each classification, and also between the two classes of genes?
- Do the results of this investigation suggest a specific mechanism of action for salmeterol as a treatment for allergic asthma?

End of Document