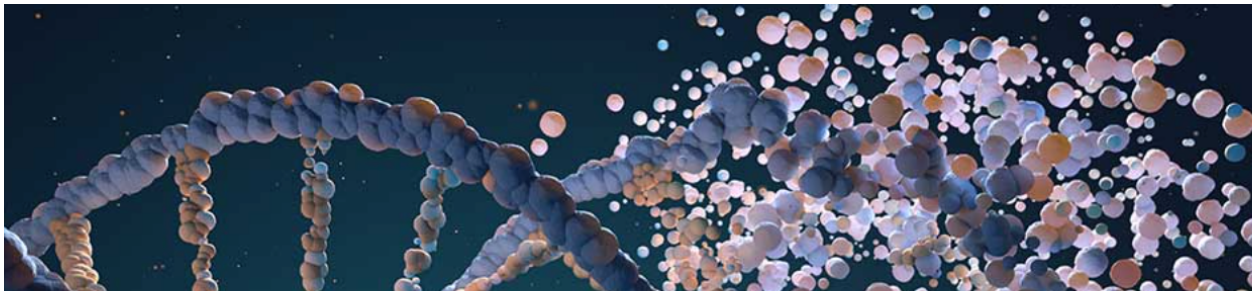
Bioinformatics: Gene-Disease Research in the Post-Genomic Era
DS3000 / DS5110 (Rachlin)



Introduction and Getting Started

Bioinformatics is an interdisciplinary field that combines biology, computer science, and data science to analyze and interpret biological data. In this assignment, you are a data scientist tasked with exploring gene-disease association catalogued by DisGeNET: a platform for studying the connection between genes and human disease. (To say that a gene and a disease are *associated* means that there is believed to be some connection between a genetic variant and the risk of developing that disease.)

First, take the tables we have given you and construct a relational database. I recommend you build a simple two-table SqLite database, using the **csvsql** command that comes with **csvkit** – a library you can install using **conda**. Next, open DBeaver or a database client of your own choosing and establish a connection to your database. Write and refine queries that answer the questions listed below. Pull the resulting tables into a Pandas DataFrame to build a visualization of your data. To carry out this assignment, you should know basic SQL syntax: **SELECT, DISTINCT, FROM, JOIN, WHERE, GROUP BY, ORDER BY, LIMIT, IN,** and **AND**. You may do some of your processing with SQL, and some using Python. How you split up the work is up to you. But I found it easiest to do most of the work with SQL.

Documentation on the table columns can be found online:

DisGeNET: https://www.disgenet.org/dbinfo
Gene Ontology (GO): http://geneontology.org/docs/go-annotations/

Research questions (10 points each)

1. Investigate the pace of progress in biomedical research. How many gene-disease associations were cataloged and how many *total* papers were published between 1960 and the present? Plot the *cumulative total* number of associations and published papers year by year from 1960 to the present. (Hint: use the last_year column). Write a short essay explaining how the completion of the Human Genome Project in 2003 (which began the "Post-Genomic Era") accelerated research in gene-disease associations. (This can be embedded in your code). Your visualization should reinforce your historical summary.

2. What 10 genes have the greatest number of associations? Report the gene symbol and name, and the number of associations and total number of publications. No plot is required – a simple table is sufficient. Repeat your analysis – but now find the *diseases* with the most number of associations. Exclude **disease_type = 'group'** from your analysis. (Do you notice a common theme?)

3. If we think of genes and diseases as vertices and associations as edges linking genes to disease, we can apply the tools of network analysis to explore the topology of this vast network. The *degree* of a node is the number of associations that gene or disease is involved in. Plot the degree distribution of genes on a log-log scale to show that it is a *scale-free* distribution. Overlay the degree distribution of diseases. They, too, follow a scale-free distribution!

4. Identify the 300+ genes that are strongly associated with Alzheimer's Disease. We define "strongly associated" as having an Evidence Index (EI) >= 0.667 based on 11 or more publications. (The evidence index is the fraction of publications that conclude that there IS a link between the gene and the disease.). Output the table of genes – it is ok to show the top 10 genes with the most publications.

5. For these 300+ genes, plot the Disease Pleiotropy Index (DPI) vs. the Disease Specificity Index (DSI) as a scatter plot. Use **plt.text(x,y,label)** to annotate your scatter plot with the symbols for four genes: APOE, APP, MAPT, and CALHM1. The first three are mentioned in the most publications. CALHM1 has the highest DSI – is very specific to Alzheimer's – and stands out as an interesting outlier among these 300 genes! Setting the size of the marker to the number of publications, and the color to the Evidence Index (EI) makes for a very interesting plot – give it a try. My call to the scatter plot function looks like this:

   plt.scatter(alz.DSI, alz.DPI, c=alz.EI, s=alz.num_pubs, cmap='viridis')

6. According to GO, what biological processes are these Alzheimer's-linked genes most frequently involved in? Provide a table with the GO ID, the qualifier, the Go Term, and

the number of genes that are involved in that process. You'll want to use the **DISTINCT** keyword in your select, because the same gene may have been annotated with the same GO Term multiple times according to difference evidence codes.

7. Take the 300+ Alzheimer's-linked genes you found earlier and now figure out what *other* diseases these genes are also associated with. Retain the requirement that EI >= 0.667 and num_pubs >= 11. Also, again exclude diseases where **disease_type = 'group'**. Rank the diseases by how many Alzheimer's-linked genes are involved and report the top 10 genes in a table.

8. Visualize Alzheimer's genes and the top-10 Alzheimer's-related diseases as a graph using the NetworkX library. I recommend you rank associations by number of publications and limit yourself to the first 200 associations – otherwise your network will start to look like a real hairball! Here is my code to draw the network.

```
net = pd.read_sql_query(query, con)
plt.figure(figsize=(15,15), dpi=100)
G = nx.from_pandas_edgelist(net, 'gene_symbol', 'disease_name', create_using=nx.Graph())
nx.draw_networkx(G, with_labels=True, node_size=50)
```

Finding the query is up to you! You might find it helpful to use **VIEW**s.

Open research initiative (20 points)

Ask a research question of this amazing catalog, produce a visualization of your answer, and caption your visualization inside a single **ONE-page** PowerPoint slide. (A template will be provided.) You will be graded on how interesting we find your question, how effectively you visualize your answer, and the clarity of your caption. These PowerPoint slides will be compiled for a future class discussion.

What to submit

- Your Python code (.py) or Jupyter Notebook (.ipynb)

- All required tables and visualizations (if not already visible in your Jupyter Notebook.)

- The single PowerPoint slide (.pptx) for your open research initiative. Please also submit this slide as a one-page .PDF so that the TAs can read it from within GradeScope without having to download it to their local drive.