CAMM 535 – Fundamentals of Biological Databases

FINAL PROJECT

Final Report 1 Submission Due Date: January 13, 2026, at 23:59

Final Report 2 Submission Due Date: January 20, 2026, at 23:59

Project Demo Session: January 21, 2026 (attendance is required for the final project to be graded.)

For the Project Demo Session, reserve your team's spot at
CAMM535 - Demo Session - Time Slots

Each team has been assigned a specific disease for this project (refer to the list provided at the end of this document).

The aim of this project is to use multiple biological databases and data analysis tools and connect them together to design a new database.

Live Demo Session Information

- You will present and demonstrate your database in a **one-on-one live demo session** with the instructor on **January 21**.
- The demo session will have a **maximum duration of 15 minutes**.
- During the session, your database must be fully operational and capable of executing the assigned SQL queries in real-time.

Evaluation

- The instructor will assess your performance based on the functionality, accuracy, and efficiency demonstrated during the live session.
- **Note**: The instructor reserves the right to determine the final score based on your demo performance.

Important: Groups that do not attend the demo session on January 21 will not receive any credit for their submitted reports, regardless of the quality of the work.

Please prepare ONE slide discussing the following to be presented (max 2 minutes) in the demo session on January 21:

- Part 1: Brief description of the phenotype/disease
- Part 2: Database scheme (tables and links) and data sources (for each table)
- Part 3: Problems/difficulties during the project

Notes on Final Report 1 and Final Report 2

While preparing your Final Report 1 and Final Report 2, be careful about the points listed below;

- Your final report should cover the description of data retrieval steps and database construction steps.
- Add the conceptual design of the database to your report.
- Describe database, tables and table details properly.
- All files you have used during visualization and database preparation steps must be submitted with your final report.
- Each figure should have a proper figure legend.
- Add screenshots from the database interface to your report to describe the database better.
- Please make your report informative and clear for a balanced and appropriate assessment. The report should be understandable and self-explanatory.
- Copying or reviewing the database or copying the design of another, joint development/debugging and sharing your design are not permitted.
- In the final report submission, 40 points deductions will be applied for one day late. Reports submitted more than one day late will not be graded.

PROJECT DESCRIPTION

In this project, you will prepare two reports; one is for data acquisition and conceptual design of the database (Report 1) and the other is for physical design and testing of your database (Report 2).

Final Project Report 1

Detailed Instructions – Understanding the Disease, Data Acquisition and Conceptual Design

Part 0: (10 points) Report format (including figure numbering, figure captions, properly citing the figures in the report)

Part 1: (10 pts) Begin your project by writing a brief introduction of the assigned disease phenotype that outlines its clinical characteristics, biological relevance, and any known genetic associations. This step sets the stage for your subsequent analyses. Use credible scientific sources such as peer-reviewed journal articles, textbooks, or reputable online databases (e.g., PubMed). Include **at least one figure** (e.g., a pathway diagram, disease prevalence) to visually support your introduction. Add a descriptive figure caption below each figure and briefly explain what the figure shows.

Part 2: (10 pts)
Retrieve a list of gene symbols associated with your phenotype using **BioMart**. Then, use your gene list as a multiple input in the **STRING** database (human) and expand your network using

the first shell with a maximum of 100 interactors. Next, download all genes in your expanded network using an appropriate export option. Use this expanded gene set to retrieve and analyze short genetic variants such as single nucleotide polymorphisms (SNPs) and small insertions/deletions (indels) associated with these genes.

This step involves using **Ensembl/BioMart** and leveraging associated human genes and retrieving variations from **dbSNP** data associated with these human genes. Select attributes like gene symbols, gene IDs (from external resources), chromosome locations, functional annotations, and variation IDs (and maybe more) to ensure your mapping is comprehensive. Think critically about the downstream use of this table, ensuring it integrates seamlessly with tables created in later steps. (Hint: Use Ensembl Genes Database to retrieve phenotype associated genes and the variants on them)

Double-check that all variations are correctly mapped to human genes that you exported from STRING.

Document any challenges or ambiguities encountered during data retrieval or mapping.

Additional Notes:

- Use the filters and options provided in Ensembl/BioMart to refine your query and retrieve high-quality, gene list specific variation data.
- Clearly describe your data retrieval and mapping methodology in your report to ensure reproducibility.

Part 3: (10 pts) In this step, you will retrieve genomic coordinates and related information for the RefSeq genes identified in the previous step. Using the **UCSC Table Browser**, you will extract precise data about these genes and organize it into a structured table in your database.

Go to the **UCSC Table Browser** tool. Input the list of RefSeq IDs or gene symbols retrieved in the previous step to filter your query. Specify the output format (e.g., tab-separated values) and download the resulting table. Select Relevant Attributes, such as: Gene Symbol, RefSeq ID, Chromosome, Genomic Start and End Positions, Strand (+/-), Exon Count, Transcript Length and maybe more depending on your database design.

Review the requirements of downstream steps to ensure you include all necessary fields.

Part 4: (10 pts) Mapping Phenotype-Related Genes to Human Proteins

In this step, you will connect the genes to their corresponding proteins using the **UniProt** database. This will provide a critical link between genetic and proteomic information, enriching your dataset with key protein attributes. Proteins are the functional entities encoded by genes, and understanding their attributes can provide insight into the molecular mechanisms underlying the phenotype.

Use the UniProt ‘ID Mapping’ Tool to limit your search space and specify the source ID type (e.g., RefSeq, Gene Symbol) and map it to UniProt/SwissProt IDs for human proteins.

Then, select and download the following attributes for the mapped proteins: UniProt/SwissProt

ID, Protein Length, Primary Gene Name, PDB ID, Molecular Mass (in Daltons), Protein Function (e.g., description of biological roles), Subcellular Location (e.g., cytoplasmic, nuclear), Protein Family/Pathway Information.

Double-check the mappings to ensure there are no missing or mismatched entries.

If some genes do not map directly to a protein, note these cases in your report and consider potential reasons (e.g., non-coding genes or incomplete annotations).

Clearly describe the mapping process in your report, including: Mapping tool settings and filters and any challenges encountered (e.g., ambiguous mappings) and how you addressed them.

Part 5: (20 pts) Do an advanced search for the phenotype that has been assigned to your project in **GEO datasets**. Consider human datasets released since Jan 1, 2005 to present and select one of the sets from the resulting list. Describe the selected dataset in your report. Analyze it with GEO2R. Create a table in your database to deposit the result of GEO2R analysis.

Part 6: (10 pts) To make your database more comprehensive, collect at least one additional data table from other databases and describe them in your report. (Hint: If you are having trouble linking these tables, you can add an extra “cross-reference” table obtained from either BioMart or UCSC Table Browser)

Part 7: (20 pts) Conceptual Design: Explain the conceptual design of your database by creating an **Entity-Relationship (ER) diagram** that visualizes the relationships between tables (Cardinality ratios must be added). Also, define the primary and foreign keys in each relation in the database.

Final Project Report 2

Detailed Instructions – Physical Design, Testing and Visualization

(10 points) Report format (including figure numbering, figure captions, properly citing the figures in the report)

Part 1: (10 pts) Recap the objectives and summarize the key findings from the first report. You are free to add figures to better illustrate the first report as an introduction.

Part 2: (30 pts) Construct your database with the collected data table and conceptual design in the first reporting period. Use phpMyAdmin for database construction and add screenshots for each step in the database construction. Define the primary and foreign keys in each relation in the database.

Part 3: (5 pts) Explain the database constraints available in the physical design of your database.

Part 4: (9 pts) Write down at least 3 queries and SQL commands that will require at least 2

tables and include their results to demonstrate that your database has been constructed properly.

Part 5: (6 pts) Write an SQL command that will give an output of variation name, PDB IDs, Uniprot/SwissProt IDs, RefSeq IDs, exon counts, logFC value of a given gene symbol. Add a demo of your database with the query of different gene symbols to your final report. In the demo, select three case studies which can be a protein or a gene.

Part 6: (3 pts) Write down an SQL query to retrieve a list of genes having PDB IDs and exon count is greater than or equal to 4.

Part 7: (7 pts) (a) Write down an SQL query to count a list of genes whose adjusted p-value in GEO2R is smaller than 0.05. How many genes are significant? If the number of genes is less than 30, find alternative ways to increase this number (e.g, using less stringent cutoffs, selecting top 100 genes etc.) and use this larger set for further analysis.

(b) Write down a query to retrieve a list of Uniprot IDs of the genes found in part (a) and show the output screenshot in your report.

Part 8: (20 pts) Find the protein interactions among the resulting list in Part 7(a) by referring to **STRING**. Visualize these protein interactions in **Cytoscape** in force-directed layout. Color down-regulated genes in blue and up-regulated ones in red. (Hint: Here, suppose that genes having negative logFC values are down-regulated, positive logFC values are up-regulated.) Arrange the sizes of the proteins using the length attribute so that the longer the protein, the lower its size (e.g., inversely correlated). Add your network figure both to your presentation and to your final report.

Deliverables:

Your final report must include all required sections outlined in the project description. The report should be formatted and include clear and concise explanations for each step of your project, figures and screenshots, such as network visualization, screenshots of your database schema and queries, along with query results.

Export your database as an SQL file and submit it along with your report.

Team No	Disease
Team 1	Type 2 Diabetes Mellitus
Team 2	Myelodysplastic Syndrome
Team 3	Rhabdomyosarcoma
Team 4	Schwannoma
Team 5	Ulcerative Colitis
Team 6	Cystic Fibrosis
Team 7	Behcet Disease
Team 8	Hodgkins Lymphoma
Team 9	Diamond-Blackfan Anemia
Team 10	Langerhans Cell Histiocytosis
Team 11	Leber congenital amaurosis
Team 12	Fanconi anemia