Project Narrative: Respiratory Pathogen Information Hub

Background

Researchers working with respiratory pathogens often need to navigate through numerous sources to obtain detailed gene annotation data and general pathogen information. As someone who works in the field of respiratory disease diagnostics, I find this process to be at times repetitive and time-consuming. As a result, I recognized the need for a centralized platform where all this essential information is easily accessible. Therefore, the goal of this project is to bridge this gap by creating a comprehensive information hub for various respiratory pathogens. This platform aims to make general information about pathogens, such as symptoms of infection, and detailed gene annotation data, including coordinates and coding sequences, readily available with just a few clicks, providing quick information access for professionals and anyone seeking knowledge about respiratory diseases.

Implementation & Challenges

Database

Since this project is based on a respiratory pathogen database, an Entity Relationship Diagram (ERD) was created to define the structure and relationships within this database (Figure 1). Throughout the implementation process, I continuously optimized the design to improve simplicity and efficiency. For instance, I initially structured the database with separate tables for "Other_names" and "Major_subtypes," but later changed them into attributes within the "PATHOGEN" table for simpler data retrieval for the pathogen search function.

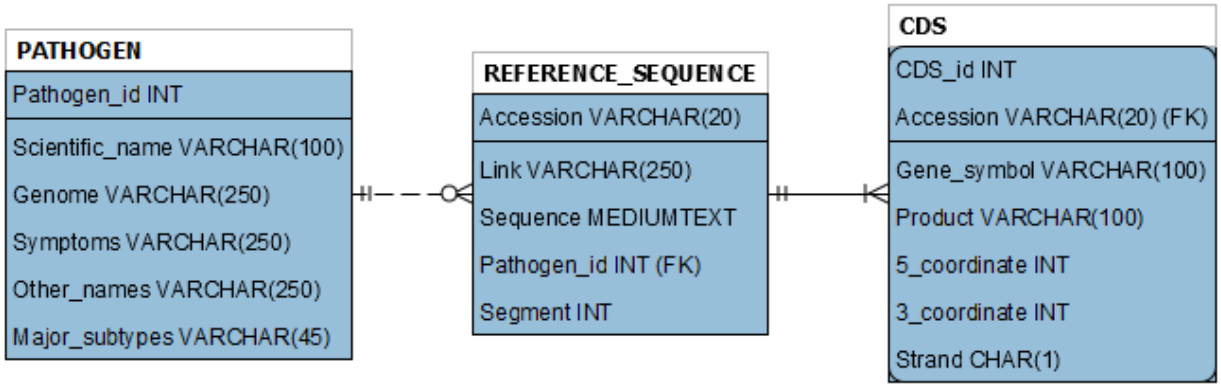

Figure 1. Final version of the ERD for the MySQL database schema.

One of the main challenges I encountered during the database implementation was the population of the “REFERENCE_SEQUENCE” and “CDS” tables. Although I manually inserted 26 respiratory pathogens (Table 1) into the “PATHOGEN” table, it was almost impossible to do the same for their corresponding gene annotations. Bordetella pertussis alone, for example, has over 3000 coding sequences within its 4-million-base-pair genome. Therefore, I developed an ‘extract.py’ script to extract the accession number, reference sequence, coding

sequence coordinates, gene symbols, products and strands from GenBank files and insert them into the MySQL database. One problem I ran into during this process was that GenBank file layouts are not always consistent. For example, some GenBank files have product names spanning multiple lines, while most others have product names only taking up one line. In addition, since I wanted to use CDS entries as references for gene annotations, it was important to prevent the script from mistakenly extracting information from other entries like gene or mat_peptide. To address this issue, I reviewed and tested many GenBank files and was able to refine the 'extract.py' script to accurately extract data from GenBank files with all types of layouts.

Pathogen_id	Scientific Name	Pathogen_id	Scientific Name
1	Human bocavirus 1	14	Chlamydia pneumoniae
2	Human rhinovirus A	15	Influenza A virus
3	Human rhinovirus B	16	Influenza B virus
4	Human rhinovirus C	17	Human coronavirus HKU1
5	Middle East respiratory syndrome-related coronavirus	18	Legionella pneumophila
6	Human coronavirus NL63	19	Mycoplasma pneumoniae
7	Human respirovirus 1	20	Human coronavirus OC43
8	Human respirovirus 3	21	Human orthorubulavirus 2
9	Severe acute respiratory syndrome coronavirus 2	22	Human orthorubulavirus 4
10	Human mastadenovirus B	23	Human metapneumovirus
11	Human mastadenovirus C	24	Human coronavirus 229E
12	Bordetella pertussis	25	Human respiratory syncytial virus A
13	Bordetella parapertussis	26	Human respiratory syncytial virus B

Table 1. The 26 respiratory pathogens in the current database of the Respiratory Pathogen Information Hub.

Web Application

The web application was developed using Python CGI scripts, HTML, CSS, and JavaScript to provide a user-friendly interface for accessing pathogen information and gene annotations.
One challenge I encountered during this process is to establish a link from the highlighted "View Gene Annotation" text to the respective gene annotation page. The solution I came up with is to have 'pathogen_search.cgi' return a URL to the 'gene_annotation.cgi' script containing the appropriate pathogen ID and scientific name along with other pathogen information.
Another challenge was that, unlike other respiratory pathogens, influenza viruses have segmented genomes. Therefore, to accurately and comprehensively display the gene annotation data of influenza viruses, the 'gene_annotation.cgi' and 'gene_annotation.html' files were updated to evaluate whether a pathogen has segmented genomes. They can now dynamically

adjust the layout of the gene annotation page to include an additional “Segment” column and allow multiple accessions to be displayed when required.

Validation

Extensive validation was conducted to verify the accuracy of the information stored in the MySQL database. I cross-referenced the gene annotations shown in this web application with the NCBI database to validate coordinates, strands, gene symbols, products, and coding sequences. All functions of the web application, including pathogen search, gene annotation viewing, and NCBI link redirection, were tested with different combinations of inputs and selections.

Discussion

There have been some minor adjustments from the project proposal to the actual implementation. For example, the name of this web application has been changed from “Respiratory Disease Information Hub” to “Respiratory Pathogen Information Hub” to better reflect its scope and purpose. Moreover, I decided to not proceed with the “Show All” function for the sequences in the gene annotation table. Instead, I prioritized the addition of the “Strand” column, which was initially overlooked, as well as the processing and display of the reverse complement sequences for CDS on the negative-sense strand.