**Project Title:** Gene Annotation and Pathway Analysis Tool

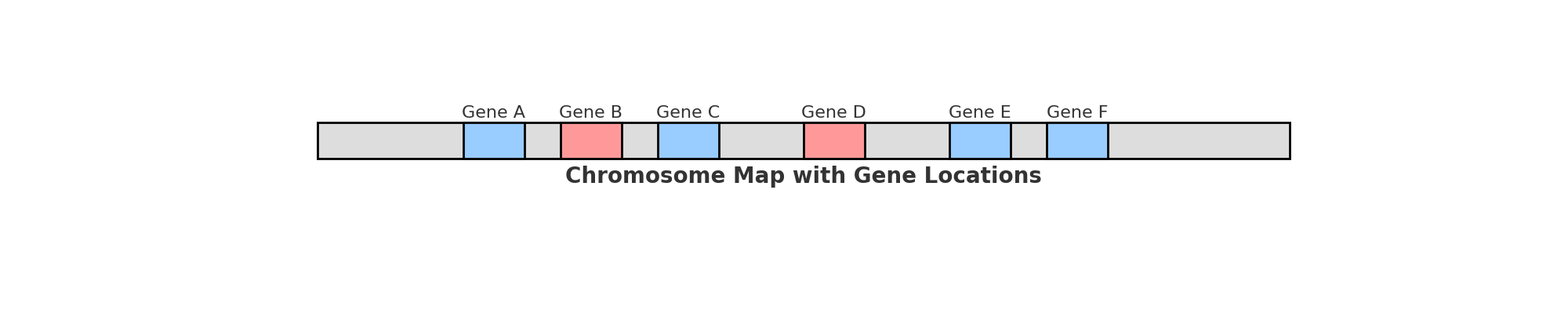
**Tool Background**  
Genomic data analysis is essential for understanding the structure, function, and evolution of genes, especially in bacterial genomes where functional annotations often require intensive study. The Gene Annotation and Pathway Analysis Tool aims to support bioinformatics students and researchers by providing an accessible, intuitive interface to search, view, and analyze gene annotations and pathway information within a specific bacterial genome. This tool enables users to perform keyword-based searches, filter results by pathways, and visualize gene locations within the genome, enhancing the overall gene exploration experience. By offering both basic search and visualization capabilities, this tool reduces the technical overhead of analyzing genome data and helps users make biological inferences based on pathway involvement and gene functions.

**Tool Functionality**

A diagram of a research process

Description automatically generated

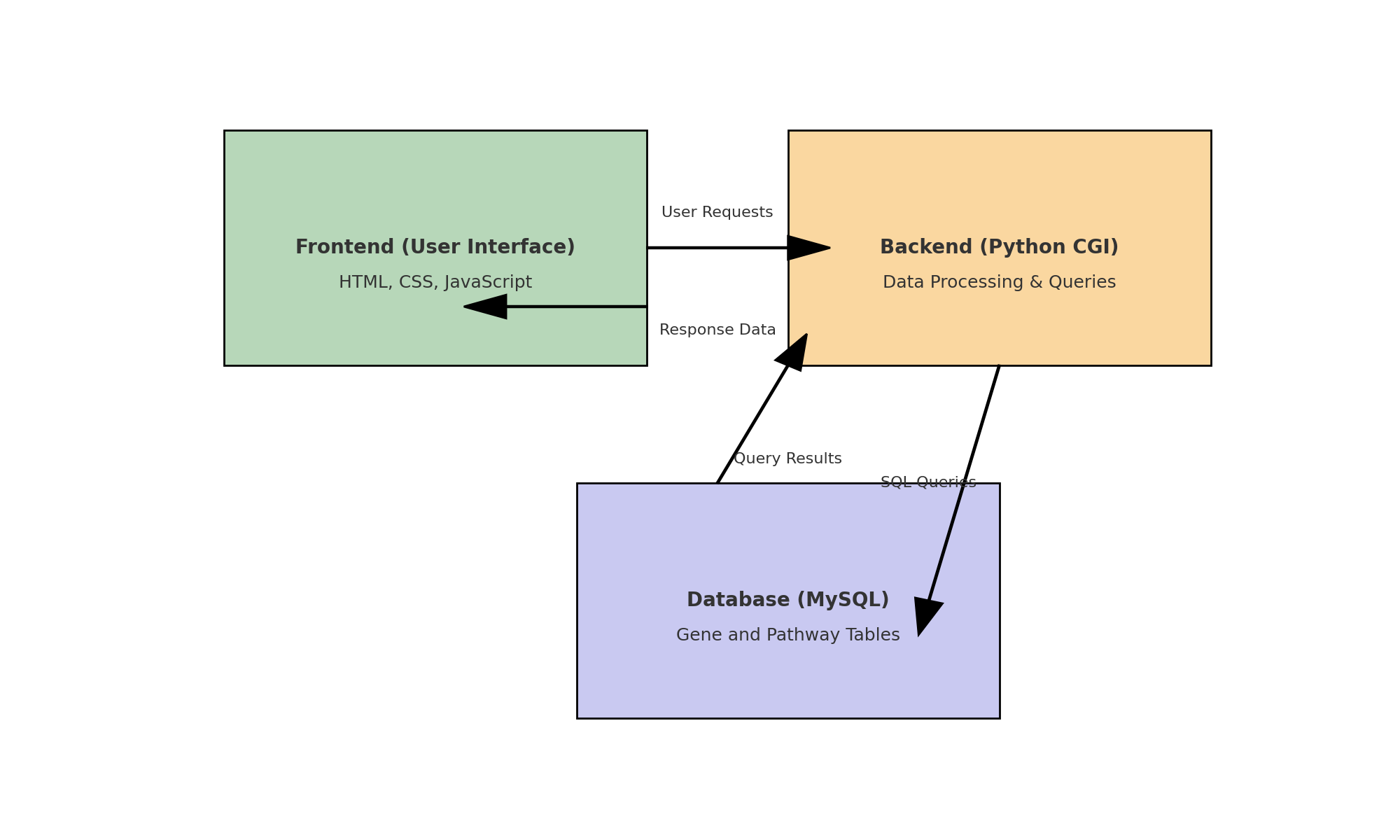
1. **Gene Search**
   * **Basic Search**: Users can search for genes by entering gene IDs, gene names, or relevant keywords, such as pathway names or functional descriptions.
   * **Result Filtering**: Search results can be refined based on pathway information or specific gene functions. For example, users can select genes associated with particular metabolic pathways or filter by functional attributes like "DNA repair" or "oxidative stress response."
   * **Search Suggestions**: To enhance user experience, the tool offers suggestions as users type, based on available gene names and pathways within the database.
2. **Pathway and Functional Analysis**
   * **Pathway Overview**: Users can access a list of pathways (e.g., glycolysis, oxidative phosphorylation) with an overview of each pathway’s role in cellular metabolism.
   * **Pathway-Gene Mapping**: By selecting a pathway, users can view all genes involved in that pathway, organized by function or position within the pathway. Each gene entry is clickable for detailed information, such as gene sequence and start-end coordinates.
   * **Functional Annotation Filtering**: Users can filter the displayed genes within a pathway by functional annotation, helping them focus on specific areas of interest (e.g., enzymes in metabolic pathways or regulatory genes).
3. **Gene Details**
   * **Detailed View**: For each gene, users can view comprehensive information, including start and end positions, nucleotide sequence, amino acid sequence (if applicable), gene description, and pathway associations.
   * **Associated Interactions**: If the gene interacts with other genes or proteins within the pathway, these interactions are also displayed, providing context for gene functionality within larger cellular networks.
   * **Expand/Collapse Details**: Users can expand or collapse individual sections (e.g., gene description, nucleotide sequence) to customize the view, reducing clutter and improving readability.
4. **Basic Visualization**



* + **Chromosome Map**: Gene locations are represented on a simple linear chromosome map, allowing users to see spatial relationships within the genome. This feature provides an intuitive way to understand the physical arrangement of genes.
  + **Highlighted Regions**: Genes related to the selected pathway or function are highlighted on the chromosome map, helping users quickly identify where relevant genes are located.
  + **Interactive Zoom**: Users can zoom in on specific chromosome regions for a closer view of gene clusters, providing insights into potential operon structures or gene neighborhoods.

1. **Interactive Interface**
   * **Asynchronous Loading**: Searches, filters, and data retrieval are managed asynchronously, allowing the interface to load information dynamically without requiring page reloads.
   * **Expandable Sections**: Pathway details, gene annotations, and interaction information can be expanded or collapsed based on user preference, providing a customizable interface.
   * **User-Friendly Design**: The interface is designed with simplicity in mind, utilizing CSS and JavaScript for a clean, intuitive layout that makes navigation straightforward for users with minimal bioinformatics experience.

**Technical Implementation**



1. **Database Design (MySQL)**
   * **Schema Structure**: The MySQL database comprises two primary tables:
     + **Gene Table**: Stores essential gene attributes, including gene ID, name, start and end positions, nucleotide sequence, amino acid sequence, functional description, and pathway associations.
     + **Pathway Table**: Stores pathway names, descriptions, and functional summaries. This table includes foreign key references to genes in the Gene table, creating a relational structure that allows efficient querying and filtering by pathway.
   * **Indexed Fields**: The database uses indexed fields (e.g., gene ID, pathway name) to speed up searches, making data retrieval efficient even with large datasets.

+------------------------+--------------+------+-----+---------+----------------+

| Field | Type | Null | Key | Default | Extra |

+------------------------+--------------+------+-----+---------+----------------+

| gene\_id | int | NO | PRI | NULL | auto\_increment |

| gene\_name | varchar(255) | NO | MUL | NULL | |

| start\_position | int | NO | | NULL | |

| end\_position | int | NO | | NULL | |

| nucleotide\_sequence | text | NO | | NULL | |

| amino\_acid\_sequence | text | YES | | NULL | |

| functional\_description | text | YES | | NULL | |

| pathway\_id | int | YES | MUL | NULL | |

+------------------------+--------------+------+-----+---------+----------------+

+---------------------+--------------+------+-----+---------+----------------+

| Field | Type | Null | Key | Default | Extra |

+---------------------+--------------+------+-----+---------+----------------+

| pathway\_id | int | NO | PRI | NULL | auto\_increment |

| pathway\_name | varchar(255) | NO | MUL | NULL | |

| pathway\_description | text | YES | | NULL | |

| functional\_summary | text | YES | | NULL | |

+---------------------+--------------+------+-----+---------+----------------+

1. **Backend Logic (Python CGI)**
   * **Database Connection**: The backend uses mysql.connector to connect Python CGI scripts with the MySQL database. Each CGI script performs a specific function, such as handling search queries, retrieving gene details, or filtering results.
   * **Query Optimization**: SQL queries are optimized to retrieve only necessary fields. For instance, a search query will return only gene names, IDs, and pathway associations initially, with additional details loaded on demand.
   * **JSON Output Formatting**: The CGI scripts format data as JSON, enabling JavaScript on the front end to dynamically load content. This separation reduces server load and improves responsiveness.
2. **User Interface (HTML5, CSS, JavaScript, JQuery)**
   * **HTML Templates**: HTML templates are structured for various sections (e.g., search results, gene details, pathway views), ensuring that CGI code remains separate from presentation logic.
   * **Asynchronous Data Loading with JQuery**: JQuery handles asynchronous requests, enabling the interface to update in real time based on user actions, like filtering or pathway selection, without requiring page reloads.
   * **CSS Styling and Layout**: CSS is used to create a clear and readable layout, with sections styled for optimal usability. CSS also manages the chromosome map, allowing for simple yet effective visualization of gene positions.
   * **Interactive Elements**: JavaScript enables dynamic elements, such as expandable gene details and pathway lists, as well as chromosome map interactions, enhancing the overall user experience.

**Code Organization**

1. **Separation of Concerns**
   * **Modular Backend Scripts**: Each Python CGI script performs a single function, such as processing searches or retrieving gene details. This modular structure enhances code readability and maintainability.
   * **Organized File Structure**: HTML templates, JavaScript, and CSS files are stored in separate directories, allowing for easy updates and adherence to web development best practices.
2. **Documentation**
   * **README**: Includes setup instructions, descriptions of each script and file, and usage guidelines. It also provides a brief overview of the database schema and sample queries.
   * **Inline Comments**: Scripts include comments explaining each section, improving readability and aiding future maintenance.

**Technical Justification**

1. **Database Design**
   * **Relational Schema**: The MySQL relational schema efficiently organizes gene and pathway information, allowing complex queries like pathway-specific filters or keyword searches to be executed with minimal performance impact.
2. **Modularity and Content Separation**
   * **Structured HTML Templates**: HTML templates separate presentation from logic, enhancing modularity and readability.
   * **JQuery for Asynchronous Requests**: Asynchronous capabilities allow smooth and responsive user interactions, reducing server load and providing a seamless experience.
3. **Usability and Visualization**
   * **Chromosome Mapping**: Basic chromosome visualization offers spatial context to gene annotations, handled by CSS and JavaScript to maintain a lightweight tool.

**Expected Deliverables**

1. **Deployed Tool on Class Server**
   * The tool will be fully deployed on the class server, providing search, filtering, and pathway analysis features as described.
2. **Code Submission**
   * A GitHub repository will include all project files, such as CGI scripts, HTML templates, JavaScript, CSS, and MySQL setup scripts.
3. **Documentation (README)**
   * A README file will provide setup instructions, usage guidelines, and descriptions of the project’s file structure, as well as sample data for testing.

**Future Extensions**  
This tool could be expanded in the future to support additional bacterial genomes or more complex visualizations, such as interactive pathway maps. It could also incorporate cross-referencing with external databases, providing more comprehensive insights into gene function and pathway interactions.