THE UNIVERSITY
of EDINBURGH

Report on

Biological Database Class assignment 2024

Student No:B269797

Introduction

The integration of gene annotation data, protein information, and pathway analysis plays a crucial role in understanding biological systems, identifying disease mechanisms, and exploring therapeutic targets(Jones, Sternberg, & Thornton, 2006). In modern bioinformatics, effective integration of various datasets is essential for enhancing the depth of biological research and driving personalized medicine. This report focuses on the process of integrating gene annotations from Ensembl, protein information from UniProt, and pathway data from KEGG into a MySQL database for comprehensive analysis.

The goal of this project is to develop a bioinformatics pipeline that allows the retrieval of gene annotations, protein data, and pathway information, and stores them in a relational database. This enables further exploration and analysis of gene function, protein interactions, and the involvement of genes in specific biological pathways.

By leveraging publicly available databases such as Ensembl, UniProt, and KEGG, we can retrieve rich biological data, including gene descriptions, protein sequences, enzyme functions, and pathway associations. This report discusses the methodology used for data retrieval, integration, error handling, and the challenges faced during the implementation. Additionally, we describe how the data is structured in a MySQL database to allow for efficient querying and data analysis.

Objectives

Retrieve gene annotations from Ensembl: Extract gene-related data including gene names, chromosomal locations, and functions.

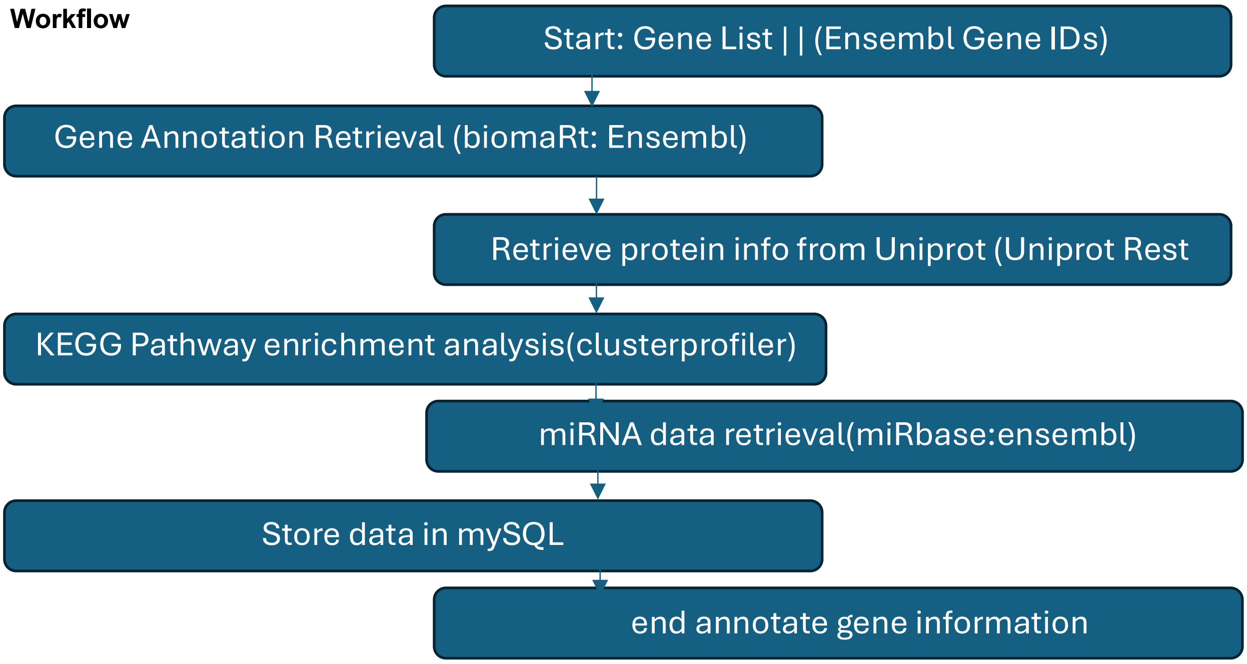
Integrate protein data from UniProt: Fetch protein sequences, functional annotations, and related protein-protein interaction data.

Integrate miRNA data from miRBase: Obtain miRNA annotations and their associated target genes.

Perform KEGG pathway enrichment analysis: Identify enriched biological pathways associated with the genes of interest.

Store and manage the data in a MySQL database: Organize all the data in a relational database for easy querying and analysis.

Workflow


Start: Gene List | | (Ensembl Gene IDs)
Gene Annotation Retrieval (biomaRt: Ensembl)
Retrieve protein info from Uniprot (Uniprot Rest
KEGG Pathway enrichment analysis(clusterprofiler)
miRNA data retrieval(miRbase:ensembl)
Store data in mySQL
end annotate gene information

Materials and Methods

1. Software and Tools

The following tools and R packages were utilized in the implementation of the pipeline:

RMySQL: For connecting to MySQL and storing the data.

biomaRt: For accessing Ensembl gene annotations.

UniProt REST API: For retrieving protein data.

miRBase REST API: For fetching miRNA data and their associated targets.

KEGGREST: For pathway analysis from KEGG.

httr, jsonlite: For handling API requests and parsing JSON data.

clusterProfiler: For performing pathway enrichment analysis.

2. Database Setup

The MySQL database was designed to store data from various sources. It includes the following tables:

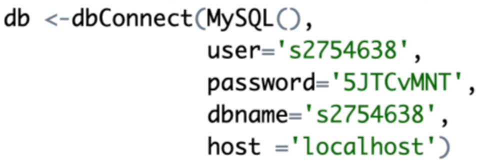
gene_annotations: Stores Ensembl gene data.

uniprot_data: Stores protein data retrieved from UniProt.

kegg_data: Stores KEGG pathway data.

mirna_data: Stores miRNA annotations and target genes from miRbase as missing_gene as there was only one micro RNA.

# Establishing connection to MySQL

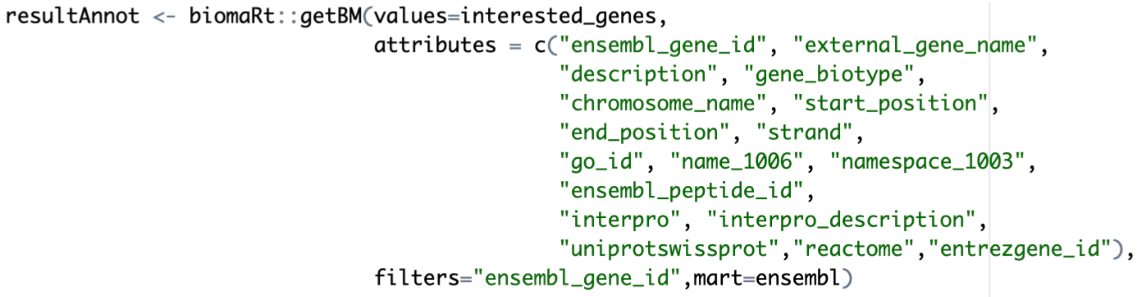


```
db <-dbConnect(MySQL(),
               user='s2754638',
               password='5JTCvMNT',
               dbname='s2754638',
               host ='localhost')
```

3. Data Retrieval

Gene Annotations from Ensembl: Gene annotations were retrieved using the biomaRt package, which accesses the Ensembl database to provide gene-related data such as names, chromosomal positions, and gene functions.

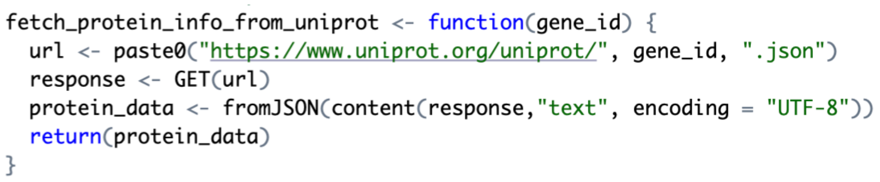
# Fetching gene annotations from Ensembl



```
resultAnnot <- biomaRt::getBM(values=interested_genes,
                              attributes = c("ensembl_gene_id", "external_gene_name",
                                             "description", "gene_biotype",
                                             "chromosome_name", "start_position",
                                             "end_position", "strand",
                                             "go_id", "name_1006", "namespace_1003",
                                             "ensembl_peptide_id",
                                             "interpro", "interpro_description",
                                             "uniprotswissprot","reactome","entrezgene_id"),
                              filters="ensembl_gene_id",mart=ensembl)
```

Protein Data from UniProt: Protein data was fetched from UniProt using their REST API. The httr package was used to send GET requests and process the returned data in JSON format.

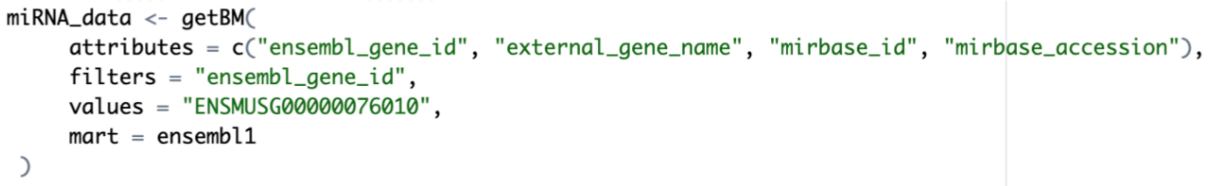
# Function to fetch protein data from UniProt



```
fetch_protein_info_from_uniprot <- function(gene_id) {
  url <- paste0("https://www.uniprot.org/uniprot/", gene_id, ".json")
  response <- GET(url)
  protein_data <- fromJSON(content(response,"text", encoding = "UTF-8"))
  return(protein_data)
}
```

miRNA Data from miRBase: miRNA data, including information on miRNAs and their target genes, was retrieved from miRBase using the miRBase REST API. miRBase provides curated data on miRNAs and their roles in gene regulation.

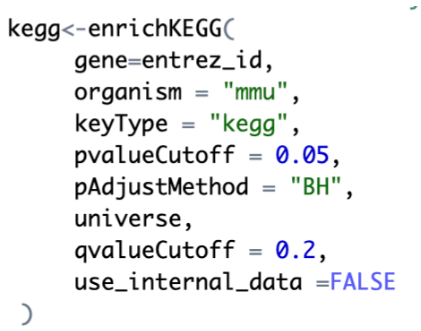
# Function to fetch miRNA data from miRBase



```
miRNA_data <- getBM(
     attributes = c("ensembl_gene_id", "external_gene_name", "mirbase_id", "mirbase_accession"),
     filters = "ensembl_gene_id",
     values = "ENSMUSG00000076010",
     mart = ensembl1
 )
```

KEGG Pathway Data: KEGG pathway analysis was performed using the clusterProfiler package. The tool was used to identify enriched biological pathways associated with the selected genes.

# Perform KEGG enrichment analysis



```
kegg<-enrichKEGG(
     gene=entrez_id,
     organism = "mmu",
     keyType = "kegg",
     pvalueCutoff = 0.05,
     pAdjustMethod = "BH",
     universe,
     qvalueCutoff = 0.2,
     use_internal_data =FALSE
 )
```

Results and Discussion

Gene Annotations from Ensembl

The retrieval of gene annotations from Ensembl provided essential information, including gene IDs, gene names, descriptions, and chromosomal locations. These annotations are fundamental for understanding gene functions and identifying the roles of specific genes in various biological processes. The data facilitated the classification of genes by their functions, which is critical for elucidating gene-disease associations and regulatory mechanisms.

UniProt Protein Data

The integration of protein data from UniProt significantly enriched the dataset with protein sequences, functional annotations, and protein-protein interaction networks. This information is pivotal for understanding the functional relationships between genes and their encoded proteins, enabling the identification of key molecular pathways and potential therapeutic targets. The detailed protein annotations offer a deeper insight into protein structure, function, and interactions, providing valuable information for further analysis in proteomics and drug discovery.

miRNA Data from miRbase

The integration of miRNA data from miRBase revealed important regulatory interactions between miRNAs and their target genes. miRNAs play a crucial role in post-transcriptional gene regulation, influencing gene expression at various stages. By linking miRNA annotations with their associated target genes, this dataset enhances our understanding of gene regulation, especially in the context of diseases such as cancer, where miRNAs are often implicated in the regulation of oncogenes and tumor suppressors. This integration helps identify miRNAs that may serve as biomarkers or therapeutic targets.

KEGG Pathway Analysis

KEGG pathway enrichment analysis was performed to identify biological pathways significantly associated with the genes of interest. This analysis highlighted key metabolic and signaling pathways involved in cellular processes such as metabolism, cell signaling, and disease mechanisms. KEGG pathway data provides a comprehensive view of the biological context in which the genes operate, helping researchers identify key pathways that are dysregulated in diseases and could be targeted for therapeutic intervention. This enrichment analysis is essential for understanding complex biological systems and guiding future experimental research.

Error Handling and Problem-Solving

Error handling is implemented to ensure smooth execution:

1. **API Request Validation:**

   The response from external APIs (e.g., UniProt) is validated to check if the request was successful. If a failure occurs, the system retries or logs the error.

2. **Database Insertion:**

   Before inserting data into MySQL, the script verifies whether the data already exists, ensuring no duplication and handling insertion errors.

3. **Missing Data Handling:**

   The script accommodates missing data by ensuring only available information is inserted into the database, preventing pipeline halts due to incomplete annotations.

Generalization for Different Gene Lists

The pipeline is adaptable to any gene list by simply changing the input list (interested_genes) or reading from external files:

1. Dynamic Gene List Input:

   The gene list can be passed from a file or an external source.

2. Flexible Species Support:

   Currently configured for *Mus musculus* (mouse), but can be adapted for other species (e.g., human) by modifying the dataset in the useEnsembl() function.

   ```
   genes_mart <- useEnsembl(biomart = "genes", host = ensembl_host)

   ensembl <- useDataset(dataset = "hsapiens_gene_ensembl", mart = genes_mart)
   ```

Challenges and Solutions

1. Handling API Rate Limits

To manage API rate limits imposed by UniProt and KEGG, the script included pauses between API calls, ensuring compliance with the usage guidelines. For large datasets, batch requests were implemented.

2. Data Gaps

Some genes did not have corresponding data in all the databases. These missing entries were logged, and the analysis proceeded with available data, ensuring that the project moved forward without interruptions.

3. MySQL Database Optimization

To ensure fast data retrieval and storage, the database was indexed on key fields like ensembl_gene_id, mirna_id, and uniprotswissprot. This optimized the performance when handling large datasets.

Conclusion

The successful integration of gene annotations from Ensembl, protein data from UniProt, miRNA data from miRBase, and KEGG pathway information into a MySQL database establishes a robust and comprehensive bioinformatics pipeline. This integrated system provides a powerful platform for the exploration of gene function, protein interactions, and regulatory networks, which are crucial for understanding complex biological processes and disease mechanisms. By combining diverse biological datasets, this pipeline facilitates detailed analysis and insights that can drive advancements in personalized medicine, functional genomics, and systems biology.

Despite facing challenges such as incomplete data entries and API rate limits, the approach demonstrates scalability and efficiency. The error handling mechanisms, database optimization, and adaptability to different gene lists and species further enhance the pipeline's usability. Ultimately, this system represents a critical tool for bioinformatics research, enabling seamless integration and analysis of large-scale biological data for a wide range of applications in genomic studies and therapeutic discovery.

Reference:

1. Biological Database- M.Sc Bioinformatics Lecture notes by Dr.Simon Tomlinson
2. Rstudio help files
3. https://www.uniprot.org/help/api_queries
4. ChatGPT for error corrections and grammar check
5. Jones, D. T., Sternberg, M. J., & Thornton, J. M. (2006). Introduction. Bioinformatics: from molecules to systems. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *361*(1467), 389–391. https://doi.org/10.1098/rstb.2005.1811