Databases & Data Integration Exercises

MISB, ISB101 Genomics & Databases course

Venkata P Satagopam (venkata.satagopam@uni.lu)
Schneider Group (Bioinformatics core) LCSB, Uni. Luxembourg

This tutorial is only meant to do exercises in the class room, it is NOT an assignment

Biological data is stored in many different databases throughout the world, e.g. see the list on the ExPASy website (http://www.expasy.org), the Nucleic Acids Research site (http://www.oxfordjournals.org/nar/database/cat/1) or Pathguide from Memorial Sloan-Kettering Cancer Center (http://www.pathguide.org) for interactions and pathway databases. Obviously, for these databases to be useful, one needs to know what kind of data present in these resources and able to query them efficiently. The purpose of this tutorial is to help you improve your knowledge about such databases and improve ability to retrieve the necessary information for your studies and research.

Part 1 Databases:

Exercise 1

Q1a: Search NCBI Gene (also called EntrezGene) database for gene 'NR1H2' and report how many gene entries in this database?

Q1b: Does EntrezGene contain only human specific genes or does it contains information about other organisms as well?

Q1c: How many human specific NR1H2 entries are in EntrezGene database?

Q1d: What are the other names for NR1H2 in human?

Q1e: Where this gene (NR1H2) located on human genome?

This exercise aims to demonstrate:

- How to search a gene of interest and retrieve information from EntrezGene database

Exercise 2

Q2a: What is the abbreviation of HUGO and HGNC? Why HGNC is important of us?

Q2b: Search HGNC for gene 'Brca1' and find out in which disease this gene playing an important role?

Q2c: Find out UniProt identifier for 'Brca1'

This exercise aims to demonstrate:

- To know what is HGNC and how to query this resource

Exercise 3

Now we will focus on protein database UniProtKB

Q3a: How many human specific proteins are deposited in SwissProt database and how it is differs from TrEMBL?

Q3b: Get human Brca1 protein sequences in FASTA format, does it has any splice variants?

Q3c: Get the FASTA sequence of protein ACES_HUMAN and find out 10 best similar sequences from UniProtKB/Swiss-Prot database?

This exercise aims to demonstrate:

- How to search uniprot database and retrieve protein sequence(s).
- How to perform blast search

Exercise 4

Q4a: Mention two important chemical databases?

Q4b: What is use of drug 'Omeprazole'?

Q4c: What is its mechanism of action?

This exercise aims to demonstrate:

- How to use drugbank information

Exercise 5

Q5: Accelerated aging syndrome 'Progeria' is caused by mutations in which gene?

This exercise aims to demonstrate:

- How to use knowledge in OMIM database

Exercises 6

Q6a: If you publish a paper, do you know where the scientific abstracts are deposited?

Q6b: How many paper are related to HIV?

Q6c: How many papers James D.Watson and Francis Crick have together? Get the title of one these papers.

This exercise aims to demonstrate:

- How to do simple and advance searches with PubMed

Exercises 7

Q7: Get the protein-protein interactions of p53 in human?

This exercise aims to demonstrate:

- How to use protein-protein interaction databases like STRING

Exercises 8

Q8: Name the list of the KEGG pathways where LAMNA gene is involved

This exercise aims to demonstrate:

- How to use KEGG pathway database

Part 2 Data Integration and Applications - I:

Exercises 9 - 10

Q9: Use the provided human specific Ensembl gene ids and query the Ensembl biomart and get the corresponding gene name, chromosomal location and associated Gene Ontolgoy (GO) terms.

Q10: Use the same gene and get the yeast orthologs.

This exercise aims to demonstrate:

- How to query the BioMart and get the required information

Part 3 Data Integration and Applications - II:

Exercises 11 - 12

Q11a: Use the same gene list and analyze by using bioCompendium

Q11b: Use the provided PDF file and find out the human specific genes mentioned in this document, and list the enriched pathways.

Q12: Do the comparative analysis with all the 3 gene lists and PDF document and report your findings

This exercise aims to demonstrate:

- How to use the bioCompendium to analyze a single gene list
- How one can extract the gene/proteins mentioned in a document with out typing them manually
- How to compare a gene list with other genes lists even they are represented by using different database identifiers and coming from different organisms, and also with a document (PDF) and further analyzing the interesting gene sub-sets