Querying data from biological data sources

Introduction

The aim of this first lab is to make you navigate in the maze of public biological data sources (available through the Web). We have selected the databases the most frequently used by our biologist collaborators (such sources are, more generally speaking, representative of the sources used in molecular biology).

We will more particularly focus on one given use case, namely, the search for information related to the *Long QT syndrome,* a heart disease studied by a team of pediatricians at CHOP, the *Childrens' Hospital of Philadelphia.*

Our goal is to find information on this disease while focusing on molecular biological data and more particularly on genes and proteins associated to *Long QT syndrome.*

The data sources considered are the following:

- **NCBI** sources: they can all be queried by using the **portal** available at: http://www.ncbi.nlm.nih.gov/gquery/

  Entrez NCBI is the name of the tool you can access to when clicking on the link above. It provides access to the most important data sources from North America (USA) stored in a large collection of databases. Among others, there are RefSeq/GenBank/EntrezGene for genomics data (nucleotides and genes), Entrez/Protein for proteins, OMIM for diseases and PubMed for Publications…

- **UniProt** : which gathers information proteins and associated genes: http://www.uniprot.org/

Part 1 - Using NCBI

Please go to the main page of Entrez NCBI: http://www.ncbi.nlm.nih.gov/gquery/
Please write *Long QT syndrome* as the keywords for your query.

1. Number of results
   a. How many databases are queried?
   b. How many entries describing genes have been retrieved?
   c. How many proteins?
   d. How many entries from OMIM?

2. Deeper look within the structure and content of entries

   - Click on « OMIM » (in the « Health » section) and inspect one of the results obtained, what do you think about the structure of each entry (is it easy to put into a relational DB)?

- Click on « Gene » (in the « Genes » section, click on the second line)
  - Filter the species to filter out only Human genes (select « Homo sapiens ») and click on « See also 2 discontinued or replaced items. » to get all the results.
  - What do you think about LQT4? What do you think about its position in the ranking of the results?
  - Look deeper at the results obtained. What do you think about the structure of an Entrez Gene entry (is it easy to put into a relational DB)?
- Go back to the main page of Entrez (where you have indicated *Long QT syndrome*) and now click on « Protein »
  - Look at the first results: what is the content of such entries? What do you think about the structure?
- Go back to the main page of Entrez and click on « Nucleotide » (section « Genomes »)
  - Look at the first results: what is the content of such entries? What do you think about the structure?

Part 2 – Using UniProt

Go to UniProt (http://www.uniprot.org/) and use the keywords *Long QT syndrome* for your query.

- How many proteins are obtained?
- Click on the first results, what is the kind of relationship between genes and protein(s): one-to-one or one-to-many? What is the difference with NCBI/Entrez Protein?

Comparing gene names.

Please consider Q12809 in UniProt and 3757 in EntrezGene

a. Are the gene names listed the same? Elaborate hypothesis on why it may be different for some genes.

b. In the section « Names and Taxonomy » click on the link to HGNC:6251. Inspect the file to look at the gene names. Follow the link to OMIM, what do you think? Return to the HGNC entry: 6251 and follow the link to GeneCards: What do you think of the structure and content of this entry from GeneCards?