AS.410.712: Advanced Practical Computer Concepts for Bioinformatics
Project Proposal
Rayna Hamilton
April. 15, 2022

Introduction

Clear cell renal cell carcinoma, or ccRCC, is the most common form of kidney cancer, accounting for 2% of both cancer diagnoses and deaths worldwide [1],[2]. With a 12% late-stage 5-year survival rate, it is the deadliest urological cancer in the US [3]. Informing future treatment modalities significantly depends on elucidating differences in expression profiles between control and affected individuals. For example, it has been recently determined by vom Roemeling et al. that neuronal pentraxin 2 (NPTX2) overexpression in ccRCC cells contributes significantly to cell migration and viability, thus qualifying as a potential therapeutic target [1]. Ideally, scientists would be able to browse the results of such expression studies in a readily accessible online format. This process would both reduce the amount of redundant analysis which needs to be repeated and enable access by non-bioinformatics experts. Such expression data browsers have been generated previously for colon [4] and liver [5] cancer, but such a format could not be found for renal cancer data. This project aims to create a web interface which enables querying ccRCC differential expression data by searching keywords such as gene name and Gene Ontology term.

Data acquisition and storage

A pre-normalized Affymetrix expression dataset originally generated by von Roemeling et al. [1] will be downloaded from the Curated Microarray Database [6]. Probe annotation information such as gene symbol, name and Gene Ontology (GO) terms will be determined in R using the hgu133plus2.db annotation package. After calculating fold changes and t-test statistics for each probe, information will be output into tsv files. These files will be parsed and loaded into a MySQL database using python's MySQL module.

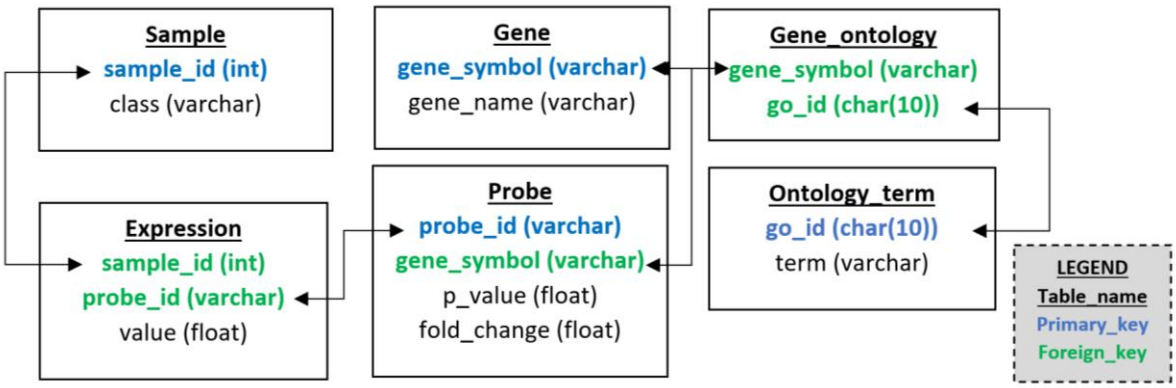

Figure 1. MySQL database schema for ccRCC microarray expression dataset. Sample class can be either "Normal" or "ccRCC". The tables Sample, Expression and Probe collectively contain all information found in the original dataset. Gene annotation information is split up into Gene_ontology, which links gene symbols to GO ids, and Ontology_term, which links GO ids to GO term descriptions. It is possible to combine the Gene and Gene_ontology tables with no loss of database functionality, however as gene_symbol: go_id is a many-to-many relationship this would require redundant duplication of the gene_name attribute. Gene_ontology and Ontology_term are similarly kept separate to avoid duplicating ontology term descriptions.

Web interface and methodology

On the main page, users will enter a search term and specify in a drop-down menu whether it refers to a gene name, gene symbol or GO term. Once this form element has been submitted, the search term will be sent to a python CGI script through Javascript AJAX [7]. This script will query the appropriate table of the MySQL database and return the resultant JSON data [8]. Javascript will then parse the JSON results and, using AJAX, display gene symbols and names of all results in a table below the search bar [9]. Each result row will also have a button which stores the associated row's gene symbol. When this button is clicked, a CGI script will query the database with the associated gene symbol and send results to an HTML template on a separate page [10]. This page will display fold changes, p-values and boxplots comparing affected to control individuals for each probe associated with the gene. The gene's GO ids will also be displayed as a set of buttons which, when clicked, will send the user to the search results page for this id. This functionality enables users to investigate genes related to a gene or ontology term of interest.

References

1. von Roemeling, C. A., Radisky, D. C., Marlow, L. A., Cooper, S. J., Grebe, S. K., Anastasiadis, P. Z., Tun, H. W., & Copland, J. A. (2014). Neuronal Pentraxin 2 Supports Clear Cell Renal Cell Carcinoma by Activating the AMPA-Selective Glutamate Receptor-4. *Cancer Research*, *74*(17), 4796–4810. https://doi.org/10.1158/0008-5472.can-14-0210
2. Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, *68*(6), 394–424. https://doi.org/10.3322/caac.21492
3. Padala, S. A., Barsouk, A., Thandra, K. C., et al. (2020). Epidemiology of Renal Cell Carcinoma. *World J Oncol*. *11*(3):79-87. https://doi.org/10.14740/wjon1279
4. Moreno, V. (2021). *Expression Browser | Colonomics*. Colonomics. Retrieved April 10, 2022, from https://www.colonomics.org/data-browser/expression-browser/
5. Kaur, H., Bhalla, S., Kaur, D., & Raghava, G. P. (2020). CancerLivER: a database of liver cancer gene expression resources and biomarkers. *Database*, *2020*. https://doi.org/10.1093/database/baaa012
6. Feltes, B. C., Chandelier, E. B., Grisci, B. I., & Dorn, M. (2019). CuMiDa: An Extensively Curated Microarray Database for Benchmarking and Testing of Machine Learning Approaches in Cancer Research. *Journal of Computational Biology*, *26*(4), 376–386. https://doi.org/10.1089/cmb.2018.0238
7. Orvis, J. (2022). *CGI and HTML Forms.* Personal Collection of J. Orvis, Johns Hopkins University, Baltimore MD.
8. Orvis, J. (2022). *Interfacing with Databases using Python.* Personal Collection of J. Orvis, Johns Hopkins University, Baltimore MD.
9. Orvis, J. (2022). *Javascript and jQuery.* Personal Collection of J. Orvis, Johns Hopkins University, Baltimore MD.
10. Orvis, J. (2022). *CGI and HTML Templates.* Personal Collection of J. Orvis, Johns Hopkins University, Baltimore MD.