– Advanced methods for signature analysis: signature meta scoring using annotated gene sets

In transcription, the DNA sequence of a gene is copied (i.e., “expressed”) to make an RNA molecule which can be directly functional or be the intermediate template for a protein that performs a function. Gene expression data contains the expression levels of all genes in the organism. Geneset analysis is crucial for interpreting the results of such complex data.

Apply until **16 April 2023** / Xplorers Camp on **27 April 2023**

Question to be solved

- How can we make better use of the rich collection of signatures available both externally and internally?
- How should signatures be tagged (annotated) to best reflect the biological theme they represent?

General background

RNASeq experiments measure the expression levels of thousands of genes at the same time. A RNASeq experiment is typically performed to investigate the transcriptomic state in a biological sample (tissue or cell line) by measuring the messenger RNA level of each gene. Such measurements potentially provide an understanding of the biological processes ongoing in the sample.

A signature or a geneset is a group of genes that represents a biological theme (signaling pathway, disease subtype or cell type). Most signatures are underutilized due to non-existing or poor annotations. Often, the names of the signatures do not provide an intuitive link to the biological theme they represent. This reduces their usefulness in geneset analysis, which is important for interpreting the results of gene expression data.

Broad Institute took the approach of generating a single “hallmark” signature for each biological theme, by combining and refining all signatures belonging to that biological theme [1]. One limitation of this approach is that the arduous process of generating all hallmark signatures has to be repeated with the addition of new signatures. Another is the need for manual curation in order to determine the biological theme that a signature belongs to. Moreover, the number of biological themes covered is limited to fifty. Finally, the hallmark signatures could be biased by the selection of the gene expression datasets against which their member genes were picked.

This project proposes a different approach which does away with the aforementioned limitations. The output of a gene set analysis method on an unlimited number of signatures (for example, enrichment scores from GSEA for all signatures in MSigDB) is fed to a subsequent step. This step, called the metascoring step, then gives a final score to each biological theme. To enable this approach, the signatures are to be annotated using natural language processing (NLP) methods and terms from standard ontologies such as cell and pathway ontologies on OBO Foundry and BioPortal.

Data types & technologies

- Data Types
  - Text (signature names, metadata, gene symbols)
- Technologies/platform
  - MongoDB
  - Python
  - High Performance Computing (Unix)
  - FastAPI

Supporting material or links

- Reference
  - The Molecular Signatures Database (MSigDB) hallmark gene set collection. Liberzon et al., Cell Syst. 2015 Dec 23; 1(6): 417–425.
- Work already done on our internal signature database and annotation pipeline (https://docs.google.com/presentation/d/1HkqR8YlyY8WIMK_56bu9-clP-I7z35xM/edit)

Needed skills

- Python programming experience and skills
- Natural language processing (NLP)

Mentor

Dr. Chiahuey Ooi
Principal Scientist, Predictive Modeling and Data Analytics, Pharmaceutical Sciences, Roche Pharma Research & Early Development, Roche Innovation Center Basel

Form of cooperation

Preferred scale: 6 months full-time in the year 2023
Possible format: Internship

How to present your idea

Preferred presentation medium: Powerpoint/Google slide – please show how you would approach the problem in 5-6 slides (10 minutes). Note: We expect the intern to improve a pipeline already initiated by a previous RAAN intern for annotating and managing the signatures (based on NLP, Python, MongoDB and FastAPI).

Specific skills I will check during the pitch session: Python knowledge, familiarity with NLP