COMP39/9900 Computer Science/IT Capstone Project
School of Computer Science and Engineering, UNSW

Project Number: P13

Project Title: Developing a bioinformatics pipeline for immunome analysis

Project Clients: Sara Ballouz

Project Specializations: Software development; Computer Science and Algorithms; Big data analytics and visualization; Bioinformatics/Biomedical.

Number of groups: 2

Background:

Identification of the antigens associated with antibodies is vital to understanding the immune response. However, high-throughput analysis of these antigens is still limited as it is expensive, and thus current software tools to analyse this data are lacking.

The aims of this project are to implement several algorithms to analyse data from the Serum Epitope Repertoire Analysis (SERA) platform that can be used as a standalone package or part of a bioinformatics pipeline.

Requirements and Scope:

The students will develop a software tool or package that will parse and process the data from the SERA platform. This will include quality control, at least two algorithms, visualisation reports and processed data for downstream analysis.

Multiple algorithms have been written to analyse this data but are not freely available. An in-house developed tool that can be shared/used will aid in the analysis of this data by non-bioinformatics researchers. Most of this analysis involves sequence processing and matching, and statistical tests to determine significance and likelihoods.

Data:

SERA data is the processed output from next-generation sequencing of the antigens that have been extracted from a sample. This data is in the form of short peptides/protein k-mers (k=5 and k=6), from a restricted alphabet of amino acids (~20 letters), along with a Log-Enrichment score (LE) of that peptide. Each sample will have on average 2.7million 5-mer and 8.7million 6-mer peptide sequences. The data is in the form of CSV files.

E.g.,

```
CRICM,112.718522

CRVCM,107.296108

TTRAE,107.136805

CIICM,96.581866

CTICM,94.422249
```

ACDSK,91.548217

Algorithms:

- PIWAS (Protein-based Immunome Wide Association Study). https://pubmed.ncbi.nlm.nih.gov/33986742/

- PIE (Protein wide Identification of Epitopes) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10620090/

PIWAs:

PIWAs is used to identify antigen and epitope signals against a reference proteome. As input for the PIWAs algorithm, samples are labelled by condition and a target proteome selected. For each sample and protein in the proteome, PIWAs scores are calculated by tiling the kmers onto the protein sequence, smoothing over a window of these kmers, normalizing to the background signal, and calculating the maximum value. Background samples might not be available and simulated data will need to be used.

PIE:

The PIE algorithm is used to locate regions within the protein sequence that had stronger outlier signals. In this case, for each position in the protein sequence, all case and control sample values are normalized using median absolute deviation based on the distribution of the control sample values. An outlier threshold is calculated and then an outlier statistic. Values exceeding set thresholds are considered significant, and adjacent regions merged to define the region of interest.

Output and data visualisation:

The output should be the proteins (and their sequences) detected from the PIWAs, and the regions of interest from the PIE analysis.

Visualisation will include distributions or log-enrichment scores, heatmaps of tiling, plots of smoothing and outlier scores.

Further/exact details will be discussed.

Depending on time, additional methods may be incorporated into the package.

These include and are not limited to:

- A shiny app
- Unsupervised analysis of the output (PCA, linear regression, hierarchical clustering)

Required Knowledge and skills:

Students should know how to use either python or R for statistical and visualisation.

Scripting languages (as simple as bash) and workflow tools such as Docker or Snakemake.

Basic biology is not necessary but useful (ie what are proteins, sequencing).

Expected outcomes/deliverables:

Expected outcome:

Final product should be a software package for either R or python (or both!) that implements the tools for the analysis of SERA data.

Deliverables:

- Source code (scripts/software) that can run as part of a next-generation sequencing pipeline
- Implementation of the PIWAs algorithm as a standalone and one that can run as part of the package
- Implementation of the PIE algorithm as a standalone and part of the package
- Visualisation plots of the outputs that can be used with the output from the software/tool
- Documentation on how to use the scripts and examples
- Any additional functionality that are within the scope of the project that will help in its use.

Supervision:

Sara Ballouz

Additional resources:

Additional references and resources
https://pubmed.ncbi.nlm.nih.gov/34811480/
https://serimmune.com/
https://www.uniprot.org/
https://blast.ncbi.nlm.nih.gov/Blast.cgi