Product Documentation

BINF6112 - Computational Biology Engineering Design Workshop

Group

Epitope Peptide Microarray Analysis Platform (Lee/Gaeta/Shwe)

Team Members

Leonie Dickson	z5215454
Aravind Venkateswaran	z5208102
David Nguyen	z5166106
Alisa Zhou	z5210644

Table of Contents

Product Introduction

Peptide microarrays are commonly used to analyse antibody binding to peptides derived from allergens. This involves printing of overlapping amino acid 15-mers on a glass slide which is then hybridised with sera from allergic patients to identify which peptides bind antibodies in the sample.

While this allows high-throughput measurement of how antibodies bind to isolated peptides, it ignores the fact that in the native allergen, some peptides may not be accessible due to the 3D folding of the protein. However incorporating information about the allergen protein's 3D structure in the microarray analysis currently requires a lot of manual work. The Epitope Microarray Analysis Platform, developed by Team Lee (UNSW BINF6112 2020 T3) aims to address this issue to aid microarray data analysis. The platform allows for easy visualisation and analysis of peptides and their antibody binding together with accessibility information and chemical properties from the allergen's 3D structure.

The platform also uses results from microarray experiments to get relevant epitopes from individual peptide data. Summary tables for peptide and epitope data are displayed, alongside visualisations in the form of both a 3D structural view and a linear representation of the binding data.

These features provide scientists a more comprehensive understanding of antibody-allergen binding from microarray experiments.

User Documentation

Usage

1. Access the platform on an OS supported computer via an internet browser*.
    a. User can access the platform through https://microarray-analysis.herokuapp.com/
    b. Users can otherwise generate a local version of the platform. Please follow installation instructions.

*Browser should be compatible with WebGL (see below) - eg. Chrome.

2. Two options to upload samples:
    a. The ‘Single Sample’ tab allows the user to upload and analyse single sample data (up to 2 microarray data files permitted, eg. one IgE and one IgG4 file)
    b. The ‘Multiple Sample Analysis’ tab allows the user to upload and analyse multiple sample files.
3. Upload microarray data files in either excel or gpr format by clicking the browse button.
4. Upload a .pdb file representing the structure of the allergen of interest by clicking the browse button. Any .pdb file can be used (eg. PDB files downloaded from the Protein Data Bank or from SWISS-MODEL are both compatible).
5. Excel/gpr files can be previewed by clicking the auto generated buttons for each file.
6. The PDB structure can be previewed before submitting in the image block below the form.
7. Click the Submit button once after uploading required files for analysis.
8. Three ways to visualize the analysed data:
    a. Linear Representation Plot:
        i. Click the buttons on top to change the binding data displayed on the graph (eg. “Foreground Median” or “Calculated SNR”).
        ii. File names displayed on top of the graph are colour coded corresponding to each line in the graph. These names also act as a toggle to hide/show their specific data in the graph.
        iii. Chemical properties of the peptide sequence can be viewed in a tooltip by hovering over the point on the graph. The chart’s x-axis shows the residue ID and first three residues of the peptide whose binding data is displayed at that point. (A dash indicates a gap in the mapping or structure, where no data was mapped to that position).
        iv. Use the sliders below the sequence to adjust the window size. This option is useful to visualize a large number of peptides in the same graph window.
        v. Click the top right logo (eye) to view the graph in full screen and press esc to exit full screen.
    b. 3D Structural View:
        i. A 3D structural view of the supplied PDB file is displayed. Mouse events such as zooming, rotating and hovering on a residue are supported.

ii. On clicking a row in a summary table, the peptide's starting residue (ID column in the table) will be highlighted in red on the 3D protein model. Deselect a row to unhighlight a selection.
iii. Clicking a point in the linear representation will highlight the associated residue in the 3D protein model in red. Click the same point again to deselect and unhighlight the residue.

c. Epitope and Peptide Summary Tables:
   i. Hide or show the column of table fields by clicking the corresponding yellow buttons above each table.
   ii. Click column headers to sort the table by that column.
   iii. Download a .csv file of the table by clicking the Export button below the desired table.
   iv. Get a description of the table fields by scrolling to the bottom of the page.

Information about Data Processing

Peptides in the microarray data are mapped to the allergen sequence generated from the DSSP output. If certain peptides belonging in the microarray data do not show up on the platform, this may be because of a mismatch (100% match to the peptide sequence is required for the peptide to be mapped) or because the peptide sequence falls outside the range of the allergen sequence.

In order to obtain peptide chemical properties using the Biopython SeqUtils.ProtParam module, ambiguous amino acid codes are replaced (B -> N, Z -> Q, J -> L).

The DSSP output is used to obtain values for:

- Relative accessible surface area: the solvent accessible surface area (calculated per residue by DSSP) relative to the maximum accessible surface area for the residue (as determined by Sander and Rost (1994)). Relative ASA values for peptides are calculated as the sum of the residues' ASA values, divided by the sum of the residues' maximum ASA values - values lie between 0 and 1.
- Secondary Structure: Mode of residue secondary structure assignments generated by DSSP.

Binding data is smoothed for overlapping peptides, calculated as: smoothed value for p_2 = 0.125 * p_1 + 0.75 * p_2 + 0.125 * p_3, for overlapping peptides p_1 and p_3 offset from p_2 by 3 residues. Weightings are configurable.

A peptide is predicted to be the center of an epitope if its binding data forms a local peak relative to its two neighbours, and this peak is above a given threshold. Neighbouring peptides are included in the epitope if their binding data is greater than a given percentage threshold of the peak. Thresholds are configurable.

Technical Documentation

Product Build Prerequisites

- Node 12.*
- Npm (Node package manager)
- Python 3+
- DSSP
- Biopython

Installation From Source

Prerequisite Installation

DSSP

1. Clone repo https://github.com/cmbi/dssp/tree/2.3.0
2. Follow instructions in README to satisfy DSSP prerequisites
3. In the repo directory run
    - ./autogen.sh
    - ./configure
    - make
    - sudo make install

Biopython

Installed as per https://biopython.org/wiki/Download

WebGL

The NGL Viewer (Protein Model) requires your browser to support WebGL. To see if your browser supports WebGL and what you might need to do to activate it, visit the Get WebGL page.

Generally, WebGL is available in recent browser versions of Mozilla Firefox (>29) or Google Chrome (>27). The Internet Explorer supports WebGL only since version 11. The Safari Browser since version 8 (though WebGL can be activated in earlier versions: first enable the Develop menu in Safari's Advanced preferences, then secondly in the now visible Develop menu enable WebGL).

Platform Installation

Users can load a local version of the platform with the following steps:

1) Clone the repo from: https://github.com/davenyen/BINF6112_PROJECT
2) In the project's root directory run:
    a) npm start
3) Open server on the port displayed in preferred internet browser (default: localhost:5000)

Platform Configuration

A configuration file is provided at: /client/src/Config.json

Parsing and Column Name Configuration

Within the config file, a JSON object configuring column names is provided for each of "excel" or "gpr" file types.

Regex to identify columns containing the peptide sequence (necessary for application to run) and peptide names (eg. "Arah1_overlapping3) (not necessary to run) can be configured in the corresponding fields for each file type.

To add a new binding data type to the platform:

1) Add a regex to the corresponding file type's "column-regex" list to allow the application to identify the correct column in the uploaded excel/gpr files.

2) Add your desired data type label (eg. “Foreground Median”) to the corresponding file type’s “column-display-names” list. This is how the data type will be labelled across the platform (in tables and in the linear representation).

   The order of columns in “column-display-names” should correspond to the order in “column-regex” list. Please ensure the “column-regex” and “column-display-names” lists have the same number of elements.

The platform can be configured to display or not display the calculated SNR value. To do this, change the “calculateSNR” field in Config.json to true or false. To calculate SNR, the platform uses the raw mean and background mean fields of an excel or gpr file - to ensure that the platform identifies these (if microarray data files from another manufacturer are being used), adjust the “rawMean” and “backgroundMean” fields in either file type’s “calculateSNR” field with the appropriate regex.

Parsing .gpr files also requires a set number of rows (containing manufacturer information etc. before the data column headers begin) to be skipped. This is currently set to 11, and can be configured in the gpr.rowsToSkip field.

Smooth Binding Data Configuration

In Config.json, the “overlap” field contains the parameters used to smooth the binding data. The “amount” field is the number of residues each overlapping peptide is offset by. The “weightPrev”, “weightCurr”, and “weightNext” fields determine the weighting of the previous, current and next overlapping peptide respectively.

Epitope Prediction Configuration

The “epitopes” field controls the epitope prediction methods. The “dataTypeColumnIndex” is the index in the “column-regex” list which corresponds to the binding data type used to determine epitopes (eg. if “Foreground Median” is the first item in these lists, to base predictions off the foreground median, dataTypeColumnIndex is set to 0).

The “threshold” field is the minimum of the configured binding data type required to consider a peak to be an epitope. “relativeIncludeThreshold” is the percentage of the peak’s binding data that neighbouring peptides are required to have to be included in the epitope.

Other Configurable Parameters

Config.json also has fields to configure:

- “bury_threshold”: Peptides with a relative accessible surface area below this threshold are considered buried, with cells in tables and corresponding x-axis labels in the linear representation colored grey.
- “protein_structure_coloring”: Change coloring of the allergen molecule in the 3D structural view. “main” sets the base color of the whole molecule, while “selected” sets the color of residues that have been selected by clicking points on the linear representation or table.
- “decimal_places”: Decimal places of numerical data on platform.