OXFORD

Sequence analysis

ImmunomeBrowser: a tool to aggregate and visualize complex and heterogeneous epitopes in reference proteins

Sandeep Kumar Dhanda[1],*, Randi Vita[1], Brendan Ha[1], Alba Grifoni[1], Bjoern Peters[1,2] and Alessandro Sette[1,2]

[1]Division of Vaccine Discovery, La Jolla Institute for Allergy and Immunology, La Jolla, CA 92037, USA and
[2]Department of Medicine, University of California, San Diego, CA 92093, USA

*To whom correspondence should be addressed.
Associate Editor: John Hancock

Abstract

Motivation: Datasets that are derived from different studies (e.g. MHC ligand elution, MHC binding, B/T cell epitope screening etc.) often vary in terms of experimental approaches, sizes of peptides tested, including partial and or nested overlapping peptides and in the number of donors tested.

Results: We present a customized application of the Immune Epitope Database's ImmunomeBrowser tool, which can be used to effectively aggregate and visualize heterogeneous immunological data. User provided peptide sets and associated response data is mapped to a user-provided protein reference sequence. The output consists of tables and figures representing the aggregated data represented by a Response Frequency score and associated estimated confidence interval. This allows the user to visualizing regions associated with dominant responses and their boundaries. The results are presented both as a user interactive javascript based web interface and a tabular format in a selected reference sequence.

Availability and implementation: The 'ImmunomeBrowser' has been a longstanding feature of the IEDB (http://www.iedb.org). The present application extends the use of this tool to work with user-provided datasets, rather than the output of IEDB queries. This new server version of the ImmunomeBrowser is freely accessible at http://tools.iedb.org/immunomebrowser/.

Contact: sdhanda@lji.org

1 Introduction

Datasets that are derived from different studies (e.g. MHC ligand elution, MHC binding, B/T cell epitope screening etc.; Li Pira *et al.*, 2010) will often vary in terms of the experimental approaches used, the size of peptides that were tested and in the number of donors tested.

To enable simplified visualization of the aggregated data present in the Immune Epitope Database (IEDB), we implemented the application called the 'ImmunomeBrowser' in the IEDB (Vita *et al.*, 2015). This tool aggregates all data relevant to the user query and allows one to visualize the known immune response to a specific antigen, as well as illustrating knowledge gaps in a reference protein.

It provides the immune reactivity in terms of response frequency (RF) and the number of subjects tested/responded and/or number of independent assays performed along the length of reference protein. The tool was originally implemented in the results page of the database section of the IEDB. To further extend the usability to predicted epitopes and propriety epitopes or non-IEDB data, the online tool described herein was developed.

The utility of the approach was demonstrated by Kim *et al.* who performed a meta-analysis of Hepatitis C virus (HCV) data available in the IEDB, to present a bigger picture of the immune reactivity and knowledge gaps in the reference protein sequences of the virus (Kim *et al.*, 2012).

Currently, the Immunomebrowser can only be used with data derived from IEDB queries, and not with user datasets. To overcome this problem, we implemented the ImmunomeBrowser as a stand-alone tool to allow users to analyze and visualize immunodominant regions within their own dataset.

2 Materials and methods

2.1 Data input

The user provides peptide sequences, the response data for each, the protein sequence/s of interest and their desired sequence identity threshold in specified formats. The peptide response can be either pasted or uploaded as a file in whitespace separated format with three columns, corresponding to peptide sequence, number of subjects tested and/or number of assays performed, and the number of subjects responded and/or number of assays resulting in a positive response. In cases, where the number of subjects tested or responded or assays performed are not provided, the program will automatically fill in a value of '1' for the number of subjects tested or assays performed, as well as for the number of subjects responding or positive assays. Protein sequences/s must be provided in 'Fasta' format and sequence identity is selected from a drop-down menu that varies from 10–100% with an interval of 10%.

2.2 Mapping of epitopes

Each peptide is mapped to a user provided reference protein sequence according to the provided identity threshold. The degree of identity is calculated based on the alignment of the peptide within reference sequence. Only peptides with sequence identity above the threshold are selected for further calculations.

2.3 RF and confidence interval calculations

The RF for a given peptide and for each source protein position is calculated as the total number of subjects that responded to that particular peptide and/or independent assay performed for which a positive response was noted (R) divided the total number of subjects tested and/or number of assays performed (N).

$$RF = \frac{R}{N} \qquad (1)$$

A Confidence interval (CI) is calculated to weight the RF reliability as a function of the number of subjects tested. CI is calculated using the binomial cumulative distribution function and Wilson score.

For large sample size ($N>=50$), lower and upper bound were calculated using following equation.

$$CI = (((R/N) + 1.96*1.96/(2*N) \pm 1.96*sqrt(((R/N)*(1-(R/N)) + 1.96*1.96/(4*N))/N))/(1+1.96*1.96/N)) \qquad (2)$$

For small small sample sizes ($N < 50$), lower and upper bounds are calculated using binomial cumulative distribution function.

2.4 Aggregation of RF data from different overlapping peptides

Aggregation of data is required to identify the most frequently recognized epitopes, which can reflect the overall frequency of recognition of peptide sequence containing a given residue. This approach is useful to identify the RF at each position in the reference sequence. To calculate the aggregated RF data, the number of subjects tested

and/or assay performed and number of subjects responded and/or number of assays resulted a positive response were summed up for each mapped position in a given source protein. The CI of RF is calculated using the equations described above.

2.5 Result display

The results are presented in two steps, where the first step provides the summary of epitopes and assays mapped back to the reference protein sequence (Fig. 1).

For each protein, a table lists all the epitopes, its mapped position, the number of subjects responded/positive assays, the number of subjects tested/assays performed and the RF along with its upper and lower bounds at 95% CI (Fig. 1A). The second step provides the aggregate plot of the mapped RF for each region of the reference protein, in two different plots representing the cumulative RF (upper and lower bound of RF) and total number of results (positive and negative) along the length of the selected reference protein (Fig. 1B).

3 Applications

The customized application of the ImmunomeBrowser lends itself to several applications. As mentioned above, Kim et al. has performed a meta-analysis of HCV data available in the IEDB, (Kim et al., 2012; Vita et al., 2015). The tool can now be utilized by users to collate and perform meta-analysis of data generated in multiple related studies. For example, the ImmunomeBrowser can be applied to natural ligand elution data containing largely overlapping peptides, and which are studied in different donors expressing different HLA molecules (Schellens et al., 2015; Shastri et al., 2002). For this purpose, the data needs to be combined for response frequencies from different donors and for each HLA molecule (Alvarez et al., 2018). In this context, Vaughan et al. analyzed naturally processed data curated within the IEDB to characterize the overall general features of the known processed data and to highlight existing knowledge gaps (Vaughan et al., 2017). The Immunomebrowser is also useful to analyze the immunogenicity testing of therapeutic proteins, where the overlapping peptides from a therapeutic protein are tested for immunogenicity to evaluate the unwanted immune response (Asgari et al., 2015; Dhanda et al., 2018; Jawa et al., 2013; Salvat et al., 2017). The Immunomebrowser, can thus aggregate the immune response data from different peptides and/or peptide analogs spanning through the length of the specified reference protein, even when tested in different donors and derived from different clinical studies. This allows users to easily view their data in a more meaningful and useful manner.
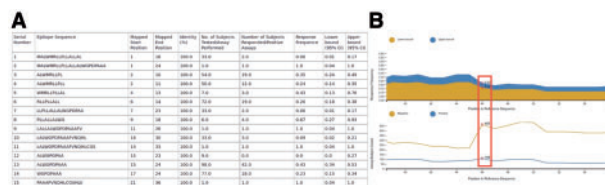

Fig. 1. Screenshots for the example output of the customized application of 'ImmunomeBrowser'. (**A**) Tabular format listing all the different epitopes mapped to the given reference protein sequence. (**B**) Area plot for upper and lower bound CI for RF. The line plot shows the number of positive and negative assays or number of responder and not-responder subjects along the positions in reference protein. Hovering the mouse over any position in the reference protein in any of these plots will display the lower and upper bounds of the RF and number of assays/subjects count found as positive and negative (as shown in red rectangle)

Acknowledgement

Funding

References

Alvarez,B. *et al*. (2018) Computational tools for the identification and interpretation of sequence motifs in immunopeptidomes. *Proteomics*. doi: 10.1002/pmic.201700252.

Asgari,S. *et al*. (2015) Rational design of stable and functional hirudin III mutants with lower antigenicity. *Biologicals*, **43**, 479–491.

Dhanda,S.K. *et al*. (2018) Development of a strategy and computational application to select candidate protein analogues with reduced HLA binding and immunogenicity. *Immunology*, **153**, 118–132.

Jawa,V. *et al*. (2013) T-cell dependent immunogenicity of protein therapeutics: preclinical assessment and mitigation. *Clin. Immunol*., **149**, 534–555.

Kim,Y. *et al*. (2012) A meta-analysis of the existing knowledge of immunoreactivity against hepatitis C virus (HCV). *PLoS One*, **7**, e38028.

Li Pira,G. *et al*. (2010) High throughput T epitope mapping and vaccine development. *J. Biomed. Biotechnol*., **2010**, 325720.

Salvat,R.S. *et al*. (2017) Computationally optimized deimmunization libraries yield highly mutated enzymes with low immunogenicity and enhanced activity. *Proc. Natl. Acad. Sci. USA*, **114**, E5085–E5093.

Schellens,I.M. *et al*. (2015) Comprehensive analysis of the naturally processed peptide repertoire: differences between HLA-A and B in the immunopeptidome. *PLoS One*, **10**, e0136417.

Shastri,N. *et al*. (2002) Producing nature's gene-chips: the generation of peptides for display by MHC class I molecules. *Annu. Rev. Immunol*., **20**, 463–493.

Vaughan,K. *et al*. (2017) Deciphering the MHC-associated peptidome: a review of naturally processed ligand data. *Expert. Rev. Proteomics*, **14**, 729–736.

Vita,R. *et al*. (2015) The immune epitope database (IEDB) 3.0. *Nucleic Acids Res*., **43**, D405–D412.