# DiscovEpi: Automatic whole proteome epitope prediction and visualization of epitope binding maps.

## Abstract

Antigenic peptide presentation by surface bound MHC class I molecules is a major key event to integrate CD8+ T cells in the adaptive immune response. Prediction of the best putative binding sites is of great importance in the development of immune-diagnostics and vaccine components. Especially, in the case of vaccines against pathogens the holistic overview of the immunogenicity of a pathogen’s proteome is of great importance. Here, DiscovEpi displays its great advantage in automatically predicting epitope binding sites of all proteins fetched from a proteome. From this point on, any in vitro approach could focus on the putatively most immunogenic proteins.

The newly developed DiscovEpi connects the protein database UniProt and epitope predictor NetMHCpan in a way that one can easily get immunogenicity predictions for whole proteomes. The proteins within the proteome are ranked by their density of predicted epitope binding sites including the average epitope binding score. In addition, the whole epitope map of every protein is visually represented for the interactive usage. An epitope map enables finding epitope-rich regions in a protein and thus paves the way for drug design approaches.

The usage of DiscovEpi is represented at the example of Staphylococcus aureus. Epitope prediction is performed with cytoplasmatic proteins of strain USA300 for HLA-A\*02:01 and nine amino acids long binding sites. Overall, it took only 7 minutes to perform epitope prediction for the cytosolic subproteome of 222 proteins and process the results to the final heatmap. Running completely in the background the user is not required to do any sequence handling. Additionally, the results are displayed protein-wise which is of advantage ranking proteins by their immunogenicity compared to NetMHCpan’s current epitope-wise table which prevents assigning an epitope to its protein.

## Introduction

Epitope-based vaccines focus on stimulating the immune system by presenting these epitopes to it, without the need for using the entire pathogen. This targeted approach enhances the safety profile and the efficiency in design and manufacturing of the vaccine. Additionally, the vaccine may have a broader coverage when targeting shared conserved epitopes and reduce the risk of immune escape avoiding targeting non-neutralizing epitopes. Epitopes are small peptides derived from proteins of a pathogen, such as a virus or bacterium, that are recognized by the immune system. They are usually short amino acid sequences that can trigger an immune response, leading to the production of antibodies or the activation of T-cells (Parvizpour et al. 2020).

A hallmark of the adaptive immune response is the interaction between cell surface bound MHC class I peptide complexes and the T cell receptor of CD8+ T cells. There is evidence that intracellular infection by bacteria such as Staphylococcus aureus is accompanied by presentation of specific epitopes at the cell surface (Bröker et al. 2016). Building the complex, selection of the pathogen derived peptide is a key event (Neefjes et al. 2011). In addition to experimental epitope mapping there are *in silico* approaches to predict epitopes from a given amino acid sequence for specific MHC molecules.

Vaccine development against rapidly evolving pathogens or those with a great number of proteins requires fast proteome-wide approaches. Currently available algorithms like NetMHCpan provide epitope prediction for a list of proteins but do not assign the epitopes found to the protein derived from. Thus, comparison of proteins based on immunogenicity or identification of epitope-rich regions within proteins is hardly possible. These approaches could reveal proteins or parts of as candidates for vaccine design without much effort. There are programmatically accessible databases and epitope prediction algorithms providing reliable proteome data like the UniProt Knowledgebase and NetMHCpan with state of the art prediction benchmarks

In a first step DiscovEpi connects the UniProt protein database with NetMHCpan. One can retrieve protein data for whole proteomes, say all annotated proteins of Staphylococcus aureus and automatically get epitope predictions with rank factors epitope density and average binding score. Finally, DiscovEpi produces a visual of the protein set, where every epitope is highlighted at its respective position.

## Methods

DiscovEpi algorithm is based on Python and the Qt framework using PySide for integration. Protein data from UniProt and prediction data from NetMHCpan are retrieved via RESTful API's from UniProt and IEDB, respectively. Seaborn visualization library is used to produce an epitope map in form of a heatmap. The GUI is created with the Python bindings of Qt and an executable for Microsoft Windows including all packages and dependencies is created with PyInstaller.

The binding score provided by NetMHCpan is converted to a value between 0 and 1 based on the given binding score threshold so that the best binding epitopes have values close to 1 and weak binding ones close to 0.

To rank the proteins there are two values calculated. On the one hand, there is the epitope density describing the ratio of predicted epitopes to the number of possible meres of the respective length. The downside of the density is leaving the binding score of each epitope aside, which gives weight to strong or weak binding epitopes. On the other hand, there is the average score, which cannot differentiate between a few well binding epitopes or many weak binding.

## Results

DiscovEpi is free of charge and enables automatic proteome-wide epitope prediction. It requires least input and manual handling by the user. One has only to keep in mind the UniProt annotation for organism and cellular location to specify the proteome. The IEDB offers different prediction algorithms while to this date only the recommended one is available in DiscovEpi to guarantee state of the art results. Over the last years NetMHCpan has outperformed its competitors and is declared as recommended.

Protein and prediction data are independently processed and put out as XLSX files before passed to produce the visual epitope map created as PNG file.

Comparison of epitope densities at the level of species

The results can be used to compare pathogens according to their immunogenicity overall. Taking the average epitope density of all proteins of a pathogens proteome gives a hint of the immune system's ability to clear out a pathogen. In this case, strains of Staphylococcus aureus have been compared to strains of Influenza A and Coronaviruses. The results indicate that as a mechanism of immune evasion due to long ongoing coexistence Staphylococcus aureus and Influenza A were able to reduce their epitope density compared to Coronaviruses.

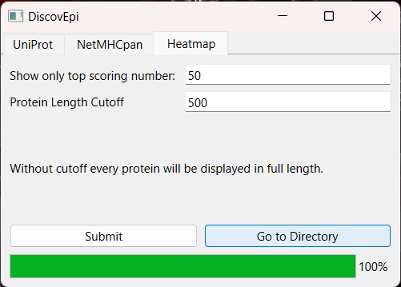
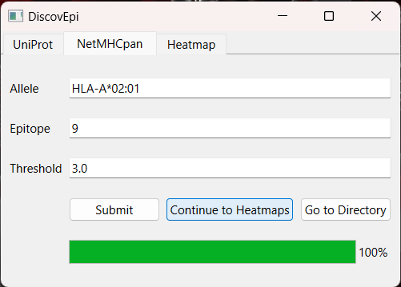
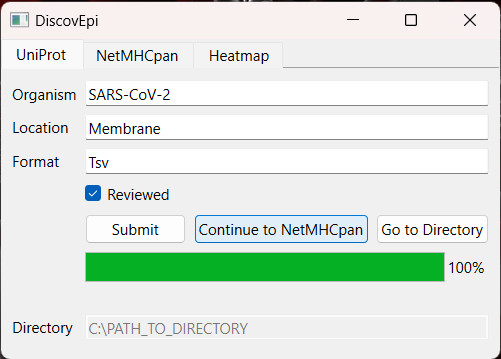
Comparison of epitope densities at the level of proteins

Epitope maps allow a closer look at the immunogenicity of proteins. Ordered by their epitope density the heatmap displays every protein of a proteome with highlighted positions where epitopes are located. If there are hundreds of proteins one can cut the list at a given number to get a clear look. Using the visuals, it is possible to identify hotspots of epitopes and thus to identify proteins containing epitope-rich areas. These areas could be of interest when it comes to antigen processing, MHC loading and presentation for T cells. Training of T cells with new identified epitope-rich short sequences of antigens from a pathogen could lead to new targets in e. x. vaccine design.

Ein Bild, das Text, Screenshot, Grafikdesign, Grafiken enthält.

Automatisch generierte BeschreibungEin Bild, das Text, Screenshot, Diagramm, Schrift enthält.

Automatisch generierte Beschreibung



Literaturverzeichnis

Bröker, Barbara M.; Mrochen, Daniel; Péton, Vincent (2016): The T Cell Response to Staphylococcus aureus. In: *Pathogens (Basel, Switzerland)* 5 (1). DOI: 10.3390/pathogens5010031.

Neefjes, Jacques; Jongsma, Marlieke L. M.; Paul, Petra; Bakke, Oddmund (2011): Towards a systems understanding of MHC class I and MHC class II antigen presentation. In: *Nature reviews. Immunology* 11 (12), S. 823–836. DOI: 10.1038/nri3084.

Parvizpour, Sepideh; Pourseif, Mohammad M.; Razmara, Jafar; Rafi, Mohammad A.; Omidi, Yadollah (2020): Epitope-based vaccine design: a comprehensive overview of bioinformatics approaches. In: *Drug discovery today* 25 (6), S. 1034–1042. DOI: 10.1016/j.drudis.2020.03.006.