# 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 (Paul et al. 2020).

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

Databases  
Protein data are retrieved directly from the UniProt Knowledgebase (UniProt: the Universal Protein Knowledgebase in 2023 2023) by using the RESTful API with a search query for organism and subcellular location given by UniProt. (Reynisson et al. 2020)

For the putative epitopes within the protein sequences DiscovEpi is using the extracted information from NetMHCpan (Reynisson et al. 2020). For each from UniProt fetched protein sequences epitope predictions are performed using the RESTful API provided by IEDB (Vita et al. 2019).

DiscovEpi Algorithm  
DiscovEpi algorithm (https://github.com/cmahncke/DiscovEpi) is based on Python (Python Software Foundation. Python Language Reference, version 3.10. Available at http://www.python.org) and the Qt framework using PySide for integration (The Qt Company, Qt for Python project. PySide, version 1.2.4. Available at https://pypi.org/project/PySide). Seaborn visualization library is used to produce an epitope map in form of a heatmap.

The score of predicted strength of binding provided by NetMHCpan (Reynisson et al. 2020) is based on a trained neural network calculating the strength of binding considering amino acid properties, peptide-MHC interactions and experimentally measured binding affinities. This score is compared to a large, maintained reference distribution of scores then providing a percentile rank to assess the relative strength of peptide-MHC binding predictions. For example, a percentile rank of 10.0 means that the binding score is better than 90% of the scores in the reference set.

DiscovEpi uses the percentile rank for further computation and sets a threshold keeping only at least weak binding peptides. By default, DiscovEpi discards all peptides with a percentile rank greater than 3 because they seem to be practically irrelevant but the value is also free to set by the user (Reynisson et al. 2020). To compare epitopes in the limited set and allow scoring each protein, the DiscovEpi epitope score is computed by normalizing the percentile rank between 0 and 1 given the threshold so that values close to one represent better binding probabilities:

The ranking of the proteins is based on their DiscovEpi protein score , a value between 0 and 1 to assess the immunogenicity of each protein. It is calculated as the average epitope score of all possible epitopes:

where is the amino acid sequence length and the epitope length. The protein score does not differentiate between few well binding or many weak binding epitopes. Hence, DiscovEpi also provides the ratio of predicted epitopes to the number of possible meres of the respective length, the epitope density , to easily assess the distribution of the epitope scores:

## Results

DiscovEpi enables the automatic proteome-wide epitope prediction and extends the approach of single proteins to the whole proteome or limited parts of it. The usage is based on a GUI (Fig1a) allowing the user to add the taxonomic ID or the name of the organism of choice from the UniProt (UniProt: the Universal Protein Knowledgebase in 2023, 2023, https://www.uniprot.org/taxonomy).

Further, the user has the option to define a specific cellular location or use the entire proteome by adding the ID or name and to choose about including unreviewed database entries or not (UniProt: the Universal Protein Knowledgebase in 2023, 2023, https://www.uniprot.org/locations). When submitting, the user is asked to delete redundant sequences. This is because if the organism value is not set for one unique strain, the same proteins are retrieved multiple times for multiple strains. Data about the set of proteins, including UniProt ID, protein and gene name, length of the signal and whole sequence, the subcellular location, and the amino acid sequence, is automatically downloaded, processed, and saved to a location specified in the GUI.

Protein sequences are then used in NetMHCpan (Reynisson et al. 2020) to predict all putative epitopes scoring better than the set threshold. In this step the user is required to specify the HLA allele to be predicted for, the length of the epitopes and the threshold mentioned above. Equally to the protein data, the prediction results, including UniProt ID, protein name, DiscovEpi protein score, epitope density, density inside the signal sequence and the epitope data, is saved to the earlier specified location.

Finally, protein and epitope data are used to produce a visual epitope map. As input, the last step uses cutoff values for firstly only showing a specific number of the top scoring proteins and secondly only showing a specified length of the proteins. That is because showing a 2000 amino acid long protein decreases the resolution of epitope rich areas.

### Application of DiscovEpi

Interested in the definitions of the overall immunogenicities of *Staphylococcus aureus*, *Influenza A* and *SARS-CoV-2* we used DiscovEpi. *SARS-CoV-2* is used here as an example for applying DiscovEpi. The first step is to retrieve the proteome data via submitting the variables as shown in figure 1A. Next, the prediction parameters have been set and submitted as shown in figure 1B. And finally, the heatmap has been produced using the parameters shown in figure 1C. During this process two spreadsheets about the protein and epitope data and one picture showing the map have been produced.

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, the just mentioned workflow has been applied for all the organisms mentioned in figure 1E. Thus, strains of *Staphylococcus aureus* have been compared to strains of *Influenza A* and *Coronavirus*. Here, the epitope density has been used as the comparing factor. 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* (Fig. 1E).

On the other hand, we were interested in the immunogenicity of each individual protein and how the putative epitopes are distributed. The past pandemic enhanced the motivation to find the most immunogenic area in the proteome to be used for further clinical applications.

Ordered by their epitope density, the heatmap (Fig. 1D) displays every protein of the retrieved proteome until the 500th amino acid. The first cutoff value from Figure 1c is not used since the proteome contains only 17 proteins (Fig. 1D). Adding up the DiscovEpi epitope scores on the respective position in the protein, there are highlighted areas revealing epitope rich regions. Here, the first 4 proteins really seem to contain immunogenic peptides which could be considered candidates in e. x. vaccine design (Fig. 1D).

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.

Paul, Sinu; Croft, Nathan P.; Purcell, Anthony W.; Tscharke, David C.; Sette, Alessandro; Nielsen, Morten; Peters, Bjoern (2020): Benchmarking predictions of MHC class I restricted T cell epitopes in a comprehensively studied model system. In: *PLoS computational biology* 16 (5), e1007757. DOI: 10.1371/journal.pcbi.1007757.

Reynisson, Birkir; Alvarez, Bruno; Paul, Sinu; Peters, Bjoern; Nielsen, Morten (2020): NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. In: *Nucleic acids research* 48 (W1), W449-W454. DOI: 10.1093/nar/gkaa379.

UniProt: the Universal Protein Knowledgebase in 2023 (2023). In: *Nucleic acids research* 51 (D1), D523-D531.

Vita, Randi; Mahajan, Swapnil; Overton, James A.; Dhanda, Sandeep Kumar; Martini, Sheridan; Cantrell, Jason R. et al. (2019): The Immune Epitope Database (IEDB): 2018 update. In: *Nucleic acids research* 47 (D1), D339-D343. DOI: 10.1093/nar/gky1006.

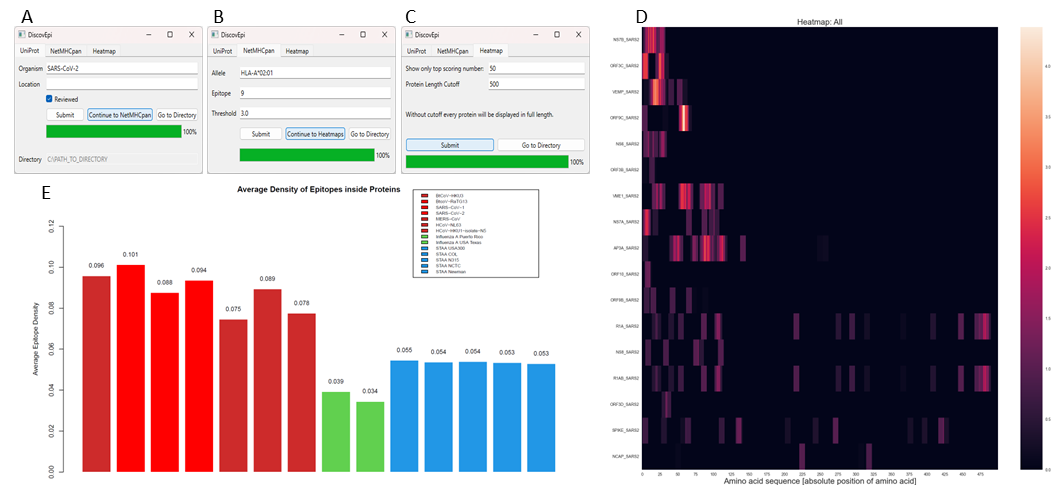


Figure : The Workflow of DiscovEpi: Step one is to retrieve the protein dataset, here the SARS-CoV-2 proteome without specifying the location but limiting to only reviewed proteins (A). The prediction of putative epitopes needs an HLA allele, epitope length and NetMHCpan-score threshold, here HLA-A\*02:01, 9meres and 3.0 (B). Visualization of the prediction needs a maximum number of proteins and maximum length of the sequences to be shown on the map, here 50 proteins shall be shown at a maximum length of 500 amino acids (C). The resulting epitope map is shown under D, where the retrieved proteins are ordered by their epitope density. The usage of proteome-wise average epitope densities is shown comparing the immunogenicity of different organisms, where relatively young viruses show higher values than long with the human coexisting viruses and bacteria (E).