

A Framework for Automating the Generation and Evaluation of Tools Predicting Peptide Binding to MHC class I Molecules

Bjoern Peters¹, Huynh-Hoa Bui², Ole Lund³ and Alessandro Sette⁴

Keywords: MHC, peptide binding, epitope prediction, machine learning

1 Introduction.

MHC class I molecules present peptides derived from intracellular proteins on the cell surface, allowing T-cells of the immune system to detect pathogen infections and cancerous cells[4]. Presented peptides that trigger an immune response in this way are called T-cell epitopes. Predicting which peptides contained in the proteome of a pathogen are likely to bind to a given MHC molecule is an efficient way to identify potential T-cell epitopes. Knowledge of T-cell epitopes is an important step in development of new vaccines and diagnostics.

A large number of computational tools have been developed to predict peptide binding to MHC molecules, with more than 20 tools currently freely available on the internet. For an experimental immunologist, it is hard to make an informed choice which tool to use for a specific problem. This makes a thorough evaluation of the prediction quality of different tools highly desirable.

We here propose a framework which will automate the evaluation of MHC binding prediction tools using experimental testing data. This effort is part of the Immune Epitope Database and Analysis Resource (IEDB) [2]. This recently established NIH funded project is intended to make experimental data obtained from several large scale epitope discovery contracts [3] publicly available, and also to provide and evaluate epitope related analysis tools such as MHC binding predictions. The IEDB already contains more than 50,000 data points of MHC:peptide binding data, many of them never published before. This and the steady stream of data expected from the epitope discovery contracts will allow to evaluate the quality of existing prediction tools, and to compare different methods used to generate them.

2 Evaluation Framework.

One part of the framework is the evaluation of existing prediction tool servers. Peptides for which new binding data is available will be submitted for prediction to all appropriate tool servers before the binding data is made public in the IEDB. Where necessary, a translation server will be set up, which transforms a common prediction input and output protocol into the format required by the existing tool (Figure 1). The second part of the framework compares different methods generating prediction tools from a set of training data. Training data in a standardized form is sent to a

¹ La Jolla Institute for Allergy and Immunology, 3030 Bunker Hill Street, Suite 326, San Diego, CA 92109. E-mail: Bjoern_Peters@gmx.net

² La Jolla Institute for Allergy and Immunology, 3030 Bunker Hill Street, Suite 326, San Diego, CA 92109. E-mail: hbui@liai.org

³ BioCentrum-DTU, Technical University of Denmark, Building 208, Lyngby, Denmark DK-2800. E-mail: lund@cbs.dtu.dk

⁴ La Jolla Institute for Allergy and Immunology, 3030 Bunker Hill Street, Suite 326, San Diego, CA 92109. E-mail: alex@liai.org

prediction method server, which then starts a method specific training process. The result of this training process - for example a trained artificial neural network (ANN) - is then made accessible through a tool server. This allows comparing the predictive performance of tools generated from the same training data, which in effect compares the methods used to generate them. Also, this design is meant to allow any interested scientist to plug his own prediction method into this framework and benefit from the expected continuous increase of data in the IEDB.

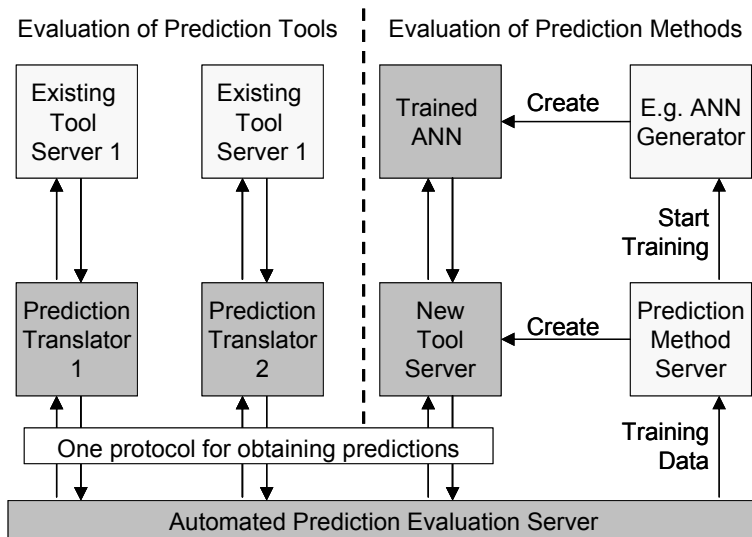


Figure 1: Scheme for the evaluation of prediction tools and methods

Here we present the framework of protocols chosen by us, which is based on a REST [1] style exchange of XML documents over HTTP. This is designed to make the integration of prediction tools and methods independent of soft- or hardware platform as simple as possible, and make each prediction server accessible as a public web service.

4 References.

- [1] Fielding RT, 2000. Architectural Styles and the Design of Network-based Software Architectures. *Doctoral dissertation*, University of California, Irvine.
- [3] Peters B, Sidney J, Bourne P, Bui HH, Buus S, Doh G, Fleri W, Kronenberg M, Kubo R, Lund O, Nemazee D, Ponomarenko J, Sathiamurthy M, Schoenberger S, Stewart S, Surko P, Way S, Wilson S and Sette A, 2005. The Immune Epitope Database and Analysis Resource: From Vision to Blueprint. *PLoS Biology* 3(3): [Epub ahead of print].
- [4] Sette A, Fleri W, Peters B, Sathiamurthy M, Bui H-H, and Wilson S, 2005. A Roadmap for the Immunomics of Category A–C Pathogens. *Immunity* 22: 155-161.
- [5] Shastri N, Schwab S, Serwold T, 2002. Producing nature's gene-chips: the generation of peptides for display by MHC class I molecules. *Annu Rev Immunol.*20:463-93