Project proposal: epitope classification using support vector machines

Bornika Ghosh and Andrew Parker

April 21, 2009

This is our project proposal.

The human body immune response to a vaccine or protein therapeutic largely determines the utility of that drug. Promising medicines which cause an immune reaction are unfortunately not usually viable. An immune function begins when so called antigen presenting cells (APC) ingest and digest exocellular biological molecules and display short parts of the cleaved antigen, an epitope, on the cell surface by binding the epitope to a Class II major histocompatibillity complex (MHC). The binding between an MHC protein and an epitope is the first stage, the detection step, in the bodies adaptive immune response to an antigen. [1] If protein engineers could predict which epitopes bind to MHC, then they could design therapeutic proteins that are less likely to elicit an immune response from patients, which as mentioned, typically renders a medicine useless. The high cost of experimental identification of binding epitopes has posed an urgent need for computational methods for predicting MHC binding epitopes.The accumulation of large quantities of data about MHC protein structure and function has allowed the application of bioinfomatics and machine learning techniques to make realistic models, and hence predictions, of which epitopes bind and do not bind to MHC. Indeed many reviews describe the problem motivation and current work in the field. [3, 8]

We shall employ support vector machine (SVM) approach for epitope binding prediction. The training data set comprises binding and non-binding epitopes for MHC-II alleles. We plan to build SVM models for 8 MHC-II alleles, and use test data sets to measure the power of each model. Due to the scientific interest and practical importance of these predictions a literature exists to assess the accuracy of prediction tools. Not only can we compare our results to published results, but we can also directly use publicly available data sets and webservers from this literature to test and compare our implementation. [2, 6] One popular tool in particular to compare against is the ProPred webserver. [7] ProPred uses ”virtual” matrices which are calculated with mostly bioinformatics tools such as sequence and structure alignment of different MHC Class II alleles. In this fashion biological data available from one allele can be applied towards modeling the other. It is interesting to compare and contrast purely machine learning tools, such as those in cited reviews or our SVM implementation, with the ProPred bioinformatics approach.

Steps in the Project: 1. We plan to design a function that maps the variable length amino acids sequences of epitopes to a fixed length feature vector. Here we wish to use the primary structure

information, i.e. the amino acid sequence, and get a vector of real-valued features, where the vector space dimensions correspond to different physical and chemical properties of amino acids. 2. The next step in the project will be to determine the parameters of the maximum margin hyperplane by solving the optimization problem. 3. Since we have 8 different alleles, each with its own set of binding and non-binding epitopes, we plan to build 8 different SVM models. It should be noted here that though there are 8 different alleles, the number of classes remains two for each. It would be an interesting experiment to train on the data sets of one allele, and test on the data set of another allele. A particular epitope may bind to more than one allele, including every allele. We plan to do a cross validation when training and testing the SVM models. The cross validations can be done using examples exclusively binding to a particular allele, and in addition, using examples binding to different alleles. 4. We will use Receiver-Operating-Characteristic (ROC) curves to determine the prediction power of our SVM models. We also plan to compare our model against other existing techniques for predicting epitope binding, like the ProPred virtual matrices as mentioned above.

Datasets: Several curated databases contain thousands of epitopes characterized through biological assays. We plan to use data sets from databases IEDB and MHCBN. [4, 5] A typical entry in a database contains the epitope primary structure, the MHC Class II allele to which the epitope binds and the binding affinity. The binding affinity can be expressed as a discrete class, for example as none, mild, moderate, or high; or it can be expressed as a positive real number, usually as the IC50 value.

We plan to build an SVM model for at least one of the eight MHC-II alleles by the time for the milestone submission (12th May, 2009). Once we generate the SVM model for one allele, the other models can also be generated, using the same code, but different data sets. The milestone submission shall mark the accomplishment of the most important part of the project, i.e. step 1 and step 2 mentioned above. In the week after the milestone submission, we plan to accomplish steps 3 and 4 mentioned above, i.e. generate ROC curves to assess the power of our models, and compare them against other existing models.

References

[1] A. L. DeFranco, R. M. Locksley, and M. Robertson. *Immunity*. Oxford University Press, 2007.

[2] Y. EL-Manzalawy, D. Dobbs, and H. Vasant. On evaluating mhc-ii binding peptide prediction methods. *PLoS ONE*, 3:e3268, 2008.

[3] A. S. De Groot and L. Moise. Prediction of immunogenicity for therapeutic proteins: State of the art. *Current Opinion in Drug Discovery and Development*, 10(3):332–340, 2007.

[4] S. Lata, M. Bhasin, and G. P. S. Raghava. MHCBN 4.0: A database of MHC/TAP binding peptides and T-cell epitopes. *BMC Research Notes (In press)*, 1:1, 2009.

[5] B. Peters, J. Sidney, P. Bourne, H. H. Bui, S. Buus, G. Doh, W. Fleri, M. Kronenberg, R. Kubo, O. Lund, and et al. The immune epitope database and analysis resource: from vision to blueprint. *PLoS Biol*, 3:e91, 2005.

[6] G. Raghava. MHCBench: Evaluation of MHC binding prediction algorithms. available at http://www.imtech.res.in/raghava/mhcbench/.

[7] H. Singh and G.P.S. Raghava. ProPred: prediction of HLA-DR binding sites. *Bioinformatics*, 17:1236–1237, 2001.

[8] S. Vivona, J. L. Gardy, S. Ramachandran, F. S. L. Brinkman, G. P. S. Raghava, D. R. Flower, and F. Filippini. Computer-aided biotechnology: from immuno-informatics to reverse vaccinology. *Trends in Biotechnology*, 26(4):190–200, 2007.