Using novel Generic String kernel to predict peptide-MHC binding affinity

Denis V. Antonets

State Research Center of Virology and Biotechnology "Vector", Koltsovo, Russia, antonec@yandex.ru

Dmitry S. Grudin

Novosibirsk State University, Novosibirsk, Russia, orinelin@gmail.com

CD8+ T-cell epitopes play crucial role in antiviral and anticancer immunity. Reliable prediction of T-cell epitopes remains one of the most important tasks of immunoinformatics since accurate *in silico* identification of potent epitopes could drastically reduce materials and time consumption in comparison to traditional experimental approaches of epitope discovery.

The main goal of this work was the development of new statistical models for predicting peptide binding to different allomorphs of human MHC class I molecules using novel GS (Generic String) kernel function [1] to measure oligopeptides similarity. The main advantage of GS kernel is its unique ability to consider both substring position uncertainty and physicochemical properties of amino acids, making GS able to compare peptides of different lengths while considering mutual similarity of their amino acid residues. Since the choice of parameterization scheme substantially influence the accuracy of the model, the following amino acid similarity matrices were used: PMBEC [2], BLOSUM62, modified THREADER_NORM matrix (THDR) [3] and amino acid identity matrix. New regression models were built using support vector regression and relevance vector regression techniques. The latter yields less complex models and thus it is expected to be more resistant to overfitting. Whereas GS allows one to train models, using datasets containing peptides of different lengths, our current models were built using the same training and testing sets of nonameric peptides that were used to produce TEpredict sparse partial least squares-based models (http://tepredict.sourceforge.net). It was done to allow straightforward comparison of new models with TEpredict.

Methods and Algorithms: GS kernel was implemented as it was described by the authors [1]. To optimize computational performance it was implemented using C++ and integrated into R

via Rcpp package. Predictive models were built using support vector regression and relevance vector regression algorithms implemented in R package kernlab. Performance of produced models was assessed using ROCR. All programs were written in R (http://r-project.org).

Results: Almost all the models demonstrated good quality of predictions: median AUC (Area Under a ROC Curve) values for BLOSUM62-, PMBEC-, THDR- and identity matrix-based models were 0.9118, 0.9102, 0.9031 and 0.9025, respectively. Median Pearson's correlation coefficients between predicted and experimentally determined pIC50 values of MHC:peptide binding were about 0.71, reaching 0.85-0.97 for several well studied HLA alleles. Both BLOSUM62- and PMBEC-based models surpassed identity matrix-based ones both in terms of AUC and Pearson's correlation coefficient, as expected. Parameterization of amino acids with PMBEC was recently shown to be superior to BLOSUM62 encoding for predicting peptide:MHC binding [2]. Using GS with SVR we were unable to find statistically significant advantages of PMBEC-based encoding over BLOSUM62-based one. However, encoding amino acids with PMBEC yields significantly less complex models with RVR algorithm in comparison to other parameterization schemes tested. Comparative testing of our new models built with GS kernel against recent TEpredict SPLS-based models have shown that new models are superior both in terms of AUC and Pearson's correlation coefficient. More detailed testing results and ROC curves could be found at http://tepredict.sourceforge.net/GSK-SVR/. In near future new regression models will be included into updated version of TEpredict.

This work was supported by Russian Foundation for Basic Reasearch grant #12-04-31746 mol_a.

1. S. Giguere et al. (2013) *BMC Bioinformatics*, **14**:82.

2. Y. Kim et al. (2009) *BMC bioinformatics*, **10**:394

3. Z. Dosztanyi and A.E. Torda (2001) *Bioinformatics*, **17**:686–699.