DTU

*Algorithms in Bioinformatics*

*Mini project no. 01 2010*
*Pages 1–2*

# Stabilization matrix method - Mini project

Sigmar Stefansson and Francesco Favero

Danmarks Tekniske Univeristet

22 October 2010

Supervisor: Morten Nielsen

## ABSTRACT

**Summary** We look at two Stabilization matrix methods [1] used for predicting binding affinity of immunogenic peptides to major histocompatibility complex [3] (MHC) molecules.

The testing data available was from 35 different MHC molecules from HLA-A and HLA-B, with number of peptides ranging from 59 to 3089.

In order to avoid Overfitting [4] We studied the relsults of the Monte Carlo and the gradient decent implementations of the SMM, to find the optimal setting of the algorithms.

## INTRODUCTION

### Stabilisation Matrix Method

Blabla bla.

$$E = \frac{1}{2}\sum_i (O_i - t_i)^2 + \lambda \sum_l w_i^2 \qquad (1)$$

### Cross Validation

Bla bla bla. BlaBlaBla.

### Overfitting

## MATERIAL AND METHODS

We were given semi-completed code for the implementation of the Stabilization matrix method. The minimization procedures were gradient descent in one case and monte carlo in the other.

The testing data available was from 35 different MHC molecules from HLA-A and HLA-B, with number of peptides ranging from 59 to 3089. In addition to finishing the code we added some optimizations like skipping unnecessary memory allocations at performance critical locations.

We ran the matrix creation on each 35 MHC molecules to get the broadest perspective.

The data for each MHC molecule was splited into five parts, equally (or nearly equally) sized. Each part was then used for the validation by the Leave One Out cross validation method (LOO-CV) [2]. So each part was trained on the rest (4/5) of the data, resulting in 5 iterations to be used for the evaluation.

In order to find the optimal $\lambda$ value we evaluate the correlation of all the 35 molecules with both algorithms with different $\lambda$ parameter. To reduce the running time of the procedure we splitted the list of the 35 dataset in 7 different jobs taking care of 5 dataset each.

As noted in Figure 1 and in Figure 2 the use of the $\lambda$ value suppresses the effect of noise in the measured data. In the case of zero value $\lambda$, the optimal entries for the weight vector $w$ minimize the difference between predicted and measured values. Minimizing with a non-zero value for $\lambda$ results in a shift of the optimal entries in $w$ towards values closer to zero. It should be noted here that the $\lambda$ value used as a parameter in the code is the per-target normalized $\lambda$ value, not the global one.

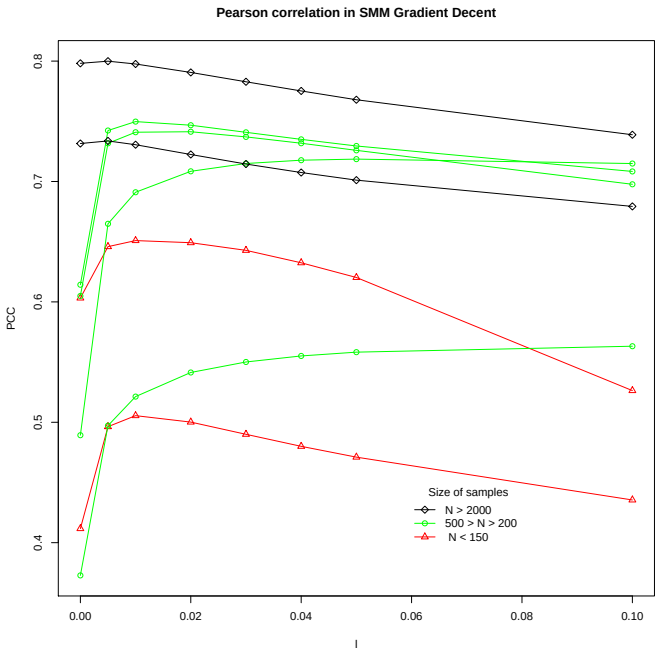


**Fig. 1.** Pearson's correlation coefficent from the SMM Gradient Decent,of different sample using increasing values of $\lambda$. The sample were grouped by the size of the dataset. We can see how dataset of similar size show a similar trend of the Pearson correlation $\lambda$ depending.

## RESULTS

Figure 1 shows the trend for Pearson's correlation coefficient (PCC, prediction accuracy) using different $\lambda$ values for various data sizes.

As a general rule, as the data size gets larger the less important the effect of a non-zero $\lambda$ becomes. For large data sets a $\lambda$ value of zero seems to give the best results. A large dataset should average

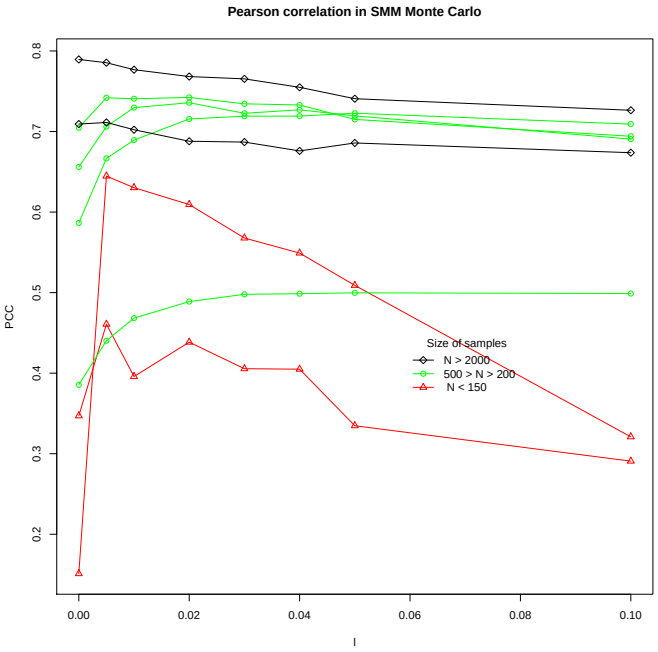


**Fig. 2.** Pearson's correlation coefficent from the SMM Monte Calro, of different sample using increasing values of $\lambda$. The sample were grouped by the size of the dataset. We can see how dataset of similar size show a similar trend of the Pearson correlation $\lambda$ depending.


**Fig. 3.** To put Sigmar fig

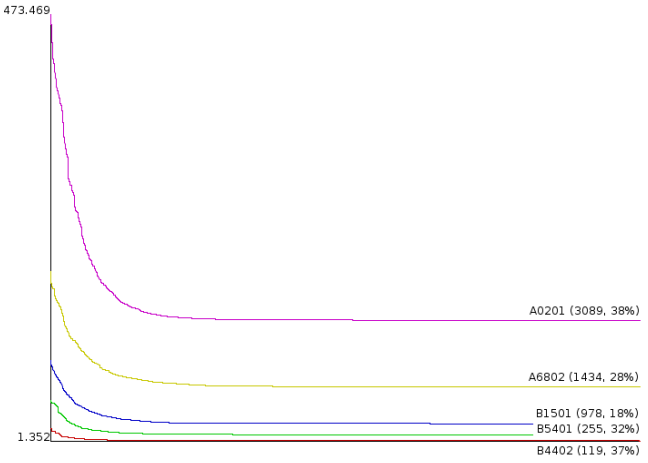


**Fig. 4.** To put Sigmar fig2

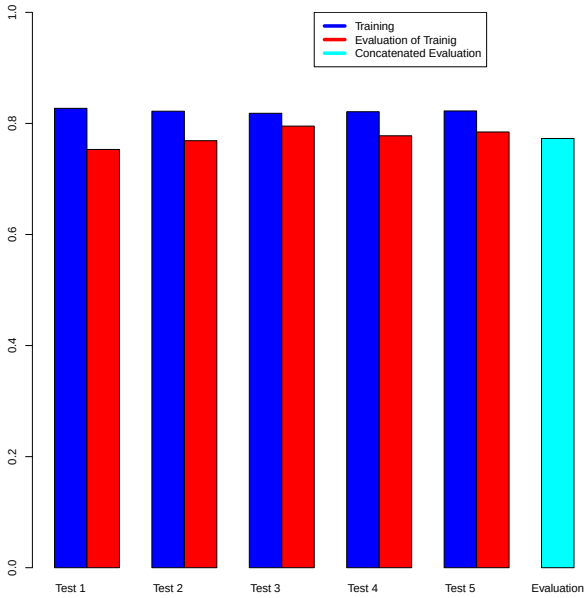


**Fig. 5.** Barplot showing the Pearson Correlation for each training set, for each evaluation on the test set and the correlation of the concatenate evaluation among the 5 test set

out the noise. Similar graph showing results from the Monte Carlo version says the same story (Figure 2).

For the small datasets an optimal $\lambda$ value is less than 0.01. Next we run the smm algorithms on a data from five different kinds of MHC molecules (with different data sizes) showing the error as a function of iterations. For the smm algorithm using gradient descent and 500 iterations, a default lambda value of 0.05 is used (Figure 3). Similar chart is shown for the monte carlo version of the algorithm in Figure 4 showing 12000 iterations. The reason for showing different data sizes on the same chart is just for showing the trends in the two different algorithms. The gradient descent method converges much faster to an optimal value than the monte carlo method. The data sizes and percentage of number of binding peptides are shown in the parantheses right by the names of the lines representing each MHC molecule.

## REFERENCES

[1]B. Peters and A. Sette. Generating quantitative models describing the sequence specificity of biological processes with the stabilized matrix method. *BMC Bioinformatics*, 6(132), 2005.

[2]Wikipedia. Cross-validation (statistics), 2010. [Online; accessed 28-September-2010].

[3]Wikipedia. Major histocompatibility complex, 2010. [Online; accessed 18-October-2010].

[4]Wikipedia. Overfitting, 2010. [Online; accessed 15-October-2010].