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 [2] (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.

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.

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) **??**. So each part was trained on the rest (4/5) of the data, resulting in 5 iterations to be used for the evaluation. As noted in Figure 1 in and Figure 2 the use of the λ value suppresses the effect of noise in the measured data. In the case of zero value λ, the optimal entries for the weight vector w minimize the difference between predicted and measured values. Minimizing with a non-zero value for λ results in a shift of the optimal entries in w towards values closer to zero. It should be noted here that the λ value used as a parameter in the code is the per-target normalized λ value, not the global one.

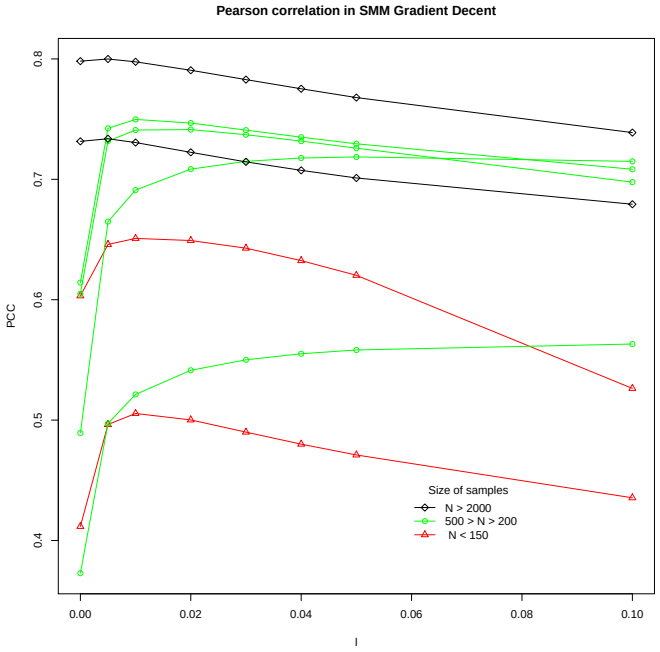

Fig. 1. Pearson's correlation coefficent from the SMM Gradient Decent,of different sample using increasing values of λ. The sample were grouped by the size of the dataset. We can see how dataset of similar size show a similar trend of the Pearson correlation λ depending.

RESULTS

Figure 1 shows the trend for Pearson's correlation coefficient (PCC, prediction accuracy) using different λ values for various data sizes. As a general rule, as the data size gets larger the less important the effect of a non-zero λ becomes. For large data sets a λ value of zero seems to give the best results. A large dataset should average out the noise. Similar

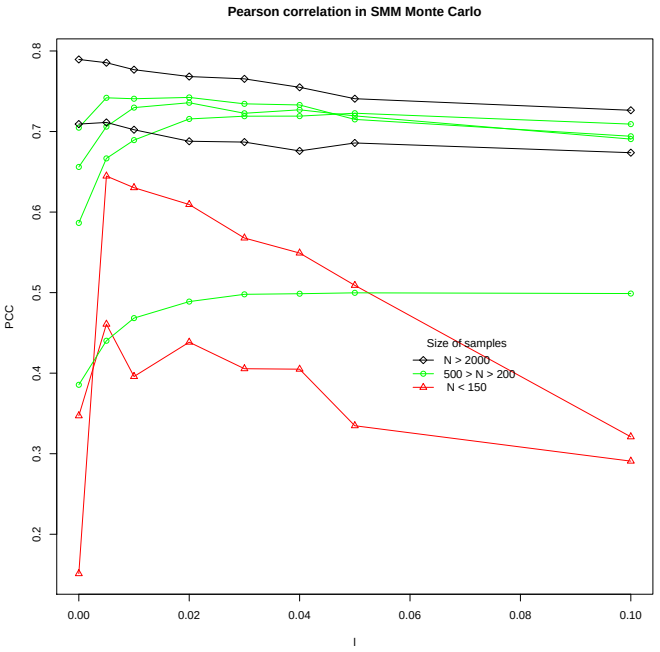

Fig. 2. Pearson's correlation coefficent from the SMM Monte Calro, of different sample using increasing values of λ. The sample were grouped by the size of the dataset. We can see how dataset of similar size show a similar trend of the Pearson correlation λ depending.

graph showing results from the Monte Carlo version says the same story (Figure 2). For the small datasets an optimal λ value is less than 0.01. Next we run the smm algorithms on a data from four different kinds of MHC molecules showing the error as a function of iterations. Here the default λ value of 0.05 is used (Figure 3).

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

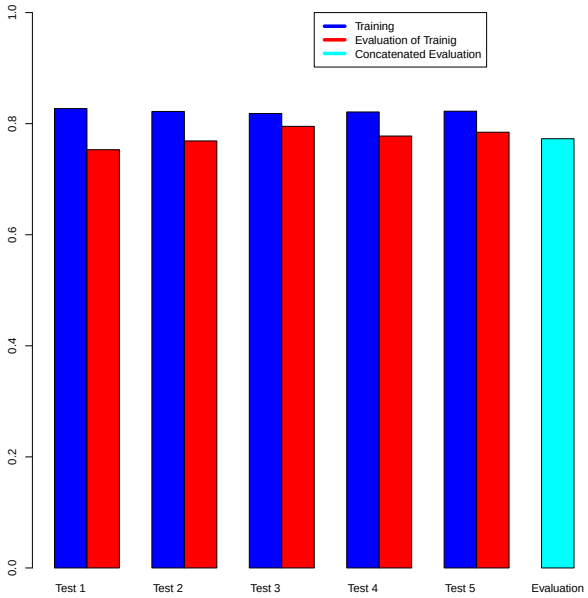

Fig. 3. Barplot showing the Pearson Correlation for each training set, for each evaluation on the test set and the correlation of the concatenate evaluation among the 5 test set