Unbiased Optimization of Microarray Pre-processing

Najmeh Abiri[1], Payam Delfani[2], Mattias Ohlsson[1], Christer Wingren[2], Patrik Edén[1]
1: Computational Biology & Biological Physics, Dept. of Astronomy and Theoretical Physics, Lund University
2: Affinity Proteomics, Dept. of Immunotechnology, Lund University

Corresponding author: Patrik Edén, patrik@thep.lu.se

The objective

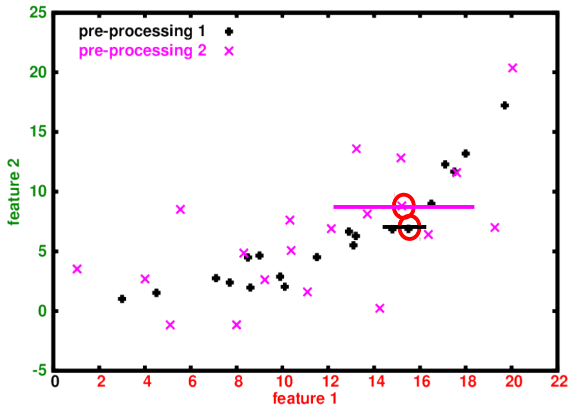
Standard statistical methods, preferably involving test sets, can control false discovery rates in the enormously flexible microarray data analysis. However, it is normally assumed that a similar flexibility in pre-processing (e.g. quality control, normalization and variance filter) was not exploited with knowledge of sample annotations. This leaves the typical research group with the unpleasant choice to either abstain from pre-processing optimization or lose formal control of their statistical tests.

We develop new computational tools that optimizes pre-processing without any use of sample annotations, or any use of sample cluster structure.

The Tool: Validated Imputation

High-throughput microarray data is expected to be rich on correlations. This explains the success of imputation algorithms, that exploit the correlations to estimate missing values. Imputation algorithms are usually evaluated by artificially removing known data, impute them, and check the error to the true values.

Our approach, which we call **Validated Imputation (VI)**, is to use this imputation test "backwards". Instead of testing imputation algorithms with benchmark data, we test pre-processing options with benchmark imputation algorithms.



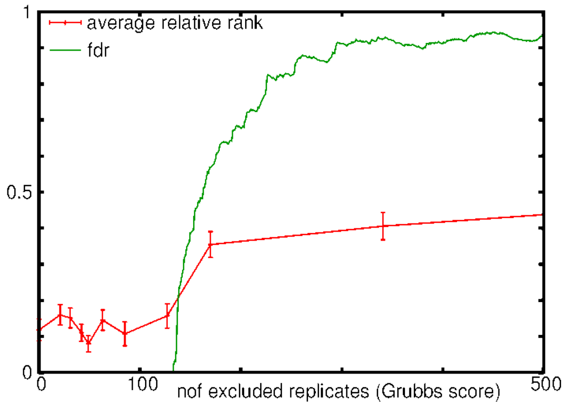
The principal idea is that proper quality filters and noise reduction give better imputation.

Artificial example: The same data in two pre-processings. The feature 1 value of the encircled sample is artificially removed (for both pre-processings) and re-imputed, ending up somewhere on the line. In the noisy (purple) option, that may be far away from the correct value.

A Test Case

Our protein affinity array includes 3-8 replicate spots, where technical errors may appear as outliers. This introduces an outlier threshold as a pre-processing parameter. Outliers among triplicates can be estimated using the Grubbs score (maximal distance to sample mean, divided by sample estimate of standard deviation).

Under normality assumption, the *p*-value and a false-discovery rate (fdr) is calculable. High fdr implies that many points rejected as outliers actually carried useful information.

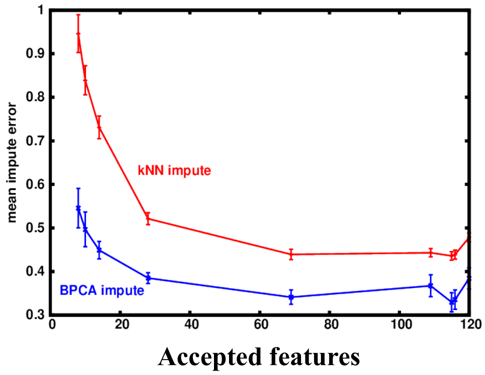


We have run validated imputation on data with various outlier thresholds, and checked which alternative that most often performed best. All threshold options imputed the same set of 5% artificially removed values, and they were ranked according to mean squared error. The rank was scaled to a number between 0 and 1 (relative rank). This was done multiple times, and the average relative rank was recorded. As seen in the figure, VI agrees in conclusion with the fdr analysis. Both methods suggest that the outlier threshold should be set to assign roughly 120 outliers. With a more stringent threshold, the fdr curve shows that a large fraction of excluded data carries useful information, and the VI test shows that imputation becomes less successful.

Discussion

The test case is a promising result, suggesting that validated imputation works: it can rank different pre-processing options. In this case, the suggested optimum agrees with an alternative approach (the fdr based on normality assumption). Note that VI provides slightly more information than the statistical test: The fdr curve shows when a large fraction of extra excluded values carries useful information, but fdr cannot tell if it is worth the prize, in order to exclude a few more technical errors. VI settles the question.

One other merit with VI is that it can be used also when there is no simple statistical test available, and it can be used to evaluate more or less heuristic procedures for, *e.g.*, background correction or normalization. For the protein recombinant antibody array in particular, the relatively low number of features measured means that one must look for normalization strategies other than standard approaches for mRNA and DNA arrays.



We have examined if VI can be run with "any" benchmark imputation algorithm, by comparing the relatively simple and quick knn-impute algorithm [1] with the more elaborate but slow bpca-impute algorithm [2]. Here, we examined a variance filter, removing low-varying features. Overall, BPCA performs better, but the important message is that both algorithms agree on the *conclusion*. (In this case that a variance filter is not really needed, except perhaps to remove 3-5 of 120 features.)

Conclusion and outlook

Validated Imputation is a promising tool to allow pre-processing optimization of high-throughput data without being influenced by any final data analysis results. We will continue to develop it within the framework of protein antibody array data, to examine quality control filters and normalization strategies. The basic method is in principle applicable to any high-throughput data with inherent correlations, for which imputation algorithms outperform simple row-average imputation.

References:

[1] Troyanskaya et al., Bioinformatics 2001, 17 p.520
[2] Oba et al., Bioinformatics 2003, 19 p.2088