

Title: Missing value estimation methods for DNA Microarrays.

Authors:

Olga Troyanskaya<sup>1</sup>, Michael Cantor<sup>1</sup>, Orly Alter<sup>2</sup>, Gavin Sherlock<sup>2</sup>, Pat Brown<sup>3,6</sup>, David Botstein<sup>2</sup>, Robert Tibshirani<sup>4</sup>, Trevor Hastie<sup>5</sup>, Russ Altman<sup>1</sup>

<sup>1</sup>Stanford Medical Informatics, Stanford University School of Medicine  
Departments of <sup>2</sup>Genetics, <sup>3</sup>Biochemistry, <sup>4</sup>Health Research & Policy and Statistics, <sup>5</sup>Statistics and Health Research & Policy, and <sup>6</sup>Howard Hughes Medical Institute, Stanford University

e-mail contacts: Olga Troyanskaya [olgat@smi.stanford.edu](mailto:olgat@smi.stanford.edu)  
Michael Cantor [mcantor@smi.stanford.edu](mailto:mcantor@smi.stanford.edu)  
Russ Altman [altman@smi.stanford.edu](mailto:altman@smi.stanford.edu)

Gene expression microarrays often generate data containing multiple missing expression values. Missing data can result from dust or scratch on the slide, weak fluorescence signal, insufficient difference in signal between the sample and the control, or a contaminated sample. Many algorithms for gene expression analysis (e.g. principle components analysis, independent components analysis) require a complete matrix of gene array values as input. In addition, other methods such as hierarchical clustering and K nearest neighbors may benefit from eliminating the need to ignore dimensions with missing data in the analysis. Therefore, automated methods for estimating missing data would be a useful addition to the array analysis tool kit.

We present a comparative study of several methods for estimating missing values in gene microarray data. We implemented and evaluated four methods: A Singular Value Decomposition (SVD) based method, weighted K-nearest neighbors, and K nearest neighbors in eigen space, and gene row averaging. The methods were evaluated using a variety of parameter settings and over different data sets (high and low noise in data, different types of experiments) and the robustness of the imputation methods to the amount of missing data was assessed. We report results of the comparative experiments and provide recommendations for accurate estimation of missing microarray data under a variety of conditions.