

***In Silico* Microdissection of Microarray Data from Heterogeneous Cell Populations**

Harri Lähdesmäki,¹ Ilya Shmulevich,² Valerie Dunmire,²
Olli Yli-Harja,¹ & Wei Zhang²

¹Institute of Signal Processing, Tampere University of Technology,
Tampere, Finland

²Cancer Genomics Laboratory, The University of Texas M. D. Anderson
Cancer Center, Houston, Texas, USA

Motivation: Very few approaches have been reported to account for the variability in the biological sample preparation for microarray measurements. A fundamental problem concerns sample heterogeneity. For example, tissue samples used in cancer studies are usually contaminated with the surrounding cell types. This heterogeneity in the sample preparation hinders further statistical analysis steps, significantly so if different samples contain different proportions of these surrounding cell types. Thus, sample heterogeneity can result in the identification of differentially expressed genes that may be unrelated to the biological question being studied. Similarly, irrelevant gene combinations can be discovered in the case of gene expression based classification.

Results: We propose a computational framework for removing the effects of sample heterogeneity by “microdissecting” microarray data *in silico*. The computational method provides estimates of the expression values of the pure (non-heterogeneous) cell samples. The inversion of the sample heterogeneity can be facilitated by providing accurate estimates of the mixing percentages of different cell types in each measurement. For those cases where no such information is available, we develop an optimization-based method for joint estimation of the mixing percentages and the expression values of the pure cell samples. The efficiency of the proposed methods is illustrated by applying them to cDNA microarray data obtained from heterogeneous samples.