In Silico Microdissection of Microarray Data from Heterogeneous Cell Populations

Harri Lähdesmäki[1], Ilya Shmulevich[2], Valerie Dunmire[2], Olli Yli-Harja[1] & Wei Zhang[2]

[1]Institute of Signal Processing, Tampere University of Technology, Tampere, Finland
[2]Cancer Genomics Laboratory, The University of Texas M. D. Anderson Cancer Center, Houston, TX, USA

THE UNIVERSITY OF TEXAS MD ANDERSON CANCER CENTER

Introduction & Motivation

Biological samples rarely consist of homogeneous cell populations. For example, tissue samples used in cancer studies are usually contaminated with the surrounding cell types. The complexity of this problem is that different tumor tissues contain different proportions of these non-tumor cells. Therefore, if tumor tissues are used without consideration of such a mixing phenomenon, biological conclusions will certainly be confounded by the heterogeneous cell populations.

Traditionally this problem has been solved by either excluding the heterogeneous samples or by using laser capture microdissection — both approaches are problematical. In this study, we develop a computational method for removing the effect of mixing from heterogeneous samples. Similar analytical approaches have been previously proposed in (Lu *et al.*) and (Stuart *et al.*). Lu *et al.* focused on estimating the fraction of cells in different phases of the cell cycle whereas Stuart *et al.* considered the problem of estimating the cell type specific expression patterns over all samples. Here we focus on estimating both the sample and cell type specific expression values using carefully controlled microarray experiments.

Materials

4704 genes of the lymph node sample (normal) and the colon cancer sample (RKO) were measured in duplicate. In total, five different heterogeneous mixtures of normal and RKO were measured using five different cDNA microarrays. The measured mixing percentages are shown in Table 1.

TABLE 1: The measured mixing percentages (RKO/normal) in the five heterogeneous samples.

	sample #1	sample #2	sample #3	sample #4	sample #5
mixing %s	100/0	80/20	56/44	30/70	0/100

Methods

Preprocessing

The replicated background-subtracted signal intensities were averaged, $\log_2$-transformed, and lowess normalized. After the correction of the dye-bias, the data were transformed back to the original domain and single-channel data were used for further analysis. The data were further standardized (mean and variance-equalized) for each array and the two channels separately.

Modeling sample heterogeneity

The two samples are mixed at the extracted RNA level. Therefore the model can be assumed to be linear. Let x_i^c and x_i^l denote the expression level of the ith gene in the colon cancer (RKO) and in the lymph node (normal) samples, respectively. Assuming only two different cell types are mixed, the sample heterogeneity is modeled by a simple linear model

$$y_i^k = \alpha_k x_i^c + (1-\alpha_k)x_i^l. \tag{1}$$

where y_i^k denotes the expression value of the ith gene in the kth heterogeneous sample, and $0 \le \alpha_k \le 1$ denotes the fraction of the colon cancer cells in the kth mixture. Note that the expression levels x_i^c and x_i^l are assumed to be "fixed" and do not change between heterogeneous measurements. Also note that the linear mixing model can be extended to incorporate more than just two cell types using a straightforward extension:

$$y_i^k = \sum_j \alpha_k^j x_i^j, \quad \text{where } 0 \le \alpha_k^j \le 1 \text{ and } \sum_j \alpha_k^j = 1 \tag{2}$$

Inversion of sample heterogeneity

The first objective is to obtain estimates for the expression values of the pure cell types x_i^c and x_i^l. In order to avoid making additional modeling assumptions, we prefer to use a general purpose least squares method. Let the number of genes be n and assume that one has measured the expression values for K different heterogeneous mixtures, i.e., y_i^k, $1 \le i \le n$, $1 \le k \le K$. Let us also assume for now that the mixing percentages (α_k) are known or have been measured. For the ith gene the sample heterogeneity can be expressed as (excluding all noise terms)

$$\begin{pmatrix} \alpha_1 & 1-\alpha_1 \\ \vdots & \vdots \\ \alpha_K & 1-\alpha_K \end{pmatrix} \begin{pmatrix} x_i^c \\ x_i^l \end{pmatrix} = A\mathbf{x}_i = \mathbf{y}_i = \begin{pmatrix} y_i^1 \\ \vdots \\ y_i^K \end{pmatrix}. \tag{3}$$

When including all the n genes, the above model can be rewritten as

$$\begin{pmatrix} A & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & A & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & A \end{pmatrix} \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_n \end{pmatrix} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_n \end{pmatrix}, \tag{4}$$

where $\mathbf{0}$ denotes the K-by-2 zero matrix. Let the block matrix in Equation (4) above be denoted as $\tilde{A}$. Assuming the column rank of A is full, the well-known least squares solution is given by

$$\hat{\mathbf{x}} = (\tilde{A}^T\tilde{A})^{-1}\tilde{A}^T\mathbf{y}, \tag{5}$$

where $\mathbf{y} = (\,\mathbf{y}_1^T\ \mathbf{y}_2^T\ \dots\ \mathbf{y}_n^T)^T$.

Because of the inherent variability of individual gene expression values, we chose to examine the performance of our method globally, by comparing the measured and estimated expression values of all the genes simultaneously using the standard principal component analysis (PCA). Results of the inversion of the sample heterogeneity are shown in Figure 1.

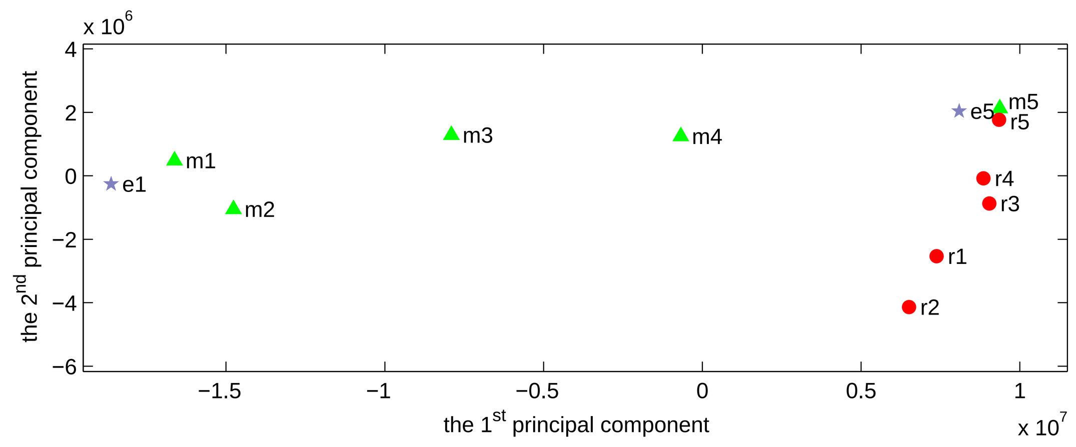

FIGURE 1: Results in the 2-dimensional PCA space. All five heterogeneous samples are used to estimate the expression profiles of the pure cell types. Symbols: estimated expression profiles (gray stars), mixture samples (green triangles), and reference samples (red circles).

Less outlier sensitive results can be obtained with robust regression methods, such as the iteratively reweighted least squares or median-based regression methods. The robust methods provided similar global results, but improved results for some individual genes that contained one or more outliers.

Optimization of mixing percentages

Since the mixing percentages must be measured by some means, they are also likely to contain some error. Assuming the model in Equation (4) is sufficiently over-determined, the mixing parameters can be adjusted computationally, too. This results in the following optimization problem

$$\begin{aligned} &\min_{\tilde{A},\mathbf{x}} \quad \|\tilde{A}\mathbf{x} - \mathbf{y}\| \\ &\text{subject to } 0 \le \alpha_k \le 1 \quad \text{for all } 1 \le k \le K \end{aligned} \tag{6}$$

In general, any iterative optimization method can be used to get a solution. Due to the large number of parameters, we use a two-step approach in the optimization (see Figure 2).

1. Initialize $\hat{A}^{(1)}$ and set $j = 1$.
2. Minimize $\|\hat{A}^{(j)}\hat{\mathbf{x}}^{(j)} - \mathbf{y}\|$ for $\hat{\mathbf{x}}^{(j)}$:
$$\mathbf{x}^{(j+1)} := (\hat{A}^{(j)T}\hat{A}^{(j)})^{-1}\hat{A}^{(j)T}\mathbf{y}.$$
3. Minimize $\|\hat{A}^{(j)}\hat{\mathbf{x}}^{(j+1)} - \mathbf{y}\|$ for $\hat{A}^{(j)}$ (subject to constraints $0 \le \alpha_k \le 1$ for all $1 \le k \le K$). Increase the iteration index $j := j+1$.
4. Repeat steps 2 and 3.

FIGURE 2: Details of the two-step algorithm used for the optimization problem shown in Equation (6).

The convergence of the optimization method is illustrated in Figure 3, the found optimal values of the mixing percentages are shown in Table 2, and the obtained expression estimates for the pure colon cancer and lymph node samples are shown in Figure 4.

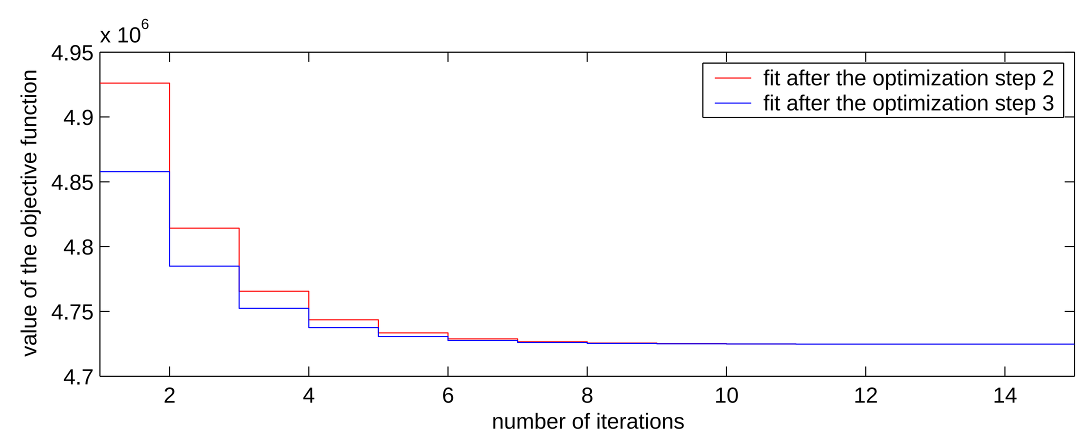

FIGURE 3: Evolution of the value of the objective function.

TABLE 2: The found optimal values of the mixing percentages.

	sample #1	sample #2	sample #3	sample #4	sample #5
mixing %s	100/0	92.96/7.04	65.06/34.94	37.96/62.04	0/100

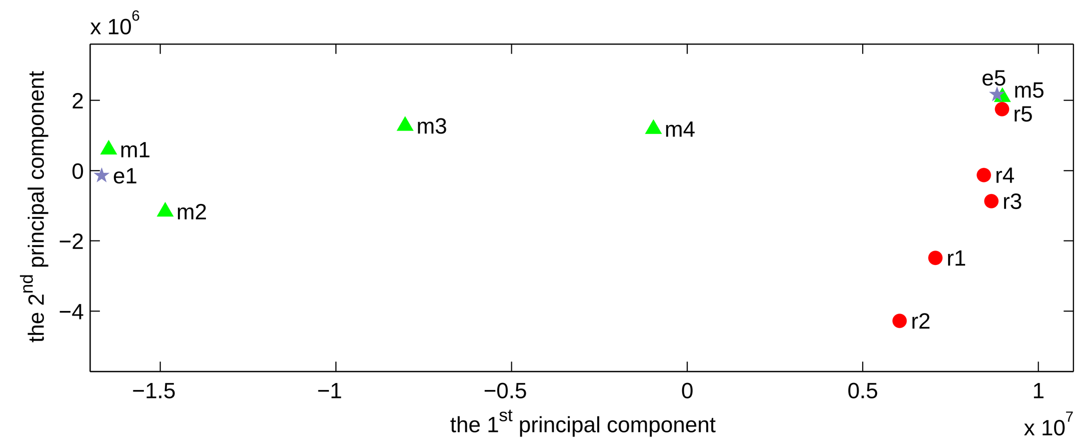

FIGURE 4: Results in the 2-dimensional PCA space. All five heterogeneous samples are used to estimate the most likely values of the mixing parameters and the expression profiles of the pure cell types. Symbols: estimated expression profiles (gray stars), mixture samples (green triangles), and reference samples (red circles).

Selection of the Number of Cell Types

There may be other experimental settings where the number of cell types may be unknown. Then it is useful to assess the validity of the model (i.e., the # of the cell types) as well. Since the standard regression-based statistical tests apply only to Gaussian noise we recommend using a general purpose cross-validation for model selection. Here we consider the leave-one-out cross-validation (LOOCV) and test the one, two, and three cell type models. The optimized (fixed) mixing parameters are used for each model. Table 3 shows the relative LOOCV errors for all the models. The results suggest that the two cell type model is indeed the correct one.

TABLE 3: The relative LOOCV errors for the one, two, and three cell type models.

	1 cell type model	2 cell types	3 cell types
relative LOOCV error	1.79	1.00	2.28

Examples

Consider the following (hypothetical) experimental setting: Given the three middle mixture measurements (#2, #3, and #4), identify a set of differentially expressed genes between the RKO and the normal samples. In a simple approach, the most heterogeneous sample would be discarded because it does not provide any direct discriminative information about the underlying two samples (mixing percentages are 56% and 44%). For illustration purposes, let us measure the expression difference of a given gene between these two samples using the 2 fold change. Thus, without any inversion of the mixing effect, one must compare the mixture measurement y_i^2 and y_i^4. Figure 5 below show some example genes whose expression difference does not meet the given threshold, but after the 'in silico microdissection,' the expression difference exceeds even 4 (or is below 1/4).

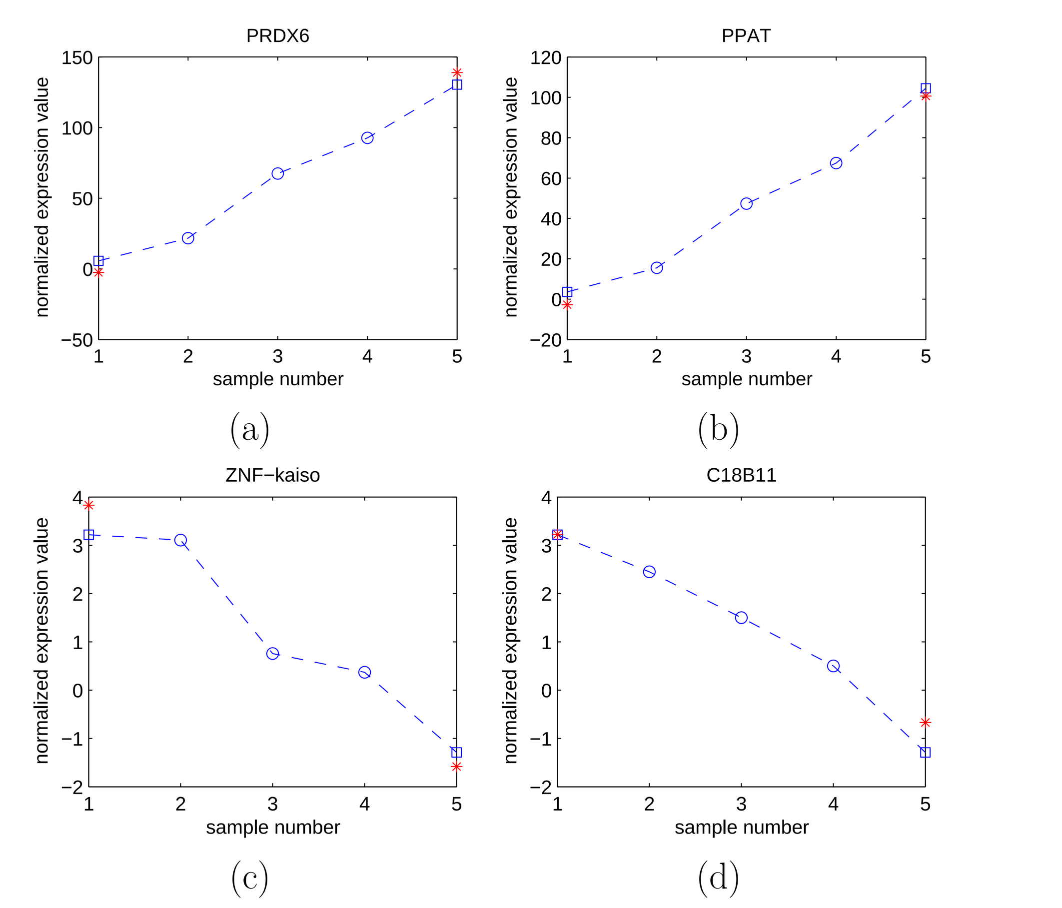

FIGURE 5: A set of genes which are not found to be significantly differentially expressed based on the heterogeneous measurements (samples #2 and #4, blue circles). After the inversion of the mixing effect, however, the expression difference between the estimated pure colon cancer cells and lymphocytes (red stars) meet even a more stringent criterion of differential expression. The measured expression values of the pure colon cancer cells (sample #1) and lymphocytes (sample #5) are also shown (blue squares).

Conclusions & Discussion

- Figures 1 and 4 clearly show that the heterogeneous samples ('m1' through 'm5') are located almost on a straight line parallel to the first principal component, suggesting that the most significant variation in the data is due to the linear mixing effect.
- The estimated expression profile of the pure colon cancer cells and lymphocytes are also close to samples number #1 and #5, respectively, indicating that the inversion of the mixing phenomenon produces reasonable results.
- Figure 4 shows that better fitting can be obtained by simultaneously estimating the mixing parameters. Note that estimation of the mixing parameters may also compensate for some other errors/biases in the data than just the mixing percentages.
- The proposed inversion methods were also tested on standard non-replicated microarray data. The obtained results were qualitatively similar with the ones shown above and only slightly more variable: thus, replicated data is not necessary for the method to work.
- The general purpose model selection method (CV) seems to work.
- As is evident from Figure 5, heterogeneity in the biological sample preparation can hinder further statistical analysis steps, but the proposed method is able to invert the mixing effect. Indeed, the conclusions one can draw based on the estimated expression profiles are consistent with those that are based on the true homogeneous samples (blue squares).
- Similar ideas have also been introduced in the context of composite microarrays (Shmulevich, I., *et al.*), and deconvolution of time-series gene expression measurements (Lähdesmäki, H., *et al.*).

References

Lu, P. *et al.* (2003) Expression deconvolution: a reinterpretation of DNA microarray data reveals dynamic changes in cell populations. *PNAS*, Vol. 100, No. 18, pp. 10370–10375.

Lähdesmäki, H., *et al.* (2003) Estimation and inversion of the effects of cell population asynchrony in gene expression time-series. *Signal Process.*, Vol. 83, No. 4, 835–858.

Shmulevich, I., *et al.* (2003) Data extraction from composite oligonucleotide microarrays. *Nucleic Acids Res.*, Vol. 31, No. 7, e36.

Stuart, R.O., *et al.* (2004) *In silico* dissection of cell-type-associated patterns of gene expression in prostate cancer. *PNAS*, Vol. 101, No. 2, pp. 615–620.