Incorporating spatial information of heterogeneous cell populations into Bayesian mRNA- and miRNA-expression analysis

Timo Erkkilä[1,2], Pekka Ruusuvuori[1,2], Ilya Shmulevich[1,2], Harri Lähdesmäki[1]

[1]*Department of Signal Processing, Tampere University of Technology, P.O. Box 553, FI-33101 Tampere, Finland;*

[2]*Institute for Systems Biology, 1441 North 34th Street, Seattle, WA 98103-8904, United States*

Microdissection of tissue samples is preferred when samples contain significant proportions of cell types not wanted to be measured in a microarray experiment. However, performing manual purification on each tissue sample can be very time-consuming, and reduced yield of mRNA or miRNA may consequently become a bottleneck in array hybridization. On the other hand, biological studies often concentrate on finding differential expression between cell types being spatially connected, e.g., cancer, stromal, and epithelial cells, so that whatever becomes discarded from dissected samples may contain, in fact, information relevant to the whole study.

Several authors have already addressed this problem of sample heterogeneity by proposing statistical models for expression profiles extracted from mixed cell populations. With such approaches no manual dissection is necessary; the idea is to computationally reverse-engineer the cell type specific expression profiles from the mixture profiles.

We follow the footsteps of previous authors by proposing a linear model called “Dsection” for the same reverse-engineering problem, but which is built fully Bayesian. Not only is our model capable of incorporating prior knowledge of proportions of cell types being, say, extracted from digital images of H&E stained tissues in an automatic manner, but the model also allows for taking multiple biological conditions, e.g., treatments, into account simultaneously. This makes it possible to assess scores for differential expressions between any tissue-condition pairs by utilizing Markov Chain samples from the posterior distributions, and by simulations we show that that our score, called “D-score”, outperforms assessments based on simple fold-change of expressions.

We show with simulated and real data that by incorporating prior information on model parameters one is able to obtain more accurate estimates for cell type specific expression profiles than without such information. Furthermore, prior densities can be tuned to reflect the quality of any incorporated, additional information in a natural way.