SAMPLING GENES OR SAMPLING PATIENTS? STRATEGIES FOR FINDING DIFFERENTIALLY EXPRESSED GENE SETS

J. J. Goeman[†1]

[1] *Leiden University Medical Center, The Netherlands*

[†] E-mail: *j.j.goeman@lumc.nl*

An analysis of a microarray experiment typically results in long lists of differentially expressed genes. However, for the biologist, this is not the end point of the analysis, but the start of a time-consuming search in databases and literature to make sense of the results obtained, because the relevant biological question is usually not which genes are differentially expressed, but which biological processes are involved.

Recently, therefore, many statistical methods have been designed which help the biologist to make sense of the data in terms of the biological processes involved. Most of these methods start from the list of significant genes, testing whether these genes are significantly often associated with a certain biological process. Others start from the ordering of the genes provided by the p-values, testing whether the genes associated with a biological process tend to have higher ranks than expected under the null. Some other methods analyze association of biological processes directly from the raw data, without the intermediate step of single gene testing.

An interesting, but hidden methodological issue in these methods is the way the methods look at the data. The majority of the methods analyzes the data as if the gene were the sampling unit, instead of the patient. These methods base their p-values on the hypothetical experiment of drawing genes at random from an urn of genes, for example when deciding whether genes that are associated with a biological process are more often among the significant genes than genes that are not. Only a minority of methods analyzes biological processes in a classical way, using patients as the sampling unit.

This paper reviews the various approaches to analysis of microarray data in terms of the biological processes involved. We argue that users should carefully consider whether they want to use genes or patients as the sampling unit, because different choices may lead to widely different interpretations and results. The most profitable analysis combines the two approaches, but puts most emphasis on the classical patient-centered analysis.