Sergiusz Wesolowski
Dept. of Mathematics, Computer Science, Mechanics; Warsaw University, Poland
e-mail: wesserg@gmail.com
Piotr Kraj
Cancer Center, Medical College of Georgia, Georgia Health Science University, USA
e-mail: pkraj@georgiahealth.edu

Improving statistical models for discovering cell type specific genes

Analysis of gene expression is one of the fundamental methods of characterizing cell populations. One of the major cells in the immune system are "helper" T cells expressing CD4 surface marker. The majority of these cells constitutes a population of conventional CD4+ T cells which supports functions of other cells of the adaptive and innate immune system. A smaller population, called regulatory CD4+ T cells (Treg), has opposite function and suppresses immune response and is responsible for the homeostasis of the immune system. The most characteristic gene expressed by Treg cells is a transcription factor Foxp3. Both conventional and Treg cells are generated in the thymus from bone marrow-derived progenitors. Treg cells produced in the thymus are called natural Treg cells. Under certain conditions, conventional CD4 T cells can express Foxp3 and acquire suppressor function. These Treg cells are called adaptive Treg.

One of the methods of investigating different subsets of CD4 T cells is to compare their gene expression profiles. This approach allows insight into cellular functions of individual cell subsets and allows for analysis of functions of differentially expressed genes. Analysis of the global expression profiles is commonly done using microarrays.

To reveal genetic control of various subsets of CD4 T cells we compared gene expression profiles of resting and activated conventional CD4 T cells, resting and activated natural Treg cells and adaptive Treg cells. RNA was isolated from the respective T cell populations and hybridized to Affymetrix GeneChip M430 2.0 Plus microarrays. Three individual samples of each kind were processed.

In order to make our data set more representative, followin a similar approach described in [1], we included microarrays from the respective CD4 T cell subsets from other laboratories. These data were obtained from the GEO database: www.ncbi.nlm.nih.gov/geo.

To deal with the problem we produced a framework combined from several available statistical approaches: Linear models for Microarray data, Bayesian approach, Non-Negative Matrix Factorisation [2].

Comparisons of data from multiple laboratories introduces additional levels of variability which need to be accounted for during data normalization.

Normalization attempts that adjusted mean values and standard deviation of gene expression resulted in the sets of differentially expressed genes that differed between laboratories instead of between different T cell populations. Our computations indicated that lab origin has more influence on gene expressions then investigated cell types among laboratories.

To account for multi-dimensionality of the normalization problem we developed a heuristic approach.

References

[1] Jean-Philippe Brunet, Pablo Tamayo, Todd R. Golub, Jill P. Mesirov, *Metagenes and molecular pattern discovery using matrix factorization* PNAS **12** 4164–4169.

[2] Franz-Josef Müller, Louise C. Laurent, Dennis Kostka, Igor Ulitsky, Roy Williams, Christina Lu, In-Hyun Park, Mahendra S. Rao, Ron Shamir, Philip H. Schwartz, Nils O. Schmidt Loring, Jeanne F. Loring, *Regulatory networks define phenotypic classes of human stem cell lines* Nature (18 September 2008) **455** 401–405.