# Unsupervised automated population detection and immunophenotypisation tool for analysis of multiparameter flow cytometric data

Podolská Tereza1, Sieger Tomáš2, Fišer Karel1,3

*1 CLIP - Childhood Leukaemia Investigation Prague, Department of Paediatric Haematology and Oncology, Second Faculty of Medicine, Charles University and University Hospital Motol, Prague, Czech Republic.*

*2 Department of Cybernetics, Faculty of Electrical Engineering, Czech Technical University in Prague, Prague, Czech Republic*

*3 Department of Bioinformatics, Second Faculty of Medicine, Charles University, Prague, Czech Republic*

The multiparameter flow cytometry (MFC) is an important part of Acute lymphoblastic leukaemia diagnostics and disease monitoring. The MFC data are conventionally analysed manually by experts in the field. However, such analysis is subjective and relies on high level of expertise. While a number of computational tools for MFC have been developed, no fully automated unsupervised method spanning population identification, phenotypisation and reporting has been proposed yet.

Here we propose a computer-assisted analytical protocol for analysis of MFC files. The analysis consists of four steps: 1) preprocessing, 2) clustering by MFC tailored hierarchical clustering analysis (Fišer et al., 2012), 3) population selection, 4) phenotypisation of the detected populations (Sieger et al., unpublished work). All four steps are performed without human input by issuing a single command.

The population identification step is based on hierarchical clustering and subsequent cluster selection. We developed a cluster selection approach based on cluster compactness evaluation metric (Sieger et al., unpublished work). In our approach the compactness of each dendrogram node is evaluated. In each step a cluster with the highest compactness value is selected and all its parent and children clusters are excluded from further selection. This selection is repeated until all cells belong to a cluster. Thus, the clusters are identified at different dendrogram tree heights.

We tested the proposed automated analytical protocol on a set of fcs files (n = 101) from the same 8-parameter antibody panel. This was a cohort of patients indicated as suspectly having childhood acute leukemia. The resulting immunophenotypes of leukemic populations represented by quantitative metrics were selected and compared with phenotypes reported by expert cytometrists. For six out of eight markers measured in the analysed files we achieved a statistically significant agreement (P < 0.01) of phenotypes reported by our automated protocol and phenotypes reported by expert, according to Spearman’s correlation coefficient.

Our protocol for an automatic MFC data analysis can produce biologically relevant results corresponding with expert reports. Our approach is unsupervised, automatic, fast and therefore is better scalable.

This work was supported by GAUK number 352922.