# Measures of quality of clusters in hierarchical clustering of flow cytometry data

Tomáš Sieger1, Tereza Podolská2, Karel Fišer2,3

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

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

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

**Introduction**

Hierarchical clustering enables unsupervised analysis of multidimensional data, yielding a dendrogram, a hierarchical tree of clusters of data samples. However, the dendrogram does not readily specify the quality of individual clusters in it. If users need to choose „good“ clusters out of the dendrogram, traditionally, they cut the dendrogram at a specific height and pick top-level clusters, or manually cherry pick some clusters. Unfortunately, this would miss „good“ clusters appearing at different heights in the dendrogram, and could be subjective.

**Objectives**

We attempted to find measures of quality of individual clusters in hierarchical clustering that could be used to guide selection of clusters in flow cytometry data both in manual and automated fashion.

**Methods**

We defined theoretical requirements of two measures of quality of clusters in a dendrogram: the compactness, assessing how tightly each cluster is connected to the hierarchy below it, and the separation, assessing how well each cluster is separated from the hierarchy above it. Notably, these quality measures do not intentionally rely on original data, but purely on the dendrogram itself. We devised and implemented nontrivial measures fulfilling the requirements mentioned above and validated them on flow cytometry data (n=10) with annotation of „good“ clusters from two independent researchers.

**Results**

The median intra-rater classification accuracy was 95.0%, and the median inter-rater accuracy was 92.2%. We confirmed that clusters with high values of the theoretically derived measures of quality often corresponded to „good“ ground truth clusters, and that the quality measures enabled to select clusters from the dendrogram automatically with the accuracy of 88.9%.

**Conclusion**

We devised measures of quality of clusters in hierarchical clustering that can be used to guide selection of meaningful clusters in flow cytometry data. Our unsupervised method enables automated processing of large amounts of data without a need of costly and subjective manual intervention. In future, we need to study the general applicability of our measures of cluster quality and validate them on other data sets.

This work was supported by GAUK number 352922.