Characterizing Cancer Subtypes Using Dual Analysis in Caleydo StratomeX

Cagatay Turkay, Alexander Lex, Marc Streit, Hanspeter Pfister, Helwig Hauser

CT: giCentre, Department of Computer Science, City University London,
Cagatay.Turkay.1@city.ac.uk
AL: School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA,
alex@seas.harvard.edu
MS: Institute of Computer Graphics, Johannes Kepler University Linz, Linz, Austria ,
marc.streit@jku.at
HP: School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA,
pfister@seas.harvard.edu
HH: Department of Informatics, University of Bergen, Norway, Helwig.Hauser@uib.no

The comprehensive analysis and characterization of cancer subtypes is an important problem to which significant resources are devoted. Thoroughly analyzing how samples relate to a given subtype and which genes are descriptive for a particular cancer subtype is highly challenging and not completely addressed by existing work.

In our work, we address this challenge by enabling our new dual analysis approach, which integrates statistics and interactive visual data analysis to describe both the dimensions and the rows of a high-dimensional dataset, within StratomeX, a Caleydo view, tailored to cancer subtype analysis. We introduce significant difference plots for showing the elements of a candidate subtype that differs significantly from other subtypes. We also enable analysts to investigate how samples relate to both the subtype they are assigned to and other subtypes. Our approach gives analysts the ability to identify and create well-defined candidate subtypes based on expressive statistical properties. In this respect, this work improves the state of the art in the semi-automatic analysis of subtypes by providing a deeper understanding of stratifications, which, in turn, leads to a better understanding of the unique properties of cancer subtypes and how they differ from each other.

In the future we will enable complex comparisons, for instance, between more than two groups, or between several datasets. Moreover, we plan to investigate the integration of statistical models that automatically derive interpretable rules characterizing a cluster, and provide visual means to choose and represent these rules.