StratomeX: Enabling Visualization-Driven Cancer Subtype Analysis

Alexander Lex* Graz University of Technology
Marc Streit Johannes Kepler University Linz
Hans-Jörg Schulz University of Rostock
Christian Partl Graz University of Technology
Dieter Schmalstieg Graz University of Technology
Peter J. Park Harvard Medical School
Nils Gehlenborg Harvard Medical School & Broad Institute

ABSTRACT

Many types of cancers are known to have subtypes that derive from common biomolecular alterations. Identifying and characterizing these subtypes is important, since knowledge about the alterations in these subtypes can have serious implications on the treatment and prognosis of patients. The lack of appropriate tools, however, makes the complex process of analyzing subtypes a tedious endeavor. We present StratomeX, an application that supports analysts in identifying and characterizing cancer subtypes using interactive visualizations. StratomeX was first published at EuroVis 2012, but we expect that the BioVis community can benefit from learning about this extensible open source system.

1 INTRODUCTION

Cancers are usually classified by the tissue in which they arise. However, most cancer types are not homogeneous within their class, but can differ significantly both in their histology and in their biomolecular profiles. These differences are systematic and are characterized in cancer subtypes. Cancer subtypes are highly relevant as they have a profound impact on prognosis and treatment of the affected patients. Therefore, significant efforts have been undertaken to uncover the genetic causes of cancer subtypes (*e.g.,* [2]). *The Cancer Genome Atlas (TCGA)*[1] project, for example, is generating genomics data from patients suffering from various types of cancer on a large scale. For 20 different tumor types and around 500 patients for each type, a wide range of genome-wide data is collected, including mRNA, microRNA, methylation, copy number and mutation status data. Additionally, clinical parameters, such as age, survival time, treatment information, *etc.* are available.

The analysis of cancer subtypes is based on the integrated analysis of these datasets. For example, candidate subtypes can be determined by clustering an mRNA expression dataset, which results in a *stratification* of the dataset into mutually exclusive groups of patients. In the ideal case, these groups correspond directly to subtypes. However, it is more likely that subtypes are characterized by a combination of stratifications of several datasets. The validity and relevance of candidate subtypes can be evaluated by considering clinical data. For example, a distinct pattern in the survival chances of patients within a particular candidate subtype can be interpreted as supporting evidence for the presence of the subtype. Similarly, the presence of a subtype can manifest itself in particular gene activation patterns in certain pathways. A notable difference of a candidate subtype's effect on a pathway can also support a given stratification.

However, such analyses are difficult and inefficient to conduct using traditional means, for example, statistical analysis and the generation of static ad hoc plots. To make the analysis of cancer subtypes more efficient and easier to interpret, we have developed **StratomeX** [1], a visualization technique targeted specifically at cancer subtype analysis. StratomeX is part of Caleydo[2], an open source tool for visualizing biomolecular data. While StratomeX can be used with arbitrary data, it also ties in with the *Firehose* data analysis pipeline[3] developed for TCGA, which preprocesses and analyzes the data and computes candidate stratifications through clustering. Firehose is run monthly on the most recent data collected by TCGA and a Caleydo project is automatically generated for each tumor type, which is made available for download[4].

2 THE STRATOMEX VISUALIZATION TECHNIQUE

We explain the visualization technique using the TCGA glioblastoma multiforme (a brain cancer) dataset. As shown in Figure 1, StratomeX visualizes datasets as columns, which are split into *bricks* based on a stratification. At the very top of each column, a small summary view shows information about the whole dataset. Below the summary view, each group of patients is shown in a separate brick. Relationships between the columns are visualized with ribbons between the bricks. Wide ribbons indicate many shared elements, whereas thin ribbons indicate outliers. The first two columns in Figure 1 show the same mRNA dataset. The only difference is that the dataset was stratified into four groups of patients in the first column, while the second column contains five groups. For each group of patients, the mRNA data is shown in a heat map, so that the homogeneity of expression profiles within the group can be evaluated. Based on the patterns of the ribbons, we find that these two stratifications only partially support each other. For example, the groups at the bottom seem to correlate strongly, while the second group from the top in the first column is split among three groups in the second column. The first group in the second column is almost completely contained in the first group of the first column.

When we look at the clinical data for this group, we can observe striking differences. The small multiples of Kaplan-Meier plots in the third column show the data for the clinical variable "days to death" (indicating survival time after diagnosis), using the same stratification as its neighboring column. We observe that the prognosis for patients in the topmost group seems to be especially poor, with less than 20% to survive the second year after the diagnosis. While such detailed information is not available in the overview shown, each group can be shown in a detail mode where more information, such as axes and scales, is visible.

The fourth column shows pathways, overlaid with mRNA expression data of the associated groups. We can see several differences in the mapped data. For example, the gene denoted by the arrow in Figure 1 is highly expressed as indicated by the dark red color, while the same gene seems to be low in all other groups.

Finally, the rightmost column shows the copy number variation data of the gene *CDKN3*. The light and dark blue bricks encode that one respectively both copies of this gene were deleted in the genomes of these patients, which at this scale is unlikely to be random. For more information visit the StratomeX website[5].

*e-mail: lex@icg.tugraz.at

[1] http://cancergenome.nih.gov/

[2] http://caleydo.org

[3] http://gdac.broadinstitute.org

[4] http://tcga.caleydo.org

[5] http://stratomex.caleydo.org

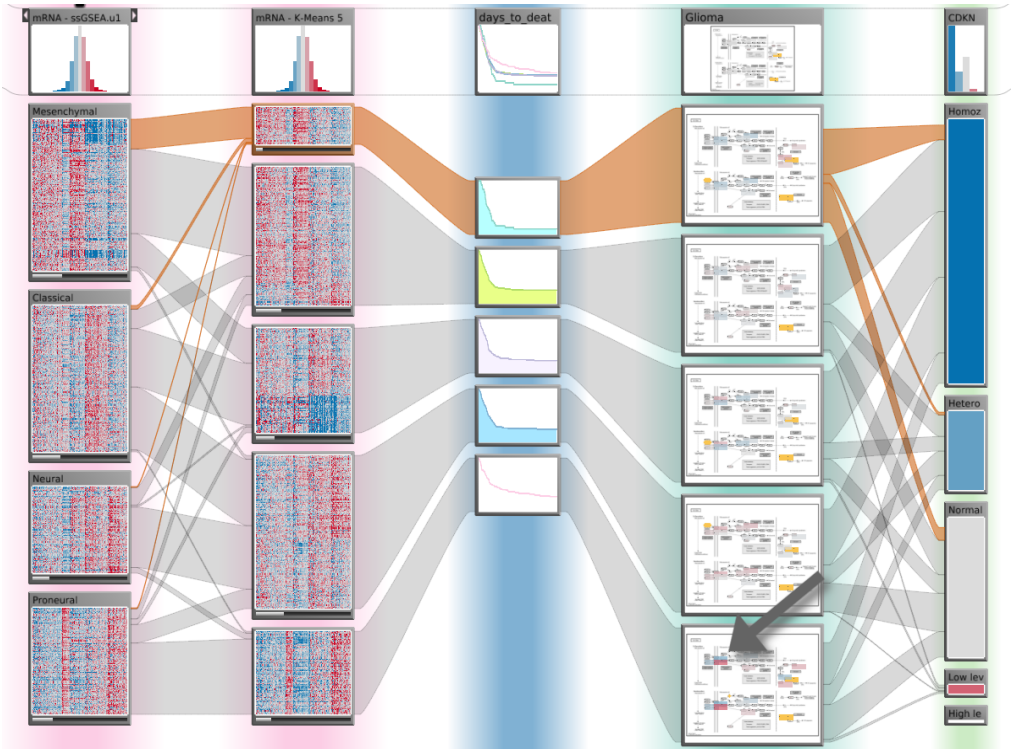

Figure 1: The StratomeX visualization technique. Each stratification of a dataset is presented as a column with a summary view at the top and bricks containing groups of patients below. Overlap between bricks is visualized as ribbons.

Dealing with many different datasets and stratifications is a challenging task by itself. To support users in assigning data to views and to provide an overview of the data shown, we developed the Data View Integrator (DVI), shown in Figure 2. The DVI enables users to specify which stratifications out of several alternatives they would like to explore (not shown in Figure 2) and to assign them to the view. Both, datasets and views, are represented as nodes that are connected if a dataset is shown in a view.

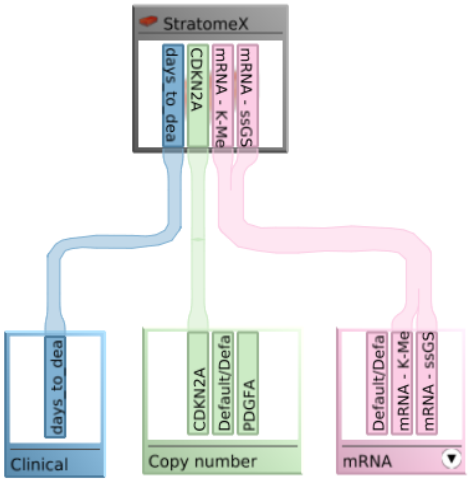

Figure 2: The Data-View Integrator used to assign datasets and stratifications to the StratomeX view.

3 Future Work

Currently, the choice of stratifications is purely user-driven with no support provided by the software. By integrating statistical and computational methods, we will provide a visualization-focused guidance system that lets users choose and preview the most relevant stratifications determined algorithmically. This is especially important when analyzing hundreds or thousands of stratifications, which is common when working with datasets where each gene can be interpreted as a stratification, such as mutation status and copy number data.

Furthermore, we will improve the ribbons used to visualize correlations and provide statistical data on the visualized effects. Ribbon width is a very salient feature but an imprecise measure of correlation. Especially for very large or very small groups, the ribbons can be deceiving. Color-coding the ribbons using the magnitude of deviation from the expected value of correlation can improve the accuracy, while statistical data can be presented for a selected ribbon.

4 Acknowledgements

This project is supported by the NCI (U24 CA143867), the FWF (P22902) and the state of Styria (GZ:A3-22.M-5/2012-21).

References

[1] A. Lex, M. Streit, H. Schulz, C. Partl, D. Schmalstieg, P. J. Park, and N. Gehlenborg. StratomeX: visual analysis of Large-Scale heterogeneous genomics data for cancer subtype characterization. *Computer Graphics Forum (EuroVis '12)*, 31(3):1175–1184, 2012.

[2] R. G. Verhaak, K. A. Hoadley, E. Purdom, V. Wang, and et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell*, 17(1):98–110, 2010.