SANJAY: Automatically Synthesizing Visualizations of Flow Cytometry Data using Decision Procedures

Faraz Hussain†, Zubir Husein†, Neslisah Torosdagli†
Narsingh Deo†, Sumanta Pattanaik†, Chung-Che (Jeff) Chang¶, Sumit Kumar Jha†

† Dept. of Electrical Engineering and Computer Science, University of Central Florida, Orlando, Florida.
Email: {fhussain, zhusein, neslisah, deo, sumant, jha}@eecs.ucf.edu

¶ Department of Pathology, Florida Hospital, Orlando, Florida.
Email: c.jeff.chang.md@flhosp.org

***Abstract*—Polychromatic flow cytometry is a widely used technique for gathering and analyzing cellular data. The data generated is high-dimensional, and therefore notoriously difficult to visualize by a human expert. The traditional method of plotting every pair of observables of the original high-dimensional data leads to a combinatorial explosion in the number of visualizations. The usual solution is to *project the data into a lower-dimensional space while approximately preserving key properties and relationships among data points*. The lower dimensional data can then be easily analyzed with the help of specialized data visualization software.**

We introduce SANJAY, a new method for automatically generating visualizations of high-dimensional flow cytometry datasets. Our technique uses symbolic decision procedures to *algorithmically synthesize 2D and 3D projections of the (original) high-dimensional data, with minimal distortion*.

We compare our approach to the popular MDS algorithm on a representative set of flow cytometry benchmarks, and show that the projections synthesized by SANJAY have distortions that are, on average, about two times smaller than those produced by MDS.

***Index Terms*—decision procedures; visualization; flow cytometry; automated synthesis.**

SUMMARY

Polychromatic flow cytometry is a revolutionary technique for analyzing biological samples by identifying multiple phenotypic properties of individual cells, including DNA content, RNA, and cell-surface proteins [1]. Besides applications in translational research, flow cytometry is routinely used to gain a deeper understanding into the fundamental biology of cellular processes. Unlike traditional methods that only compute statistical summaries of large populations of cells (e.g., the average concentration of a protein in a cell sample), flow cytometry allows measuring various phenotypic properties of *each cell* in a sample, thereby providing more detailed subject data [2].

This extensive data allows experts to identify even small groups of cells that are different from others (often signifying a clinical abnormality) whose presence could not have been detected by simply studying average phenotypic properties. There remain, however, two long-standing barriers to more widespread adoption of flow cytometry for diagnosing diseases:

- Cognitive processing studies have shown that our data analysis capacity is limited to about four dimensions. Therefore, flow cytometry techniques that often produce *high-dimensional data* (often up to 50 dimensions) cannot be easily visualized.
- In addition, flow cytometry produces *very large datasets*, typically with millions of data points per sample, which is well beyond our cognitive memory limits [3]. Hence, statistical summarization of this data that causes the loss of small, but potentially biologically significant details, has been considered a necessary evil. This often leads to an inability to detect rare events, risking significant harm to the subject.

TABLE I: Distortions produced by MDS and SANJAY on 10 randomly chosen data points from 5 flow cytometry benchmarks. We considered a total of 30 datasets and found that, on average, the maximum distortion produced by SANJAY was 2.14 times less than that produced by MDS.

Dataset ID	Max. distortion MDS	Max. distortion SANJAY	Ratio of maximum distortions (MDS/SANJAY)
6	2151.264	1500	1.434
8	3137.625	1500	2.092
21	2825.082	1500	1.883
26	3579.171	1000	3.579
28	2591.432	1500	1.728

To address these problems we have designed SANJAY, a new approach for synthesizing low dimensional visualizations of flow cytometry data. We highlight the main features of SANJAY:

- It is a new algorithmic technique for automatically synthesizing 2D and 3D visualizations of high-dimensional flow cytometry data, by using symbolic decision procedures [4] to search for low-dimensional projections of the original dataset.
- SANJAY avoids statistical summarization and stochastic search and provides a *deterministic method for complete visualization* of massive datasets with *minimal loss of information*.
- Table I shows results of our experiments comparing SANJAY to the multidimensional scaling (MDS) algorithm. It demonstrates that our technique produced projections with distortions that were on average 2.14 times less than those produced by MDS.

I. ACKNOWLEDGMENT

We acknowledge support from Leidos Biomedical Research Inc., the NSF Software and Hardware Foundations Project #1422257, NVIDIA Corporation, CyberRiskPartners LLC, the Royal Bank of Canada, the Oak Ridge National Laboratory, and the NSF Exploiting Parallelism and Scalabilty (XPS) project #1438989.

REFERENCES

[1] M.-C. Shih, S.-H. S. Huang, R. Donohue, C.-C. Chang, and Y. Zu, "Automatic b cell lymphoma detection using flow cytometry data," *BMC genomics*, vol. 14, no. Suppl 7, p. S1, 2013.
[2] A. L. Givan, *Flow cytometry: first principles*. John Wiley & Sons, 2013.
[3] A. Baddeley, "Working memory," *Science*, vol. 255, no. 5044, pp. 556–559, 1992.
[4] S. Jha and S. A. Seshia, "A theory of formal synthesis via inductive learning," *arXiv preprint arXiv:1505.03953*, 2015.

978-1-4673-9963-9/15/$31.00