

SamSPECTRAL: Efficient spectral clustering on flow cytometry data

Habil Zare^{1,2}, Parisa Shooshtari^{1,2}, Arvind Gupta³, and Ryan Brinkman^{1,4}

¹Terry Fox Laboratory, BC Cancer Agency, 675 W 10th Ave., ²Department of Computing Science, University of British Columbia, ³Faculty of Science, University of British Columbia, ⁴Department of Medical Genetics, University of British Columbia, Vancouver, BC, Canada

Spectral clustering is a non-parametric clustering method that avoids the problems of estimating probability distribution functions by using a heuristic based on graphs. Not only does it not require a priori assumptions on the size, shape or distribution of clusters, but it is not sensitive to outliers, noise or shape of clusters; it is adjustable so that biological knowledge can be utilized to adapt it for a specific problem or dataset; and there is mathematical evidence to guarantee its proper performance. However, spectral clustering cannot be directly applied to flow cytometry datasets due to time and memory limitations. To address this issue, we modified spectral clustering by adding an information preserving sampling procedure and applying a post-processing stage. We call this entire algorithm SamSPECTRAL. It has significant advantages in the proper identification of populations with non-elliptical shapes, low density populations close to dense ones, minor subpopulations of a major population and rare populations. In particular, the performance of SamSPECTRAL was tested in identifying a rare population in 34 samples from our stem cell dataset.

The rare population comprised between 0.1% to 2% of total events and SamSPECTRAL could distinguish it correctly in 27 (79%) samples.

An implementation of our algorithm as an R package is freely available through BioConductor.

Supported by the MITACS Network of Centres of Excellence, Canadian Cancer Society grant #700374 (Statistical and bioinformatics approaches to the classification of clinical lymphoma and leukemia data), and NIH/NIBIB grant EB008400 (The statistical and computational analysis of flow cytometry data).

Reference:

Zare, H. and Shooshtari, P. and Gupta, A. and Brinkman R.B: Data reduction for spectral clustering to analyze high throughput flow cytometry data. *BMC Bioinformatics* 2010,11:403.