GPU accelerated Bayesian mixture models for FCM analysis

Cliburn Chan1,5, Jacob Frelinger1,2, Adam Richards1, Lin Lin3, Ioanna Manolopoulou3, Andrew Cron3, Quanli Wang3, Janet Ottinger5, Kent Weinhold5, Marc Suchar4, Mike West3,

1Program in Computational Biology and Bioinformatics, 2Department of Biostatistics and Bioinformatics, 3Department of Statistical Science, 5Center for AIDS Research, Duke University, 4Departments of Biomathematics, Biostatistics and Human Genetics, UCLA

We describe the integration of Bayesian non-parametric mixture models, massively parallel computing on GPUs and software development in Python to provide an extensible toolkit for automated statistical analysis in high-dimensional flow cytometry (FCM). The use of standard Bayesian non-parametric Dirichlet process mixture models allows the flexible density estimation of the posterior distribution (MCMC) or modes (EM) of high-dimensional FCM data, and provides a coherent statistical framework for data analysis and interpretation (1,2,3). By exploiting the massively parallel nature of GPUs to achieve greater than 100 fold speed-ups over serial code, it is now realistic to perform large scale data analysis using these methods (4). To facilitate dissemination to the computational cytometry community, the statistical and computational foundations have been wrapped into fcm (5), a Python library that makes it simple to design and implement analysis and visualization pipelines for FCM (6). Used interactively, Python is ideal for rapid prototyping and algorithm development. In addition, Python has robust support for databases, networking, XML processing, GUIs and web development, facilitating the development of full stack applications (7). Together, statistical mixture models, GPU computing and Python “glue” provide a principled, efficient and extensible foundation for research into automated FCM analysis (8).

This research was supported by funding from the National Institutes of Health (RC1AI086032, UL1 RR024128-01 and P30-AI064518-05) and the National Science Foundation (DMS-0342172).

References

2. Selection sampling from large data sets for targeted inference in mixture modeling, I. Manolopoulou, C. Chan and M. West, Bayesian Analysis (2010), 05:429, DOI:10.1214/10-BA517