

GPU accelerated Bayesian mixture models for FCM analysis

Cliburn Chan^{1,5}, Jacob Frelinger^{1,2}, 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 *fc*m (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

1. Chan C, Feng F, Ottinger J, Foster D, West M and Kepler TB, Statistical mixture modeling for cell subtype identification in flow cytometry, *Cytometry A*, (2008), 73A:693-701 [PMID: 18496851]
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
3. Optimization of a highly standardized CFSE flow cytometry panel and gating strategy design using Discriminative Information Measure Evaluation (DIME), C. Chan, L. Lin, J. Frelinger, V. Hebert, D. Gagnon, C. Landry, R-P. Sékaly, J Enzor, J. Ottinger, KJ Weinhold, M. Jaimes, M. West, *submitted*, *Cytometry A*.
4. Understanding GPU programming for statistical computation: Studies in massively parallel massive mixtures (2010), M.A. Suchard, Q. Wang, C. Chan, J Frelinger, A. Cron & M. West, *Journal of Computational and Graphical Statistics*, DOI: 10.1198/jcgs.2010.10016
5. <http://code.google.com/p/py-fcm>
6. <http://code.google.com/p/cytostream>
7. Frelinger, J, Kepler TB and Chan C, *Flow: Statistics, visualization and informatics for flow cytometry*, *Source Code for Biology and Medicine*, (2008), 3:10.
8. Modeling flow cytometry data for cancer vaccine immune monitoring, J. Frelinger, J. Ottinger, C. Gouttefangeas, C. Chan, *Cancer Immunology Immunotherapy* (2010): 59:1435, DOI 10.1007/s00262-010-0883-4