FlowCAP SUMMIT 2011

U.S. NIH Campus

Natcher Auditorium
Bethesda, MD.
22 - 23 September 2011

Flow Cytometry:

Critical Assessment of Population Identification Methods (FlowCAP)

To advance the development of computational methods for the identification of cell populations of interest in flow cytometry data

http://flowcap.flowsite.org

FlowCAP Organizing Committee

- **Ryan Brinkman,** British Columbia Cancer Agency
- **Raphael Gottardo,** Fred Hutchison Cancer Research Center
- **Tim Mosmann,** University of Rochester Medical Center
- **Richard Scheuermann,** University of Texas Southwestern Medical Center

Sponsored by the National Institute of Allergy and Infectious Diseases

FlowCAP-II Summit 2011

Sept 22-23, 2011
NIH Campus, Bethesda, Maryland
Natcher Auditorium (Balcony A & B).

FlowCAP-II Summit 2011 will assemble the key stakeholders in field to present and discuss the results from the FlowCAP-II competition, and to discuss how automated methods are being used to address biological questions. The FlowCAP project (http://flowcap.flowsite.org/) was established to provide a mechanism to compare and contrast the utility of these novel computational approaches as applied to a common set of reference datasets.

The objectives of this two-day workshop are to:

- Allow developers of novel computational algorithms to describe their methods and present the results from their initial analysis of the seven FlowCAP -2 Challenges
- Allow key representatives of the flow cytometry user community to review and critique the results;
- Discuss how the new computational methods could be improved to meet the needs of the user community;
- Discuss opportunities for additional analytical solutions to solve problems for other data related to flow cytometry.

Day 1:

8:30 Meet and greet (light refreshments provided)
9:00 Call to order - (Ryan Brinkman)
9:10 Welcome address (Daniel Rotrosen, DAIT/NIAID)

Session 1 – Informatics Challenges in Flow Cytometry

This session will explore some of the current challenges faced by the flow cytometry community in analyzing flow cytometry data, including rare population detection, assessment of cellular activation responses, evaluation of disease states, therapeutic responses and longitudinal analysis of temporal changes.

9:30 Informatics Challenges in High Throughput Flow Cytometry -
I - FITMaN/HIPC Standardization of Immunophenotyping (McCoy, J. Philip (NIH/NHLBI))
II - TBA (Deborah Phippard - ITN)

Session 2 FlowCAP-II Challenges & Results

Talks by data providers discussing the datasets, their evaluation criteria and results (FlowCAP-CC). Each challenge and its results will be discussed in turn, with added discussion at end. We hope the participants will discuss features they used to discriminate between groups to perhaps help us understand potential batch effect issues.

10:30 Challenge 1 - HIV-Exposed-Uninfected (HEU) versus Un-exposed (UE)

The goal of this challenge was to find cell populations that can be used to discriminate between HEU (n = 20) and UE (n = 24) infants. Blood samples were taken at 6 months after birth and were left unstimulated (for control) or stimulated with 6 Toll-like receptor molecules.

Dataset description and analysis results: Nima Aghaeepour

11:10 Break

11:20 Challenge 2: Acute Myeloid Leukaemia

The goal of this challenge was to find cell populations that can be used to discriminate between AML positive (n = 43) and healthy donor (n = 316) patients. Peripheral blood or bone marrow aspirate samples were collected over a 1 year period using 8 tubes with different marker combinations.

Dataset description and challenge: Wade Rogers
Analysis 1 (FlowCAP) - Nima Aghaeepour
Analysis 2 (DREAM Initiative) - Raquel Norel

12:30 Lunch (on your own)

1:30 Challenge 3 - Intracellular Cytokine Staining of Post-HIV Vaccine Antigen Stimulated T-cells

3A Identification of Antigen Stimulation Group

The goal of this challenge was to correctly label the antigen stimulation group of post-HIV vaccine T-cells. The data set contains samples from 48 individuals (column ”pub-id” in the metadata). Each individual received an experimental HIV vaccine. Samples were collected approximately 10 months later and T-cells challenged with two antigens (ENV-1-PTEG and GAG-1-PTEG, column ”antigen” in the meta- data). The response of CD4+ and CD8+ T-cells was measured by flow cytometry for each of these groups. The cells were found to respond differently to the two antigen stimulations.

3B Identification of Responders and Non-Responders in Intracellular Cytokine Staining of Post-HIV Vaccine Antigen Stimulated T-cells

The dataset is identical to challenge 3A. However, the goal is to use automated methods to identify responders and non-responders to the antigen stimulations, as defined by manual analysis. Important to this challenge are additional positive controls for each sample, in which positive staining cells can be detected for each cytokine following stimulation with an antigen that is known to produce a cytokine response. Also provided is a matched negative control, where the sample was not challenged with antigen. This negative control is used as a patient-matched baseline. The goal of this challenge was to identify each sample as either a responder or non-responder to the antigen stimulation.

Dataset and analysis: Raphael Gottardo

2:15 Challenge 4: Multiple Sclerosis Treatment

The goals of this challenge were to i) recapitulate the results of expert manual gating using automated algorithms and ii) to find cell populations that correlate with treatment arms between healthy controls, multiple sclerosis patients treated with Copaxone and multiple sclerosis patients treated with interferon beta. Peripheral blood mononuclear cells were isolated from blood specimens taken before (t0) and 4 hr, 12 hr, 24 hr and 72 hr after the initiation of treatment. Samples were stained with an 8-marker T cell reagent panel. Data was collected on a viable lymphocyte gate based on FSC/SSC, with 100,000 gated events collected/sample.

Dataset and analysis: Richard Scheuermann

2:35 Challenge 5: Omalizumab and Rush Immunotherapy

The goal of this challenge was to find cell populations that correlate with treatment arms between placebo, omalizumab alone, rush immunotherapy alone, and the combination of omalizumab and rush immunotherapy (see Casale TB, et al., J Allergy Clin Immunol 2006;117:134-40 and Casale TB, et al., J Allergy Clin Immunol 2007;120:688-95 for experiment design details and outcome results). Peripheral blood mononuclear cells were isolated from blood specimens taken before and after treatment at various timepoints. Samples were stained with 7 different 4-marker reagent panels.

Dataset and analysis: Richard Scheuermann

3:00 Coffee &snacks

3:15 Challenge 6: Stimulation of Influenza-specific cytokine-expressing T cells
The goal of this challenge was to identify rare populations (e.g. 25 cells per million) that are induced by influenza antigen stimulation of PBMC. Human PBMC were incubated in triplicate with pools of influenza peptides (G1 to G6), DMSO control medium, tetanus peptides and Staph Enterotoxin B (SEB). After 10 hours, cells were stained with a panel of antibodies specific for surface antigens and cytokines, then fixed and analyzed. Algorithms were to identify the rare cytokine-secreting activated T cell populations that are induced by influenza and tetanus antigen stimulation.

Dataset and analysis: Tim Mosmann

3:35 Challenge 7: Dilution of activated cells in control populations.
The goal of this challenge was to determine the sensitivity with which your algorithm can identify SEB activated T cell populations. Human PBMC were incubated with control medium, or SEB. After 10 hours, cells were stained with a panel of antibodies specific for surface antigens and cytokines, fixed and analyzed. Stimulated and unstimulated cell populations were then combined electronically in different ratios. Participants identified the activated T cell populations that are induced by SEB stimulation, analyzing each sample independently.

Session 3(1) – New Algorithms for FCM Analysis and their application to FlowCAP-II

This session will include presentations from those groups who have developed novel methods for FCM data processing and analysis not presented in FlowCAP-I and groups that have developed significant modifications to previous methods that have resulted in substantial improvements. Each presentation will be 15 minutes plus time for short questions, with a longer Q&A session at the end.

4:00 Phenotyping (flowType) and Robust Feature Selection (FeaLect) for Flow Cytometry Data - Nima Aghaeepour

4:20 FIND: A new software tool and development platform for enhanced multicolor flow analysis - Shareef M. Dabdoub

4:40 Analysing Flow Cytometry ICS Data Using a BioConductor Pipeline - Mike Jiang

5:00 Wrap up - Scientific Session Day 1

6:30 Dinner (not subsidized) - Meet in DoubleTree Inn (Bethesda) Lobby to head out for group dinner

Day 2

8:30 Meet and greet (light refreshments provided)
8:55 Call to order - (Ryan Brinkman)

Session 3(2) – New Algorithms for FCM Analysis and their application to FlowCAP-II

9:00 flowBin: A Complete Pipeline for Feature Extraction and Classification of Multi-tube Flow Cytometry Data - Kieran O'Neill

9:20 Automated identification of cell population changes using cross-sample comparison with FLOCK - Yu Qian

9:40 Extracting a cellular hierarchy from high-dimensional single-cell data - Peng Qiu

10:00 Automated Identification of Differential Signatures in Cellular Populations - Robert Bruggner

10:20 Applying K-means (or flowPeaks) and Support vector machines to the sample classification problem using the flow cytometry data - Yongchao Ge

10:40 flowMatch:A tool to create feature--□preserving templates by population matching - Ariful Azad

11:00 Break

Session 4 – Future of Flow Informatics

11:10 FlowCAP-1 Results & Discussion (Richard Scheuerman)

12:00 FlowCAP Future: An open session dedicated to the review of the state-of-the-art, planning of FlowCAP-III, etc.

1:00 Meeting ends

FlowCAP is supported by grants from the NIH/NIAID and NIB (EB008400)

Phenotyping (flowType) and Robust Feature Selection (FeaLect) for Flow Cytometry Data

Nima Aghaeepour, Habil Zare, and Ryan Brinkman

We have prepared two sets of results for challenges 1, 2, and 3(a): flowType: This pipeline uses the flowMeans algorithm for cell population identification [1]. Briefly, flowMeans identifies a large number of clusters in the data and merges them based on the Mahalanobis distance between them until the desired number of clusters is reached. For each of the markers in a given dataset, flowMeans was used to identify a partition that divides the cells into a positive and a negative population. This is based on the assumption that the cells either express a given marker or not (i.e., there are two distinct cell populations). For N markers this results in 2^N phenotypes. To allow exclusion of markers from population identification (which later enabled us to identify the "important" markers), each marker was also allowed to be "neutral" (i.e., that marker was excluded from the clustering); thus, for any single subset, each marker could be negative, positive, or neutral (ignored). This increases the number of cell populations to 3^N. These phenotypes are then evaluated using ROC analysis, t-test with Bonferroni correction, and bootstrapping-based sensitivity analysis. These tests result in a hit list of "statistically significant" features (with the exception of the HEUvsUE challenge were non of the phenotypes remained significant after p-value correction). These phenotypes are then divided to several groups, based on the Pearson correlation between them and the markers required for defining the phenotypes in each group are identified. The final representative phenotype with maximum area under the ROC on the training set is used to label the samples in the test set. A more detailed description of the pipeline is available elsewhere[2]. The flowType R package is available through Bioconductor.

FeaLect: This pipeline builds a multivariate model using the phenotypes measures by flowType. A bagging technique is used to score the features for the linear classifier. Robustness of the model is measured by both cross-validation and holdout-validation on the training-set. The model is then used to label the samples in the test-set. A more detailed description of FeaLect is available elsewhere[3,4]. The FeaLect R package is available through CRAN.

[1] Nima Aghaeepour, Radina Nikolic, Holger Hoos, and Ryan Brinkman. rapid cell population identification in flow cytometry data. Cytometry Part A, 79(1):6–13, 2011.

[2] Nima Aghaeepour, Pratip K. Chattopadhyay, Anuradha Ganesan, Kieran O'Neill, Tess M. Brodie, Habil Zare, John R. Mascola, Adrin Jalali, Armstrong Murira, Celsa A. Spina, Jamie Scott, Holger H. Hoos, Nelson Michael, Ryan R. Brinkman, and Mario Roederer. Early immunologic correlates of hiv protection can be identified from computational analysis of complex multivariate t-cell flow cytometry assays. Status: Manuscript circulated to co-authors.

[3] Habil Zare, Ali Bashashati, Josef Connors, Nima Aghaeepour, Arvind Gupta, Randy Gascoyne, Ryan Brinkman, and Andrew Weng. Automated analysis of multidimensional

flow cytometry data improves diagnostic accuracy between mantle cell lymphoma and small lymphocytic lymphoma, American Journal of Clinical Pathology, 2011.

[4] Habil Zare, Gholamreza Haffari, Andrew Weng, Randy Gascoyne, Arvind Gupta, Ryan Brinkman, Statistical analysis of overfitting features. Status: manuscript in preparation.

FIND: A new software tool and development platform for enhanced multicolor flow analysis

Shareef M. Dabdoub[1,2], William C. Ray[1,2], Sheryl S. Justice[2]
1 The Ohio State University Biophysics Program
2 The Research Institute at Nationwide Children's Hospital

Flow Cytometry is a process by which cells, and other microscopic particles, can be identified, counted, and sorted mechanically through the use of hydrodynamic pressure and laser-activated fluorescence labeling. As immunostained cells pass individually through the flow chamber of the instrument, laser pulses cause fluorescence emissions that are recorded digitally for later analysis as multidimensional vectors.

Most widely adopted analysis software limits users to manual separation of events based on visualization of two or three dimensions. While this may be adequate for experiments using four or fewer colors, advances have lead to laser flow cytometers capable of recording 24 different colors. In addition, mass-spectrometry based machines capable of recording at least 100 separate channels are being developed. Analysis of such high-dimensional data by visual exploration alone can be error-prone and susceptible to unnecessary bias. The last few decades have seen a good deal of research activity into creating new tools and adapting existing algorithms for automated group classification of multi-dimensional data. However, the majority of this research has not been made available to users through widely adopted analysis software packages and, as such, are not in common use.

Here we present a new software application for analysis of multi-color flow cytometry data. The main goals of this effort are to provide a user-friendly tool for automated gating (classification) of multi-color data as well as a platform for development and dissemination of new analysis and visualization tools. With this software, users can easily load single or multiple data sets, perform automated event classification, and compare results between experiments. We also make available a simple plugin system that allows researchers to implement and share their data analysis and classification algorithms. This will greatly reduce development time as well as provide a common platform for distribution of new techniques to flow cytometry users around the world.

Analysing Flow Cytometry ICS Data Using a BioConductor Pipeline.

Mike Jiang, Greg Finak,_Raphael Gottardo

We applied existing BioConductor packages for flow cytometry data analysis to gate and analyze an intracellular cytokine staining assay of T-cells from HIV-vaccinated individuals, (challenge three of flowCAP II). The goals of the challenge were two-fold: a) predict the antigen stimulation group of each sample, b) identify whether the CD4 and CD8 T-cell subpopulations for each subject were responders or non-responders to each stimulation. Using the flowStats and flowCore packages, we applied a knowledge-driven gating approach and a new sequential normalization strategy by alternately gating and normalizing subpopulations to identify cytokine-positive, CD4 and CD8 T-cells in each sample. Normalization allowed us to use a common set of gates for each subject across stimulations, whereas cytokines were gated in a sample-specific manner to account for variation in the peak width of the cytokine-negative population. For the classification challenge, the negative control was used to compute the background adjusted proportion of cytokine positive cells for each subject, and a decision tree classifier was trained using the marginal cytokine features, under 10-fold cross validation. For the responder / non-responder calls, we fit a Beta-Binomial model to the raw cytokine-positive counts of stimulated and unstimulated samples, then estimated the posterior probability that the proportion of cytokine positive cells for the stimulated sample is larger than the proportion of cytokine positive cells for the unstimulated sample. The training data were used to calibrate these probabilities in a decision tree classifier.

flowBin: A Complete Pipeline for Feature Extraction and Classification of Multi-tube Flow Cytometry Data

Kieran O'Neill[1;2] and Ryan Brinkman[1;3]
1 Terry Fox Laboratory, BC Cancer Agency, Vancouver, BC, Canada
2 Bioinformatics Program, University of British Columbia, Vancouver, BC, Canada
3 Department of Medical Genetics, University of British Columbia, Vancouver, BC, Canada

Multiplexing flow cytometry experiments across tubes containing different combinations of markers is a common solution to the problem of measuring the expression of more markers than a particular flow cytometer can handle in one run. Data from such experiments produces unique challenges, particularly for cross centre and retrospective analysis, since markers are often assayed in different combinations.

One solution is Pedreira et al's method of combining tubes via K-nearest neighbours (KNN) applied across parameters shared by all tubes, to create a very high-dimensional single file.[2] However, this method implies imputation, and can lead to spurious artificial populations.[1] To solve this problem, we instead binned data using overfitted K-means clustering in the shared parameters, and mapped these bins across tubes using KNN. We then extracted summary statistics (e.g., median fluorescence intensity) for each bin in terms of each parameter. Although this approach involved some data reduction, it avoided imputation.

Binning within patients raised the problem of linking features across patients for classification. To solve this, we took each bin from each sample as a separate training instance, labelled with the sample label, and then trained a support vector machine (SVM) classifier. For class prediction, we took the majority vote of the predicted labels for a given sample's bins. Classification with parameter optimization and cross-validation was implemented using the ksvm R package, but could in theory be made to work with any modern classification method.

References

[1] G. Lee, W. Finn, and C. Scott. Statistical file matching of flow cytometry data. Journal of Biomedical Informatics, 44(4):663–676, 2011.
[2] C E Pedreira, E S Costa, S Barrena, Q Lecrevisse, J Almeida, J J M van Dongen, and A Orfao. Generation of flow cytometry data files with a potentially infinite number of dimensions. Cytometry Part A, 73(9):834–846, 2008.

Acknowledgements

This work was funded by NIH/NBIB grant EB008400.

Automated identification of cell population changes using cross-sample comparison with FLOCK

Yu Qian[1, 2], John Campbell[3], Yue Liu[3], Megan Kong[2], and Richard H. Scheuermann*[1, 2]
[1]Division of Biomedical Informatics, [2]Department of Pathology, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA
[3]Health Information Systems, Northrop Grumman, Inc., Rockville, MD 20850, USA

Multi-dimensional flow cytometry (FCM) brings in challenges not only in identifying individual cell populations but also with population mapping and interpretation across different samples and treatment groups. FLOCK (Flow Clustering without K) is an automated software system we have developed for the identification of cell populations from multi-dimensional FCM data [Qian 2010], which has been made publicly available at the Immunology Database and Analysis Portal - ImmPort (http://www.immport.org). This presentation will focus on the design of a general cross-sample comparison method and how this method can be used with FLOCK to map populations and detect their changes across different samples. Based on population characteristics identified by FLOCK, we successfully model the similarity between populations across different samples using F-measure. Then a novel relative distance model is proposed to capture the position pattern and relative order of cell populations, so that their slight shifts between different samples could not affect the mapping. A meta-clustering of cell populations is performed based on their relative distances to identify whether a population in one sample can be found in another sample or it is a new population. When the number of samples is large, the meta-clustering is done incrementally and the problem is converted to a constrained maximum bipartite matching problem which can be efficiently solved by existing graph algorithms. The proposed method is general and can be combined with other existing automated population identification methods to map populations across samples. This presentation will also briefly discuss a semi-supervised approach to encode user knowledge to assist population interpretation and identify rare populations. Results we generated based on FlowCAP2 datasets will be demonstrated and discussed.

Supported by U.S. NIAID Bioinformatics Integration Support Contract (N01-AI40076)

References:

Qian, Y., Wei, C., Eun-Hyung Lee, F., Campbell, J., Halliley, J., Lee, J. A., Cai, J., Kong, Y. M., Sadat, E., Thomson, E., Dunn, P., Seegmiller, A. C., Karandikar, N. J., Tipton, C. M., Mosmann, T., Sanz, I. and Scheuermann, R. H. (2010), Elucidation of seventeen human peripheral blood B-cell subsets and quantification of the tetanus response using a density-based method for the automated identification of cell populations in multidimensional flow cytometry data. Cytometry Part B: Clinical Cytometry, 78B: S69–S82. doi: 10.1002/cyto.b.20554

Extracting a cellular hierarchy from high-dimensional single-cell data

Peng Qiu[1], Erin F. Simonds[2], Sean C. Bendall[2], Kenneth D. Gibbs Jr.[2], Michael D. Linderman[3], Karen Sachs[2], Garry P. Nolan[2], Sylvia K. Plevritis[4]

[1]Department of Bioinformatics and Computational Biology, University of Texas M.D. Anderson Cancer Center; [2]Department of Microbiology and Immunology, [3]Computer Systems Laboratory, [4]Department of Radiology, Stanford University

Flow cytometry and the next-generation mass cytometry technologies capture the heterogeneity of biological systems by providing multiparametric measurements of single cells. Even as cytometry technology is rapidly advancing, methods for analyzing this complex data lag behind. Traditional flow cytometry analysis is often a subjective and labor-intensive process that requires users' deep understanding of the cellular phenotypes underlying the data. Furthermore, the advent of mass cytometry is quickly increasing the dimensionality of the data, making the traditional analysis approaches a critical bottleneck. To objectively explore the richness of such high-dimensional single-cell data, new computational methods are needed.

We present a novel analytical approach, Spanning-tree Progression Analysis of Density-normalized Events (SPADE), to explore high-dimensional cytometry data in a robust and unsupervised manner, and reveal a likely underlying cellular hierarchy. Briefly, SPADE views a cytometry dataset as a high-dimensional point cloud of cells, and uses topological methods to reveal the geometry of the cloud. We applied SPADE to an 8-parameter flow cytometry dataset of normal mouse bone marrow, and a 31-parameter mass cytometry dataset of normal human bone marrow. In both datasets, SPADE detected a hierarchy which recapitulates well-described patterns of hematopoiesis.

SPADE is a versatile tool for cytometric data analysis, facilitating identification of cellular hierarchy, identification of rare cell types, and automated comparison of functional markers that enables new biology discoveries.

FlowCAP Methods

Maria Chikina

The method is based on computing 2D histograms of all possible pairs of stains (forward and side scatter were ignored.). Each feature in the dataset represents the proportion of cells that fall into a specific bin on a 2 dimensional scatter plot. The bin width was set to 0.1 and bins spanned the full range of data in each dimension.

The resulting features were used for classification with SVMstruct [1] using precision-recall break-even point as the loss function. The tradeoff constant was empirically optimized over the range of 10^{-5} to 10^{3}. The number of cross-validation trials was 50 for the AML challenge and 22 for the HIV challenge (in this case this represents leave-one-out cross-validation). Classifications of unlabeled results were obtained from all cross-validation models and the results were averaged to produce final classification values. As SVM produces continuous classification values a threshold was chosen to achieve best classification accuracy on held out examples.

References

[1] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support vector machine learning for interdependent and structured output spaces.
In *Intl Conf. on Machine Learning*, 2004.

Automated Identification of Differential Signatures in Cellular Populations

Authors:

Robert Bruggner (presenting), Rachel Finck, Robin Jia, Noah Zimmerman, Michael Linderman, David Dill, Garry Nolan

Abstract:

Nuanced behavior of phenotypically distinct cellular populations plays a critical role in both immune response to and development of cellular diseases (i.e. cancer). Furthermore, recent work has highlighted the utility of subpopulation profiling in patient prĺognosis. Accordingly, the ability to discern and identify condition-relevant populations can potentially play a critical role in disease diagnosis and treatment. To facilitate analysis of complex mixtures of cells, instrumentation technologies such as flow cytometry have emerged that enable high-throughput, simultaneous measurement of intra- and extra-cellular molecules within a single cell. However, the high-dimensionality of such data coupled with normal biological variation make comprehensive manual identification of phenotype-relevant subpopulations unfeasible.

Recent work on cell population-finding algorithms has enabled automated identification of clusters of cells in multidimensional space. We utilize these algorithms in conjunction with supervised learning models and present here a method for automated discovery of differential cell populations. Given multiple samples from patients belonging to two or more phenotypic classes, we automatically identify sub-populations of cells within each sample, extract meta-features describing each population, and train a supervised classifier for identification of a sample class. It follows that the stratifying features of a successful classifier correspond to class-differentiating populations. We demonstrate our method to by identifying differential populations in the blood of HIV patients challenged with two different antigens.

As technologies such as mass cytometry continue to increase the number of simultaneous measurements per cell, automated approaches such as the one described here will play a crucial role in the unbiased discovery and identification of populations involved in both disease mechanism and response.

Applying K-means (or flowPeaks) and Support vector machines to the sample classification problem using the flow cytometry data

Yongchao Ge[1] and Stuart Sealfon[1]

1. Department of Neurology, Mount Sinai School of Medicine, New York, NY, 10029

In the sample classification problem when each sample is assayed by high-dimensional flow cytometry data, there are generally two stages involved. The first is the data summarization to reduce the data complexity, and the second is applying a traditional machine learning algorithm to build a good classifier with a low cross-validation error rate.

For the data complexity reduction stage, the K-means algorithm becomes our first choice. K-means algorithm was proposed more than 50 years ago and is still one of the most widely used algorithms for clustering [Jain 2010]. K-means has been served very well in the data compression technique such as vector quantization widely used in signal processing. In the application of the sample classification for the flowCAP II challenges, we used K-means with a large K to prototype the data, the determination of the exact cluster number K is not as crucial as in the traditional clustering so we fix it to be 300. The seeds of the K-means algorithm are generated by the kmeans++ [Arthur and Vassilvitskii 2007], and the algorithm is implemented by using k-d tree [Kanungo and Mount 2002] to speed up computation.

For the machine learning stage, we used the support vector machine (SVM) due to its ability to handle high dimensional data, in which many variables can be correlated. We have used the SMO algorithm [Platt 1998] implemented at WEKA [Hall, Frank, Holmes and et al, 2009] to train the SVM machine.

The combination of Kmeans and SVM works quite well for Challenge 3A with between zero and two cross-validate errors out of 54. For Challenge 2, we found that our own developed algorithm flowPeaks (manuscript in progress), which combines the clusters of the K-means into a larger cluster based on the density peaks, gives a slightly better cross-validation accuracy than using the plain K-means algorithm. None of the two approaches can give reasonable cross-validation accuracy for Challenge 1. In the end, we submitted Kmeans and SVM for Challenge 3A, and flowPeaks and SVM for Challenge 2.

Research supported by NIH/NIAID HHSN272201000054C

References

1. Arthur D and Vassilvitskii S. k-means++: the advantages of careful seeding. in *Proceedings of the eighteenth annual ACM-SIAM symposium on discrete algorithms*, 1027–1035, 2007
2. Hall M, Frank E, Holmes G and et al. The WEKA Data Mining Software: An Update; *SIGKDD Explorations* **11**(1), 2009
3. Jain AK. Data clustering: 50 years beyond K-means, *Pattern Recognition Letters* **31**:615-666, 2010
4. Kanungo T, Mount D, Netanyahu N and et al. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24**(7), 881–892, 200
5. Platt J. Fast training of support vector machines using sequential minimal optimization. In *Advances in Kernel Methods - Support Vector Learning*, Schoelkopf, Burges, and Smola eds., MIT Press, 1998

flowMatch: A tool to create feature-preserving templates by population matching

Ariful Azad, Saumyadipta Pyne and Alex Pothen
Computer Science, Purdue University; Medical Oncology, Dana Farber Cancer Center, Harvard Medical School

Cell populations in a multiparametric flow cytometric sample can be characterized with finite mixture models of multivariate probability distributions. For automated registration and comparison of cell populations across samples, new algorithms are necessary to match these distributions in high-dimensional marker space. With increasing number of markers and large cohort sizes, such approaches must be both efficient and systematic. Towards this, we present flowMatch, a robust solution to the matching problem based on templates that summarize all samples from a given class, thus allowing us to study the population-level changes across different conditions and time points.

We designed flowMatch as a hierarchical template-construction algorithm where each step uses a Generalized Edge Cover for optimally matching populations across samples, which are then used to create meta-clusters for a new class template. We applied the algorithm on samples obtained across different time points from healthy controls and multiple sclerosis patients treated with Copaxone and interferon beta. At each time point we created distinct templates for each of the three classes of samples and matched them across time points to follow the progression of populations defined by the corresponding meta-clusters. We detected certain correlated populations across treatment arms by comparing the progression paths (converging, diverging or parallel) of the meta-clusters between the two treatment arms. To demonstrate that meta-clusters in a template preserve the common features of populations from initial samples, we applied flowMatch on 30 anti-CD3-antibody stimulated blood samples, formed templates from samples before and after stimulation, and identified a clear shift at the meta-cluster level after stimulation.