flowCAP: Participant Guide

Ryan Brinkman, Raphael Gottardo, Richard H. Scheuermann,
Jill Schoenfeld, Nima Aghaeepour, Alireza Khodabakhshi

May 18, 2010

Introduction

This document provides detailed instructions for participation in the flowCAP project. For more information please visit: `http://flowcap.flowsite.org/` or join the Google Group and mailing list at `http://groups.google.com/group/flowcap`.

Challenge Structure

flowCAP will compare automated gating algorithms against manual analysis based on a number of properties of cell populations. These properties will include, but are not limited to, misclassification rate[2], Fmeasure, homogeneity, completeness, purity, and the number of cell populations[3][1]. Participants may choose to participate in one or more of the following subchallenges:

Challenge 1: Automated Algorithms

The main goal is to compare automated gating algorithms for exploratory analysis on a wide range of FCM samples. **Software used in this challenge should not have any free parameters (if you have a free parameters it must be set to a single value for all of the datasets).**Please make sure that all of the provided dimensions are included in the study (you may not exclude FSC/SSC for some of the datasets. For this challenge, participants will use software that, given only a FCS file and no other information, produces a population membership label (or set of labels with likelihoods) for each event. The participants will submit the output of their algorithm (i.e., predicted labels, run time, and specifications of their system) on the provided datasets. The organizers will evaluate the submitted results against a benchmark of the same FCS files analyzed manually.

Challenge 2: Tuned Algorithms (in the Absence of Example Human-Provided Gates)

The main goal is to compare automated gating algorithms for exploratory analysis on a wide range of FCM samples. **Software used in this challenge may have free parameters that can be manually adjusted before running (i.e., you can submit an algorithm with some free parameters for each dataset).** For this challenge, participants will use software that, given only a FCS file and no other information, produces a population membership label (or set of labels with likelihoods) for each event. The participants will submit the output of their algorithm (i.e., predicted labels, parameter values, and specifications of their system) on the provided datasets. The organizers will evaluate the submitted results against a benchmark of the same FCS files analyzed manually.

Challenge 3: Assignment of Cells to Populations with Predefined Number of Populations

The main goal is to compare the ability of the algorithms to assign correct labels to cells when the number of expected populations is known. **In this challenge the number of populations within each FCS file will be provided to participants however, similar to challenge one, free parameters must be set to a single value and all of the dimensions must be included.** For this challenge, the participants will use software that, given a folder of FCS files and the number of population in each sample, produces a folder of population membership labels (or set of labels with likelihoods) for each event. The participants will also submit the output of their algorithm (i.e., predicted labels, run time, and specification of their system) on the provided datasets. The organizers will evaluate the submitted results against a benchmark of the same FCS files analyzed manually.

Challenge 4: Supervised Clustering Approaches Trained using Human-Provided Gates

In this challenge a few files with manual gates (i.e., membership labels) will be provided to the participants for tuning their algorithms for each dataset. The tuned software can then be run on the remaining data files; the results will be evaluated using these remaining data files. Participants will submit the output of their algorithm (i.e., predicted labels, run time for training set and overall, and specification of their system) on the provided datasets. The organizers will evaluate the submitted results against a benchmark of the same FCS files analyzed manually.

Handling Outliers

Every cell that was not included in the manual analysis by the human expert (due to noise or lack of biological interest) will be considered as an *outlier* for

the purpose of this challenge. Algorithms will not be penalized for assigning an incorrect label to cells that are marked as outliers by these criteria. However, predicting biologically relevant (i.e., non-outlier) cells as outliers (with not assigning that cell to a cluster) will penalize the algorithm. Therefore, our advice is that the algorithms should analyze all the cells in every sample and assign a label to as many as possible.

Time line

- Release of materials for challenges 1 and 2: *01 MAR 2010*
- Submission deadline for challenges 1 and 2: *30 JUN 2010*
- Release of materials for challenge 3: *30 JUN 2010*
- Submission deadline for challenge 3: *21 JUL 2010*
- Release of materials for challenge 4: *21 JUL 2010*
- Submission deadline for challenge 4: *15 AUG 2010*
- Public release of the results: *15 SEP 2010*
- NIH/NIAID-sponsored flowCAP summit at NIH campus: *21-22 SEP 2010*

Data Standards

Flow Cytometry Cells

The flow cytometry events are available in the Flow Cytometry Standard (**FCS**) format. For languages that do not have the libraries required for parsing this format the Comma-Separated Values (**CSV**)[4] can be used.

Cell Populations (Clusters)

Using manual analysis by FCM experts, a label (a *0 to N* integer value where N is the number of populations and 0 indicates an outlier cell) is assigned to each cell. This array of the labels can be found in the CSV files in the *Labels* directory of each dataset. Your software must generate a similar CSV file for evaluation (please closely follow the CSV[4] description). For example:

A hard-cluster must have a cluster label per line were the $i'th$ line shows the label of the $i'th$ cell:
1
1
2
3

A Soft cluster must have N probabilities per line were the $i'th$ line shows the membership probability of the $i'th$ cell. Sum of these numbers must be equal to one except for outliers for which all of the values must be zero:
0.5, 0.4, 0.1
0.3, 0.2, 0.5
0.4, 0.1, 0.6
0.2, 0.6, 0.2

Please note that in these labels ordering is irrelevant (i.e., $1, 1, 2, 3$ is equivalent to $2, 2, 1, 3$).

How to Participate

Please take the following steps to participate in a challenge:

- Join email discussion list at `http://groups.google.com/group/flowcap`
- Send an email to flowcap@flowsite.org to register for the challenges that you are going to participate and to recieve instructions for downloading the data.
- Send in the required materials by due date to flowcap@flowsite.org

Materials

The following materials should be submitted for each challenge separately:

- The output of the algorithm (CSV files containing membership labels) for **all of the datasets** provided for that challenge.
- Run time of the algorithm for each dataset and specifications of their machine.
- A executable version of their program and a shell script to reproduce their results (if the software is publicly available, provide a link to the specific version).
- A short document, describing the methodology used in the software.

Example

An example for preparing the K-means algorithm for submission to flowCAP using R and Linux is provided at:
`http://flowcap.flowsite.org/download/flowCAP/Example.pdf`

Agreement

Publishing the datasets provided by flowCAP is prohibited until the project publishes the results. The datasets and results of the flowCAP project will be publicly available for any use after the summit. Software submitted to flowCAP will remain confidential. Participant won't be identified (by name, group name, etc) in any materials without their approval.

References

[1] Nima Aghaeepour, Alireza Hadj Khodabakhshi, and Ryan R. Brinkman. An empirical study of cluster evaluation metrics using flow cytometry data. Whistler, British Columbia, Canada, December 2009. Clustering Theory Workshop, Neural Information Processing Systems (NIPS). `http://clusteringtheory.org/papers/empiricalmetrics.pdf`.

[2] F. Greg, B. Ali, B. Ryan, et al. Merging Mixture Components for Cell Population Identification in Flow Cytometry. *Advances in Bioinformatics*, 2009. `http://www.hindawi.com/journals/abi/2009/247646.abs.html`.

[3] Andrew Rosenberg and Julia Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. *In Proc.of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 410–420, 2007. `www.aclweb.org/anthology/D/D07/D07-1043.pdf`.

[4] Y. Shafranovich. Common Format and MIME Type for Comma-Separated Values (CSV) File. *Internet Society: Request for Comments (4180)*, 2005. `http://tools.ietf.org/html/rfc4180`.