FlowCAP-III Dataset Summary

September 26, 2012

Introduction

This document provides detailed instructions for participation in the FlowCAP-III challenges. For more information please visit http://flowcap.flowsite.org/, join the Google Group and mailing list at http://groups.google.com/group/flowcap, or contact rbrinkman@bccrc.ca. To receive instructions for participation please contact rbrinkman@bccrc.ca

Data Usage Agreement

Publishing or using the datasets provided by FlowCAP for other purposes is prohibited until the project publishes the results. The datasets and results of the FlowCAP project will be made publicly available or any use after publication. Once submitted to FlowCAP, the software and results cannot be withdrawn, but can be de-identified upon submitter's request.

Challenge 1: EQAPOL Proficiency Test

This dataset is produced by multiple laboratories participating in the External Quality Assurance Program Oversight Laboratory (EQAPOL) project. The goal of the study is to automatically identify two rare cell populations. The results will be evaluated against expert manual analysis. The dataset consists of 405 FCS files. A proportion (202) of the manual gating results are provided to the participants as a training-set through Comma Separated Value (CSV) files. In these files, 0, 1, and 2 correspond to "ungated", cell population 1, and cell population 2. Your task is to produce similar files (identical data format is required) for the samples in the test-set (those for which CSV files have not been provided). The result files provided in this challenge are derived from samples that have been subject to several potentially sources of variation (different stimulations, sample sources, and machine configurations). In the second part of this challenge (after the initial results have been submitted) information about the sources of variability between samples will be provided to the participants, and the methodologies that can benefit from such information can provide a secondary submission (see the timeline below).

Challenge 2: Survival Analysis

The goal of this challenge is to identify cell populations in high-dimensional FCM that correlate with a clinical outcome. The dataset includes FSC-A and FSC-H (for doublet removal), V-Amine/CD14 for gating live lymphocytes, SSC, and 11 other markers. The meta-data spreadsheet links the FCS files to the survival times of each subject (in days) and the observed clinical status (0: censored, 1: death). Half of the labels are provided for training purposes. Your task is to complete the missing values in the "Survival Time" column of the spreadsheet. Your results will be primarily evaluated using a cox proportional hazards regression and a log-rank test.

Challenge 3: Vaccination Time Points

This is a subset of an ICS (intracellular cytokine staining) data set produced by the HVTN (HIV Vaccine Trials Network), consisting of data from 74 subjects at two time points. All subjects have received a vaccine dose and have had blood samples drawn and analyzed for response to stimulation with a POL-3 peptide pool at time point A (visit code 2, before vaccination) and at time point B (visit code 12, day 183, primary immunogenicity time point). Each subject at each time point also has two negative controls, which are samples that have not been stimulated with POL-3 antigen. Thus, there are three files per time point and two time points per subject for a total of six files per subject and 444 FCS files in all. The challenge is to classify the samples into the correct visits / time points (either before or after vaccination). The data have been split into a training set (37 subjects, 222 FCS files) and a testing set (37 subjects, 222 FCS files). Two CSV (comma separated value) files are provided for each of the testing and training sets (filelist.train.csv, filelist.test.csv) that provide the visit numbers (2 or 12) for each FCS file in the training set, or for the testing set, contain a numeric code ("subjectmatchedfiles") that identifies which three FCS files (negative controls and stimulated) are from the same visit, but doesn't provide the visit code. Results will be evaluated for the accuracy of classification of the *samples* into visits (n=74) based on features extracted from the FCS files.

Challenge 4: Standardized Lyoplate Panels

This data set consists of standardized Cytotrol cells from Beckman-Coulter that were distributed amongst nine participating centers, and subjected to flow cytometry analysis using standardized "lyoplate" staining panels in quadruplicate. The panels

available for the FlowCAP challenge are the T-cell and the B-cell staining panels. There are 34 FCS files per panel, four from each center (except for one center which generated only two replicates). Cell populations of interest are defined a-priori for each staining panel. The data from all centers have been subjected to centralized, consensus manual gating to identify these cell populations.

The goal of the challenge is to use automated methods to gate the data and identify these same cell populations. PDFs of the gating scheme applied to each panel will be provided. These will identify the channels, shape and sequence of gates used to define the cell populations in the consensus manual gating so that participants fully understand how the target populations are defined.

Algorithms will be evaluated on their ability to gate these pre-defined cell populations (in a data-driven manner) with minimal bias (compared to the consensus manual gates), and minimal variance (within and between centers). Participants must provide cell population statistics (i.e. the number of cells in the population, and the proportion of cells relative to the parent population) for all cell populations in each panel from each sample and center. While some algorithms may define cell populations that do not have a strict hierarchical relationship, participants should endeavour to match the populations defined by their algorithms to the populations defined in the consensus manual gating. Note that FlowCAP will not perform any cell population matching for this challenge, as the goal is to accurately identify these pre-defined cell populations. All submissions must have cell populations labelled and matched to those in the manual gating consensus (Maecker, McCoy, and Nussenblatt 2012). Participants are also asked to submit cluster result (clr) files. These should be in the format of one column per population, and one row per event, with a 1 in the row if the cell is in the population. Only the populations defined in the results template need to be included. Columns should be labelled with the population name.

Our goal is to identify algorithms that can be utilized "out-of-the-box" to gate standardized lyoplate data in the future.

Reproducibility

A well-documented source code that can be used to reproduce your results from the raw data is mandatory. Please note that FlowCAP no longer accepts pseudocodes or binary files. Please provide detailed instructions for proper configuration of your scripts (paths, libraries for parallel processing, etc). If internal variables are altered for different challenges, please provide this information. **The FlowCAP evaluation committee reserves the right to re-run any analysis to reproduce submission results.**

Timeline

- Call for participants: 26.Sept.2012
- Primary deadline for submitting results to challenge 1: 23.Oct.2012
- Deadline for submitting results to challenge 4: 01.Nov.2012
- Deadline for submitting results to challenges 2, 3, and the second part of challenge 1: 08.Nov.2012
- The third FlowCAP summit will be held on NIH's Campus, Bethesda, MD, 29-30.Nov.2012 (http://palladianpartners.cvent.com/d/jcqwb4). A limited number of travel awards will be available to our best participants.

Maecker, Holden T, J Philip McCoy, and Robert Nussenblatt. 2012. "Standardizing Immunophenotyping for the Human Immunology Project.." *Nature Reviews. Immuno*logy 12 (3): 191–200. doi:10.1038/nri3158.