BIOINFORMATICS APPLICATIONS NOTE

Vol. 27 no. 5 2011, pages 732–733
doi:10.1093/bioinformatics/btr003

Systems biology

Advance Access publication January 5, 2011

flowPhyto: enabling automated analysis of microscopic algae from continuous flow cytometric data

Francois Ribalet†, David M. Schruth†, E. Virginia Armbrust*

School of Oceanography, University of Washington, Seattle, WA 98195, USA

Associate Editor: Olga Troyanskaya

ABSTRACT

Motivation: Flow cytometry is a widely used technique among biologists to study the abundances of populations of microscopic algae living in aquatic environments. A new generation of high-frequency flow cytometers collects up to several hundred samples per day and can run continuously for several weeks. Automated computational methods are needed to analyze the different phytoplankton populations present in each sample. Software packages in the programming environment R provide powerful tools for conducting such analyses.

Results: We introduce *flowPhyto*, an R package that performs aggregate statistics on virtually unlimited collections of raw flow cytometry files and provides a memory efficient, parallelized solution for analyzing high-throughput flow cytometric data.

Availability: Freely accessible at http://www.bioconductor.org

Contact: armbrust@u.washington.edu

Received and revised on November 2, 2010; accepted on December 30, 2010

1 INTRODUCTION

Phytoplankton are microscopic algae that are at the foundation of the marine food web and can influence Earth's climate. Flow cytometry has provided new insights into phytoplankton ecology in freshwater and marine environments. It creates 'fingerprints' of phytopkankton cells based on their ability to scatter or re-emit specific wavelengths of light. This light can be detected and used to estimate cell size and the composition of photosynthetic pigments that are used to distinguish different phytoplankton populations.

In the last decade, several high-frequency flow cytometers have been developed to study phytoplankton community structure at very fine spatial and temporal scales, collecting several hundreds of samples per day for several weeks. One of the largest flow cytometry datasets publicly available to marine biologists is produced by a new generation of flow cytometer created at the University of Washington, called SeaFlow, that continuously measures phytoplankton composition and abundance. The instrument generates the equivalent of 6700 samples, representing a dataset of 35–135 GB, after a typical 2-week long oceanographic cruise (Ribalet *et al.*, 2010). Software tools for automated data analysis and visualization of phytoplankton populations is therefore essential. We describe the R package *flowPhyto*, which provides a disk-based, parallelized solution to the analysis of high-throughput flow cytometric data.

2 DATA STRUCTURE

The *flowPhyto* package is compatible with conventional Flow Cytometry Standard (FCS) files (Spidlen *et al.*, 2010), as well as those from the SeaFlow repository. SeaFlow data are stored in a custom binary file (EVT file) created every 3 min and consists of eight 16-bit integer channels (see Ribalet *et al.*, 2010 for more details). The acquisition time and location (longitude and latitude) of EVT files is written into a log file (SDS file). The SeaFlow repository is composed of julian day labeled directories, each containing chronologically-ordered EVT files and an SDS file. The partioning of the data circumvents hardware memory constraints and enables parallelization over a beowulf class computer cluster.

As an example, we use a subset of SeaFlow data available at http://seaflow.ocean.washington.edu that represents 1 day (480 files) of an oceanographic cruise that took place in Puget Sound, WA, USA, in November 2009. The goal of the study was to identify hotspots of phytoplankton diversity and abundance.

3 COMPUTATIONAL ANALYSIS

The *flowPhyto* package consists of a collection of functions that turn the high volume of flow cytometry files into customizable cytograms, georeferenced images and a summary data table. Figure 1 depicts the analysis pipeline from raw flow cytometry files to end statistics through four batch processing steps, namely `filter`, `classify`, `census` and `summarize`. Each step makes a separate pass over the entire repository and processes the samples in a highly parallel fashion using file-based wrapper functions.

Unlike a traditional flow cytometer, SeaFlow directly analyzes a raw stream of seawater using two detectors that determine the position of a particle in the focal region of the instrument optical system (Swalwell *et al.*, 2009). The `filter` function selects optimally positioned particles (OPP) in each EVT file and creates an OPP file used to distinguish the different phytoplankton populations.

Because the characteristics of each phytoplankton population vary according to environmental conditions and instrument settings, a table of customizable parameters (pop.def.tab) is used to define the pre-gating regions and statistical priors of phytoplankton population clusters. The `classify` function uses these predefined parameters and inputs one or more OPP files (3 by default) to classify individual phytoplankton cells into different populations using statistical clustering methods (Lo *et al.*, 2009). The function can also input FCS files after conversion to OPP files. For each group of OPP

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First authors.

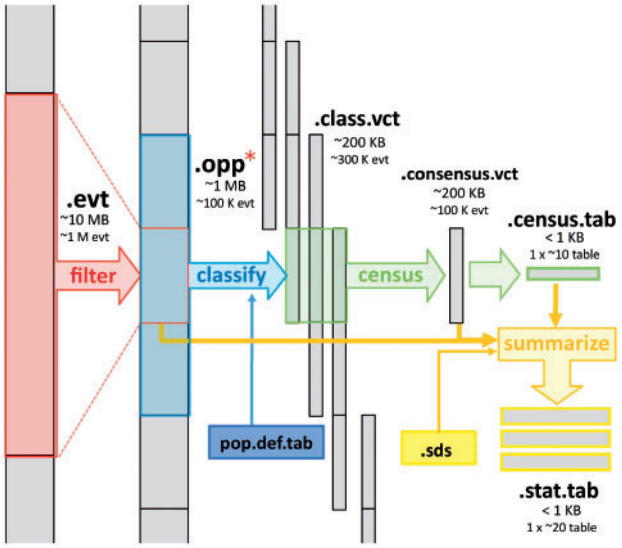

Fig. 1. Diagram of the *flowPhyto* package analysis pipeline. The four main functions `filter`, `classify`, `census` and `summarize` convert raw flow cytometry files into a summary table for post-analysis. Stacked vertical gray bars represent the different files. The asterix (*) indicates that `classify` can input FCS files and OPP files from SeaFlow repository.

files, `classify` outputs corresponding vector files (consensus.vct) that contain the population identification of the cells. `classify` is run in single file increments to provide multiple passes over a single cell and strengthen the clustering analysis.

During the `census` step, these multiple-pass vector files are collapsed into one consensus vector, which represents the most likely population classification of the different phytoplankton cells. In addition, `census` produces a one-row census tab file that contains the number of cells per population for each file. The concatenation of these census tab files is used to create a per-population resampling scheme that calculates the number of OPP files necessary so a sufficient number of cells (500 by default) is present in the resampled population.

The `summarize` function performs per-population aggregate statistics using the resampling scheme and associates the corresponding acquisition time and location. It outputs a summary table (stat.tab) of the entire dataset that can easily be exported by the user.

4 VISUALIZATION

The `filter` step optionally outputs a quality control plot for the position-sensitive detectors specific to SeaFlow technology. `plotCytogram` outputs a series of customizable 2-D cytograms to visualize the various phytoplankton populations present in the sample for each OPP or FCS file.

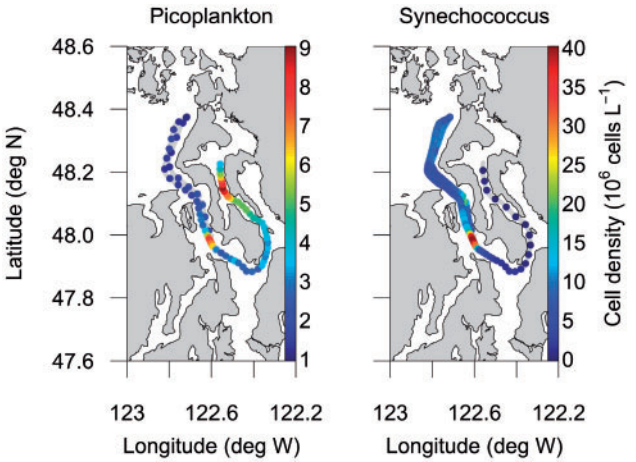

Fig. 2. Distribution of two phytoplankton populations, namely picoplankton and *Synechococcus*, in Puget Sound, WA, USA on 9 November 2009. Note that the spatial resolution of the two populations improves as their relative cell density increases.

The `plotStatMap` creates customizable plots of the geo-referenced data created by `summarize`. A combination of the different parameters per population or a single parameter over different populations can be selected depending on the purpose of the analysis. Hotspots of phytoplankton diversity and abundance are shown in Figure 2.

5 CONCLUSION

flowPhyto provides an efficient approach for the analysis of high-throughput flow cytometry data. Although this approach provides an unique solution for the analysis of SeaFlow data repository, it is also compatible with standard flow cytometry data.

ACKNOWLEDGMENTS

We thank V. Iverson for his comments on the manuscript.

Funding: Gordon and Betty Moore Foundation Marine Microbiology Investigator Award (to E.V.A.).

Conflict of Interest: none declared.

REFERENCES

Lo,K. *et al*. (2009) flowclust: a bioconductor package for automated gating of flow cytometry data. *BMC Bioinformatics*, **10**, 145.

Ribalet,F. *et al*. (2010) Unveiling a phytoplankton hotspot at a narrow boundary between coastal and offshore waters. *Proc. Natl Acad. Sci. USA*, **107**, 16571–16576.

Spidlen,J. *et al*. (2010) Data file standard for flow cytometry, version fcs 3.1. *Cytom. Part A*, **77**, 97–100.

Swalwell,J. *et al*. (2009) Virtual-core flow cytometry. *Cytom. Part A*, **75**, 960–965.