Analysis of Multiple Datasets

February 7, 2021

$Revision: 1688 $

1 Summary

This document describes statistical analysis of segments obtained from multiple datasets.

2 Introduction

Segment properties obtained form segment analysis from multiple experiments are separated into experimental groups. Experimental groups are typically based on the experimental conditions (such as pharmacological treatments or genetic manipulations), but other criteria can also be used (such as the presence of certain organelle). Thus, each experimental group contains data from several experiments (often 5-10). Segment properties are analyzed within groups and the results (such as means) are then statistically compared between groups. This allows identification of properties that differ significantly between the groups and are then likely to be caused by the experimental conditions.

In this document only the general concepts are presented. For the presynaptic workflow, more detailed description that includes specific parameter values is presented in workflows.pdf. For other applications, the user is still advised to start from the presynaptic workflow and customize the files mentioned there.

2.1 Terminology

- An experiment (or observation) contains the data obtained by segmentation and analysis of one dataset (one biological object of interest).
- A group is a collection of experiments obtained under the same experimental conditions (repetitions). These groups correspond directly to the experimental groups used in the standard statistical analysis. A group is typically defined by a specific pharmacological manipulation or genetic background, but it can be also based on biological content (such as containing a specific organelle).

3 Procedure

3.1 Prerequisites

3.1.1 Individual analysis pickles

All individual datasets have to be segmented and analyzed as explained in the segmentation and analysis guide (segmentation.pdf). This procedure creates pickle files that are dataset and processing step dependent (individual analysis pickles in Fig 1).

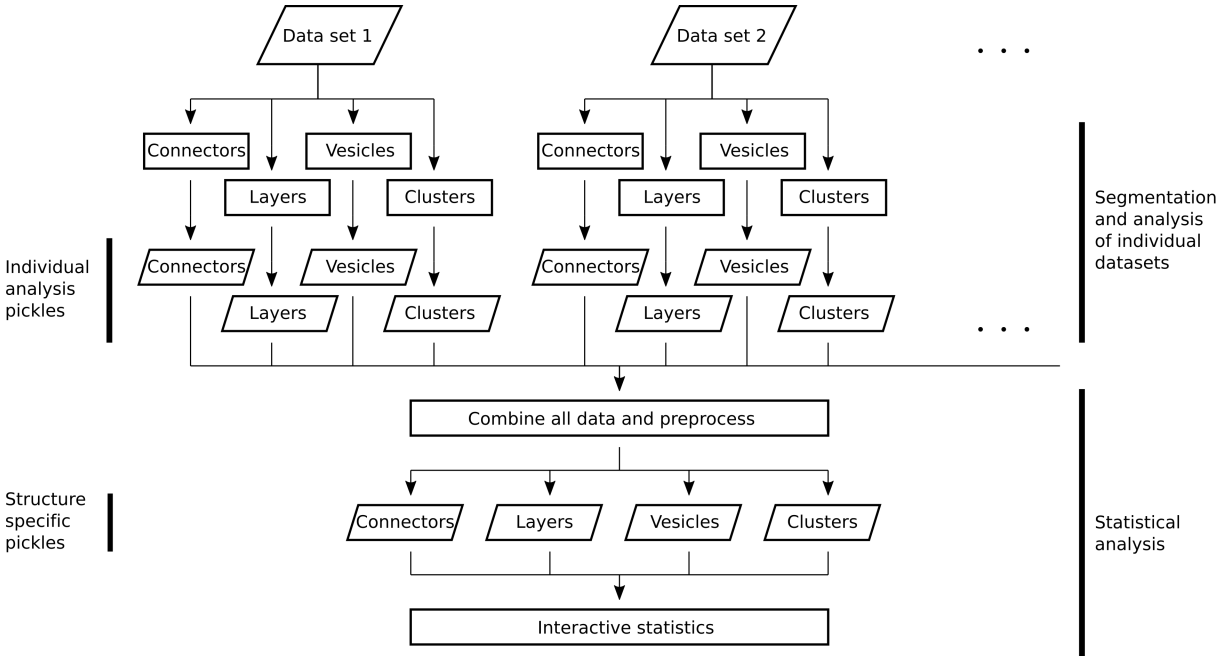

Figure 1: The Pyto workflow for presynaptic segmentation and analysis. Processing steps are represented by rectangles and data by rhomboids. The upper part shows processing (segmentation and analysis) of individual datasets. It consists of processing steps that perform the hierarchical connectivity segmentation ("Connectors") and the analysis typically used for the presynaptic terminals ("Connectors", "Layers", "Vesicles" and "Clusters"). Results of the processing steps are saved in Python pickle files indicated by the same names. The bottom part represents the statistical analysis of all data. First, all single dataset analysis results are combined and preprocessed and the results are saved in biological structure-specific pickle files ("Connectors", "Layers", "Vesicles" and "Clusters"). Second, the data from the structure-specific pickles are combined to allow statistical analysis between datasets organized in experimental groups and the interactive display of this data.

3.2 Creating catalog files

It is recommended to first make unique identifiers for experiments and names for experimental conditions (treatments) first. For example, use 'wt, 'protein-x_ko', 'untreated', 'compound-y' for experimental conditions and: 'wt_1', 'wt_2', 'protein-x_ko_1', 'untreated_1', 'compound-y_1', 'compound-y_4', 'compound-y_5', ...

The user has to make catalog files that contain data about experiments (metadata). Catalog files are dataset specific, that is each dataset has to have a corresponding catalog file. All catalog files need to reside in the same directory. Typically, catalog files are very simple, they only contain statements of the type `variable = value`. Ideally, all catalog files should define the same variables. However this is not an absolute requirement, but it may cause errors.

More precisely, catalog files have to be valid Python files. They are imported by Python `import` statement, so they may contain more complicated Python statements.

In any case, the following variables have to be defined in catalog files:

- `identifier` (str): unique experiment identifier
- `treatment` (str): specifies experimental condition.
- file names of all individual analysis pickles
- `pixel_size`: pixel size in nm
- (optional) arbitrary number of other variables can be defined, such as those that describe tomographic acquisition, processing or biological content

3.3 Statistical analysis of multiple datasets

3.3.1 Generation of the structure-specific pickles

In this step, the analysis data from all individual analysis pickles and the metadata from the corresponding catalogs are combined. This data is internally organized in the experimental groups to facilitate statistical analysis between the groups.

In addition, new properties are calculated. These include:

- The conversion of properties that are in pixel units to the corresponding properties given in nm (such as lengths, distances, surface area, volume).
- Combinations between the existing properties. For example, for the presynaptic terminals the number of connectors is calculated for each synaptic vesicle.

All these results are separated according to the structures to which they refer (core) and saved as the structure specific pickles (Fig 1). Each of these pickles focuses on one core structure, such as segments or layers, and contains data for all experimental groups.

3.3.2 Interactive statistical analysis

This analysis can be separated in the following steps:

- Loading the individual structure specific pickles
- Further preprocessing
- Execution of predefined functions that statistically analyze data between experimental groups and display the results

These scripts are provided for the analysis of the presynaptic terminal (presynaptic_stats.py) and the synaptic cleft (cleft_stats.py). In these cases the script require only minimal modifications (see workflows.pdf). For other applications these scripts should be used as initial templates, but they might require more extensive modifications.

3.4 Including additional datasets

When a new datasets is added, a corresponding catalog file needs to be created and the complete statistical analysis of multiple datasets needs to be repeated to generate the structure specific pickles that also contain the data from the new dataset. New pickles have to be created also when something is changed in the segmentation and analysis of individual datasets resulting in the modification of one or more individual analysis pickles, as well as when one or more catalogs are modified.

4 Applications

Practical details about of the segmentation and analysis procedure are available for the following specific applications:

- Presynaptic terminal (see workflows.pdf)
- Synaptic cleft (work in progress)

5 Citing

Please consider citing us if you use segmentation and analysis in Pyto:

Lučić V, Fernández-Busnadiego R, Laugks U and Baumeister W, 2016. Hierarchical detection and analysis of macromolecular complexes in cryo-electron tomograms using Pyto software. J Struct Biol. 196(3):503-514. doi: 10.1016/j.jsb.2016.10.004.

Thank you.