**ARI\_IoU\_calculation readme**

1. **Installing R (3-5 minutes)**

The code runs in the R programming language. To install, download R from: <http://www.r-project.org/>

We then recommend the use of R-Studio, which provides a graphical user interface for R. This is also freely available at:

<http://www.rstudio.com/>

The first time the code is run, additional R plugins are required and can be installed using the tools -> install packages menu. The plugins required are:

* mclustcomp
* dbscan
* gtools
* imager
* spatstat
* contoureR
* sp
* RColorBrewer
* grDevices

This code was written and run using R version 3.6.1 and will run on PC and Mac with at least 4GB RAM.

1. **Data format.**

We have included an example data set, however, should you wish to test the code with your own data it needs to be in a specific format. – the same as our example

1. Create a folder, e.g., with the name of the algorithm or experiment, in our example it is “test”, and place the “ARI\_IoU\_source.R” file inside.
2. Within this folder you may place different folders pertaining to the different cluster simulation scenarios you want to analyse; in our example we have “simulation\_1” and “simulation\_2”. Within there you can also generate separate folders, if you have may different ROIs form the same simulation scenario; “ROI\_1” AND “ROI\_2” in our example.
   * N.B. The code currently is set to find folders with the file name pattern “simulation\_”, set in the pfolders variable. If you name your folders differently, please modify this accordingly.
3. Within each folder you place the ground truth data:
   * It must contain 3 columns: x coordinates (in nm), y coordinates (in nm) and ground truth index for each molecule, in that order. The format must also be .csv
   * The headings must be “x”, “y”, and “index”, like in the example file (data.csv).
4. Create another folder within the test folder called “classes”, and place in here the result indexes from your clustering algorithm.
   * You can place multiple result index files in here, for example, if you used several different user settings for your algorithm.
   * The file must be a single column (heading – “result”), with the cluster result indexes for that particular clustering result, and in .csv format.
     + For example, if there are 600 points in total in the ground truth data, then there must be 600 indexes in the file, i.e., all points must be assigned an index. In our example case, for both the ground truth and the cluster results, noise/non-clustered points are assigned a zero as an index, and clustered points get non-zero indexes.
5. **Running the ARI\_IoU\_calculation R script (~ 3mins on test data set)**

Once the data are in place the ARI and IoU values can be calculated using “ARI\_IoU\_calculation.R”. The code will batch process the data, provided the format and file structure are as above.

1. Open RStudio
2. Open the file named “ARI\_IoU\_calculation.R”
3. Enter the file path where the “ARI\_IoU\_source.R” and data were placed. For example:
   1. “~Desktop/ARI\_IoU\_Indexes/test/”
4. In line 6, enter the size of one dimension of your region of interest in nanometers.
   1. We recommend square regions no larger than 2000\*2000 nm.
5. Click “Source” to run the code, and the code will then run. Time to complete will depend on the number of files placed on the class folder.
6. Once the code has finished it will have created a new timestamped folder within each separate data folder. In this folder will be:
   1. A single .csv with the ARI and IoU values for each of the files within the class folder (filenames retained for easy identification after analysis)
   2. A .PDF with plots of the polygons extracted for the IoU calculations for each class (plots: ground truth point pattern, ground truth polygons, class file results polygons, overlay of ground truth and results polygons).
   3. A .RData file, which can be opened in R Studio if you wish to use the output and source data for any further analysis, processing or plotting.

The code should take approximately 3 minutes to run on the test data set on a system with at least 4 GB RAM.