27.10..2023 by Laura Martin

CLUSTER\_ANALYSIS\_main. In details.

**>defineFindClustersStruct**

**>FindClusterStruct** Set analysis **parameters**

**>Select1DataGroup**

**>FindClusters**

Identifies Islands and clusters

**Calculates clusters’ features**:

* n° localizations
* Areas
* NND (of In-Island only)

**>Insight3**

**>Inisght3io**

**Saves** output file (.bin, .xyn)

**>saveClusterMetricData**

**>calcMaskAreas**

**>loadI3data**

**>Locs2Mask**

**>create\_xynFilename**

**>extractClusterStats**

**>XYN**

**>DDC**

**>DistanceDualColor**

Calculates the **global NND**

between ALL clusters

**Saves** output file (.mat, .ddc)

**>compareClusterMetricData**

**>calc\_plot\_ClusterMedians**

**>plotCompareData**

**>plotClusterStat**

**>calcMaskAreas**

**>plotStairs**

**Saves** output files (.fig, .png)

|  |  |  |  |
| --- | --- | --- | --- |
| **NAME** | **INPUT** | **OUTPUT** | **FUNCTION** |
| **CLUSTER\_ANALYSIS\_main** |  |  |  |
| **defineFindClustersStruct** | pix2nm | params | Create GUI to set analysis parameters |
| **FindClusterStruct** | --- | --- | Defines the defaults parameters |
| **Select1DataGroup** | .bin | files=structure(filename, file\_location) | Create GUI to select .bin files |
| **FindClusters** | findClustersStruct | results, i3Out, pathstr, subFolder, fileName | - Creates 'cluster\_analysis' folder.  - Perform the majority of cluster analysis.  - Creates an array ‘ClusterResults’ of 12 columns of cluster metrics  -**Saves the .bin and the incompete .xyn** (only 12 columns, no global NND) |
| **Insight3** | .bin | Data inside .bin | reads .bin localizations files |
| **Inisght3io** |  |  |  |
| **saveClusterMetricData** | params, ClusterResults, pathstr, subFolder, fileName | xynData, mask, ClustStats, ddcData, savefile | **Overwrites** complete **.xyn** (adding the 13rd column of NND Global**), and saves .ddc and .mat** files |
| **calcMaskAreas** | LL, params, true | mask | Calculates conventional & S.R. areas from localization list |
| **loadI3data** | LLname = .bin or .txt localization list | LL=.bin or .txt | Checks that localization list has the right format |
| **Locs2Mask** | xy, finalPix, pix2nm, showText | mask,area,pixLocs | Calculates area occupied by total locs |
| **create\_xynFilename** | findClustersStruct | xynFilename | Creates .xyn file inside ‘cluster\_analysis’ |
| **extractClusterStats** | filexyn, params, ClusterResults | InIslandData, GlobalData, ClustStats | Organizes clusters info from .xyn |
| **XYN** | .xyn | Data inside .xyn | Reads the .xyn file |
| **DDC** | .DDC | Data inside .DDC | Reads the .DDC file |
| **DistanceDualColor** | filexyn, filexyn, Inf, params.original\_pixel\_size,...  params.minimum\_molecules\_per\_cluster-1, false | cat, counts | Calculates the global NND between ALL clusters |
| **compareClusterMetricData** | .xyn or .mat files from cluster\_analysis folder | .fig,.png,xlsx of statistics and bar plots | Saves .fig,.png,xlsx of clusters metrics comparison |
| **calc\_plot\_ClusterMedians** | type,startlocn,FileExtension,saveFileLoc,barPosn,alphaValue | export | Calculates median values of clusters metrics from .xyn |
| **plotCompareData** | plotdat,barPosn,alphaval,… strIn,figH,ax\_h,nrc,subIdx | p\_values, ax\_h, figH | Creates box plot with statistical test |
| **plotClusterStat** | filename, xynData, mask, ClustStats, ddcData, params | fg | **Saves plot.fig** and plot.png |

**Detailed description of CLUSTER\_ANALYSIS\_main workflow**

In **CLUSTER\_ANALYSIS\_main** we can define “rootdir”, if to “plotstats” or not, and the “pix2nm” size of the microscope we used to acquire the images.

Then, CLUSTER\_ANALYSIS\_main calls **defineFindClustersStruct**.

defineFindClustersStruct defines the GUI for entering the analysis parameters. Here we can change the size of the GUI (ln 12) and the names of the parameters to insert (e.g. ln 52).

defineFindClustersStruct calls **FindClustersStruct**.

FindClustersStruct defines the values of the parameters which will appear in the GUI as default.

CLUSTER\_ANALYSIS\_main calls **FindClusters**.

**FindClusters.m:**

Extracts information from localization list.

Resizes the image of a chosen zoom\_factor = original\_pixel\_size/analysis\_pixel\_size;

Xn are the locs coordinates changes for the new zoom

M is the matrix of the number of locs inside each pixel (if px=160nm, 4096x4096 px)

Mf = imfilter(full(M),ones(roi)); imfilter sums the number of molecules of each pixel of a roi to the central pixel of the roi. Thus, Mf is the matrix of all pixels values after the summation of the number of localizations.

Mf2 is an empty matrix of the same size of M. It is filled with only the pixels which contain a number of locs >threshold.

* Alert! If Mf2 is not filled, it means that the code found no pixels with locs >threshold. Thus the threshold values is too high and should be reduced.
* If we put True for Use Iterative Segmentation, go to ln 920-1104, function segment\_nn(Mf,Mf2,BW,th,max\_area). If there is any Island with an area >max\_segmentation\_area, then this Island is segmented to reduce it.

Mf2 is converted to a Black & White image (BWL). Pixels ≠ 0 are White (1), pixels = 0 are Black (0).

Regionprops(Mf2) automatically identifies Islands in the B&W image of Mf2.

m is the matrix containing all the Islands (j).

BB=m(j).BoundingBox identifies each white block [=Island, =m(j)] and isolates it by the function Bounding Box (square).

Mi=m(j).Image. Each Island [m(j)] is converted into an image.   
Mi=px x px (=a subgroup of 4096x4096 px)

xi\_b is the matrix containing the coordinates of localizations of each Island.

[iok, jok] is the matrix of the pixels of each Island [only the White (1) pixels of each Bounding Box].

* If an Island has a number of pixels (iok) > ignoreNumPeakThreshold, this Island will be discarded from the analysis. If you want to analyze every Island, increase the ignoreNumPeakThreshold (keep in mind that the analysis will take very long).

xi\_c is the matrix of localization coordinates belonging to each Island (Mi), updated after the filtering of Islands [(iok) > ignoreNumPeakThreshold]. See function xi\_c in ln 866-918.

Mi\_c is a new empty matrix. It has the size of the Mi image, after resizing of a chose factor (defined in the parameters at the beginning).  
(if Mi was 400x400 px, and factor is 5, Mi\_c will be 2000x2000 px)

Mi\_c=zeros(ceil(size(Mi)\*factor));

If the size of the island is small and the factor is very small too, the size of Mi\_c could be as small as 1x1.

* If an Island has a number of locs (xi\_c) > ignoreNumIslandThreshold, this Island will be discarded.

xo, yo are the new coordinates of locs of each Island (Mi), once resized of the factor.

sigma = precision / (analysis\_pixel\_size/factor). Sigma is defined as the Localization precision (in nm) (defined at the beginning) normalized to the new pixel size.

minx, maxx, miny, maxy. These coordinates define a square roi around each localization.  
The extremities of this square are:  
 - mix x = left = max(1, floor(x0-numSigma\*sigma))  
 - max x = right = min(size(Mi\_c,2), ceil(x0+numSigma\*sigma))  
 - min y = up = max(1, floor(y0-numSigma\*sigma))  
 - max y = down = min(size(Mi\_c,1), ceil(y0+numSigma\*sigma))

These extremities are located at numSigma\*sigma from the coordinates of the localization.  
numSigma is set as 3 but can be changed. The smaller the numSigma, the smaller the area associated to each localization.

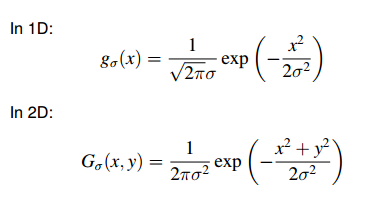
For how the roi is defined, if Mi\_c size is small, this could lead to maxx<minx and maxy<miny 🡪resulting in xval and yval in empty matrices  
  
🡪in ths case the Mi\_c will become 0 = no signal in that Island= Island discarded..

In each square roi localization, a kernel Gaussian is drawn, with the peak corresponding to the coordinates x,y of the localization.  
Mi\_c(yval,xval) = Mi\_c(yval,xval) + G(xval, yval, x0, y0, sigma).

The formula for the kernel 2D Gaussian is the following:  
G= @(x,y,x0,y0,rx) (1./((2\*pi)\*(rx\*rx))).\*exp(-(((x-x0)).^2)/(2\*rx^2).\*exp(-(((y-y0)).^2)/(2\*rx^2))

Function handle

Formula



All kernel Gaussias belonging to the same Island are summed together, generating a local density map of Gaussians. The values are carried in the matrix called Mi\_c.

im\_bw(find(Mi\_c>((1-kkk(k))\*max(max(Mi\_c)) )))=1;

kkk=(linspace((1/10),(9/10),10))=

k=1:length(kkk)

kkk(k)=0.9

(1-0.9)\*max(max(Mi\_c))= 0.1\*maxGaussianvalue in matrix Mi\_c = THRESHOLD

im\_bw is a matrix assigning 1 to the cells with values> than threshold.

All the others becomes 0.

IM\_BW is a new empty matrix of the same size of Mi\_c.

im\_bw is copied onto the empty matrix IM\_BW.

The functions bwlabel and regionprops are used to extract the feature of the Centroids from the Mi\_c. This process is repeated for each Island (Mi\_c).

Function lik\_sig associates the localizations (N) to each Centroid previously automatically found through regionprops. See function lik\_sig in ln 784-862.

Then N (localizations belonging to one cluster) are filtered. Only clusters having N ≥ minCluster are kept, the others are discarded. minCluster is defined in the parameters at the beginning as minimum\_molecules\_per\_cluster.

The Centroid of each cluster (previously automatically extracted by the function regionprops) is re-calculated as the mean of the coordinates of N localizations. These new coordinates should agree with the ones previously found by regionprops.

Then, all features calculated are summarized:

* X = x coordinate of the centroid
* Y = y coordinate of the centroid
* NN = Number of total localizations of the cluster
* SIGX = Standard deviation of x coordinates of cluster localizations
* SIGY = Standard deviation of y coordinates of cluster localizations
* SIG = (xstd+ystd)/2. Average of sd of x and y coordinates.
* SIGQUAD = sqrt(xstd^2 + ystd^2).Sd of the centroid.
* Z = z coordinate of the centroid.   
  (Calculated as the average of z- coordinates of localizations belonging to the cluster.)
* SIGZ = Standard deviation of z coordinates of cluster localizations
* xyzNND = X and Y of centroid(s) of the cluster(s) of one Island, in nm. (Temporary).

Only for clusters inside Islands (thus, NOT for single clusters), it calculates the Nearest Neighbour Distance (NND).

* NCI = index of the clusters inside the Island (1, 2, 3, ...) or Single (1).
* IslandIdx = index of the Island to which the cluster belongs to

xyzNND is a matrix containing the x-y-z-coordinates of the centroid(s) of one Island.  
If ncisland is the dimension of xynNND matrix. If ncisland >1 it means that the Island contains more than >1 cluster, so the NND can be computed, as:  
squareform(pdist(xyzNND)) + diag(Inf\*ones(size(xyzNND,1)  
Otherwise, it means that xynNND has only one cluster, so it is not an Island but a Single Cluster, and thus NND is “Inf”.

Features of every cluster identified will be summarized in columns in an Excel.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| X | Y | NN | SIGX | SIGY | SIG | SIGQUAD | Z | SIGZ | NND | NCI | IslandIdx |

Finally, the code extracts few other parameters:

* medN = Median of number of localizations per cluster
* medArea = Median of cluster area.   
  Cluster area is calculated as: π\*(SIG)2.
* Median density = medN / medArea

In ln 691 of **FindClusters.m** the code generates a .bin file for clusters visualization with the Insight3 software.  
In details:

- if in the initial GUI we put “False”, all localizations belonging to the same cluster will be displayed with the same color, randomly chosen among the 9 available (with no correlation with the z coordinate of each centroid). The centroids are displayed in white.  
- if in the initial GUI we put “True” for drawUsingZRange instead, the localizations of each cluster will be assigned to 9 different channels, based on the z-coordinate of their centroid. The z- range is defined as the interval between the max and min z-coordinate of all centroids, divided per 9. z-min correspond to magenta color, while z-max corresponds to red color. So, localizations will be displayed with a different color depending on the z-coordinate of the centroid of the cluster to which they belong. Localizations of different clusters will be displayed with the same color if the z-coordinate of their centroids belong to the same interval.

Then CLUSTER\_ANALYSIS\_main calls **saveClusterMetricData.**

SaveClusterMetricData creates a temporary .xyn file containing the parameters used for the analysis (ln 18). It then calls **extractClusterStats**.

extractClusterStats calls **DistanceDualColor**,and defines parameters of cluster files, categories and minimum number of molecules per cluster in ln 151-152.

**DistanceDualColor:**

It was originally created to measure the distance between two different protein clusters (labelled with 2-colors). Inside this code, it is used for 1-color cluster images.

It takes one cluster file (.xyn) generated from FindClusters.m and doubles it, as they were two different lists of clusters (=data1; data2).

Clusters listed in data1 and data2 are checked for dimension. They have to contain > minimum\_localizations to be further analysed.

It calculates the distance in x-y-z dimensions between data1 and data2 in nm:

Dist = sqrt( ((data1(i,1)-data2(j,1))\*pixel\_size\_nm)^2 + ((data1(i,2)-data2(j,2))\*pixel\_size\_nm)^2 + (data1(i,8)-data2(j,8))^2);

Every time the same cluster is picked, distance is set as “Inf” and ignored.

if isSameFile dist(j,k) = Inf;

All distances are sorted (function srt) in order to pick the closest neighbour for every cluster.

The clusters can be separated into categories, depending on their distance. For this code, categories are not defined (ln 181-182), so NNDs are listed all together.

{Otherwise, categories can be defined (e.g. [ 70 250 ] so that NND are divided into 3 groups: NNDs < 70 nm, 70 < NNDs < 250 nm, NNDs > 250 nm), see from ln 184}.

**extractClusterStats** saves the list of NND global distances in a ddc file called “ddcData.nndXY”.

**SaveClusterMetricData** appends the results “ddcData.nndXY” to the results generated in FindClusters.m as a 13th column.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| X | Y | NN | SIGX | SIGY | SIG | SIGQUAD | Z | SIGZ | NND | NCI | IslandIdx | NNDglobal |

This table of 13 columns is temporarily saved as the matrix “allResults”.

The .xyn file previously generated in ln 18 is temporarily saved with a name containing the original name of the list file plus the abbreviations of the parameters used for the analysis (ln 55 or ln 57).

* Warning! If the original list file has a long name, the code could give error when saving the final .xyn. Apparently .xyn cannot be open correctly when they have a name > 256 characters. If .xyn file gives problem, rename the .xyn (reducing the number of characters) or rename the original list and re-run the cluster analysis.

Inside the .xyn file, the results of “medianNum, “median Area[nm^2]” and “medianDensity” are appended (ln 79, 81, 82).

Finally, the 13 columns table with all clusters feautures is appended (ln 85). The .xyn in now completed and saved for the last time.

At the very end, all the results calculated up until now are saved in a .mat file (ln 103).

Back in **CLUSTER\_ANALYSIS\_main** the plot statistics (if we had put “true”in ln 5) are saved as .fig and .png files.