frontiers | Frontiers in Bioinformatics

TYPE Mini Review
PUBLISHED 24 November 2023
DOI 10.3389/fbinf.2023.1237551

OPEN ACCESS

EDITED BY
Hannah S. Heil,
Gulbenkian Institute of Science (IGC), Portugal

REVIEWED BY
Sarah Aufmkolk,
Harvard Medical School, United States
Gerhard J. Schütz,
Vienna University of Technology, Austria

*CORRESPONDENCE
Juliette Griffié,
juliette.griffie@dbb.su.se

RECEIVED 09 June 2023
ACCEPTED 28 September 2023
PUBLISHED 24 November 2023

CITATION
Panconi L, Owen DM and Griffié J (2023), Cluster analysis for localisation-based data sets: dos and don'ts when quantifying protein aggregates. *Front. Bioinform.* 3:1237551. doi: 10.3389/fbinf.2023.1237551

Cluster analysis for localisation-based data sets: dos and don'ts when quantifying protein aggregates

Luca Panconi[1], Dylan M. Owen[1] and Juliette Griffié[2]*

[1]School of Mathematics, Centre of Membrane Proteins and Receptors (COMPARE), Institute of Immunology and Immunotherapy, University of Birmingham, Birmingham, United Kingdom, [2]Department of Biochemistry and Biophysics, Stockholm University, Stockholm, Sweden

Many proteins display a non-random distribution on the cell surface. From dimers to nanoscale clusters to large, micron-scale aggregations, these distributions regulate protein-protein interactions and signalling. Although these distributions show organisation on length-scales below the resolution limit of conventional optical microscopy, single molecule localisation microscopy (SMLM) can map molecule locations with nanometre precision. The data from SMLM is not a conventional pixelated image and instead takes the form of a point-pattern—a list of the x, y coordinates of the localised molecules. To extract the biological insights that researchers require cluster analysis is often performed on these data sets, quantifying such parameters as the size of clusters, the percentage of monomers and so on. Here, we provide some guidance on how SMLM clustering should best be performed.

KEYWORDS

cluster analysis, single molecule localisation microscopy (SMLM), protein aggregates, image quantification, bioinformactics, spatial point pattern (SPP)

Introduction

Cellular processes heavily rely on the ability of key proteins to form aggregates, also called clusters. Immune cells for instance are regulated through subtle variations in signalling protein clustering characteristics. These clusters have now been shown to involve only a small number of proteins and range from 10 nm to 50 nm in size (Griffié et al., 2015). Until the development of super resolution microscopy (SRM), light microscopy, bound by the diffraction limit (>200 nm), was unable to resolve cells' nanoscale architecture, including clusters. SRM in comparison encompasses imaging techniques with a spatial resolution below 200 nm. SMLM in particular achieves a resolution close to molecular scale (typically 10 nm) in cells (Lelek et al., 2021). It enabled, for the first time, the visualisation and quantification of protein nanoscale organisation including clusters, pores and filaments.

SMLM relies on the separation in time of fluorophores' emission (i.e., blinks), which are collected over thousands of frames. Ideally, on each frame only a very small subset of well spatially separated fluorophores are emitting, allowing to extract from their diffraction limited point spread function their precise localisation. Therefore, the output of an SMLM acquisition does not consist of a conventional pixelated image, but rather of a spatial point pattern (SPP, i.e., scatter plot of collected localisations (x, y) in 2D (x, y, z) in 3D). Ultimately, it is the estimated uncertainty associated to every localisation that is often used as a proxy for spatial resolution. SPPs require totally different statistical tools for their analysis compared to

pixelated images. Although relatively uncommon in microscopy, this type of dataset has been analysed for decades in fields such as ecology and epidemiology. This has provided a baseline for the development of a vast range of cluster analysis tools dedicated to SMLM (Khater et al., 2020). Here, we will discuss the dos and don'ts when analysing SMLM SPP for cluster quantification.

Dos

Pre-processing matters!

SMLM data sets require comprehensive curation before analysis (Figure 1). This includes drift correction and filtering badly localised emitters. Multiple blinking, which translates into localising the same fluorophore multiple times over the acquisition, also has to be addressed as it leads to self and artificial clustering (Annibale et al., 2011; Baumgart et al., 2016). In the case of PALM, there are 2 powerful existing strategies to account for multiple blinking. The first relies on the relative short timescale over which the reblinks tend to happen in order to merge them (Annibale et al., 2011). The second more recent solution is model-based and has been shown to provide artifact free point distribution for cluster analysis with unprecedented reliability (Jensen et al., 2022). In both cases, multiple blinking correction should be included in the analysis pipeline upstream of the cluster analysis. When it comes to STORM however, the issue has not been fully addressed and remains an open avenue of research (Bohrer et al., 2021). A good practise however remains to merge blinks that extend over multiple frames.

Multiple blinking is not the only pre-processing requirement for most cluster analysis tools. A field of view is typically made of millions of localisations, with strong variations in overall point density (e.g., in and outside the cell) and clustering (e.g., sizes), for which most methods fail to provide accurate clustering descriptors. This results from widespread cluster analysis methods [e.g. DBSCAN (Ester et al., 1996), localised Ripley's K (Owen et al., 2010)] reliance on fixed user parameters which cannot fit all noise and clustering variations in the same ROI as well as across ROIs. The computational cost of handling millions of points is also a key limitation, apart from graph-based methods specifically repurposed for large data sets (e.g., whole cell, field of view) (Levet et al., 2015). Both factors indicate the overall need to define regions of interest (ROI) in which the number of points does not limit the computation and the clusters within the picked ROI display similar characteristics. In cases where broad cluster density/size range (within the same ROI or across the data sets) as well as uneven background cannot be addressed with ROI picking, a Bayesian-based cluster analysis tool will typically provide more reliable results as it picks the best parameters pair, in regard to a realistic model on protein aggregation in cells, independently for each ROI (Rubin-Delanchy et al., 2015; Griffié et al., 2016). In all cases, ROI should also be associated to edge correction strategies. Most cluster analysis tools now include edge correction directly in the analysis pipeline (e.g., symmetry, framing) but the user must be aware of which strategy is implemented to provide an ROI of bigger size for instance if required.

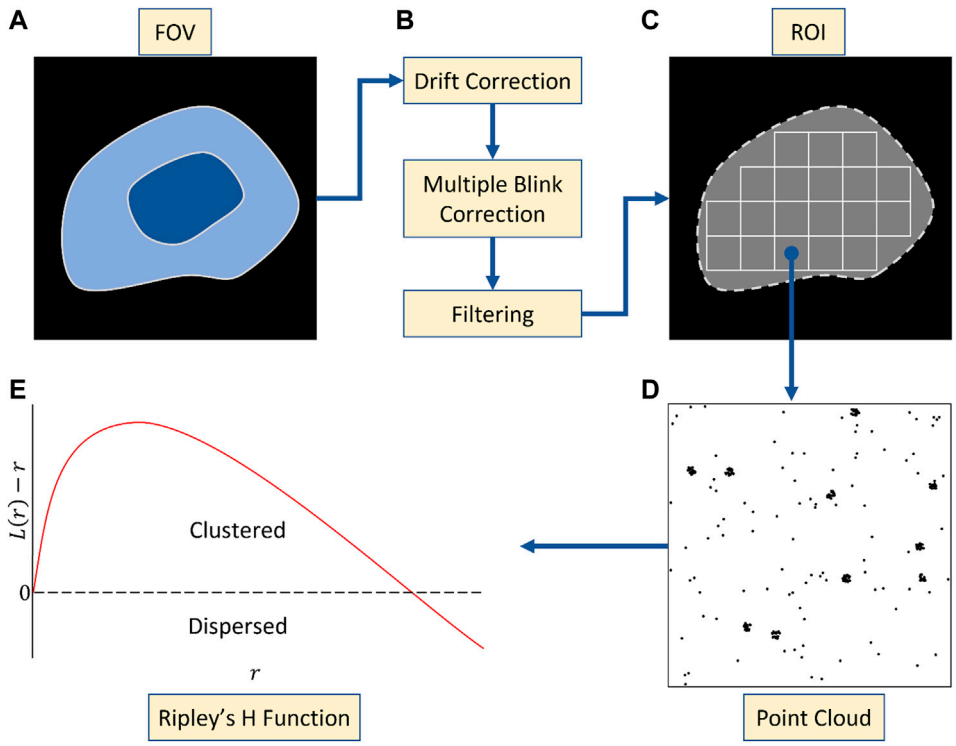

FIGURE 1
Pre-processing workflow. **(A)** Whole cell acquisition across a field of view (FOV). **(B)** Artefact correction following acquisition including drift correction, multiple blinking correction and filtering. **(C)** Segmentation of FOV into ROIs of fixed size. **(D)** Point clouds arising from one such ROI. **(E)** Use of the Ripley's H function to determine the degree of clustering across the ROI. Non-negativity across radial values suggests clustering at that scale.

Is cluster analysis needed?

Randomness in biology is hardly ever uniform, but rather consist of heterogenous distribution in which each point has an equal probability to be at any location in the ROI, also called complete spatial randomness (CSR). CSR can manifest in SPPs in which clustering like structures can be seemingly present. When looking for clusters, it is therefore crucial to avoid fitting your data and analysis to your prior and hypothesis. An efficient way to differentiate a CSR from a clustered distribution is to use Ripley's K curve analysis (Kiskowski et al., 2009; Owen et al., 2010). These curves are calculated by averaging over the point population and ROI, the local density for various scales and are today implemented as pre-made functions in most coding interfaces. It provides a very robust statistical tool to differentiate with high confidence if the protein distribution studied is CSR or clustered and hence suited to further cluster analysis (assuming multiple blinking has been accurately corrected for). Ripley's K curves however do not provide cluster identification and visualisation or detailed information on clusters' composition and sizes. In cases where multiple blinking cannot be easily addressed (e.g., STORM), the clustering landscape may still be accessible through experimental SMLM approaches as suggested by Arnold et al. (2020).

Finally, most available cluster analysis tools focus on circular shapes and are overall unsuitable for multiscale or shape independent segmentation. As a result, virtually all available cluster analysis tools are unsuitable to quantify filamentous structures, and only some can accommodate for elongated aggregates or rings (Ester et al., 1996; Pike et al., 2020). Whilst novel dedicated approaches have started to emerge for filamentous mesh quantification in SMLM data sets (Peters et al., 2018), it overall remains an active topic of research.

Optimise and report user defined parameters

There are a wide variety of cluster analysis algorithms available, even when only considering those that have been tested and validated for use on SMLM data (Khater et al., 2020). While they all have advantages and disadvantages compared to each other, one property that almost all share is the use of user-defined analysis settings. These are numbers that the user must enter into the algorithm to dictate what kind of structural features should be highlighted in the data. A common necessity is for two parameters—one somehow related to the spatial scale of objects of interest and the other related to the density of points within the clusters, i.e., is the user looking for big or small clusters and are they looking for sparse or dense clusters? Naturally, the choice of these parameters can strongly influence the output of the analysis. In some sense, there is no right or wrong answer to the choice because which clusters in the data are most relevant depends on the biological questions being asked. A minimum requirement is therefore to simply report the choice of parameters when describing the method so the results can be reproduced. However, if something about the data is known *a priori*, it is possible to optimise the choice of analysis parameters. For example, using a success or performance metric, analysis can be performed while scanning the values of analysis parameters and the best performing parameters chosen for continued use. This can be done especially if one can simulate data that closely recapitulates the experimental case (Nieves et al., 2023) or if prior knowledge can be summarised about the expected clustering properties (Rubin-Delanchy et al., 2015). Furthermore, Nieves et al. (2023) provides a detailed performance assessment on the vast majority of the cluster analysis tools described in this minireview in order to help user identifying which algorithm may be best suited to their data sets.

Don'ts

Analyse blindly

Most researchers will typically seek a completely automated analysis pipeline. This not only saves time but there is also a perception that it reduces user bias if a human makes no decisions within any particular analysis. While these are worthy goals, SMLM clusters analysis algorithms are not yet capable enough to warrant this level of confidence. A frequent occurrence in SMLM are unexpected features in the data sets. These might be real but rare biological structures, misplaced artefacts of the sample such as fiducial markers or other contaminants or unexpected artefacts of the imaging and analysis such as cell edges, uncorrected drift or sparse data. Inputting such data sets into the algorithms will produce meaningless results and bias biological conclusions. All images undergoing analysis should therefore be inspected visually to ensure the data structure is compatible with the proposed analysis. In addition, several algorithms exist which can help the user locate and assess potential data artefacts and allow them to therefore perform analysis with confidence. These include HAWKMAN (Marsh et al., 2021) and SQUIRREL (Culley et al., 2018).

Treat results as absolute quantification

SMLM data sets are artifact prone. This results both from the sample preparation stage and the processing stage. For sample preparation, most SMLM acquisitions rely on immunolabelled sample or transfection. Both strategies come with sampling issues. At the processing stage, STORM does not have reliable strategies to account for multiple blinking to date and, for both PALM or STORM, a subset of the fluorophore population will not be detected at all. For all these reasons any quantification extracted from cluster analysis tools should be treated as relative rather than absolute. Typically, this translates into using the term "localisation" when talking about the clusters' composition, rather than protein. Cluster analysis tools are thus suited for relative comparison in between conditions rather than the description of the exact protein composition of identified aggregates. If absolute quantification is required for the biological issue at hand, there are today emerging

experimental and statistical means to by-pass this issue with PAINT (Simoncelli et al., 2020) but they remain very low throughput.

Discussion

Overall, SMLM is a powerful tool for obtaining the precise locations of membrane proteins on the cell surface. However, to derive biologically meaningful conclusions such as describing the nanoscale clustering of those proteins, that data must be processed and analysed. We propose that users adopt a standardised analysis pipeline for their analysis which is broken down into a number of stages. First, data curation in which imaging artefacts such as drift and multiple-blinking can be corrected and data formatted into standardised ROIs. Second, data validity. A visual inspection of the ROIs to identify artefacts and the use of image quality algorithms will ensure data passed down the pipeline is valid. In particular Ripley's K-function should be used to confirm the presence of clustering. Third: Cluster analysis. Using prior or preliminary data, data analysis parameters should be optimised and reported and the results subjected to a secondary visual inspection. Finally, interpretation. Keeping in mind that none of the above steps can be completed perfectly. Each will add some uncertainty and bias to the final output and will sit on top of the artefacts arising during sample preparation, imaging and localisation. As a general rule therefore, users should be wary of treating outputs as absolute and SMLM cluster analysis is most powerfully utilised to compare between experimental conditions.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. DDLS fellowship (K&A Wallenberg foundation) granted to JG. EPSRC Centre for Doctoral Training in Topological Design granted to LP.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Annibale, P., Vanni, S., Scarselli, M., Rothlisberger, U., and Radenovic, A. (2011). Quantitative photo activated localization microscopy: unraveling the effects of photoblinking. *PLoS One* 6 (7), e22678. Epub 2011 Jul 26. PMID: 21818365; PMCID: PMC3144238. doi:10.1371/journal.pone.0022678

Annibale, P., Vanni, S., Scarselli, M., Rothlisberger, U., and Radenovic, A. (2011). Identification of clustering artifacts in photoactivated localization microscopy. *Nat. Methods* 8 (7), 527–528. PMID: 21666669. doi:10.1038/nmeth.1627

Arnold, A. M., Schneider, M. C., Hüsson, C., Sablatnig, R., Brameshuber, M., Baumgart, F., et al. (2020). Verifying molecular clusters by 2-color localization microscopy and significance testing. *Sci. Rep.* 10, 4230. doi:10.1038/s41598-020-60976-6

Baumgart, F., Arnold, A. M., Leskovar, K., Staszek, K., Fölser, M., Weghuber, J., et al. (2016). Varying label density allows artifact-free analysis of membrane-protein nanoclusters. *Nat. Methods* 13 (8), 661–664. Epub 2016 Jun 13. PMID: 27295310; PMCID: PMC6404959. doi:10.1038/nmeth.3897

Bohrer, C. H., Yang, X., Thakur, S., Weng, X., Tenner, B., McQuillen, R., et al. (2021). A pairwise distance distribution correction (DDC) algorithm to eliminate blinking-caused artifacts in SMLM. *Nat. Methods* 18 (6), 669–677. Epub 2021 May 31. PMID: 34059826; PMCID: PMC9040192. doi:10.1038/s41592-021-01154-y

Culley, S., Albrecht, D., Jacobs, C., Pereira, P. M., Leterrier, C., Mercer, J., et al. (2018). Quantitative mapping and minimization of super-resolution optical imaging artifacts. *Nat. Methods* 15, 263–266. doi:10.1038/nmeth.4605

Ester, M., Kriegel, H., Sander, J., and Xu, X. (1996). "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," in Proc. of 2nd International Conference on Knowledge Discovery, Portland, Oregon, 2-4 August 1996, 226–231.

Griffié, J., Burn, G., and Owen, D. M. (2015). The nanoscale organization of signaling domains at the plasma membrane. *Curr. Top. Membr.* 75, 125–165. Epub 2015 Apr 15. PMID: 26015282. doi:10.1016/bs.ctm.2015.03.004

Griffié, J., Shannon, M., Bromley, C. L., Boelen, L., Burn, G. L., Williamson, D. J., et al. (2016). A Bayesian cluster analysis method for single-molecule localization microscopy data. *Nat. Protoc.* 11 (12), 2499–2514. PMID: 27854362. doi:10.1038/nprot.2016.149

Jensen, L. G., Hoh, T. Y., Williamson, D. J., Griffié, J., Sage, D., Rubin-Delanchy, P., et al. (2022). Correction of multiple-blinking artifacts in photoactivated localization microscopy. *Nat. Methods* 19 (5), 594–602. PMID: 35545712. doi:10.1038/s41592-022-01463-w

Khater, I. M., Nabi, I. R., and Hamarneh, G. (2020). A Review of Super-Resolution Single-Molecule Localization Microscopy Cluster Analysis and Quantification Methods. *Patterns (N Y)* 1 (3), 100038. PMID: 33205106; PMCID: PMC7660399. doi:10.1016/j.patter.2020.100038

Kiskowski, M. A., Hancock, J. F., and Kenworthy, A. K. (2009). On the use of Ripley's K-function and its derivatives to analyze domain size. *Biophys. J.* 97 (4), 1095–1103. PMID: 19686657; PMCID: PMC2726315. doi:10.1016/j.bpj.2009.05.039

Lelek, M., Gyparaki, M. T., Beliu, G., Schueder, F., Griffié, J., Manley, S., et al. (2021). Single-molecule localization microscopy. *Nat. Rev. Methods Prim.* 1, 39. doi:10.1038/s43586-021-00038-x

Levet, F., Hosy, E., Kechkar, A., Butler, C., Beghin, A., Choquet, D., et al. (2015). SR-Tesseler: a method to segment and quantify localization-based super-resolution microscopy data. *Nat. Methods* 12 (11), 1065–1071. Epub 2015 Sep 7. PMID: 26344046. doi:10.1038/nmeth.3579

Marsh, R. J., Costello, I., Gorey, M. A., Ma, D., Huang, F., Gautel, M., et al. (2021). Sub-diffraction error mapping for localisation microscopy images. *Nat. Commun.* 12, 5611. doi:10.1038/s41467-021-25812-z

Nieves, D. J., Pike, J. A., Levet, F., Williamson, D. J., Baragilly, M., Oloketuyi, S., et al. (2023). A framework for evaluating the performance of SMLM cluster analysis algorithms. *Nat. Methods* 20 (2), 259–267. Epub 2023 Feb 10. PMID: 36765136. doi:10.1038/s41592-022-01750-6

Owen, D. M., Rentero, C., Rossy, J., Magenau, A., Williamson, D., Rodriguez, M., et al. (2010). PALM imaging and cluster analysis of protein heterogeneity at the cell surface. *J. Biophot.* 3, 446–454. doi:10.1002/jbio.200900089

Peters, R., Griffié, J., Burn, G. L., Williamson, D. J., and Owen, D. M. (2018). Quantitative fibre analysis of single-molecule localization microscopy data. *Sci. Rep.* 8 (1), 10418. doi:10.1038/s41598-018-28691-5

Pike, J. A., Khan, A. O., Pallini, C., Thomas, S. G., Mund, M., Ries, J., et al. (2020). Topological data analysis quantifies biological nano-structure from single molecule localization microscopy. *Bioinformatics* 36 (5), 1614–1621. doi:10.1093/bioinformatics/btz788

Rubin-Delanchy, P., Burn, G. L., Griffié, J., Williamson, D. J., Heard, N. A., Cope, A. P., et al. (2015). Bayesian cluster identification in single-molecule localization microscopy data. *Nat. Methods* 12 (11), 1072–1076. Epub 2015 Oct 5. PMID: 26436479. doi:10.1038/nmeth.3612

Simoncelli, S., Griffié, J., Williamson, D. J., Bibby, J., Bray, C., Zamoyska, R., et al. (2020). Multi-color Molecular Visualization of Signaling Proteins Reveals How C-Terminal Src Kinase Nanoclusters Regulate T Cell Receptor Activation. *Cell Rep.* 33 (12), 108523. ISSN 2211-1247. doi:10.1016/j.celrep.2020.108523