ORIGINAL PAPER

# Method for co-cluster analysis in multichannel single-molecule localisation data

Jérémie Rossy · Edward Cohen · Katharina Gaus · Dylan M. Owen

Accepted: 28 February 2014 / Published online: 19 March 2014 © Springer-Verlag Berlin Heidelberg 2014

Abstract We demonstrate a combined univariate and bivariate Getis and Franklin's local point pattern analysis method to investigate the co-clustering of membrane proteins in two-dimensional single-molecule localisation data. This method assesses the degree of clustering of each molecule relative to its own species and relative to a second species. Using simulated data, we show that this approach can quantify the degree of cluster overlap in multichannel point patterns. The method is validated using photo-activated localisation microscopy and direct stochastic optical reconstruction microscopy data of the proteins Lck and CD45 at the T cell immunological synapse. Analysing co-clustering in this manner is generalizable to higher numbers of fluorescent species and to three-dimensional or live cell data sets.

**Keywords** Cluster analysis · Super-resolution · Co-localisation · PALM · STORM

J. Rossy · K. Gaus (🖂) Centre for Vascular Research and Australian Centre for NanoMedicine, University of New South Wales, Sydney, Australia e-mail: k.gaus@unsw.edu.au

#### E. Cohen Department of Mathematics, Imperial College London, London, UK

#### D. M. Owen (🖂)

Department of Physics and Randall Division of Cell and Molecular Biophysics, King's College London, London, UK e-mail: dylan.owen@kcl.ac.uk

#### Introduction

Super-resolution fluorescence imaging methods have created enormous excitement in recent years since they break the resolution barrier imposed by optical diffraction. Among these methods are a class of single-molecule localisation microscopy (SMLM) techniques such as photo-activated localisation microscopy (PALM) (Betzig et al. 2006) and (direct) stochastic optical reconstruction microscopy ((d)STORM) (Rust et al. 2006; Heilemann et al. 2008). Here, signal from closely spaced fluorophores, which would otherwise be unresolvable, are separated temporally from each other by employing photo-activatable or photo-switchable fluorophores (e.g. fluorescent proteins in PALM) or by exploiting long-lived dark states to cycle molecules on and off (dSTORM). Once a sparse set of molecules has been activated, for example by using a very low-powered UV-activating laser, individual molecules are imaged on a sensitive CCD camera and the molecules are subsequently bleached before a new subset of molecules are activated and the process repeated. The raw images are, therefore, sets of individual, approximately Gaussian, point spread functions (PSFs), each one derived from a single molecule. The centre of each PSF can then be found by fitting a two-dimensional Gaussian function; a process that can typically achieve an uncertainty on the localisation of individual molecules of only 20-30 nm. Repeating this process for all molecules in the data set results in a list of the x and y coordinates of all the fluorescent molecules within the field-of-view. In order to generate good signalto-noise, the illumination is often provided in the form of a total internal reflection fluorescence microscope evanescent wave that restricts the axial dimensions of the imaged volume to below around 100 nm, meaning the final detected coordinates are typically treated as two-dimensional point patterns, although recently schemes for the three-dimensional acquisition, and analysis of data have been developed (Huang et al. 2008; Pavani et al. 2009; Xu et al. 2012).

Unlike other forms of super-resolution microscopy imaging such as stimulated emission depletion microscopy (STED) (Hell and Wichmann 1994; Vicidomini et al. 2011), the data derived from SMLM is not in the form of a conventional fluorescence image-that is, a pixelated array of intensity values. Instead, the data consist of a list of the x and y coordinates of individual molecules, each with an individual associated uncertainty, which is usually calculated theoretically based on the photon count, PSF width and background (Thompson et al. 2002). It is, therefore, necessary to develop new techniques to analyse such data sets and quantify cellular structures. One of the most common requirements is to detect molecular clustering since biological function can be inferred from protein complexes. Analysis methods for clustering include Ripley's K-function (Owen et al. 2010; Ripley 1977) and the related technique of pair-correlation analysis (Sengupta et al. 2011; Veatch et al. 2012), both of which can provide information on the length scale of molecular clustering. Using local point pattern analysis such as Getis and Franklin's (G&F) method (Getis and Franklin 1987), it is possible to construct 'maps' of molecular clustering from which a richer variety of cluster parameters such as shape can be extracted (Rossy et al. 2013; Williamson et al. 2011). This method has been used extensively to characterise clustering of proteins at the cell surface, in particular during T cell signalling in which the formation of protein microclusters is a key feature. The method was also recently demonstrated for three-dimensional data sets (Owen et al. 2013).

In the G&F method, the number of molecules within a specified distance of each molecule is counted. This number is normalised by the total molecular density of the analysed region, meaning each molecule can be assigned a clustering value, termed the L value, according to the relative local molecular density at that position, which is independent of the overall average molecular density. This is useful as it accounts for variations such as the number of collected frames or the protein expression level. These values, together with the x and y coordinates, can then be interpolated to generate a two-dimensional cluster 'heat map' to which a threshold can be applied in order to extract key cluster statistics. There have been a number of recent reports of algorithms designed to test for co-localisation in SMLM data sets all of which offer a number of advantages and disadvantages. These include the cross-variant of the pair-correlation (PC) method previously discussed and the coordinate-based co-localisation analysis (CBC) of Malkusch et al. (Sengupta et al. 2011; Malkusch et al. 2012; Annibale et al. 2012). The PC analysis method has the advantage of intrinsically accounting for the multiple blinking (re-excitation) of fluorescent molecules whereas here, this blinking is corrected for before the co-localisation analysis. Our method also allows co-localisation to be measured without fitting cross-correlation curves and requiring calibration data for the number of blinks per molecule. The method presented here is more similar to the CBC method, however, the use of a specific search radius rather than a weighted distribution function allows the user to use a biologically driven hypothesis to interrogate the data for specific scale clusters. In addition this, coupled with the application of a cluster threshold, allows intuitive parameters such as the number of molecules of one species overlapping with clusters of the other to be extracted. Here, we show that when analysing two-colour SMLM data, a combined univariate and bivariate G&F analysis cannot only detect clustering of each species individually as described above but also quantify the co-clustering behaviour of the two species. This is demonstrated using both simulated and cellular data.

#### Method

Values of L(r) are calculated for each species by counting the number of its own species contained within a circle of radius r nm centred on each molecule. Similarly, values of L(r)cross are calculated for each species by counting the number of the opposite species contained within a circle of radius r nm centred on each molecule (Fig. 1). The value of L(r) or L(r)cross is given by:

$$L(r)_j = \sqrt{A \sum_{i=1}^n \left(\frac{\delta_{ij}}{n}\right) / \pi} \quad \text{where } \delta_{ij} = \begin{cases} 1 \text{ if } d_{ij} < r \\ \text{else } 0 \end{cases}$$

where A is the analysed area, n is the total number of points in that area and  $d_{ii}$  is the distance between point *i* and point *j*. In this example, clustering will be examined on a spatial scale of r = 50 nm. To account for edge effects, a 50 nm buffer region was calculated around each region which was cropped out from the final analysis. The user-selected value for r (in this case 50 nm) determines the size scale of structures which the point pattern is interrogated for by the analysis algorithm. In this example, therefore, the method will search for co-localisation of points on 50 nm length scales. Smaller length scales can be selected, however, since the localisation precision of each point is of the order of 20–30 nm, this sets the lower limit for length scales that can be investigated. The method can also be applied to larger length scales, for example to detect cluster co-localisation on 100 or 200 nm length scales that may be of interest depending on the biological application. Pseudocoloured cluster maps were generated by interpolating



**Fig. 1** Calculations of L(50) and L(50)cross. L(50) for each of the *green* and *red* species are calculated by counting the number of the same species within a 50 nm radius while ignoring the second species. L(50)cross is calculated by counting the number of the opposite species within a 50 nm radius while ignoring the self-species

the L(r) values onto a 5 nm resolution grid as previously described (Owen et al. 2010). Linear trend lines in the scatter plots were generated using linear least squares fitting (MATLAB).

Clustered data for each channel was simulated by generating 100 points per cluster in a Gaussian distribution overlaid with a completely spatially random (CSR) background (400 points). Four cases were considered—that clusters of the two species completely overlap, that clusters of each species completely avoid each other, that one species is clustered but the other is a CSR distribution or two CSR distributions.

To demonstrate the method with real data, two cases were examined. Jurkat T cell synapses were generated on coverslips coated with anti-CD3 and anti-CD28 antibodies according to previously published protocols suitable for SMLM imaging (Rossy et al. 2013). Jurkat T cells were cultured in RPMI1640 media supplemented by 10 % foetal calf serum at 37 °C in a 5 % CO<sub>2</sub> atmosphere. Cells were dropped onto glass coverslips coated with anti-CD3 and anti-CD28 and allowed to form synapses for 10 min. Synapses were then fixed for 10 min at 37 °C using a 4 % paraformaldehyde solution. To demonstrate high co-localisation, cells were transfected with a fluorescent fusion construct of the kinase Lck simultaneously fused to the photo-switchable fluorescent protein PS-CFP2 and the

photoactivatable protein PA-mCherry using an Invitrogen NEON transfection system. As a negative co-localisation control, Lck-PS-CFP2 was imaged with the phosphatase CD45 labelled with Dylight 639 by immunostaining as previously described (Rossy et al. 2013). For Dylight 639 labelling, cells were fixed and permeabilized with 100 µg/ml lysolecithin (Sigma). Labelling was then performed using an anti-CD45 primary antibody (rabbit, cat# 10559, 1:200, Abcam) followed by a DyLight 639 labelled anti-rabbit antibody (cat# 111-495-047, 1:100, Jackson ImmunoResearch). To achieve blinking of DyLight 639 for dSTORM imaging, cells were placed in an oxygenscavenging, reducing buffer composed of 25 mM HEPES, 25 mM glucose, 5 % glycerol, 0.05 mg/ml glucose oxidase and 0.025 mg/ml horseradish peroxidase supplemented with 75 mM cysteamine (all from Sigma).

Cells were imaged on a Zeiss Elyra PS-1 PALM microscope with a 1.46 NA objective with 20 % of 488 nm laser power, 15 % of 561 nm and 0.5 % of 405 nm for PS-CFP2 and PA-mCherry and 100 % of 641 nm for Dylight 639. For PS-CFP2 and PA-mCherry, the signal was split and collected on two EMCCD cameras simultaneously (Andor iXon). Dylight 639 and PS-CFP2 images were recorded sequentially. Channel alignment was performed with Zen (Zeiss) using at least three 200 nm gold colloids per image.

Molecular blinking from re-excited molecules can cause cluster artefacts in single-molecule microscopy. Small organic molecules are known to undergo repeated cycles of activation during dSTORM imaging, and the fluorescent protein mEos2 was recently shown to also display reexcitation behaviour (Annibale et al. 2010), which is likely also the case for PS-CFP2. To overcome these effects, we applied the off-gap method in which molecules that are reexcited in the same location within a certain time (the offgap) are grouped and counted as the same molecule. Here, the total number of localisations in a field-of-view is plotted as a function of off-gap for a representative data set. As the off-gap increases, the total number of localisations falls as individual blinks are grouped into a single localisation. The optimal off-gap is then selected as the value where a further increase no longer has any significant effect in decreasing the total number of localisations. This method has been shown to be effective in removing re-excited molecules from further analysis (Annibale et al. 2011a, b). In this case, the selected off-gap was 300 ms for PS-CFP2 and PA-mCherry and 1 s for DyLight 639.

## **Results and discussion**

We first analysed four simulated data sets consisting of one data set with overlapping and one data set with excluding clusters on a background of randomly distributed



Fig. 2 Univariate G&F analysis applied to four simulated data sets overlapping clusters, excluding clusters, clusters paired with a random distribution and two random distributions. a Images showing the locations of individual-simulated molecules, pseudo-coloured cluster

molecules, a data set where one species is clustered and the other randomly distribution and a data set where both species are randomly distributed (Fig. 2). Using only the standard univariate G&F analysis, cluster maps were generated, pseudo-coloured and merged (Fig. 2a). These were then analysed using Pearson's correlation coefficient as is often used in standard two-colour fluorescence microscopy data (Bolte and Cordeliere 2006) (Fig. 2b). In the case of perfectly overlapping clusters, this method generates a high Pearson's correlation coefficient of 0.74. However, for the case of excluding clusters where a negative correlation coefficient would be expected, only a very small negative value is generated (-0.03). Clusters of one channel against a CSR distribution in the other generate a positive value of 0.18 whereas two random distributions resulted in a value close to zero (0.02). Hence, this method is biased towards detecting co-clustering. In addition, it does not generate data on the number of molecules that reside within clusters of the other species.

Next, we calculated the L(50) and L(50)cross values for each molecule and plotted these as a scatter plot. Figure 3 is taken from a representative data set of a T cells expressing Lck with two different fluorescent proteins. As previously maps of the *red* and *green* channels and merged images showing the degree of cluster overlap (*yellow*). Areas represent  $2 \times 2 \,\mu$ m simulated regions **b** Pearson's correlation coefficient of the four distributions



**Fig. 3** Illustration of a combined univariate and bivariate G&F analysis at a spatial scale of r = 50 nm for Lck-PS-CFP2 and Lck-PA-mCherry transfected into T cells. L(50) represents clustering within the same species and L(50)cross clustering with the second species. The pseudo-colour represents the number of individual molecules at each position. Clustering thresholds of L(50) and L(50)cross of 50 are indicated by *black lines* that segments the plot into four quadrants (A-D)



Fig. 4 Combined univariate and bivariate G&F analysis of four simulated possible cluster configurations showing the scatter plots for each channel together with trendline

reported, Lck is not randomly distributed (Rossy et al. 2013) in these cells resulting in L(50) values above random distributions. Here, we plotted the L(50) and L(50)crossvalues in such a manner that the x-axis (L(50)) denotes how clustered a molecule is with its own species (Lck-PS-CFP2 in this case) and the y-axis (L(50)cross) shows to what extent it is within a cluster of the opposite species (Lck-PS-CFP2 with Lck-PA-mCherry in this case). The colour code denotes the number of molecules with the same L(50) and L(50) cross value. As previously demonstrated, these values can be thresholded to categorise molecules as either being in or out of clusters. In this example, for Lck co-transfected into T cells as either a PS-CFP2 (green) or a PA-mCherry (red) construct, a threshold of 50 is applied to L(50) and L(50) cross in order to divide the plot into four quadrants. Note that like the value for r, which determines the size scale of clusters that are examined, the value of the threshold applied to the data is also a user-defined parameter. This selection determines the value of the local molecular density above which molecules are considered as being inside clusters. Selecting a lower threshold will assign more molecules to clusters, and less dense objects will be counted as clusters. Conversely, a high threshold will only select the densest regions as being clusters. The value for the threshold used here is simply an example—whether less dense objects (lower threshold), or only high density areas (high threshold) should be selected depends on the biologically informed hypothesis.

Figure 3 shows these data for the PS-CFP2 molecules only. For this data set, 47 % of green molecules exist in green clusters (Quadrants B and D) whereas 39 % of green molecules exist within red clusters (C + D). Where these overlap, 17 % of green molecules exist simultaneously in both green and red clusters (Quadrant D). Fifty-two per cent of green molecules exist in green OR red clusters (B + C), whereas 69 % exist in green AND/OR red clusters (B + C + D). This leaves the remainder, 31 %, which are not in green or red clusters (A).

We next applied this method to the four simulated data sets previously examined (Fig. 4). In the case of



Fig. 5 Cluster maps and scatter plots for T cells transfected with Lck-PS-CFP2-PA-mCherry (a) or Lck-PS-CFP2 and CD45-Dylight 639 (b)

overlapping clusters, applying a threshold of L(50) = 200revealed that 46 % of red molecules were found in red clusters (quadrant B + D) (simulated data was generated with 50 % of red molecules in clusters). With this threshold, 46 % of red molecules were also found in green clusters (quadrant C + D), consistent with there being complete overlap. Specifically, 45 % of molecules were in quadrant D, indicating they were associated simultaneously with both red and green clusters, 54 % were not associated with clusters at all leaving only 1 % which were associated with red OR green clusters (B + C). Similar results are found when examining the statistics for the green molecules. The co-clustering can also been observed in the significant positive correlation (trendline) between the values of L(50) and L(50)cross.

In the case of excluding clusters, the method correctly determined that no red molecules (0 %) exist simultaneously in both red and green clusters (D). Red molecules were still detected in red clusters (28 %) (B + D), however, only a small fraction (1 %) was found where the red CSR background overlaps with green clusters (C). Also, in this case, no positive correlation between the values of L(50) and L(50) cross is observed in the trendline. Similar results are found when examining the statistics for the green molecules. In the third case where green molecules were clustered and red ones randomly distributed, only green molecules were found to be clustered, whereas red molecules were not (Quadrant B). In this case, the scatter plots for the red and green channels were now asymmetric. In the green case, high self-clustering was still observed above L(50) = 200 (Quadrant B), however, no green molecules were found in red clusters and indeed quadrants C + D contained 0 % of green molecules. For the red CSR case, Quadrant B contained 0 % of molecules as expected, however, 2 % of the red CSR distribution happened to overlap with the green clusters (C). In the final case of two CSR distributions, all points were correctly found in Quadrant A and the plots for the red and green channel were again similar.

Finally, we demonstrate the method in T cells. Figure 5a shows data derived from the tandem construct Lck-PS-CFP2-PA-mCherry. The individual cluster maps from the PS-CFP2 and PA-mCherry channels are also shown. These two maps have a Pearson's correlation coefficient of 0.05, indicating again the difficulty of applying this analysis to SMLM data. Also shown are the univariate and bivariate scatter plots for the PS-CFP2 (green) channel and the PAmCherry (red) channel. Using a threshold of L(50) = 80, 7 % of PS-CFP2 molecules reside simultaneously in both PS-CFP2 and PA-mCherry clusters (D). The number is similar (6 %) when analysing the PA-mCherry species. Five per cent of PS-CFP2 molecules reside in clusters of one OR another colour (B + C). The value for PA-mCherry is 4 %. The degree of cluster overlap is also illustrated by the positive correlation between L(50) and L(50) cross shown in the trendline.

Next, we examined data derived from Lck-PS-CFP2 and CD45 labelled with Dylight 639 (Fig. 5b). This showed that the Pearson correlation coefficient was also 0.05. However, it was found that only 2 % of PS-CFP2 molecules reside

simultaneously in both PS-CFP2 and Dylight 639 clusters (D), almost thrice lower than the Lck-PS-CFP2-PAmCherry case. These represent just 10 % of the total clustered PS-CFP2 molecules. For the Dylight 639 molecules. 2 % reside simultaneously in pS-CFP2 and Dylight 639 clusters. In this case, the trendline does not show any significant correlation between L(50) and L(50) cross. While the absolute numbers of co-localised molecules may seem low, in interpreting these data, one ought to consider that it is unlikely that two fluorescent proteins or indeed two fluorophores have the identical detection probability. Therefore, co-clustering of a fusion protein containing two fluorophores also reflects the relative detection probability of these fluorophores (including the probabilities that both are folded correctly, generate sufficient photons to be detected and so on). It should be noted that these values also depend on the cluster assignment of each species individually (and therefore on the choice of threshold) and on the spatial scale being discussed—on what length scale are they co-localised (and therefore depends on the choice of r). In the case of PS-CFP2 + DyLight 639, the physical size of the primary and secondary antibody must be considered together with the labelling efficiency of both antibodies. Hence, Lck-PS-CFP2 co-clustering with CD45 Dylight 639 has to be interpreted in the context of Lck-PS-CFP2-PAmCherry. Taking into account that the clustering of Lck-PS-CFP2-PA-mCherry is relatively low (A + B + C = 93)and 94 % for the green and red channels, respectively), we can conclude that Lck and CD45 co-clustering is lower than Lck clustering with self. Hence, it is unlikely that Lck and CD45 cluster together.

#### Conclusion

Super-resolution SMLM microscopy methods such as PALM and dSTORM generate pointillistic data sets of molecular coordinates rather than traditional fluorescence intensity images. These data sets, therefore, have to be statistically interrogated in order to detect features such as molecular clusters. For single channel data, clustering has previously been analysed using pair-correlation techniques (Sengupta et al. 2011; Sengupta and Lippincott-Schwartz 2012), Ripley's K-function (Owen et al. 2010) or Getis and Franklin's local point pattern analysis (Rossy et al. 2013; Williamson et al. 2011). Here, we have demonstrated a combined univariate and bivariate version of Getis and Franklin's analysis to quantify the degree of co-clustering in two channel point pattern data. The method was applied to simulated data that showed either co-localised clusters, excluding clusters or CSR distributions and was found to be more reliable in detecting co-clustering than the use of Pearson's correlation coefficient applied to the individual channel cluster maps. The method was also applied to data derived from activated T cells transfected with fluorescent fusion constructs of the kinase Lck and the phosphatase CD45. In principle, the method is extensible to more than two fluorescent species, to three-dimensional data sets and to live cell PALM experiments.

Acknowledgments D.M.O. is supported by a Marie Curie Career Integration Grant (CIG) Ref 334303. KG is supported by the National Health and Medical Research Council of Australia and the Australian Research Council.

### References

- Annibale P et al (2010) Photoactivatable fluorescent protein mEos2 displays repeated photoactivation after a long-lived dark state in the red photoconverted form. J Phys Chem Lett 1:1506–1510
- Annibale P et al (2011a) Identification of clustering artifacts in photoactivated localization microscopy. Nat Methods 8(7):527–528
- Annibale P et al (2011b) Quantitative photo activated localization microscopy: unraveling the effects of photoblinking. PLoS ONE 6(7):e22678
- Annibale P et al (2012) Identification of the factors affecting co-localization precision for quantitative multicolor localization microscopy. Opt Nanoscopy 1(9):1–13
- Betzig E et al (2006) Imaging intracellular fluorescent proteins at nanometer resolution. Science 313:1642–1645
- Bolte S, Cordeliere FP (2006) A guided tour into subcellular colocalization analysis in light microscopy. J Microsc 224:213–232
- Getis A, Franklin J (1987) Second-order neighbourhood analysis of mapped point patterns. Ecology 68(3):473–477
- Heilemann M et al (2008) Subdiffraction-resolution fluorescence imaging with conventional fluorescent probes. Angew Chem Int Ed 47(33):6172–6176
- Hell SW, Wichmann J (1994) Breaking the diffraction resolution limit by stimulated emission: stimulated-emission-depletion fluorescence microscopy. Opt Lett 19(11):780–782
- Huang B et al (2008) Three-dimensional super-resolution imaging by stochastic optical reconstruction microscopy. Science 319(5864):810–813
- Malkusch S et al (2012) Coordinate-based colocalization analysis of single-molecule localization microscopy data. Histochem Cell Biol 137(1):1–10
- Owen DM et al (2010) PALM imaging and cluster analysis of protein heterogeneity at the cell surface. J Biophotonics 3(7):446–454
- Owen MD et al (2013) Quantitative analysis of three-dimensional fluorescence localization microscopy data. Biophys J 105(2):L05–L07
- Pavani SRP et al (2009) Three-dimensional, single-molecule fluorescence imaging beyond the diffraction limit by using a double-helix point spread function. Proc Natl Acad Sci 106(9):2995–2999
- Ripley BD (1977) Modelling spatial patterns. J R Stat Soc B 39:172–192
- Rossy J et al (2013) Conformational states of the kinase Lck regulate clustering in early T cell signalling. Nat Immunol 14(1):82–89
- Rust MJ, Bates M, Zhuang X (2006) Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). Nat Methods 3(10):793–796
- Sengupta P, Lippincott-Schwartz J (2012) Quantitative analysis of photoactivated localization microscopy (PALM) datasets using pair-correlation analysis. BioEssays 34(5):396–405

- Sengupta P et al (2011) Probing protein heterogeneity in the plasma membrane using PALM and pair correlation analysis. Nat Methods 8(11):969–975
- Thompson RE, Larson DR, Webb WW (2002) Precise nanometer localization analysis for individual fluorescent probes. Biophys J 82(5):2775–2783
- Veatch SL et al (2012) Correlation functions quantify super-resolution images and estimate apparent clustering due to over-counting. PLoS ONE 7(2):e31457
- Vicidomini G et al (2011) Sharper low-power STED nanoscopy by time gating. Nat Methods 8(7):571–573
- Williamson DJ et al (2011) Pre-existing clusters of the adaptor Lat do not participate in early T cell signalling events. Nat Immunol 12(7):655–662
- Xu K, Babcock HP, Zhuang X (2012) Dual-objective STORM reveals three-dimensional filament organization in the actin cytoskeleton. Nat Methods 9(2):185–188