SPHERICAL SPATIAL STATISTICS FOR 3D FLUORESCENCE VIDEO-MICROSCOPY

A.Chessel, J.Dodgson, R.E.Carazo-Salas

Gurdon Insitute, Cambridge university, Cambridge, UK

ABSTRACT

Recent work has revealed that, contrary to expectations, the protein machineries responsible for regulating cellular polarity in fission yeast cells localises cortically to the semi-spherical tips of each cell as punctual structures. To quantify this discovery and try and reconcile it with previous findings, a statistical co-localisation framework is set up. We use spatial statistics to analyse collections of punctual objects detected from deconvolved wide-field fluorescence microscopy images as point patterns from their respective spatial point processes. The theory is described for spherical data, as the structures of interest reside on a half-sphere in 3D. Uses include test of complete independence with Monte-Carlo hypothesis testing and comparison between pairs of point collections using bootstrap hypothesis testing. Evaluation on simulated data are performed on top of the targeted biological studies.

***Index Terms*—** Fluorescence microscopy, Co-localisation, Spatial Statistics, Schizosaccharomyces pombe, Cell biology, Cell Polarity, Fission yeast

1. INTRODUCTION

Cell polarity refers to the capacity of cells to behave in an anisotropic and spatially differentiated manner. It is a fundamental cell process and its understanding is crucial in a number of biomedical applications. The unicellular fission yeast *S. pombe* is a perfect model organism to study polarity as its cells grow cylindrically-shaped from their tips in a highly polarised manner. Polarity factors, the proteins involved in regulating polarity, are known to localize to the cell cortex at the tips of cells and have traditionally been thought to do so in a mostly homogeneous manner (Fig 1). Unexpectedly, our recent work imaging the areas of cell polarity 'head-on' (frontally) [1] revealed a striking spatial organization of polarity factors in punctual ('node') structures (Fig 1 right). Moreover, by simultaneously imaging pairs of polarity factors, we have found that some pairs thought to be part of a same complex may have a more complicated spatial relationship (Fig 2, left). This work concerns itself with the development of a well defined framework allowing for the study of the co-localisation of those pairs of polarity factors at fission yeast cell tips.

Fluorescence co-localisation, i.e. the spatial relationship between two proteins expressed in cells as seen by fluorescence imaging, has been the focus of much work since the onset of two colour fluorescence microscopy. Object based co-localisation techniques, describing the co-localisation of punctual or localised structures, classically include the computation of various metrics [2], such as overlapping area, proximity based ratio, etc. However, in part because of the variety of experimental protocols and biological questions they aim to answer, co-localisation still is a study dependent problem.

This work was funded by a European Research Council Starting Independent Researcher Grant (ERC SYSGRO) and by a Human Frontier Science Program Young Investigator Grant (HFSP RGY0066/2009-C).

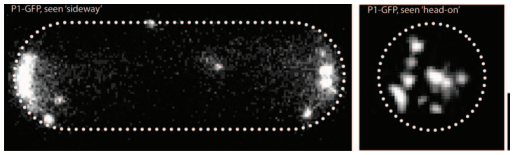


Fig. 1. A *S. pombe* cell expressing a polarity factor fused to Green Fluorescent Protein (GFP), imaged sideways (left) and 'head on' (right)

Point processes can be defined informally as random point pattern generators. Their study is initially tied to ecological data studies that have focused on the spatial distribution of species of animals or plants. The theoretical foundation of spatial statistics, linking point processes to measure theory, is well defined and consists now of a collection of tools adapted to the study of fluorescence microscopy images [3]. These methods have been previously applied to biological problems [4, 5, 6], including the study of signalling receptor spatial distribution [7], and recently the use of models to study spatial interaction [8] . Given the increased expressivity of point process models and the availability of a large array of statistical tools, spatial statistics are a good candidate for an object-based co-localisation framework grounded in theory.

The data of interest lives at the cell tip cortex, which is in first approximation a semi-spherical surface. In the first two sections we will develop and adapt the necessary point process theory for that particular case, adjusting for spherical coordinates and geodesic distance. The last two sections will be devoted to answering the two questions we are interested in. First whether two node populations are significantly correlated in space, using Monte-Carlo hypothesis testing. Secondly how to quantify the relative co-localisation of two pairs of protein nodes, using bootstrap to evaluate the variability within each dataset and assess whether they are significantly different. Additional evaluation is provided by fully artificially simulated point patterns with controlled known spatial correlation.

2. SPHERICAL POINT PATTERNS OF PROTEINS IN 3D

Point patterns and point processes A point pattern P in a window W is simply a collection of points $X_i,\ i = 1 \dots n$ in W. A spatial point process is a random point pattern. A rigorous presentation of the point process theory is out of the scope of this paper but can be found in [3] and references therein. For our purpose, we will define a point process as a random counting measure on bounded sets.

Definition 1 (Point process)**.** A point process in a space S is defined by a random counting measure N, i.e. a collection of random

variables $N(B)$ for all bounded set $B \subset S$ such that $N(B)$ is a nonnegative finite integer.

The idea is not to randomly draw points, but to equivalently randomly draw the number of points present in any possible set. One of the simplest examples of a point process is the spatial Poisson process, which we will use later to define spatially independent point patterns.

Definition 2 (Inhomogeneous Poisson Process)**.** The inhomogeneous Poisson process on $\mathbb{R}^2$ with density λ, $\mathcal{P}_\lambda$, is a point process such that:

- for any bounded compact set $B \in \mathbb{R}^2$, $N(B)$ is Poisson distributed with mean $\int_B \lambda dS$
- if $B_1 \dots B_m$ are disjoint bounded sets in $\mathbb{R}^2$ then the point counts $N(B_1) \dots N(B_m)$ are independent

If the density λ is uniform, we simply have that the average point count is $\hat{N}(B) = \lambda \times \text{area}(B)$. It can then be shown that the Poisson process is equivalent to the intuitive random point process obtained by drawing points with a uniform probability on their coordinates.

Point process on a sphere While these definitions are generic, some care needs to be taken when dealing with actual data on a sphere. Indeed classical spherical coordinates, which for a half sphere amount to the radius r, the latitude $\theta \in [0 \dots \pi/2]$ and the longitude $\phi \in [0 \dots 2\pi]$, are biased, i.e. the area of a surface element d^2S depends on its position on the sphere.

Proposition 1. *The area of a surface element d^2S is unbiased with respect to the modified coordinates u, ϕ, with $u = cos(\theta)$.*

The surface element is $d^2S = r^2 \sin(\theta) d\theta d\phi$, which is dependent on θ. We have $dud\phi = sin(\theta) d\theta d\phi$. Thus $d^2S = r^2 dud\phi$ is independent of the position on a sphere of given radius. □

Hence defining a uniform Poisson process in u, ϕ will indeed result in uniformly scattered points on a sphere, and consequently an inhomogeneous Poisson process defined with densities in u, ϕ will work as expected. From here on, only point patterns in u, ϕ will be considered; we will now see how our experimental data of interest can be expressed in that framework.

From images to point patterns Two polarity factor proteins of interest were tagged with GFP and mCherry (green-fluorescent and red-fluorescent fluorophores, respectively) and endogenously expressed in a S. pombe cell strain. Two-colour fluorescence images of cell tips were imaged in 3D (x,y,z) using an OMX wide-field epifluorescence microscope (Applied Precision, Issaquah), and image stacks were systematically deconvolved. A $4.5 \times 4.5 \mu$m square was manually cropped around each cell tip of interest, resulting in $57 \times 57 \times 17$ pixel 3D volumes. Object detection was performed using wavelet thresholding [9], implemented in the ICY open source platform [10], giving us two sets of (x,y,z) coordinates of the detected nodes, one in each colour. Let us call, for a given experiment, $T_i = (T_i^g, T_i^r)$ $i = 1 \dots n$ the coordinates of the green and red nodes detected respectively for the n cell tips. Those coordinates were imported in the R statistical environment where the rest of the computations were carried out, using in particular the SpatStat package [11].

The θ, ϕ coordinates, and hence u, ϕ coordinates, were obtained for each cell tip T_i by fitting a half-sphere on each collection of points and projecting them on it individually (Fig. 2, right). Each protein can be modelled by a inhomogeneous Poisson process in u, ϕ space, with the individual density estimated using kernel density estimation. Kernel size is preset and kept constant. (Fig 3).

To evaluate the proposed algorithms, we simulated point patterns of predictable spatial correlations. A first channel was drawn randomly with the second channel being generated based on the first. The first channel was drown from an inhomogeneous Poisson process in u, ϕ space, modelled after one particular protein. The points of the second channel were the points of the first, randomly moved by one step of a diffusion process of set diffusion constant σ around their original location.

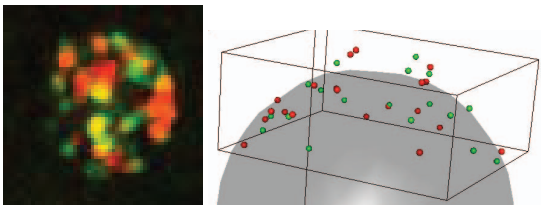


Fig. 2. (left) Two-colour cell tip showing two polarity factors. (right) Detected points in 3D (red and green) and the fitted sphere (gray)

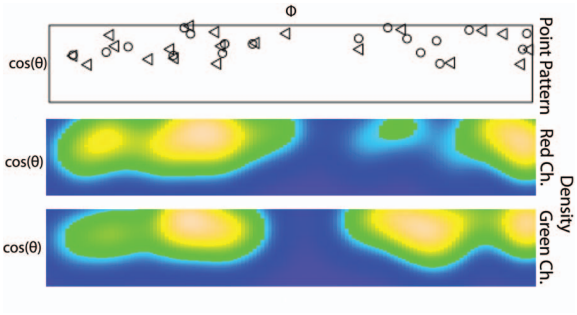


Fig. 3. (above) A point pattern shown in $cos(\theta)$, ϕ coordinates. (below) Computed densities via kernel density estimation

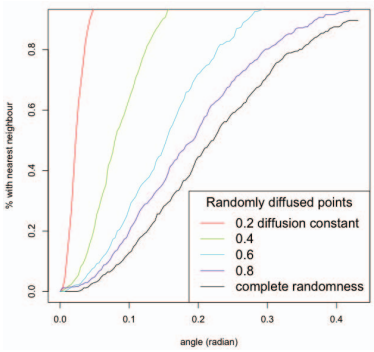


Fig. 4. Correlation in randomly generated point patterns of various known correlations, parametrised by a set diffusion constant

3. SUMMARY STATISTICS ON A SPHERE

Summary statistics using geodesics The use of summary statistics to study point patterns is well established and a number of them have been proposed [3, 11]. We will only consider here summary statistics across channels. The best used is the K function, it seemed

however not well suited to our dataset as it had a tendency to diverge on individual tips. We will use instead the so-called G function, the cumulative distribution of the distance to the nearest neighbour. As a rigorous definition, involving conditional probabilities on point processes, is outside the scope of this paper, we will pose $G_{\mathbf{X}}(r) = \mathbb{P}\{d(x, \mathbf{X}^g) < r | x \in \mathbf{X}^r\}$, with $d(x, \mathbf{X})$ the distance between a point x and the point of the point pattern $\mathbf{X}$ nearest to it. Replicated point patterns, i.e. several experimental point patterns corresponding to the same data, have been relatively less studied than single point patterns. Here, we will pull the various point patterns together and compute G as

$$G_{\mathbf{X}}(r) = \mathbb{P}\{d(x, \mathbf{X}_i^g) < r | i \in [1 \dots n]\ x \in \mathbf{X}_i^r\}.$$

The empirical nearest neighbour distance distribution, defined as $G(r) = \frac{1}{|\sum \mathbf{X}_i^r|} \sum_i \sum_{x_j \in \mathbf{X}_i^r} \mathbb{1}\{d(x_j, \mathbf{X}_i^g < r\}$ is a biased estimator of G because of edge effects. An edge corrected estimator, provided by Spatstat, is used in practice.

An adaptation to spherical data involves replacing the Euclidian distance with geodesics on the sphere. In u, ϕ (or, equivalently for distances, θ, ϕ) coordinates, the geodesic distance is $d(x_i, x_j) = \arccos(\sin(\theta_1)\sin(\theta_2)\cos(\phi_1 - \phi_2) + \cos(\theta_1)\cos(\theta_2))$. Thus, a spherical $G_{\mathbf{X}}$ is defined by replacing the distance in the above definition. Let us note that, as $G_{\mathbf{X}}$ is defined using the nearest neighbours, any distance that conserves neighbour order would be equivalent; the Euclidian distance in 3D could have been used as well.

Use on real data G_X is first evaluated on the simulated data (Fig 4, multicolour curves). As expected, the curves are steeper and steeper as the two point patterns are more correlated. Fig 5 shows in black the experimental results on five pairs of proteins. For simplicity and confidentiality, they will be named P1 through P6. The results are varied and seem to span the whole range of possible correlations, from very high correlation of P1-P2 and P1-P3 to very low correlation of P1-P6. However, drawing any kind of conclusion seems problematic without a reference.

4. MONTE-CARLO HYPOTHESIS TESTING OF SPATIAL INDEPENDENCE

To test for independence between the point patterns of the two channels, Monte-Carlo simulations will be used.

Theory The null hypothesis is "$\mathcal{H}_0$: the point patterns of the two channels are statistically independent". The idea of Monte-Carlo hypothesis testing is to use simulations of data following the null hypothesis to compute its distribution. Here, uncorrelated data will be simulated by independently drawing two populations of points according to two inhomogeneous Poisson processes.

Specifically, let $T_i = (T_i^g, T_i^r),\ i = 1 \dots n$ be the u, ϕ coordinates of the detected points of a two colour experiment in n cell tips. The per channel density of points $\lambda_i = (\lambda_i^g, \lambda_i^r),\ i = 1 \dots n$ is computed using kernel density estimation. The null hypothesis for that dataset is then modelled by the collection of point processes $\mathcal{P}_i = (\mathcal{P}_{\lambda_i^g}, \mathcal{P}_{\lambda_i^r}),\ i = 1 \dots n$. A first set of m collections of point patterns are generated according to those point processes and their summary statistic $G_{\mathbf{X}}$ is evaluated to estimate the average G_{avg} for uncorrelated point patterns for that dataset. A second set $X_i,\ i = 1 \dots m$ of m collections of n point patterns are then generated, and their summary statistic curves $G_i, i = 1 \dots m$ are used to compute $g_{max} = \max_i \max_r |G_{avg} - G_{\mathbf{X}_i}|$.

We then say that an experimental collection of point patterns $T_i,\ i = 1 \dots n$, assumed to be independent and identically distributed from the same point process do not follow $\mathcal{H}_0$ if, for any r, $G_T(r)$ is outside of the range $[G_{avg} + g_{max},\ G_{avg} - g_{max}]$.

Several polarity factors complexes coexist at the tip The results on the five protein pairs of interest are shown Fig. 5. We see that P1-P6 are essentially uncorrelated in space at the cell tip, which is expected as they are quite distant in terms of cell function. For the 4 other couples, the observed curve statistic falls outside the grey area, hence they all are spatially correlated.

P1-P2 and P1-P3 are very correlated (the experimental curve is far from random), and correspondingly a lot of actual co-alignment of the nodes can be seen. This is expected as both of those proteins have been shown to interact with P1 and depend on it for their localisation. The correlation of the two other pairs P1-P4 and P1-P5 seems weaker, i.e. the observed statistic seems closer to randomness; correspondingly, in the images the nodes do not seem to actually co-align with one another. P1 and P4 have been shown to interact and to be co-dependent for their localisation. P1 and P5 have not been shown to interact with one another, but P5 depend on P1 for its localisation and they both interact with P3.

The fact that node populations can be closer together than true randomness would allow, but without being actually on top of one another, would suggest a more complex spatio-temporal mechanism, that could include unseen partners or partial or dynamical interactions. Those comparisons are still quite qualitative however and a quantitative way to compare the spatial correlation of two protein pairs is needed.

5. COMPARING POINT PATTERNS USING BOOTSTRAP HYPOTHESIS TESTING

Comparing comparable point patterns The first point to clarify is whether two point patterns are comparable, as the function G used for comparison is obviously very dependent on number and density of points.

Definition 3 (Comparable point patterns)**.** Two pairs of point patterns $(T_i^1)_{i=1\dots n_1}$ and $(T_i^2)_{i=1\dots n_2}$ are said to be comparable if their corresponding *uncorrelated* point processes, modelled using inhomogeneous Poisson processes, can be seen as identically distributed, i.e. if their nearest neighbour cumulative distance distributions G_{avg}^1 and G_{avg}^2 fall within each other's simulated variation band as defined above.

This ensures that the number and density of points for each pair are similar, and hence that eventual differences in their observed nearest neighbour distributions actually correspond to differences in relative co-localisation. To test whether those differences are significant, and given that acquiring additional experimental images is usually a limiting factor, we use bootstrapping resampling.

Let $T^1 = (T_i^1)_{i=1\dots n_1}$ and $T^2 = (T_i^2)_{i=1\dots n_2}$ be two comparable pairs of point patterns. Let $\mathcal{H}_0$ be "T^1 and T^2 are samples from the same population". Let $\hat{\tau} = \max_r |G_{T^1}(r) - G_{T^2}(r)|$ be the test statistic under scrutiny. A bootstrap sample $B = (B^1, B^2)$ is comprised of two sets of respectively n_1 and n_2 tips randomly drawn with replacement out of $T^1 \cup T^2$. If G_{B^1} and G_{B^2} are the respective cumulative distance distribution corresponding to the bootstrap sample, we pose $\tau_B = \max_r |G_{B^1}(r) - G_{B^2}(r)|$.

Then, if m bootstrap samples B_i are drawn, the p-value of the test is defined as ([12]) $p = \frac{1}{m} \sum_i \mathbb{1}_{\tau_{B_i} > \hat{\tau}}$. A low p-value means that the value of the observed statistic τ of samples following $\mathcal{H}_0$, i.e where all tips are pulled together and assumed to be identically distributed, are most often lower than the observed one; that value then is not adequately described by the null distribution and $\mathcal{H}_0$ can be rejected.

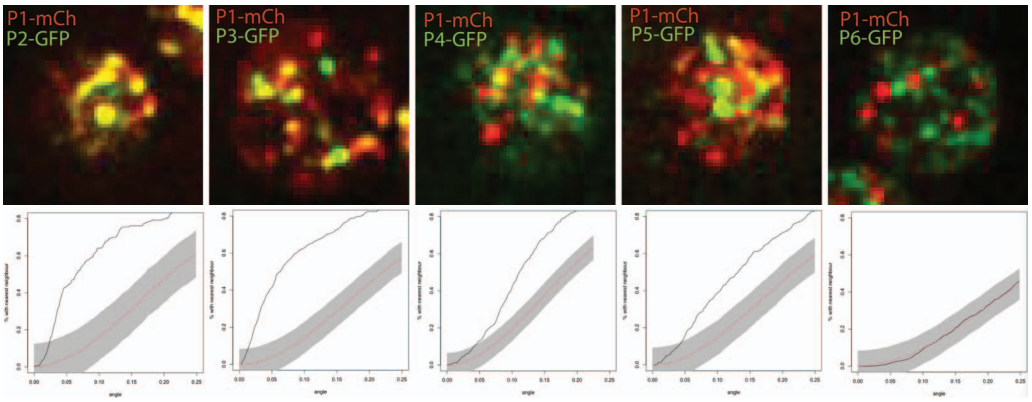


Fig. 5. (above) Example of 5 pairs of proteins, named P1 to P6 for simplicity and confidentiality. (below) G function for at least 20 cells (black line), with the corresponding uncorrelated simulated point patterns (red and gray).

Comparing polarity pairs On the simulated point patterns, the curves compared two by two $\sigma = 0.2$ and $\sigma = 0.4$, $\sigma = 0.4$ and $\sigma = 0.6$, $\sigma = 0.6$ and $\sigma = 0.8$ are all seen as significantly different ($p < 0.01$). When compared to uncorrelated point patterns, randomly displaced point patterns are seen as significantly different at a 5% significance for $\sigma \leq 0.8$. Those results do validate the proposed test, and seem to indicate that its power is similar across the range of correlations.

It can readily be seen that P1-P6 is not comparable to the other pairs. The number of P6 nodes, in particular, is significantly below the others. The other pairs are all comparable however. P1-P2 and P1-P3 are of equivalent spatial correlation ($p = 0.375$), and are both significantly more correlated than P1-P4 ($p < 0.001$), corroborating the qualitative comparison of the previous section. The fact that different couples of proteins can be of significantly different, but non-random, spatial correlation is again a hint at the presence of several differing underlying physical process giving rise to them.

6. DISCUSSION AND CONCLUSION

In this work, a spatial statistic framework was adapted to 3D fluorescence microscopy image data and used to study the co-localisation of punctual protein structures at the tip of yeast cells. Using that framework we proved the existence of multiple distinct populations of protein nodes at the cortex of fission yeast cells, hinting at the existence of complex spatio-temporal mechanisms. To elucidate those mechanisms, future work will be aimed at extending the current framework to use time information for the study of video microscopy data, careful use of more involved point pattern interaction models and/or the characterisation of sub-populations of cells within a given protein pair.

The developed framework provides a quantitative and statistical way to deal with object co-localisation, a crucial yet still challenging problem. Importantly, it does so without imposing ad hoc models, but using non-parametric data-based modelling. As a number of biological processes involve the spatio-temporally relative position of punctual structures, the proposed methodology may be considered as a good candidate for a generic analysis framework in microscopy.

7. REFERENCES

[1] J. Dodgson, A Chessel, and al., manuscript submitted.

[2] S. Bolte and F. P. Cordelières, "A guided tour into subcellular colocalization analysis in light microscopy.," *Journal of microscopy*, vol. 224, no. Pt 3, pp. 213–32, Dec. 2006.

[3] A. Baddeley, "Spatial Point Processes and their Applications," in *Lecture Notes in Mathematics: Stochastic Geometry 1892*, vol. 3, pp. 1–75. 2007.

[4] T. Mattfeldt, "Explorative statistical analysis of planar point processes in microscopy.," *Journal of microscopy*, vol. 220, no. Pt 3, pp. 131–9, Dec. 2005.

[5] E. Díaz, Guillermo A., M. E. Díaz, L.-W. Gong, and D. Toomre, "Automatic detection of large dense-core vesicles in secretory cells and statistical analysis of their intracellular distribution.," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 7, no. 1, pp. 2–11, 2010.

[6] B. Zhang, N. Chenouard, J.-C. Olivo-Marin, and V. Meas-Yedid, "Statistical Colocalisation in Biological Imaging with False Discovery Control," in *IEEE ISBI*, 2008, pp. 1327 – 1330.

[7] J. Zhang, K. Leiderman, J.R. Pfeiffer, B.S. Wilson, J.M. Oliver, and S.L. Steinberg, "Characterizing the topography of membrane receptors and signaling molecules from spatial patterns obtained using nanometer-scale electron-dense probes and electron microscopy.," *Micron (Oxford, England : 1993)*, vol. 37, no. 1, pp. 14–34, Jan. 2006.

[8] J.A. Helmuth, G. Paul, and I.F. Sbalzarini, "Beyond co-localization: inferring spatial interactions between sub-cellular structures from microscopy images.," *BMC bioinformatics*, vol. 11, pp. 372, Jan. 2010.

[9] J.C. Olivo-Marin, "Extraction of spots in biological images using multiscale products," *Pattern Recognition*, vol. 35, no. 9, pp. 1989–1996, Sept. 2002.

[10] F. de Chaumont, S. Dallongeville, and J.-C. Olivo-Marin, "ICY: a new open-source community image processing software," in *IEEE ISBI*, 2011.

[11] A. Baddeley and R. Turner, "spatstat: An R Package for Analyzing Spatial Point Patterns," *Journal of statistical software*, vol. 12, no. 6, 2005.

[12] B. Efron and R. Tibshirani, *An introduction to the bootstrap*, Chapman & Hall, 1993.