A STATISTICAL ANALYSIS OF SPATIAL COLOCALIZATION USING RIPLEY'S K FUNCTION

Thibault Lagache [*]*, Vannary Meas-Yedid, Jean-Christophe Olivo-Marin*

Institut Pasteur, Quantitative Image Analysis Unit, F-75015 Paris, France
CNRS URA 2582, F-75015 Paris, France

ABSTRACT

Proteins colocalization in fluorescence microscopy is a key quantitative tool to decipher cellular processes at a molecular level. Most colocalization analysis are based on intensity correlation between different colour channels, which can either lead to colocalization overestimates when proteins point spread functions (PSF) are large, or mis-colocalization when signals of spatially close proteins do not strictly overlap. Similarly, methods based on Monte-Carlo simulations are very time consuming and are generally out of the reach of biological labs. In this paper, we present a new object-based method that is both fast and easy to implement. Using an asymmetric Ripley based statistic, we develop an analytical method that permits statistical quantification of proteins colocalization and accounts asymptotically for ROI boundaries. Tests against Monte-Carlo simulations and synthetic data show that our method is both sensitive and specific.

Index Terms— Colocalization, Ripley's K function, Asymptotic analysis, Object-based method.

1. INTRODUCTION

In fluorescence microscopy, colocalization proteins reveals the molecular organization of cellular processes. For example, the analysis of colocalization of viral proteins with markers of specific cellular compartments such as endosomal Rabs is commonly used to decipher early stages of pathogen entry into cells [1]. Similarly, molecular orchestration of clathrin mediated endocytosis has been revealed by the analysis of spatio-temporal colocalization of different proteins involved [2]. There is an increasing need for quantitative approaches in colocalization studies to reject non-significant colocalization coming from randomly distributed proteins that are close each other by chance.

Most of existing colocalization methods are based on the spatial overlap between the (denoised) signal that is emitted from two (or more) different fluorescent labels. In particular, **intensity-correlation-based methods** propose a global image similarity coefficient that measures pixel coincidences, and compute some correlation score of the intensity values in a dual-channel image. Common scores include Pearson's [3] and Manders' coefficients [4]. Yet, these methods present some limitations such as strong dependence on the PSF width and on the denoising method. For example, using wide and overlapping PSF will lead to false positive colocalizations. Conversely, reducing PSF with super-resolution methods can lead to missing spatially close but not overlapping signals. Consequently, **object-based methods**, that first segment and identify objects (proteins) with elaborate detection algorithms such as wavelet-based methods [5] or patch-based methods [6], and then account for objects inter-distances to analyze possible colocalization, are now increasingly developed ([7], [8], [9], [10], [11]). However, it remains difficult to discriminate a real *versus* a false positive colocalization. Indeed, proteins can be close to each other just by chance (null hypothesis), through their spatially random distribution and determining a level of statistical significance in colocalization analysis has become a key methodological issue leading to many publications in both intensity-correlation-based methods ([3]) and object-based methods ([7], [9]). In both cases, the null hypothesis model in which the distribution of the distance (respectively overlap) between two independently randomly drawn proteins spots (respectively pixel blocks) is obtained with extensive Monte-Carlo simulations in the specific region of interest (ROI). However, these simulations depend on ROI shape and new computations are required for each given ROI. In addition, due to high computational time, they cannot be used on large set of images, or in real-time colocalization assessment.

Hereafter, we present new analytical tools that account for ROI shapes and provide closed form formula for statistical levels of significance in object-based colocalization, with no need for Monte-Carlo simulations. More precisely, we first define in sub-section 2.2 a statistics $\tilde{K}_{12}$ based on Ripley's K function, capturing information about objects inter-distances. We then show in sub-section 2.3 the convergence of $\tilde{K}_{12}$ towards the normal law when the number of objects is sufficiently large and develop a new asymptotic analysis to account for ROI boundaries in sub-section 2.4. We further quantify the minimum number of objects that are needed to verify the asymptotic normality of $\tilde{K}_{12}$ in sub-section 2.5. Finally, in section 3, we test our analytical tool against Monte-Carlo simulations and perform analysis on synthetic data where proteins are either partially colocalized or randomly distributed. We find that $\tilde{K}_{12}$ is both specific and sensitive, detecting accurately either the null hypothesis of proteins random distribution or their partial colocalization.

2. A COLOCALIZATION STATISTICS BASED ON RIPLEY'S K FUNCTION

2.1. Using the Ripley's K function to measure spatial colocalization: state of the art

Most object based methods use Ripley's K function to study objects spatial distribution, whose standard expression for a single population of n objects $\mathbf{x}$ in ROI Ω is

$$K(r) = \frac{|\Omega|}{n(n-1)} \sum_{\mathbf{x} \neq \mathbf{y}} \mathbf{1}_{\{|\mathbf{x}-\mathbf{y}| \leq r\}} f(\mathbf{x}, \mathbf{y}), \tag{1}$$

[*]Corresponding author: thibault.lagache@pasteur.fr. T.L.'s and V. M.-Y.'s research is supported by Institut Pasteur, PTR 387.

Table 2. Statistics on synthetic data, $n_2 = 100$

	$\alpha = 0$		$\alpha = 0.2, \sigma = 0.1$		$\alpha = 0.2, \sigma = 0.3$		$\alpha = 0.5, \sigma = 0.1$		$\alpha = 0.5, \sigma = 0.3$	
	$\tilde{K}_{12}(r)$	p-value	$\tilde{K}_{12}(r)$	p-value	$\tilde{K}_{12}(r)$	p-value	$\tilde{K}_{12}(r)$	p-value	$\tilde{K}_{12}(r)$	p-value
$r = 0.3$	-0.57	0.72	**4.34**	$\mathbf{7 \times 10^{-6}}$	1.49	0.07	**8.46**	$\mathbf{< 10^{-16}}$	**4.44**	$\mathbf{4.5 \times 10^{-6}}$
$r = 0.5$	-1.33	0.90	1.97	2.5×10^{-2}	**2.42**	$\mathbf{7.8 \times 10^{-3}}$	**4.54**	$\mathbf{2.78 \times 10^{-6}}$	**5.24**	$\mathbf{7.9 \times 10^{-8}}$
$r = 1.0$	-1.56	0.94	0.96	0.17	**2.52**	$\mathbf{5.8 \times 10^{-3}}$	**3.16**	$\mathbf{7.88 \times 10^{-4}}$	**2.96**	$\mathbf{1.5 \times 10^{-3}}$

Table 3. Statistics on synthetic data, $n_2 = 1000$

	$\alpha = 0$		$\alpha = 0.2, \sigma = 0.1$		$\alpha = 0.2, \sigma = 0.3$		$\alpha = 0.5, \sigma = 0.1$		$\alpha = 0.5, \sigma = 0.3$	
	$\tilde{K}_{12}(r)$	p-value	$\tilde{K}_{12}(r)$	p-value	$\tilde{K}_{12}(r)$	p-value	$\tilde{K}_{12}(r)$	p-value	$\tilde{K}_{12}(r)$	p-value
$r = 0.1$	-1.24	0.89	**16.1**	$\mathbf{< 10^{-16}}$	1.90	2.9×10^{-2}	**36**	$\mathbf{< 10^{-16}}$	**7**	$\mathbf{< 10^{-16}}$
$r = 0.3$	-1.56	0.94	**12.5**	$\mathbf{< 10^{-16}}$	**5.23**	$\mathbf{8.56 \times 10^{-8}}$	**30**	$\mathbf{< 10^{-16}}$e	**14**	$\mathbf{< 10^{-16}}$
$r = 0.5$	1.01	0.16	**6.74**	$\mathbf{7.78 \times 10^{-12}}$	**5.80**	$\mathbf{3.37 \times 10^{-9}}$	**19**	$\mathbf{< 10^{-16}}$	**16**	$\mathbf{< 10^{-16}}$
$r = 1.0$	1.27	0.10	**5.44**	$\mathbf{2.63 \times 10^{-8}}$	**4.20**	$\mathbf{1.33 \times 10^{-5}}$	**12**	$\mathbf{< 10^{-16}}$	**13**	$\mathbf{< 10^{-16}}$

Table 1. Test against Monte Carlo simulations

	$q_{0.99}$	$\frac{\lvert q_{0.99} - z_{0.99}\rvert}{z_{0.99}}$	$q_{0.999}$	$\frac{\lvert q_{0.999} - z_{0.999}\rvert}{z_{0.999}}$
$n = 1$	2.40	3%	3.13	1.5%
$n = 10$, uniform	2.43	4.5%	3.32	7.4%
$n = 10$, cluster	2.37	2%	3.19	3%

where $f(\mathbf{x}, \mathbf{y})$ is a boundary correction term. The main goal of these estimators is to test whether a given point distribution is a realization of a homogeneous Poisson process, that is, for any given subset $A \subset \Omega$, $\Pr\{\mathbf{x} \in A\} = n|A|/|\Omega|$. A lack of theoretical results associated with bias that is induced by ROI edges complicate the analysis and most studies are actually based on Monte-Carlo simulations [12]. Yet, some recent theoretical results on the asymptotic normality of Ripley's K function [13] for large n, coupled with the estimation of their mean and variance accounting for edges effects [13, 14] pave the way to analytical tests of uniform distribution.

In multivariate cases, when m species A_i, for $1 \leq i \leq m$, are spatially distributed in ROI Ω, Ripley's K cross function $K_{ij}(r)$ have been extensively used to study spatial colocalization between points sets A_i and A_j [8, 9]. In the bivariate case, when A_1 and A_2 are both homogeneous Poisson processes, arguments of [13] apply, demonstrating that K_{12} is asymptotically normal as n_1 and n_2 tends to infinity. In addition, formula for the mean $\mathbb{E}\{K_{12}\}$, and the variance $var\{K_{12}\}$ when Ω is a rectangle, have been derived [15]. However, there is neither general formula for unspecific shape of Ω, nor for an arbitrary spatial distribution of A_1. The latter case is a very important practical issue in cellular biology where proteins are rarely uniformly distributed inside the whole cytoplasm or nucleus but rather confined to cellular micro-domains. Thus, given positions of A_1 points, with no hypothesis on their spatial distribution, it is important to test unilaterally whether some A_2 points appear to be close to A_1 points juste by chance, A_2 being a realization of a homogeneous Poisson process, or if this proximity is statistically relevant, revealing molecular interactions.

Consequently, we build hereafter a Ripley based, unilateral estimator $\tilde{K}_{12}(r)$ to test whether the vicinity between A_1 and A_2 is statistically relevant, with no hypothesis on A_1 spatial distribution. Then, we demonstrate that for any n_1, $\tilde{K}_{12}(r)$ tends in law towards the normal distribution when n_2 is sufficiently large, and we compute an asymptotic formula for $\text{var}\{\tilde{K}_{12}(r)\}$ that accounts for edge effects.

2.2. Building an asymmetric, Ripley-based statistics

We accounted for possible unobserved points at distance $|\mathbf{x}-\mathbf{y}|$ from $\mathbf{x}$ due to ROI boundaries by using isotropic Ripley's correction [14] $f(\mathbf{x}, \mathbf{y}) = \frac{|\partial b(\mathbf{x}, |\mathbf{x}-\mathbf{y}|)|}{|\partial b(\mathbf{x}, |\mathbf{x}-\mathbf{y}|) \cap \Omega|}$ that represents the inverse proportion of circumference $b(\mathbf{x}, |\mathbf{x} - \mathbf{y}|)$ that falls inside ROI Ω. Assuming that the edge of the ROI is straight where it intersects $b(\mathbf{x}, |\mathbf{x} - \mathbf{y}|)$, $f(\mathbf{x}, \mathbf{y})$ can be determined analytically [16], and is given by:

$$f(\mathbf{x}, \mathbf{y}) \approx \left(1 - \frac{1}{\pi} \arccos\left(\frac{\min(|\mathbf{x} - \mathbf{y}|, |\mathbf{x} - \partial\Omega|)}{|\mathbf{x} - \mathbf{y}|}\right)\right)^{-1}. \quad (2)$$

We highlight that the straight boundary approximation holds as soon as the ROI boundary is sufficiently smooth, ensuring that its local radius of curvature $R \gg |\mathbf{x} - \mathbf{y}|$. To test unilateraly whether A_2 spots are significantly close to A_1 spots, we then used the asymmetric Ripley's K function

$$K_{12}(r) = \frac{|\Omega|}{n_1 n_2} \sum_{\mathbf{x} \in A_1} \sum_{\mathbf{y} \in A_2} \mathbf{1}_{\{|\mathbf{x}-\mathbf{y}| \leq r\}} f(\mathbf{x}, \mathbf{y}), \quad (3)$$

and considered the reduced statistics

$$\tilde{K}_{12}(r) = \frac{K_{12}(r) - \mathbb{E}\{K_{12}(r)\}}{\sqrt{\text{var}\{K_{12}(r)\}}}, \quad (4)$$

that is centered ($\mathbb{E} = 0$) and normalized (var $= 1$).

2.3. Asymptotic normality of $\tilde{K}_{12}(r)$

We hereafter show that for any A_1 points distribution, $\tilde{K}_{12}(r)$ converges in law towards the normal distribution when $n_2 \gg 1$:

$$\tilde{K}_{12}(r) \underset{n_2 \to \infty}{\xrightarrow{\mathcal{L}}} \mathcal{N}(0, 1). \quad (5)$$

First, denoting S_k the surface of Ω that is exactly covered by k balls $b(\mathbf{x}, t)$, for $\mathbf{x} \in A_1$ and $k = 1 \dots n_1$ (see Fig. 1), and $A_{1,k} \subset A_1$ the subset of A_1 points that is such that $\bigcap_{\mathbf{x} \in A_{1,k}} b(\mathbf{x}, t) = S_k$, we decompose Ripley's K function $K_{12}(r)$ as

$$K_{12}(r) = \frac{|\Omega|}{n_1 n_2} \sum_{k=1}^{n_1} k \sum_{\mathbf{x} \in A_{1,k}} \sum_{\mathbf{y} \in A_2} \mathbf{1}_{\{\mathbf{y} \in S_k\}} f(\mathbf{x}, \mathbf{y}). \quad (6)$$

Under the hypothesis that A_2 is an homogeneous Poisson process in Ω, we have

$$\sum_{\mathbf{y} \in A_2} \mathbf{1}_{\{\mathbf{y} \in S_k\}} = \text{Bin}\left(n_2, p_k = |S_k|/|\Omega|\right), \quad (7)$$

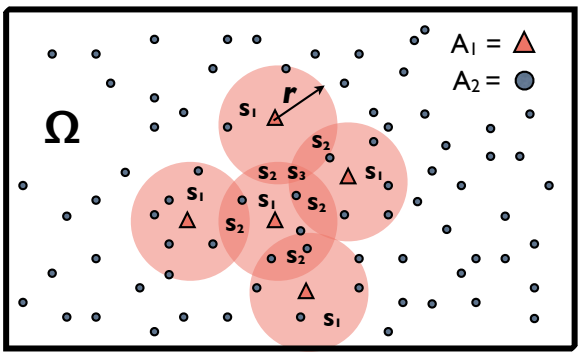

Fig. 1. Ripley's function $K_{12}(r)$ in Ω is proportional to the number of A_2 points that are in balls centered at each point of A_1 with radius r. These balls are represented in red. S_1 represents the total surface of Ω that is covered by exactly one ball, S_2 is the surface covered by the intersection of exactly 2 balls...And S_{n_1} is the surface covered by all balls. In many cases, $S_{n_1} = 0$, except when A_1 is very clustered and/or for sufficiently large r.

and we can write the following approximation

$$\sum_{\mathbf{y}\in A_2} \mathbf{1}_{\{\mathbf{y}\in S_k\}} f(\mathbf{x},\mathbf{y}) \approx \text{Bin}\,(n_2, p_k)\, \frac{\int_{S_k} f(\mathbf{x},\mathbf{u})d\mathbf{u}}{|S_k|}. \tag{8}$$

Denoting, $\delta_k(\mathbf{x}) = \frac{k}{|S_k|}\int_{S_k} f(\mathbf{x},\mathbf{u})d\mathbf{u}$, we can then rewrite Eq. 6

$$K_{12}(r) = \frac{|\Omega|}{n_1 n_2}\sum_{k=1}^{n_1}\left(\sum_{\mathbf{x}\in A_{1,k}} \delta_k(\mathbf{x})\right) \text{Bin}\,(n_2, p_k)\,, \tag{9}$$

Finally, convergence of the binomial distribution towards the normal distribution ensures that $K_{12}(r)$ is asymptotically normal as a sum of independent normal laws, and $\tilde{K}_{12}(r)$ converges towards the standard normal law $\mathcal{N}(0,1)$ as claimed in Eq.5. Thus, under the hypothesis that A_2 is uniformly distributed, and denoting q_γ and z_γ the quantiles at level γ of $\tilde{K}_{12}(r)$ and $\mathcal{N}(0,1)$ respectively, we have that

$$q_\gamma \underset{n_2\to\infty}{\longrightarrow} z_\gamma, \tag{10}$$

Consequently, for n_2 sufficiently large (we will see in subsection 2.5 what "sufficiently large" means), we can use $\tilde{K}_{12}(r)$ as a statistical test of protein colocalization: If

$$\tilde{K}_{12}(r) > z_\gamma, \tag{11}$$

then we can reject the null hypothesis of A_2 uniform distribution with a confidence level of $1-\gamma$.

2.4. Computation of var$\{K_{12}(r)\}$

Because $\mathbb{E}\{K_{12}(r)\} = \pi r^2$ ([14] page 39), $\tilde{K}_{12}(r)$ simplifies to

$$\tilde{K}_{12}(r) = \frac{K_{12}(r) - \pi r^2}{\sqrt{\text{var}\{K_{12}(r)\}}}, \tag{12}$$

and the analytical computation of var$\{K_{12}(r)\}$ for an arbitrary ROI Ω is the final step leading to a closed form expression for the colocalization statistics $\tilde{K}_{12}(r)$. Accounting for edge effects, we have that (see Appendix)

$$\text{var}\{K_{12}(r)\} = \frac{|\Omega|}{n_1^2 n_2}\left(\sum_{\mathbf{x}_1\in A_1}\beta(\mathbf{x}_1) + \sum_{\mathbf{x}_2\neq\mathbf{x}_1} A_{12}\right) - \frac{\pi^2 r^4}{n_2}. \tag{13}$$

where $\beta(\mathbf{x}_1)$ is a function of the distance $|\mathbf{x}_1 - \partial\Omega|$ of each point $\mathbf{x}_1$ to the boundary $\partial\Omega$, and A_{12} is the area of balls intersection $A_{12} = |b(\mathbf{x}_1,t)\cap b(\mathbf{x}_2,t)|$. We give a semi-analytical expression for $\beta(\mathbf{x}_1)$ in Appendix, while A_{12} is equal, for $d_{12} = |\mathbf{x}_1 - \mathbf{x}_2|$, to [17]

$$A_{12} = \mathbf{1}_{\{d_{12}<2r\}}\left(2r^2\cos^{-1}\left(\frac{d_{12}}{2r}\right) - \frac{d_{12}}{2}\sqrt{4r^2 - d_{12}^2}\right). \tag{14}$$

2.5. Convergence criterion

Berry-Essen theorem [18] ensures that asymptotic normality of Bin(n,p) is controlled by $C/\sqrt{np(1-p)}$, where C is a constant. Here, $K_{12}(r)$ is a sum of binomial processes with different probabilities $p_k = \frac{S_k}{|\Omega|}$ that depend on A_1 points inter-distances and parameter r. Thus, convergence criterion is not well defined. However, denoting $p(r) = \frac{|A_1(r)|}{|\Omega|}$ with $|A_1(r)| = |\bigcap_{\mathbf{x}_1\in A_1} b(\mathbf{x}_1,r)|$, we assume that convergence of $K_{12}(r)$ is controlled by

$$\sqrt{n_2 p(r)(1-p(r))}. \tag{15}$$

In particular, for a single A_1 point and r such that $p(r) = \frac{\pi r^2}{|\Omega|} = 0.5$, that is $r = \sqrt{\frac{|\Omega|}{2\pi}}$, we find that for $n_2 > n_2^0 \approx 30$, the relative error $\frac{|q_\gamma - z_\gamma|}{z_\gamma}$ was less than 5%. More generally, given n_1, $|\Omega|$ and r we can approximate $p(r)$, for small r and quite well separated A_1 points, with $p(r) \approx \frac{1}{|\Omega|}\left(n_1\pi r^2 - \frac{1}{2}\sum_{\mathbf{x}_i\neq\mathbf{x}_j\in A_1} A_{ij}\right)$, where A_{ij} is given by Eq.14. Then, based on convergence criterion (15), we can deduce an approximated minimum value of n_2 ensuring that $\frac{|q_\gamma - z_\gamma|}{z_\gamma} < 5\%$:

$$n_2 \geq \frac{n_2^0}{p(r)(1-p(r))}, \tag{16}$$

where $n_2^0 \approx 30$. We test convergence criterion (16) for $n_1 = 100$ A_1 points that are uniformly distributed in a square Ω with side 10 and $r = 0.3$, leading to $p(r) \approx 0.3$, and find that for $n_2 > 100$, $\frac{|q_\gamma - z_\gamma|}{z_\gamma} < 5\%$, which is in agreement with criterion (16) that predicts $n_2 > 140$.

3. TEST AGAINST MONTE-CARLO SIMULATIONS AND SYNTHETIC DATA

3.1. Test against Monte-Carlo simulations

To verify the specificity of our Ripley-based statistics when A_2 is a homogenized Poisson process, that is the accuracy of the rejection zone (11), we performe Monte-Carlo simulations for three different A_1 spatial distributions (see Fig. 2a,b and c). More precisely, we either consider that $n_1 = 1$ or $n_1 = 10$ A_1 points are uniformly distributed in Ω, which is a 10 by 10 unit square (Fig. 2a-b), or that $n_1 = 10$ A_1 points are clustered following a two dimensional Gaussian process $\mathcal{N}(\mathbf{P}, \sigma = 1)$ where $\mathbf{P}$ is a random location in Ω. r is determined such that $p(r) \approx 0.3$ that is $r = \sqrt{\frac{|\Omega|0.3}{n_1\pi}}$ when

A_1 points are uniformly drawn in Ω, and $r = \sqrt{\frac{|\Omega|0.3}{\pi}}$ in clustered condition. We then performe $N = 10^6$ Monte-Carlo simulations, where we draw uniformly, at each simulation, $n_2 = \frac{n_2^0}{p(r)(1-p(r))} \approx 140$ A_2 points in Ω and compute the corresponding $\tilde{K}_{12}^j(r)$, for $1 \leq j \leq N$. Finally, we obtain the quantile q_γ of $\tilde{K}_{12}^j(r)$ at level $\gamma = 0.99$ and $\gamma = 0.999$ by sorting the $\tilde{K}_{12}^j(r)$ and choosing

$$q_\gamma = \tilde{K}_{12}^{\lfloor \gamma N \rfloor}(r) \tag{17}$$

with $\lfloor \gamma N \rfloor$ the floor function of γN. In table 1, we compare quantiles obtained numerically with the quantiles of the standard normal law $z_{0.99} = 2.32$ and $z_{0.999} = 3.09$. As expected theoretically (see Eq. 10), q_γ is very close to z_γ in each test condition. The major discrepancy, $\frac{|q_\gamma - z_\gamma|}{q_\gamma} = 7.4\%$, is obtained for $n_1 = 10$ uniformly distributed A_1 points and $\gamma = 0.999$. In this case, $\Phi(q_\gamma) = \Pr\{\mathcal{N}(0,1) < q_\gamma\} = 0.9995$, which is still very close to 0.999.

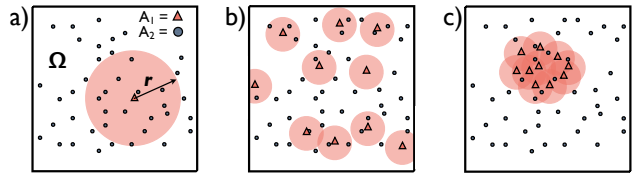

Fig. 2. We test the specificity of our statistical test $\tilde{K}_{12}(r)$ against the null hypothesis of A_2 uniform distribution by verifying the accuracy of the rejection zone (11) for 3 different A_1 patterns: a) a single A_1 point is considered, b) 10 A_1 points are uniformly distributed in Ω and c) 10 A_1 points are clustered following a two dimensional Gaussian process. In all three cases, r is determined such that $p(r) \approx 0.3$.

3.2. Test against synthetic data

Hereafter, we test the accuracy of colocalization detection with our method on synthetic data, where A_2 points are either uniformly distributed in Ω or partially colocalized with A_1 points. More precisely, we draw uniformly $n_1 = 100$ A_1 points in Ω and part αn_2 ($\alpha = 0, 0.2$ or 0.5, $n_2 = 100$ or 1000) of the A_2 points is normally distributed among A_1 points (standard deviation $\sigma = 0.1$ or 0.3), the others $(1-\alpha)n_2$ A_2 points being uniformly distributed in Ω. Points distributions that have been used in our computations are represented in the supplementary Fig. S1 of the Appendix. Standard deviation σ is used here to simulate proteins interactions with typical length scale $l \approx 2\sigma$. Indeed, for a particle $\mathbf{x}_2$ interacting with $\mathbf{x}_1$, we have that $\Pr\{|\mathbf{x}_1 - \mathbf{x}_2| < 2\sigma\} \approx 99\%$. We then measure in tables 2 and 3, $K_{12}(r)$ for $r = 0.1, 0.3, 0.5$ and $r = 1$ and deduce corresponding $\tilde{K}_{12}(r)$ with Eq. 4 and p-values $= \Phi(\tilde{K}_{12}(r))$, where Φ is the cumulative density function of the standard normal law $\mathcal{N}(0,1)$. We emphasize that $r = 0.1$ is only considered when $n_2 = 1000$ because for $r = 0.1$, $p(r) \approx n_1 \pi r^2 = 0.03$ and convergence criterion (15) imposes that $n_2 \geq 1000$. We find that $\tilde{K}_{12}$ is highly specific and sensitive, detecting accurately null hypothesis ($\alpha = 0$) and proteins colocalization even for $\alpha = 0.2$. Interestingly, we find that test accuracy increases with the number of colocalized particles (α and n_2), decreases with interaction length scale (σ) and is maximal for $r \approx l$.

In many practical applications, statistical colocalization can be performed on multiple ROIs Ω_i, $1 \leq i \leq N$, and variance of the mean statistics $\bar{K}_{12}^N(r) = \frac{1}{N}\sum_{i=1}^N \tilde{K}_{12}^i(r)$ is inversely proportional to N: $\mathrm{var}\{\bar{K}_{12}^N(r)\} = \mathrm{var}\{\tilde{K}_{12}(r)\}/N$, increasing the sensitivity of colocalization detection. We plot $\bar{K}_{12}^N(r)$, for $N = 10$ and $\alpha = 0, 0.2$ and 0.5 in the supplementary Fig. S2. Denoting $z_\gamma^N = \Pr\{\mathcal{N}(0, 1/N) \leq \gamma\}$, we observe that mean statistics $\bar{K}_{12}^N(r)$ is still specific with $z_{0.01} < \bar{K}_{12}^N(r) < z_{0.99}$ for all $0.3 < r < 1.0$ when $\alpha = 0$, and highly sensitive: $\bar{K}_{12}^N(r) \gg z_{0.99}^N$ for $\alpha = 0.2, 0.5$.

4. CONCLUSION

Quantitative colocalization is a key methodological issue in fluorescence microscopy, revealing molecular orchestration of cellular processes. We have constructed here an analytical object-based method to test statistically whether a population of detected proteins (spots) is spatially close to another population in an arbitrary ROI, accounting asymptotically for edges effects. We have tested our method against Monte-Carlo simulations and synthetic data, demonstrating the high specificity and sensitivity of our method.

5. REFERENCES

[1] M. Lakadamyali, M. J. Rust, H. P. Babcock, and X. Zhuang, "Visualizing infection of individual influenza viruses," *Proc Natl Acad Sci U S A*, vol. 100, no. 16, pp. 9280–5, Aug 2003.

[2] M. J. Taylor, D. Perrais, and C. J. Merrifield, "A high precision survey of the molecular dynamics of mammalian clathrin-mediated endocytosis," *PLoS Biol*, vol. 9, no. 3, p. e1000604, Mar 2011.

[3] S. V. Costes *et al.*, "Automatic and quantitative measurement of protein-protein colocalization in live cells," *Biophys. J.*, vol. 86, pp. 3993–4003, 2004.

[4] E. Manders *et al.*, "Dynamics of three-dimensional replication patterns during the S-phase, analysed by double labelling of DNA and confocal microscopy," *J. Cell Sci.*, vol. 103, pp. 857–862, 1992.

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

[6] J. Boulanger, A. Gidon, C. Kervran, and J. Salamero, "A patch-based method for repetitive and transient event detection in fluorescence imaging," *PLoS ONE*, vol. 5, no. 10, p. e13190, 10 2010.

[7] B. Zhang, N. Chenouard, J.-C. Olivo-Marin, and V. Meas-Yedid, "Statistical colocalization in biological imaging with false discovery control," in *ISBI*, 2008, pp. 1327–1330.

[8] G. Ayala *et al.*, "Analysis of spatially and temporally overlapping events with application to image sequences," *IEEE Trans. on PAMI*, vol. 28, no. 10, pp. 1707 –1712, oct. 2006.

[9] E. Diaz *et al.*, "Measuring spatiotemporal dependencies in bivariate temporal random sets with applications to cell biology," *IEEE Trans. on PAMI*, vol. 30, no. 9, pp. 1659 –1671, sept. 2008.

[10] E. Lachmanovich *et al.*, "Co-localization analysis of complex formation among membrane proteins by computerized fluorescence microscopy: application to immunofluorescence co-patching studies," *J. Microsc.*, vol. 212, pp. 122–131, 2003.

[11] F. Jaskolski, C. Mulle, and O. J. Manzoni, "An automated method to quantify and visualize colocalized fluorescent signals," *J. Neurosci. Meth.*, vol. 146, pp. 42–49, 2005.

[12] D. Nunez *et al.*, "Hotspots organize clathrin-mediated endocytosis by efficient recruitment and retention of nucleating resources," *Traffic*, vol. 12, no. 12, pp. 1868–78, Dec 2011.

[13] G. Lang and E. Marcon, "Testing randomness of spatial point patterns with the ripley statistic," *ESAIM: Probability and Statistics*, 2012 (accepted).

[14] B. Ripley, *Statistical inference for spatial processes*. Cambridge University Press, 1988.

[15] H. Lotwick and B. Silverman, "Methods for analysing spatial processes of several types of points," *J. Roy. Statist. Soc. B*, vol. 44, pp. 406–413, 1982.

[16] A. Getis and J. Franklin, "Second-order neighborhood analysis of mapped point patterns," *Ecology*, vol. 68, pp. 473–477, 1987.

[17] E. Weisstein. Circle-circle intersection. [Online]. Available: http://mathworld.wolfram.com/Circle-CircleIntersection.html

[18] A. Berry, "The accuracy of the gaussian approximation to the sum of independent variates," *Transactions of the American Mathematical Society*, vol. 49, no. 1, pp. 122–136, 1941.