STATISTICAL COLOCALIZATION IN BIOLOGICAL IMAGING WITH FALSE DISCOVERY CONTROL

*B. Zhang, N. Chenouard, J.-C. Olivo-Marin, V. Meas-Yedid**

Unité Analyse d'Images Quantitative Institut Pasteur 75015 Paris France

ABSTRACT

In this paper, we present a novel object-based statistical colocalization method. Our colocalization relies on multiple hypothesis tests on the distances between all pairs of the (spot-shaped) objects from the two markers. We wish to test among all these pairs how many are significantly close to each other such that they cannot occur just "by chance". Two objects are decided to be colocalized if the test on their distance is significant. For this purpose, we first extract the objects by applying a wavelet-based spot detection approach which fully takes into account the mixed-Poisson-Gaussian noise process of confocal fluorescence images. Then, we build a null hypothesis model in which the distribution of the distance between two independently randomly drawn detections in the cell is estimated by a kernel method. The observed distances are tested against this null model. Our tests control the false discovery rate (FDR) of the colocalizations. Simulations show that this approach has a good specificity. Furthermore, our method has been successfully applied in a real problem of protein colocalization analysis during the endocytic process.

***Index Terms*—** colocalization, false discovery rate, protein association

1. INTRODUCTION

Colocalization is widely considered as an important and useful analysis in cell biology. For example, colocalization of a protein with specific markers of cellular functional compartments contributes to the understanding of the role of the protein in biological processes; colocalization between proteins also has implications on their interactions. However, large biological data sets usually prohibit manual analysis of colocalization, which is tedious and unreliable.

A range of computational colocalization approaches have been proposed in the literature (see [1] for a comprehensive review), which can be classified into **1) intensity-correlation-based methods:** In these methods, some correlation score of the intensity values in a dual-channel image is calculated (the two channels will be respectively denoted as I_A and I_B). A large score implies a high colocalization degree between the two channels. Common scores include the Pearson's coefficient [2], Manders' coefficient [3], cross correlation [4], and Li's coefficient [5]. Statistical validation was also introduced in [2] through assessing the statistical significance of the score by estimating its distribution from images of randomly permuted pixel blocks. **2) object-based methods:** Unlike intensity-correlation-based methods which rely on some global image similarity measure and a pixel coincidence analysis, in the object-based approaches, the structures of interest in colocalization are explicitly explored. The objects are first segmented and identified, and the colocalization events are established by using object information such as their locations. For example, [6] considers a colocalization of two objects if their centers are separated below the microscope resolution. A similar method is studied by [7] termed the "nearest-neighbor distance" approach. A statistical validation is also carried out by computing the significance of the colocalizarion degree defined as the ratio of the number of colocalized objects and the total object number in one channel. The distribution of the colocalization degree is estimated from the degrees computed between the reference image crop in I_A and image crops in I_B serving as randomized images. This estimation is valid only if the object spatial distribution in I_B is homogeneous and inter-crop independent. Besides this nearest-neighbor distance method, an overlap approach is also introduced in [7]. A method combining the intensity- and object-based approaches is proposed in [8], where a correlation analysis is performed in the region of interest obtained from Sobel pre-filtering. However, no statistical validation was provided.

In this paper, we present a novel object-based statistical colocalization method. Our biological application context is as follows. We are interested in how an unknown protein X is localized during the endocytic process. The protein X shows vesicular staining in the image. Because such staining is reminiscent of proteins involved in intracellular trafficking, we wished to determine if X could play a role in endocytosis. For this purpose, 5 different proteins (P_1, P_2, P_3, P_4, P_5)[1] were used as markers of cellular compartments corresponding to different steps of the endocytic process. Co-

*This work is funded by CNRS and Institut Pasteur of France. E-mail: {bzhang,jcolivo,vmeasyed}@pasteur.fr

[1]The real protein names are masked due to confidential reasons.

immunofluorescence labeling was performed between X and each of the five proteins, with X labeled in I_A, and endocytic marker in I_B.

Our colocalization relies on multiple hypothesis tests on the distances between all pairs of the spots generated by the two protein markers. Two spots are decided to be colocalized if their distance is significantly small. For this purpose, we first extract the spots by applying a wavelet-based spot detection approach which fully takes into account the mixed-Poisson-Gaussian noise process of confocal fluorescence images. Then, we build a null hypothesis model in which the distribution of the distance between two independently randomly drawn spots of the two proteins is estimated by a kernel method. The observed distances are tested against this null model. Our tests control the false discovery rate (FDR) of the colocalizations. Simulations show that this approach has a good specificity; in our biological application, the method detects that X strongly colocalizes with one of the five proteins during the endocytic process.

2. STATISTICAL COLOCALIZATION

Our colocalization procedure consists of three main steps, i.e., spot detection, determination of the cellular supports of the proteins, and hypothesis tests, which are presented in the following sections.

2.1. Spot detection

In each image, the fluorescein-marked proteins (aggregates) are present as a number of bright spots. Our first step aims at localizing these protein spots. This is accomplished by applying the spot extractor proposed in [9]. The basic idea is to reconstruct the image from the thresholded wavelet bands such that the spots are denoised and enhanced. The wavelet scaling (approximation) band is set to zero so that the smooth background is not reconstructed. By introducing a multiscale variance stabilizing transform (MS-VST) [9], the thresholds in the wavelet domain detect significant wavelet coefficients derived from the observed data which have a mixed-Poisson-Gaussian (MPG) statistical nature. Indeed, our images are from a confocal microscope, which are contaminated by both photon noise (Poisson) and camera readout noise (Gaussian), together forming an MPG process. The MS-VST allows us to Gaussianize and stabilize the noise in the wavelet bands. This indeed brings a complex MPG problem to the Gaussian denoising case, which has been well studied.

The reconstructed image is then binarized by a threshold $T \geq 0$. In our case, $T = 0$, i.e., the positive part of the image is retained and all negative pixels are set to zero. Then, all connected components as putative bright spots are extracted, and their intensity-weighted centers are computed. Note that the estimated centers do not heavily depend on the threshold T. In the ideal case of isolated isotropic spots, their computed centers will be independent of T (unless a too high T is used such that the spot is missed). Clearly, this is not the case for some classical colocalization approaches based on overlapping surface of the detected spots. Suppose $S_A := \{c_{A,i} := (x_{A,i}, y_{A,i}, z_{A,i}), 1 \leq i \leq N_A\}$ and $S_B := \{c_{B,i} := (x_{B,i}, y_{B,i}, z_{B,i}), 1 \leq i \leq N_B\}$ to be respectively N_A and N_B spot centers computed from I_A and I_B. We further define the distance between each center $c_{A,i}$ and each $c_{B,j}$ to be $d_{i,j} := |c_{A,i} - c_{B,j}|$. We have thus in total $N_A \cdot N_B$ distances $(d_{i,j})_{1 \leq i \leq N_A, 1 \leq j \leq N_B}$, which will be hypothesis-tested. We intend to see how many distances are significantly small such that they can not be observed just "by chance".

2.2. Cellular supports of the proteins

In order to test $(d_{i,j})_{i,j}$, we need to formulate our null hypothesis H_0 which is as follows: under the null, we suppose that $(d_{i,j})_{i,j}$ are observed from the object centers S_A and S_B which are independently and uniformly randomly distributed in the cellular supports of protein A (R_A) and of protein B (R_B), respectively. These supports are cellular regions where the proteins can be present (under H_0). The probability distribution under H_0, i.e., the density of the distance of two random points, one in R_A and the other in R_B, is denoted as $p(d)$. To estimate $p(d)$, we need to first estimate the supports R_A and R_B.

As R_A and R_B are protein supports under H_0, they can not be estimated from the observed data as the proteins could have undergone interactions (so would have been significantly colocalized) such that their observed supports would differ from those under H_0 (i.e., when they can only be colocalized by chance). For our study which is still preliminary, R_A and R_B are assumed to be the cell support which is pre-segmented by a biologist. In near future, we will use specific fluorescence markers to determine the regions (such as nuclei) where the proteins are known to be absent from prior knowledge. In this way, we will be able to refine our support estimations by excluding these regions.

2.3. The hypothesis testing framework

The protein supports being determined, we are at the point to estimate the distance distribution under H_0. Toward this purpose, we randomly draw a large number of uniformly distributed points in R_A and R_B, and compute their distances. In our experiments, approximately 3×10^6 distances are drawn. Then, we use the Gaussian kernel density estimator [10] (Parzen window method) to derive the density $p(\log(d))$. Note that it is the density of $\log(d)$ which is actually estimated. This is because a kernel estimator is known to be inaccurate at the boundaries. As d is always nonnegative, direct estimation of $p(d)$ will have undesirable side-effects around $d = 0$. A logarithmic transform does not modify the

problem since the logarithmic function is monotone, and it avoids the boundary instability of a kernel estimator.

Now, we can test the $N_A \cdot N_B$ log-transformed observed distances $(\log(d_{i,j}))_{i,j}$ against H_0. This should be done in a multiple hypothesis testing framework in order to correct multiple comparisons. For example, we may use the Bonferroni over-conservative correction to control the probability of erroneously rejecting even one of the true null hypotheses, i.e., the Family-Wise Error Rate (FWER). Alternatively, one can carry out the Benjamini and Hochberg procedure [11] to control the False Discovery Rate (FDR), which is the average fraction of false detections over the total number of detections, i.e.,

$$\text{FDR} := \mathbb{E}[|\text{FP}|/(|\text{FP}| + |\text{TP}|)]$$

Here $|\text{FP}|$ and $|\text{TP}|$ respectively stand for the number of false positives and that of true positives. The control of FDR has the following advantages over that of FWER: 1) it usually has a greater detection power; 2) it can easily handle correlated data [12]. The latter point is crucial for us because the $N_A \cdot N_B$ distances are dependent statistics. As a result, the FDR test is applied in our problem.

Suppose that the FDR is controlled at level β, i.e., FDR $\leq \beta$. This indicates that the true discovery rate (TDR)

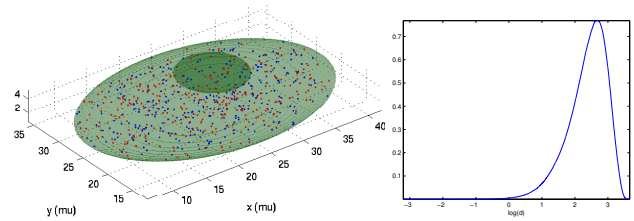
$$\text{TDR} := \mathbb{E}[|\text{TP}|/(|\text{FP}| + |\text{TP}|)]$$

will be at least $(1-\beta)$. Thus, supposing K hypotheses among $N_A \cdot N_B$ have been rejected (i.e., the K distances are judged as significantly small), we will have at least $K(1-\beta)$ correct decisions on average. Thus, from this value, we can finally calculate the colocalization ratio r_c, which is the fraction between the number of colocalized pairs and the total number of pairs, i.e., $r_c := K(1-\beta)/(N_A N_B)$. The higher r_c is, the higher is the degree of colocalization between S_A and S_B.

3. RESULTS

3.1. Colocalization specificity

We first evaluate the specificity of the colocalization under our null model. For this purpose, we simulated a cell volume with a nucleus, both having ellipsoidal shapes (Fig. 1(a)). R_A and R_B are both supposed to be the cellular region except for the nucleus. Fig. 1(b) shows the density $p(\log(d))$ estimated from the distances of randomly drawn point-pairs in the protein supports (approximately 3×10^6 distances are drawn). Furthermore, about 300 virtual center detections for each protein type are randomly generated in their support with a uniform spatial distribution (see Fig. 1(a), blue and red points are respectively from protein A and B). Thus, in one simulation we have in total approximately 10^5 distances to test. Our FDR level is set to $\beta = 0.5$. Clearly, as the simulation follows the scenario of our null model, any detected colocalization represents a false positive. We have carried out 10 replications



(a) (b)

Fig. 1. Colocalization specificity. (a) A cell volume with a nucleus, both having ellipsoidal shapes, is simulated. Blue and red points are virtual detections for protein A and B, respectively ($N_A \approx N_B \approx 300$); (b) Estimated $p(\log(d))$. With 10 simulations (each having about 10^5 tests) and β set to 0.5, not a false positive was observed.

of the simulation and not a single colocalized pair (false positive) was observed. This shows that our method has a good specificity.

3.2. Colocalization in real cells

In our application, we are interested in the colocalization of the reference protein X with the 5 target proteins. Fig. 2(a) and (b) show two confocal slices of a cell with markers for X and for P_1, respectively. The cell support is manually segmented. The protein detections within the support are presented in Fig. 2(c) and (d). The colocalization procedure has been carried for X and each of the target proteins on at least 5 cells. The average colocalization ratios for different proteins are shown in Fig. 3. We can see that the target proteins can be roughly classified into three families according to their colocalization ratios (r_c = mean $\pm$ standard deviation): (1) P_1 ($r_c = 10^{-3} \pm 4.67 \times 10^{-4}$), which exhibited the highest average colocalization ratio; (2) P_2 ($r_c = 1.67 \times 10^{-4} \pm 7.19 \times 10^{-5}$) and P_3 ($r_c = 6.49 \times 10^{-5} \pm 9.15 \times 10^{-5}$), which had ratios much lower than P_1 but much higher than (3) P_4 ($r_c = 1.29 \times 10^{-5} \pm 3.35 \times 10^{-6}$) and P_5 ($r_c = 4.78 \times 10^{-6} \pm 3.53 \times 10^{-6}$), which presented very low degrees of colocalization.

4. CONCLUSION

In this paper, we have presented a novel statistical colocalization approach where significantly colocalized spots are detected by FDR tests on the distances between all spot pairs from the two markers. This method has a good specificity and has been successfully applied in a real study of protein association. Our future work could involve the refinement of the protein support estimation, more validations on real data, and theoretic investigations such as point-process modeling for colocalization.

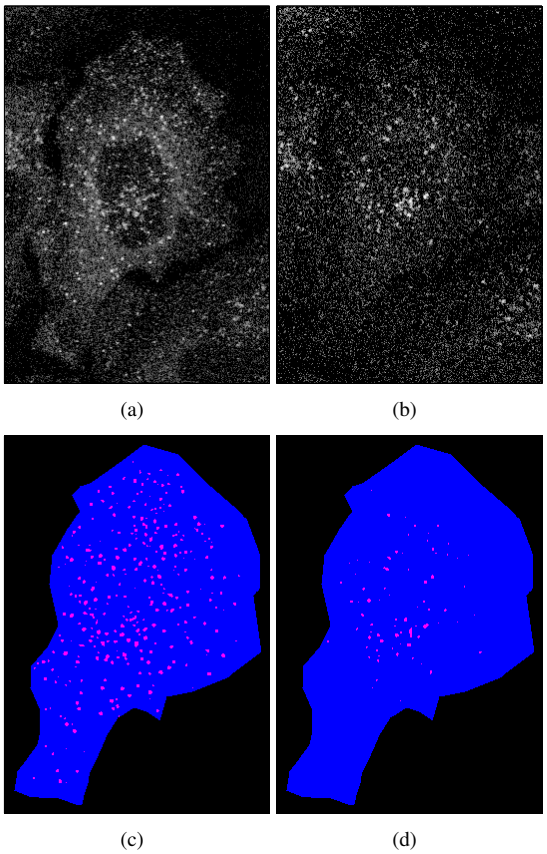


Fig. 2. Spot detection for protein X and P_1 in a cell. (a) a slice of the image of protein X; (b) a slice of the image of protein P_1; (c) detected spots (red) in (a) within the cell (blue); (d) detected spots (red) in (b) within the cell (blue).

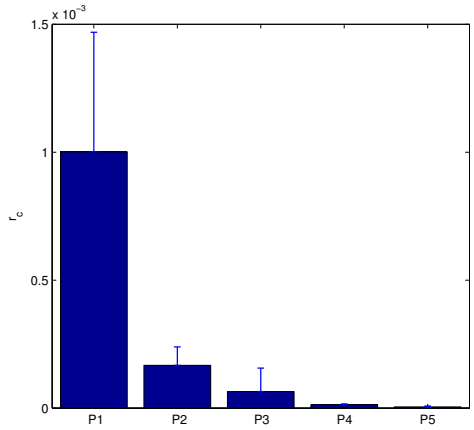


Fig. 3. Colocalization ratios between the reference protein X and the 5 different proteins (r_c = mean $\pm$ standard deviation): P_1 ($r_c = 10^{-3} \pm 4.67 \times 10^{-4}$), P_2 ($r_c = 1.67 \times 10^{-4} \pm 7.19 \times 10^{-5}$), P_3 ($r_c = 6.49 \times 10^{-5} \pm 9.15 \times 10^{-5}$), P_4 ($r_c = 1.29 \times 10^{-5} \pm 3.35 \times 10^{-6}$), and P_5 ($r_c = 4.78 \times 10^{-6} \pm 3.53 \times 10^{-6}$). The FDR detection level is set to $\beta = 0.5$.

ACKNOWLEDGMENTS

The authors are grateful to T. Duong (Institut Pasteur), I.-M. Viklund (Institut Pasteur and Karolinska Institute), G. Tran Van Nhieu (Institut Pasteur), and S. Pettersson (Karolinska Institute) for valuable discussions.

5. REFERENCES

[1] S. Bolte and F. P. Cordelieres, "A guided tour into subcellular colocalization analysis in light microscopy," *J. Microscopy*, vol. 224, no. 3, pp. 213–232, 2006.

[2] S. V. Costes, D. Daelemans, E. H. Cho, Z. Dobbin, G. Pavlakis, and S. Lockett, "Automatic and quantitative measurement of protein-protein colocalization in live cells," *Biophys. J.*, vol. 86, pp. 3993–4003, 2004.

[3] E. Manders, J. Stap, G. Brakenhoff, R. van Driel, and J. Aten, "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.

[4] B. van Steensel, E. van Binnendijk, C. Hornsby, H. van der Voort, Z. Krozowski, E. de Kloet, and R. van Driel, "Partial colocalization of glucocorticoid and mineralocorticoid receptors in discrete compartments in nuclei of rat hippocampus neurons," *J. Cell Sci.*, vol. 109, pp. 787–792, 1996.

[5] Q. Li, A. Lau, T. J. Morris, L. Guo, C. B. Fordyce, and E. F. Stanley, "A syntaxin 1, G_{α_o}, and N-type calcium channel complex at a presynaptic nerve terminal: analysis by quantitative immunocolocalization," *J. Neurosci.*, vol. 24, pp. 4070–4081, 2004.

[6] Y. Boutte, M. T. Crosnier, N. Carraro, J. Traas, and B. Satiat-Jeunemaitre, "The plasma membrane recycling pathway and cell polarity in plants: studies on PIN proteins," *J. Cell Sci.*, vol. 113, pp. 1255–1265, 2006.

[7] E. Lachmanovich, D. E. Shvartsman, Y. Malka, C. Botvin, Y. I. Henis, and A. M. Weiss, "Co-localization analysis of complex formation among membrane proteins by computerized fluorescence microscopy: application to immunofluorescence copatching studies," *J. Microsc.*, vol. 212, pp. 122–131, 2003.

[8] 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.

[9] B. Zhang, M. J. Fadili, J.-L. Starck, and J.-C. Olivo-Marin, "Multiscale variance-stabilizing transform for mixed-Poisson-Gaussian processes and its applications in bioimaging," in *ICIP*, 2007, vol. VI, pp. 233–236.

[10] A. W. Bowman and A. Azzalini, *Applied Smoothing Techniques for Data Analysis*, Oxford University Press, 1997.

[11] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *J. Roy. Statist. Soc. ser. B*, vol. 57, no. 1, pp. 289–300, 1995.

[12] Y. Benjamini and D. Yekutieli, "The control of the false discovery rate in multiple testing under dependency," *Ann. Statist.*, vol. 29, no. 4, pp. 1165–1188, 2001.