Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

Analysis of spatial dependencies of endocytic proteins based on temporal random sets

Rafael Sebastian, Maria Elena Diaz, Guillermo Ayala, Roberto Zoncu and Derek Toomre

Abstract— **Statistical image analysis has emerged as a basic methodology in the study of many real phenomena, which can be represented by sequences of binary images. Recent techniques, such as Total Internal Reflection Fluorescent Microscopy, allow us to image simultaneously two fluorescent-tagged proteins which are relevant in important cellular processes, such as endocytosis. Here, we model these biological pairs of image sequences as realizations of a 3D non-isotropic bivariate random set, in which one dimension corresponds to time. We analyze their second-order properties by means of the cross-covariance and the pair correlation function in order to study colocalization between proteins. Results show the proposed methodology allows us to study spatial dependencies in a formal and robust way.**

I. INTRODUCTION

Images can be usually obtained either as planar samples from a truly two-dimensional process, or as a sampled from two-dimensional sections of a binary structure that is indeed three-dimensional. In particular, we are interested in the study of the inter-dependencies between clathrin-GFP protein and other proteins which are involved in the endocytosis process [7], [4], a cellular mechanism by which cells carry traffic from the plasma membrane to internal cell compartments. Endocytosis has recently been imaged in real time by conjugating clathrin proteins to fluorescent molecules and imaging cells with specialized microscopy techniques such as Total Internal Reflection Fluorescence Microscopy (TIRFM) [18]. Under TIRFM, the assembly of fluorescently-labeled protein at a site of ongoing endocytosis results in the appearance and steady growth of a diffraction-limited spot. The areas of fluorescence associated to the different endocytic spots overlap, forming random clumps of different sizes, shapes and durations. The distribution of the endocytic events in space is influenced by many biological factors. Knowing the spatial relationship between proteins in living cells is essential. Why? Although biochemical assays can show that given proteins can in theory interact - in the cellular context what matters is if these proteins do targeted to the same regions. This spatial cross-correlation is not proof of an interaction, although can be incredible suggestive and when combined with additional manipulations can prove that such interactions are meaningful. For instance, protein X can lead to the recruitment of proteins Y and Z to a specific location. For clathrin mediated endocytosis, many of the key proteins players are known [12], but understanding their spatial interdependencies in normal and perturbed conditions will allow new insight into their function. Traditionally, obtaining this kind of data has been very time-consuming and much of the work had to be done manually, making it virtually impossible in large image sequences.

Many papers can be found in the literature concerned with the study of random sets defined in the 2D or 3D Euclidean space. In particular, Boolean models provide good descriptions for very irregular patterns observed in Microscopy, Material Sciences [5], [6], Biology, Medicine [8], [11], [9], Chemistry, Geostatistics [14], and more recently in Cellular Communications Networks [3]. They formalize the configuration of independent and randomly located particles. Some results concerning the estimation on Boolean models can be found in [16], [17], [13]. The temporal Boolean models, an extension to model randomly placed particles with random durations was defined and applied in [1], [15] to count endocytosis and to estimate their duration distribution from image sequences.

In this paper, we consider the pairs of binary image sequences as a realization of a spatio-temporal bivariate stochastic process, in particular, regions of fluorescence of endocytic proteins are modelled as bivariate temporal random sets. Firstly, we will examine some second-order properties for this model. Secondly, we will apply our proposed estimators to investigate spatial dependencies of pairs of proteins from several image sequences.

Section II describes the statistical spatial analysis of these binary sequences by means of some second-order characteristics, i.e. the covariance and the pair correlation function. In Section III we present the results obtained for six pairs of image sequences corresponding to four different protein combination. Conclusions and further developments are summarized in Section IV.

II. SECOND-ORDER ANALYSIS OF TEMPORAL RANDOM SETS

The binary image sequences will be considered as realizations of a stationary non-isotropic random set in the 3D space. We will not assume any parametric model for the random sets. In particular, we will use the covariance $\mathbb{C}(\mathbf{s})$ of the random set Ξ, defined as the probability that two points, separated by a vector $\mathbf{s}$, lie simultaneously in Ξ i.e., $\mathbb{C}(\mathbf{s}) = P[\{x \in \Xi, x + \mathbf{s} \in \Xi\}]$. If Ξ is isotropic as well as

This paper has been supported by project HFSP RGY40/2003.

R. Sebastian and Maria Elena Diaz are with Department of Computer Science, University of Valencia. Avda. Vicent Andrés Estellés, 1. 46100-Burjasot, Spain. {rafa.sebastian, elena.diaz}@uv.es.

G. Ayala is with Departamento de Estadística e Investigación Operativa, University of Valencia. Avda. Vicent Andrés Estellés, 1, 46100-Burjasot, Spain. guillermo.ayala@uv.es.

R. Zoncu and D. Toomre are with Department of Cell Biology, Yale University. P.O. BOX 208002, New Haven, Connecticut, 06520-8002 USA. {roberto.zoncu,derek.toomre}@yale.edu.

stationary, the spatial temporal covariance only depends on the length $s = \|\mathbf{s}\|$.

Let first introduce some basic notation. If we observe a sequence of binary images over time and denote by $\Xi(t)$ the pixels valued one at time t (a frame or cross-section) then the set $\Xi = \cup_{t\geq 0}\Xi(t)$ is a subset of $\mathbb{R}^2 \times \mathbb{R}_+$. If for each time t, we have a random set $\Xi(t)$ then $\Xi = \cup_{t\geq 0}\Xi(t)$ is a *temporal random set* in $\mathbb{R}^2 \times \mathbb{R}_+$ in such a way that the temporal cross-section of Ξ at time t will be $\Xi(t)$. Each $\Xi(t)$ is contained in the product space $W \times [0, T]$, where W is the observation window and $[0, T]$ is the time interval observed. It will be assumed temporal and spatial stationarity.

If two random sets Ξ_i and Ξ_j are jointly considered in the spatial temporal space then the cross-covariance function is

$$\mathbb{C}_{ij}(s,t) = P[\{x \in \Xi_i(0), (x+s) \in \Xi_j(t)\}]. \quad (1)$$

This function gives the probability that an arbitrary point at an arbitrary time belongs to the i-th temporal random set and its translated by a distance s and a time t belongs to the j-th temporal random set. We are interested in the estimation of the spatial covariance between two random closed sets Ξ_i and Ξ_j from temporal cross-sections,

$$\hat{\mathbb{C}}_{ij}(s) = \frac{1}{n}\sum_{k=1}^{n} \frac{\nu_2\Big[\big(\Xi_i(k)\cap W\big)\cap\big(\big(\Xi_j(k)\cap W\big)+s\big)\Big]}{\nu_2\big[W\cap\big(W+s\big)\big]}, \quad (2)$$

where $\Xi_i(k)$ denotes the k-th frame of the sequence Ξ_i, ν_2 stands for the area, n is the number of frames and W is the sampling window. Given $\mathbb{C}_{ij}(s)$, we can estimate other interesting functions as the pair correlation function $g_{ij}(s)$ and the Ripley $\mathbb{K}_{ij}$-function. The pair correlation function is estimated as the ratio of the covariance to the square of the area fractions, which makes samples with different area fraction more comparable than by simple use of the covariance. Large values of $g_{ij}(s)$ show that point pairs of distance s appears frequently, small values that are rare. Thus, values of g_{ij} close to 1 suggest independence, whereas values exceeding 1 suggest positive association or clustering. In the case of the $\mathbb{K}_{ij}$-function, it measures the area covered by the $j-th$ component within a specific distance s of a randomly chosen point of the $i-th$ component. Specifically, the estimators are

$$\hat{g}_{ij}(s) = \frac{\hat{\mathbb{C}}_{ij}(s)}{\hat{p}_i\hat{p}_j}, \quad (3)$$

$$\hat{\mathbb{K}}_{ij}(s) = \int_{B(0,s)} \frac{\hat{\mathbb{C}}_{ij}(u)}{\hat{p}_i\hat{p}_j}du, \quad (4)$$

where $\hat{p}_i$ is the estimator of the area fraction of Ξ_i. The natural estimator of the area fraction p_i from image sequences is the arithmetic mean of the area fractions observed at each frame (see [10] for an improved estimator),

$$\hat{p}_i = \frac{1}{n+1}\sum_{k=0}^{n} \frac{\nu_2[\Xi_i(k\delta)\cap W]}{\nu_2[W]}. \quad (5)$$

$\mathbb{K}_{ij}$-function is an integrated version of g_{ij} within a disk, therefore is more robust against noise. Traditionally, information provided by both summary statistics has been combined to test spatial interaction.

A. Testing spatial independency

In this section we explore the spatial dependency between some pairs of proteins which take part during the process of endocytosis. We propose to use g_{ij} and $\mathbb{K}_{ij}$-function to test the spatial independency of temporal random closed sets. If the random sets Ξ_1 and Ξ_2 are spatially independent then $\mathbb{C}_{12}(s) = P(x \in \Xi_1, x+s \in \Xi_2) = P(x \in \Xi_1)P(x+s \in \Xi_2) = p_1p_2$ does not depend on s and from (3) and (4) we have $g_{12}(s) = 1$ and $\mathbb{K}_{12}(s) = \pi s^2$, respectively.

To test the complete spatial independence, we apply a sequence of independent random toroidal shifts on W. A toroidal shift, T_h, is a simultaneous, parallel displacement of all points of the set by the same displacement vector $\mathbf{h}$, which is generated at random with equal probability for all possible displacements and uniformly distributed in W. From the original bivariate sequence observed $\{\Xi_1(k)\cap W, \Xi_2(k)\cap W\}_{k=1,\ldots,n}$, we generate the displacement vector $\mathbf{H} = \mathbf{h}$ and consider the new bivariate sequence $\{(\Xi_1(k)\cap W, T_{\mathbf{h}}(\Xi_2(k)\cap W)\}_{k=1,\ldots,n}$. A toroidal random set has the property of breaking the relationship between the components Ξ_1 and Ξ_2. We conduct a Monte Carlo test of independence by comparing the value of a suitable test statistic for the observed data with values generated under the randomizations. The cross $\mathbb{K}$-function is estimated for the original pair of sequences, $\hat{\mathbb{K}}_{12,0}$, and for its randomizations $\hat{\mathbb{K}}_{12,i}$ with $i = 1,\ldots,m$. Afterwards, the lower and upper envelopes obtained from $\hat{\mathbb{K}}_{12,i}$ with $i = 1,\ldots,m$ are compared with $\hat{\mathbb{K}}_{12,0}$ which should be contained within the envelopes under the null hypothesis of complete spatial independence [2]. The statistic used was $d_n = \int_0^{+\infty}(\mathbb{K}_{12,n}(s,t) - \bar{\mathbb{K}}_{12,n}(s,t))^2 dsdt$, where $\bar{\mathbb{K}}_{12,n}(s,t) = \sum_{r=0,r\neq n}^{m} \frac{\mathbb{K}_{12,r}(s,t)}{m}$ and $r = 0,1,\ldots,m$. All rankings of d_0 are equiprobable under the null hypothesis. We will reject the null hypothesis on the basis that d_0 ranks the k-th largest. It gives an exact, one-sided test with p-value $1 - \frac{k}{m+1}$. Here we used $m = 99$ simulations.

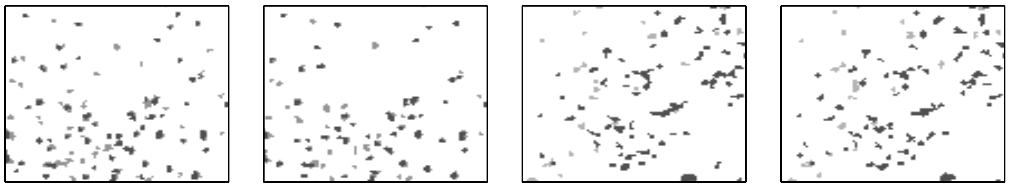
III. Results

Our data comprise six pairs of sequences of proteins for six different cells (See Table I). One channel of the microscope was used to image clathrin protein and the other to acquire a second protein. $g_{12}(s)$ and $\mathbb{K}_{12}(s)$ were estimated for the six bivariate sequences as described in Section II. Fig. 1 (a) and (b) show two images of pair #1 and Fig. 1 (c) and (d) correspond to pair #3.

TABLE I

DESCRIPTION OF BIOLOGICAL IMAGE SEQUENCES

Pair	Protein #1	Protein #2	# of frames	Size	p-value	Correlated
1	Clathrin-RFP	Hip1R-GFP	151	180×153	0	Yes
2	Clathrin-RFP	Epsin-GFP	203	213×185	0	Yes
3	Clathrin-RFP	Caveolin-GFP	78	135×281	0.596	No
4	Clathrin-RFP	Caveolin-GFP	66	110×221	0.778	No
5	Clathrin-RFP	Clathrin-GFP	50	122×186	0.020	Yes
6	Clathrin-RFP	Clatrhin-GFP	143	166×210	0	Yes



(a) (b) (c) (d)

Fig. 1. **Snapshots from pairs of proteins**. (a) and (b) show Clathrin-RFP (light gray) and Hip1R-GFP (dark gray) which colocalize. (c) and (d) correspond to Clathrin-RFP (light gray) and Caveolin-GFP (dark gray) which are spatially independent.

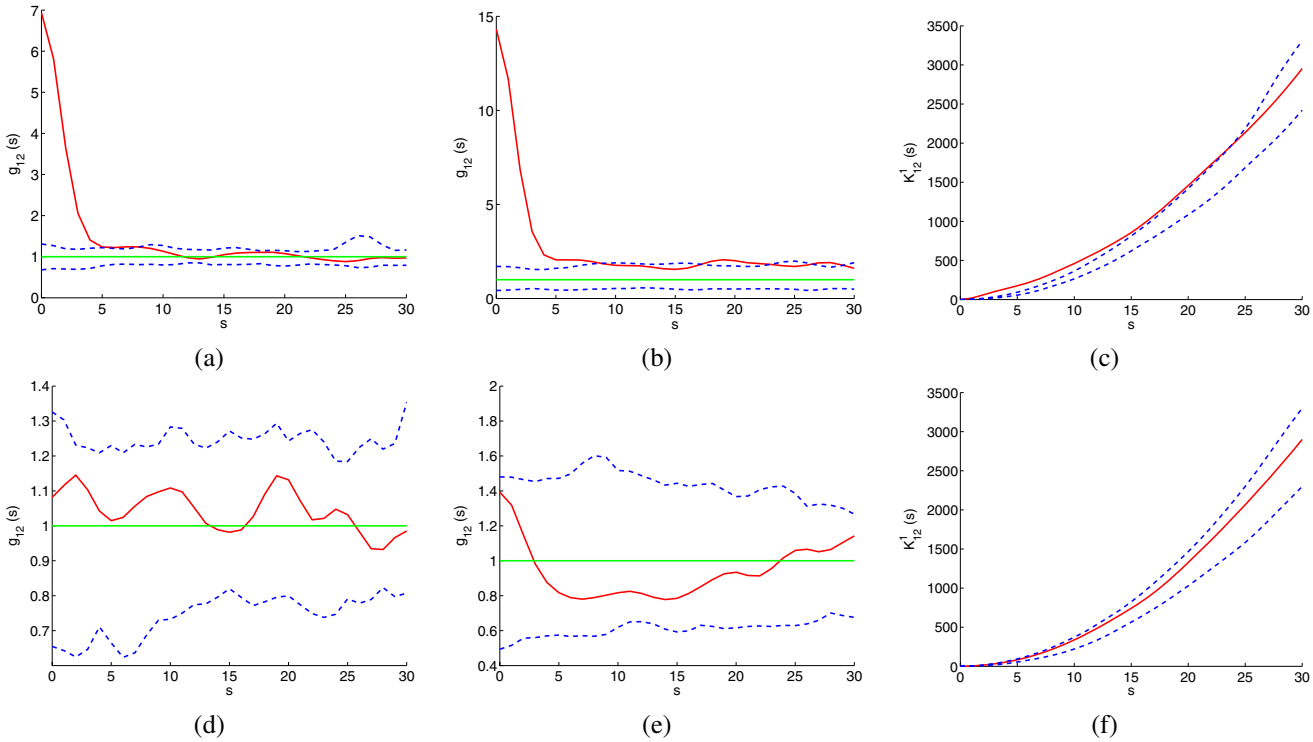


Fig. 2. **Results of second-order analysis on** 4 **biological pairs of image sequences.** g_{12} of pairs from 1 to 4 are shown in plots (a), (b), (d) and (e) respectively. Solid lines represent the empirical pair correlation function (flat line represents the value under complete independence $g_{ref} = 1$), dashed lines are the envelopes obtained from 99 randomizations. In particular, (a) and (b) display those pairs (#1 and #2) which clearly show colocalization since the empirical g_{12} is outside the envelopes for small distances. Those which were expected not to be related (#3 and #4) are plotted in (d) and (e). Plots (c) and (f) show the cross $\mathbb{K}_{12}$-functions obtained for pairs #1 and #3, respectively.

Time-lapse images were acquired by TIRFM. In this technique, a laser beam is sent to the sample at a critical angle. In this technique only objects which are within $100 - 200$ nm of the bottom plasma membrane of the cell are illuminated, while the nucleus, inner cytosol and upper plasma membranes are left in the dark [18]. In this way, it is possible to image membrane-associated events, such as endocytosis, with superior signal-to-noise. The setup employed for this study was an objective-based TIRFM (63X magnification) implemented on an inverted IX70 microscope (Olympus) and coupled to a 488-nm laser line (Melles Griot). The laser power was between 80 and 100 mW. The image processing was based on the application of the following steps: a top-hat transform to subtract the background and extract peaks of fluorescence, a template matching to remove eventual noise, followed by a region growing technique in order to delineate each marked object.

The method applied on the pairs of sequences $\#5$ and $\#6$ (clathrin against itself) provided a good starting point for validation. The small p-values obtained clearly pointed out the rejection of the null hypothesis of independence (see Table I). As it was expected [12], a high correlation was obtained for pairs $\#1$ and $\#2$ (p-value$= 0$). Figs. 2 (a)-(c) corroborate these results. Regarding pairs $\#3$ and $\#4$, high p-values were obtained and therefore we can not reject the null hypothesis of spatial independence. The empirical g_{12} and $\mathbb{K}_{12}$ plots lie within the envelopes for every distance analyzed.

The pair correlation function $g_{12}(s)$ was the best discriminator because it takes into account the differences in the area fractions of the different channels of the cells. Curves representing the pair correlation function were expected to reach the reference line $g_{ref} = 1$ for larger distances, since the relationship between proteins, when it exist, is thought to be at small distances (see Fig. 2).

IV. Conclusions

We have proposed and applied a methodology for analyzing spatial dependencies and, in general, carry out a second-order analysis on proteins from biological image sequences. This methodology could be used as a robust screening tool to study the dynamics of in-vivo cells expressing certain proteins under different treatments. Classical techniques based on visual inspection, (see Fig. 1 (c) and (d)), could lead to wrong conclusions. Our results matched the finding obtained by other techniques. Here the analysis is done at the single unit level in contrast to ensemble methods such as traditional biochemistry. Future extensions to the proposed methodology are the study of temporal dependencies and spatio-temporal dependencies and the fitting of an interaction model.

V. ACKNOWLEDGMENTS

This paper has been supported by project HFSP RGY-40/2003.

References

[1] G. Ayala, R. Sebastian, M. Diaz, E. Diaz, R. Zoncu, and D. Toomre, "Analysis of spatially and temporally overlapping events with application to image sequences," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. in press, 2006.

[2] P. Diggle, *Statistical Analysis of Spatial Point Patterns*. London: Arnold, 2003.

[3] O. Dousse, P. Mannersalo, and P. Thiran, "Latency of wireless sensor networks with uncoordinated power saving mechanisms," *MobiHoc*, 2004.

[4] M. Ehrlich, W. Boll, A. van Oijen, R. Hariharan, K. Chandran, M. Nibert, and T. Kirchhausen, "Endocytosis by random initiation and stabilization of clathrin-coated pits," *Cell*, vol. 118, pp. 591–605, 2004.

[5] J. Handley, "Discrete approximation of the linear boolean model of heterogeneous materials," *Physical Review*, vol. 60, pp. 6150–6152, 1999.

[6] D. Jeulin, "Random models for morphological analysis of powders," *J. Microsc.*, vol. 172, pp. 13–22, 1993.

[7] M. Kaksonen, C. Toret, and D. Drubin, "A modullar design for the clathrin- and actin-mediated endocytosis machinery," *Cell*, vol. 123, pp. 305–320, 2005.

[8] T. Mattfeldt, H. Frey, and C. Rose, "Second-order stereology of benign and malignant alterations of the human mammary gland," *J. Microsc.*, vol. 171, pp. 143–151, 1993.

[9] T. Mattfeldt, V. Schmidt, D. Reepschlager, C. Rose, and H. Frey, "Centred contact density functions for the statistical analysis of random sets. a stereological study on benign and malignant glandular tissue using image analysis." *J. Microsc.*, vol. 183, pp. 158–169, 1996.

[10] T. Mattfeldt and D. Stoyan, "Improved estimation of the pair correlation function of random sets," *J. Microsc.*, vol. 200, pp. 158–173, 2000.

[11] T. Mattfeldt, U. Vogel, H.-W. Gottfried, and H. Frey, "Second-order stereology of prostatic adenocarcinoma and normal prostatic tissue," *Acta Stereol.*, vol. 12, pp. 203–208, 1993.

[12] M. Metzler, V. Legendre-Guillemin, L. Gan, V. Chopra, A. Kwok, P. McPherson, and M. Hayden, "Hip1 functions in clathrin-mediated endocytosis through binding to clathrin and adaptor protein 2," *Journal of Biological Chemistry*, vol. 276, pp. 39 271–39 276, 2001.

[13] I. Molchanov, *Statistics of the Boolean Model for Practitioners and Mathematicians*. Chichester: John Wiley and Sons, 1997.

[14] ——, *Stochastic Geometry Likelihood and Computation*, ser. Monographs on Statistics and Applied Probability. Chapman & Hall /CRC, 1999, vol. 80, ch. 7, pp. 285–331.

[15] R. Sebastian, E. Diaz, G. Ayala, M. Diaz, R. Zoncu, and D. Toomre, "Studying endocytosis in space and time by means of temporal boolean models," *Pattern Recognition*, vol. in press, 2006.

[16] J. Serra, *Image Analysis and Mathematical Morphology*. Academic Press, 1982, vol. 1.

[17] D. Stoyan, W. Kendall, and J. Mecke, *Stochastic Geometry and its Applications*, 2nd ed. Berlin: Wiley, 1995.

[18] D. Toomre and D. Manstein, "Lighting up the cell surface with evanescent wave microscopy," *Trends Cell Biol.*, vol. 11, pp. 298–303, 2001.