INSTITUT PASTEUR

Colocalisation

V. Meas-Yedid

Quantitative Image Analysis Unit

Institut Pasteur, France

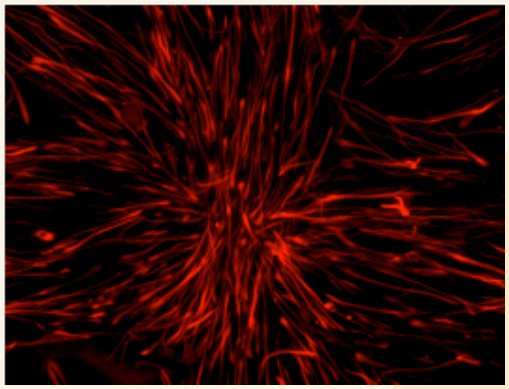
Colocalization Definitions

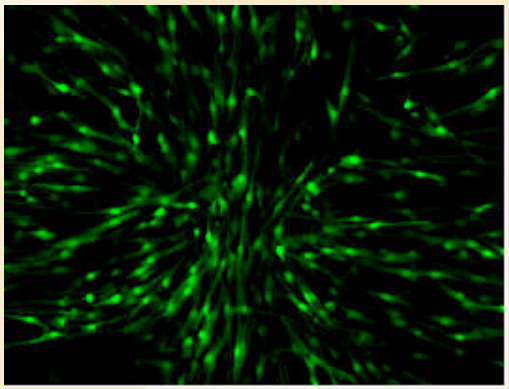
Colocalization refers to different data analysis methods to characterize the degree of overlap between two different fluorescent labels, each having a separate emission wavelength, to see if two different cellular "targets" are located in the same area or very near to one another. (Wikipedia)

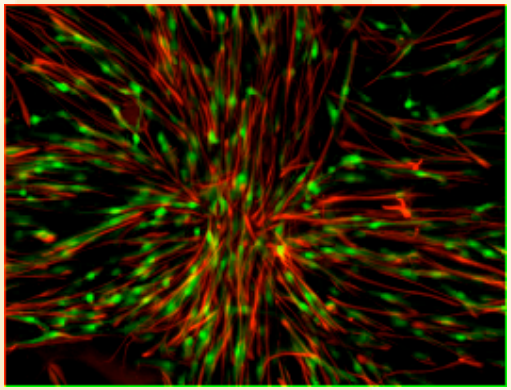
In cell biology:
two proteins are at the same location
At statistical level:
at the observed resolution, we can not exclude that two proteins are at the same location

- Why is colocalization investigated ?
  - to show an association between two molecules
  - to show the recruitment of a molecule during a process

Definitions







How can colocalization be present ?

- chance colocalization
- **apparent** colocalization due to inappropriate image analysis methods
- **real** colocalization where the assumption of direct or indirect interactions between molecules is correct.

Different «colocalizations»

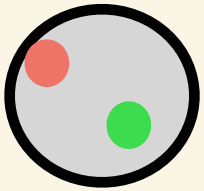
- Direct interaction

distance ~ 1-10 nm

- Indirect interaction

distance ~ 50-100 nm

- Cellular microdomains

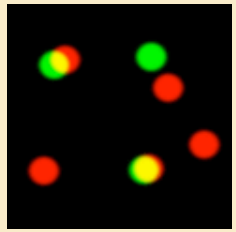


distance ~ 100-500 nm

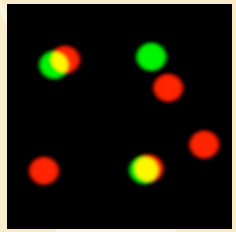
membrane domains, endosomes...

We want coefficients

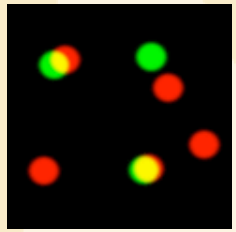
sensitive to the extra non-colocalized signal



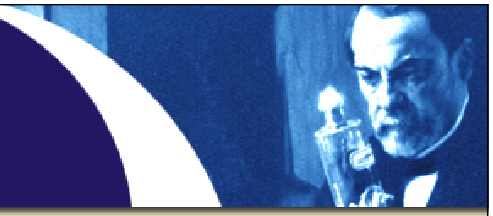
robust to relative variations of intensities



robust to the presence of Background

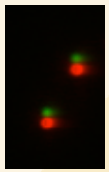


Problems related to the acquisition

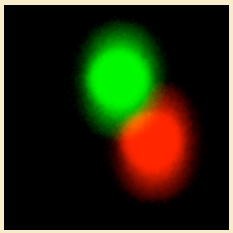


Effect of imaging conditions

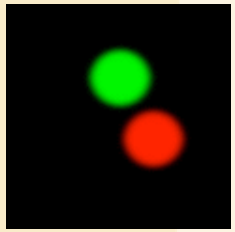
Color shift



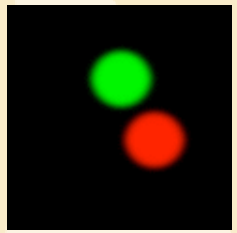
Background: weak SNR



Blur: out of focus



Cross-talk



Colocalization Overview

- A guided tour into subcellular colocalization analysis in light microscopy ImageJ plugin: Jacob [Botte et al. 2006]
  - Intensity-correlation based approaches [Costes et al. 2004][Manders et al. 1992] [Van Steensel et al. 1996][Li et al. 2004]
    - Global intensity similarity measure
    - Mixed information; noise contamination
  - Object-based approaches [Boutte et al. 2006][Lachmanovich et al. 2003]
    - Information of the objects of interest which are explicitly explored
  - Our method
    - Statistical reliable object detection adapting the microscope noise nature
    - Statistical reliable colocalization controlling a false discovery rate

Quantitative colocalisation

- intensity-based methods

2 approaches

Bolte S. and Cordelières F.P. (2006) J. Microscopy

- object-based methods

- correlation score of the intensity values in a dual-channel image is calculated
- Pearson's, Manders', cross correlation and Li's coeff., Costes: statistical validation

- Object identification by segmentation
- Analysis of object position

Quantitative colocalization

Quantification

image as a collection of pixels:
look for pixels which are linearly linked between the two channels.

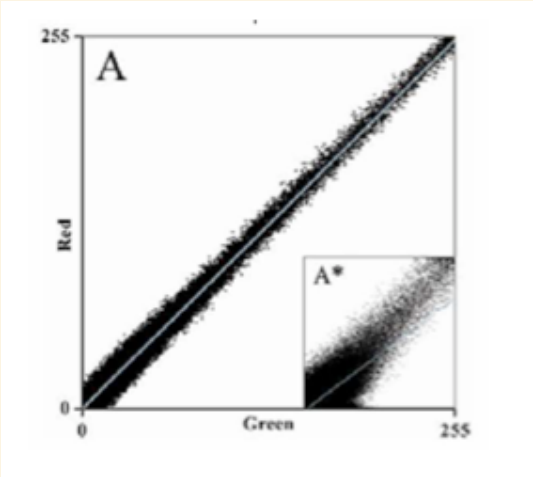
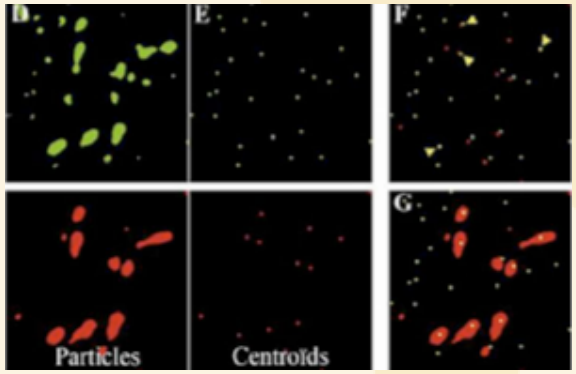


image as a collection of objects:
look for partial/total overlap of objects



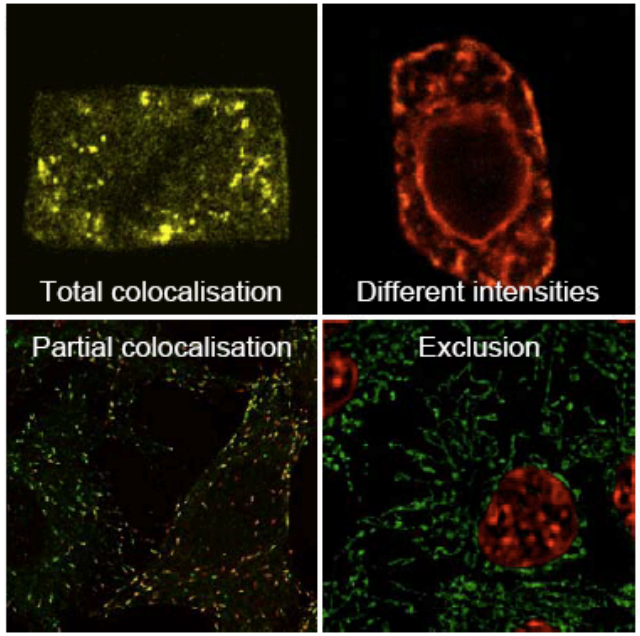
Intensity based: well adapted for a **global intensity** analysis, but not for **local spatial** analysis

Intensity-based methods

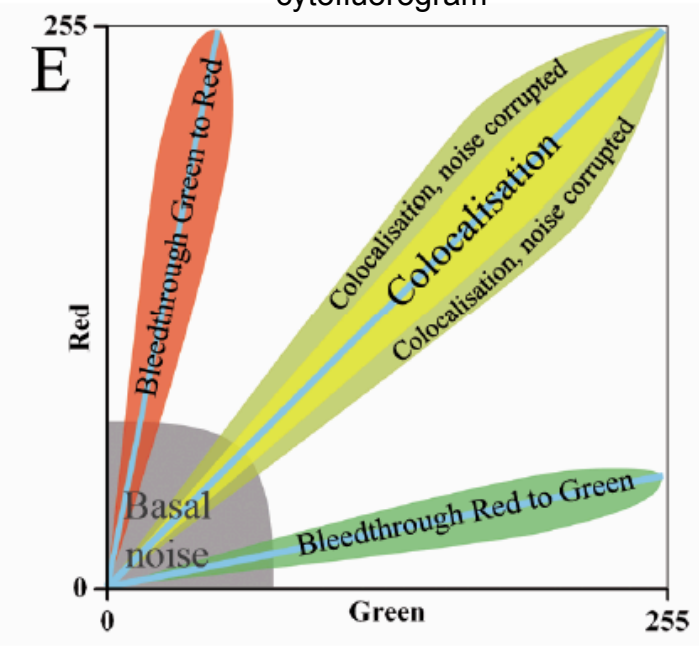
Pearson's coefficient: evaluate the linearity of relationship

$$R_r = \frac{\sum (R_i - \bar{R}) \times (G_i - \bar{G})}{\sqrt{\sum (R_i - \bar{R})^2 \times \sum (G_i - \bar{G})^2}}$$

$R_r \approx -1$, no conclusion or exclusion
$R_r \approx 0$, no correlation
$R_r \approx 1$, high correlation



cytofluorogram



Intensity-based methods

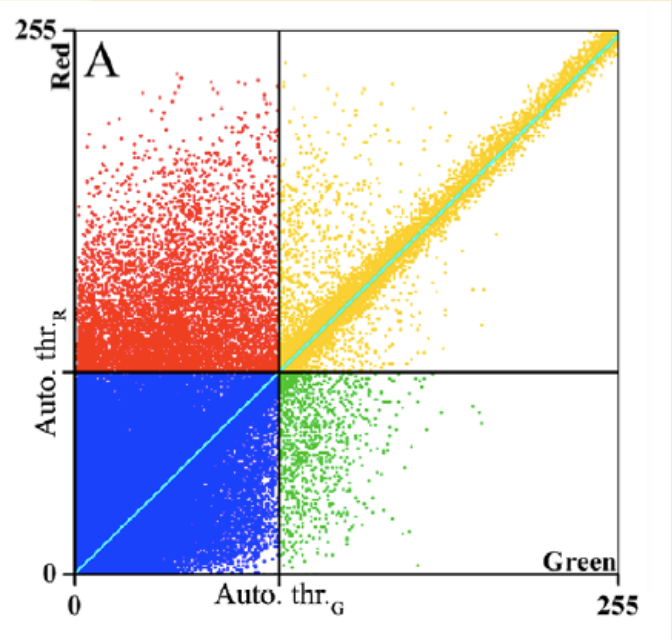
Noise is a major drawback which may decrease the Pearson's coefficient value
Remove noise, but how ? thresholding

Costes' method: progressively decrease the threshold until R_P below threshold is equal or below zero

Manders' coefficients:

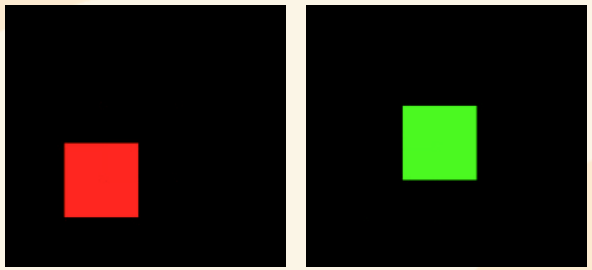
$$M_{red} = \frac{\sum_i R_{i,coloc}}{\sum_i R_i} \qquad M_{green} = \frac{\sum_i G_{i,coloc}}{\sum_i G_i}$$

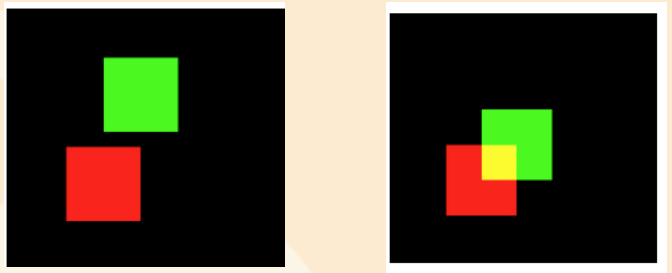
$R_{i,coloc} = R_i$, *if* $G_i > 0$; $G_{i,coloc} = G_i$, *if* $R_i > 0$



Manders *et al. (1992)* J. Cell Sci 103, 857-862.
Costes *et al.* (2004). Biophys J. 86, 3993-4003

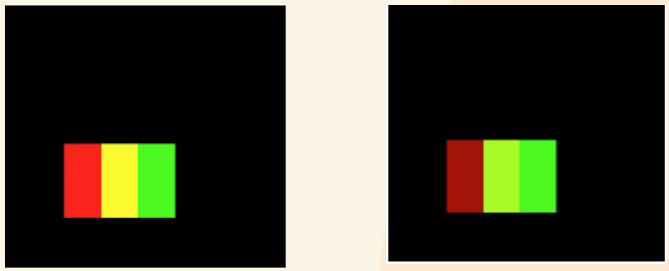
Intensity-based methods





PC:-0.108

PC:0.169



PC:0.446

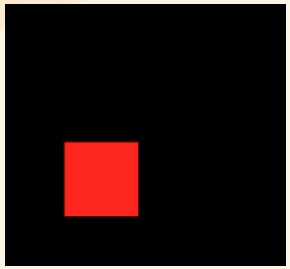
PC:0.446

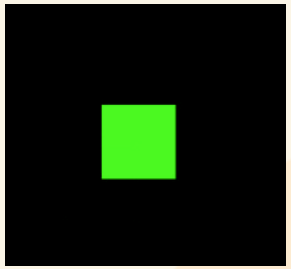
PC:0.446

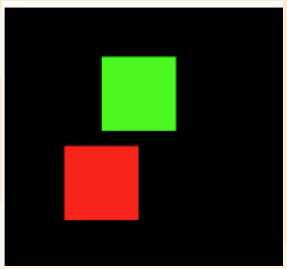
PC:0.228

INSTITUT PASTEUR

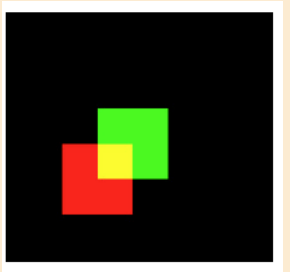
Intensity-based methods



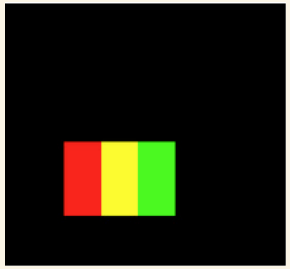




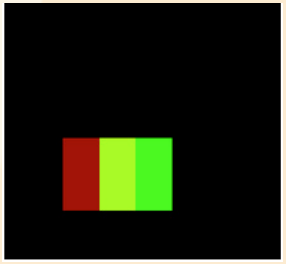
PC:-0.108
M1=0, M2=0



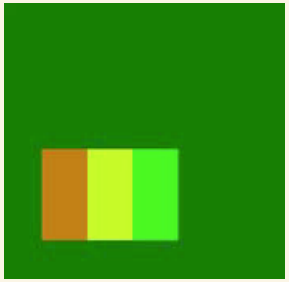
PC:0.169
M1=0.25, M2=0.25



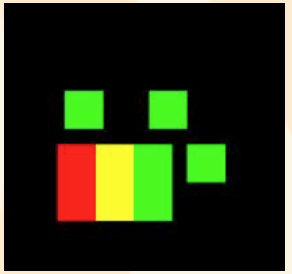
PC:0.446
M1=0.5, M2=0.5



PC:0.446
M1=0.5, M2=0.5



PC:0.446
M1=1, M2=0.163



PC:0.228
M1=0.5, M2=0.286

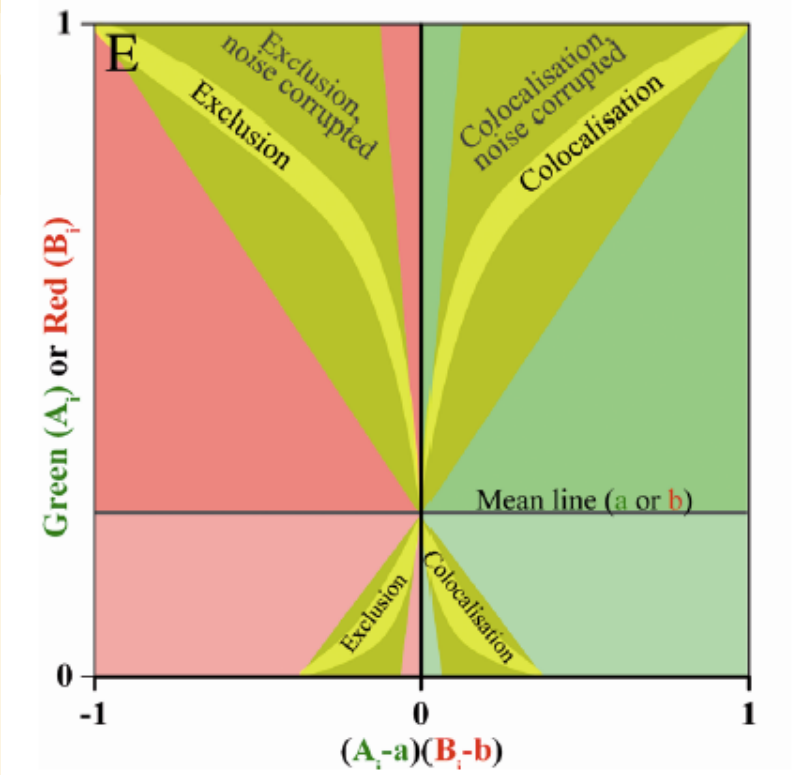
Intensity-based methods (3/4)

Li's approach:

Normalize both channel intensities to a [0, 1] range
Postulate: if colocalization exists, both intensities are on the same side of the mean value (Ai-a)(Bi-b)>0
Intensity Correlation Quotient, ICQ
reflects the proportion of covarying pixels

The ICQ values are distributed between [-0.5;0.5]
Random staining: ICQ ≈ 0
Segregated staining: -0.5<ICQ<0
Dependent staining : 0<ICQ<0.5

A good mean to highlight exclusion and absence of colocalization

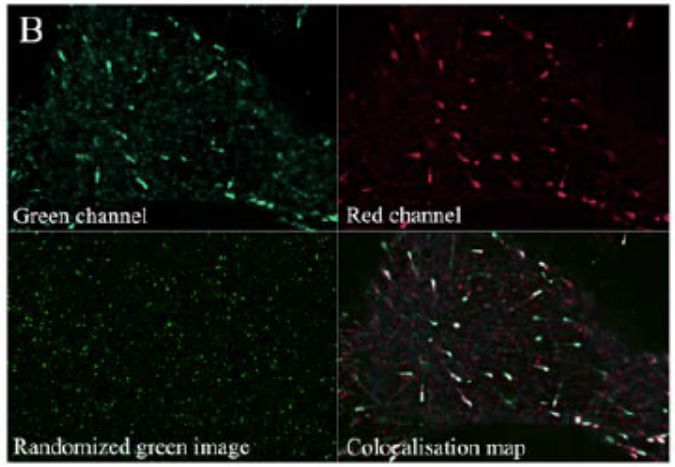


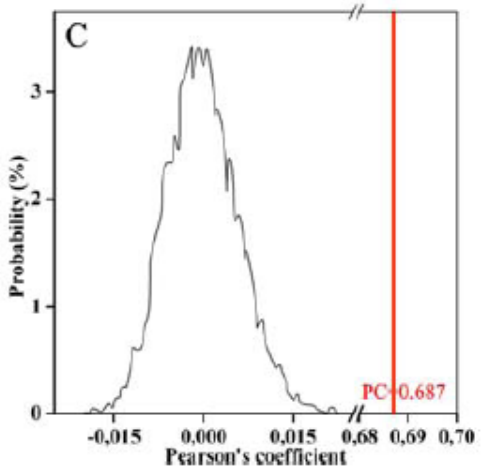
Li *et al.* (2004). J. Neurosci. 24, 4070-4081

Intensity-based methods (4/4)

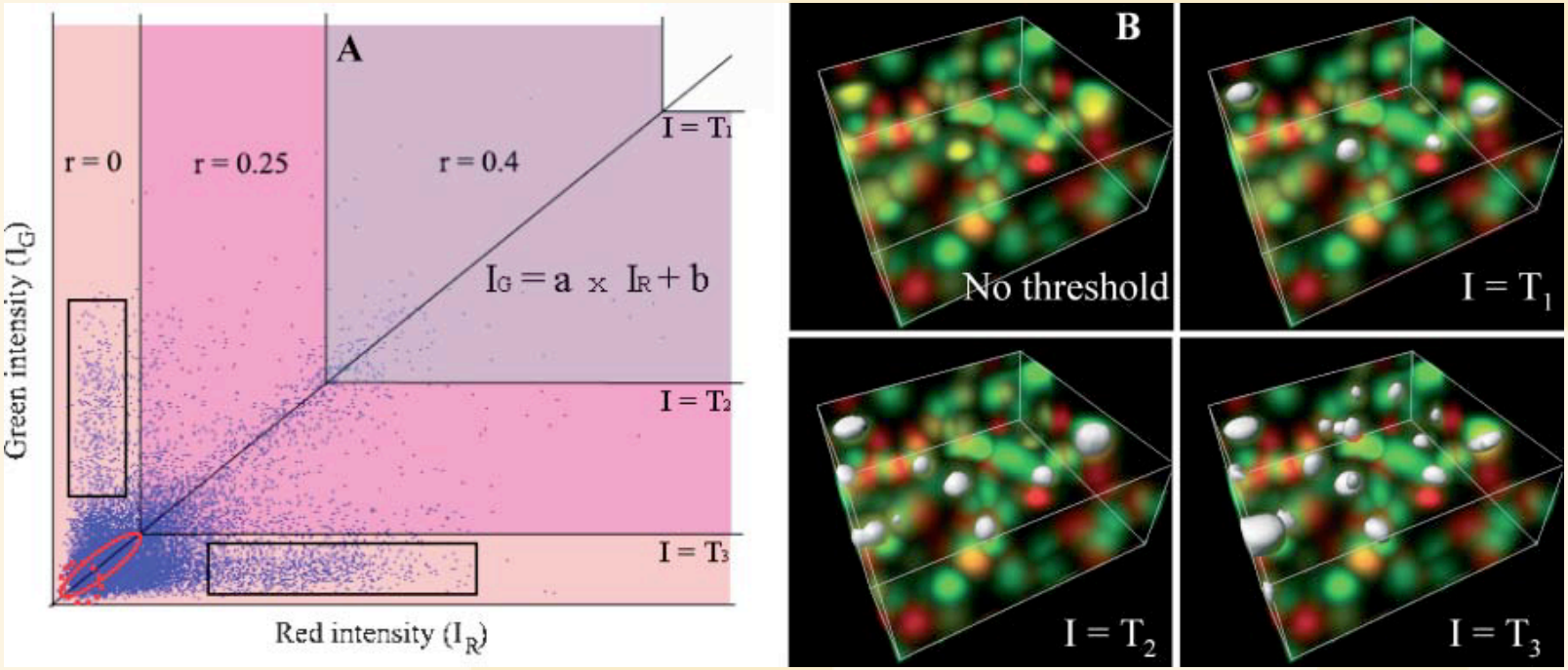
While Pearson's coefficient put a number on two images, the significance of the number remains unclear.
Costes proposes to confront the observed R_p to a distribution of R_p obtained from randomized images

Statistical approach





Costes *et al.* (2004). Biophys J. 86, 3993-4003


A
r = 0
r = 0.25
r = 0.4
I = T1
I = T2
I = T3
IG = a × IR + b
Green intensity (IG)
Red intensity (IR)
B
No threshold
I = T1
I = T2
I = T3

Summary on intensity-based approaches

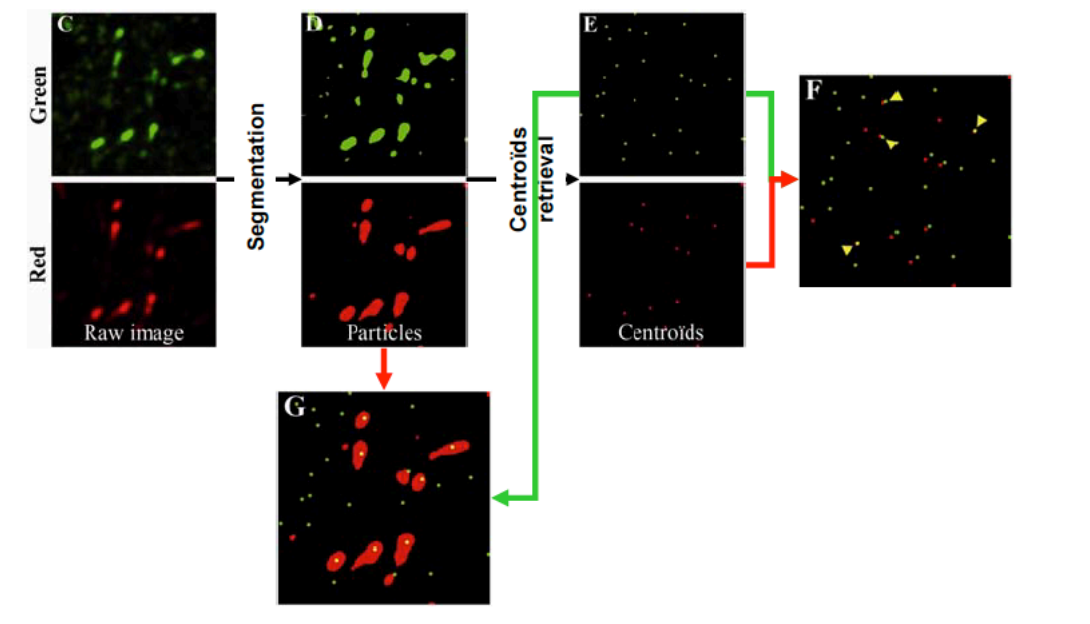
Pearson's coefficient:
- higly sensitive to noise
- Not appropriate if several ratios of proteins exist on the same image
- not easy to highlight exclusion
- not comparable from one couple of images to another

Manders coefficient:
- threshold hard to set, not always accurate if automatically set
- not always comparable from one couple of images to another

Li's approach:
+ easy to highlight exclusion/no colocalization
+ A good visual estimate of the proportion of colocalization
+ A first approach to image normalisation

Coste's approach:
+ The first statistical approach
+ No need to compare a couple of image to another: absolute statement for colocalization
- Need to minimize noise

Object-based analysis



1.Segmentation
2.Centroid retrieval
3.Colocalisation

[Lachmanovich et al. 2003]
[Jaskolski et al. 2005]

Colocalisation Method

Main idea

- 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 statistically significantly small

$$S_A := \{c_{A,i} := (x_{A,i}, y_{A,i}, z_{A,i}), 1 \le i \le N_A\}$$
$$S_B := \{c_{B,i} := (x_{B,i}, y_{B,i}, z_{B,i}), 1 \le i \le N_B\}$$

$$(d_{i,j})_{1 \le i \le N_A, 1 \le j \le N_B}$$

Spot detection

I_A, I_B

Null hypothesis model

R_A, R_B

$p(\log(d))$

Multiple hypo. tests

colocalization ratio

$$r_c := \frac{N_S}{N_A N_B}$$

Microscopie confocale *vs* microscopie à champ large

Microscope confocal

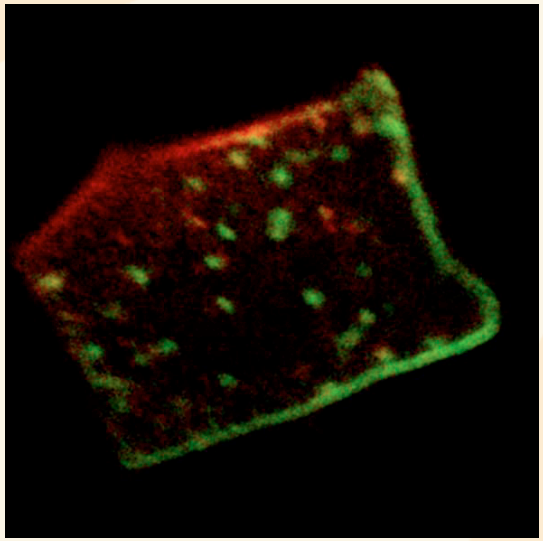


image brute

Microscope à champ large

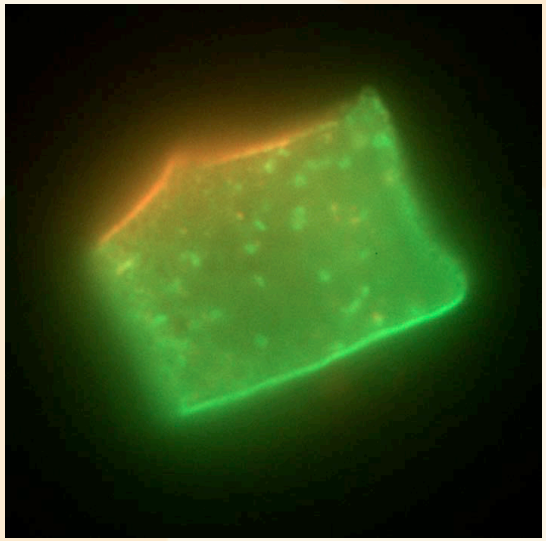


image brute

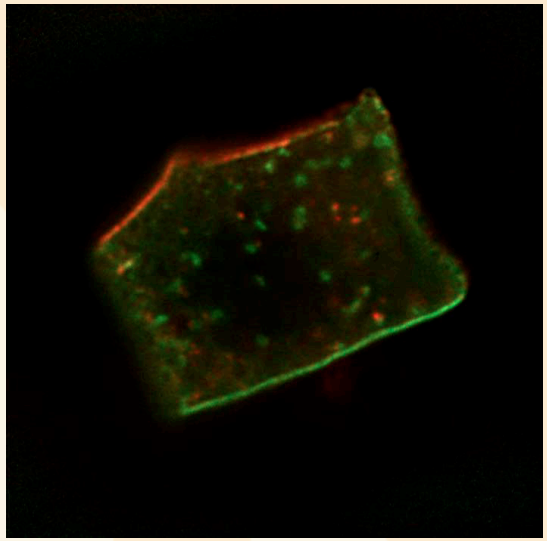
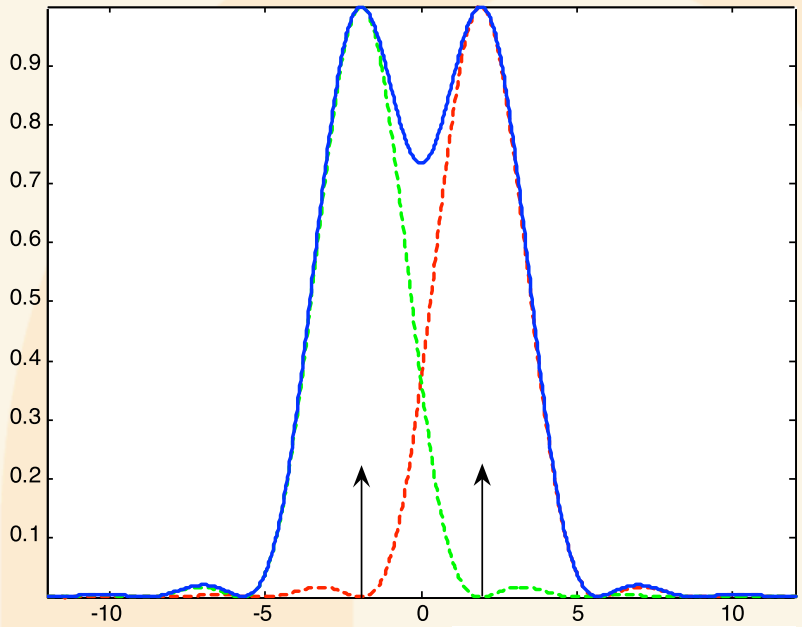


image déconvoluée

Rayleygh's Resolution Criterion

- The "minimal" resolvable distance between **two** symmetrically-placed incoherent point sources of equal amplitudes



- Wide-field PSF (Airy function)

$$d_R := 0.61 \frac{\lambda_{em}}{\mathrm{NA}}$$

Confocal imaging

- **Confocal microscope**
- **The small pinhole: photon-limited imaging**
- **Data model** $\mathbf{X} := (X_i)_{i \in \mathbb{Z}^q}$
  - **Photon counting mode (Poisson)**

$$X_i \sim \mathcal{P}(\lambda_i)$$

  - **Analog mode (Poisson+Gaussian)**

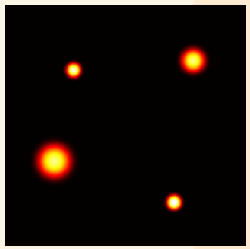
$$X_i = \alpha U_i + V_i, \quad U_i \sim \mathcal{P}(\lambda_i), \quad V_i \sim \mathcal{N}(\mu, \sigma^2)$$

α » *is the overall gain of the detector*

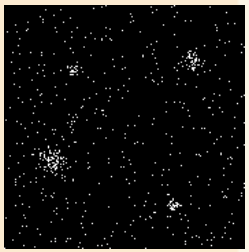
$(U_i)_i$ » *models the photon counting*

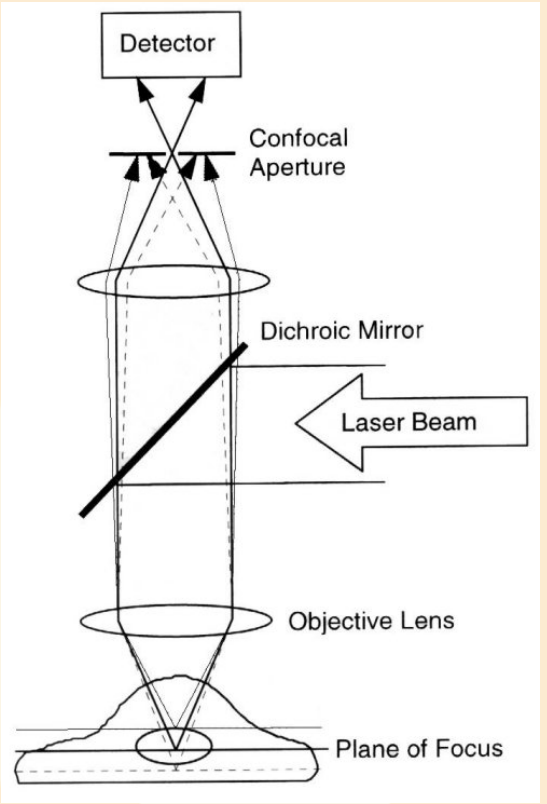
$(V_i)_i$ » *models the readout noise*

intensity image



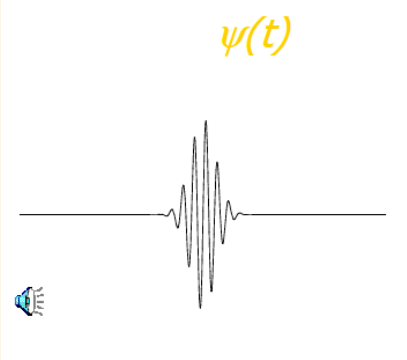
counts (observed)

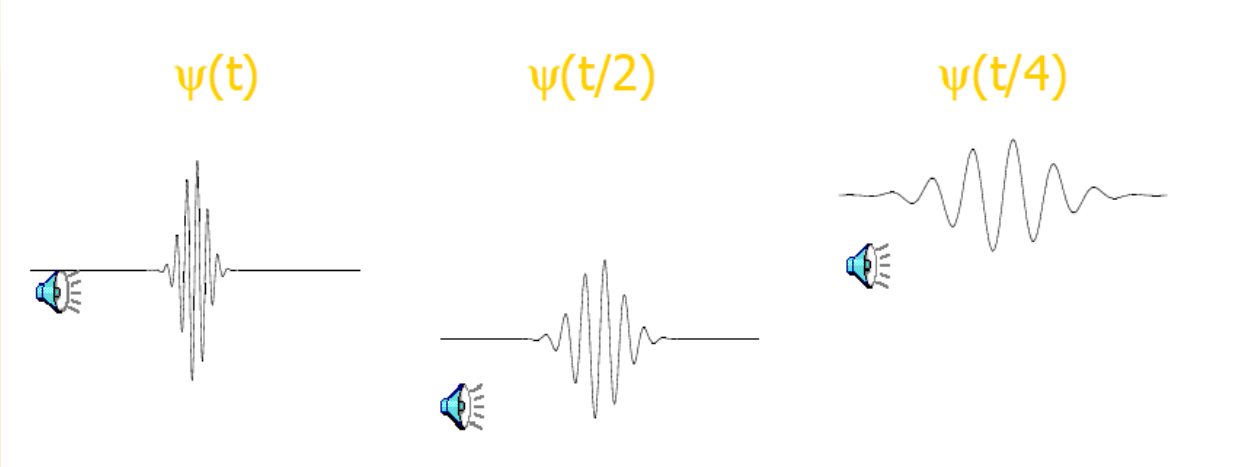




Wavelet transform

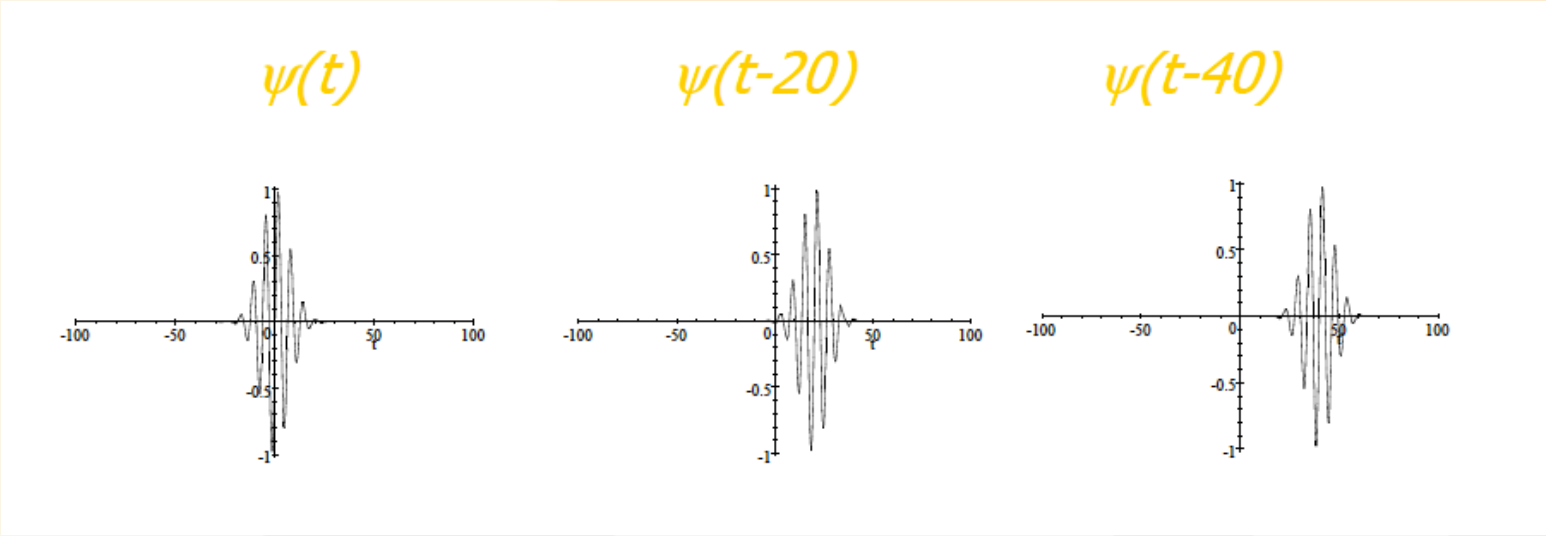
Une ou plusieurs fonctions mères qui engendrent par dilatation et translation la famille d'ondelettes





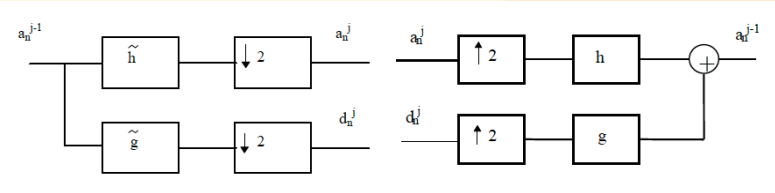
Wavelet transform

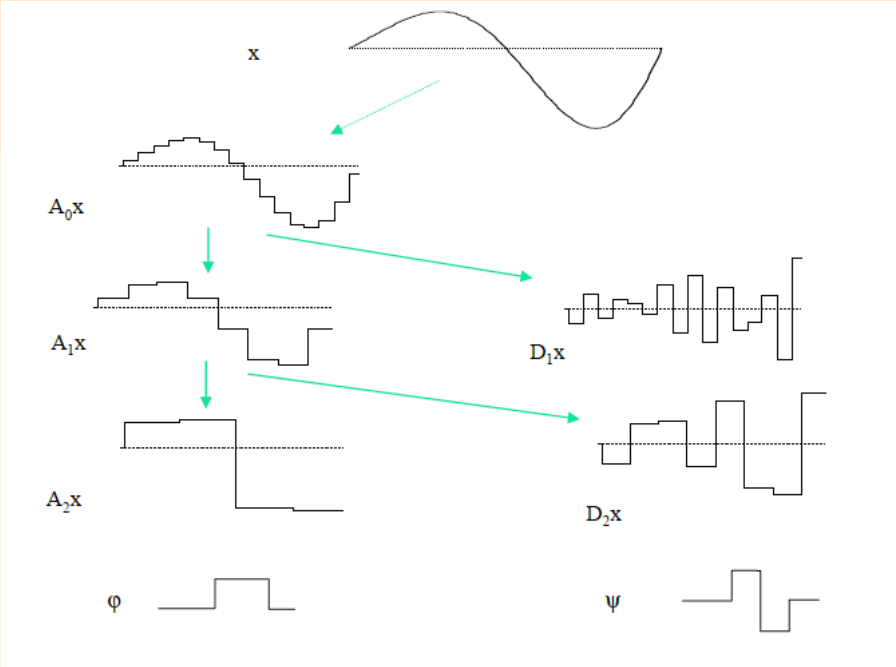
translation



Analyse multi-résolution: base orthogonale

$$A_{i-1}x(t) = A_i x(t) + D_i x(t)$$



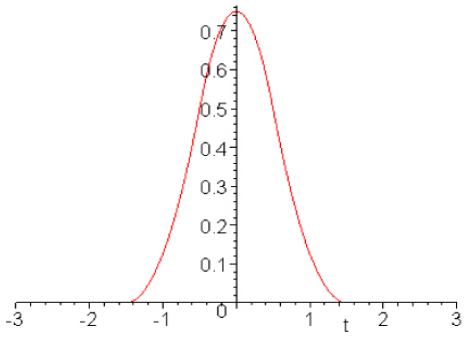


INSTITUT PASTEUR

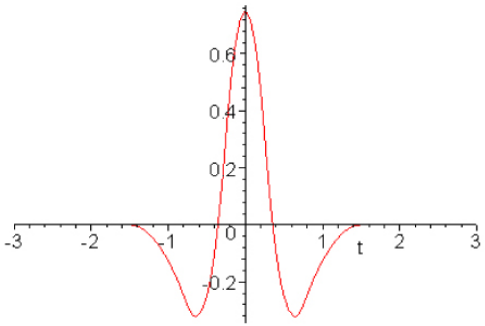
Spot Detection by Wavelet Transform

Features of the wavelet transform

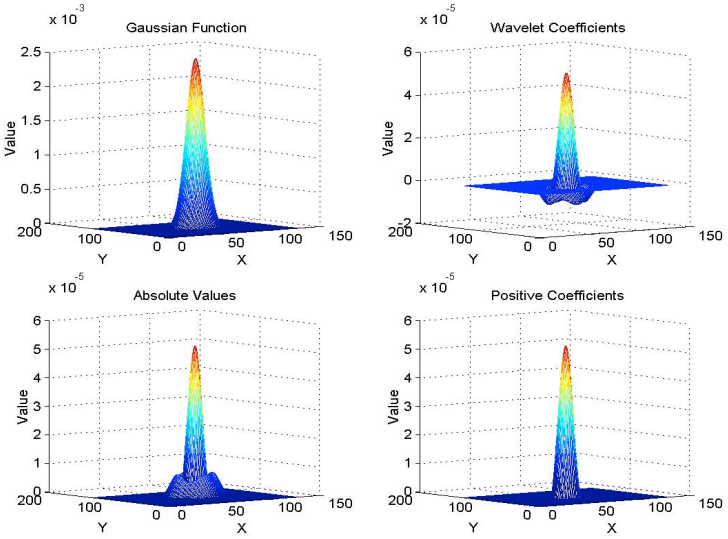
- Based on the convolution of the signal with a family of functions derived from a "mother" function by translation and dilation



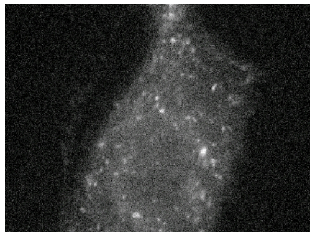
B2 - scale function



B2 - wavelet function

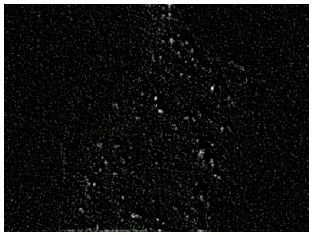


- Analysis tool that cuts up data into different frequency components and studies each with a resolution adapted to its scale

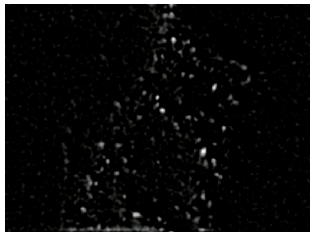


Original

1st scale



2nd scale

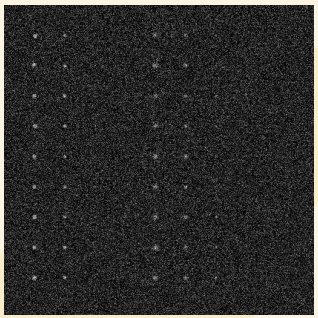


3rd scale

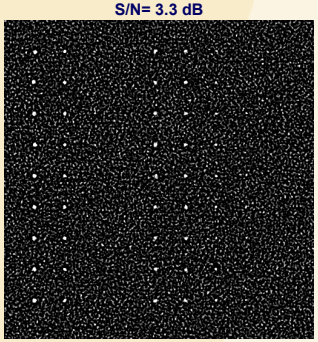
Olivo-Marin,J.-C., ***Pattern Recognition*** (2002)

Spot detection by wavelet Transform

Noisy original image

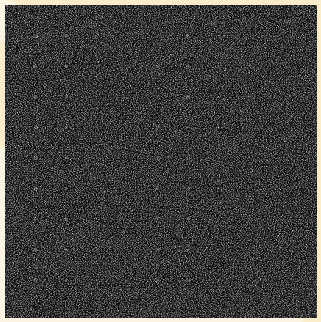


S/N= 3.3 dB

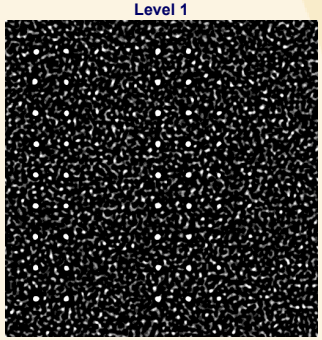


Level 2

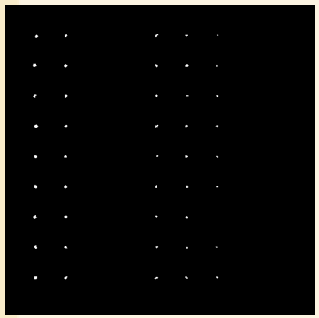
Wavelet bands



Level 1

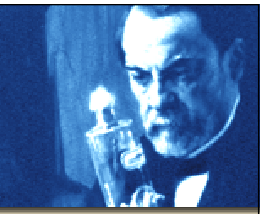


Level 3

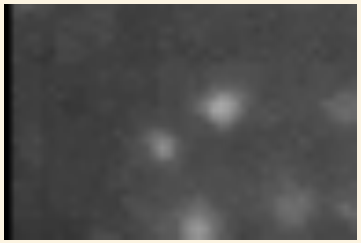


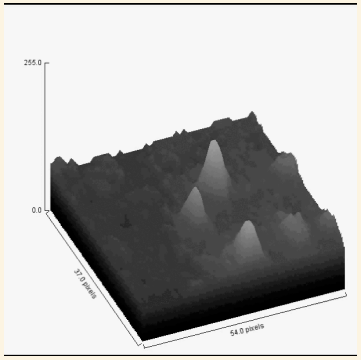
Olivo-Marin,J.-C., *Pattern Recognition* (2002)

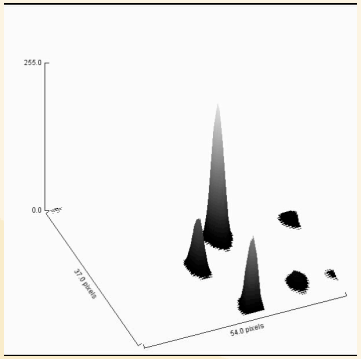
Spot Detection

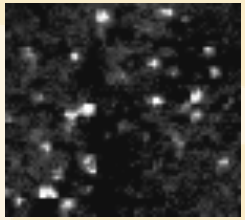


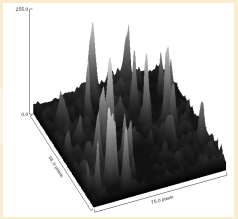
Noise reduction

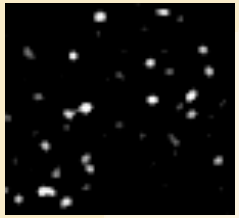


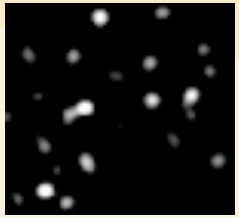


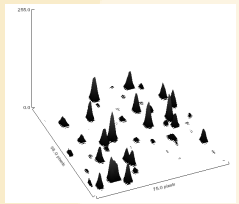


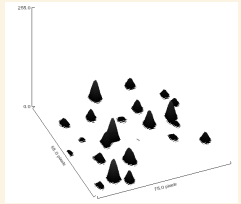


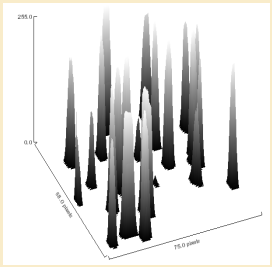




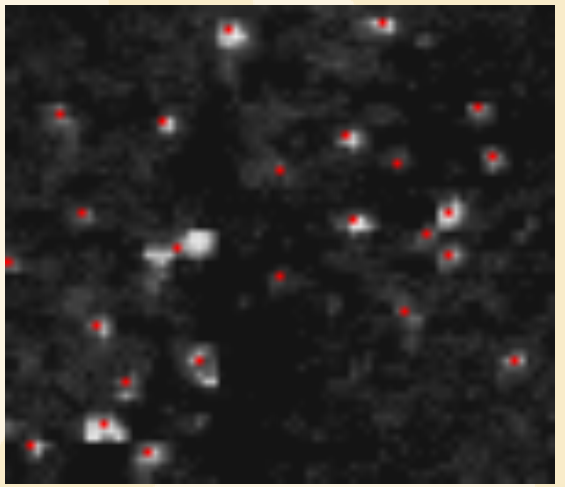








Thresholding



Olivo-Marin,J.-C., *Pattern Recognition* (2002)

Zhang et al. 2007

Feature Adapted Detection with Wavelet Transform

Step 2: Null Hypothesis Model

- **Simulation from the detections**
- **Null model:** $(d_{i,j})_{i,j}$ are observed from the object centers S_A and S_B which are independently and uniformly randomly distributed in the supports of protein A (R_A) and of protein B (R_B), respectively
- **Null model describes the situation where colocalizations can occur only by chance**
- Currently, R_A and R_B are both supposed to be the cell support, but can be refined using prior information
- The null distribution of d is estimated by a Parzen window method (Gaussian kernel estimator) applied on distances drawn from the null model
  - $p(\log(d))$ is estimated instead of $p(d)$ to avoid instability at the boundary (d = 0)

Step 3: Multiple Hypothesis Tests

- Test the observed distances against the null distribution
- Multiple hypothesis tests
  - Controlling FamilyWise Error Rate (FWER)
    - The probability of erroneously rejecting even one null hypothesis
    - Highly conservative
  - Controlling False Discovery Rate (FDR) [Benjamini and Hochberg 1995, Benjamini and Yekutieli 2001]

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

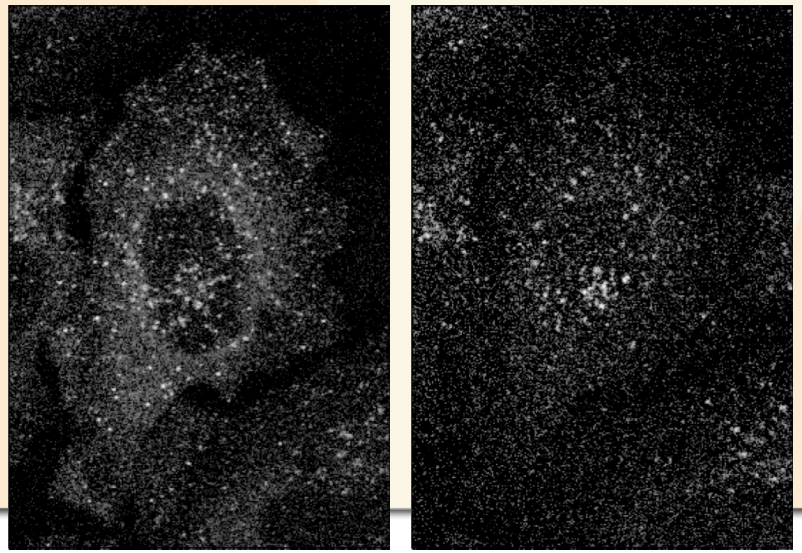
    - Usually have a high detection power
    - Can easily handle dependent statistics (e.g. the observed distances)
- Colocalization ratio computation
  - If FDR ≤ β then and if K hypotheses have been rejected, we will have at least $K(1-\beta)$ correct decisions on average

$$r_c := K(1-\beta)/(N_A N_B)$$

Applications Context

- How an unknown protein X is localized during the endocytic process
  - Physiological function and the location of a protein are highly related
  - Location reveals the information on the protein's biological role
  - 5 proteins (P_1, P_2, P_3, P_4, P_5) as markers of cellular compartments
  - Co-immunofluorescence labeling for (X, P_i), showing bright "spots"
  - Protein-protein colocalization (association) analysis

Hela cells (a slice view of the confocal volume)



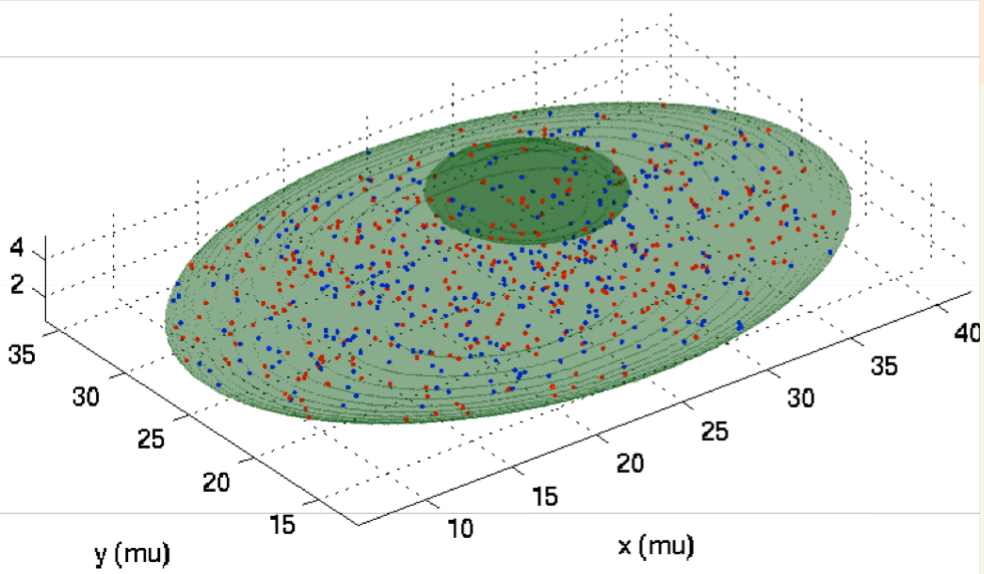
I_A: protein X I_B: protein P_1

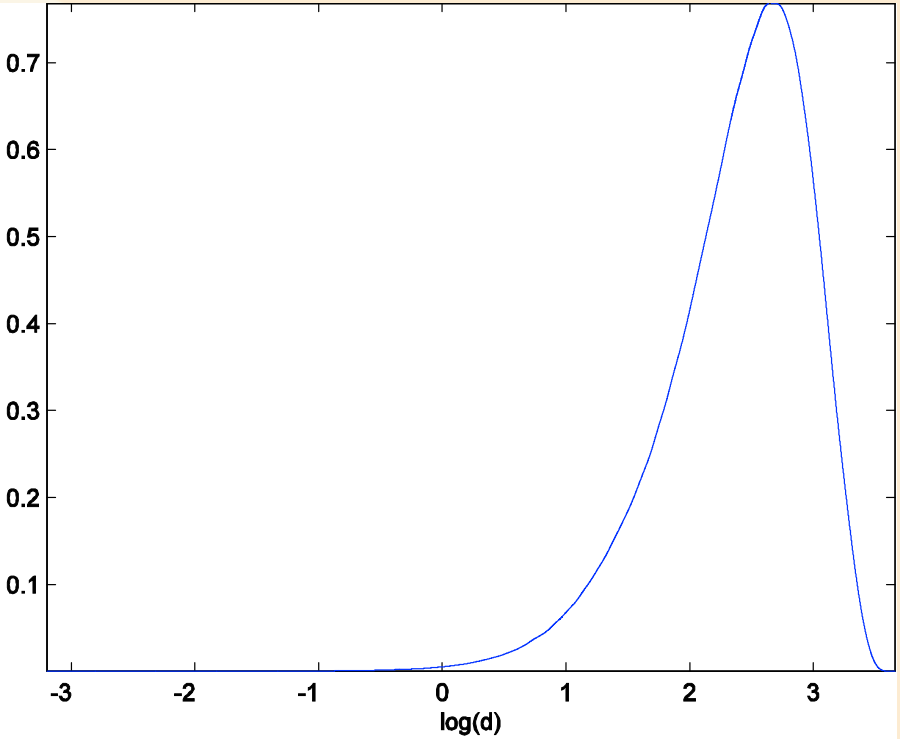
INSTITUT PASTEUR

Results (Simulated 3D images)

Colocalization specificity under the null model

Simulated Hela cell





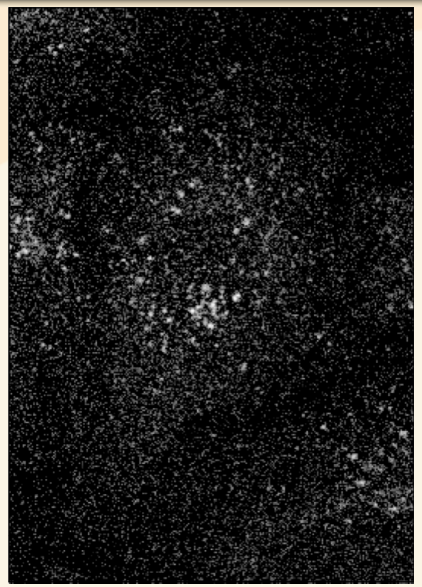
$p(\log(d))$

- About 300 virtual detections for each of the 2 proteins;
- 10 replications with each having approximately 10^5 tests
- FDR controlled at $\beta = 0.5$

Not a single colocalization (false positive) detected

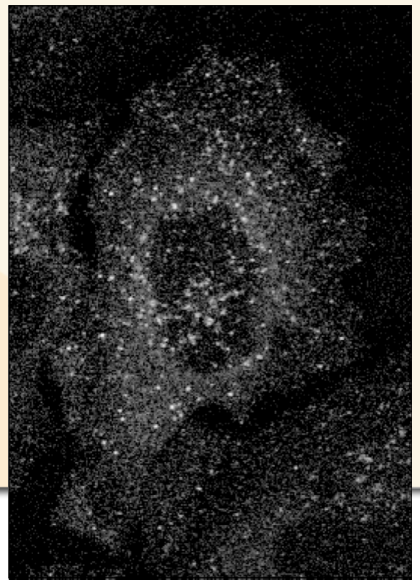
INSTITUT PASTEUR

Results (Real 3D images)

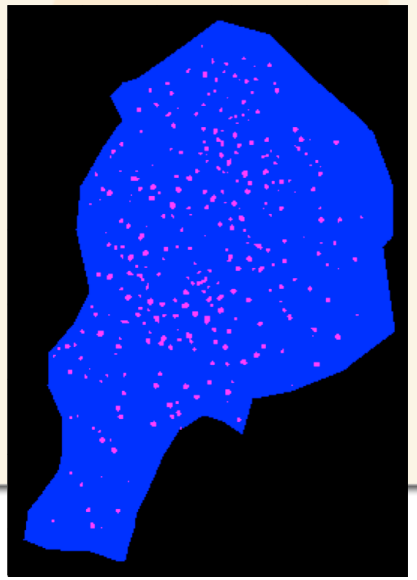


Protein P_1

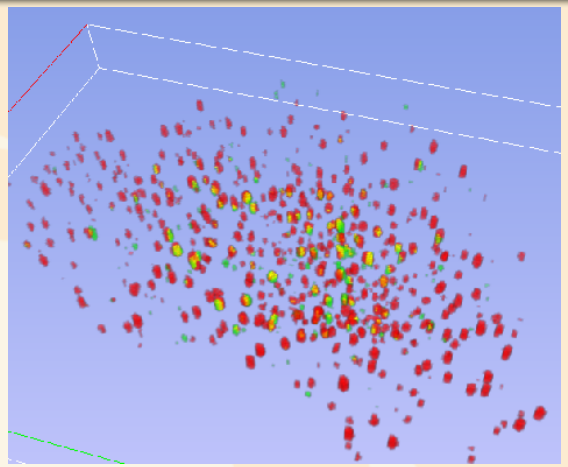
Detected spots

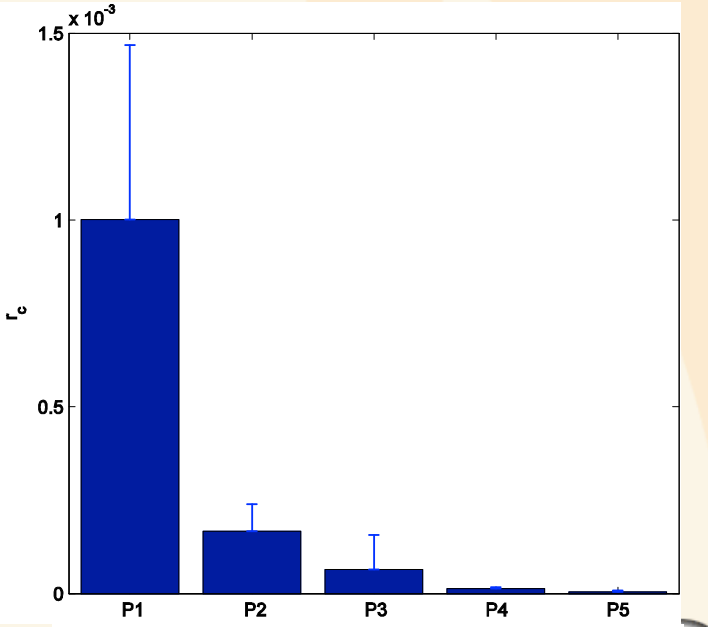


Protein X



Detected spots





Colocalization ratios with different proteins
FDR controlled at $\beta = 0.5$

INSTITUT PASTEUR

Total Internal Reflection Fluorescence Microscopy

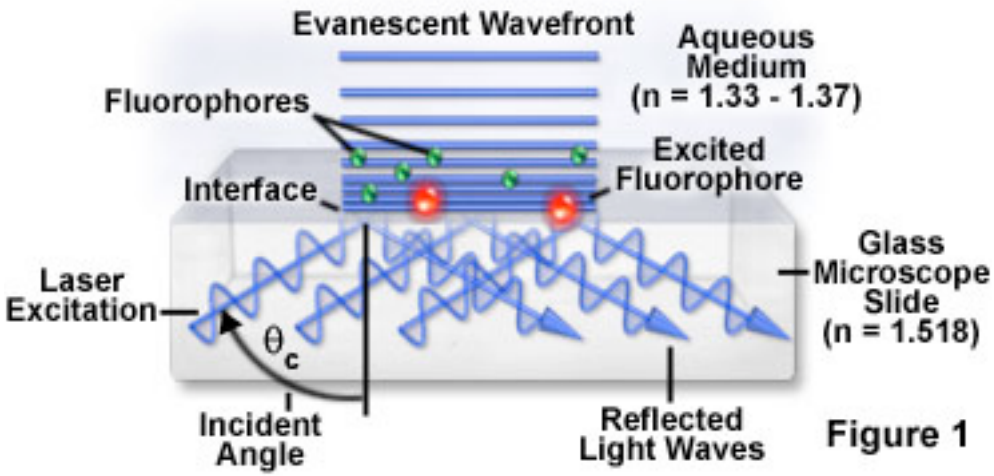
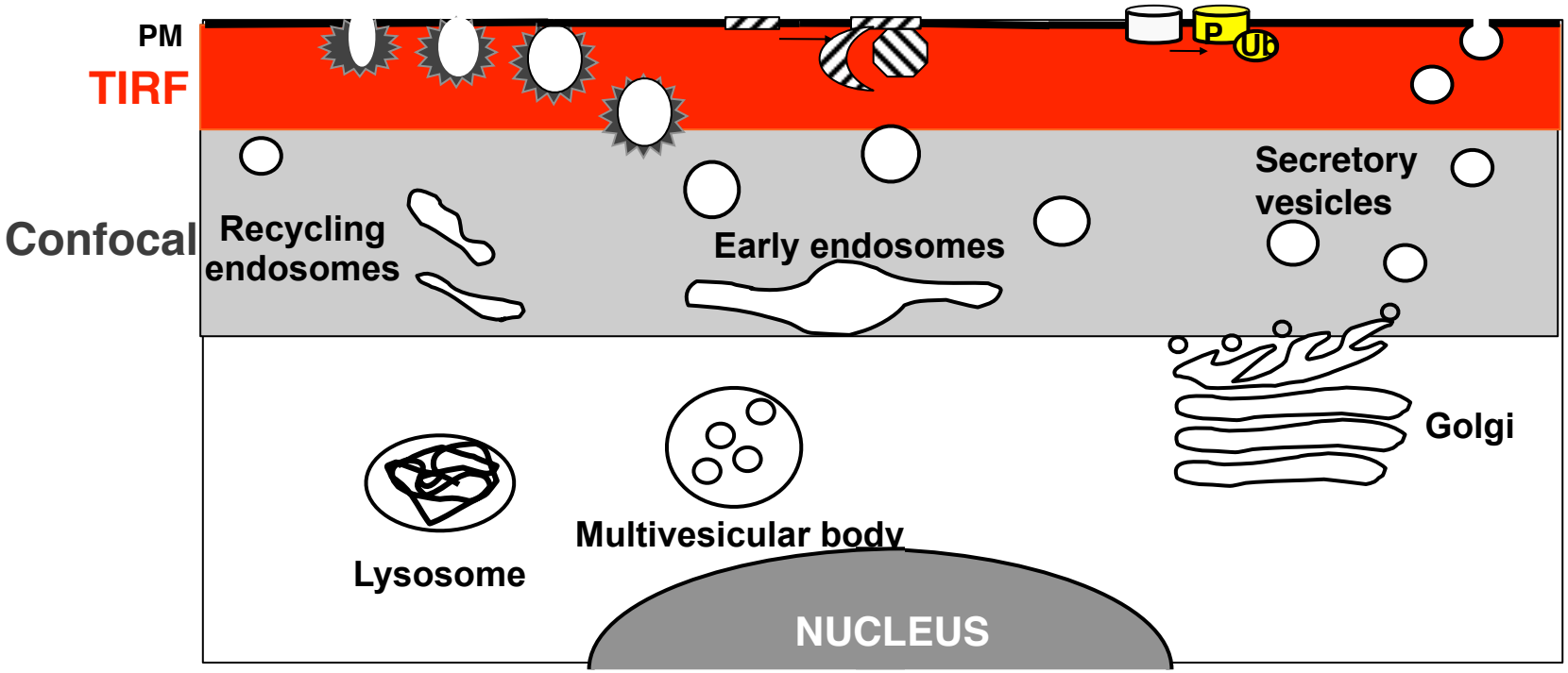
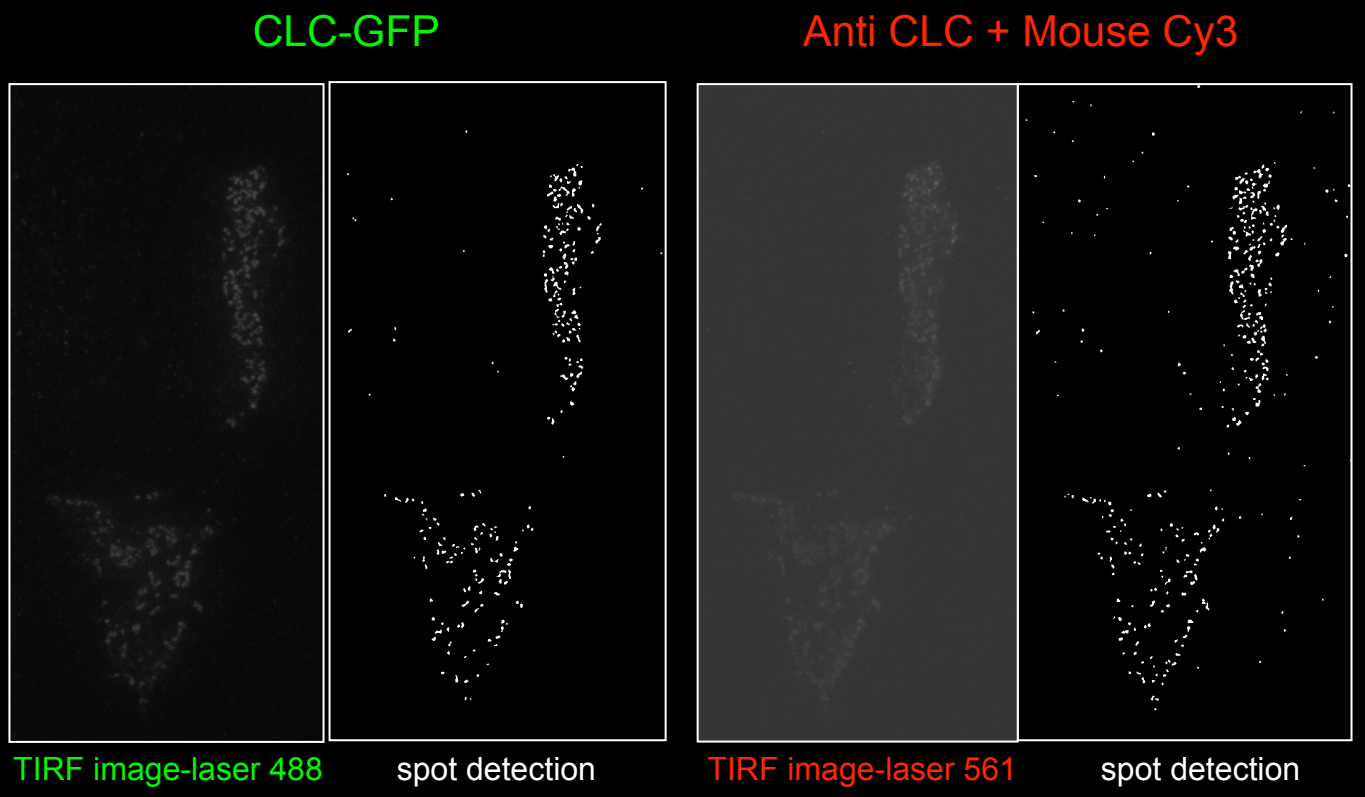


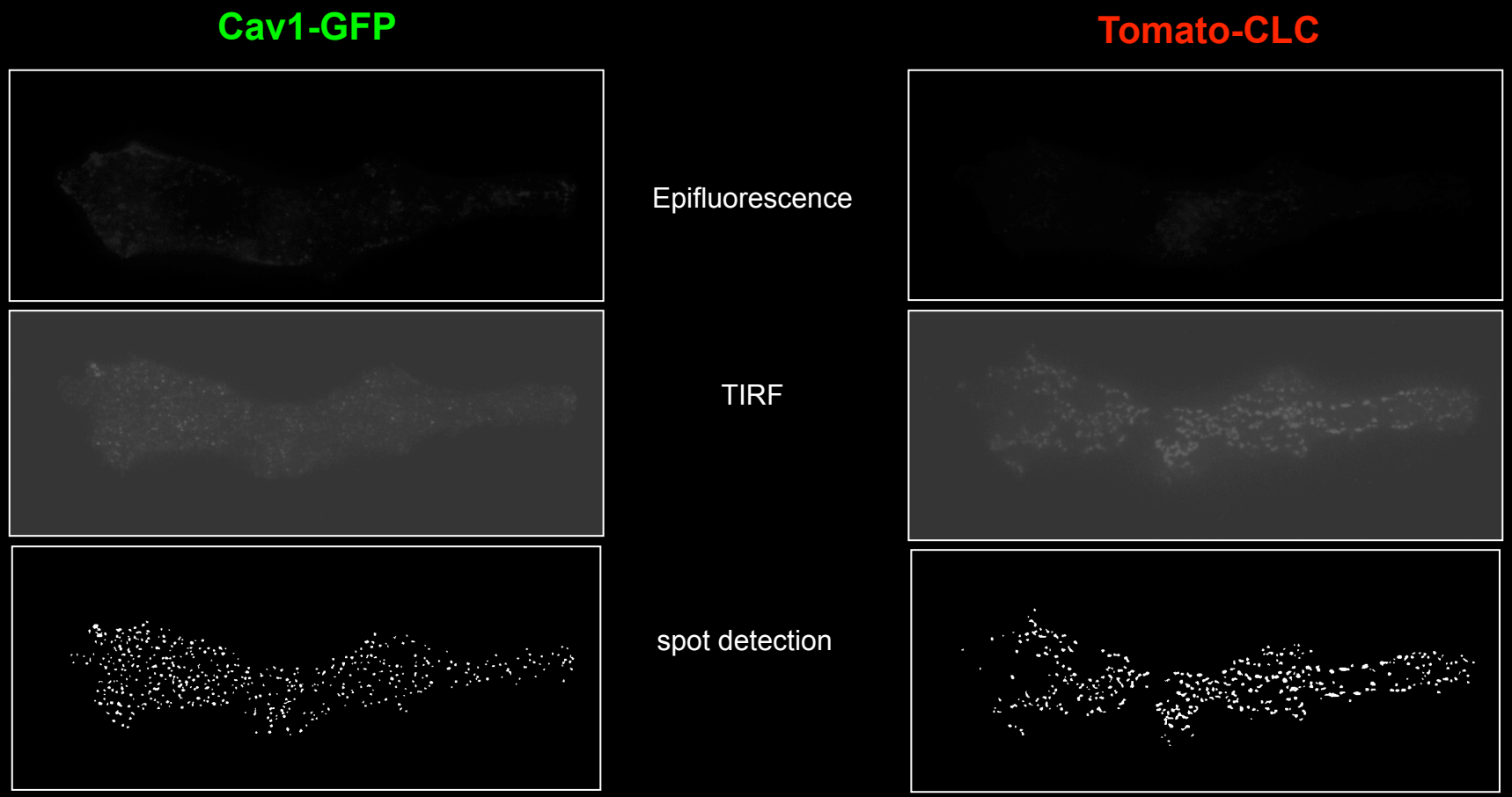
Figure 1



Our analysis of the positive control: CLC-GFP/anti-CLC

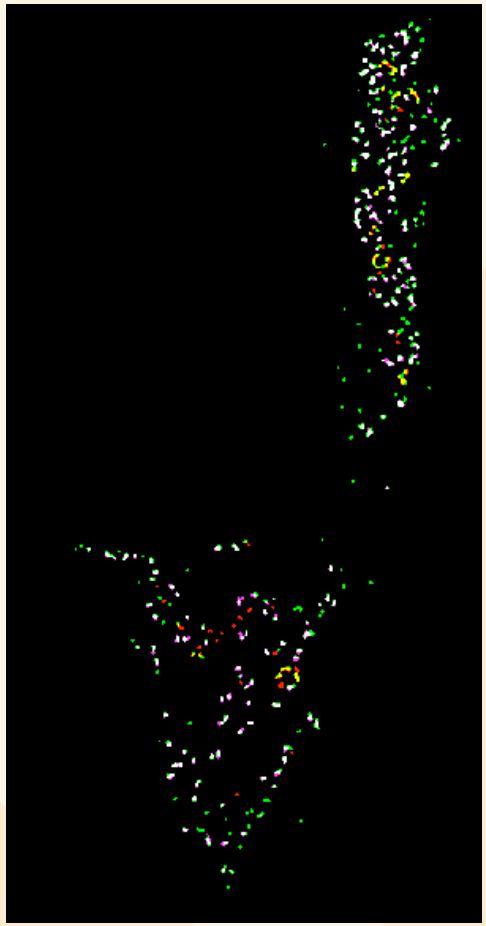


Our Analysis of the negative control: Caveolin and clathrin light chain(CLC)



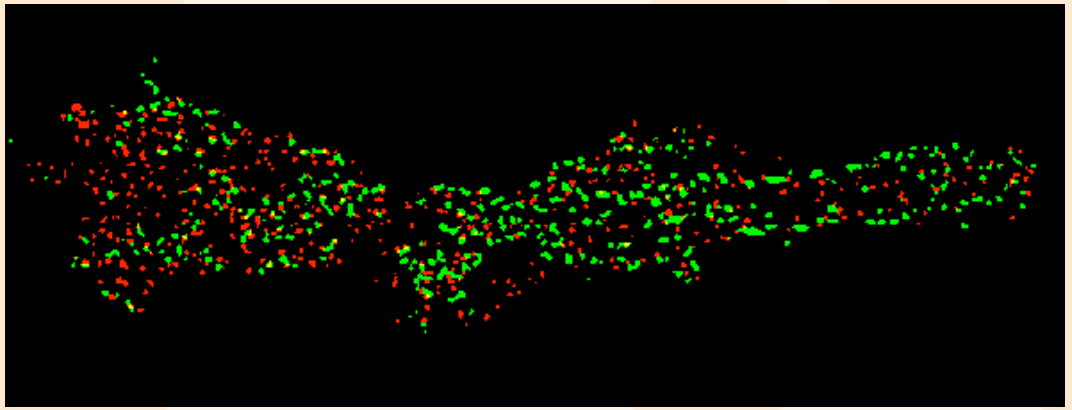
Results

Positive: >80%



Colocalization Controls

Negative: <5%



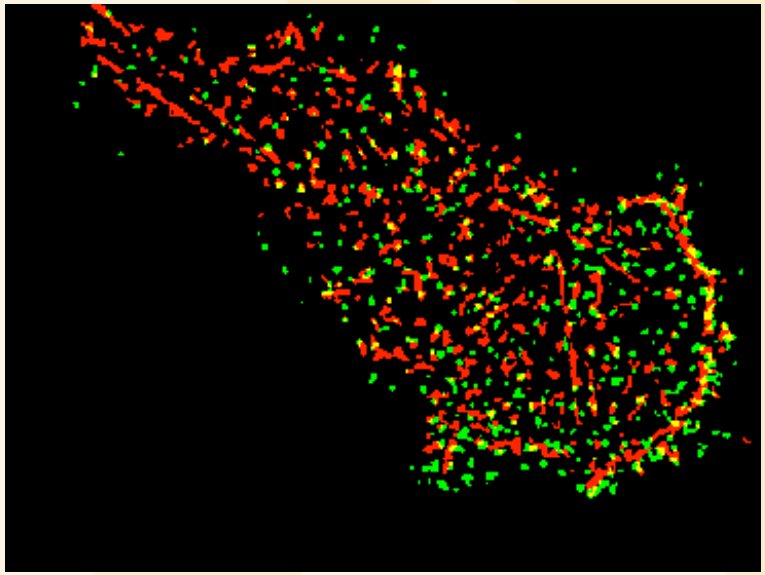
INSTITUT PASTEUR

Conclusion 1

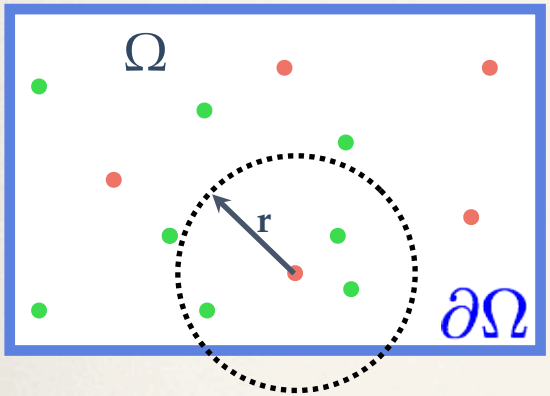
- A novel statistical colocalization approach
  - Multiscale spot detection
  - Null model generation
  - Multiple hypothesis tests controlling FDR
- The method has a good specificity

• This approach is well adapted for microscopic point-like object, but not for anisotropic objects.
• Time consuming

- Ripley'K cross function



Statistics: Ripley's K cross function $K_{12}(r)$



A_1 *spots*, A_2 *spots*

2 proteins: A_1 and A_2.
Statistics on the number of A_2 spots **closer than** ***r*** from A_1 spots.

Average number of A1-A2 neighbors closer than r

$$K_{12}(r) = \frac{|\Omega|}{n_1} \sum_{\mathbf{x} \in A_1} \frac{1}{n_2} \sum_{\mathbf{y} \in A_2} \mathbf{1}_{\{|\mathbf{x}-\mathbf{y}| \leq r\}} b\left(|\mathbf{x} - \partial\Omega|, |\mathbf{x} - \mathbf{y}|\right)$$

boundary correction

Problems: -Specificity (with no extensive simulations !) ?
-Interpretation: number of real colocalizations ? Length scale?

Specificity: Asymptotic normality of $K_{12}(r)$

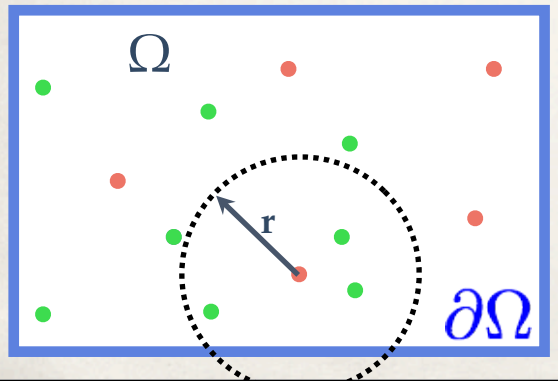
If A_2 spots are randomly distributed (for any A_1 spots)

$$\hat{K}_{12}(r) \underset{n_2 \gg 1}{\longrightarrow} \mathcal{N}\left(\pi r^2, \sqrt{\mathrm{var}\{\hat{K}_{12}(r)\}}\right)$$

with

$$\mathrm{var}\left\{\hat{K}_{12}(r)\right\} = \frac{|\Omega|}{n_1^2 n_2}\left(\sum_{\mathbf{x}_1 \in A_1} H_1\left(|\mathbf{x}_1 - \partial\Omega|, r\right) + \sum_{\mathbf{x}_1 \neq \mathbf{x}_2 \in A_1} H_2\left(|\mathbf{x}_1 - \mathbf{x}_2|, r\right)\right)$$

Boundary correction ***Correlations between A_1 spots***



A_1 spots, A_2 spots

A statistical test of proteins colocalization

$$\frac{\hat{K}_{12}(r) - \pi r^2}{\sqrt{\mathrm{var}\left\{\hat{K}_{12}(r)\right\}}} = \tilde{K}_{12}(r) > q_\delta \longrightarrow \textbf{Pr\{«real» colocalization\} > } \delta$$

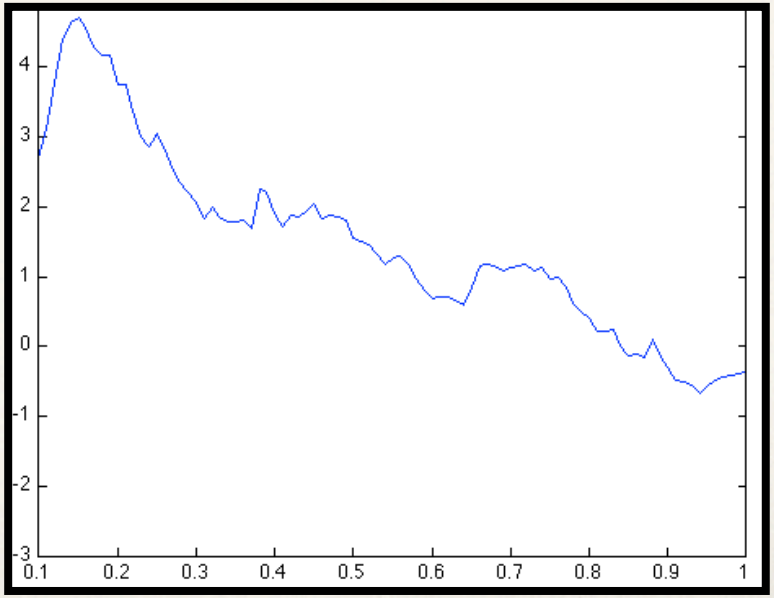
quantile of standard normal law N(0,1) at level δ

A statistical test of proteins colocalization

$$\frac{\hat{K}_{12}(r) - \pi r^2}{\sqrt{\mathrm{var}\left\{\hat{K}_{12}(r)\right\}}} = \tilde{K}_{12}(r) > q_\delta \longrightarrow \textbf{Pr\{«real» colocalization\} > } \delta$$

quantile of standard normal law N(0,1) at level δ

$\tilde{K}_{12}(r)$



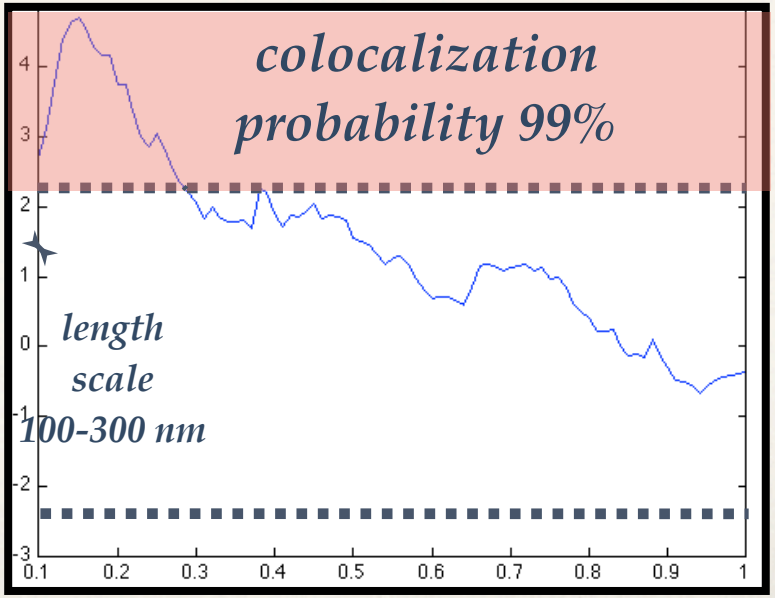
r

A statistical test of proteins colocalization

$$\frac{\hat{K}_{12}(r) - \pi r^2}{\sqrt{\mathrm{var}\left\{\hat{K}_{12}(r)\right\}}} = \tilde{K}_{12}(r) > q_\delta \longrightarrow \Pr\{\text{«real» colocalization}\} > \delta$$

quantile of standard normal law N(0,1) at level δ

$\tilde{K}_{12}(r)$

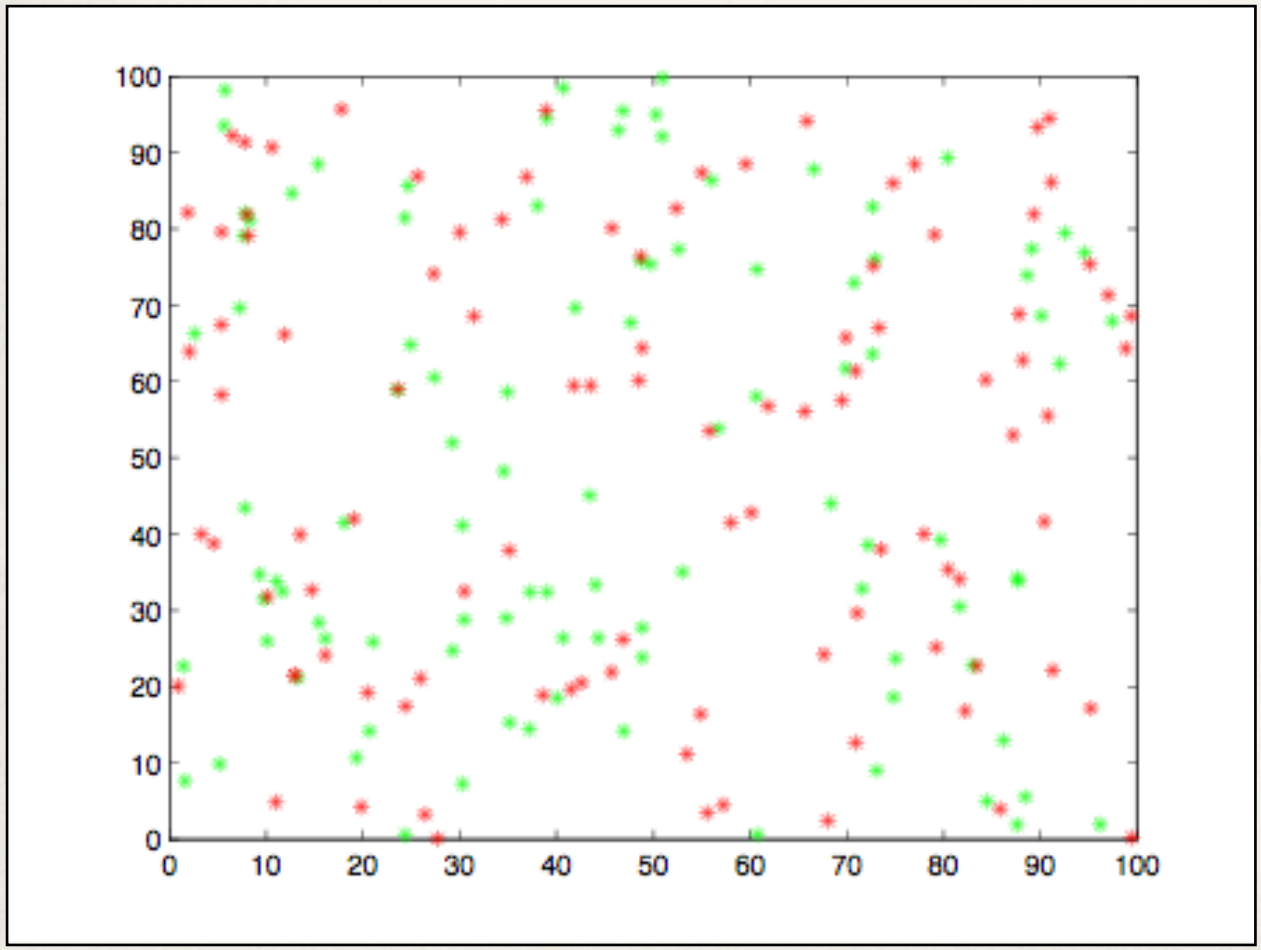


$q_{0.99}$=2.32

$q_{0.01}$=-2.32

r

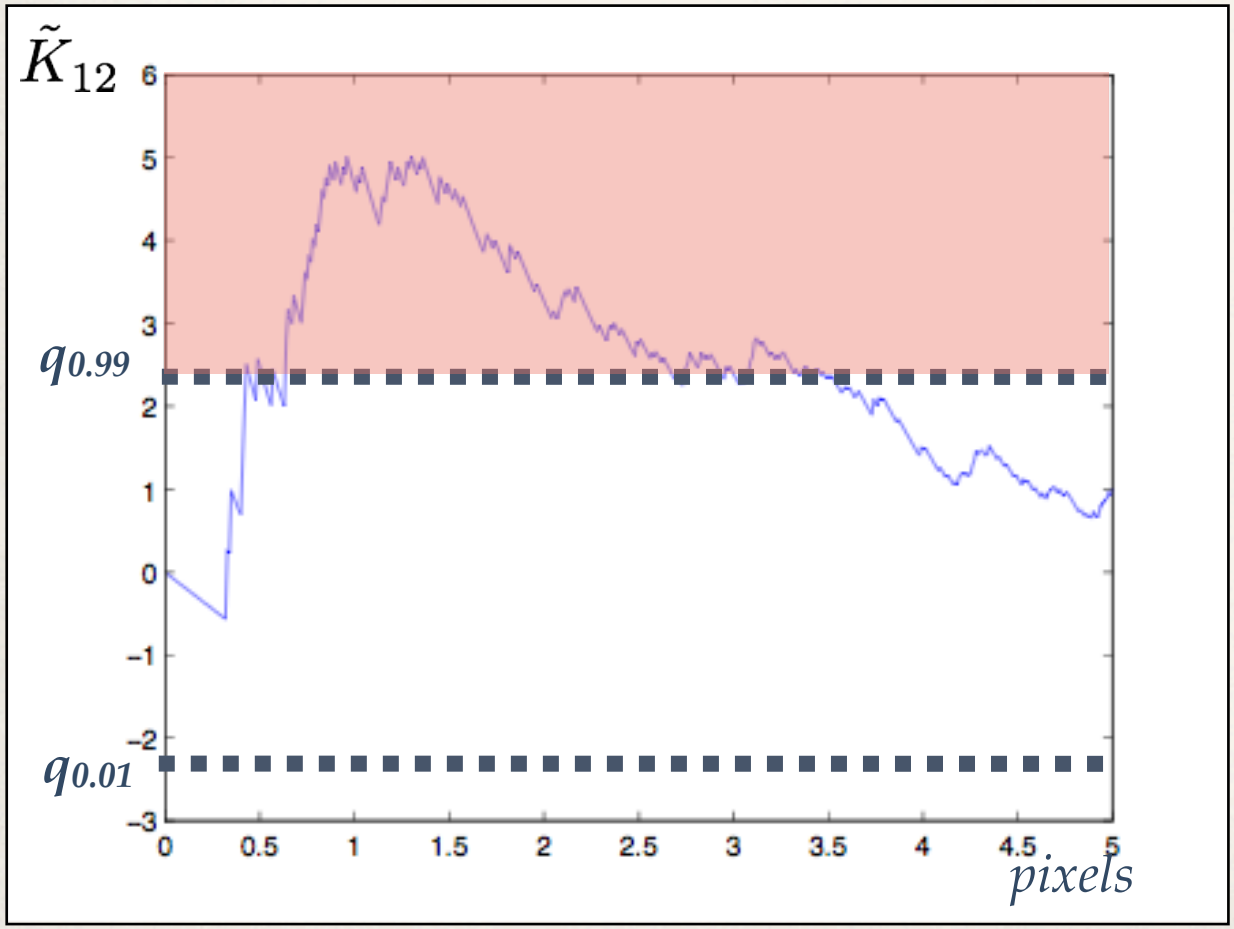
Test against synthetic data



Colocalization ?

$\alpha > 0$?

Test against synthetic data

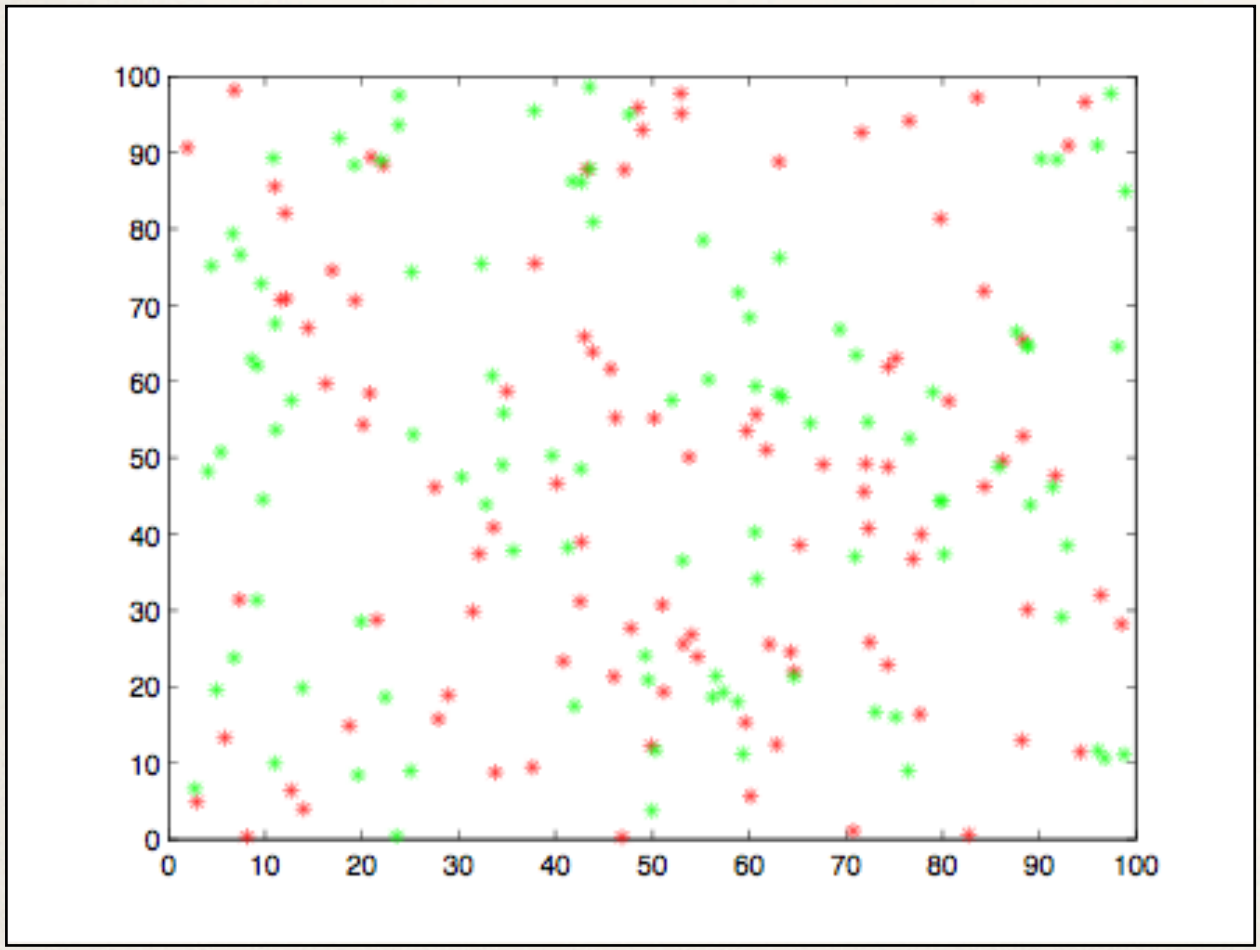


Colocalization ?

YES !

$\alpha = 0.2$

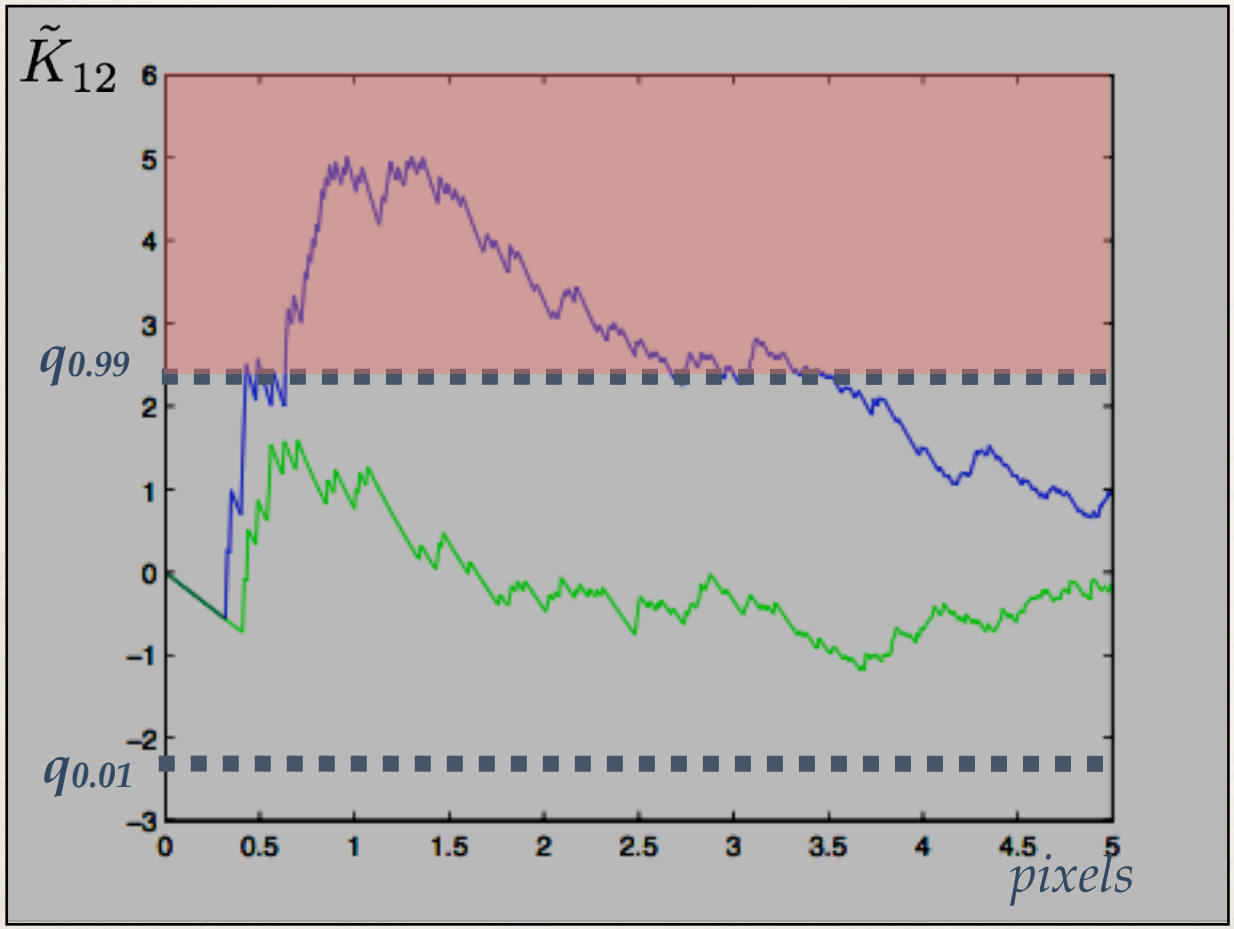
Test against synthetic data



Colocalization ?

$\alpha > 0$?

Test against synthetic data

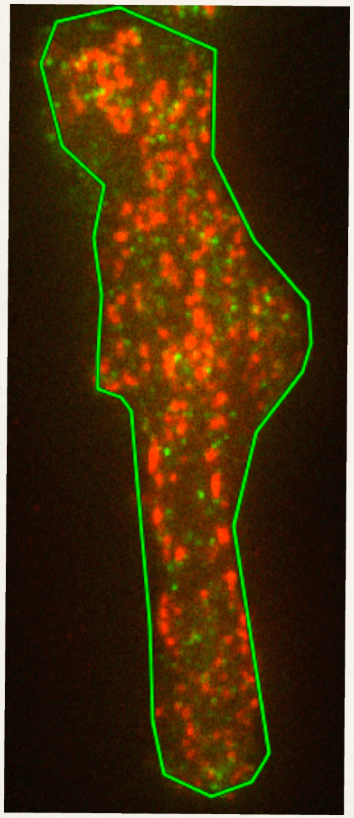


Colocalization ?

NO !

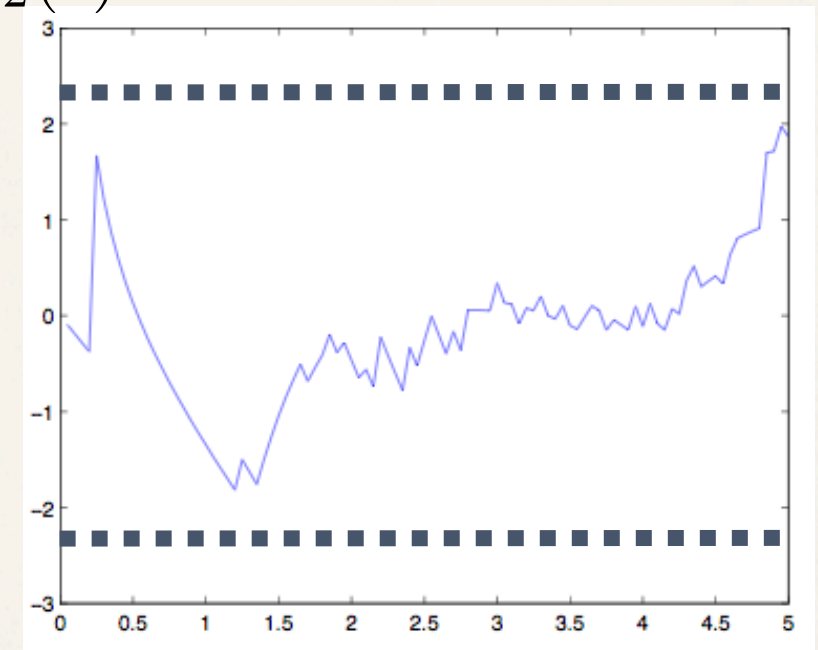
$$\alpha = 0$$

Negative control: Caveolin (fluo. Antibody) *vs* Clathrin (GFP)



Caveolin (fluo. Antibody) – Clathrin (GFP)

$K_{12}(r)$

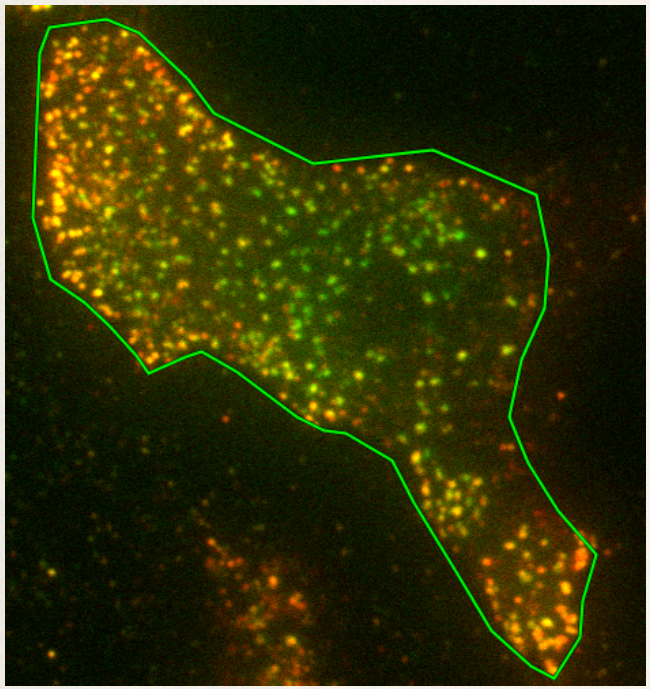


$z_{0.99}$=2.32

$z_{0.01}$=-2.32

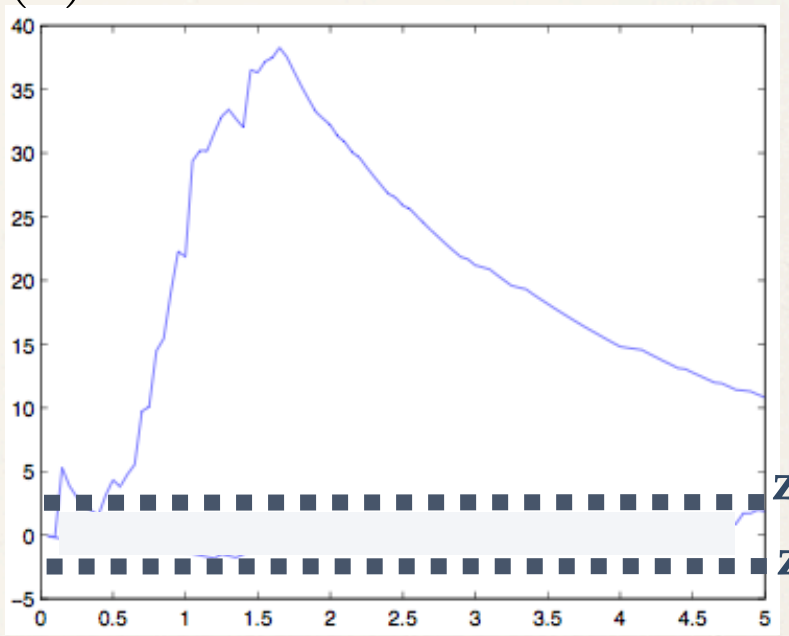
No colocalization

Positive control: Clathrin (fluo. Antibody) *vs* Clathrin (GFP)



Clathrin (fluo. Antibody) –Clathrin (GFP)

$K_{12}(r)$



Strong colocalization

Conclusion: Problems of pixel-based methods and solutions

Problem 1 - Sensitivity to noise

Solution 1 - Object-based method: proteins spots detection with elaborate algorithms and statistics on spots inter-distances

Problem 2 - Specificity (true *vs* false colocalizations)

Solution 2- Analytical formula for the level of significance of $K_{12}(r)$

Problem 3 - Colocalization parameters (distance scale and stœchiometry)

Solution 3- Fitting parametric models to $K_{12}(r)$ curve. **In progress !**

Distance scale $d \sim \mathcal{N}(\mu, \sigma)$, or ...