A Probabilistic Method to Quantify the Colocalization of Markers on Intracellular Vesicular Structures Visualized by Light Microscopy

Yannis Kalaidzidis*,†, Inna Kalaidzidis* and Marino Zerial*

*Max Planck Institute of Molecular Cell Biology and Genetics, 01307, Dresden, Germany
†Faculty of Bioengineering and Bioinformatics, Moscow State University, 119991, Moscow, Russia

Abstract. The intracellular localization of proteins to their specific compartments is a rich source of information for the study of biological processes. The colocalization of a protein with an established compartment marker is routinely measured from multi-color fluorescent microscopy images. Unfortunately, the apparent colocalization is a mixture of real and random colocalization. Random colocalization results from the limited resolution of the light microscope and the close location or occlusion of objects in the crowded cytoplasmic environment. Commonly used methods for the correction of random colocalization work well if the random colocalization is significantly smaller than the real colocalization. In the case where they have comparable values, the final result could be a senseless negative value. To solve this problem, we first developed a probabilistic model for the estimation of random colocalization and demonstrated that it produces results that coincide with the standard scramble method. Second, we developed a probabilistic model for the correction of random colocalization for the double and multiple colocalization of intracellular markers on vesicular structures. Our probabilistic method of estimation of real colocalization has two main advantages: 1) it never gives a negative colocalization value and 2) provides the estimation of colocalization uncertainty.

Keywords: Quantitative microscopy, Colocalization, Multi-channel colocalization
PACS: 87.10.Mn

INTRODUCTION

The temporal-spatial distribution of proteins in cells can be deduced from the visualization of fluorescent protein chimeras and fluorescently labeled antibodies. The intracellular localization with respect to established subcellular compartment markers is an important parameter to determine the potential site of function. Colocalization is typically quantified as a spatial correlation of marker intensities on microscopy images [1-3] or as number of the overlapped pixels of binarized by threshold images [4]. However, the parameters used in most systems biology models are not correlations between markers A and B but the proportion of marker A colocalized with marker B. In the case of distinct vesicular structures, e.g, endosomes, the event of interest is the presence of proteins A and B on the same structure (Fig.1A). However, the two proteins may be present in different amounts on the endosomes and only partially overlap on microscopy images. The partial overlap is the result of different stoichiometry of markers, their domain localization in the endosome (Fig.1B), and optical aberrations of the microscope
The influence of optical aberration on the apparent object colocalization can be measured experimentally by imaging sub-micrometer-size multi-color beads. Optical aberrations, difference in the size of diffraction spots, difference in laser intensity and sensitivity of microscope detectors decrease the overlap of images of beads on different channels (color planes) (Fig.2B). In order to handle the partial overlap of markers on vesicles, we used object-based colocalization [5, 6], where the whole marker spot is considered to be colocalized if the relative overlap (ratio of the overlapped area between the spot of protein A and that of protein B to the total area of spot of protein A) is above a given threshold. Then, the object-based colocalization is defined as the ratio of the integral fluorescence intensity of colocalized objects to the total integral intensity of all objects. Clearly, the colocalization depends on the threshold value. For example, the threshold equal to 1 results in zero colocalization even for multi-color beads (Fig.2). A reasonable value for threshold can be estimated from the dependency of the calculated colocalization on threshold values for multi-colour beads (Fig.2B), where the real colocalization is known to be 1.

**FIGURE 1.** Fluorescent image of endosomes. Endosomes are labelled by two specific endocytic machinery proteins (APPL1 and EEA1) and two cargos (EGF - signalling molecule, and LDL - nutrients). **A.** Images of HeLa cells with zoomed split channels (upper right panel) and segmented objects (bottom right panel). In the zoomed image one can see that markers which apparently share one endosome do not overlap completely. **B.** Super-resolution dSTORM image of single endosomes, where the distinct localization of markers EEA1 and EGF on endocytic membrane, as well as bulk localization of LDL is clear seen. Nevertheless, from the point of view of the endocytic traffic these markers are colocalized (on one endosome).

In addition to the aforementioned optical aberrations that lead to decrease in colocalization, the diffraction limit of light microscopy (especially in depth) and the presence of background fluorescence result in a significant apparent (by-chance) increase in colocalization in the highly crowded environment of the cytoplasm (Fig.1A). The colocalization resulting from close apposition and occlusion of objects as a matter of
chance is referred in the literature as “random colocalization” [7, 8]. In order to estimate the real colocalization value, one has to estimate the random colocalization and then use it for the correction of the measured colocalization.

Common practice for random colocalization estimation is to calculate the colocalization of the randomly scrambled blocks of images [7]. The problem for an automated scrambling algorithm is the high heterogeneity of object density within cells. If a scrambling procedure replaces blocks with significantly different densities of objects, then the estimation of random colocalization will be biased. Therefore, in order to avoid the scrambling approach, we developed a theoretical model for random colocalization estimation.

Suppose that a 2D image has two types of independent objects: “red” and “green” and we want to estimate their random colocalization. The “red” object is colocalized to the “green” one if the ratio of overlapped area to the total area of “red” object is above given threshold. For the sake of simplicity, we considered round objects. Suppose that the “red” and “green” objects have size distribution $r_i$ and $R_j$, respectively. The overlap of two objects is above the threshold $T$ if the distance between their centres satisfies to constrain $x < x_0$, where $x_0$ (Fig.2C) is root of the system:

\[
\begin{align*}
x_0 &= r \cdot \cos\left(\frac{\theta}{2}\right) + R \cdot \cos\left(\frac{\varphi}{2}\right) \\
r \cdot \sin\left(\frac{\theta}{2}\right) &= R \cdot \sin\left(\frac{\varphi}{2}\right) \\
(\phi - \sin(\phi)) \cdot r^2 + (\varphi - \sin(\varphi)) \cdot R^2 &= T \cdot \pi \cdot r^2
\end{align*}
\]

If the image size is $S$ and objects are homogenously distributed, then the probability for two objects to overlap above the threshold $T$ is:

\[
P(x \leq x_0 | r_i, R_j) = \frac{\pi \cdot x_0^2(T, r_i, R_j)}{S}
\]

So, if the following two assumptions hold:

a) objects are placed independently
b) two objects of the same colour are non-overlapping

then the expectation of number of colocalized objects is:

\[
N_{col} = \sum_i n_i \cdot \left(1 - \prod_j \left[1 - \frac{\pi \cdot x_0^2(T, r_i, R_j)}{S_j}\right]^{k_j}\right)
\]

where $k_j$ - is the number of “green” objects with size $R_j$, $n_i$ - is the number “red” objects with size $r_i$, and $S_j = S - \sum_{m=0}^j \pi \cdot R_m^2$

The prediction of equation (3) was verified by comparison with result of stochastic simulation (Fig.3) and demonstrated its validity.

In reality, objects are not homogeneously distributed inside cells. To take in account this inhomogeneity formula (3) will be modified to
FIGURE 2. Object-based colocalization as a function of relative overlapping area. **A.** Images of 500nm multi-colour beads. **B.** Colocalization of image with 647nm excitation to the images with excitation by 405nm, 488 nm and 568 nm lasers are drawn by blue, green and red curves respectively. The colocalization was calculated as a ratio of integral intensity of colocalized objects to the total integral intensities of objects of given channel. **C.** Relative overlapped area T is ratio of brown area to the area of green circle.

\[ N_{col} = \sum_i n_i \left( 1 - \prod_j \left[ 1 - \int_0^{\phi_j(T,R_i,R_j)} 2\pi x \cdot P(x) \,dx \right] \right)^{k_j} \]  

(4)

where \( P(x) \) - is distribution of object-to-object distances.

In the case of homogeneous distribution \( P(x) = \frac{1}{S_j} \) and (4) coincides with (3).

**CORRECTION FOR RANDOM COLOCALIZATION**

The behaviour of experimentally measured and estimated random colocalizations of two endosomal markers, APPL1 to EEA1 [9], are presented in Fig.4. In case of overlap threshold 0.8 the random colocalization is about 10% of measured one and direct subtraction is a good way of correction:

\[ c = m - r \]  

(5)

where \( c \) is the corrected colocalization, \( m \) is the measured apparent colocalization and \( r \) is the estimated random colocalization.

However, at overlap threshold equal 0.35 chosen on the basis of beads calibration...
We have tested the formula (4) by randomly placing in the image 1412 “red” and 1973 “green” round objects with log-normal size distribution to mimic the size distribution of endosomes (typical intracellular organelles). Red circles present the result of the simulation; the black line presents the model prediction.

(Fig.2B, red curve), the contribution of random colocalization becomes significant (40% of measured one). In such case, the common practice [4-7] is to use the correction formula

\[ c = \frac{m - r}{1 - r} \]  

which came from set consideration (Venn diagram).

Unfortunately, in real measurements there are cases where the estimated random colocalization is larger than the experimentally measured apparent colocalization. Then formula (6) leads to the absurd result of negative colocalization. In order to handle the cases where the random colocalization is close to the apparent colocalization, we applied the following probabilistic approach.

Suppose we measured colocalization \( m, \sigma_m \) and made estimation of random colocalization \( x, \sigma_x \) value, SEM. Then the apparent colocalization \( \mu \) is \( \mu = c + \xi - c \cdot \xi \), where \( c \) is the real colocalization, \( \xi \) is the random colocalization and we assume that real colocalization and random colocalization are independent.

\[
p(c|x, m) = \frac{1}{Z} \int \int p(\mu|m) \cdot p(\xi|x) \cdot \delta(c + \xi - c \cdot \xi - \mu) d\mu d\xi \tag{7}
\]

Suppose that the error of colocalization measurement and random colocalization estimation are normally distributed, then

\[
p(\mu|m) = \frac{1}{\sqrt{2\pi}\sigma_m} e^{-\frac{1}{2} \frac{(\mu - m)^2}{\sigma_m^2}} \quad \text{and} \quad p(\xi|x) = \frac{1}{\sqrt{2\pi}\sigma_x} e^{-\frac{1}{2} \frac{(\xi - x)^2}{\sigma_x^2}} \tag{8}
\]
From (7) and (8) we derive:

\[
p(c|x, m) = \frac{1}{Z} \int e^{-\frac{1}{2} \frac{(\mu - m)^2}{\sigma_m^2}} e^{-\frac{1}{2} \frac{(\xi - x)^2}{\sigma_t^2}} \cdot \delta(c + \xi \cdot (1 - c) - \mu) d\mu d\xi = \frac{1}{Z} \int e^{-\frac{1}{2} \frac{(\xi - x)^2}{\sigma_t^2}} e^{-\frac{1}{2} \frac{(1-c)(\frac{m-c}{m} - \xi)^2}{\sigma_t^2}} d\xi = \frac{1}{Z} \int e^{-\frac{1}{2} \left( \frac{(x-\xi)^2}{\sigma_t^2} + \frac{(m-c-\xi)^2}{\sigma_t^2} \right)} d\xi
\]

(9)

After the combination of variables, this expression can be written as:

\[
p(c|x, m) = \frac{1}{Z} e^{-\frac{1}{2} \left( \frac{x^2}{\sigma_m^2} + \frac{y^2}{\sigma_t^2} - \frac{z^2}{\sigma_t^2} \right)} \left( \text{erf} \left( \frac{1-z}{\sqrt{2}} \right) + \text{erf} \left( \frac{z}{\sqrt{2}} \right) \right)
\]

(10)

where \( \sigma_m = \frac{\sigma_m}{1-c} \), \( s^2 = \frac{\sigma_m^2 \sigma_t^2}{\sigma_m^2 + \sigma_t^2} \), \( y = m-c \), \( z = \frac{x^2 \sigma_t^2 + y^2 \sigma_t^2}{\sigma_m^2 + \sigma_t^2} \).

Then the expectation of colocalization can be found as

\[
\langle c \rangle = \int_0^1 c \cdot p(c|x, m) dc \quad \text{with variance} \quad \sigma_c^2 = \int_0^1 c^2 \cdot p(c|x, m) dc - \langle c \rangle^2
\]

(11)

The comparison of naive subtraction of random colocalization (5) and correction by (6) and (11) are presented in Fig.5.

**RELATIVE (MULTI-COLOUR) COLOCALIZATION**

Suppose we want to find the proportion of a cargo molecule, e.g. EGF that is colocalized with another cargo, LDL, on EEA1-positive endosomes. This is the ratio of integral intensities of triple-overlapped objects (EGF overlaps with EEA1 and LDL) to the
FIGURE 5. We calculated the correction for random colocalization by 3 formulas: 1) naive subtraction by formula (5) (black curve); 2) subtraction by formula (6), which takes into account that some really colocalized objects can be randomly colocalized as well (blue curve); 3) subtraction by formula (11) (red curve). Panel A presents an example, where the measured colocalization is $m = 0.6 \pm 0.05$ and the estimation of random colocalization varies from 0.0 to 0.75 ($r \in [0.0, 0.75]$) with $\sigma_r = \frac{r}{10}$. Panel B presents an example, where measured colocalization is $m = 0.2 \pm 0.01$ and random colocalization varies from 0.0 to 0.3 ($r \in [0.0, 0.3]$) with $\sigma_r = \frac{r}{20}$.

Integral intensities of double overlapped objects (EGF overlaps with EEA1). If we neglect the uncertainty of colocalizations, then relative (multi-color) colocalization $c_r$ is defined by formula $c_r = \frac{c_t}{c_d}$, where $c_t$ is a triple colocalization and $c_d$ is a double colocalization. The problem of the relative colocalization is that both the nominator and denominator are subjects of random colocalization. In other words, both of them are defined with some uncertainty. It is obvious that independent correction for random colocalization of nominator and denominator could result in the value above one, which is nonsense.

To handle this problem, we generalized the probabilistic approach for the case of multi-color colocalization. For the sake of simplicity, we approximated the uncertainty of colocalization by the truncated normal distribution (as amount of EGF colocalized to LDL and EEA1 cannot be more than the amount of EGF colocalized to EEA1 only):

$$p(c = \frac{x}{y} | x, y) = \frac{1}{Z} \int_0^1 e^{-\frac{1}{2} \left( \frac{(c - \eta)^2}{\sigma_t^2} + \frac{(c - \eta)^2}{\sigma_d^2} \right)} dy = \frac{1}{Z} \int_0^1 e^{-\frac{1}{2} \left( \frac{(c - \eta)^2}{\sigma_t^2} + \frac{(c - \eta)^2}{\sigma_d^2} \right)} dy$$

(12)

After transformation we get:

$$p(c = \frac{x}{y} | x, y) = \frac{1}{Z} e^{-\frac{1}{2} \left( \frac{c^2}{\sigma_t^2} + \frac{c^2}{\sigma_d^2} - \eta^2 \right)} \left( erf \left( \frac{1 - \eta}{\sqrt{2s}} \right) + erf \left( \frac{\eta}{\sqrt{2s}} \right) \right)$$

(13)
where $s^2 = \frac{\sigma^2 \sigma^2}{c^2 \sigma^2 + \sigma^2}$, $\eta = \frac{c \sigma^2 + \sigma^2}{c^2 \sigma^2 + \sigma^2}$.

Then the expectation and variance of colocalization are:

\[
\langle c \rangle = \int_0^1 c \cdot p(c|x,y) \, dc \quad \text{with variance} \quad \sigma^2_c = \int_0^1 c^2 \cdot p(c|x,y) \, dc - \langle c \rangle^2 \quad (14)
\]

**SUMMARY**

We developed a probabilistic model for the estimation of double and multiple colocalization of intracellular markers on vesicular structures. The estimated value is the proportion of the marker localized on specific vesicles and, as such, can be directly used as parameter in models of biological processes. The proposed model has two main advantages: 1) it never gives a negative colocalization value and 2) provides the estimation of colocalization uncertainty.

**ACKNOWLEDGMENTS**

We acknowledge K. Diamantara and H.A.Morales Navarrete for discussions and comments on the manuscript and Yury Bodrov for help in manuscript preparation.

**REFERENCES**

5. J.C. Rink, E. Ghigo, Y.L. Kalaidzidis, M. Zerial, *Rab Conversion as a Mechanism of Progression from Early to Late Endosomes*, Cell, 122, 735-749 (2005)