Probabilistic model of random co-colocalization in a cell

IGH - Floric Slimani

March 12, 2026

1 Aim

For biological studies purpose we image a cell in 3D using fluorescent microscopy and we are interested in quantifying the interactions between different types of single molecules. From those images we have no means of quantifying directly the interaction between single molecules, instead we want to see how frequently they localize together in space. Our goal is to build a general model that will modelise the likelyness that a single molecule from distribution i localize together or **co-localize** with a single molecule from distribution j in a cell of volume V **assuming single molecules take random postions within the cells independently of one another**.

Our images are produced through a sequential fish microscope in fixed cells that divides, according to its resolution cells in a set of v voxels localized with their set of coordinates *(z,y,x)* in the volume V.

2 Co-localization events

2.1 Self co-localization event

We note *C(i,i)* a **self co-localization** event between two single molecules of distribution i occupy the same voxel.

2.2 Co-localization event

We note *C(i,j)* a **co-localization** event between two single molecules of distribution i and *(*j) when a single molecule from distribution i occupy the same voxel as a single molecule from distribution j.

3 Probabilistic model

3.1 Presentation

The process of assigning a position to single molecules amongst the v possible positions can be modeled as a probabilistic game where positions are uniquely numbered balls placed in an urn. Assigning coordinates to a single molecule is drawing a ball in the urn, before each draw balls are replaced in the urn.

3.2 Unique distribution

To begin, let us consider a system containing a unique distribution I of single molecules randomly placed amongst the v positions of the cell.

3.2.1 Voxel occupancy by a distribution

First let us consider how single molecules from an unique distribution occupy the voxels forming volume V. We call a voxel from this distribution m, the probability that this voxel m is drawn at any draw ($i \in I$) is:

$$p_i(m) = \frac{1}{v} \qquad \forall i \in I; \forall m \in V$$

Then the probability to not draw the specific location m is:

$$p_i(\overline{m}) = 1 - \frac{1}{v} \qquad \forall i \in I; \forall m \in V$$

After k draws, the probability to never draw a specific location m follows a bionomial law and is :

$$p_k(X_m = 0) = (1 - \frac{1}{v})^k \qquad k \in \mathbb{N}^*$$

On the contrary the probability that the location m was drawn at least once is :

$$p_k(X_m \geq 1) = 1 - (1 - \frac{1}{v})^k \tag{1}$$

We define the observation variable ϵ to study occupancy of positions $m \in V$:

$$\forall m \in V, \epsilon(m) = \begin{cases} 0 \text{ if } X_m = 0 \\ 1 \text{ if } X_m \geq 1 \end{cases}$$

The expected number of **different** positions $\mu_{uniquepos}$ drawn is then the expectancy of ϵ in the volume after k draws.

$$\mu_{uniquepos} = \sum_{m \in V} p_k(X_m = 0).0 + p_k(X_m \geq 1).1$$

$$\mu_{up} = v(1 - (1 - \frac{1}{v})^k) \tag{2}$$

Where μ_{up} is the mean number of occupied position. The expected number of different positions drawn is, in other words, the expected number of draws **that discovered a new position**.

3.2.2 Self co-localization expectancy

From the voxel occupancy expectancy defined as the number of draws that discovered a new position, we can find the number of duplicates draw by removing the number of draws that discovered a new position to the total number of draws.

$$\mu_{duplicatepos} = k - v(1 - (1 - \frac{1}{v})^{k})$$

However, to properly count the number of self co-localization we have to take into account the multiplicity of single molecules in voxel when making a duplicate draw at this position. For example if a position m is drawn $Xm = 2$ times, the number of duplicate draw is 1 while the number of molecules that self-colocalize is at m is 2, that is to say the first duplicate draw at a position will create 2 self co-localization events but any subsequent draw will add only one. To adress this we create a new observation variable $s_m(X_m)$:

$$s_m(X_m) = \begin{cases} 0 \text{ if } X_m < 2 \\ X_m \text{ if } X_m \geq 2 \end{cases}$$

The expectancy of self colocalizing singe molecules at m is:

$$\mu_{selfcolocalization} = E(s_m(X_m)) = \sum_{X_m=0}^{k} s(X_m)p_m(X_m)$$

$$\Rightarrow \mu_{sc} = 0 + 0 + \sum_{X_m=2}^{k} s(X_m)p_m(X_m)$$

$$=$$

3.2.3 Self dependency of any distribution within the volume

Although all distributions i_i are independent from each other, X_m and X_l are **not independent** for any pair of voxels (m, l).ϵ follows a Bernoulli law, usually, summing this variable in the volume should lead to a variable following a bionomial law. However we argue in this section that there is a co-dependency between the voxels, meaning we can't consider the sum of $\epsilon(m)$ as abionomial distribution. To understand this let us consider the probability that a single molecule i is found at voxel m and at voxel $l \neq m$. Starting with the inclusion-exclusion principle:

$$P(X_m \geq 1 \cap X_l \geq 1) = P(X_m \geq 1) + P(X_l \geq 1) - P(X_m \geq 1 \cup X_l \geq 1)$$

$$P(X_m \geq 1 \cap X_l \geq 1) = 2p(X \geq 1) - P(X_m \geq 1 \cup X_l \geq 1)$$

The probability that a single molecule localize at l or m is the chance to draw one of 2 voxels out of the volume V.

$$P(m \cup l) = \frac{2}{v}$$

Thus from:

$$P(X_m \geq 1 \cup X_l \geq 1) = 1 - (1 - \frac{2}{v})^k$$

We can already conclude X_m and X_l are not independent since their intersection is non null, it also hilights that the dependency is made through the volume of voxels. To conclude calculus using above and **(1)** :

$$P(X_m \geq 1 \cap X_l \geq 1) = 1 - 2(1 - \frac{1}{v})^k + (1 - \frac{2}{v})^k \tag{3}$$

Conclusion : All disitributions **are** independent from each other, but the occupancy of a voxel **is not** independent of the occupancy of other voxels. This result is important for the computation of variance for the occupancy of voxel, since it means we must introduce a covariance term when summing up variables over voxels.

3.2.4 Covariance between I_m and I_l

Covariance is defined as :

$$Cov(\epsilon(m); \epsilon(l)) = E(\epsilon(m) \cap \epsilon(l)) - E(\epsilon(m))E(\epsilon(l))$$

where $E(\epsilon(m) \cap \epsilon(l))$ is the expectancy that positions m **and** l were drawn at least once in k draws. Let us start with left term and using **(4)**:

$$E(\epsilon(m) \cap \epsilon(l)) = P(X_m \geq 1 \cap X_l \geq 1).1 = 1 - 2(1 - \frac{1}{v})^k + (1 - \frac{2}{v})^k$$

For right term we have $E(\epsilon(m)) = E(\epsilon(l)) = E(\epsilon)$:

$$E(\epsilon) = P(X \geq 1).1 = 1 - (1 - \frac{1}{v})^k$$

Finally :

$$Cov(\epsilon(m); \epsilon(l)) = 1 + (1 - \frac{2}{v})^k - 2(1 - \frac{1}{v})^k - [1 - (1 - \frac{1}{v})^k]^2$$

After simplification :

$$Cov(\epsilon(m); \epsilon(l)) = (1 - \frac{2}{v})^k - (1 - \frac{1}{v})^{2k} \tag{4}$$

Notations:

- Let $p_i = (1 - (1 - \frac{1}{v})^{k_i})$, the probability that voxel m is occupied with at least one molecule i.
- Let $q_i = (1 - 2(1 - \frac{1}{v})^{k_i} + (1 - \frac{2}{v})^{k_i})$, the probability that voxel m and $l \neq m$ are both occupied with at least one molecule i.
- Let $c_i = q_i - p_i^2$ which can be shown equal to $Cov(\epsilon(m); \epsilon(l))$ the covariance between the occupancy of two different voxels by the same distribution i.

The previously demonstrated results can be re-written as:

$$\mu_{selfcolocalization} = k - vp$$

3.2.5 Standard deviation

To fully determine the probability density of voxel occupancy, let us find its standard deviation. We are computing the variance from the sum of dependent variables thus we have to use the law of total variance in its covariant form:

$$Var(N_{up}) = \sum_{m \in \mathbb{J}} Var(\epsilon(m)) + 2 \sum_{m<l} Cov(\epsilon(m); \epsilon(l))$$

Where $\sum_{m<l}$ means we sum on all different **unordered** pairs for $m \neq l \in \mathbb{J}^2$. I_m is a Bernoulli event of variance $p(1-p)$:

$$Var(\epsilon_m) = (1 - (1 - \frac{1}{v})^k)(1 - \frac{1}{v})^k$$

The number of unordered (m, l) pairs for $l \neq m$ is the number of pairs of voxel we can fit in the v voxels:

$$2 \sum_{m<l} Cov(I_m; I_l) = 2 \binom{v}{2} Cov(I_m; I_l) = 2 \frac{v(v-1)}{2} Cov$$

So total variance is:

$$\sigma_{up}^2 = v[1 - (1 - \frac{1}{v})^k](1 - \frac{1}{v})^k + v(v-1)[(1 - \frac{2}{v})^k - (1 - \frac{1}{v})^{2k}]$$

And standard deviation:

$$\sigma_{up} = \sqrt{v[1 - (1 - \frac{1}{v})^k](1 - \frac{1}{v})^k + v(v-1)[(1 - \frac{2}{v})^k - (1 - \frac{1}{v})^{2k}]}$$

which simplyfies to:

$$\sigma_{up} = \sqrt{vp_i(1 - p_i) + v(v-1)c_i} \tag{5}$$

Notes: σ_{up} is also the standard deviation for self-colocalization since the number of self co-localization is the also the number of draws that didn't discover a new location: $\mu_{sc} = k_i - v * p_i$ so variance arises only from p_i term.

3.3 Two distributions

We now consider positions assignement of a second distribution I of abundancy k_i after the assignement of a first distribution J of abundancy k_j with $I \neq J$. The number of unique positions J occupy can be estimated with **(2)**.

$$N^j_{uniquepos} = v(1 - (1 - \frac{1}{v})^{k_j})$$

A colocalisation event is drawing during assignement of I a position that was already drawn when assigning positions to J. To understand this in our probabilistic game, let us consider again a pool of v uniquely numbered balls where all balls that were drawn while assigning postions to J distribution have been colored in **red**. Again all balls are replaced in the pool after each draw. Then the co-localization probability $p(C(i,j))$ is the probability to draw a red ball.

3.3.1 Probability of co-localization and co-localization expectancy

Drawing ball in the urn is a sequence of independent and uniformly random events. The probability of picking one of the $N^j_{uniquepos}$ red ball amongst the v balls is :

$$p(X \in \mathbb{J}) = \frac{N^j_{up}}{v}$$

where $\mathbb{J}$ is the set of positions drawn for the J distribution. If we consider distribution J as fixed, co-localization events exactly follow a bionomial of parameters $(p(X \in \mathbb{J}), k_i)$. However, to consider a more generalistic model, let us consider a distribution distribution J of know abundancy k_j with mean number of unique positions occupied following a normal law $N(\mu_{sc}, \sigma_{sc})$. In such a case we use the law of total expectation :

$$\mu_{Cij} = E(E(C(i,j)|J))$$

$E(C(i,j)|J)$ is the expectancy from the bionomial law mentioned above.

$$\rightarrow E(k_i \frac{N^j_{up}}{v}) = \frac{k_i}{v} E(N^j_{up})$$

$$\mu_{Cij} = \frac{k_i}{v}(1 - (1 - \frac{1}{v})^{k_j}) \tag{6}$$

Notes:

- $\mu_{Cij} \neq \mu_{Cji}$

3.3.2 Co-localization standard deviation

Similarly to previous paragraph, let us find the standard deviation for number of co-localization events when considering J not fixed. To do so the calculus of

variance should take into consideration the variance of $N^j_{uniquepos}$ and use the law of total variance, which states :

$$Var(C(i,j)) = E[Var(C|J)] + Var(E[C|J])$$

$C|J$ is the case where distribution J is fixed and co-localization events follow bionomial law (§3.3.1) :

$$Var(C|J) = np(1-p) = k_i \frac{N^j_{uniquepos}}{v}(1 - \frac{N^j_{uniquepos}}{v})$$

$$E(C|J) = k_i \frac{N^j_{uniquepos}}{v}$$

Now let us consider the first term with notation $N_j = N^j_{uniquepos}$

$$\begin{aligned} E[Var(C|J)] &= E(k_i \frac{N_j}{v}(1 - \frac{N_j}{v})) \\ &= \frac{k_i}{v} E(N_j - \frac{N_j^2}{v}) \\ &= \frac{k_i}{v}[E(N_j) - \frac{E(N_j^2)}{v}] \\ &= \frac{k_i}{v}(\mu^j_{up} - \frac{\sigma^2_{up} + \mu^2_{up}}{v}) \end{aligned}$$

Using $Var(X) = E(X^2) - E^2(x)$.
The other term is straightforward :

$$Var(E(C|J)) = k_i \frac{\sigma^{2_j}_{up}}{v}$$

Total variance is :

$$\sigma^2_{Cij} = \frac{k_i}{v}(\mu^j_{up} - \frac{\sigma^2_{up} + \mu^2_{up}}{v}) + k_i \frac{\sigma^{2_j}_{up}}{v}$$

With μ^j_{up} and σ^j_{up} defined in equations **(2)** and **(6)**.
In conclusion we can write the standard deviation of the number of co-localization events as:

$$\sigma_{Cij} = \sqrt{\frac{k_i}{v}(\mu^j_{up} - \frac{\sigma^2_{up} + \mu^2_{up}}{v}) + k_i \frac{\sigma^{2_j}_{up}}{v}} \tag{7}$$

Which simplifies as :

$$\sigma_{Cij} = \sqrt{k_i((1 - \frac{1}{v})(p_j(1-p_j) - c_j) + c_j)}$$

3.3.3 Voxel occupancy by a pair (i,j) - Expectancy

In section 3.1 we deduced the probability that a specific location m to be drawn at least once in k draws as :

$$p_k(X_m \geq 1) = 1 - (1 - \frac{1}{v})^k \qquad \textbf{(1)}$$

To know how many red balls have been drawn at least once (**i.e. number of unique pair (i,j)**) let us use again the observation variable ϵ.

$$N_{up}^{(i,j)} = \sum_{m \in \mathbb{J}} p(C_m(i,j) \geq 1).1 + 0. \ \textbf{...}$$

Which mean value can be found using the law of total expectation :

$$\mu_{up}^{(i,j)} = N_{up}^{j}(1 - (1 - \frac{1}{v})^{k_i}) \tag{8}$$

or for J not fixed:

$$\mu_{up}^{(i,j)} = \mu_{up}^{j}(1 - (1 - \frac{1}{v})^{k_i})$$

$$\mu_{up}^{(i,j)} = v(1 - (1 - \frac{1}{v})^{k_j})(1 - (1 - \frac{1}{v})^{k_i})$$

Notes:

- Though $\mu_{Cij} \neq \mu_{Cji}$, explicit writting in last equation shows that $\mu_{up}^{(i,j)} = \mu_{up}^{(j,i)}$.
- Last equation also seems to show it will be easy to generalise the count of unique position occupied by any combination of molecules c of dimension n as $N_{uniquepos}^{c} = v \prod_{i \in c} \mu_c$.

3.3.4 Voxel occupancy by a pair (i,j) - Standard deviation

Notes & notations:

- Let's recall $p_i = (1 - (1 - \frac{1}{v})^{k_i})$, the probability that voxel m is occupied with at least one molecule i.
- The probability that i and j are in voxel m can be written $P(X_m^i \cap X_m^j) = p_i p_j$ since I and J are independent.
- Recall aswell $q_i = (1 - 2(1 - \frac{1}{v})^{k_i} + (1 - \frac{2}{v})^{k_i})$, the probability that voxel m and $l \neq m$ are both occupied with at least one molecule i.
- The probability that i and j are in voxel m and l can be written $P((X_m^i \cap X_m^j) \cap (X_l^i \cap X_l^j)) = q_i q_j$ since I and J are independent.

For this calculus we have to consider again the variance of the voxel occupancy of the J distribution in case it is not fixed. To do so we use again the total variance formula in its generalistic form:

$$Var(N_{up}^{(i,j)}) = E[Var(N_{up}^{(i,j)}|J)] + Var(E[N_{up}^{(i,j)}|J])$$

To make this less dense, let us break this in a few terms.
The simplest term is obtained from **(9)**:

$$E(N_{up}^{(i,j)}|J) = N_{up}^{j} p_i$$

$$\Rightarrow Var(E(N_{up}^{(i,j)}|J)) = vp_i Var(N_{up}^{j})$$

For the $Var(N_{up}^{(i,j)}|J)$ term we will have to take into consideration covariance term that arise from voxel-voxel dependancy (§3.2.2).

$$Var(N_{up}^{(i,j)}|J) = \sum_{m \in V} Var(\epsilon_{ij}(m)) + 2 \sum_{m<l \in V} Cov(\epsilon_{ij}(m); \epsilon_{ij}(l))$$

However we can take advantage that in this case the J distribution is **fixed**, meaning that if we restrict to the voxels having at least one j particle (**i.e.** $\mathbb{J}$) the covariance arises only from the distribution I and is exactly the same that the one computed in §3.2.3 when substuing v with N_{up}^{j}.

$$\begin{aligned} 2 \sum_{m<l \in V} Cov(\epsilon_{ij}(m); \epsilon_{ij}(l)) &= 2 \sum_{m<l \in \mathbb{J}} Cov(\epsilon_{ij}(m); \epsilon_{ij}(l)) \\ &= N_{up}^{j}(N_{up}^{j} - 1)[(1 - \frac{2}{v})^k - (1 - \frac{1}{v})^{2k}] \end{aligned}$$

The last term we should look at before summing-up is $\sum_{m \in V} Var(\epsilon_{ij}(m))$ using again J fixed, we know $\epsilon_{i,j}(m) = 0 \ \ \forall m \notin \mathbb{J}$:

$$\begin{aligned} \sum_{m \in V} Var(\epsilon_{ij}(m)) &= \sum_{m \in \mathbb{J}} Var(\epsilon_{ij}(m)) \\ &= \sum_{m \in \mathbb{J}} p_i(1 - p_i) \\ &= N_{up}^{j} p_i(1 - p_i) \end{aligned}$$

Let us group everything together :

$$\begin{aligned} Var(N_{up}^{(i,j)}|J) &= \sum_{m \in V} Var(\epsilon_{ij}(m)) + 2 \sum_{m<l \in V} Cov(\epsilon_{ij}(m); \epsilon_{ij}(l)) \\ &= N_{up}^{j} p_i(1 - p_i) + N_{up}^{j}(N_{up}^{j} - 1)[(1 - \frac{2}{v})^{k_i} - (1 - \frac{1}{v})^{2k_i}] \end{aligned}$$

Applying expectancy to this term:

$$\begin{aligned} E(Var(N_{up}^{(i,j)}|J)) &= vp_j p_i(1 - p_i) + E((N_{up}^{2j} - N_{up}^{j})[(1 - \frac{2}{v})^{k_i} - (1 - \frac{1}{v})^{2k_i}]) \\ &= vp_j p_i(1 - p_i) + (vp_j(1 - p_j) + v^2 p_j^2 - vp_j)[c_i] \\ &= vp_j p_i(1 - p_i) + (vp_j(1 - p_j)c_i + v^2 p_j^2 c_i - vp_j c_i) \end{aligned}$$

Notes & notations

- Let $c_i = (1 - \frac{2}{v})^{k_i} - (1 - \frac{1}{v})^{2k_i}]$
- Note that $c_i + p_i^2 = q_i$
- From §3.2.4 we can write $Var(N_{up}^j) = vp_j(1 - p_j)v(v - 1)[q_j - p_j^2]$

Summing-up the total variance:

$$\begin{aligned} Var(N_{up}^{(i,j)}) = vp_jp_i(1 - p_i) + (vp_j(1 - p_j)c_i + v^2p_j^2c_i - vp_jc_i) \\ + vp_iv(v - 1)[q_j - p_j^2]) \end{aligned}$$

After simplification, we have the total variance:

$$Var(N_{up}^{(i,j)}) = vp_ip_j(1 - p_ip_j) + v(v - 1)[q_iq_j - (p_ipj)^2] \tag{9}$$

3.4 Generalization to n Distributions

Consider n independent molecule distributions labeled $1, 2, \ldots, n$, with abundances k_i in a volume discretized into v voxels.

3.4.1 Single-Distribution Quantities

For each distribution i, define

$$p_i \;=\; \Pr(X_{i,m} \geq 1) = 1-\big(1-\tfrac{1}{v}\big)^{k_i}, \quad q_i \;=\; \Pr(X_{i,m} \geq 1,\; X_{i,\ell} \geq 1) = 1-2\big(1-\tfrac{1}{v}\big)^{k_i}+\big(1-\tfrac{2}{v}\big)^{k_i},$$

for any two distinct voxels $m \neq \ell$.

3.4.2 Combination of Distributions

Let $c \subseteq \{1, \ldots, n\}$ be any nonempty subset of distributions. Define

$$p_c \;=\; \Pr(\forall\, i \in c:\; X_{i,m} \geq 1) = \prod_{i \in c} p_i, \quad q_c \;=\; \Pr(\forall\, i \in c:\; X_{i,m} \geq 1,\; X_{i,\ell} \geq 1) = \prod_{i \in c} q_i.$$

3.4.3 Random Variable

Define the count of voxels co-occupied by all distributions in c:

$$N_c \;=\; \sum_{m=1}^{v} I_m^{(c)}, \qquad I_m^{(c)} = \mathbf{1}\{\forall\, i \in c:\; X_{i,m} \geq 1\}.$$

3.4.4 Expectation

Each $I_m^{(c)}$ has marginal probability p_c, so

$$E\big[N_c\big] = \sum_{m=1}^{v} E[I_m^{(c)}] = v\, p_c.$$

Note: this is the expected number of combination without accounting for multiplicity of distributions (one distribution having more that 1 single at the same spot). It will only count the expected location number containing this combination (Co-occupancy).

3.4.5 Variance and Standard Deviation

Because the indicators $I_m^{(c)}$ are dependent across voxels, use

$$Var(N_c) = \sum_{m=1}^{v} Var(I_m^{(c)}) + 2 \sum_{1 \leq m < \ell \leq v} Cov\big(I_m^{(c)}, I_\ell^{(c)}\big).$$

Here,

$$Var(I_m^{(c)}) = p_c(1 - p_c), \quad Cov\big(I_m^{(c)}, I_\ell^{(c)}\big) = q_c - p_c^2.$$

Thus

$$Var(N_c) = v\, p_c(1 - p_c) + v(v-1)\big(q_c - p_c^2\big),$$

and the standard deviation is

$$\sigma(N_c) = \sqrt{Var(N_c)} = \sqrt{v\, p_c(1 - p_c) + v(v-1)\big(q_c - p_c^2\big)}\,.$$