FROM MOLECULAR DIFFUSION TO ORGAN FORMATION: TRACKING ENDOCYTOSIS ACROSS SCALES IN THE DEVELOPPING EMBRYO

ANDRÉ THÉO

Turing Centre for Living Systems & Aix-Marseille Université

Abstract. This report focuses on the study of the role and mechanisms of membrane trafficking, or endocytosis, in the context of the embryo development. Indeed, many approaches have been proposed to understand how membrane can control the introduction of molecules inside the cell. This work, however, has been limited to cell cultured in hard surfaces such as a glass coverslip. Very little is known about those processes in an environment where tissue-like forces and geometry at the embryonic scales regulates (and are influenced by) the endocytic process at the nano-scale. In this report, we first provide a brief overview of the morphogenetic and endocytic process. We then describe how recent breakthroughs have been possible in the study of membrane trafficking through the combination of fluorescence microscopy and image processing algorithms that enable accurate detection of endocytic events, from nanometric membrane curvature to cargo release in the cytosol and later trafficking. We particularly focus on the problem linked to endocytic motion modeling, that requires the use of stochastic filtering approaches to efficiently estimate and predict particles position for thousands of concurrent processes. We then demonstrate with numerical experiments the challenges linked to repeating those experiments in the developing embryo and propose a few solutions. This study highlights an example of the more general problem posed by the understanding of molecular processes in their physiological environment. Our preliminary suggest that the quantitative analysis of nanometric processes in the context of cellular and morphogenetic motions requires precisely calibrated imaging along with robust estimator for the spatial distribution and dynamics of molecular events across scales.

Contents

1. Understanding the role of molecular processes in physiological functions: a case-study involving morphogenesis and endocytosis 3
1.1. Morphogenesis-driven dynamics 3
1.2. Receptor-mediated endocytosis 3
1.3. A few words on long-term endocytosis 4
2. Measuring Endocytosis: a multiple particle tracking problem 4
2.1. Studying endocytosis with fluorescence microscopy 4
2.2. Overview of dynamic estimation for the study of endocytosis 6
2.3. A focus on motion modeling 6
2.4. Bayesian filtering 6
2.5. The Kalman Filter 7
2.6. Modeling motions, noise, and model approximations 7
2.7. Piecewise-stationary Motion Models 9
3. Tracking Endocytosis in the developing embryo 10
3.1. Description of the raw data 10
3.2. Preliminary results 11
3.3. Discussion and possible research axis 11
Appendix A. Fundamental results on probabilities and Markov Chains 14
Appendix B. Additional Visuals 16
References 16

1. Understanding the role of molecular processes in physiological functions: a case-study involving morphogenesis and endocytosis

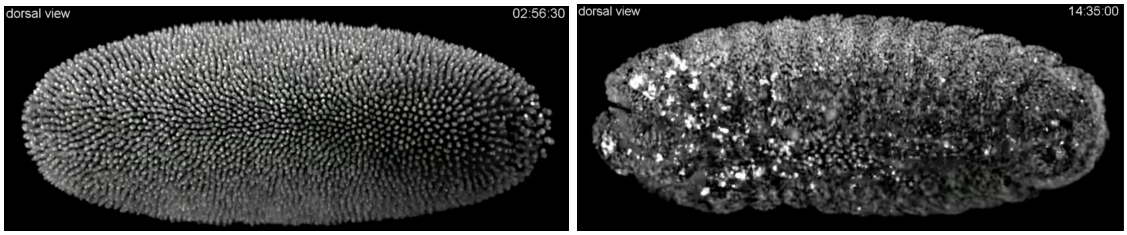


Figure 1. Two pictures of a drosophila embryo at different times t_1 (left) and t_2 (right) with $t_1 << t_2$. Credit to the full video: https://www.youtube.com/watch?v=h9RfeU5u85k

1.1. **Morphogenesis-driven dynamics.** Morphogenesis is the biological process in which an embryo sees its cells and tissues differentiate to give rise to a full-grown organism. All along the process, many cellular events occur and cells undergo several mechanical constraints, especially during gastrulation. Some of them will be involved in invagination, some will intercalate between each other, some will stretch, the list goes on... If one wishes to zoom close enough, one would observe contractions or extensions occurring at the boundaries of each cell. This movement is induced by the morphogenesis process and orchestrated thanks to Myosin II motor proteins (as an example) (see Figure 2). As such, the result of the morphogenetic process is not merely the product of genetic expression but also results from the influence of external forces and the geometry of the environment [1], Figure (3).

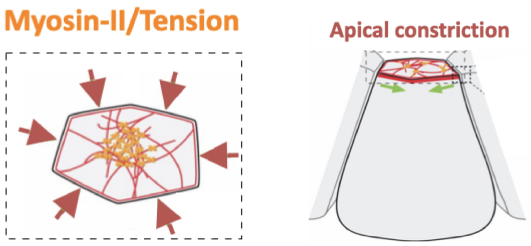


Figure 2. Apical contraction causing strong mechanical forces at the apex of a cell, due to Myosin II action: typical scenario for a cell undergoing invagination

1.2. **Receptor-mediated endocytosis.** The purpose of this section is to rapidly provide a sufficient background on the functioning of endocytosis for the rest of the report. There are several categories of endocytosis, each of them deserving a full biology course. In this report, we shall focus on clathrin-mediated endocytosis only. There is a range of proteins whose purpose is to behave as a receptor (we can mention E-Cadherin as an example). These proteins (naturally called receptor proteins) are expressed on the entire plasma membrane of the cell in small clusters in regions we call endocytic pits or clathrin-coated pits. Such a mechanism allows the cell to target specific molecules that it wishes to uptake for its proper development. Like small magnets, when a molecule passes near a cell expressing the appropriate receptor on its surface, the molecule of interest will bind to the receptor with a certain probability and send a signal to the cell that will be aware of the new binding. As long as there are enough molecules attached (this is called a cargo), the process of endocytosis starts (maturation process). If, after a short period of time, the amount of molecules at the surface is not sufficient, the process of endocytosis aborts. Please note that we distinguish early-aborted and late-aborted endocytosis event based on the amount of time waited before the process cancels.

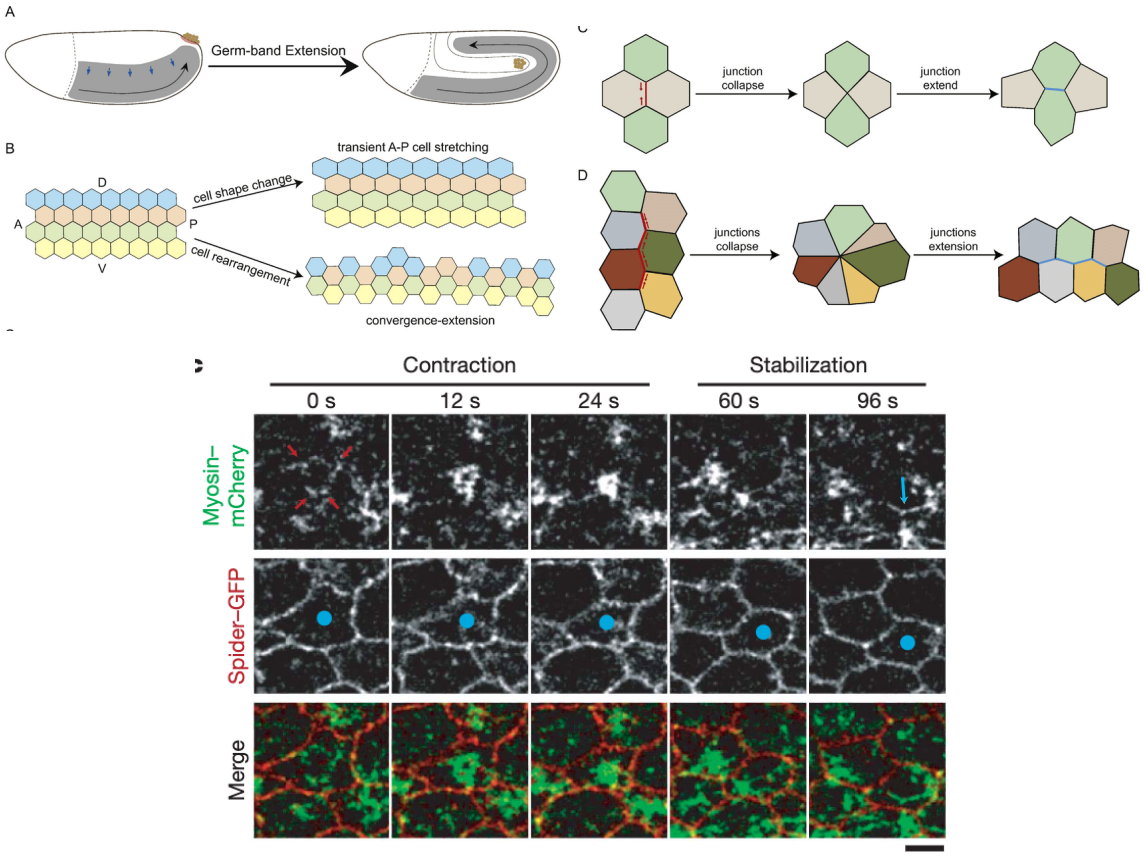


FIGURE 3. Morphogenetic activity induces topological constraints

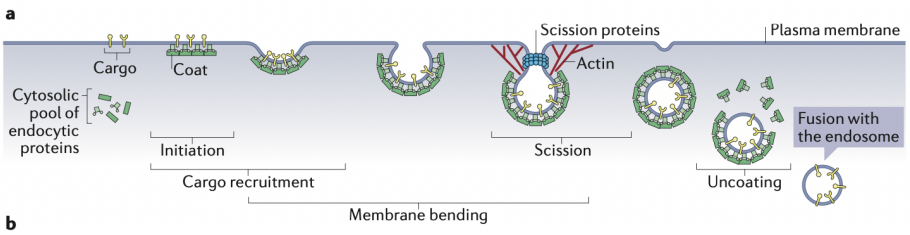


FIGURE 4. Depiction of the course of events for Clathrin-mediated endocytosis. This picture illustrates each step happening in the cell, from the call of the cytosolic pool to formation of a vesicle, provided the process of endocytosis maturation occurs.

1.3. **A few words on long-term endocytosis.** Say one wishes to study endocytic trafficking, the tracking of molecules such as Clathrin (4) is not sufficient because of the uncoating mechanism during the release of the cargo in the cytosol. Other proteins, such as Rab5 (see Figure (5)) could actually be better test-subject to understand, for instance, early-endosomes dynamics. This, even though interesting to mention, is not our topic of study, so let us conclude this small deviation and move on to the next part.

2. Measuring Endocytosis: a multiple particle tracking problem

2.1. **Studying endocytosis with fluorescence microscopy.** There are several ways to tackle down the endocytosis tracking problem: one particular choice is whether to work in a static or dynamic context... Both have been studied (see [2], [3] for static, and [1], [4] for dynamic studies) and both have their own pros and cons. Static study of endocytosis is, by all means, the most comfortable to work with. First, the image is static and therefore easier to capture than a moving subject. More channels are available: up to

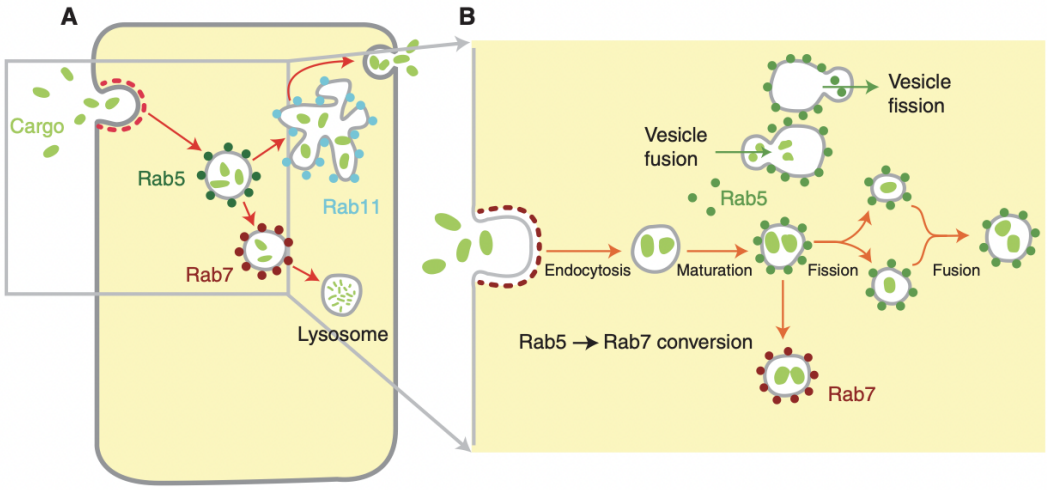


FIGURE 5. Detailed illustration endocytic pathway, extracted from

four channels in the static context against two in a dynamic one, allowing a more in-depth study of relations between molecules tagged with different fluorophores (a channel is roughly an array of pixels *e.g.,* Red, Green, Blue channels in a standard image). Some fluorophores may have fixation issues due to the dynamics of the embryo, making them exclusively usable on coverslips. When imaging dynamical movements there are some constraints on the photon budget whereas there is none in static. With all these facts in mind, why do we even care about dynamical context? The answer is that isolated cells in culture does not provide relevant information about real *in vivo* dynamics. More accurately, the dynamic study, besides all downsides, is the only way to capture the temporal hierarchy of events, as well as the role of changes in the geometry and the different forces that are exerted on each cell.

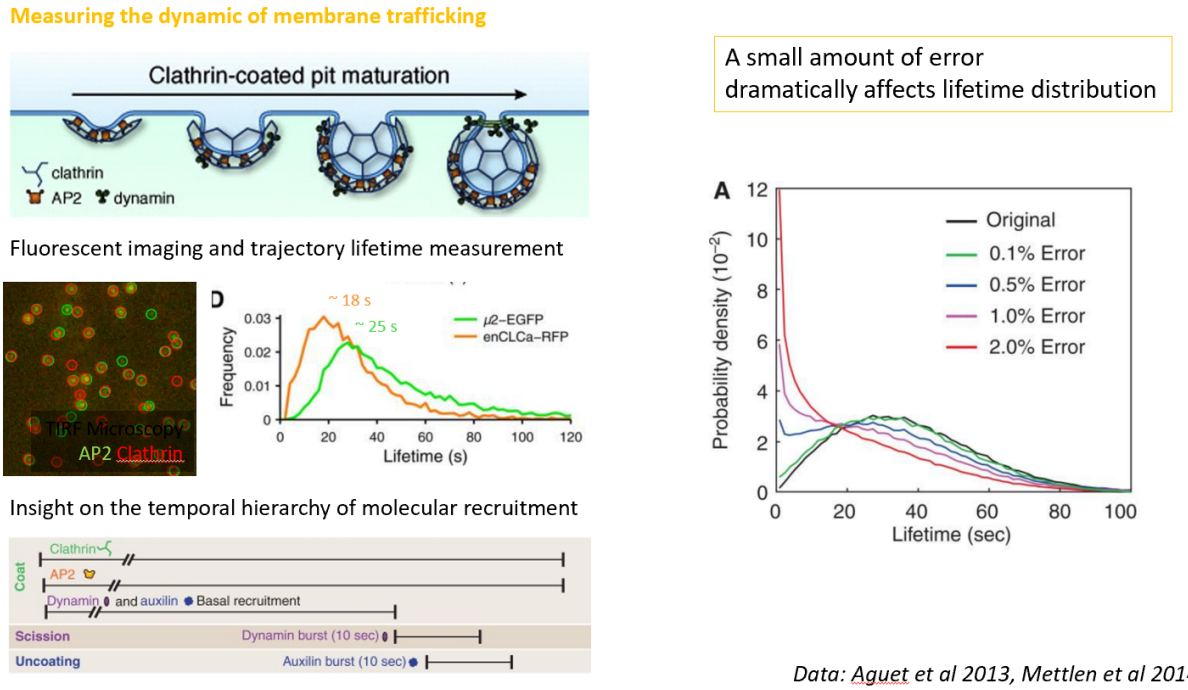


FIGURE 6. Results on dynamical lifetimes study from [2]

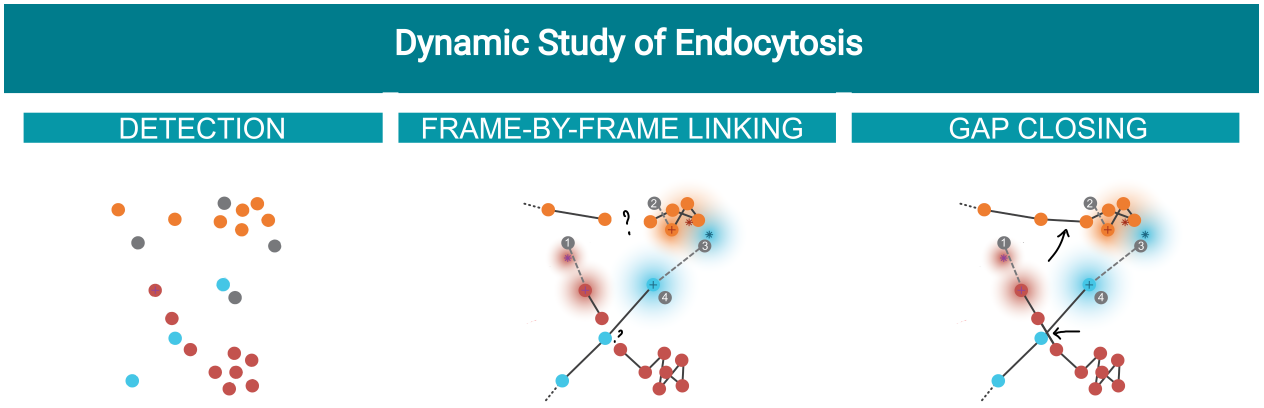


FIGURE 7. Major steps of image processing for endocytosis tracking

Independently of the context, some quantities are of higher importance than others. In particular, we are interested in two specific values: Lifetimes and trajectories. The main reason of our interest in lifetimes is that it would provide us with values that could be used to determine the kind of endocytic event occurring (early-aborted, late-aborted, maturation), see [4]. Secondly, having a look on the trajectories Would allow us to obtain information on the global movement of the embryo and used later to study internal cell movements by subtracting the flow to the movie.

2.2. **Overview of dynamic estimation for the study of endocytosis.** The goal of this section is to explain the generic workflow associated to this kind of problem: detection, linking, gap-closing and lifetime analysis. Each step can be summarized visually in the diagram presented in Figure (7).

2.3. **A focus on motion modeling.** A key challenge in the embryo is motion prediction, since live motions are much more complicated than in a cell stuck on a coverslip, we focus on that part of the approach. This would then require us to estimate an optimal set of particle trajectory, requiring the evaluation of all possible paths across detected objects. Small downside: this is an NP-hard problem. To tackle the combinatorial complexity, a temporally recursive approach is proposed based on stochastic filtering. The primary aim of stochastic filtering is to provide a temporally recursive approach to the estimation of parameters of interest (*i.e.,* new estimation per unit of time, or frame). By inferring dynamic parameters such as diffusion uncertainty or velocity, a stochastic filtering approach allows the estimation of likely locations for a particle from one image to the next. This method greatly improved tracking inaccuracy, while reducing the number of possible trajectories to evaluate. In this report, we first describe the use of Bayesian Filtering to perform this prediction, then the special case of Kalman Filtering under reasonable assumption for measurement noise and motion priors for efficient and scalable estimation, and finally an extension to multiple dynamics model that have been used to estimate molecular dynamics.

2.4. **Bayesian filtering.** Well-known results from probability theory allow us to recursively compute the distribution $p(\mathbf{x}_t|\mathbf{z}_{1:t})$ given an initial distribution $p(\mathbf{x}_0)$. We encapsulate this statement under the following property. (If a some point, there is a manipulation that is not clear, please refer to appendix A)

Proposition 1. *(Filtering Distribution) Let $\boldsymbol{\mathcal{X}}$ denote the set of all possible positions $\boldsymbol{x}$ can take. Given a set of measurements $\boldsymbol{z}_{1:t}$, and an initial distribution $p(\boldsymbol{x}_0)$, the filtering distribution of f satisfies*

$$p(\boldsymbol{x}_t|\boldsymbol{z}_t) = \frac{p(\boldsymbol{z}_t|\boldsymbol{x}_t)}{p(\boldsymbol{z}_k|\boldsymbol{z}_{1:k-1})} \int_{\boldsymbol{\mathcal{X}}} p(\boldsymbol{x}_k|\boldsymbol{x}_{k-1}) p(\boldsymbol{x}_{k-1}|\boldsymbol{z}_{1:t-1}) \, d\boldsymbol{x}_{k-1}$$

Proof. Let us start from the density and make the prior distribution $p(\mathbf{x}_{t-1}|\mathbf{z}_{1:t-1})$ appear. By Bayes rule, it holds

$$p(\mathbf{x}_t|\mathbf{z}_{1:t}) = \frac{p(\mathbf{x}_t, \mathbf{z}_{1:t})p(\mathbf{x}_t|\mathbf{z}_{1:t-1})}{p(\mathbf{z}_{1:t})} = \frac{p(\mathbf{z}_t|\mathbf{z}_{1:t-1}, \mathbf{x}_t)p(\mathbf{x}_t, \mathbf{z}_{1:t-1})}{p(\mathbf{z}_t|\mathbf{z}_{1:t-1})p(\mathbf{z}_{1:t-1})} = \frac{p(\mathbf{z}_t|\mathbf{z}_{1:t-1}, \mathbf{x}_t)p(\mathbf{x}_t|\mathbf{z}_{1:t-1})}{p(\mathbf{z}_t|\mathbf{z}_{1:t-1})}$$

We are close to the actual result, remark that, by Markov property, if $\mathbf{x}_t$ is known, it must contain all the knowledge from the past measurements: $p(\mathbf{z}_t|\mathbf{x}_t, \mathbf{z}_{1:t-1}) = p(\mathbf{z}_t|\mathbf{x}_t)$. it is then only left to marginalize the density with respect to the previous position, $\mathbf{x}_{t-1}$ which yields the wanted result

$$\frac{p(\mathbf{z}_t|\mathbf{z}_{1:t-1}, \mathbf{x}_t)}{p(\mathbf{z}_t|\mathbf{z}_{1:t-1})}p(\mathbf{x}_t|\mathbf{z}_{1:t-1}) = \frac{p(\mathbf{z}_t|\mathbf{x}_t)}{p(\mathbf{z}_t|\mathbf{z}_{1:t-1})}\int_{\boldsymbol{\mathcal{X}}} p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathbf{z}_{1:t-1})\mathrm{d}\mathbf{x}_{t-1}$$

QED

Remark: One can notice that, using marginalization again, the denominator can be replaced by the quantity $\int_{\boldsymbol{\mathcal{X}}} p(\mathbf{z}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{z}_{1:t-1})d\mathbf{x}_t$ which is nothing but a constant when $\mathbf{x}_t$ is fixed. For this reason, the denominator is often omitted in the literature and the posterior distribution $p(\mathbf{x}_t|\mathbf{z}_t)$ is said to be 'proportional to the prior times the likelihood' and is written as

$$p(\mathbf{x}_t|\mathbf{z}_t) \propto p(\mathbf{z}_t|\mathbf{x}_t)\int_{\boldsymbol{\mathcal{X}}} p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}, \mathbf{z}_{1:t-1})\mathrm{d}\mathbf{x}_{t-1}$$

2.5. **The Kalman Filter.** Introduced in the 60s by **Rudolph Kalman** and first used to estimate the trajectory of the rocket in the context of the Apollo program [5], the Kalman filter [6] is inscribed in the family of Bayesian Filters. Considered as a standard approach, essentially in domains including sensor detection such as automated cars, planes detection and road radars, the Kalman filter could be identified to a stronger linear regression technique. Given an initial distribution, the filter recursively provides 'the best' parameters estimation for our model using two steps: prediction, then update. Its efficiency is the reason we chose to work with it in the first place.

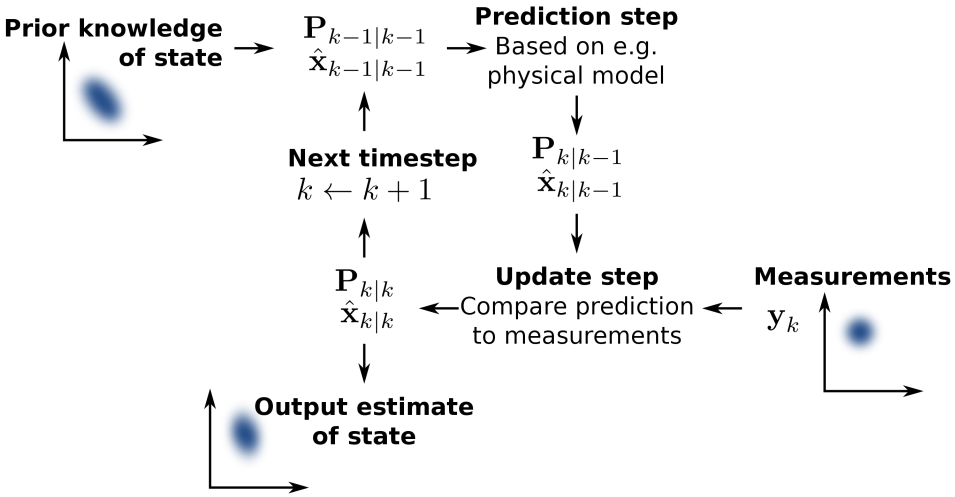


FIGURE 8. Graphical representation of the Kalman filter. The way to read the diagram is simply by following the arrows. This view is somewhat analogous to 'pseudocode' when one wishes to describe clearly the steps of an algorithm. $\hat{\mathbf{x}}_{m|n}$ is to be read 'estimation of $\mathbf{x}$ at time m, knowing measurements up to time n, the same goes for $\boldsymbol{P}_{m|n}$.

2.6. **Modeling motions, noise, and model approximations.** In what follows, all random quantities are defined on a probability space $(\Omega, \mathcal{T}, p)$. Let us start by introducing the random vectors $\mathbf{x}_t \in \mathbb{R}^6$ and $\mathbf{z}_t \in \mathbb{R}^6$ representing an arbitrary particle and its measurement at time t. Each particle is characterized by its position and velocity in space, in other words.

$$\mathbf{x}_t = (x_t, y_t, z_t, dx_t, dy_t, dz_t)$$

Additionally, introduce the two sets $\mathbf{z}_{1:t} = \{\mathbf{z}_1, \dots, \mathbf{z}_t\}$ and $\mathbf{x}_{1:t} = \{\mathbf{x}_1, \dots, \mathbf{x}_t\}$ so that we can refer to all measures and positions more conveniently. Now that all is set up, we introduce the following motion-measurement model

$$\begin{cases} \mathbf{x}_t = \boldsymbol{f}_{t-1}(\mathbf{x}_{1:t-1}) + \mathbf{w}_{t-1} \\ \mathbf{z}_t = \boldsymbol{g}_t(\mathbf{x}_{1:t}) + \mathbf{v}_t \end{cases} \quad \text{or equivalently} \quad \begin{cases} p(\mathbf{x}_t|\mathbf{x}_{1:t-1}) \\ p(\mathbf{z}_t|\mathbf{x}_{1:t}) \end{cases}$$

Where $\mathbf{w}_{t-1}, \mathbf{v}_t \in \mathbb{R}^6$ denotes exterior, zero-mean, random noise, $p(\cdot|\cdot)$ denotes the conditional density. Introduce the two families of functions $\boldsymbol{f}_{t-1} : \bigotimes_{1 \leqslant T \leqslant t-1} \mathbb{R}^6 \longrightarrow \mathbb{R}^6$ and $\boldsymbol{g}_t : \bigotimes_{1 \leqslant T \leqslant t} \mathbb{R}^6 \longrightarrow \mathbb{R}^6$ associating a new position to $\mathbf{x}_t$ given that the set of previous positions. This model being obviously too general and too complex to deploy in a practical context, we introduce further assumptions to make it less daunting. First, let us assume that $\mathbf{x}_t$ and $\mathbf{z}_t$ are hidden states of a first order hidden Markov model. This way we can get rid, of all previous information beyond order 1, *i.e.,* $p(\mathbf{x}_t|\mathbf{x}_{1:t-1}) = p(\mathbf{x}_t|\mathbf{x}_{t-1})$ and $p(\mathbf{z}_t|\mathbf{x}_{1:t}) = p(\mathbf{z}_t|\mathbf{x}_{t-1})$. Furthermore, non-linear models often give rise to unpredictable behaviors, which is why we will assume linear relations on both variables, in a way that we can find two matrices $\boldsymbol{F}, \boldsymbol{H}$ conveniently shaped (they are elements of $\boldsymbol{\mathcal{M}}_6(\mathbb{R})$ and $\boldsymbol{\mathcal{M}}_{1,6}(\mathbb{R})$ respectively), so that $\boldsymbol{f}_{t-1}(\mathbf{x}_{t-1}) = \boldsymbol{F}\mathbf{x}_{t-1}$ and $\boldsymbol{g}_t(\mathbf{x}_t) = \boldsymbol{H}\mathbf{x}_t$. Finally, we will consider that the model is Gaussian, meaning that the initial distribution of $\mathbf{x}, \mathbf{z}, \mathbf{w}$ and $\mathbf{v}$ are all normally distributed, in particular, we will assume that $\mathbf{w}_t, \mathbf{v}_t$ are independent white processes, that is $p(\mathbf{v}_t) = \boldsymbol{\mathcal{N}}(\mathbf{v}_t; 0, \boldsymbol{Q}_t)$ [1] and $p(\mathbf{w}_t) = \boldsymbol{\mathcal{N}}(\mathbf{w}_t; 0, \boldsymbol{R}_t)$. We have now reduced our model to the following one

$$\begin{cases} \mathbf{x}_t = \boldsymbol{F}\mathbf{x}_{t-1} + \mathbf{w}_{t-1} \\ \mathbf{z}_t = \boldsymbol{H}\mathbf{x}_t + \mathbf{v}_t \end{cases}$$

In the literature, the matrix $\boldsymbol{H}$ is often called **observation** matrix. Loosely, this matrix describes what characteristics of $\mathbf{x}$ can be observed thanks to our measurements. Since fluorescence Microscopy can realize three-dimensional pictures, it is possible to get estimates on the position of a given particle, but there is no way one can detect celerity of particles. Recall that $\mathbf{x}_t$ is defined by the vector whose first three components are associated to the position in space, and the three others velocity, one can easily deduce that a good choice for H is the following

$$\boldsymbol{H} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \end{bmatrix}$$

Even though simpler, this model allows some flexibility with the choice of $\boldsymbol{F}$ (the **transition** matrix). For instance, we can either choose to model the overall movement as a Brownian motion (*i.e.,* we do not take into account the velocity of the particle) or a linear movement with the choice

$$\boldsymbol{F}_{\textbf{Brownian}} = \left[\begin{array}{c|c} \boldsymbol{I}_3 & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array}\right] \quad \text{or} \quad \boldsymbol{F}_{\textbf{Linear}} = \left[\begin{array}{c|c} \boldsymbol{I}_3 & * \\ \hline \mathbf{0} & \boldsymbol{I}_3 \end{array}\right]$$

with $\boldsymbol{I}_3$ being the identity matrix of $\boldsymbol{\mathcal{M}_3}(\mathbb{R})$. By construction, this kind of models has a structure that allows one to compute the filtered distribution $p(\mathbf{x}_t|\mathbf{z}_{1:t})$ recursively.

[1] $\boldsymbol{\mathcal{N}}(x; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the multidimensional Gaussian probability distribution $x \longmapsto \mathbf{det}(2\pi\boldsymbol{\Sigma})^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x-\boldsymbol{\mu})^{\mathbf{T}}\Sigma^{-1}(x-\boldsymbol{\mu})\right)$

2.7. **Piecewise-stationary Motion Models.** When tracking multiple objects with heterogeneous behavior, a single Kalman Filter is not enough. A great alternative, proposed in paper [7] suggests we should use multiple Kalman Filters, each one labeled by an integer $\theta_t \in \{1, ..., N\} := \Lambda$, where θ_t has some distribution on a probability space $(\Upsilon, \mathcal{A}, p)$[2]. We can thus define the following (large) system of equations

$$\begin{cases} \mathbf{x}_t = \boldsymbol{F}^{\theta_t}\mathbf{x}_{t-1} + \mathbf{w}_{t-1}^{\theta_t} \\ \mathbf{z}_t = \boldsymbol{H}^{\theta_t}\mathbf{x}_t + \mathbf{v}_t^{\theta_t} \end{cases} \tag{1}$$

where each $\boldsymbol{F}^{\theta_t}, \boldsymbol{H}^{\theta_t}$ is associated to a different transition matrix to better describe the movement. Analogously to what has been done in proposition 1., let us try to derive a recursive formula and then apply filters to compute the distribution $p(\mathbf{x}_t|\mathbf{z}_{1:t})$.

Theorem 1. *(Recursive formula for the distribution) : Given process (1) satisfying all assumptions above, the filtering distribution is given by*

$$p(\boldsymbol{x}_t|\boldsymbol{z}_{1:t}) \sim \boldsymbol{\mathcal{N}}(\boldsymbol{x}_t; \hat{\boldsymbol{x}}_t, \hat{\boldsymbol{P}}_t) \quad \textit{where} \quad \begin{cases} \hat{\boldsymbol{x}}_t = \sum_{\theta_t \in \Lambda} p(\theta_t|\boldsymbol{z}_{1:t})\hat{\boldsymbol{x}}_t^j \\ \hat{\boldsymbol{P}}_t = \sum_{\theta_t \in \Lambda} p(\theta_t|\boldsymbol{z}_{1:t}) \left(\hat{\boldsymbol{P}}_t^j + \|\hat{\boldsymbol{x}}_t^j - \hat{\boldsymbol{x}}_t\|^2 \right) \end{cases}$$

and $\hat{\boldsymbol{x}}_t^j = \mathbb{E}[\boldsymbol{x}_t|\boldsymbol{z}_{1:t}, \theta_t]$ *and* $\hat{\boldsymbol{P}}_t^j = \mathbb{V}[\boldsymbol{x}_t|\boldsymbol{z}_{1:t}, \theta_t]$.

Proof. The proof goes in a very similar fashion as in the previous one. What is new here is that we have to take θ into account ! Using the fact that the set of $\{\theta_t = i : i \in \Lambda\} = \bigsqcup_{i \in \Lambda}\{\theta_t = i\}$ forms a finite partition of Υ we can marginalize our probability of θ_t (notice that Λ is countable, and thus we use the counting measure when marginalizing, transforming the integral in a sum)

$$p(\mathbf{x}_t|\mathbf{z}_{1:t}) = \sum_{\theta_t \in \Lambda} p(\mathbf{x}_t|\theta_t, \mathbf{z}_{1:t}) p(\theta_t|\mathbf{z}_{1:t})$$

Our goal is to make the prior distribution appear. In that extent, we can operate some surgery on $p(\mathbf{x}_t|\theta_t, \mathbf{z}_{1:t})$ using successive conditioning

$$p(\mathbf{x}_t|\theta_t, \mathbf{z}_{1:t}) = \frac{p(\mathbf{x}_t, \mathbf{z}_{1:t}, \theta_t)}{p(\mathbf{z}_{1:t}, \theta_t)} = \frac{p(\mathbf{z}_t|\mathbf{x}_t, \mathbf{z}_{1:t-1}, \theta_t) p(\mathbf{x}_t|\mathbf{z}_{1:t-1}, \theta_t) \cancel{p(\mathbf{z}_{1:t-1}, \theta_t)}}{p(\mathbf{z}_t|\mathbf{z}_{1:t-1}, \theta_t) \cancel{p(\mathbf{z}_{1:t-1}, \theta_t)}}$$

Once again, use Markov property to remove past information from the distribution

$$p(\mathbf{x}_t|\mathbf{z}_t) = \sum_{\theta_t \in \Lambda} p(\theta_t|\mathbf{z}_{1:t}) \frac{p(\mathbf{z}_t|\mathbf{x}_t, \theta_t)}{p(\mathbf{z}_t|\mathbf{z}_{1:t-1}, \theta_t)} p(\mathbf{x}_t|\mathbf{z}_{1:t-1}, \theta_t)$$

Expanding the last term of the summand using Chapman-Kolmogorov yields

$$p(\mathbf{x}_t|\mathbf{z}_{1:t-1}, \theta_t) = \int_{\boldsymbol{\mathcal{X}}} \mathrm{d}\mathbf{x}_{t-1} \Big(p(\mathbf{x}_t|\mathbf{x}_{t-1}, \theta_t) p(\mathbf{x}_{t-1}|\mathbf{z}_{1:t-1}, \theta_t) \Big)$$

which, once plugged back in the sum

$$p(\mathbf{x}_t|\mathbf{z}_t) = \sum_{\theta_t \in \Lambda} p(\theta_t|\mathbf{z}_{1:t}) \frac{p(\mathbf{z}_t|\mathbf{x}_t, \theta_t)}{p(\mathbf{z}_t|\mathbf{z}_{1:t-1}, \theta_t)} \int_{\boldsymbol{\mathcal{X}}} \mathrm{d}\mathbf{x}_{t-1} \Big(p(\mathbf{x}_t|\mathbf{x}_{t-1}, \theta_t) p(\mathbf{x}_{t-1}|\mathbf{z}_{1:t-1}, \theta_t) \Big)$$

Expressions starting to get long, let us give some names: $A(t, j) := p(\theta_t|\mathbf{z}_{1:t}) \frac{p(\mathbf{z}_t|\mathbf{x}_t, \theta_t)}{p(\mathbf{z}_t|\mathbf{z}_{1:t-1}, \theta_t)}$ so that the distribution is rewritten as

[2]It is not morally the same p, but this is way more convenient notation-wise

$$p(\mathbf{x}_t|\mathbf{z}_t) = \sum_{\theta_t \in \Lambda} A(t,j) \int_{\boldsymbol{\mathcal{X}}} \mathrm{d}\mathbf{x}_{t-1} \Big(p(\mathbf{x}_t|\mathbf{x}_{t-1}, \theta_t) p(\mathbf{x}_{t-1}|\mathbf{z}_{1:t-1}, \theta_t) \Big)$$

We now unfoil the second term of the integrand just like we did in the first step of our computations

$$p(\mathbf{x}_{t-1}|\mathbf{z}_{1:t-1}, \theta_t) = \sum_{i \in \Lambda} p(\mathbf{x}_{t-1}|\theta_{t-1}, \mathbf{z}_{1:t-1}) p(\theta_{t-1}|\theta_t, \mathbf{z}_{1:t-1})$$

Hence, the final result

$$p(\mathbf{x}_t|\mathbf{z}_t) = \sum_{\theta_t \in \Lambda} p(\theta_t|\mathbf{z}_{1:t}) \frac{p(\mathbf{z}_t|\mathbf{x}_t, \theta_t)}{p(\mathbf{z}_t|\mathbf{z}_{1:t-1}, \theta_t)} \int_{\boldsymbol{\mathcal{X}}} \mathrm{d}\mathbf{x}_{t-1} \Big(p(\mathbf{x}_t|\mathbf{x}_{t-1}, \theta_t) \sum_{i \in \Lambda} p(\mathbf{x}_{t-1}|\theta_{t-1}, \mathbf{z}_{1:t-1}) p(\theta_{t-1}|\theta_t, \mathbf{z}_{1:t-1}) \Big)$$

By doing so, we successfully made the term $p(\mathbf{x}_{t-1}|\mathbf{z}_{1:t-1}, \theta_{t-1})$ appear as wanted. However, notice that the formula is too complicated. At this point, the only deduction we are able to formulate is that $p(\mathbf{x}_t|\mathbf{z}_t)$ has a Gaussian distribution with $\mathrm{Card}(\Lambda)^2$ terms contributing. The distribution is entirely defined by the one at previous iteration and is still Gaussian thanks to the stability by linearity and convolution of Gaussian distributions. To shorten out this monster, we must invoke approximations based on moment matching methods (very similar computations are driven in this paper [8]). Such operations lead to the wanted conclusion. **QED**

Remark: Another approach is suggested in the same paper that proposes to use forward-backward smoothing. Since we already have all measurements before doing our prediction, why not using all the information instead of simply the one up to a certain time? In such a case, the way to proceed is not called filtering anymore, but smoothing. The forward-backward term refers to the fact that there is a back-and-forth during the frame linking phase to minimize errors.

3. Tracking Endocytosis in the developing embryo

This section is dedicated to the practical aspects of this internship, showcasing some results we obtained using tools described in section II on data described in Section I with the tracker u-track 3D [9].

3.1. Description of the raw data.

Let us recall that each movie is under the form of a single TIFF file that preserve the 3D and time structure of the recordings. The data we dispose of is made of four movies: All of them are recordings of two different zones of the embryo. Two of them are in the amnioserosa, the two others located in the lateral ectoderm, see figure (9). Because of a high resolution, it so happens that each file is quite heavy (around 4.2 gigabytes each).

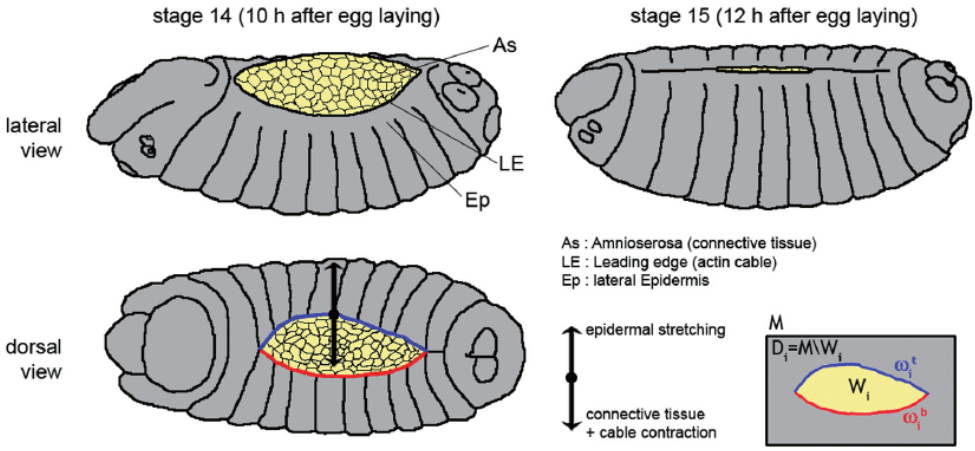


Figure 9. Different parts of the fly embryo, our data specifically focuses on the Lateral Ectoderm and the Amnioserosa

When opening the file, one can observe such images as in figure (10), the goal of the algorithm is to be able to track each white dot (up to some exceptions with the large clusters) on each frame and then assemble the data from each frame to build a reasonable trajectory for each particle. A challenge that does not clearly appear with these pictures is that, with time running, the intensity of the fluorophore decreases, making the detection even harder.

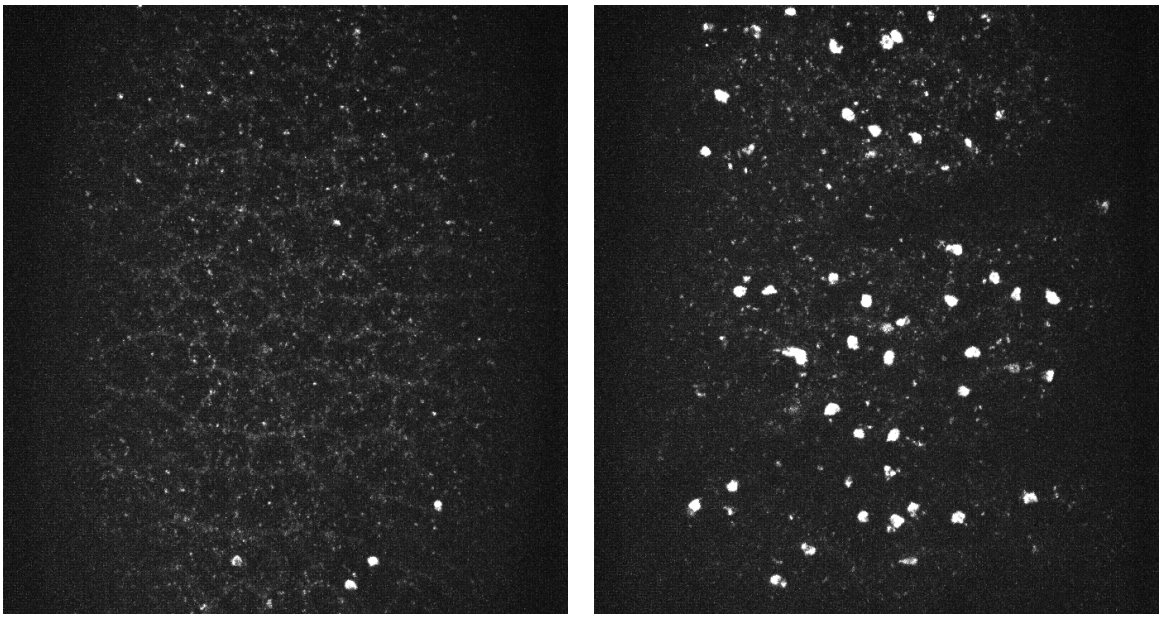


FIGURE 10. First frame coming from a movie, each from a different zone: (left) Lateral Ectoderm, (right) Amnioserosa

3.2. **Preliminary results.** By the end of the first month of this internship, we were able, so far, to detect particles and even assemble a first rendering tracking the general movement of the embryo, yielding motivating results for further computations. Such renderings can be witnessed in Figure (11)

3.3. **Discussion and possible research axis.** Independently of our results, there is room for improvement. As shown in the following picture, it so happens that in small, noisy regions, the algorithm has difficulties to track particles.

This kind of errors from the algorithm are rather unforgiving in a way that this will not allow us to compute quantities such the trajectory lifetime, a critical value to understand the degree of maturation of the endocytic process. Here is what we can improve: SNR, framerate and the algorithm.

- **Signal-to-noise ratio and framerate** The signal-to-noise ratio (SNR) is defined as the quotient of the respective powers of the signal and the noise from the film. It serves as an indicator of the quality of the signal, and by definition, a higher value means a better signal. In our case, we notice that the data remain relatively noisy, preventing the detection of some less intense yet important points. A way to impact the SNR could be by playing on the framerate of our movies. A lower framerate would reduce the trackability (as particles have more time to travel from one frame to the other), the uptake to this approach being that the signal greatly increases in quality and facilitates the detection process. Conversely, doing the opposite has symmetrical consequences (high framerate causes terrible detection due to noise, but has a great trackability), so there is an in-between to find to increase our SNR. This tradeoff issue is technically coming from photon emission restrictions, which is why a solution proposed by COLLINET and ROUDOT is to use another fluorophore, more fluorescent, longer-lasting. This approach is quite costly and complex biochemically-speaking, but it may yield great results in the end.

- **Mathematical Model:** Under all our assumptions, the model we use is designed to capture linear motion: Let $\boldsymbol{u}_t = (x_t, y_t, z_t)$ and $\boldsymbol{v}_t = (dx_t, dy_t, dz_t)$ denote the position and velocity of a particle

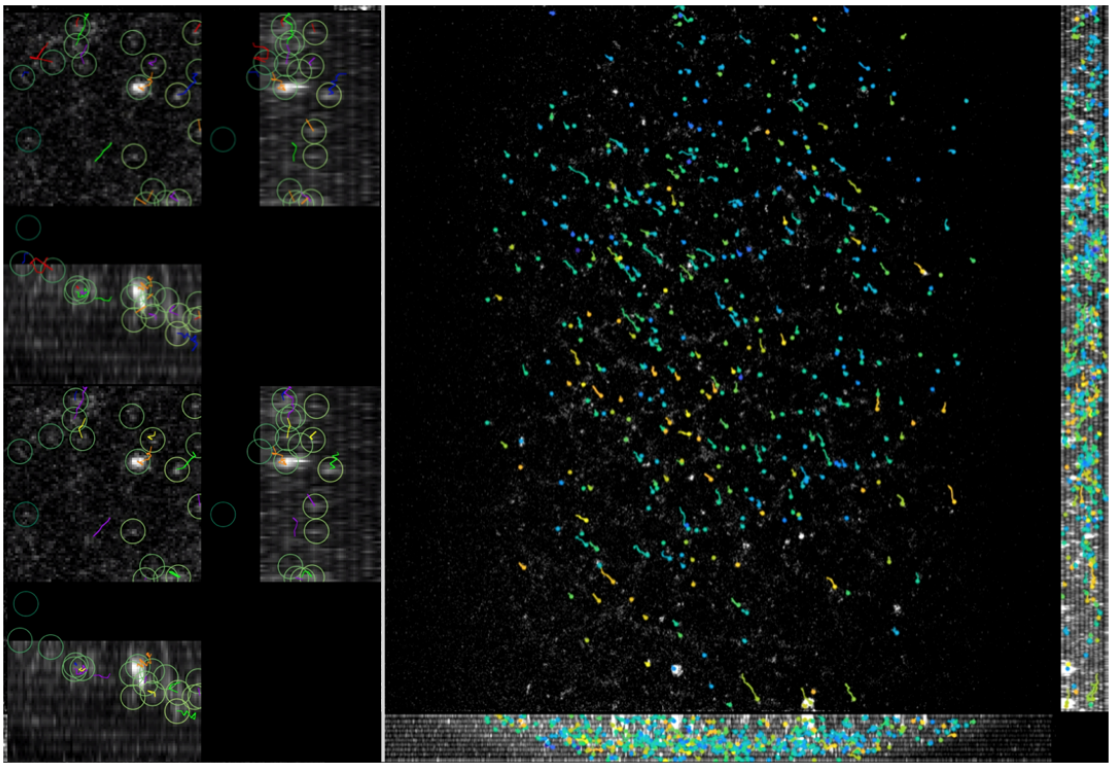


FIGURE 11. Multiple renderings of the tracks superposed onto the movie. The left-hand side represents tracking realized in small zones called ROI (regions of interest)

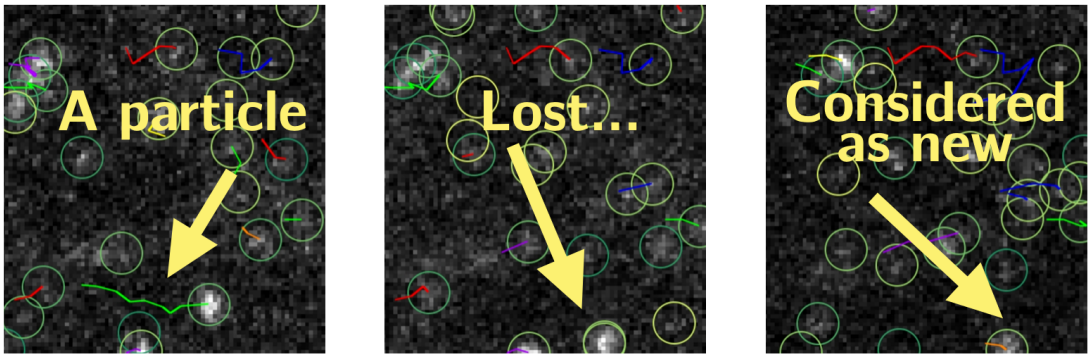


FIGURE 12. Example of a lost trajectory through three consecutive frames even though the signal was clear, such events have a tendency to provide wrong statistics (for instance, it divides molecule lifetime by roughly two each time a link is lost)

$\mathbf{x}_t$ at time t, linear motion implies that the model will only predict movements of the form $\boldsymbol{u}_{t+\delta t} = \boldsymbol{u}_t + \delta t \boldsymbol{v}_t$ (*i.e.,* directed motion with constant speed). Linear motions indeed locally reflect the flow-like structure of embryonic motion and the diffusion-like motion of endocytic process, however, the additional component coming from cell contraction and the stochastic nature of the system dynamic limit the predictive power of a constant-speed assumption. Such a phenomenon is illustrated by Figure (12) in a way that a clearly directed motion (green trajectory) is lost after a drifting event that could not have been predicted by the model. This kind of issue is particularly noticeable when an object abruptly changes direction. A possible lead to improve our tracking is by elaborating a more complex model able to capture the multiscale nature of morphogenetic and endocytic dynamics. One approach being currently explored at I2M is the multiscale decomposition of the motion field through the application of wavelets decomposition of the graph formed by all detected objects.

I wish to address a hearty thanks to Philippe Roudot and Claudio Collinet for their time, support, and immense help since the beginning of this internship.

Appendix A. Fundamental results on probabilities and Markov Chains

Throughout this report we use several of results on Markov chains and Bayesian statistics, we thus propose this section in order to recall important notions and introduce some vocabulary. For the sake of avoiding repetition, we will always work with a probability space, $(\Omega, \mathcal{T}, p)$ except stated otherwise. Since the proof of all following theorems are trivial (or well-known in the literature), we may skip them and redirect the reader to the following book [10] for some proofs)

Definition 1. *(Conditional Probability) Consider two events $A, B \in \mathcal{T}$, the probability that A occurs, knowing B is true is denoted by $p(A|B)$ and is called* ***conditional probability of*** *A* ***knowing*** *B. Given B has a non-zero probability of occurring, the following holds.*

$$p(A|B) = \frac{p(A \cap B)}{p(B)}$$

This notion of conditional probability is crucial for our analysis and will be overly used in this report.

Theorem 2. *(Bayes formula) Let $A, B \in \mathcal{T}$, two events, then they satisfy the next equation*

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}$$

the quantity $p(A|B)$ is often referred to as ***likelihood,*** *while $p(A)$ is called the* ***prior distribution****. Because the quantity $p(B)$ is constant when A is fixed, the formula is sometimes stated the following way*

$$p(A|B) \propto p(B|A)p(A)$$

with '$\propto$' denoting proportionality of the quantities

Theorem 3. *(Total probabilities / Marginalization) Consider an event $A \in \mathcal{T}$ and an at-most countable partition of Ω, $\{B_i\}_{i \in \mathcal{I}}$, then one has*

$$p(A) = \sum_{i \in \mathcal{I}} p(A|B_i)p(B_i)$$

Remark: *In the case $\mathcal{I}$ is non-countable, the previous result can be extended to a continuous version of this theorem (which we will use more often than the discrete one in this report)*

$$p(A) = \int_{\mathcal{I}} p(A|\boldsymbol{b})p(\boldsymbol{b})\, d\boldsymbol{b}$$

Theorem 4. *(Successive conditioning) Let $A_1, ..., A_n \in \mathcal{T}$ be a finite collection of events, then, it holds that*

$$p\left(\bigcap_{i=1}^{n} A_i\right) = p(A_1) \prod_{i=1}^{n} p\left(A_i \;\middle|\; \bigcap_{j=1}^{i-1} A_j\right)$$

as an example, taking $n = 3$ yields) $p(A_1 \cap A_2 \cap A3) = p(A_1|A_2 \cap A_3)p(A_2|A_3)p(A_3)$

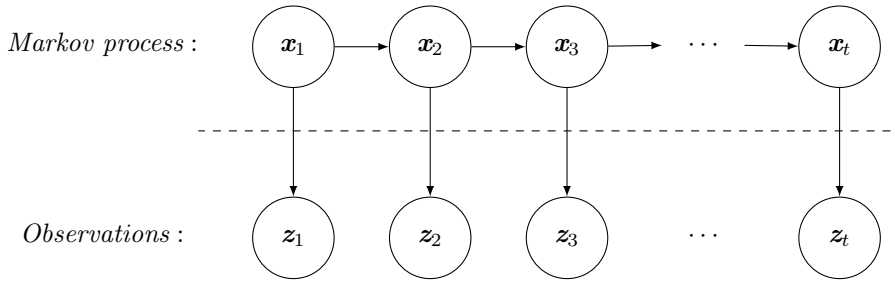
Property 1. *(Markov Processes of order 1) A Markov process is a collection of random variables* $\{X_n\}_{n=1}^{\infty}$ *(called states of the process) satisfying the Markov property (or memoryless property), i.e.: For a given time* n*, it holds*

$$p(X_n|X_1, ..., X_{n-1}) = p(X_n|X_{n-1})$$

A common interpretation of the Markov property is to say that all the information from the past is contained in the previous state of the model. This definition extends itself to higher orders $2, 3, ...$ *by considering the fact that the information is not only contained in* X_{n-1} *but also in prior states, therefore, a Markov process of order* k *would obey*

$$p(X_n|X_1, ..., X_{n-1}) = p(X_n|X_{n-1}, ...X_{n-k})$$

Property 2. *(Seeing HMM as Bayesian networks) Consider the following Hidden Markov Model, with states* $\boldsymbol{x}_{1:t}$ *and* $\boldsymbol{z}_{1:t}$



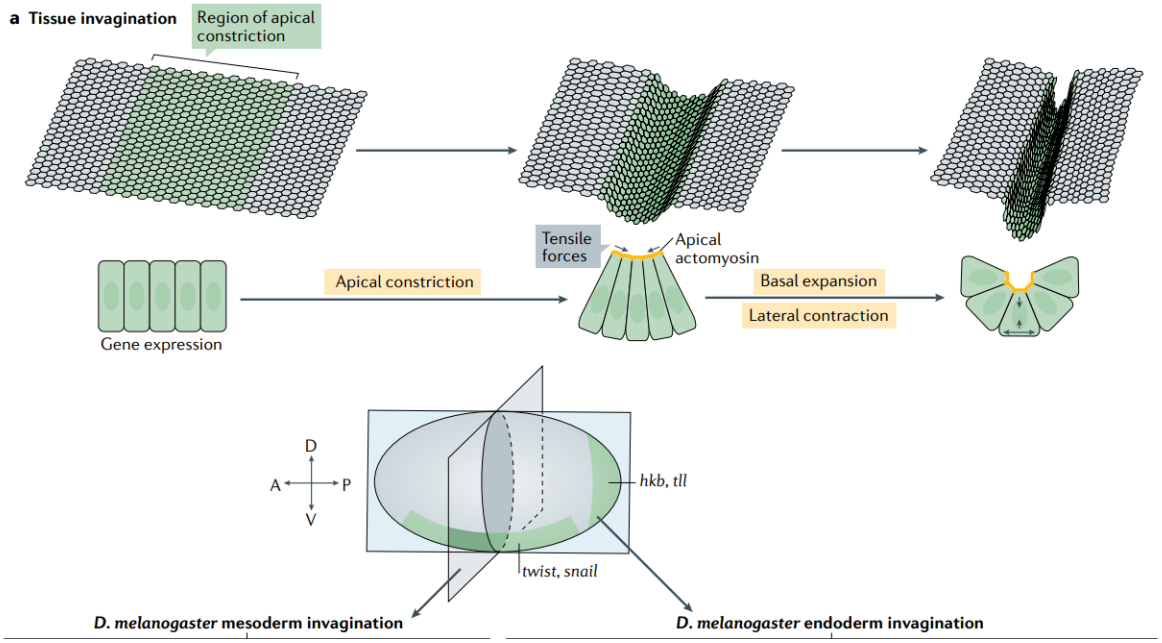
then each state of the model can be considered as a node of an oriented graph. Assume η_1, η_2 *are two nodes in the model, then an arrow from* η_1 *to* η_2 *reads* $p(\eta_2|\eta_1)$*. As an example, take* $\eta_1 = \boldsymbol{x}_1$ *and* $\eta_2 = \boldsymbol{x}_2$*, then the arrow from* $\boldsymbol{x}_1$ *pointing to* $\boldsymbol{x}_2$ *means* $p(\boldsymbol{x}_2|\boldsymbol{x}_1)$ *which can be interpreted as the probability of landing on* x_2 *at time* $t = 2$ *knowing that we are at* x_1 *at time* $t = 1$*.*

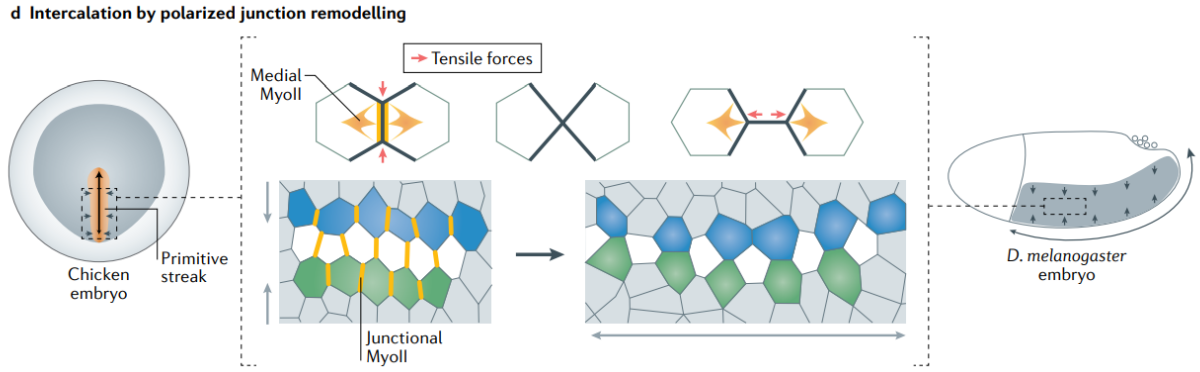
Remark: *This diagram encapsulates perfectly the idea behind Markov axiom. Under this form, one can clearly read the following property*

$$p(\boldsymbol{z}_t|\boldsymbol{x}_t, \boldsymbol{z}_{1:t-1}) = p(\boldsymbol{z}_t|\boldsymbol{x}_t)$$

Appendix B. Additional Visuals

Here are two wonderful visuals extracted from [1] that might assist one in visualizing what is actually going on in the embryo.





References

[1] C. Collinet and T. Lecuit, "Programmed and self-organized flow of information during morphogenesis," Nature Reviews Molecular Cell Biology, vol. 22, pp. 245–265, Jan. 2021.

[2] F. Aguet, C. N. Antonescu, M. Mettlen, S. L. Schmid, and G. Danuser, "Advances in analysis of low signal-to-noise images link dynamin and AP2 to the functions of an endocytic checkpoint," Developmental Cell, vol. 26, pp. 279–291, Aug. 2013.

[3] A. M. Lakoduk, P. Roudot, M. Mettlen, H. M. Grossman, S. L. Schmid, and P.-H. Chen, "Mutant p53 amplifies a dynamin-1/APPL1 endosome feedback loop that regulates recycling and migration," Journal of Cell Biology, vol. 218, pp. 1928–1942, May 2019.

[4] D. Loerke, M. Mettlen, D. Yarar, K. Jaqaman, H. Jaqaman, G. Danuser, and S. L. Schmid, "Cargo and dynamin regulate clathrin-coated pit maturation," PLoS Biology, vol. 7, p. e1000057, Mar. 2009.

[5] Wikipedia, "Kalman-filter — wikipedia, die freie enzyklopädie," 2022. [Online; Stand 27. Mai 2022].

[6] R. E. Kalman, "A new approach to linear filtering and prediction problems," Journal of Basic Engineering, vol. 82, pp. 35–45, Mar. 1960.
[7] P. Roudot, L. Ding, K. Jaqaman, C. Kervrann, and G. Danuser, "Piecewise-stationary motion modeling and iterative smoothing to track heterogeneous particle motions in dense environments," IEEE Transactions on Image Processing, vol. 26, pp. 5395–5410, Nov. 2017.
[8] "Derivation of the IMM filter." http://www.control.isy.liu.se/student/graduate/TargetTracking/IMMderivation.pdf. [Online; accessed 27-May-2022].
[9] P. Roudot, W. R. Legant, Q. Zou, K. M. Dean, E. S. Welf, A. F. David, D. W. Gerlich, R. Fiolka, E. Betzig, and G. Danuser, "u-track 3d: measuring and interrogating intracellular dynamics in three dimensions," bioRxiv, 2020.
[10] Y. Caumel, Probabilités et processus stochastiques. Springer Paris, 2011.
[11] M. Gonzalez-Gaitan and F. Julicher, "The role of endocytosis during morphogenetic signaling," Cold Spring Harbor Perspectives in Biology, vol. 6, pp. a016881–a016881, July 2014.
[12] A. C. Martin, M. Kaschube, and E. F. Wieschaus, "Pulsed contractions of an actin–myosin network drive apical constriction," Nature, vol. 457, pp. 495–499, Nov. 2008.
[13] D. C. a. Alan Bain, Fundamentals of Stochastic Filtering. Stochastic Modelling and Applied Probability n°60, Springer, 1 ed., 2009.
[14] R. Levayer, A. Pelissier-Monier, and T. Lecuit, "Spatial regulation of dia and myosin-II by RhoGEF2 controls initiation of e-cadherin endocytosis during epithelial morphogenesis," Nature Cell Biology, vol. 13, pp. 529–540, Apr. 2011.
[15] S. L. Schmid, A. Sorkin, and M. Zerial, "Endocytosis: Past, present, and future," Cold Spring Harbor Perspectives in Biology, vol. 6, pp. a022509–a022509, Oct. 2014.
[16] R. Rieger, A. Michaelis, and M. M. Green, Glossary of genetics. Berlin, Germany: Springer, 5 ed., July 1991.
[17] F. J. Paola Lecca, Ian Laurenzi, Deterministic versus stochastic modelling in biochemistry and systems biology. Woodhead Publishing Series in Biomedicine, Woodhead Publishing, 1 ed., 2013.
[18] S.-M. CHEN, Y.-S. HSU, and W. L. PEARN, "CAPABILITY MEASURES FOR m -DEPENDENT STATIONARY PROCESSES," Statistics: A Journal of Theoretical and Applied Statistics, vol. 37, pp. 1–24, Jan. 2003.
[19] D. Williams, Probability with Martingales. Cambridge University Press, Feb. 1991.
[20] M. Cavey and T. Lecuit, "Molecular bases of cell-cell junctions stability and dynamics," Cold Spring Harbor Perspectives in Biology, vol. 1, pp. a002998–a002998, Sept. 2009.
[21] K. Jaqaman, D. Loerke, M. Mettlen, H. Kuwata, S. Grinstein, S. L. Schmid, and G. Danuser, "Robust single-particle tracking in live-cell time-lapse sequences," Nature Methods, vol. 5, pp. 695–702, July 2008.
[22] Y.-J. Liu, M. L. Berre, F. Lautenschlaeger, P. Maiuri, A. Callan-Jones, M. Heuzé, T. Takaki, R. Voituriez, and M. Piel, "Confinement and low adhesion induce fast amoeboid migration of slow mesenchymal cells," Cell, vol. 160, pp. 659–672, Feb. 2015.
[23] G. Charras and E. Sahai, "Physical influences of the extracellular environment on cell migration," Nature Reviews Molecular Cell Biology, vol. 15, pp. 813–824, Oct. 2014.
[24] S. Sigismund, L. Lanzetti, G. Scita, and P. P. D. Fiore, "Endocytosis in the context-dependent regulation of individual and collective cell properties," Nature Reviews Molecular Cell Biology, vol. 22, pp. 625–643, June 2021.
[25] S. Jansen, "Markov Jump Processes." https://www.mathematik.uni-muenchen.de/~jansen/jump-processes.pdf, January 26, 2020. [Online; accessed 27-May-2022].
[26] A. Bain and D. Crisan, Fundamentals of Stochastic Filtering. Springer New York, 2009.
[27] N. Akai, S. Ohsawa, Y. Sando, and T. Igaki, "Epithelial cell-turnover ensures robust coordination of tissue growth in drosophila ribosomal protein mutants," PLOS Genetics, vol. 17, p. e1009300, Jan. 2021.
[28] J. M. Wells and J. R. Spence, "How to make an intestine," Development, vol. 141, pp. 752–760, Feb. 2014.