Probabilistic reconstruction of truncated particle trajectories on a closed surface

Yunjiao Lu, Pierre Hodara, Charles Kervrann, Alain Trubuil

Probabilistic reconstruction of truncated particle trajectories on a closed surface

Yunjiao Lu[1,2,$], *Pierre Hodara*[1,$], *Charles Kervrann*[2], *Alain Trubuil*[1 *]

[1]*INRA, UR 1404, MaIAGE, Université Paris-Saclay, Jouy-en-Josas, France*
[2]*Inria, Centre de Recherche Bretagne-Atlantique, EPC SERPICO, Rennes, France*

[$]*these authors contributed equally to this work*
[*]*corresponding author*

ABSTRACT

Investigation of dynamic processes in cell biology very often relies on the observation in two dimensions of 3D biological processes. Consequently, the data are partial and statistical methods and models are required to recover the parameters describing the dynamical processes. In the case of molecules moving over the 3D surface, such as proteins on walls of bacteria cell, a large portion of the 3D surface is not observed in 2D-time microscopy. It follows that biomolecules may disappear for a period of time in a region of interest, and then reappear later. Assuming Brownian motion with drift, we address the mathematical problem of the reconstruction of biomolecules trajectories on a cylindrical surface. A subregion of the cylinder is typically recorded during the observation period, and biomolecules may appear or disappear in any place of the 3D surface. The performance of the method is demonstrated on simulated particle trajectories that mimic MreB protein dynamics observed in 2D time-lapse fluorescence microscopy in rod-shaped bacteria.

1 INTRODUCTION

In 2D and 3D live-cell imaging, spatiotemporal events and biomolecule dynamics are frequently observed with an incomplete field of view. Very often these observations are related to regions of observation (ROO) inside a tissue, a cell, or in the neighborhood of membranes. Nevertheless, it is quite unusual to analyze 3D dynamics of biomolecules or events occurring on a closed surface and observed on a 2D plane. Our work is motivated by the study of dynamics of MreB proteins, moving close to the inner membrane during cell wall construction in rod-shaped bacteria ([**2**], [**20**]). Its dynamics can only be observed in a small region and are recorded as 2D time-lapse movies (Fig. 1a). As for 3D image acquisition, even it can solve the problem of partial observation, is not always appropriate, especially if the objective is to capture fast and temporally short events as described in [**2**]. The frame rate adapted to the scale of dynamics may be too high when compared to the period of time to acquire temporal series of 3D volume ([**5**] and [**9**]).

To our knowledge, identifying re-entrance events of the same entities inside the ROO is not addressed in the literature. In experimental data, when the unobserved region represents a significant part of the entire surface, a complete description of the dynamics on these closed surfaces becomes of paramount importance for deciphering the mechanisms of some processes. In our study of the regulation of the dynamics of MreB protein, as inputs, we consider a set of trajectories estimated by tracking algorithms (e.g. [**14**], [**8**]). These tracking algorithms are very sophisticated and allow us to handle large sets of particles, different stochastic dynamical models [**4**], [**6**], and observation models [**12**], [**16**]. They take into account birth/death events, and/or split/merge events. Particles may be unobserved or undetected for short periods of time, especially in 2D+time microscopy. However, none computational or statistical method manages the situation corresponding to a large hidden region inside the region of interest. Also, the identification of particles leaving the ROO through one border of the domain and re-entering from a far border is not addressed. Our objective is then to provide a generic approach to tackle the problem of the reconstruction of particle trajectories observed on a small part of a closed surface as illustrated in Fig. 1b.

In this paper, we focus on the design and evaluation of a self-contained mathematical framework to tackle the reconstruction of particle trajectories on cylindrical surfaces, given the tracklets observed in a small window sampled on the surface. In our study, the particles are assumed to obey a stochastic Brownian motion with drift and may appear or disappear during the observation period. Split or merge events are not considered in the modeling framework. The trajectory reconstruction problem is defined as the maximization of the likelihood function given tracklets inside the ROO. The optimization problem to be solved is formulated as an integer linear programming problem. The final algorithm is a data-driven algorithm with no hidden parameter to be set by the user. We demonstrate the performance and robustness of our computational method on simulation data, by varying the ratio of observed to unobserved region, the drift and variance of particles, as well as the rates of birth and death of particles.

The remainder of the paper is organized as follows. In Section 2, we present the problem formally and introduce notations. In Section 3, we describe the probabilistic framework, including Poisson processes used to describe birth and death events, and Brownian motion with drift to represent particle motion. We also describe the computational procedure aiming at connecting tracklets belonging to the same trajectory, and then recovering the dynamics of particles on the whole closed surface. Note that we suppose that the curvature of the cylinder is known and so that the movements are represented on a 2D unwrapped surface. In Section 4, the performance of our algorithm is evaluated on simulated data. Finally, we conclude and propose some future work. A summary of notations useful for the evaluation of the likelihood is given in Supplementary Materials (4).

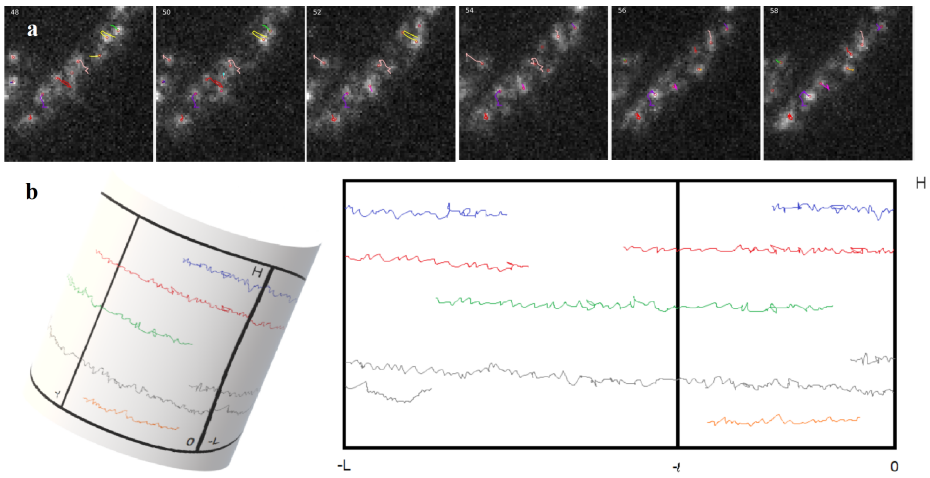

Figure 1: (a): Several consecutive images from a real TIRFM movie[**1**]. Tracks are superposed on the images.(b) left: Illustration of trajectories observed during recorded time $[0, T_S]$ on the surface of a cylinder. Only the motions inside the ROO $]-l, 0[\times[0, H]$ can be observed, even though the dynamics happen on the whole surface; right: Representation of the dynamics on a 2D unwrapped surface $]-L, 0[\times[0, H]$. The objective is to recover the dynamics on the whole surface from the partial observations, by coordinating the inputs through $\{-l\} \times [0, H]$ and the outputs through $\{0\} \times [0, H]$ in a movie during T_S, taking into account particles birth and death events.

2 PROBLEM STATEMENT AND NOTATIONS

We consider a probabilistic model to represent particles that are born, move and die on a cylindric membrane. Formally, let us denote H and L the height and perimeter of the cylinder respectively (see Fig. 1). We associate 2D coordinates $(x, y) \in [-L, 0] \times [0, H]$ to each point of the underlying cylindric manifold. The particles are "born" with a constant rate λ and appear uniformly at random on the membrane surface. We consider a Poisson process with intensity λ to statistically represent the birth events. Each particle is assumed to have the same constant rate of death τ_d such that life duration T_d of a particle follows an exponential law of parameter τ_d. During its lifetime, a particle k born at time t_0 and located at $\mathbf{Z}_0^k = (X_0^k, Y_0^k)$, moves according to Brownian motion with drift. On the set $]-L, 0[\times[0, H]$, the position of the particle at time $t \geq t_0$ prior to its death time is given by

$$\mathbf{Z}_t^k = \mathbf{Z}_0^k + \mathbf{v}(t - t_0) + \mathbf{\Sigma}\mathbf{B}_{t-t_0}^k \tag{1}$$

where $\mathbf{Z}_t^k = (X_t^k, Y_t^k)$, $\mathbf{v} = (v_x, v_y)$, $\mathbf{\Sigma} = \begin{bmatrix} \sigma_x & 0 \\ 0 & \sigma_y \end{bmatrix}$, $\mathbf{B}_t^k$ is a two-dimensional Wiener process. In order to model the topology of the cylinder as illustrated in Fig. 1, we impose deterministic jumps when the process reaches one of the two borders $\{-L\} \times [0, H]$ or $\{0\} \times [0, H]$. For any $y \in [0, H]$, the process reaching position $(-L, y)$ jumps to position $(0, y)$ and vice versa.

In y direction the initial position of a particle lies between $[0, H]$. When a particle hits the vertical borders, its following trajectory is no longer considered. Finally, we assume that each particle behaves independently from the others and that there is no fission or fusion of particles.

In the sequel, we observe the dynamics at discrete times $\Delta t, 2\Delta t, 3\Delta t \ldots$ We denote Δt the time step on the subset $[-l, 0] \times [0, H]$ with $l < L$. The observations are recorded during a time interval $[0, T_S]$. As we suppose that a particle does not change its drift direction along its trajectory, we assume that $v_x > 0$, even though particles can actually move in both directions, which requires a classification to separate them into two groups. We consider that an observed tracklet of a given trajectory is an output if the last observed point of the segment is within a neighborhood of $\{0\} \times [0, H]$. Meanwhile, we consider that it is an input if the first observed point is within a neighborhood of $\{-l\} \times [0, H]$. Our main objective is then to associate the set of tracklets exiting the observed set $[-l, 0] \times [0, H]$ with the set of tracklets entering this observation set. The challenge is to correctly match the outputs and the inputs associated to particles (see Fig. 1).

3 PROBABILISTIC MODELS AND METHODS

Let us consider a given sample S, the observation set of all the trajectories. Define the sets $O_S = \{o_1, ..., o_p\}$ and $I_S = \{i_1, ..., i_q\}$ of p outputs and q inputs. Each output $o = (t_o, y_o) \in O_S$ is characterized by its output time t_o and its position $y_o \in [0, H]$ where the particle left the observed region. Similarly each input $i = (t_i, y_i) \in I_S$ is characterized by its input time t_i and its position $y_i \in [0, H]$ where it entered the observed region. A particle "involved" in an output $o \in O_S$ either died after time t_o in the unobserved region, or is "involved" in a given input $i \in I_S$ with $t_i > t_o$. We will denote this event by $\{o \to i\}$. Similarly a particle "involved" in an input $i \in I_S$ was either born before time t_i in the unobserved region, or is "involved" in a given output $o \in O_S$ with $t_i > t_o$, which corresponds to the event $\{o \to i\}$.

Define $c = (D_c, B_c, b_c)$ with $D_c \subset O_S$, $B_c \subset I_S$ and b_c a bijection from $O_S \setminus D_c$ to $I_S \setminus B_c$ in order to describe the configuration for which all outputs in D_c died in the unobserved region, all inputs in B_c are born in the unobserved region, and the event

$$\bigcap_{o \in O_s \setminus D_c} \{o \to b_c(o)\}$$

was realized. Our aim is to determine the maximum likelihood configuration c given the sample S. The outline of the connection procedure is given in Fig. 2, to facilitate the understanding of the modeling steps.

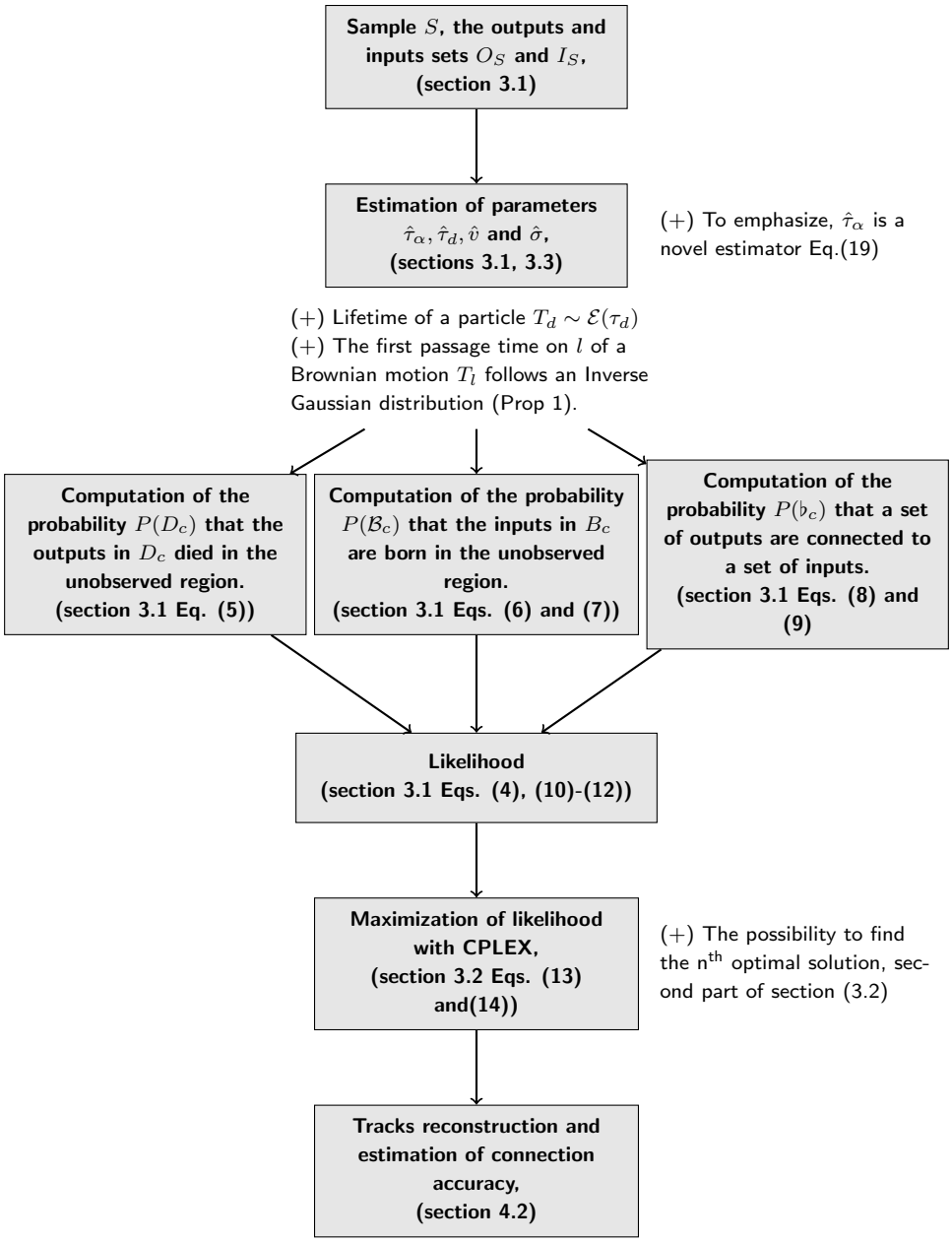

Figure 2: An outline of the connection procedure: from the estimation of the parameters to connection accuracy measurement, including likelihood formulation. All notations are defined in the corresponding sections.

3.1 *Likelihood of a configuration*

In this section, our objective is to derive an analytic expression of the likelihood $Q(c)$ of a configuration c. The aim is to find, for a given sample S, the configuration $\hat{c}$ such that $P(\hat{c}/S)$ is maximal. It is difficult to calculate directly $P(\hat{c}/S)$. Since $c \subset S \subset O_S$, we can compute $P(\hat{c}/S)$ working conditionally on O_S.

However, since the model is in continuous time and involves random variables with continuous densities with respect to the Lebesgue measure, the conditional probability $P(c/O_S)$ is equal to 0. This prevents to compute directly $P(\hat{c}/S)$ with the classical conditional formula

$$P\left(c/S\right) = \frac{P\left(c/O_S\right)}{P(S/O_S)},$$

because it gives $P(S/O_S) = \sum_{c \in \mathcal{C}_S} P(c/O_S) = 0$.

Therefore, for each input $i = (t_i, y_i) \in I_S$, we consider a spatiotemporal neighborhood $V_i^\epsilon = T_i^\epsilon \times H_i^\epsilon$ with $T_i^\epsilon = [t_i - \frac{\epsilon}{2}, t_i + \frac{\epsilon}{2}]$ and $H_i^\epsilon = [y_i - \frac{\epsilon}{2}, y_i + \frac{\epsilon}{2}]$ for some $\epsilon > 0$.

The idea is to replace a given configuration c by a set $\mathcal{C}_c^\epsilon$ of configurations where each element $c^* \in \mathcal{C}_c^\epsilon$ is similar to c but each input $i \in I_S$ is replaced by an input in V_i^ϵ. Formally, for each configuration c leading to the input set I_S, $\mathcal{C}_c^\epsilon$ is the set of configurations defined as follows: $c^* = (D_{c^*}, B_{c^*}, b_{c^*}) \in \mathcal{C}_c^\epsilon$ if and only if for each $i \in I_S$, there exist $i_\epsilon^* \in V_i^\epsilon$ satisfying:

$$\begin{cases} D_{c^*} = D_c, \\ B_{c^*} = \{i_\epsilon^*, i \in B_c\}, \\ \text{For each } i \in I_S \setminus B_c,\ b_{c^*}\left(b_c^{-1}(i)\right) = i_\epsilon^*. \end{cases}$$

With this definition, we have

$$P(c/S) = \lim_{\epsilon \to 0} P\left(\mathcal{C}_c^\epsilon/S\right) = \lim_{\epsilon \to 0} \frac{P\left(\mathcal{C}_c^\epsilon/O_S\right)}{\sum_{c' \in \mathcal{C}_S} P\left(\mathcal{C}_{c'}^\epsilon/O_S\right)}. \tag{2}$$

In what follows, we study the behavior of $P(\mathcal{C}_c^\epsilon/O_S)$ when ϵ goes to 0. We will always work conditionally on the realization of the output set O_S but we will keep this conditioning implicit and write $P(\mathcal{C}_c^\epsilon)$ instead of $P(\mathcal{C}_c^\epsilon/O_S)$ in order to simplify the notations. The study of $P(\mathcal{C}_c^\epsilon)$ will involve the probability for a particle to die in the unobserved region but also the probability that a particle born in this unobserved region enters the observed one in a given spatiotemporal neighborhood V_i^ϵ.

Furthermore, we assume that the particles born in the unobserved region, enter the observed one with a constant rate τ_α and with a uniform distribution on $\{-l\} \times [0, H]$. This is consistent with the fact that the particles are born with constant rate λ and appear uniformly at random on the membrane surface. Therefore, denote by N_α the Poisson process of intensity τ_α counting the number of inputs involved by particles born in the unobserved region.

Consider an output $o \in O_S$ and the possibility for the particle involved in o to die in the unobserved region. We have the following proposition (see [**18**], [**19**], [**21**]).

Proposition 1 *Given the particle motion model as Brownian motion with drift as described in equation 1, the first passage time noted as T_l on the entrance line $\{-l\}\times[0,H]$ of a particle starting at position $z_0=(0,y_0)$ for some $y_0\in[0,H]$, follows a law of inverse Gaussian, that is $T_l \sim IG\left(\frac{l_u}{\mathbf{v}_x}, \left(\frac{l_u}{\sigma_x}\right)^2\right)$ where $l_u := L-l$ is the length of the unobserved region.*

Recall that if $X\sim IG(\mu,\lambda)$, then $X\geq 0$ almost surely, and for each $x\geq 0$,

$$P(X\leq x)=\int_0^x \sqrt{\frac{\lambda}{2\pi y^3}}\exp\big(-\frac{\lambda(y-\mu)^2}{2\mu^2 y}\big)dy. \tag{3}$$

In our framework, the event corresponding to the death of a particle with life duration T_d following an exponential law of parameter τ_d in the unobserved region is precisely $\{T_d<T_l\}$. Hence, we can derive an explicit expression of $P(\mathcal{C}_c^\epsilon)$.

Assume ϵ small enough so that for each $i,i'\in I_S$, $T_i^\epsilon\cap T_{i'}^\epsilon=\emptyset$. For a given configuration c and a given $\epsilon>0$, we will write $\mathcal{C}_c^\epsilon=(D_c,\mathcal{B}_c^\epsilon,\flat_c^\epsilon)$ with $\mathcal{B}_c^\epsilon=\{B_{c^*},c^*\in\mathcal{C}_c^\epsilon\}$ and $\flat_c^\epsilon=\{b_{c^*},c^*\in\mathcal{C}_c^\epsilon\}$.

Due to the independent behavior of the particles, we have the following decomposition:

$$P(\mathcal{C}_c^\epsilon)=P(D_c)P(\mathcal{B}_c^\epsilon)P(\flat_c^\epsilon). \tag{4}$$

We can then compute separately the probabilities of events D_c, $\mathcal{B}_c^\epsilon$ and $\flat_c^\epsilon$. First, note that we can assume without loss of generality that each output $o\in D_c$ starts at time $t_o=0$ and that only the position $y_o\in[0,H]$ fluctuates with o, but with no influence on T_d or T_l. Moreover, the loss of memory property of the exponential law ensures that the life duration T_d of the particle after the output o still follows an exponential law of parameter τ_d.

Since all outputs behave identically and independently, we have $P(D_c)=P(T_d<T_l)^{|D_c|}$,where $|D_c|$ stands for the cardinal of D_c. According to proposition 1, and since T_d and T_l are independent, we have

$$\begin{aligned}
P(T_d<T_l) &= \int_0^{+\infty}\int_0^{t_l} f_{T_d}(t_d)f_{T_l}(t_l)dt_d\,dt_l, \\
&= \int_0^{+\infty}\int_0^{t_l} \tau_d e^{-\tau_d t_d}\frac{l_u}{\sigma_x\sqrt{2\pi t_l^3}}\exp\left(-\frac{(\mathbf{v}_x t_l-l_u)^2}{2\sigma_x^2 t_l}\right)dt_d\,dt_l, \\
&= \int_0^{+\infty}\frac{l_u\left(1-e^{-\tau_d t_l}\right)}{\sigma_x\sqrt{2\pi t_l^3}}\exp\left(-\frac{(\mathbf{v}_x t_l-l_u)^2}{2\sigma_x^2 t_l}\right)dt_l,
\end{aligned} \tag{5}$$

where f_{T_d} and f_{T_l} respectively stand for the density functions of T_d and T_l.

Now, consider the event $\mathcal{B}_c^\epsilon$. We call "spontaneous input" an input related to a particle born in the unobserved region that has never been observed. The set $\mathcal{B}_c^\epsilon$ is defined so that,

for each input $i \in B_c$, we have exactly one "spontaneous input" appearing during the time interval T_i^ϵ, with a position in H_i^ϵ. Moreover, outside $\cup_{i\in B_c} T_i^\epsilon$, there is no "spontaneous input". Formally, we have

$$\mathcal{B}_c^\epsilon = \left\{ N_\alpha \left([0,T_S] \setminus \bigcup_{i\in B_c} T_i^\epsilon \right) = 0 \right\} \bigcap \left(\bigcap_{i\in B_c} \left(\{N_\alpha(T_i^\epsilon) = 1\} \cap H_i^\epsilon \right) \right), \tag{6}$$

where N_α is a Poisson process of intensity τ_α associated to the counting of inputs involved by particles born in the unobserved region on the time interval $[0, T_S]$. In order to simplify the notations, H_i^ϵ denotes also the event of "spontaneous" appearance of an input i in H_i^ϵ. This event is independent of the process N_α, and since the "spontaneous inputs" appear uniformly on $[0, H]$, we have $P(H_i^\epsilon) = \frac{\epsilon}{H}$.

Meanwhile, for any time interval I, $N_\alpha(I)$ follows a Poisson law of parameter $\tau_\alpha |I|$ where $|I|$ denotes the length of the interval I. Since ϵ is small enough so that for each $i, i' \in I_S$, $T_i^\epsilon \cap T_{i'}^\epsilon = \emptyset$, $N_\alpha(T_i^\epsilon)$ and $N_\alpha(T_{i'}^\epsilon)$ are independent. Consequently, we can compute $P(\mathcal{B}_c^\epsilon)$ as follows:

$$P(\mathcal{B}_c^\epsilon) = e^{-\tau_\alpha(T_S - |B_c|\epsilon)} \left(\epsilon \tau_\alpha e^{-\epsilon\tau_\alpha} \frac{\epsilon}{H} \right)^{|B_c|} = \left(\frac{\epsilon^2 \tau_\alpha}{H} \right)^{|B_c|} e^{-\tau_\alpha T_S}. \tag{7}$$

Finally, consider the event $\flat_c^\epsilon$. For each input $i \in I_S \setminus B_c$, we denote by $\{o_c^i \to V_i^\epsilon\}$ the survival event of the particle involved in the output $o_c^i = b_c^{-1}(i)$ in the unobserved region which appears in the spatiotemporal neighborhood V_i^ϵ. Since the particles behave independently, we have

$$P(\flat_c^\epsilon) = \prod_{i\in I_S \setminus B_c} P\Big(\{o_c^i \to V_i^\epsilon\} \Big). \tag{8}$$

In the sequel, we consider a given input $i \in I_S \setminus B_c$ and its related output $o = b_c^{-1}(i)$. Defining $s_i = t_i - t_o$ and $h_i = y_i - y_o$ allows us to center the situation around the output o in the following way. A particle born at time 0 in position $z_0 = (0,0)$ has a life duration T_d following an exponential law of parameter τ_d. During its lifetime, the position of the particle is driven by a Brownian motion with drift $\mathbf{Z}_t = (X_t, Y_t)$: $\mathbf{Z}_t = \mathbf{v}t + \mathbf{\Sigma}\mathbf{B}_t$, where $\mathbf{B}_t$ is a two-dimensional Wiener process and $\mathbf{v}$ and $\mathbf{\Sigma}$ are given in Equation (1). Define T_l the first reaching time of $l_u = L - l$ of the process X_t. The event $\{o \to V_i^\epsilon\}$ can now be written as follows:

$$\{o_c^i \to V_i^\epsilon\} = \{T_d > T_l\} \bigcap \left\{ T_l \in \left[s_i - \frac{\epsilon}{2}, s_i + \frac{\epsilon}{2} \right] \right\} \bigcap \left\{ Y_{T_l} \in \left[h_i - \frac{\epsilon}{2}, h_i + \frac{\epsilon}{2} \right] \right\}.$$

This expression corresponds exactly to the fact that in order to realize $\{o_c^i \to V_i^\epsilon\}$ the particle needs to have a life duration longer than its first reaching time of l_u and to appear in the spatiotemporal neighborhood $\left[s_i - \frac{\epsilon}{2}, s_i + \frac{\epsilon}{2} \right] \times \left[h_i - \frac{\epsilon}{2}, h_i + \frac{\epsilon}{2} \right]$. Furthermore, T_d follows an

exponential law of parameter τ_d, Y_t follows a Gaussian law of parameters $\mathbf{v}_y t$ and $\sigma_y^2 t$ and $T_l \sim IG\left(\frac{l_u}{\mathbf{v}_x}, \left(\frac{l_u}{\sigma_x}\right)^2\right)$. Moreover, due to the fact that $\mathbf{\Sigma}$ is diagonal, the process Y_t is not only independent of T_d but also of T_l. This allows us to write

$$P\left(\{o_c^i \to V_i^\epsilon\}\right) = \int_{s_i-\frac{\epsilon}{2}}^{s_i+\frac{\epsilon}{2}} f_{T_l}(t_l) \left(\int_{t_l}^{+\infty} f_{T_d}(t_d) \left(\int_{h_i-\frac{\epsilon}{2}}^{h_i+\frac{\epsilon}{2}} f_{Y_{t_l}}(y) dy \right) dt_d \right) dt_l.$$

As the two integrals involve a small domain of size ϵ, $P\left(\{o_c^i \to V_i^\epsilon\}\right) \sim \mathcal{O}(\epsilon^2)$, and

$$\lim_{\epsilon \to 0} \frac{P\left(\{o_c^i \to V_i^\epsilon\}\right)}{\epsilon^2} = f_{T_l}(s_i) f_{Y_{s_i}}(h_i) \int_{s_i}^{+\infty} f_{T_d}(u) du \tag{9}$$

$$= \frac{l_u}{\sigma_x \sqrt{2\pi s_i^3}} \exp\left(-\frac{(\mathbf{v}_x s_i - l_u)^2}{2\sigma_x^2 s_i} \right) \frac{1}{\sigma_y \sqrt{2\pi s_i}} \exp\left(-\frac{(h_i - \mathbf{v}_y s_i)^2}{2\sigma_y^2 s_i} \right) e^{-\tau_d s_i}$$

$$= \frac{l_u}{2\pi \sigma_x \sigma_y s_i^2} \exp\left(-\frac{(\mathbf{v}_x s_i - l_u)^2}{2\sigma_x^2 s_i} - \frac{(h_i - \mathbf{v}_y s_i)^2}{2\sigma_y^2 s_i} - \tau_d s_i \right).$$

For each configuration c, we calculate the likelihood $Q(c)$ of the configuration c as follows:

$$Q(c) \;:=\; \lim_{\epsilon \to 0} \frac{P(\mathcal{C}_c^\epsilon)}{\epsilon^{2|I_S|}}.$$

From (4) and Equations (5, 7, 8 and 9), we finally obtain the likelihood

$$Q(c) \;=\; \left(\frac{\tau_\alpha}{H}\right)^{|B_c|} e^{-\tau_\alpha T_S} \left(\int_0^{+\infty} \frac{l_u \left(1 - e^{-\tau_d t_l}\right)}{\sigma_x \sqrt{2\pi t_l^3}} \exp\left(-\frac{(\mathbf{v}_x t_l - l_u)^2}{2\sigma_x^2 t_l} \right) dt_l \right)^{|D_c|}$$

$$\times \prod_{i \in I_S \setminus B_c} \left[\frac{l_u}{2\pi \sigma_x \sigma_y s_i^2} \exp\left(-\frac{(\mathbf{v}_x s_i - l_u)^2}{2\sigma_x^2 s_i} - \frac{(h_i - \mathbf{v}_y s_i)^2}{2\sigma_y^2 s_i} - \tau_d s_i \right) \right]. \tag{10}$$

Note that the limit when ϵ goes to 0 of $\frac{P(\mathcal{C}_c^\epsilon)}{\epsilon^{2|I_S|}}$ is well defined, strictly positive, and that the exponent $2|I_S|$ does not depend on the configuration c.

Recalling (2), this allows us to write

$$P(c/S) = \frac{Q(c)}{\sum_{c' \in \mathcal{C}_S} Q(c')} \tag{11}$$

and as a consequence, we have

$$\hat{c} = \underset{c \in \mathcal{C}_S}{\operatorname{argmax}} \{Q(c)\}. \tag{12}$$

3.2 *Maximum likelihood and optimal configuration*

The aim of this section is to identify the configuration c corresponding to the maximal likelihood $Q(c)$ (see Equation (10)). Define

$$\beta := -\log\left(\frac{\tau_\alpha}{H}\right),$$

$$\delta := -\log\left(\int_0^{+\infty} \frac{l_u\left(1-e^{-\tau_d t_l}\right)}{\sigma_x\sqrt{2\pi t_l^3}} \exp\left(-\frac{(\mathbf{v}_x t_l - l_u)^2}{2\sigma_x^2 t_l}\right) dt_l\right)$$

and for each configuration c and each $i \in I_S \setminus B_c$

$$\gamma_c^i := -\log\left[\frac{l_u}{2\pi\sigma_x\sigma_y s_i^2} \exp\left(-\frac{(\mathbf{v}_x s_i - l_u)^2}{2\sigma_x^2 s_i} - \frac{(h_i - \mathbf{v}_y s_i)^2}{2\sigma_y^2 s_i} - \tau_d s_i\right)\right].$$

It follows that

$$\begin{aligned} \hat{c} \;=\; \underset{c\in C}{\operatorname{argmax}}\, Q(c) \;&=\; \underset{c\in C}{\operatorname{argmin}} -\log\left(Q(c)\right) \\ &=\; \underset{c\in C}{\operatorname{argmin}} \left(\beta|B_c| + \delta|D_c| + \sum_{i\in I_S\setminus B_c} \gamma_c^i\right). \end{aligned} \tag{13}$$

This decomposition allows us to consider a linear optimization problem where β represents the cost of the spontaneous birth of an input, δ the cost of the death of an output and γ_c^i the cost of the connection between the output $b_c^{-1}(i)$ and the input i. The cost of connection can be defined for any couple $(o,i) \in O_S \times I_S$ as

$$\gamma_o^i := -\log\left[\frac{l_u}{2\pi\sigma_x\sigma_y s_{o,i}^2} \exp\left(-\frac{(\mathbf{v}_x s_{o,i} - l_u)^2}{2\sigma_x^2 s_{o,i}} - \frac{(h_{o,i} - \mathbf{v}_y s_{o,i})^2}{2\sigma_y^2 s_{o,i}} - \tau_d s_{o,i}\right)\right],$$

where $s_{o,i} := t_i - t_o$, $h_{o,i} = y_i - y_o$ and the convention $\gamma_o^i = +\infty$ if $t_i \leq t_o$.

In order to write in a canonical way this linear optimization problem, we associate to each configuration c a family of coefficients $(c^{o,i})_{(o,i)\in O_S\times I_S}$ such that $c^{o,i} = 1$ if $b_c(o) = i$ and $c^{o,i} = 0$ if $b_c(o) \neq i$. Since an output can be connected to at most one input, for each $o \in O_S$, $\sum_{i\in I_S} c^{o,i} \in \{0,1\}$ and $\sum_{i\in I_S} c^{o,i} = 0$ corresponds to the death of the output o. Similarly, for each $i \in I_S$, $\sum_{o\in O_S} c^{o,i} \in \{0,1\}$ and $\sum_{o\in O_S} c^{o,i} = 0$ corresponds to the fact that the input i is a "spontaneous input".

Our optimization problem is then equivalent to finding the family of coefficients $(c^{o,i})_{(o,i)\in O_S\times I_S}$ that minimizes the quantity

$$\beta\left(\sum_{i\in I_S}\left(1 - \sum_{o\in O_S} c^{o,i}\right)\right) + \delta\left(\sum_{o\in O_S}\left(1 - \sum_{i\in I_S} c^{o,i}\right)\right) + \sum_{o\in O_S}\sum_{i\in I_S} \gamma_o^i c^{o,i}$$

or equivalently

$$K(c) := \sum_{o\in O_S}\sum_{i\in I_S}\left(\gamma_o^i - \beta - \delta\right)c^{o,i} \text{ s.t. } \begin{cases} \forall o \in O_S,\, \forall i \in I_S,\, c^{o,i} \in \{0,1\}, \\ \forall o \in O_S,\, \sum_{i\in I_S} c^{o,i} \in \{0,1\}, \\ \forall i \in I_S,\, \sum_{o\in O_S} c^{o,i} \in \{0,1\}. \end{cases} \tag{14}$$

In order to avoid to have infinite costs γ_o^i when $t_i \leq t_o$,, we can also impose $c^{o,i} = 0$ if $t_i \leq t_o$. Actually the problem (14) is a conventional linear optimization problem which can be solved by applying the CPLEX Linear Programming solver (https://www.ibm.com/analytics/cplex-optimizer).

The configuration $\hat{c}$ is then the solution of the optimization problem (14) and corresponds to the most likely configuration given the sample S. In order to complete the study, we propose to compute the following most likely configurations in a reccurent way by solving (14) with additional constraints ensuring that the solution is different from the previous ones. In other words we define recursively the sequence $(c_n)_{n\in\mathbb{N}}$ in the following way:

- $c_1 := \hat{c}$
- $\forall n \geq 2$, c_n solves (14) with the $n-1$ additional constraints

$$\forall k \in \{1,\ldots,n-1\},\ \sum_{o\in O_S}\sum_{i\in I_S}\left[c_n^{o,i}(1-c_k^{o,i}) + (1-c_n^{o,i})c_k^{o,i}\right] \geq 1. \tag{15}$$

With this definition, c_n is then the $n-$th most likely configuration. When n is greater than the number n_S of configurations compatible with the sample S, the constraints are impossible to satisfy. In other words this sequence is well defined up to n_S.

3.3 *Estimation of parameters*

Several parameters are involved in our computational approach. In this section, we propose clues to set these parameters. First, the parameters $\mathbf{v}$ and $\mathbf{\Sigma}$ can be estimated with classical maximum likelihood estimation procedures.

Second, we propose an estimator $\hat{\tau}_d$ of τ_d as explained below. The sample S can be considered as a set of points $p = (t_p, \mathbf{Z}_p)$ observed at time t_p and position $\mathbf{Z}_p = (X_p, Y_p)$ grouped in clusters s corresponding to tracklets of trajectories. The death of a particle in the observed region is detected in S for each point $p \in S$ for which the associated tracklet s_p has no successor point at time $t_p + \Delta t$. In order to be sure that the absence of successor is effectively due to the death of a particle and not to a particle leaving the observed region, we restrict the analysis to a region excluding a neighborhood of the border. However, we can check in this neighborhood the existence of successors for points in the restricted region.

We denote by $S_r \subset S$ the sample of points in the restricted region. For each point $p \in S_r$, we denote by D_p the event corresponding to the absence of successor for p. This corresponds to the fact that the particle involved in p died during the time interval $[t_p, t_p + \Delta t]$. Since the life duration T_d of a particle follows an exponential law of parameter τ_d, and the absence of memory property of the exponential law, we have

$$P(D_p) = P(T_d \in [0, \Delta t]) = 1 - e^{-\tau_d \Delta t}. \tag{16}$$

Hence, we define our estimator $\hat{\tau}_d$ as

$$\hat{\tau}_d = \frac{1}{\Delta t |S_r|} \sum_{p \in S_r} 1[D_p], \tag{17}$$

where $|S_r|$ stands for the number of points in S_r and $1[\cdot]$ denotes the indicator function. Due to the absence of memory property of the exponential law, the random variables $1[D_p]$ are i.i.d. As $|S_r|$ goes to $+\infty$, the strong law of large numbers yields to

$$\lim_{|S_r| \to \infty} \hat{\tau}_d = \frac{1 - e^{-\tau_d \Delta t}}{\Delta t} \qquad a.s.$$

The justification of this choice for $\hat{\tau}_d$ relies in the following almost sure convergence:

$$\lim_{\Delta t \to 0} \lim_{|S_r| \to \infty} \hat{\tau}_d = \tau_d \qquad a.s. \tag{18}$$

Our estimator $\hat{\tau}_d$ is then consistent as Δt is small enough. Moreover, since the variables $1[D_p]$ are i.i.d Bernoulli random variables, we can calculate the related confidence interval. If q_α denotes the α-quantile of the standard normal distribution, we have the following confidence interval of level α for $\frac{1-e^{-\tau_d \Delta t}}{\Delta t}$:

$$CI_\alpha = \left[\hat{\tau}_d - q_\alpha \sqrt{\frac{\hat{\tau}_d \left(\frac{1}{\Delta t} - \hat{\tau}_d \right)}{|S_r|}}, \hat{\tau}_d + q_\alpha \sqrt{\frac{\hat{\tau}_d \left(\frac{1}{\Delta t} - \hat{\tau}_d \right)}{|S_r|}} \right]. \tag{19}$$

If Δt is small enough, we get a good approximation of a confidence interval of level α for τ_d since

$$\lim_{\Delta t \to 0} \frac{1 - e^{-\tau_d \Delta t}}{\Delta t} = \tau_d.$$

Now, we describe the estimation procedure for the rate τ_α of "spontaneous inputs" induced by particles born in the unobserved region $[-L, -l] \times [0, H]$ and reached the border $\{-l\} \times [0, H]$. We assume here that the parameters $\mathbf{v}$, $\mathbf{\Sigma}$ and τ_d are known, keeping in mind that in practice estimators are used instead. As introduced earlier, L is the perimeter of the cylinder, l is the length of the observed region, and $l_u = L - l$ is the length of the unobserved region. For a given length x, we denote by N_x the number of tracklets born in the region

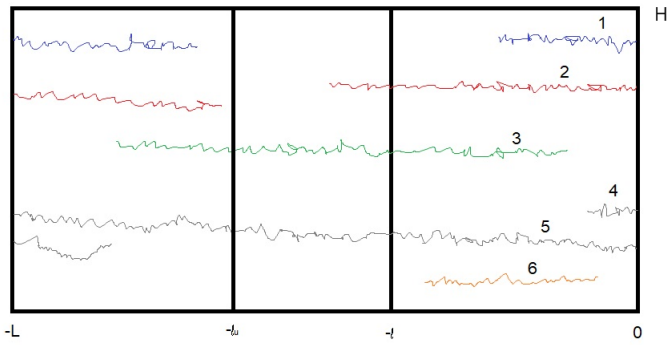

Figure 3: An artificially constructed zone $]-l_u, 0] \times [0, H]$ having the same size as the unobserved region $]-L, -l] \times [0, H]$. The observed region is $]-l, 0] \times [0, H]$; as the width of the invisible part is l_u, the extended zone has width $l_e = l_u - l$.

$]-x, 0] \times [0, H]$ and reached the border $\{0\} \times [0, H]$. Accordingly, $\frac{N_{l_u}}{T_S}$ is a consistent estimator of τ_α since the dynamics are assumed to be homogeneous on the surface of cylinder. Our aim is actually to build an estimator for τ_α in the case where $l_u > l$ which prevents us to compute directly N_{l_u}. Therefore, we compute N_l by taking the whole observed region into account, and denote by S_l^* the set of tracklets having an input in $\{-l\} \times [0, H]$ and an output in $\{0\} \times [0, H]$. For each tracklet $s \in S_l^*$ and each length $x \in [0, l_u]$, we denote by B_s^x the event corresponding to the birth of s within $]-l-x, -l] \times [0, H]$. Let $l_e = l_u - l$ be the length of the extended zone $[-l_u, -l] \times [0, H]$. We are now interested in the realization of the events $B_s^{l_e}$.

In Fig. 3, $N_l = 2$ correspond to tracks #1 and #4, $S_l^* = \{2, 5\}$, and the event $B_2^{l_e}$ is realized while $B_5^{l_e}$ is not.

Note that since the particles have the same independent dynamics, $P(B_s^x)$ does not depend on s. For $x < l$, this probability can easily be estimated as follows:

$$\hat{p}_x = \frac{N_x}{|S_o|},$$

where S_o is the set of tracklets having an output in $\{0\} \times [0, H]$. The strong law of large numbers yields a consistent estimator and allows us, in the case where $l_e < l$, to define our estimator $\hat{\tau}_\alpha$ as follows:

$$\hat{\tau}_\alpha = \frac{N_l + \hat{p}_{l_e}|S_l^*|}{T_S}. \tag{20}$$

Intuitively, this estimator amounts to counting the number of particles reaching $\{0\} \times [0, H]$ with weight 1 for each tracklet that we actually saw being born in the observed region and

with weight $\hat{p}_{l_e}$ for each spontaneous input that appeared in $\{-l\} \times [0, H]$. Note that, as $N_{l_u} = N_l + \sum_{s \in S_l^*} 1[B_s^{l_e}]$, $\hat{\tau}_\alpha$ is an unbiased estimator of τ_α. Moreover, if we assume that the number of observed tracklets grows linearly with the observation time T_S, this estimator is consistent when T_S goes to $+\infty$.

Now, we consider the case $l < l_e < 2l$ which can easily be extended to the general case $l < l_e$. Consider $s \in S_l^*$ and denote for each interval $J \subset [-L, 0]$ the event B_s^J where the tracklet s is born in the region $J \times [0, H]$. The event $B_s^{l_e}$ can be decomposed as follows:

$$B_s^{l_e} = B_s^{[-2l,-l]} \bigcup \left(\overline{B_s^{[-2l,-l]}} \cap B_s^{[-l_u,-2l]} \right).$$

The loss of memory and homogeneity properties of the dynamics lead to the following estimator $\hat{p}_{l_e}$:

$$\hat{p}_{l_e} := \hat{p}_l + (1 - \hat{p}_l)\, \hat{p}_{l_e - l}.$$

3.4 *Limits of the model*

The main assumptions in this work are homogeneity in time and space, induced by the constant death and birth rates, as well as constant speed and noise. While these assumptions lead to a simple model and allows a reasonably technical study, it is natural to question it. The main reason of this choice is that it corresponds to uniform laws when we have no reason prioritize one specific behavior in particular.

Note that a similar study can be made with different speeds among trajectories. This can be done by classifying the trajectories according to their speeds and applying the present procedure to each class. This would lead to the same estimation procedure with smaller datasets but theoretical results will still hold.

We then discuss the homogeneity in time, for which the most questionable assumption is the constant death rate that could possibly depend on the position or on the age of the particle. Concerning the dependence in space, this modification would lead to the estimation of a function of the position instead of the simple constant τ_d. From a practical point of view, this would increase the dimension of the parameter to estimate, with the same size of dataset. From a theoretical point of view a more technical study can be made as long as we assume the death function rate (depending on the position) constant on each tracklet $\{y\} \times]-L, 0]$ in order to overcome the issue of partial observation.

Concerning the dependence in time, the assumption that the death rate depends on the age of the particle prevents to propose a similar study. Indeed, due to partial observation, the age of each particle entering the observed region is unknown and can not be estimated.

3.5 *Modeling hypothesis and MreB dynamics*

The study of the dynamics of MreB patches or assemblies in the vicinity of the internal membrane of *Bacillus subtilis* bacteria reveals several subpopulations undergoing constrained, randomly or directionally moving [**2**]. Herein we are interested in the directionally moving subpopulation dynamics. This subpopulation moves possessively around the cell diameter [**11**, **10**]. Following Hussain *et al* [**13**], Billaudeau *et al* [**3**] confirmed that directionally moving filaments travel in a direction close to their main axis, perpendicularly to the long axis of the cell (angle $\gamma = 89.9° \pm 37.0°$). Hence, for some filaments, the speed vector may have a component in the main direction of the bacteria.

According to Wong et al. [**22**] a motion model (named "biased random walk") reproduces the dynamics patterns of MreB filaments. In their simulations, the speed is constant, the noise variance between several time steps depends on the duration and, possibly on the local curvature of the surface. These properties are shared with the Brownian motion model with constant drift we consider.

4 SIMULATION STUDY

In this section, we present a series of experiments performed on synthetic datasets. These experiments aim to evaluate and analyze the sensitivity of the reconstruction procedure when the characteristics of the dynamics as well as the spatio-temporal sampling resolution of observations vary. In addition to demonstrate the potential of our procedure, these experiments might also be useful for the design of the experimental setting for images acquisition. The reconstruction procedure has been implemented in *MATLAB ver. R2018b*. The codes are available on Github `https://github.com/atrubuil/ReconstructionOfTruncatedTrajectories`.

4.1 *Generation of trajectories*

Trajectories are generated on a rectangular unwrapped cylindrical surface of size $[0, L] \times [0, H]$ (Fig. 4). In our experiments, we set $L = 50, H = 30$. The initial position of each trajectory is drawn from uniform distribution on the surface. Time duration T between two births follows an exponential law with birth rate parameter λ. At each birth, the intrinsic properties of a trajectory i are given, such as velocity $\mathbf{v}_i$, variance $\mathbf{\Sigma}_i$, and lifetime T_d^i. The lifetime T_d follows an exponential law, with the same death rate τ_d for all trajectories in the whole simulated image sequence. The drift $\mathbf{v}_i = (v_{xi}, v_{yi})$ and noise $\mathbf{\Sigma}_i = \begin{bmatrix} \sigma_{xi} & 0 \\ 0 & \sigma_{yi} \end{bmatrix}$ are set to be constant along one given trajectory.

According to the assumptions made on real biological context, unless otherwise stated, it is set by default, $\theta = 0.01 (\approx 0.6°)$ is the angle between the direction of motion of particle and the X direction, $v_y = \tan(\theta) v_x$, $\sigma_x = \sigma_y = \sigma$, $v_x = 0.6$, $\sigma = 0.2$, $\lambda = 0.03$, $\tau_d = 0.005$.

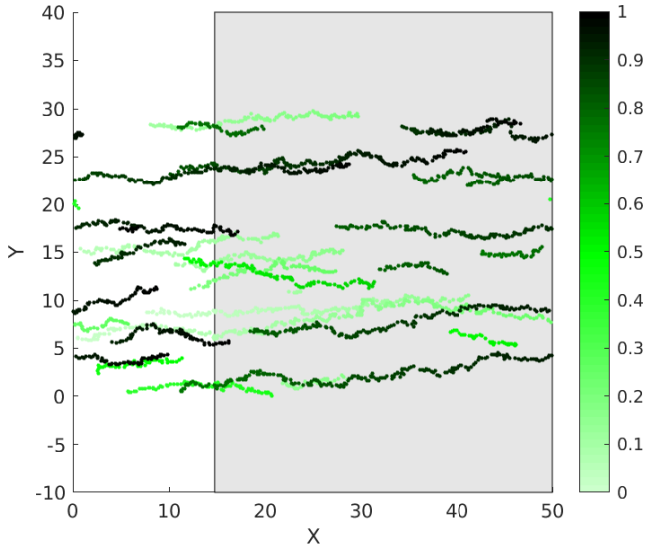

Figure 4: A set of simulated trajectories during 2.5 minutes (in stationary regime). X(resp. Y) axis represents the unfolded circumferential (resp. main) direction of the cylinder. Colors from light to dark green represent time evolution. Shadowed area corresponds to the unobserved region and white area corresponds to the ROO.

The time interval between two images $\Delta t = 0.25$. As known, the theoretical depth of the observation field of TIRFM is $200nm$, the diameter of the bacteria cell is $1\mu m$, therefore the width of the ROO l is set to 14.76 and that of the unobserved region $l_u = 35.24$ (unit in pixel, note that in TIRF images 1 pixel $\approx 64nm$).

As there is no particle on the surface at the beginning, the simulated set of trajectories needs some warm-up time to reach the stationary regime, where the law of the number of trajectories does not depend on time. The assumed dynamic process is a special case of birth and death process. As a known result[**15**], the expectation of the trajectories number N during stationary regime is $E(N) = \frac{\lambda}{\tau_d}$. To ensure that the dynamics are in a stationary regime, the images sequence is simulated long enough, for around 2 hours (Fig. 5).

4.2 *The "Adjusted Rand Index" for the evaluation of connection results*

Given the true and estimated class assignments, we compute the so called Adjusted Rand Index to evaluate similarity or consensus between the two sets. The Adjusted Rand index is the corrected-for-chance version of the Rand index. It is scored exactly 1 when the two sets are identical, close to 0 for random labeling. It could be negative when the index is lower than the expectation under random labeling. More precisely, let G and K be the true and estimated assignments respectively, let us define a and b as: a the number of pairs of elements that are in the same class in G and in the same class in K, b the number of pairs of

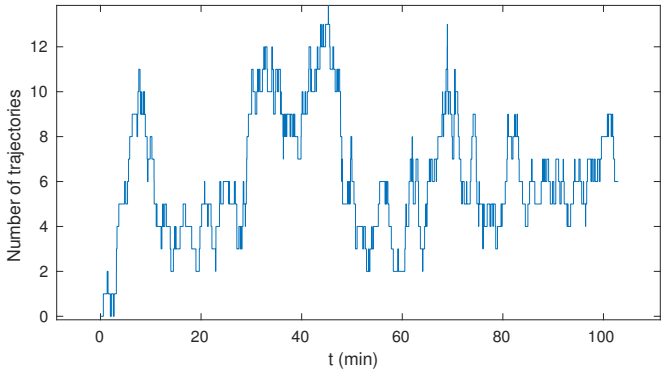

Figure 5: Fluctuations of the number of trajectories w.r.t. time. At around $t = 20$ min, the trajectories number fluctuates around the theoretical expectation value 6.

elements that are in different classes in G and in different classes in K. The raw (unadjusted) Rand index is then given by:

$$RI = \frac{a+b}{C_2^M}, \tag{21}$$

where C_2^M is the total number of possible pairs in the dataset (without ordering) of size M. The RI score does not guarantee that random assignments will get a value close to zero. This is especially true if the number of clusters has the same order of magnitude as the number of samples. To overcome this difficulty, we prefer to consider the Adjusted Rand Index defined by [**17**]:

$$ARI = \frac{RI - E(RI)}{1 - E(RI)}. \tag{22}$$

Here $E(RI)$ denotes the expectation of the Rand Index where the estimated assignment K is replaced by an assignment chosen uniformly at random. This means that the assignment procedure does not do better than random assignment if the ARI score is zero, and that it does worse than random if $ARI < 0$.

4.3 *Experimental results*

The good performance of the connection procedure relies on the estimation of the characteristics of the dynamics: the speed, $\mathbf{v}$, the diffusion variance, $\mathbf{\Sigma}$, the arrival rate τ_α and the death rate τ_d, as these quantities are used in the calculation of the likelihood (Eq. 13). Here we evaluated the impact of spatio-temporal sampling $(l/l_u, T_S)$ on the estimators and the impact of parameters of the dynamics $(\mathbf{v}, \mathbf{\Sigma}, \tau_\alpha, \tau_d)$ on the accuracy of the reconstruction.

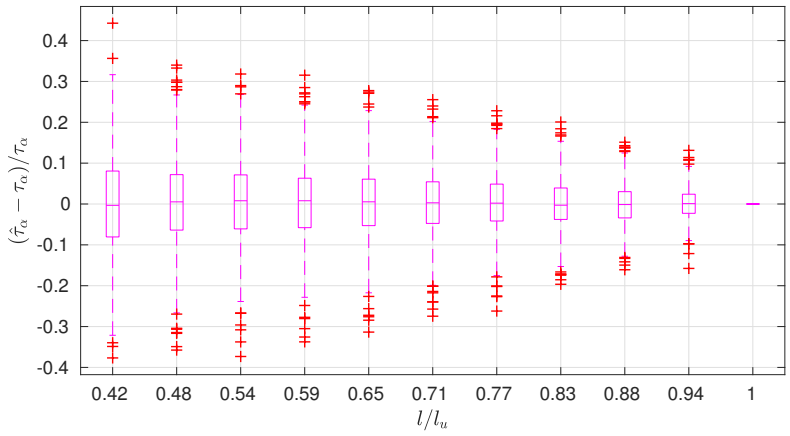

Figure 6: At $l/l_u = 0.42$, which corresponds to the realistic situation in 2D-TIRF, more than half of the trials presents a relative error smaller than 10%. The proposed estimator $\hat{\tau}_\alpha$ is unbiased and the variance decreases as l/l_u increases.

4.3.1 The estimator $\hat{\tau}_\alpha$ performs well, in the case of realistic 2D-TIRF, where $\frac{l}{l_u} \approx 0.42$

The estimator $\hat{\tau}_\alpha$ is proposed in Eq. 20. Here we test how it performs with different spatio-temporal sampling $(l/l_u, T_S)$, and different birth rate λ and death rate τ_d.

By its definition in section 3.1, τ_α, the rate of "spontaneous input" induced by particles born in the unobserved region and reach the border of the ROO, is not a preset parameter. A reference of the "true" value of τ_α is given by $\frac{N_{l_u}}{T_s}$, where N_{l_u} denotes the number of tracklets born in the region $]-l_u, 0] \times [0, H]$ and reached the border $\{0\} \times [0, H]$, l_u is the width of the unobserved region.

Next, we test the robustness of the estimator $\hat{\tau}_\alpha$ w.r.t. l/l_u (Fig. 6). To avoid the influence of T_S on the consistency of the estimator, T_S is set to be long enough as 30 min. We can conclude that, naturally, the more the observed area is larger, better is the performance of the estimator $\hat{\tau}_\alpha$. In the case of the simulation of the real situation, where $l/l_u = 0.42$, the estimator works reasonably good.

Following, we test the robustness of $\hat{\tau}_\alpha$ w.r.t. T_S (Fig. 7). This test is essential because in reality it is impossible to use a 30-min movie, because of technical issues like photobleaching of fluorophores and natual growth in living samples. At this stage, the propotion of observed and unobserved region l/l_u is set to 0.42. T_S varies from 2.5 min to 30 min. In Fig. 7, it can be noticed that the reference 'ground truth' of τ_α (blue boxes) decreases as T_S lengthens. Actually, the reference is only a pseudo 'ground truth'. It is sensible to T_S when T_S is small

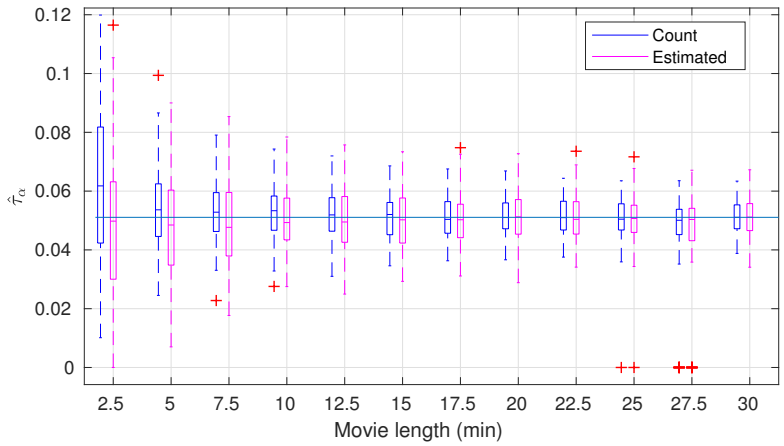

Figure 7: The counted 'ground truth' τ_α and the estimated $\hat{\tau}_\alpha$ obtained by movies of different duration, varying from 2.5 min to 30 min. Blue boxes (resp. Magenta boxes) correspond to the 'ground truth', i.e. the counted value (resp. the estimator $\hat{\tau}_\alpha$). The blue horizontal line represents the 'ground truth' value when $T_S = 30$ min.

and it converges as $T_S \to \infty$. The distributions of counted 'ground truth' and estimator become close to each other for $T_S \geq 10$ min.

The absolute value of τ_α depends on λ and τ_d. Fig. 8 displays for different combinations of λ and τ_d, the estimations of $\hat{\tau}_\alpha$ by 5-min movies (magenta) and 30-min movies (blue). It shows that τ_α increases linearly as the birth rate λ increases, and decreases slightly linearly as the death rate τ_d increases.

4.3.2 The estimator $\hat{\tau}_d$ is unbiased and performs reasonably well with 5-min movies

As explained in section 3.3, $\hat{\tau}_d$ is a rather classical estimator. Fig. 9 shows the estimator with 5-min movies (magenta) and 30-min movies (blue) respectively. It confirms that the estimator is unbiased. Black horizontal lines represent the true value of τ_d. Naturally, the variance is bigger with shorter movies.

4.3.3 The choice of T_S

According to Figs. 8 and 9, when $T_S = 30$ min, the estimators of τ_α and τ_d perform well, being converged with small variance. As 30-min movie acquisition is almost infeasible under the situation of fluorescence microscopy, we need to find a compromise with smaller T_S and

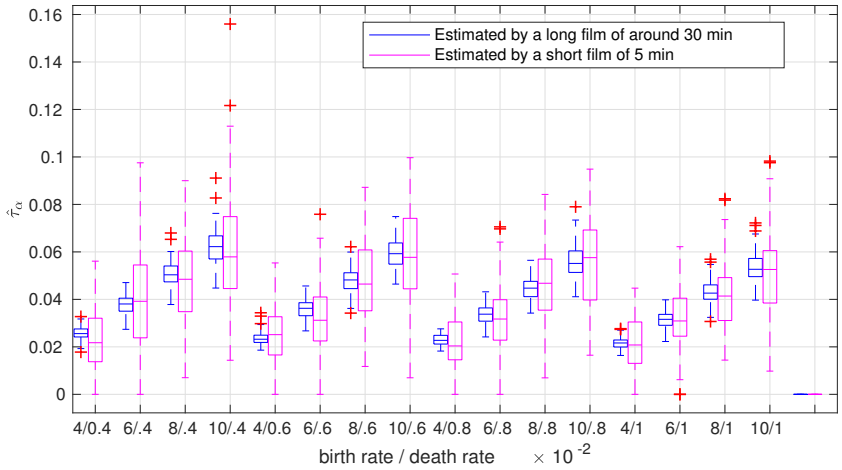

Figure 8: The estimation of arrival rate τ_α w.r.t. different λ and τ_d. For example, when $\lambda = 0.04, \tau_d = 0.004$, the median value of $\hat{\tau}_\alpha$ is around 0.025, which means that at each moment, the probability that a particle born in the invisible zone arrives at $\{-l\} \times [0, H]$ is around 0.025.

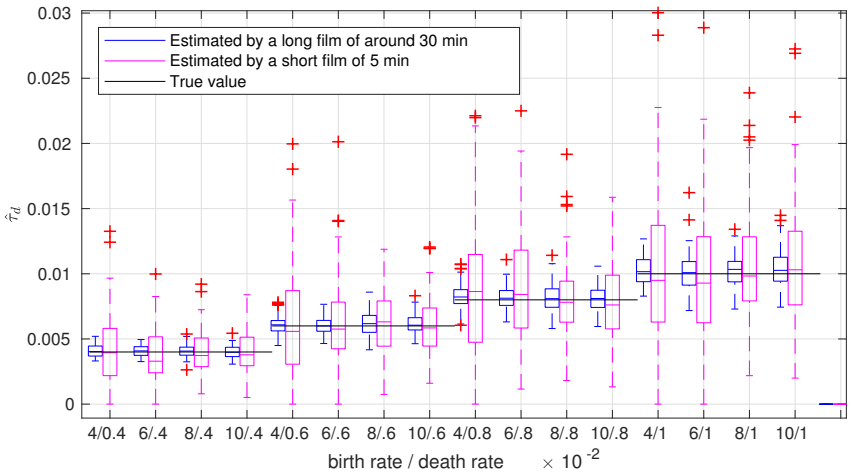

Figure 9: $\hat{\tau}_d$ with different λ and τ_d. The estimator is unbiased. The variance of the estimator is larger with shorter movies (magenta). For a given τ_d, when birth rate λ increases (e.g. the first four boxes), then the number of particles also increases, in consequence, the variance decreases.

reasonably good estimators. We tested especially T_S=2.5 min and T_S=5 min. Comparing the estimation results with 2.5-min movies, we found that $T_S = 5$ min is a good choice to limit the estimation error of τ_d and to ensure a good connection performance (more details about the experiments for the choice of T_S in Supplementary Materials 1).

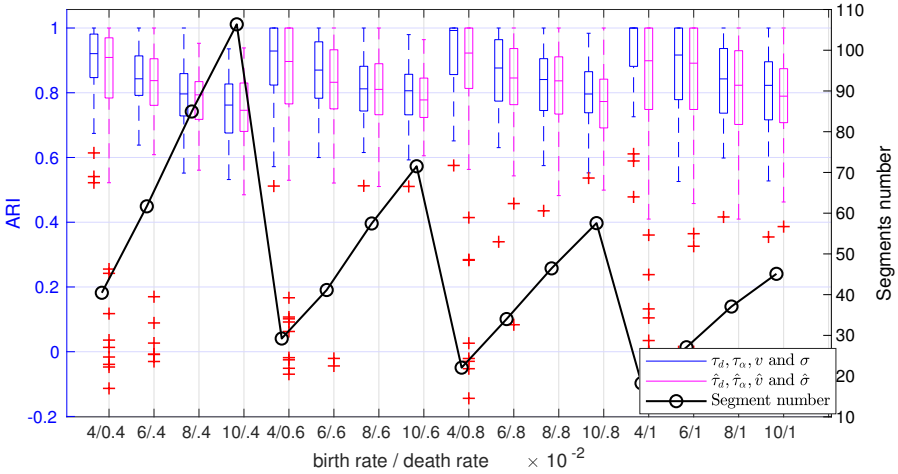

Figure 10: Connection performance comparison for different λ and τ_d, when $T_S = 5$ min. Blue (resp. magenta) boxes represents ARI values obtained with true parameters τ_d, τ_α, v and σ (resp. with estimators $\hat{\tau}_d, \hat{\tau}_\alpha, \hat{v}$ and $\hat{\sigma}$). The black line represents the mean number of tracklets in a movie.

4.3.4 The connection procedure works well, even when true parameters are unknown

In this part, we assess the performance of the connection algorithm with different parameters λ and τ_d. We evaluate as well the impact of the error of the estimator, by using in the connection procedure respectively true parameters $\tau_\alpha, \tau_d, v, \sigma$ and their estimators $\hat{\tau}_\alpha, \hat{\tau}_d, \hat{v}, \hat{\sigma}$. The duration of movies T_S is set to 5 min. The connection results measured by ARI are presented in Fig. 10.

Each pair of blue and magenta box represents the connection result of a setting of λ and τ_d. The black line represents the mean value of the number of tracklets fluctuating with different settings of λ and τ_d. The performance of connection is affected by the number of tracklets in each movie to be connected. The higher the density of tracklets is, the more difficult it is to find the right ones.

It can be noticed that the ARI value when we use the estimators $\hat{\tau}_d, \hat{\tau}_\alpha, \hat{v}$ and $\hat{\sigma}$, is almost as good as when we use true values for all the parameters. This is an encouraging result as it means that it is feasible to apply the algorithm in real image sequences. When the number of tracklets is around 20 (e.g. $\lambda = 0.04$ and $\tau_d = 0.008$), the median values of ARI are higher than 0.9, showing a promising connection performance. Even for the case with the highest particle density, when the average number of tracklets reaches 100 ($\lambda = 0.1$ and $\tau_d = 0.004$), the median value of ARI is still higher than 0.7.

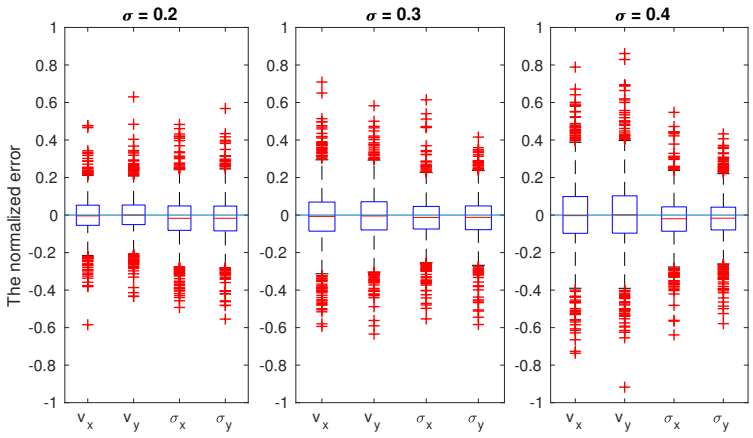

Figure 11: The NEs of v_x, v_y, σ_x and σ_y in cases where $\sigma = 0.2, 0.3$ and 0.4.

4.3.5 The connection procedure is robust even when each particle moves at different speed (but with constant speed along a trajectory). However v and σ should be well estimated

In the previous experiments, all trajectories are generated with the same speed v and standard error σ. In this section, we design experiments to test the performance of the connection algorithm, when the drift v varies from particle to particle, $v_x \sim \mathcal{U}(0.5, 0.9)$. In one movie, as all particles are in the same environment, there is no obvious reason for different particles to have different σ . Therefore the standard error σ is set to be constant for particles in one movie. However, we test in independent movies, when $\sigma = 0.2, 0.3$ or 0.4, the influence of σ on the performance of connection procedure. Other parameters to be specified are the angle between the direction of the motion and the circumferential direction of the cylinder, $\theta = 0.15(\approx 8.6°)$, $v_y = \tan(\theta)v_x$, $\sigma_y = \tan(\theta)\sigma_x$, $\sigma_x = \sigma$, birth rate and death rate are fixed, with $\lambda = 0.08$ and $\tau_d = 0.02$.

The normalized error (NE) of an estimator is defined by the error of the estimator normalized by its ground truth. For example, the NE of v_x equals to $\frac{v_x - \hat{v}_x}{v_x}$. In Fig. 11, the NEs of $\hat{v}_x, \hat{v}_y, \hat{\sigma}_x$ and $\hat{\sigma}_y$ when σ takes different values are presented. It shows that when σ increases, the variance of $\hat{v}_x$ and $\hat{v}_y$ increases.

For tracklets connection, we compare the results when true values of v and σ or when the estimated value $\hat{v}$ and $\hat{\sigma}$ respectively are taken by the connection procedure. The experiments are carried under three situations, when $\sigma = 0.2, 0.3$, and 0.4. The results in Fig. 12 shows that whether using true v and σ or estimated value $\hat{v}$ and $\hat{\sigma}$, the performance measured by ARI degrades when σ increases. When the standard error $\sigma = 0.2$, using true v and σ, the median value of ARI reaches to 1. When using the estimated $\hat{v}$ and $\hat{\sigma}$, the median value

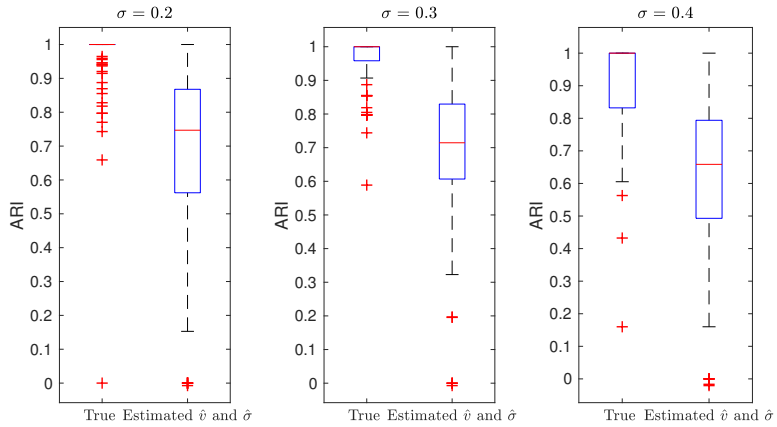

Figure 12: Comparison of connection performance measured by ARI, when the procedure takes true v and σ or estimated $\hat{v}$ and $\hat{\sigma}$, in three experiments where $\sigma = 0.2, 0.3$ and 0.4 respectively.

of ARI is approximately 0.75. It can be concluded that the estimation of v and σ has an impact on the performance of the algorithm.

4.4 *Analysis of the connection results*

4.4.1 An example of tracklets connection

Figure 13 shows, on the left, trajectories in a movie and on the right, the results of tracklets connection. The path from an output to the matched input is represented by the dashed straight line, as we don't know how exactly the particle went through the hidden zone. The only wrong connection corresponds to the bold line. Compared with the figure on the left, we can find the realization of these two tracklets. In reality, the orange bold tracklet disappeared at the hidden region and the bold purple tracklet appeared nearby and entered into the observed zone.

In fact, not only the optimal configuration can be calculated, but also the most likely alternative configurations in decreasing order of probability (Fig. 14). It should be noticed that the optimization algorithm tries to minimize $K(c) = -\log Q(c)$, instead of finding the c^* maximizing $P(c)$. It costs too much to obtain the probability $P(c^*)$, as it requires the enumeration of all the possible configurations $c \in \mathcal{C}$ (Eq. 11). However, The number of configurations can be determined in order to guarantee that the sum of the probability of these configurations will be greater than a given threshold (see Supplementary Materials 3). As a result, we can obtain lower and upper bounds for the probability.

For this example, we see that the second most likely configuration corresponds to the

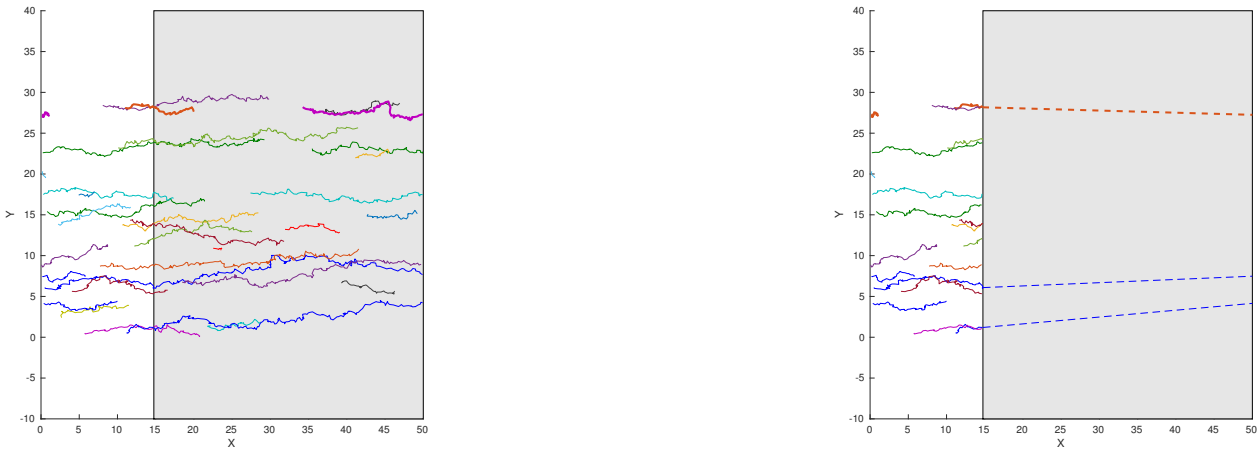

Figure 13: Left: Trajectories in one movie; Right: the connection results.

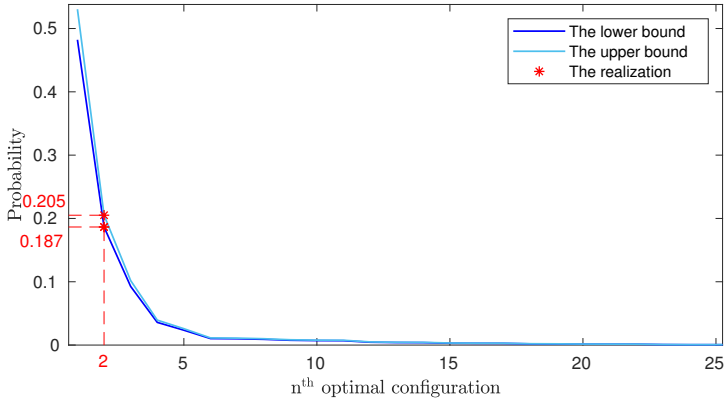

Figure 14: The probability of n^{th} optimal configuration and the probability of the realization.

realization of trajectories, $0.187 < P(c^t) < 0.205$ according to the algorithm (Fig. 14). Combining with Fig. 13, the optimal configuration found by the algorithm, committing one connection error, does not correspond to the realization. In section 2, we evaluated the connection error caused by randomness.

4.4.2 The number of rotations around the cylinder

Once the connection procedure is achieved, we can address the question of the number of rotations of a particle around the cylinder. In the context of simulation, the death rate τ_d and the dynamic velocity v_x are known. Accordingly, the value of the number of rotations is known to be equal to $\frac{v_x T_d}{L}$, where $T_d \sim \mathcal{E}(\tau_d)$ ensures a theoretical expectation value of $\frac{v_x}{\tau_d L}$. By counting the tracklets for each trajectory, we can obtain a proxy of the number of

rotations around the cylinder.

In Fig. 15, λ is set to 0.04 and different values of τ_d between 0.004 and 0.01 are evaluated. Blue bars represent the distribution of the number of rotations of true connections. The magenta bars to display the distribution of the number of rotations estimated by the connection procedure. The corresponding ARI values, indicating the connection accuracy, are given as well. The density of the theoretical values of the number of rotations is presented in green color. The vertical lines represent the median values of the corresponding distribution. Overall, when τ_d is small, the median value of number of rotations is higher and the distribution has a heavier tail. In general, the distributions with all three colors are similar to each other.

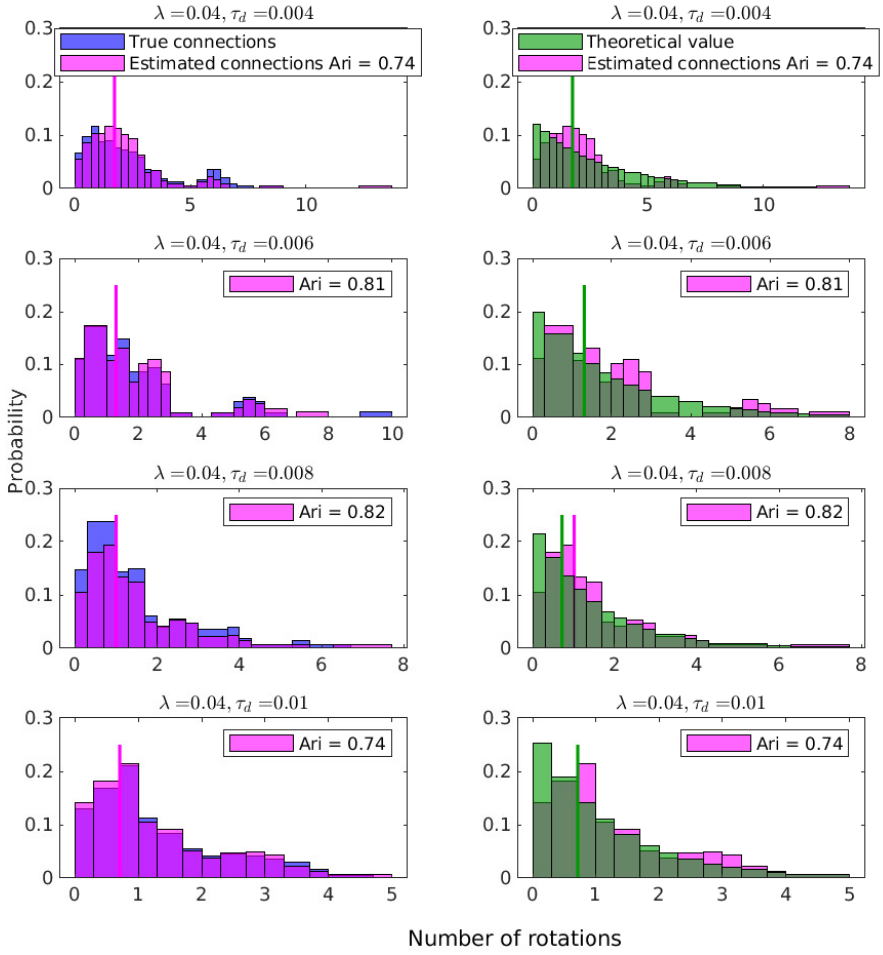

Figure 15: The density distribution of 'number of rotations'. Each row represents the result for a different τ_d value.

CONCLUSION

In this paper, we proposed a probabilistic framework and a computational approach with no hidden parameter to connect tracklets from 2D partial observations. We provided several consistent estimators of parameters to automatically drive the connection procedure. The performance of our procedure is satisfying if we consider the ARI criterion. Moreover, an ordered set of the best reconstructions could also be proposed. The robustness of the procedure has been tested for different drifts, diffusion of the dynamics, and trajectory densities. Our computational approach can be extended to the case when the drift/speed is not the same for all particles but remains constant along time. In that case, it is straightforward to estimate and classify the drifts before applying our connection procedure to each class of drift since the tracklets with different speeds are not likely to be connected.

After recovering the whole trajectory on the surface of the cylinder, we can have a better understanding of the average duration of a particle, and more accurate statistics about the spatio-temporal organization of particles. The simulation study can also serve as a guideline for the design of experiments.

The connection procedure is tested with a real TIRFM dataset. The experimental results are illustrated in Appendix A. For future works, we plan to investigate more on real TIRFM datasets. Experiments on real data show that the observed region corresponds approximately to one-third of the total surface, which is rather small. However, we have shown that we are able to cope with the hidden region of such size. Nevertheless, several assumptions and approximations need to be further investigated. For instance, we assumed spatial homogeneity, suggesting that the particles are born or die uniformly on the membrane surface. Moreover, we assumed a memoryless lifetime while dependency with respect to particle "age" could be more realistic.

A AN ILLUSTRATION OF THE CONNECTION ALGORITHM APPLIED TO REAL MREB DYNAMICS

Data obtained using TIRF microscopy of MreB aggregates in *Bacillus subtilis* ([**2**]) are considered. A typical movie from this dataset shows several MreB aggregates moving inside one or several cells (see Fig. 1 and Supplementary Materials 3). The pixel size, frame rate and duration are respectively $\Delta x = \Delta y = 64nm$, $\Delta t = 1s$, $T = 2mn$. Hereafter, we selected one cell to illustrate the application of our algorithm. First, tracklets exhibiting directed motion should be extracted from the movie data, then tracklets should be projected back on the cylinder shape of the cell and unwrapped, eventually the connection algorithm is applied and a list of likelihood decreasing ordered configurations of trajectories connections is presented to the user.

A.1 *Construction of the local cell referential*

Once MreB aggregates pixels are separated from the background inside each image of the movie, a bounding box is drawn around a given cell and a local **x-y** referential is estimated using Principal Component Analysis (PCA) on the coordinates of pixels belonging to aggregates (Fig. 16). The **z** coordinate of an aggregate is inferred using as a prior the cylinder shape of the bacteria and its radius, R, so $z(x, y) = R - \sqrt{R^2 - x^2}$.

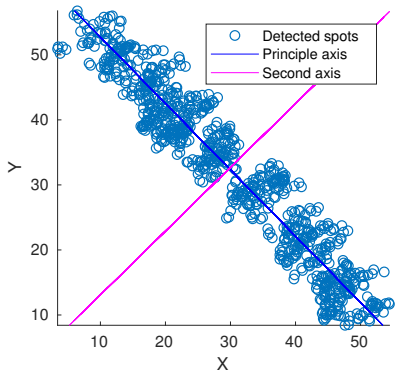

Figure 16: The estimation of the local $x - y$ referential for a cell.

A.2 *Tracking and selection of aggregates in the observed region*

Using U-Track [**14**], MreB aggregates are tracked and constitute a set of tracklets. The automatic classification of these tracklets in three classes, respectively Brownian, subdiffusive and, directed motion is done using two algorithms: the classical MSD algorithm and a recent algorithm ([**7**]). The tracklets classified as directed motion by either one of the two algorithms are selected for the application of our connection algorithm (Fig. 17). The tracklets were projected back on the cylinder and unwrapped, as explained in the technical part of the paper. As we can see, only a few aggregates crossed the borders of the visible region. Others aggregates, according to our definitions are born or die in the visible region, which is not true. When an aggregate approaches the borders, its intensity becomes weak as it is farther from the support plane, and less excitation light penetrates higher z-position in TIRF microscopy settings. As a result, the detection algorithm fails to detect the aggregates when they approach the borders.

A.3 *The connection of tracklets*

All the selected tracklets crossing the magenta lines in Fig. 17 are considered.

First, the speed and diffusion are estimated (Fig. 18) for each tracklet, respectively. Two populations of tracklets evolving in opposite x directions are identified. These two populations are considered one after the other in the connection procedure. Tracklets corresponding to speed lower than 0.4 are filtered out.

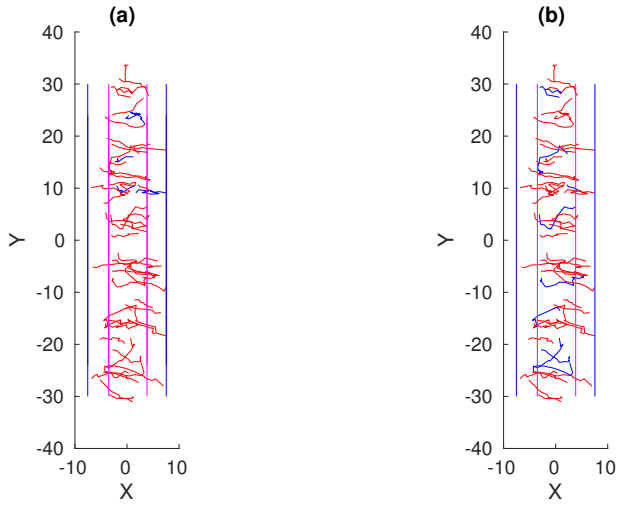

Figure 17: The tracklets classification. (a) MSD classification. (b) STP classification. Brownian tracklets (blue), Directed tracklets (red). Blue lines represent the border of the visible region. Magenta lines represent the 0.1 quantile and the 0.9 quantile of x coordinate values.

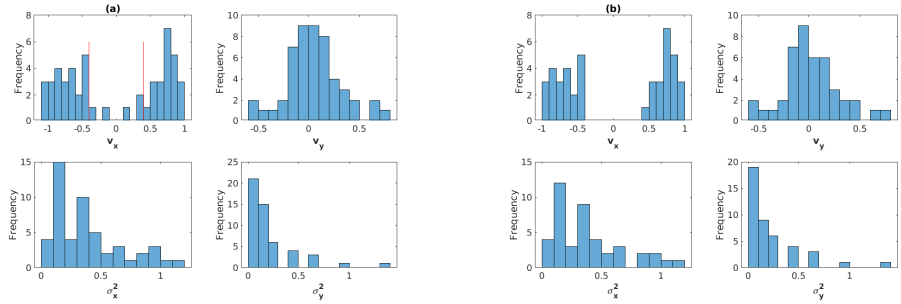

Figure 18: The distribution of drift and variance in the selected tracklets population. (a) without filtering. (b) after filtering.

For the population of tracklets associated with positive (resp. negative) v_x, death rate τ_d is estimated as 0.0691 (resp. 0.0756). The arrival rate τ_α is estimated as 0.0310 (resp. 0.0220).

† **tracklets of positive speed** v_x

The first, fifth, seventh and eighth optimal configuration suggests one connection. The second suggests that there is no connection. The third, fourth and sixth configurations suggest two connections. Some of these configurations are shown in Fig. 19.

† **tracklets of negative speed** v_x

The first optimal configuration suggests no connection. The second one suggests one

connection Fig. 20.

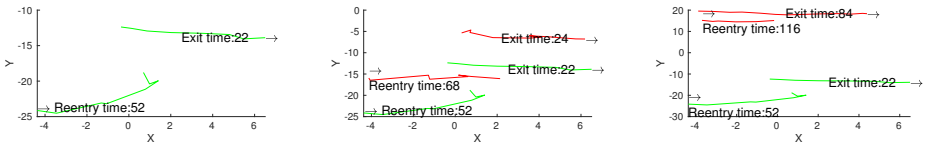

Figure 19: Some optimal configurations for the population of positive speed v_x. From left to right, first, third and sixth better configurations. Connected tracklets are drawn with the same color.

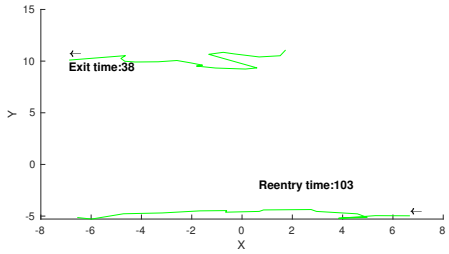

Figure 20: Second optimal configuration for the population of negative speed v_x.

In Fig. 21 we show a 3D reconstruction of the aggregates and two tracks that could correspond to aggregates doing more than one loop around the cylinder surface of the cell. For the positive (resp. negative) speed set of tracklets, the eighth (resp. second) optimal solution was selected.

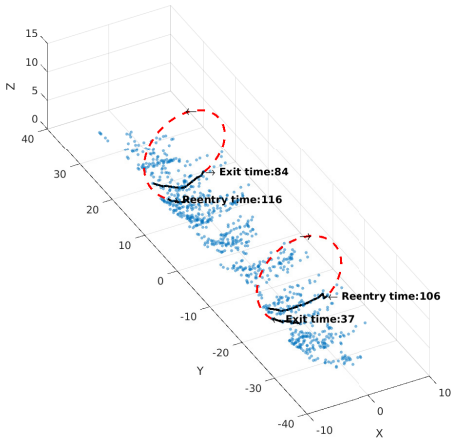

Figure 21: The 3D reconstruction of tracks. The centroids of aggregates are represented as blue squares.The arrows indicates the direction of motion. The full blue lines represent tracklets crossing the observed region. The dotted red lines represent the simplified extrapolation of aggregate motion in the invisible region.

ACKNOWLEDGEMENTS

The authors thank L. Tournier for fruitful discussions about the optimization procedure, R. Carballido-López and C. Billaudeau for their inspiring work in MreB studies which triggered this research. This work was partially supported by ANR DALLISH, Programme CES232016.

REFERENCES

[1] D. Axelrod, T. P. Burghardt, and N. L. Thompson. Total internal reflection fluorescence. *Annual review of biophysics and bioengineering*, 13(1):247–268, 1984.

[2] C. Billaudeau, A. Chastanet, Z. Yao, C. Cornilleau, N. Mirouze, V. Fromion, and R. Carballido-López. Contrasting mechanisms of growth in two model rod-shaped bacteria. *Nature communications*, 8:15370, 2017.

[3] C. Billaudeau, Z. Yao, C. Cornilleau, R. Carballido-López, and A. Chastanet. Mreb forms subdiffraction nanofilaments during active growth in bacillus subtilis. *mBio*, 10(1):e01879–18, 2019.

[4] H. A. Blom and Y. Bar-Shalom. The interacting multiple model algorithm for systems with markovian switching coefficients. *IEEE transactions on Automatic Control*, 33(8):780–783, 1988.

[5] J. Boulanger, C. Gueudry, D. Münch, B. Cinquin, P. Paul-Gilloteaux, S. Bardin, C. Guérin, F. Senger, L. Blanchoin, and J. Salamero. Fast high-resolution 3d total internal reflection fluorescence microscopy by incidence angle scanning and azimuthal averaging. *Proceedings of the National Academy of Sciences*, 111(48):17164–17169, 2014.

[6] P. C. Bressloff and J. M. Newby. Stochastic models of intracellular transport. *Reviews of Modern Physics*, 85(1):135, 2013.

[7] V. Briane, C. Kervrann, and M. Vimond. Statistical analysis of particle trajectories in living cells. *Physical Review E*, 97(6):062121, 2018.

[8] N. Chenouard, I. Bloch, and J.-C. Olivo-Marin. Multiple hypothesis tracking for cluttered biological image sequences. *IEEE transactions on pattern analysis and machine intelligence*, 35(11):2736–3750, 2013.

[9] C. Cornilleau, A. Chastanet, C. Billaudeau, and R. Carballido-López. Methods for studying membrane-associated bacterial cytoskeleton proteins in vivo by tirf microscopy. In *Cytoskeleton Dynamics*, pages 123–133. Springer, 2020.

[10] J. Domínguez-Escobar, A. Chastanet, A. H. Crevenna, V. Fromion, R. Wedlich-Söldner, and R. Carballido-López. Processive movement of mreb-associated cell wall biosynthetic complexes in bacteria. *Science*, 333(6039):225–228, 2011.

[11] E. C. Garner, R. Bernard, W. Wang, X. Zhuang, D. Z. Rudner, and T. Mitchison. Coupled, circumferential motions of the cell wall synthesis machinery and mreb filaments in b. subtilis. *Science*, 333(6039):222–225, 2011.

[12] A. Genovesio, T. Liedl, V. Emiliani, W. J. Parak, M. Coppey-Moisan, and J.-C. Olivo-Marin. Multiple particle tracking in 3-d+ t microscopy: method and application to the tracking of endocytosed quantum dots. *IEEE Transactions on Image Processing*, 15(5):1062–1070, 2006.

[13] S. Hussain, C. N. Wivagg, P. Szwedziak, F. Wong, K. Schaefer, T. Izore, L. D. Renner, M. J. Holmes, Y. Sun, A. W. Bisson-Filho, et al. Mreb filaments align along greatest principal membrane curvature to orient cell wall synthesis. *Elife*, 7:e32471, 2018.

[14] 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*, 5(8):695, 2008.

[15] S. Karlin. *A first course in stochastic processes*. Academic press, 2014.

[16] N. Li and X. R. Li. Target perceivability and its applications. *IEEE Transactions on Signal Processing*, 49(11):2588–2604, 2001.

[17] W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.

[18] E. Schrodinger. Zur theorie der fall-und steigversuche an teilchen mit brownscher bewegung. *Physikalische Zeitschrift*, 16:289–295, 1915.

[19] M. C. Tweedie. Inverse statistical variates. *Nature*, 155(3937):453, 1945.

[20] S. van Teeffelen and L. D. Renner. Recent advances in understanding how rod-like bacteria stably maintain their cell shapes. *F1000Research*, 7, 2018.

[21] A. Wald. Sequential analysis. 1973.

[22] F. Wong, E. C. Garner, and A. Amir. Mechanics and dynamics of translocating mreb filaments on curved membranes. *eLife*, 8:e40472, 2019.