AN ADAPTIVE STATISTICAL TEST TO DETECT NON BROWNIAN DIFFUSION FROM PARTICLE TRAJECTORIES

V. Briane[⋆†]*, M. Vimond*[†]*, C. Kervrann*[⋆]

[⋆] Inria, Centre Rennes – Bretagne Atlantique, Campus Universitaire de Beaulieu, 35042 Rennes Cedex, France
[†] CREST-ENSAI, ENSAI, Ker-Lann, 35172 Bruz, France

ABSTRACT

Assessing the dynamics of particles inside live cell is of paramount interest to understand cell mechanisms. In this paper, we assume that the motions of particles follow a certain class of random process: the diffusion processes. Our contribution is to propose a statistical method able to classify the motion of the observed trajectories into three groups: confined, directed and free diffusion (namely Brownian motion). This method is an alternative to Mean Square Displacement (MSD) analysis. We assess our procedure on both simulations and real cases.

***Index Terms*—** statistical test, diffusion, Brownian motion, mean square displacement, trajectory classification.

1. INTRODUCTION

A cell is composed of organelles interacting with each other. They exchange biological material, such as proteins, directly in the cytosol or via transport vesicles moving along the cytoskeleton. The study of these moving particles/objects is of main interest. As the interior of a living cell is a fluctuacting environment, we model the trajectories of particle with diffusion processes, a class of continuous time stochastic processes with continuous paths. We are particularly interested in classifying the motions into three distinct types of diffusion:

1. Directed diffusion: the particle is transported actively via molecular motors along the cytoskeleton [1, Section 4].
2. Free diffusion (or Brownian motion): the particle evolves freely inside the cytosol [1, Section 2].
3. Confined diffusion: the particle is confined in a domain or evolves in an open but crowded area [2, 1, Section 3].

Traditionally, the Mean Square Displacement $\mathrm{MSD}(t) = \mathbb{E}(\|X_t - X_0\|_2^2)$ ($\mathbb{E}(.)$ denoting the expectation, $\|.\|_2$ the Euclidean norm and t the time) is used to classify the aforementioned motions. The MSD is estimated and fitted to $t \to t^\beta$ with a least-square method. The type of motion is determined according to different rules based on the values of $\beta > 0$ [3, 4]. Unfortunately, as t increases the variance of $\widehat{\mathrm{MSD}}(t)$, the estimator of $\mathrm{MSD}(t)$, increases [5, 6]. Then $\widehat{\mathrm{MSD}}(t)$ is reliable only for small t. Consequently it becomes difficult to estimate accurately β and then to classify properly the type of motion.We propose here a systematic statistical procedure in order to determine to which type of diffusion the observed trajectory fits the best. This procedure is flexible and adapts to the length of the observed trajectory. This adaptive property allows this test to perform well even on short trajectories contrary to the MSD approach. The remainder of this paper is organized as follows: in Section 2, we set some assumptions about 2D diffusions processes necessary for the validity of our statistical procedure. In Section 3, we describe the test procedure. In Section 4, we present results on both simulation and real data.

2. STOCHASTIC DIFFERENTIAL EQUATION

A diffusion process is a continuous time stochastic process with continuous paths and which has the strong Markov property. Diffusion processes can be seen as the solutions of Stochastic Differential Equations (SDE) [7, Chapter 15, Section 14]. Heuristically, a SDE models the motion of a particle in a fluid submitted to a deterministic force due to the fluid and a random force due to random collisions with others particles. In this paper, we focus on the 2D case. Then, in 2D, the displacement of the particle between t and $t + \Delta t$ is:

$$X_{t+\Delta t} - X_t \approx \mu(X_t, t)\Delta t + \sigma(X_t, t)(B_{t+\Delta t} - B_t), \quad (1)$$

where $\mu(x, t)$ is a vector of size 2 called the drift and models the deterministic force, $\sigma(x, t)$ is a 2×2 matrix called the diffusion coefficient and models the random force and finally $B_t = \left(B_t^1, B_t^2\right)^\top$ is a 2D Brownian motion (composed of 2 independent Brownian motions). The Markov property comes from the fact that the increments of the Brownian are independent. We define a SDE as (1) when $\Delta t \to 0$ and write it in infinitesimal notations as:

$$dX_t = \mu(X_t, t)dt + \sigma(X_t, t)dB_t. \quad (2)$$

We suppose that the diffusion coefficient is the diagonal matrix $\sigma \mathbf{I}_2$ which reflects the isotropy of the random force. In the literature, they also define the diffusion coefficient D as $\sigma = 2D$. In Section 4, we present some examples of diffusion illustrating the confined and directed diffusions.

3. STATISTICAL PROCEDURE

Our procedure can be related to a statistical test with the Brownian motion as the null hypothesis and the confined and directed diffusion as the alternative hypothesis. Formally we would write the test as:

$$H_0 : X_t = \sigma B_t \quad \text{vs} \quad H_1 : (X_t)_{t>0} \text{ a } \begin{cases} \text{confined} \\ \text{directed} \end{cases} \text{ diffusion.} \quad (3)$$

3.1. An intuitive test statistic

2D Brownian motion is recurrent: it will come back to a neighbourhood, however small, of any point infinitely often. [8, Remark 3.8]. That is why we call it free diffusion. It is neither the case for the confined diffusion (the particle being trapped in small area), nor for the directed diffusion (the particle is driven by a motor in a certain direction). Then an intuitive statistic (or measure) to distinguish a confined/directed diffusion from a Brownian motion is:

$$S_X(t) = \sup_{0 \leq s \leq t} \|X_s - X_0\|_2. \quad (4)$$

This statistic allows to answer the question : how far from its initial position did the process go during the period of length t? If $S_X(t)$ is low, it means the process stayed close to its initial position whereas, if $S_X(t)$ is high, it went far from the initial position during the period $[0\,,\,t]$. Now we choose the Brownian motion of diffusion coefficient $\sigma \mathbf{I}_2$ (namely σB_t) as our process of reference. We define a region of high probability for $S_{\sigma B}(t)$:

$$\left\{ q_{\sigma,t}\Big(\frac{\alpha}{2}\Big) \leq S_{\sigma B}(t) \leq q_{\sigma,t}\Big(1 - \frac{\alpha}{2}\Big) \right\}, \quad (5)$$

where $q_{\sigma,t}(x)$ is the quantile of order x of the random variable $S_{\sigma B}(t)$. Therefore $S_{\sigma B}(t)$ has probability $1 - \alpha$ to be in the region (5). In the framework of the statistical test, α is called the level of the test, usually set to $\alpha = 0.05$. Then we classify the motions according to the decision rule:

- if $S_X(t)$ is in the interval defined in (5) we state that $(X_t)_{t>0}$ is a Brownian motion.
- if $S_X(t) < q_{\sigma,t}\big(\frac{\alpha}{2}\big)$, we state that $(X_t)_{t>0}$ is a confined diffusion.
- if $S_X(t) > q_{\sigma,t}\big(1 - \frac{\alpha}{2}\big)$, we state that $(X_t)_{t>0}$ is a directed diffusion.

We present now how to use the procedure on real data.

3.2. Continuous case

We suppose that we observe $(X_t)_{t>0}$, solution of (2) (with $\sigma(x,t) = \sigma \mathbf{I}_2$), on the continuous interval of time $[0\,,\,t]$. That is, we observe a continuous curve (or path or trajectory) on $[0\,,\,t]$. In this case, we know exactly σ [9, Lemma 4.2, p 212]. There exists an analytical form for the cumulative distribution of $S_{\sigma B}(t)$ noted $x \to F_{S_{\sigma B}(t)}(x)$ [10, Formulae.1.1.4, p. 280]. Then in order to implement our procedure, we just need to compute (at least numerically) $q_{\sigma,t}(x) = F^{-1}_{S_{\sigma B}(t)}(x)$. Now, it would be desirable to have a test statistic whose distribution does depend neither on σ nor on t. It will be useful for the implementation in the discrete case. With the remark $S_{\sigma B}(t) = \sigma S_B(t)$ and thanks to the fractal property of Brownian motion (i.e Brownian motion exhibits the same pattern at different time scale), we have:

$$F_{S_{\sigma B(t)}}(x) = F_{S_{B(1)}}\left(\frac{x}{\sigma\sqrt{t}}\right). \quad (6)$$

This result is straightforward in the continuous case, as we know the analytical form of $x \to F_{S_{B(t)}}(x)$. Interestingly we can extend this result to the discrete case, as the fractal property holds also in discrete time. To sum up, now our region of high probability defining the decision rule for the classification of motions is (using the notations of (5)):

$$\left\{ q_{1,1}\Big(\frac{\alpha}{2}\Big) \leq \frac{S_{\sigma B}(t)}{\sigma\sqrt{t}} \leq q_{1,1}\Big(1 - \frac{\alpha}{2}\Big) \right\}. \quad (7)$$

Therefore the statistic of interest becomes:

$$S_X(t)/(\sigma\sqrt{t}). \quad (8)$$

As previously, we compare (8) to the boundaries of (7) to determine which kind of motions fits the best to the process $(X_t)_{t>0}$, observed on a finite continuous interval of time.

3.3. Discrete case

Now, we suppose that we observe $(X_t)_{t>0}$ at discrete times equidistant of Δt, the resolution time of the sensor. Therefore we observe $(X_0, X_{\Delta t}, \ldots, X_{n\Delta t})$ a trajectory of length $n+1$. In discrete time $S_X(t)$ turns into:

$$S^n_X(n\Delta t) = \max_{i=0,\ldots,n} \|X_{i\Delta t} - X_0\|_2. \quad (9)$$

The statistic of interest in discrete times is analogous to (8). However, in discrete time, we can no longer assume that σ is known. With the assumption that $\sigma(x,t) = \sigma \mathbf{I}_2$ made in Section 2, we can estimate σ by:

$$\hat{\sigma} = \left[\frac{1}{2n\Delta t} \sum_{j=1}^{n} \|X_{j\Delta t} - X_{(j-1)\Delta t}\|_2^2 \right]^{1/2}. \quad (10)$$

According to [9, Lemma 4.2, p 212], this estimator converges almost surely to σ under both hypothesises H_0 and H_1. The statistic of interest is then:

$$\frac{S^n_X(n\Delta t)}{\sqrt{n\Delta t}\hat{\sigma}} = \frac{S^n_X(n\Delta t)}{\left[\frac{1}{2} \sum\limits_{j=1}^{n} \left\| X_{j\Delta t} - X_{(j-1)\Delta t} \right\|_2^2 \right]^{1/2}}. \quad (11)$$

Under H_0, the test statistic (11) converges in distribution to $S_{\sigma B}(t)/\big(\sigma\sqrt{t}\big)$ as $n \to \infty$. It validates the procedure described in Subsection 3.2. However, as in practice n can be small, the quantiles defining the region of high probability (7) in the continuous case, no longer defines a region of high probability for the test statistic (11). Therefore we replace them by $q^n_{1,1}(x)$ ($x = 1 - \alpha/2, \alpha/2$), the quantiles of the test statistic (11). We estimate them with algorithm 1. The depen-

Data: n, α
Result: $\hat{q}^n_{1,1}\left(\frac{\alpha}{2}\right)$, $\hat{q}^n_{1,1}\left(1-\frac{\alpha}{2}\right)$
for *i=1* **to** *N* **do**
  initialization $Y^i_0 = (0,0)^\top$;
  for *j=1* **to** *n-1* **do**
    Draw $\epsilon \sim \mathcal{N}(0\,, \mathbf{I_2})$;
    $Y^i_j = Y^i_{j-1} + \frac{1}{\sqrt{n}}\epsilon$;
  end
  Compute $S_i = \max_{j=1,\dots,n} \|Y^i_j\|_2$ and $\hat{\sigma}_i$ (10);
  Compute the ratio $R_i = \frac{S_i}{\hat{\sigma}_i}$;
end
Sort $\mathbf{R} = (R_1, \dots, R_N)^\top$ in increasing order ;
Set $\hat{q}^n_{1,1}\left(\frac{\alpha}{2}\right) = R_{\lfloor \frac{\alpha}{2} N \rfloor}$, $\hat{q}^n_{1,1}\left(1-\frac{\alpha}{2}\right) = R_{\lfloor (1-\frac{\alpha}{2}) N \rfloor}$;

Algorithm 1: Estimation of the quantiles by Monte Carlo simulations. $\left(Y^i_0, \dots, Y^i_n\right)$ has the same distribution as $\left(B_0, B_{1/n}, B_{2/n}, \dots, B_1\right)$ so $S_i \overset{d}{\sim} S^n_{1\times B}(1)$.

Estimated quantiles	Trajectory size			
quantile order	10	30	100	asymp
2.5%	0.725	0.754	0.785	0.834
97.5%	2.626	2.794	2.873	2.940

Table 1: Estimation of the quantiles of order $\alpha/2$ and $1-\alpha/2$ ($\alpha = 5\%$) for different size trajectory n, using algorithm 1 with $N = 1\,000\,001$. Estimations are accurate at ± 0.001.

dence on n of the new boundaries $q^n_{1,1}(x)$ ($x = 1-\alpha/2, \alpha/2$) of the high probability region shows that our test adapts to the trajectory size . In Table 1, we see there is a significant difference between asymptotic and non asymptotic quantiles. As expected, as $n \to \infty$ $q^n_{1,1}(x)$ converges to $q_{1,1}(x)$.

4. EXPERIMENTAL RESULTS

4.1. Monte Carlo study of the procedure

We assess our test procedure using two parametric diffusion processes as the two possible outcomes of the alternative hypothesis H_1 described in (3). For illustrating directed diffusion, we use the directed Brownian:

$$dX^i_t = v_i dt + \sigma dB^i_t \quad i = 1, 2, \tag{12}$$

where v_i is the constant drift parameter. For illustrating confined diffusion, we use the Ornstein-Uhlenbeck process:

$$dX^i_t = -\lambda(X_t - \theta_i)dt + \sigma dB_t \quad i = 1, 2, \tag{13}$$

with $\lambda > 0$ reflecting a restoring force directed towards the long term average $\theta = (\theta_1, \theta_2)^\top$. The test procedure discriminates well the three kind of diffusion if the parameters of the processes (12) and (13) belongs to a certain range. We derive approximately these ranges of value in (14) and (15). We detect the Ornstein-Uhlenbeck process as effectively confined diffusion if:

$$\lambda > \frac{q^{\chi^2(2)}_{1-\alpha_1}}{2K^2 n\Delta t \left(\hat{q}^n_{1,1}\left(\frac{\alpha}{2}\right)\right)^2}. \tag{14}$$

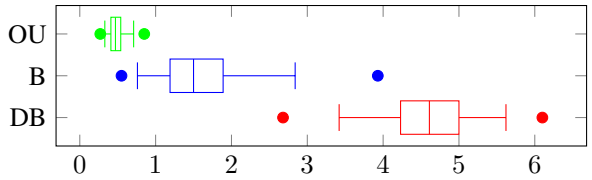

Fig. 1: Box-plots of the test statistic (11) for the processes of interest. We choose $n = 30$. For the Ornstein-Uhlenbeck process, we define λ with (14) setting $K = 0.5$ and $\alpha_1 = 0.05$. For the directed Brownian, we take $v_1 = v_2 = v$ and define v with (15) setting $A = 3$. For all processes, we take $\sigma = 2$. We simulate $N = 1000$ trajectories of each process to get the box-plots. The points represent the extrema and the bars the quantile of order $(2.5\%, 25\%, 50\%, 75\%, 97.5\%)$.

where K is a constant, recalling that these computations are approximate. Take $K < 1$ to be sure to reject H_0. We notice that there is no condition on σ. Moreover the smaller is Δt the larger must be λ to detect the process as confined. We detect directed Brownian as effectively directed diffusion if:

$$\frac{\|v\|}{\sigma} > \frac{(A-1)\hat{q}^n_{1,1}\left(1-\frac{\alpha}{2}\right)}{\sqrt{n\Delta t}}, \tag{15}$$

with $A > 1$ a constant playing the same role as K in (14). With such choices of λ and v, we see that the box-plots of the test statistic (11) of the different processes are disconnected. It means that (11) is a good statistic to discriminate these processes. We apply our test procedure on the simulations used in Fig. 1 to classify the trajectories as Brownian, confined diffusion or directed diffusion. We also classify the trajectories using the MSD method. We build MSD curve using [6, Equation 7]. For a trajectory of length n, we compute MSD(t) at $t = \Delta t, \dots, (n-5)\Delta t$ and not until $t = n\Delta t$: the estimator relies then on less than 5 observations which makes it too variable. Then we fit MSD(t) to the function $t \to Ct^\beta$ parametrized by (β, C), C a scale parameter generally not of interest. In practice, we fit $\log(\text{MSD}(t))$ to $\log(C) + \beta\log(t)$ by a linear regression to determine $(\beta, \log(C))$. In the Brownian case i.e $\beta = 1$ we have $\sigma = C$ (σ the diffusion coefficient). Then we classify the trajectories into the three types of diffusion according to the values of β [3].
Moreover, we also assess the robustness of our test procedure to noise. We model the positional noise with a Gaussian noise of variance σ_{err}. We quantify the level of noise with the signal to noise ratio defined as SNR $= (\sigma\sqrt{\Delta t})/\sigma_{err}$. We build confusion matrices to assess our test procedure and the MSD method. First, we see that our test procedure detects very well the ground truth, as confirmed by the confusion matrix Table 2 which is close to the identity matrix. Our test is also robust to noise as we get good results with a small SNR (see Table 2). By contrast we notice that the MSD does not detect well Brownian and confined diffusion (see Table 3).

4.2. Results on real data

We apply our test procedure on sequences of fluorescent images (TIRF microscopy) depicting the traffic of Langerin pro-

Test label	BR	OU	DBR	BR	OU	DBR
Ground truth	without noise			with noise		
BR	94.8	2.1	3.1	87.4	12.5	0.1
OU	1.1	98.9	0	0.4	99.6	0
DBR	0.4	0	99.6	4.9	0	95.1

Table 2: Confusion matrices of our test on the same simulation used to build Fig. 1. BR stands for Brownian, OU for Ornstein-Uhlenbeck and DBR for Directed Brownian. Results are written in %. We read 0.1% of the simulated Brownian trajectories with noise are labelled as directed Brownian by our test. For the noisy case we set $\sigma_{err} = 0.4$ which gives SNR=0.5.

MSD label / Ground truth	BR	OU	DBR	ML	NC
BR	21.8	80.9	26.2	0.7	0.4
OU	0	9.1	0	35.4	55.5
DBR	0	0	100	0	0

Table 3: Confusion matrix of the MSD method on the same simulation used to build Fig. 1. Results are written in %. ML stands for 'motionless', NC for non classified. In this latter class, we put trajectories for which $\beta < 0$ which is not expected according to [3]. For most of these trajectories β is significantly equal to 0 (t-test).

tein in micro-patterned cells. In Fig. 2, the sequence is composed of 1 199 images of size 256×283 (1 pixel=160nm) acquired at 10 frames/s ($\Delta t = 0.1s$). 5 506 trajectories are computed with the ICY tracker[11]. As we are interested in moving particles, we analyse only the trajectories which have at least 9 distinct positions and which stops less than $K = \lfloor n/10 \rfloor$ times (with n the length of the trajectory). We end up with 1 618 trajectories whose median length is $n = 84$. In Fig. 2, our preliminary results show that our approach and the MSD method do not produce similar classification results. As noticed in the simulations (Table 3), the MSD analysis labels more trajectories as confined or directed diffusion (see Fig. 2) compared too our approach. From the simulations, we know that this over-detection of confined and directed diffusion is wrong: we may conclude that it is also the case for the real Langerin sequence. With our test, we label 68% of the trajectories of the sequence as Brownian, 27% as confined diffusion and 5% as directed diffusion while with the MSD method we detect 12% of the trajectories as Brownian, 66% as confined diffusion, 14% as directed diffusion and 8% as 'motionless'.

5. CONCLUSION

We have proposed a test procedure to classify the motions of particles within cell which is statistically consistent. It is an alternative to the MSD method which gives more reliable results from Monte Carlo simulations. The method has been evaluated on sequences of real data showing Langerin protein dynamics. Future work will concern the detection of change of motion dynamic over time.

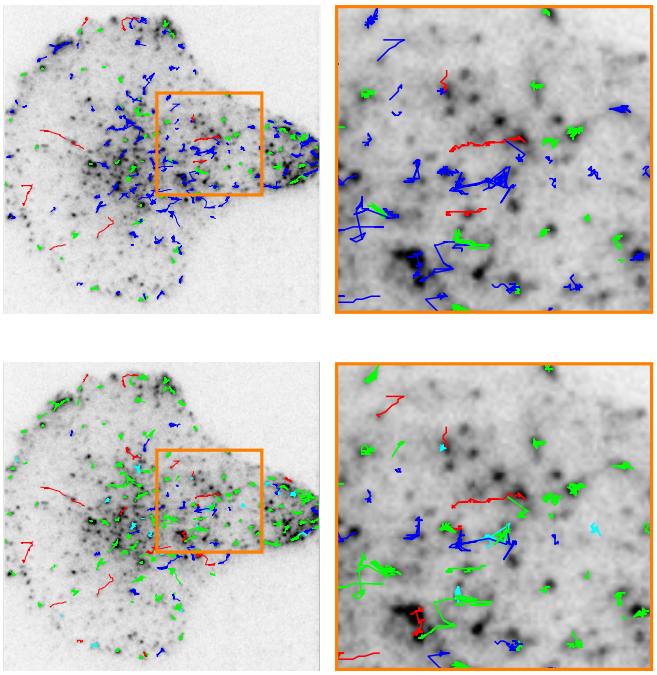

Fig. 2: Labelling of the dynamics of trajectories on the Langerin protein sequence (Courtesy of UMR 144 CNRS Institut Curie - PICT IBiSA). For more clarity, we show only the trajectories appearing on the first 100 frames and black and white have been inverted. The color code is red for directed Brownian, green for Ornstein-Uhlenbeck , blue for Brownian, cyan for 'motionless'. Top panel is labelled with our test, bottom panel with the MSD method.

REFERENCES

[1] P.C. Bressloff and J.M. Newby, "Stochastic models of intracellular transport," *Reviews of Modern Physics*, vol. 85, no. 1, pp. 135, 2013.

[2] H. Berry and H. Chaté, "Anomalous diffusion due to hindering by mobile obstacles undergoing brownian motion or orstein-ulhenbeck processes," *Phys. Rev. E*, vol. 89, no. 2, pp. 022708, 2014.

[3] T.J. Feder et al., "Constrained diffusion or immobile fraction on cell surfaces: a new interpretation," *Biophysical J.*, vol. 70, no. 6, pp. 2767–2773, 1996.

[4] F.W. Lund et al., "Spattrack: An imaging toolbox for analysis of vesicle motility and distribution in living cells," *Traffic*, vol. 15, no. 12, pp. 1406–1429, 2014.

[5] H. Qian, M.P. Sheetz, and E.L. Elson, "Single particle tracking. analysis of diffusion and flow in two-dimensional systems.," *Biophysical J.*, vol. 60, no. 4, pp. 910, 1991.

[6] X. Michalet, "Mean square displacement analysis of single-particle trajectories with localization error: Brownian motion in an isotropic medium," *Phys. Rev. E*, vol. 82, no. 4, pp. 041914, 2010.

[7] S. Karlin and H.M Taylor, *A Second Course in Stochasic Processes*, Academic, 1981.

[8] F.C. Klebaner et al., *Introduction to Stochastic Calculus with Applications*, Imperial College Press, 2012.

[9] I.V. Basawa and B.L.S. Prakasa Rao, *Statistical Inferences for Stochasic Processes*, Academic, 1980.

[10] A.N. Borodin and P. Salminen, *Handbook of Brownian Motion-Facts and Formulae*, Birkhäuser, 1996.

[11] F. De Chaumont et al., "Icy: an open bioimage informatics platform for extended reproducible research," *Nature methods*, vol. 9, no. 7, pp. 690–696, 2012.