ECMI Modelling Week 2018
University of Novi Sad, Faculty of Sciences

DIFFUSION AND ANOMALOUS DIFFUSION MODELS: SIMULATION AND APPLICATION TO BIOLOGICAL DATA

Michał Balcerek Wrocław University of Science and Technology

Nikity Belyak	Lappeenranta University of Technology
Peter Asbjørn Bysted	Technical University of Denmark
Giulia Celora	University of Oxford
Laia Domingo Colomer	Autonomous University of Barcelona
Nikolina Romić	University of Osijek
Nataša Topić	University of Novi Sad

15.10.2018

Contents

1 Introduction

In the process of protein transcription, the genetic information needs to be sent from the DNA to the ribosomes. In order for this to happen, the information is conveyed by mRNA molecules. In the work by [Golding and Cox 2004], the authors have been able to study the dynamics of these particles into living bacterial cells. The newly recorded data have been then analyzed based on a common Brownian motion model. This has been used to estimate macro-scale parameters related to the movement, such as displacement of the tethered molecule, elongation rate and diffusion coefficient. The results found agreed with know biophysical properties of the cells. However, a qualitative analyses of the videos has revealed a more complex picture. If some of the RNA molecules seem to move freely in the cell, the molecules remain often partly bounded to the DNA. This results in a limited and constrained motion, which can not be described by standard diffusion model. For this reason, we here propose a new quantitative analyses of the same data based however on the more general Fractional Brownian Motion model. This allows us to encompass a wide range of dynamics with little assumption on the actual motion of the particles. In particular, a single parameter, H, will identify the general nature of the transport mechanism: either sub-diffusion, normal diffusion or super-diffusion. Starting from a general overview of Fractional Brownian Motion in the framework of stochastic analysis, we will then present the general procedure followed in the analyses of the data-set. The results obtained will be presented in Section 4, where we will also discuss them in comparison to the study previously proposed in the literature. As we will see, the qualitative observations in [Golding and Cox 2004] agree with our mathematical results. These suggest that a sub-diffusion mechanism led the movement of mRNA molecules, supporting the idea that this is passive and impede by external constraints. The new mathematical model will finally be used to extrapolate information on the actual biophysical system, with particular focus on the diffusion coefficient.

2 Theory

As mentioned, the first part of this work is focused on a brief overview of fractional Brownian motion in the framework of stochastic process. In particular, we will present the key feature that uniquely characterize this class of models.

Throughout this section we will denote by X a general random variable. A continuous-time stochastic process is a collection of indexed random variables $\{X(t)\}_{t\in\mathcal{T}}$, where $\mathcal{T}$ is a set of index sampled on the real line.

We start from the general concept of *stable distribution*, which is defined as follows:

Definition 2.1. (Stable distribution) A random variable X is said to be stable if for the two independent copies of X, X_1, X_2 and for $a, b, c > 0$ and $d \in \mathbb{R}$

$$aX_1 + bX_2 = cX + d$$

Ȧ stable distribution is uniquely defined by 4 parameters: $\alpha \in (0, 2]$, the stability parameter, $\beta \in [-1, 1]$ the skewness parameter(not the third order moment, which is undefined), $\gamma \in \mathbb{R}_{>0}$ the scale parameter and $\delta \in \mathbb{R}$ the location parameter. For this reason, this are usually denoted as $S(\alpha, \beta, \gamma, \delta)$.

Normal distributions, Cauchy distributions and Lévy distributions are just some of the most well-known examples belonging to this class.

From the point of view of stochastic processes, stable distribution have a key advantage. If the process $X(t)$ is stable than the increment and process distribution belong to the same family. A classical example is the Brownian motion.

Definition 2.2. (Brownian motion) A brownian motion, X_t is a stochastic process characterized by the four properties

1. $X_0 = 0$
2. X_t is almost surely continuous
3. The increments are stationary and independent
4. The increments are normally distributed with $(X_s - X_t) \sim N(0, s-t)$ for $s \geq t \geq 0$

As mentioned in the introduction, this model has been successfully applied to the study of particle dynamics. However, its usage is limited. A process that encapsulates a greater set of stochastic processes is called the Lévy process.

Definition 2.3. (Lévy process) A Lévy process, X_t is a stochastic process characterized by the four properties

1. $X_0 = 0$ with probability 1.

2. X_t is continuous in probability: For any $\epsilon > 0$ it holds that $\lim_{h\to 0}\mathbb{P}(|X_{t+h} - Xt| > \epsilon) = 0$.

3. The increments are stationary and independent.

The motion is said to be a Lévy stable motion process if it is a Levý process and its increments are distributed by $L_s - L_t \sim S(\alpha, \beta, \gamma|s-t|^{1/\alpha}, 0)$. As it can be seen, this includes standard Brownian motion.

Based on the concepts here presented, we can now defined the idea of Fractional Brownian Motion. This model will then be applied to the study of our data-set.

2.1 Fractional Brownian Motion

A generalization of the Brownian motion can be made, where the increments doesn't need to be independent.

Definition 2.4. A Fractional Brownian Motion, X_t is a stochastic process characterized by the four properties

1. $X_0 = 0$

2. X_t is almost surely continuous

3. The increments of X_t are stationary

4. The increments have an autocovariance given by

$$\mathbb{E}[(X_{t+k+1} - X_{t+k})(X_{t+1} - X_t)] = \frac{1}{2}(|k+1|^{2H} + |k-1|^{2H} - 2|k|^{2H})$$

which is equivalent to

$$\mathbb{E}\left[X_s X_t\right] = \frac{1}{2}\left(s^{2H} + t^{2H} - |s-t|^{2H}\right)$$

where $H \in (0,1)$ is a parameter often called the self-similarity parameter or Hurst parameter. In particular, a Brownian motion is a Lévy stable process with $\alpha = 2$, whereas the fractional Brownian motion for $H \neq \frac{1}{2}$ is not a Lévy stable process.

2.2 Properties of FBM

Having defined FBM, we now look at the key features that characterize this class of stochastic process. First of all, we introduce the concept of *stationary process*.

Definition 2.5. A stochastic process $\{X(n)\}_{n=0,1,\cdots}$ is said to be stationary if its unconditional joint probability distribution $F(x_{t_1}, \cdots F_{x_{t_n}})$ does not change when shifted in time:

$$F(x_{t_1+\tau}, \cdots F_{x_{t_n+\tau}}) = F(x_{t_1}, \cdots F_{x_{t_n}}) \tag{1}$$

When dealing with FBM, the distribution of the increments is stationary. In other words, if we consider $\{X(n)\}_{n=0,1,\cdots}$ to be a fractional Brownian motion, we can define the process $\{\mathrm{d}X(n)\}_{n=1,2\cdots}$, where $\mathrm{d}X(n) = X(n) - X(n-1)$. Then we have that $\mathrm{d}X$ is a stationary process.

Consequently the stationarity of the increments is a necessary condition for a process to be a fractional Brownian motion. Given several realizations of the process, this requirement can be check with the use of a *quantiline test*. This technique will be presented more in detail in the next section.

Another important property of the increment of a FBM process is ergodicity. Intuitively, a stationary stochastic process is said to be ergodic if all of its properties can be deduced from only one sample of the process. When this is the case, the Bolztmann ergodic hypothesis holds: the time averages converge to ensemble averages. Given a general function g depending on the process, this translates in the following condition:

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T g\left(Y(t)\right) \,\mathrm{d}t = \mathbb{E}\left[g\left(Y(0)\right)\right] \tag{2}$$

Notice that in order to extrapolate valid information from the analysis of a sample, this property is essential. However, from a practical perspective, a more useful definition of ergodicity is the one introduced in [Burnecki et al. 2012], which first requires the definition of the *dynamical function*.

Definition 2.6. For a weak stationary process $\{Y(n)\}_{n=0,1,\cdots}$ the dynamical function $D(n)$ is defined as

$$D(n) = \mathbb{E}\left[\exp\left\{i(Y(n) - Y(0))\right\}\right] \tag{3}$$

Given this, we can defined an ergodic process as follows:

Definition 2.7. A stationary stochastic process $\{Y(n)\}_{n=0,1,\cdots}$ is said to be ergodic if

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} D(k) = |\mathbb{E}\left[\exp\left\{iY(0)\right\}\right]|^2 \tag{4}$$

Another fundamental property is mixing, that is, the asymptotic independence of two random variables $X(n)$ and $X(0)$ as n goes to infinity. Formally:

Definition 2.8. A stationary stochastic process is said to be mixing if and only if

$$\lim_{n\to\infty} D(n) - |\exp\left\{iY(0)\right\}|^2 = 0 \tag{5}$$

In some cases it is easier to check if a process has the mixing property than if this is ergodic. Since the mixing property implies ergodicity, this can be particularly useful in the analysis of data.

3 Method

Having completed a brief overview of the theory behind *Fractional Brownian Motion*, we now apply this model to the study of RNA dynamics in real cells. To do that, we have worked on real data, provided by [Golding and Cox 2004]. Before moving to actual analysis, we first focus on the description of the data. Indeed, due to the intrinsic problematic associated with real measurements, these can not be directly used, but they first need to be pre-processed.

3.1 Dataset

The dataset consists of the trajectories of 27 RNA molecules in live *Escherichia Coli* cells. For each particle the x and the y components of the position had been recording using epiflourescent microscopy at different times. An accurate description of the experimental method applied can be found in the study by Golding et al. [Golding and Cox 2004], who kindly provided us the results of their experiments. The position of each particle has been recorded for around 10 minutes, sampling every other second, with a total of $300 - 500$ samples for particles. However, out of these 27 trajectories, only 10 are complete, while the remaining present gaps from 1 up to 20 frames missing. As for any other experiment, the recording are likely to have been subjected to noise, however no information on its nature was available.

3.2 Interpolation

As mentioned previously, most of the trajectories in the data set present gaps. However, our analysis requires the sampling rate to be constant. Following a common approach, we were able to reconstruct continuous path interpolating the data at our disposal using deterministic methods. More precisely, for each trajectories we have considered only the longest interval with gaps of maximum length 3, i.e. for which up to two detection were missing. In this way, of the 27 initial particles, only 4 of them have to be discarded. Let us denote with x_i and x_{i+1} the recorded particle position at time t_i and t_{i+1} respectively. In order to find the intermediate position $\hat{x}$ at time $\hat{t}$, such that $d_i = \hat{t} - t_i > 0$ and $d_{i+1} = t_{i+1} - \hat{t} > 0$, we have used the interpolating technique known as *inverse squared distance weighting*, which is based on the formula:

$$\hat{x} = \frac{\frac{x_i}{d_i^2} + \frac{x_{i+1}}{d_{i+1}^2}}{\frac{1}{d_i^2} + \frac{1}{d_{i+1}^2}}. \tag{6}$$

We just point out that in the case of gap of $d_i = d_{i+1}$, e.g a single missing detection, Equation (6) reduces to simple linear interpolation. The same technique could also be applied to fill longer gaps, in order to recover more and longer samples than the one used in our study. However, the reader may be aware that this *deterministic* modification of the data set might introduce a bias, leading to an inaccurate mathematical description

of the particle motion. Indeed, even though the approach above discussed is commonly used, there are no qualitative or quantitative studies in the literature on how this impact on the results of time series analysis. Due to the limited amount of time, we have not had the chance to look at this aspect, thus this can be a natural extension of the work here presented.

3.3 Identifying FBM

In order to properly identify FBM, a method was proposed in [Burnecki et al. 2012]. The algorithm is as follows:

1. Use parallel quantile lines to check if the increments of the data are stationary
2. Verify the probability distribution of the increments using normality tests
3. Check if the data is mixing and thus ergodic using the dynamic functional test.
4. Find the self similarity parameter H and memory parameter d using the sample and ensemble average mean squared distance.
5. Validate the model using the generalized p-variation test.

3.4 Probability distribution of the increments

In a fractional Brownian motion $\{X_t\}_{t>0}$, the increments follow a standard normal distribution. In order to check for normality, multiple statistical tests have been computed.

1. **Kolmogorov-Smirnov test:** It computes the maximum difference between the theoretical and empirical distribution function. It is sensitive to differences in both location and shape of the empirical and theoretical cumulative distribution functions.
2. **Lilliefors test:** It is an improvement of Kolmogorov-Smirnov test when both the mean and variance are unknown. It measures the same statistic as the Kolmogorov-Smirnov test but considering that it has a different distribution.
3. **Anderson-Darling test:** It computes a weighted distance between the empirical and theoretical distribution function. This method gives a higher weight to the tails of the distribution, while the previous two test focus more on the central values.
4. **Jarque bera test:** It compares the skewness and kurtosis of both the empirical and the theoretical distribution.

3.5 Ergodicity via Mean Square Displacement

In order to check if the process is ergodic, we will calculate the estimators of the ensemble average mean square displacement ($EA-MSD$), the time average mean square displacement ($TA-EA-MSD$) and the time and ensemble mean square displacement ($EA-TA-MSD$)

$$EA-MSD(\tau)=\frac{1}{N}\sum_{k=1}^{N}(X_k(\tau)-X_k(0))^2 \tag{7}$$

$$TA-MSD(k,\tau)=\frac{1}{N+1-\tau}\sum_{t=0}^{N-\tau}(X_k(t+\tau)-X_k(t))^2 \tag{8}$$

$$EA-TA-MSD(\tau)=\frac{1}{N}\sum_{k=1}^{N}TA-MSD(\tau) \tag{9}$$

For $\tau=0,1,\cdots,N-1$.

In an ergodic process, $EA-MSD$ and $EA-TA-MSD$ should coincide in the long time and sample limit. However, we are dealing with a finite and rather short sample, both in time and ensemble, and therefore one would expect there to be differences in $EA-MSD$ and $EA-TA-MSD$ even if the process is ergodic. The problem is to decide whether the differences are significant or not. Just as it was done in [Janczura and Weron 2015], we will not only compare the curves of $EA-MSD$ and $EA-TA-MSD$ but also to see if the latest lies between the confidence interval for $EA-MSD$. If it does, we will conclude that the process is ergodic and that differences between $EA-MSD$ and $EA-TA-MSD$ come from the finite sample. As it was calculated in [Janczura and Weron 2015], the confidence interval for $EA-MSD$ with confidence α is given by

$$\left(\frac{EA-MSD(\tau)-\mathbb{E}(X^2)}{Z_{\chi^2_n,1-\alpha/2}/N},\ \frac{EA-MSD(\tau)-\mathbb{E}(X^2)}{Z_{\chi^2_n,\alpha/2}/N}\right) \tag{10}$$

Where $Z_{\chi^2_n,\alpha/2}$ is the quantile of order α of the χ^2 distribution with n degrees of freedom.

3.6 Estimation of parameter H

3.6.1 Memory parameter method

When each particle has been proven to have a fractional Brownian motion, we will estimate the self-similarity parameter H. If $H<1/2$ there is subdiffusion, if $H=1/2$ the process has a Brownian Motion, and if $H>1/2$ there is superdiffusion. After estimating the H parameter, we will be able to make different clusters of particles corresponding to different values of H.

If the data sample comes from an H-self similar process with stationary increments

belonging to the domain of attraction of the Lévy α stable law, then for a large time domain N,

$$TA - MSD(\tau) \sim \tau^{2H} \tag{11}$$

By doing a linear fit of $\log(TA - MSD(\tau))$ with $\log(\tau)$ we can obtain parameter H.

3.6.2 p-Variation test

Another method to estimate the H parameter for a is the so called p-Variation test. Given $p > 0$ and for $\tau = 0, 1, \cdots, N$ we define the function

$$V(\tau)^{(p)} = \sum_{k=0}^{N/m-1} |X((k+1)\tau) - X(k\tau)|^p \tag{12}$$

Then, we plot $V(\tau)^{(p)}$ for $p = 1/H$, where $H = 0, 0.1, 0.2, \cdots, 1$. If H_r is the real value of the H parameter in our sample, we expect that for $p < 1/H_r$ $V(\tau)^{(p)}$ is a decreasing function of time, whereas for $p > 1/H_r$ $V(\tau)^{(p)}$ increases with time. Therefore, at $p \sim 1/H_r$ the plot of $V^{(p)}$ should become flat. This method is not as precise as the memory parameter method, as there is a wide range of p that give similar behaviours of $V^{(p)}$. However, it is a good tool to check the results obtained in the Memory parameter method.

4 Results

4.1 Simulations

In order to establish an intuition about the effect of the Hurst parameter, different FBM sample paths were simulated using different values of H. The variance of the increments were kept the same for all the three paths. For the low value of H, it can

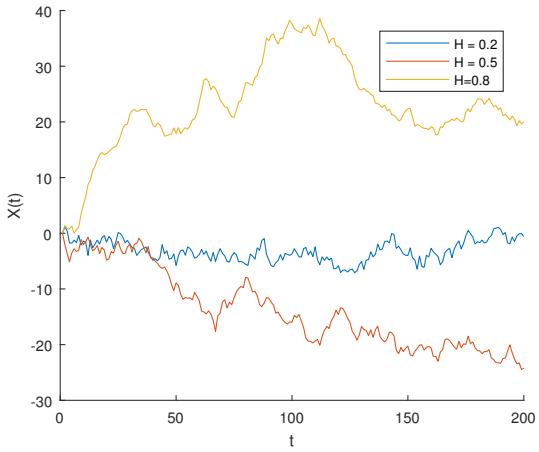


Figure 1: Simulation of three different sample paths, using different values of H

be seen that the motion has a tendency to regress to the mean, meanwhile for large H the motion has a tendency to cluster in a direction, and for $H = 0.5$ the motion is standard Brownian motion. This can be confirmed by inspecting the autocovariance function of the increments of the FBM given in definition 2.4. The correlation will be negative for $H < 0.5$, zero for $H = 0$ and positive for $H > 0.5$.

4.2 Gaussianity tests

In order to verify the probability distribution of the trajectory increments, we will perform multiple normality tests as described in section 3.4. We have performed these tests with both the interpolated data in the initial coordinates and the one obtained after the Principal Component Analysis. We obtained similar results for both sets of data. Here we show the p-values obtained for the four tests and for both the principal and orthogonal principal component.

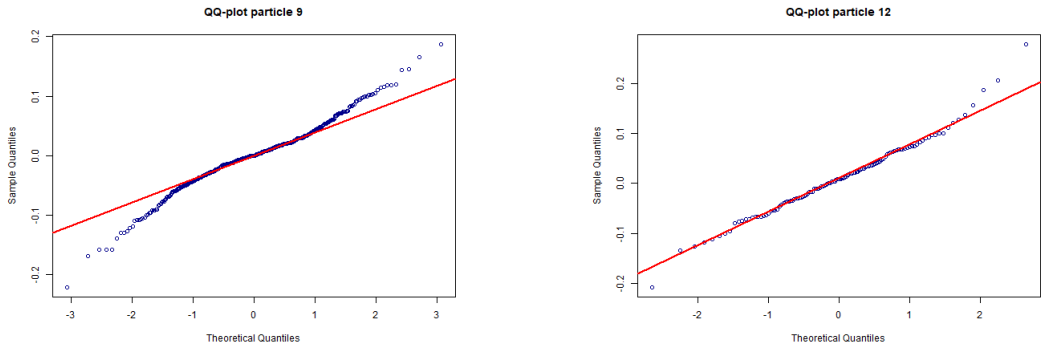
Particle	Kolmogorov-Smirnov test	Anderson-Darling test	Lilliefors test	Jarcque-Bera test
1	0.972	0.357	0.812	$\sim 10^{-16}$
2	0.282	0.00120	0.0200	$\sim 10^{-16}$
3	0.102	0.000268	0.00134	0.00815
4	0.0591	$\sim 10^{-14}$	$\sim 10^{-7}$	$\sim 10^{-16}$
5	0.6795	0.172	0.237	$\sim 10^{-10}$
6	0.404	0.0291	0.0535	0.00235
7	0.392	0.0136	0.0491	0.000469
8	0.826	0.0514	0.440	0.000641
9	0.0123	$\sim 10^{-9}$	0.0217	$\sim 10^{-14}$
10	0.0894	$3.68 \cdot 10^{-5}$	0.000688	$\sim 10^{-16}$
11	0.329	0.0114	0.0298	$\sim 10^{-16}$
12	0.476	0.342	0.0837	0.232
13	0.101	0.00148	0.000968	$\sim 10^{-16}$
14	0.545	0.00292	0.00542	0.00163
15	0.534	0.0392	0.115	0.0462
17	0.955	0.537	0.753	0.146
18	0.0193	$\sim 10^{-14}$	$\sim 10^{-9}$	0.688
21	0.861	0.305	0.505	0.904
22	0.087	$\sim 10^{-7}$	$\sim 10^{-6}$	0.00296
23	0.127	0.528	0.0216	0.0265
25	0.0650	$\sim 10^{-13}$	$\sim 10^{-6}$	$\sim 10^{-16}$
26	0.0729	0.0575	0.00381	0.0286
27	0.0838	$\sim 10^{-6}$	0.0650	$\sim 10^{-7}$

Table 1: p-value for the Gaussianity tests of the principal component increments.

Particle	Kolmogorov-Smirnov test	Anderson-Darling test	Lilliefors test	Jarcque-Bera test
1	0.9167	0.0552	0.00670	0.00962
2	0.251	0.00876	0.0144	$\sim 10^{-66}$
3	0.146	0.0009288	0.00321	0.00405
4	0.0418	$\sim 10^{-16}$	$\sim 10^{-11}$	$\sim 10^{-16}$
5	0.105	0.172	0.0121	0.268
6	0.182	0.0291	0.00572	0.00173
7	0.0958	0.0136	0.00912	0.00656
8	0.415	0.0514	0.0578	0.00267
9	0.255	0.0151	0.0155	$\sim 10^{-14}$
10	0.0897	$\sim 10^{-6}$	0.00695	$\sim 10^{-13}$
11	0.0173	$\sim 10^{-8}$	$\sim 10^{-5}$	0.0133
12	0.853	0.2896	0.491	0.421
13	0.233	0.00122	0.0109	$\sim 10^{-5}$
14	0.858	0.373	0.501	$\sim 10^{-8}$
15	0.607	0.149	0.166	0.602
17	0.695	0.0517	0.253	0.00114
18	0.935	0.517	0.684	0.152
21	0.892	0.505	0.573	0.303
22	0.0297	$\sim 10^{-7}$	$\sim 10^{-5}$	0.00109
23	0.178	$\sim 10^{-6}$	0.00563	$\sim 10^{-8}$
25	0.0377	$\sim 10^{-10}$	$\sim 10^{-8}$	$\sim 10^{-8}$
26	0.193	0.0575	0.653	0.310
27	0.0856	$\sim 10^{-6}$	0.00694	$\sim 10^{-7}$

Table 2: p-value for the Gaussianity tests of the orthogonal principal component increments.

A p-value smaller than 0.01 means that if the increments followed a Gaussian distribution, the probability of obtaining the experimental values would be less than 1%. In that case, we would reject the hypothesis of normality. Otherwise, if the p-value is greater than 0.01 we would accept that the increments are normally distributed. If we look at tables 1 and 2, we can see that there are particles, such as particle 12 which pass all four tests of normality. However, there are other particles which fail only some of the tests. We noticed that most of the particles seem to have worse results in the Anderson-Darling and Jarque-Bera tests. These two tests give more value to the tails of the distribution than to the central values. This motivates us to look at the QQ-plot of these distributions. The following figures show an example of these QQ-plots for two different particles.



(a) Particle 9 (b) Particle 12

Figure 2: QQ-plots of the increments' distribution of the principal component.

The QQ-plot show the quantiles of the experimental distribution against the theoretical distribution. If the two distributions are equal, the experimental data should coincide with the theoretical one. This happens with particle 9 of figure 2. If we look at the QQ-plot of particle 12, we can see that the experimental distribution coincides with the theoretical one in the central values, whereas in the tails the experimental distribution has more extreme values. The presence of heavy tails is usually caused by the existing noise of the data. As we do have noisy data, it was expected to have heavy tails and therefore worse results in the normality tests. In order to check if this is the case, we could repeat the experiments with longer trajectories and see if the distribution resemble more to a normal one.

Since all particles seem to pass the Kolmogorov-Smirnov test, and the failure of some of the other tests is caused by the presence of heavy tails, which might be attributed to the noise, we can accept that the general trend of the increments follows a Gaussian distribution. Therefore, we accept the hypothesis of normality and move on to check the other properties of the fractional Brownian motion.

4.3 Estimation of the H parameter

Now that we have checked for the normality of the increments, we will estimate the Hurst parameter for each particle and each motion component. Assuming that the particles follow a fractional Brownian motion, we can use the *memory parameter method* to estimate the value of H. As presented in Section 3.6.1, this method is based on the asymptotic behaviour of the mean square displacement. As illustrated in Figure 3, in order to estimate H a linear approximation of the logarithm of the mean square displacement is considered. According to the theory, the slope of the is $\sim 2H$.

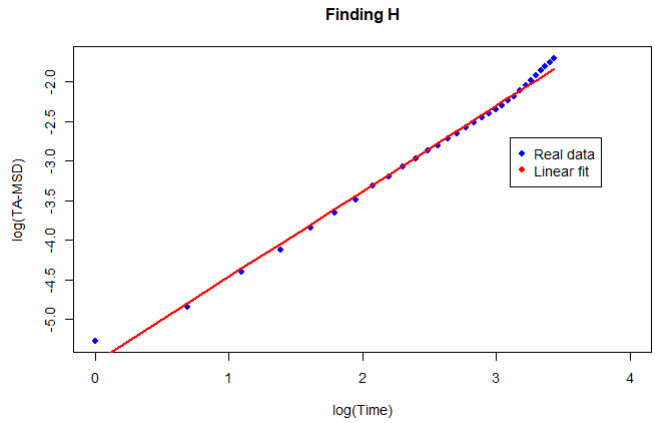


Figure 3: Memory parameter method to estimate the Hurst parameter for particle 1 in the principal component.

Using this technique, we have calculated the H parameter for the original x and y coordinates. The result are illustrate in Figure 4.

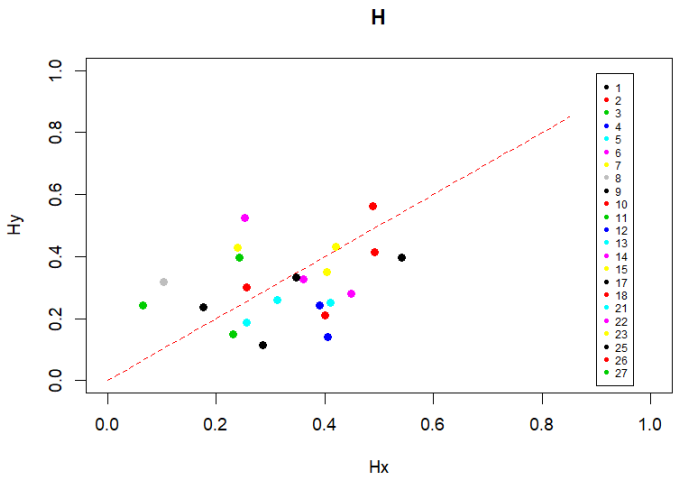


Figure 4: H parameter for all particles in the x and y direction.

We noticed that there are pairs of particles that have symmetric H values, that is,

that the value of H in the x coordinate for one particle is similar to the value of H for the y coordinate for another particle. Moreover, looking at the original video of the detection [Golding and Cox 2004], we noticed the bacteria cells are not aligned, but disposed with a random inclination. This facts motivate us to perform the principal component analysis. In this way we were able to identify for each particle the direction of movement with the largest variability. By changing the coordinates in this way, we would expect that the value of H for the principal component is larger than the value of H for the orthogonal principal component, as shown in Figure 5. We can now discern clearly a common trend in the behaviour of different particles.

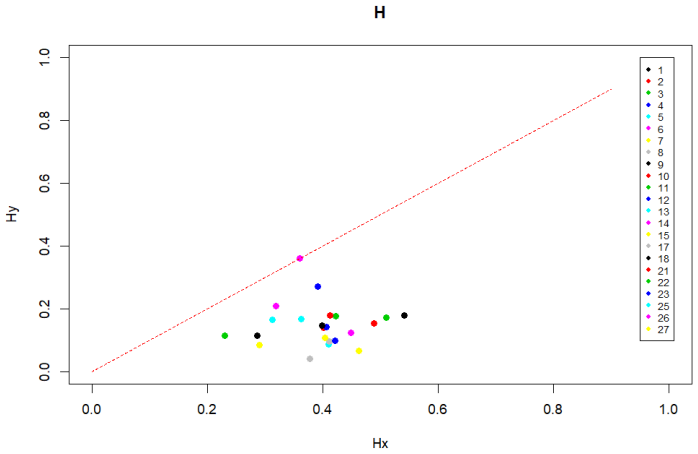


Figure 5: H parameter for all particles in the principal component and orthogonal principal component.

In this case, most of the particles have a value of $H < 0.5$, and only three of them have $H \sim 0.5$. In other words, most of the particles seems to be subjected to sub-diffusion, meanwhile only few behaves according to standard Brownian motion.

4.4 Clustering

By looking at Figure 5 we have clustered particles according to the value of H. Since we consider the motion in the principal and orthogonal component independent, we form different clusters as listed in Table 3 and Table 4. We now assume that the trajectories in the same cluster are realization of the same process, so that this can be treated as an ensemble.

Principal component:

Cluster	Particles	mean H
1	3,4,5,6,7,8,10,11,13, 14,15,17,18,21,23,25,26	0.385
2	1,2,22	0.513
3	9, 27	0.288

Table 3: Clusters for the principal component.

Orthogonal principal component:

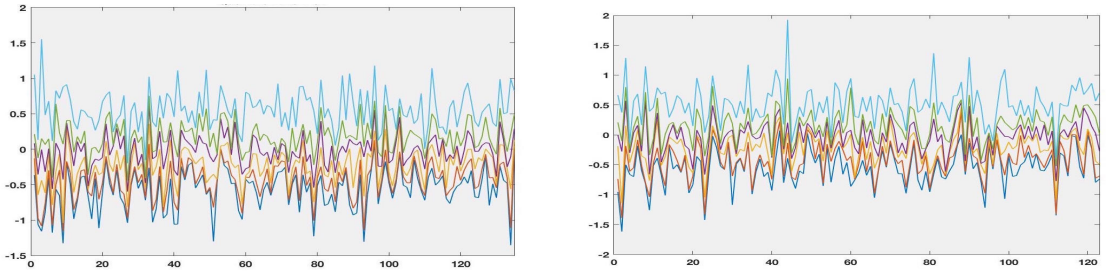
Cluster	Particles	mean H
1	1,2,3,6,9,10,11,12,13, 15,18,21,22,25,26	0.150
2	5,7,8,17, 23,27	0.0782
3	4,14	0.315

Table 4: Clusters for the orthogonal principal component.

4.5 Stationarity

Following the procedure outlined in Section 3.3, we can now use each cluster identified as an ensemble. In particular, we will consider Cluster 1 in Table 3 and Table 4 as independent process defining the motion of the particle in the principal and orthogonal component respectively.

The first step consists in the analysis of the quantile lines for the increments. As specified previously, for a stationary process as FBM, these are expected to be parallel. However, as illustrated in Figure 6, this is not the case for neither of the samples considered. Instead of straight lines, the plots present highly oscillating curves.



(a) Quantile lines Principal component

(b) Quantile lines Orthogonal component

Figure 6: Quantile lines plots with the following color convention: 0.2 (dark blue), 0.3 (orange), 0.4 (yellow), 0.5 (purple), 0.6 (green) and 0.75 (light blue).

However, the same trend characterizes Figure 7. The latter is obtained using samples of the same size as the one we have been using but generated from an exact distribution.

Even though we know that the underline process is stationary, the limited size of the ensemble make the quantile-line test inconclusive.

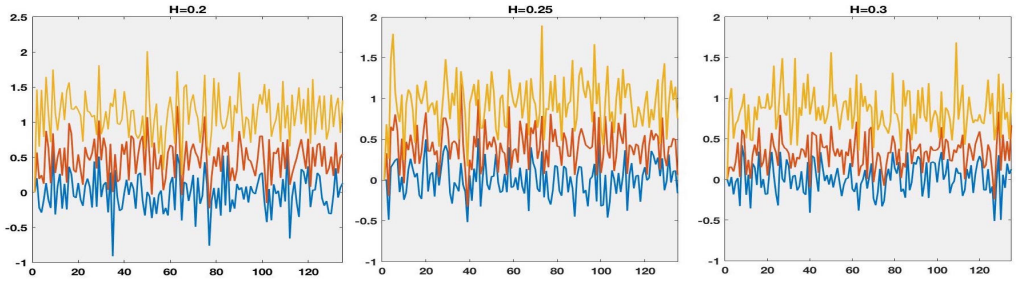
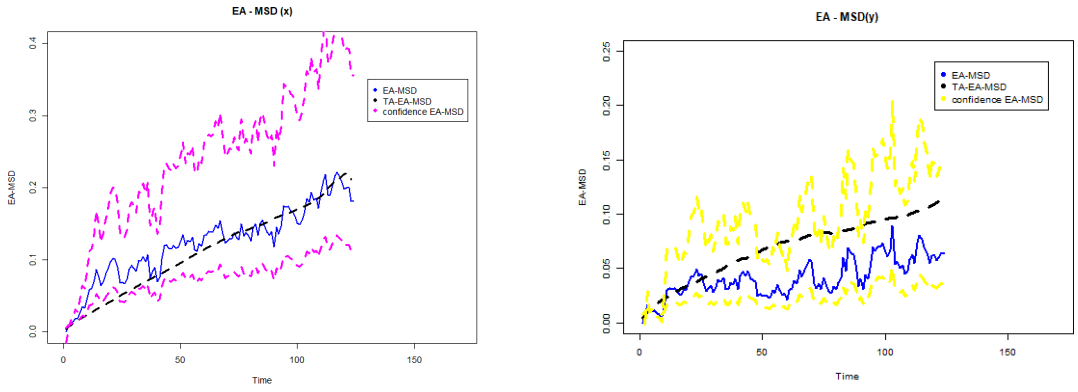


Figure 7: Quantile lines obtained with 20 samples generated from an exact Fractional Brownian Motion: $H = 0.2$ (left), $H = 0.25$ (centre) and $H = 0.3$ (right). Each color stands for a different quantile: 0.25 (blue), 0.5 (orange) and 0.75 (yellow).

Given this, in order to proceed with our analysis, we will just assume that the processes are indeed stationary.

4.6 Mean square displacement for ergodicity

As described in Section 3.5, we will now check for ergodicity using each cluster as an ensemble of particles of the same distribution. The following figures show the values of EA-MSD, TA-MSD and EA-TA-MSD as a function of time, and also the confident intervals for EA-MSD, for the first clusters.



(a) Ergodicity for the principal component. (b) Ergodicity for the orthogonal component.

For the principal component, the $EA - TA - MSD$ fits perfectly inside the confidence interval of the $EA-MSD$, with a confidence of $1-\alpha = 0.95$. Therefore, we conclude that the process is ergodic. For the orthogonal principal component, the $TA - EA - MSD$ does not always fit in the confidence interval of the $EA - MSD$. However, with the

99% confidence, the values of $EA - TA - MSD$ do lie in the confidence interval, as seen in Figure 9.

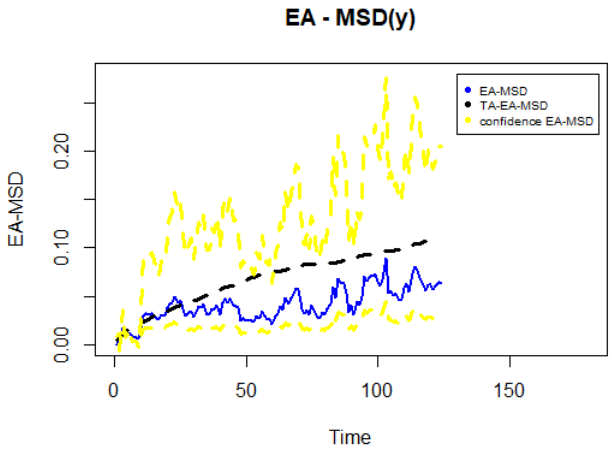


Figure 9: Ergodicity for the orthogonal principal component with the $\alpha = 0.01$ interval of confidence.

Therefore, we conclude that the ergodicity is mostly satisfied for both motion components but the results are better for the principal component.

4.7 A deterministic point of view

As known, diffusion can also be describe deterministically with the use of Partial Differential Equation. In particular, the motion is characterized by the value of the diffusion coefficient D. This macroscopic constant can be either estimated experimentally or determined based on microscopic properties of the motion. In particular, the following relation holds:

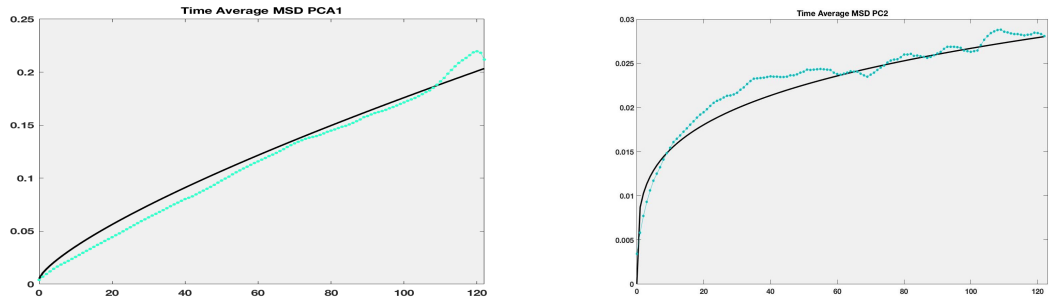
$$TA - EA - MSD(\tau) \sim 6\, D\, \tau^{2H} + C \tag{13}$$

Looking at equation (13), we notice that the left hand side is the same quantity estimated in Section 4.6. On the other hand, the right hand side has three unknown parameters, D, C and H. We have already discussed the meaning of H and D. The constant C is instead a measure of the noise present in the measurements. The estimate obtained using MATLAB package for interpolation are listed in Table 5. In Figure 10, we have compared the estimated quantities with the corresponding analytic fit. For both the components, we notice that the value of C is approximately zero. This means the results are unlikely to be affected by a systematic error. However, random error may still plays a role. Despite the difference, the value of H is below 0.5 in both cases. This means that along both direction the motion is not completely random. Remember indeed that $H = 0.5$ corresponds to a standard Brownian motion. We are instead dealing with

sub-diffusion. Despite this indicates a form of passive transport as well as standard diffusion, this is impeded by the presence of obstacles in environment. In particular, the lower the H the more constrained the motion. As expect from the results obtained in Section 4.3, the value of H is higher in the principal component than in the orthogonal one. Consequently, we can conclude that the motion is constrained along one direction, where the particle moves almost randomly. This is in line with the properties of the intra-cellular fluid, or *cytoplasmic matrix*, the particles are diffusing in. Besides its viscosity, this contains several organelles which represent obstacles for the particles. The nature of the media is related also to the diffusion coefficient D. In both directions, the value is of the order of 10^{-3}, which is in the range estimated by [Golding and Cox 2004]: from 10^{-3} to 3×10^{-2}.

	Principal	Orthogonal	Units
D	0.001	0.001446	$\mu m^2/sec$
H	0.3825	0.1220	–
C	0.005186	~ 0	μm^2

Table 5: Estimation of the parameters in Equation (13). Cluster 1 in Table 3 and 4 have been used respectively for the Principal and Orthogonal components.



(a) Principal component

(b) Orthogonal principal component

Figure 10: Plot of $TA - EA - MSD(\tau)$ as defined by 9 and based on the analytic fit given by Equation 13.

5 Conclusion

Starting from a brief overview of fractional Brownian and its properties, we have introduced the methodology proposed by [Burnecki et al. 2012]. This has then been applied to the analyses of RNA molecule dynamics in bacterial cells. Using the principal component analysis, we have first set a 2 dimensional local coordinate system for each trajectory in the data-set. Treating the principal and orthogonal component independently, we have analyzed the motion in the two directions separately. Based on the value of the parameter H, as defined in 4, we have identified 3 clusters. To get a better insight in the nature of RNA motion, each have been considered as an ensemble of realizations of the same process. In agreement with the qualitative observation [Golding and Cox 2004], the results obtained suggests that there is also a quantitative evidence that particles are mainly subjected to sub-diffusion. This supports the idea that RNA molecules are not subjected to active transport mechanism, but they are instead impeded in their motion even though at different degree. However, the small size of the sample of trajectories and the gaps in the position recorded significantly narrowed the reliability of our conclusion. A pre-processing of the data can be applied to recover more and longer trajectories and consequently draw more reliable conclusions. This is a natural extension of the work together with a more detailed analysis of the noise involved in the experimental measurements.

6 Group work dynamics

During the Modelling week 2018 we developed the project regarding diffusion and anomalous diffusion, under the mentorship of our instructor Michał Balcerek. Afterwards, the students group wrote a final report containing the methodology and results obtained during the stay in Novi Sad.

The group dynamics during the modelling week evolved as follows. In the first one an a half days, our instructor taught us some lectures including the fundamental concepts we needed to know and some supplementary information. Afterwards, all together we started getting familiar with the data set and planning the possible steps that we could follow throughout the stay. Once we had a clearer view of the problem, we distributed some tasks between the students and began working. We discussed all together the new trajectories that could lead the project as well as the results and methodology followed by each partner. We distributed the work so that everyone could develop their strengths and provide a useful insight to the project. Since everyone explained their tasks and how they performed them, we could all learn from the others' experience and give suggestions for improvements. In this way, all of us took part in all the parts developed during the stay.

We started writing the report the last day in Novi Sad. All together we decided what we should write and how it should be structured. Afterwards, we distributed the tasks so that everyone could take part in the writing of the report. We wrote

each part individually but we all read the others' part work and give suggestions of improvements.

7 Instructor's assessment

During ECMI Modelling Week in Novi Sad the students were given a task to explore topic connected with bio-mathematics. Specifically, the task was to analyze the well-known data set of *Escherichia Coli* cells, presented in Golding and Cox article. The plan was to propose a mathematical model which could describe movement of the E. Coli bacteria.
Given given such a advanced and complicated problem the students performed really well. During the modelling week they obtained thorough mathematical intuition and knowledge concerning the studied problem. They considered different stochastic models, starting from ones describing classical diffusion, usually used in mathematical biology, going through some of its modifications, and finally examining anomalous diffusion models in form of Fractional Brownian Motion and ARFIMA models. They demonstrated a handful of fresh and interesting ideas, such as using principal component analysis presented in this report. I am considering using such a approach myself!

References

Burnecki, Krzysztof et al. (2012). "Universal algorithm for identification of fractional Brownian motion. A case of telomere subdiffusion". In: *Biophysical journal* 103.9, pp. 1839–1847.

Golding, Ido and Edward C Cox (2004). "RNA dynamics in live Escherichia coli cells". In: *Proceedings of the National Academy of Sciences* 101.31, pp. 11310–11315.

Janczura, Joanna and Aleksander Weron (2015). "Ergodicity testing for anomalous diffusion: Small sample statistics". In: *The Journal of chemical physics* 142.14, 04B603_1.