Hurst parameter estimation on fractional Brownian motion and its application to the development of the zebrafish

Menglong FU
Supervisor: Paul Bourgine

July 7, 2013

Abstract

In order to distinguish different morphogenic fields automatically, fractional Brownian motion is used in this paper and the Hurst parameter is used as an important measure to distinguish different morphogenic fields in embryo of zebra fish.

1 Introduction

It is known from embryology that cellular differentiation is one of the principal characteristics of the early development of an animal. In the later stages of the life of the animal this differentiation is evident, as seen by the formation of different tissues and organs with different functions and organizations. However, the identification of different types of cells in the beginning of its lifetime, and specially during its formation, is a complicated activity whose complete details are yet unknown. What is presently possible to do is to obtain images from the whole development of the embryo and to follow this differentiation step by step, having continuous access to the whole formation of tissues. This allows the identification of cellular types, but it is currently done in a complicated and laborious way.

In the BioEmergences platform, from where we have access to a large database of microscopy of in the vivo development of the Zebra fish embryo and to the data of the tracking of cells, this procedures currently mainly by backtracking. The different morphogenetic fields are separated visually by a biologist at the end of the development of the animal visually and the tracking data is used to come back to the past and see the initial position of these different fields. This

procedure is clearly not predictive and demands a heavy load of manual work, without going deep on the origins of the separation of morphogenic fields.

In this project we want to study an automatic procedure that allows the identification of different morphogenic fields (e.g. organs) coming from the tracking of the cells done by the BioEmergencies platform that can be used in a predictive way.

2 Segmentation model

Our approach to this problem is to have a model for the relative movement of cells (that is, the movement that is independent of the hydrodynamic flow of cells) that depends on a given parameter, and to use this parameter to differ different groups of cells.

Some previous evidence from unpublished work internal to the laboratory shows that the relative movement of neighboring cells, that can be considered as stochastic, is subdiffusive. For this reason, the model we choose for this project is fractional Brownian motion(fBm), this being a minimal generalization of usual Brownian motion having an index which measures diffusion.

When two cells are in the same morphogenetic field, they tend to move in similar behavior, that is, have a similar hydrodynamic flow. This phenomenon inspires us to use the assumption that the relative movement of neighboring cells represents a purely stochastic process, which will follows a fractional Brownian motion with a given Hurst parameter H. Different morphogenetic fields are going to be found by local similarity of these Hurst parameters.

The assumption that the relative movement of cells is purely stochastic is clearly broken in the case where two neighboring cells correspond to two different morphogeinc fields, these having two deterministic destinations that are not similar. In this case, high variations are going to be identified, with large Hurst parameters, that correspond to balistic movement. These large variation should allow us to identify the boundary between different fields.

The software we used in this paper is R. We also use the package "dvfbm" in the simulation and modified functions in the real data.

3 Model and parametric estimation

3.1 Fractional Brownian motion

Let $B_H(t)$ be a Gaussian process on $[0, T]$ which has expectation zero for all t and it satisfies

$$E[B_H(s)B_H(t)] = \frac{1}{2}(|t|^{2H} + |s|^{2H} - |t-s|^{2H})$$

where H is the Hurst parameter in $(0, 1)$. Then we call B_H a fractional Brownian motion of parameter H. The form of this correlation function shows easily that is increments are stationary and that:

$$E[(B_H(t) - B_H(s))^2] = |t - s|^{2H}$$

which is the main interest of this process. This property allows us to classify the movement for different parameters:

- For $H < \frac{1}{2}$, the behaviour is *subdiffusive* with non-independent increments.
- For $H = \frac{1}{2}$, it is a standard Brownian motion whose increments are independent.
- For $H > \frac{1}{2}$, the behaviour is *superdiffusive* with non-independent increments.

Here we show some simulations of fBm for different parameters. It is worth to remark that low Hurst parameters represent steps that are negatively correlated, and high ones present positive correlation, with the extremes of $H = 0$ (equivalent to an alternating process) and $H = 1$ (equivalent to balistic movement).

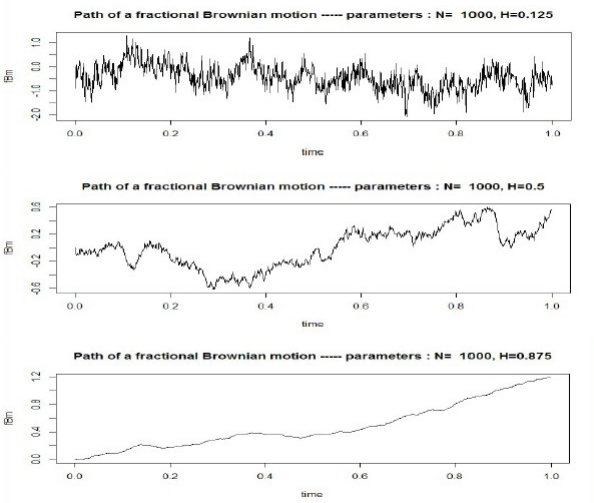

Figure 1: Comparation of different realizations of fBm for the cases $H = 0.125$, $H = 0.5$ and $H = 0.875$ respectively.

3.2 Estimation of Hurst parameter

There are many possible methods of estimation of the Hurst parameter in the literature. Some examples are:

- Discrete variation method, with a rate of convergence of $O(\frac{1}{N})$;
- Whittle's method, with a similar rate of convergence, but that is very sensitive to additive noise and outliers;
- Log-periodogram method, whose rate is $O(\frac{1}{log(N)})$.

The comparison of various different methods composed a first part of this project, which finished by the choice of the first method, due to its good convergence rate, adaptability to different conditions and robustness to noise.

3.2.1 Discrete variation method

We assume the process is sampled on $\{0, \frac{1}{N}, \cdots, \frac{N-1}{N}, 1\}$ and B_H denote a path of fBm. A vector a of length $l+1$ is a filter of length l and of order p if it satisfies $\sum_{j=0}^{l} j^r a_j = 0$ and $\sum_{j=0}^{p} j^r a_j \neq 0$, where $0 \leqslant r < p$. Denote $V(\frac{i}{N}) = a * B_H = \sum_{q=0}^{l} a_q B_H(\frac{i-q}{N})$, namely V is the convolution of a and B_H. For example, $a = (1, -1)$ that is a filter with order 1, we have $V\left(\frac{i}{N}\right) = B_H\left(\frac{i+1}{N}\right) - B_H\left(\frac{i}{N}\right)$, a discrete variation.

The empirical k-th absolute moment of discrete variations is defined by

$$S_N = \frac{1}{N-l} \sum_{i=l}^{N-1} \| V\left(\frac{i}{N}\right) \|^k$$

Let π_H denote the covariance function of V, given by $\pi_H(j) = \mathbb{E}[V((i+j)/N)V(i/N)]$. From the stationarity of V, we have

$$\mathbb{E}(S_N) = \frac{1}{N^{kH}} \pi_H(0) E_k$$

where $E_k = 2^{k/2}\Gamma(k+1/2)\Gamma(1/2)$.
Then we denote function $g(t) = \frac{1}{N^{kt}}\pi_H(0)E_k$. Using S_N as an estimator for $\mathbb{E}(S_N)$, we obtain the estimator:

$$\hat{H} = g^{-1}(S_N)$$

Remark 3.1. *In fact, the behavior of this estimator in terms of different k has been studied by Coeurjolly[2] and the results underline the optimality of the value $k = 2$ since it has smaller variance. In the following work, we always use $k = 2$.*

In practice, it is not possible to get an analytical expression for $\hat{H}$ when $p > 1$. We use a numerical procedure to get it by minimizing $g(t) - S_N$. It is proven the result that [2]:

$$\hat{H}_N \xrightarrow{a.s.} H. \quad as \quad N \to +\infty$$

In general, the process we want to model is not a pure fBm, it is modulated by an scale coefficient. We define the data vector $D(i) = C \cdot B_H\left(\frac{i}{N}\right)$ for $i = 0, \cdots, N-1$, where $B_H\left(\frac{i}{N}\right)$ is a process as the one estimated before.

Firstly, we introduce new filter a^m which is defined by:

$$a_i^m = \begin{cases} a_j, i = jm \\ 0, otherwise. \end{cases}$$

In the case $m = 2$ and $a = (1, -2, 1)$, we get $a^m = (1, 0, -2, 0, 1)$.
Then, the second empirical moment of the corresponding discrete variations is

$$S_N(a^m) = \frac{1}{N - ml} \sum_{i=l}^{N-1} \| V^{a^m}\left(\frac{i}{N}\right) \|^2$$

where we defined $V^{a^m} = a^m * D$.
Noting that:

$$\mathbb{E}S_N(a^m) = m^{2H}\mathbb{E}S_N(a)$$

We get that $\log \mathbb{E}(S_N(a^m))$ is linear in H. Then we use the regression method to estimate H.

3.3 Different filters for standard model and its robustness

The result that is proven is independent of the filter that is used, so the question in practice is that if there is a filter that is more suitable than others.

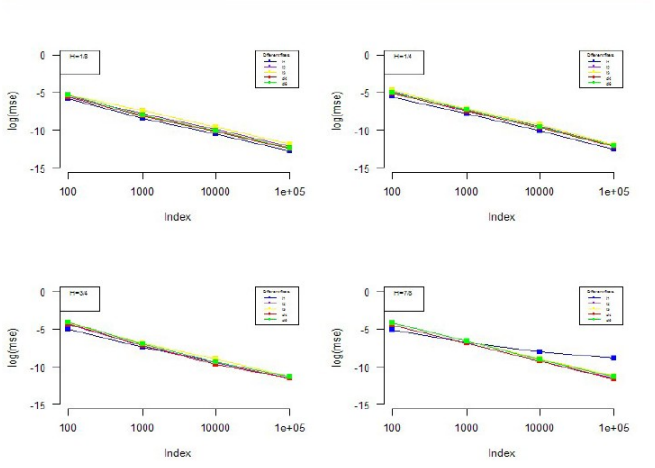

Figure 2: log(mean square error) and different data size corresponding to different Hurst parameter

We use several kinds of filters:
$i1 = (1, -1)$
$i2 = (1, -2, 1)$
$i3 = (1, -3, 3, -1)$
$d4 = (-0.09150635, -0.15849365, 0.59150635, -0.34150635)$
$d6 = (0.02490875, 0.060416105, -0.9546721, -0.3251825, 0.57055846, -0.235232605)$
where i-th are the increment filter of order i. $d4$ and $d6$ are Daubechies wavelet filters of order 4 and 6 respectively.

Effect of the use of different filters is on the same data set. Although $i1$ works only when $H < \frac{3}{4}$ [6], the filter $i1$ has been chosen as the Hurst parameter we get from real data is less than $\frac{3}{4}$ and the error is the smallest one. When we calculate the boundary cells, we need to change filter and $d4$ can be used that is good as well from our result.

3.4 Robustness of this model

The robustness of this model and its convergence has been tested on the literature on the presence of an additive gaussian noise and also on the presence of outliers [4][6]. The convergence is kept unchanged in the $N \to \infty$ limit, and is verified computationally to keep very well in practice.
We studied the robustness of this model in terms of different filters. Three kinds of noisy model are discussed:

- Model with additive outliers
  the model used here comes from [4](Beran (1994))

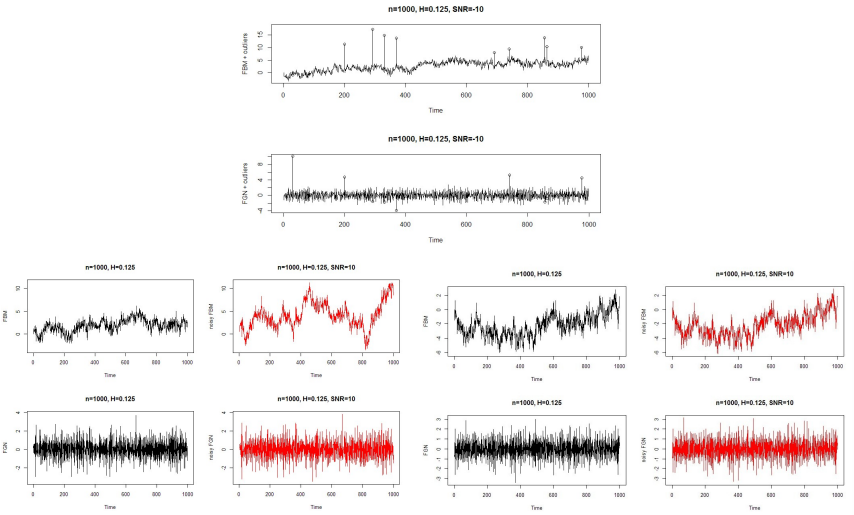

Figure 3: model with additive outliers B^0 andB^1

- Model with additive noise
  $X(i) = B_H(i) + \sigma B^0(i)$
  $X(i) = B_H(i) + \sigma B^1(i)$
  where B^0 is a standard Brownian motion and B^1 are i.i.d standard Gaussian variables.

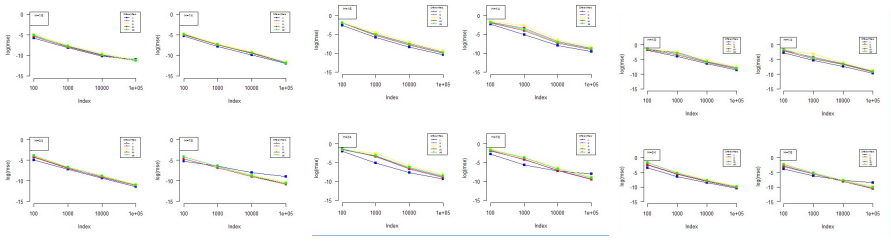

Figure 4: different filters for model with noise AO, B^0 andB^1

The result [Figure 4] shows in the model with additive noise this method works well. Due to the high presence of noise in our data, the robustness is a very important property.

4 Application to microscopy data

When using the data coming from the microscopy of the embryo, many difficulties appear. The first point to be noted is that the quantity of data that is available for one single cell is very limited, which restrains a lot the quality of results to be obtained. This is one of the principal reasons for the deep study on different methods of estimation, since we need as much quality on the analysis as possible. In a more practical sense, there are problems of identification of trajectories. For

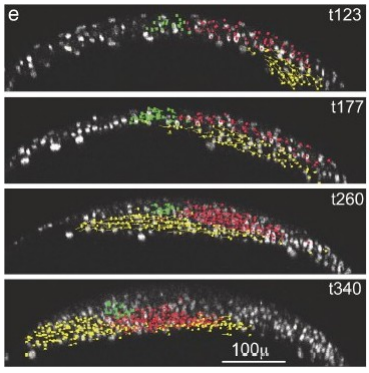

Figure 5: Example of the selection made by the biologists on the BioEmergences platform.

a given pair of cells, it is necessary to find the exactly the time interval where

both cells are alive, that is, from the moment they are born to the moment where they divide, which can be computationally costly.

Another point that is worth noting is that even if the tracking quality is very high (going up until 99% of precision for each time step), any error in the tracking costs a lot, as this may interfere with our first assumption, that says that the two cells that are being analyzed follow the same hydrodynamics flow.

In this first study of segmentation by Hurst parameter estimation, that data that is going to be used is a gold quality one that has been verified by hand for the tracking, minimizing the errors coming from this source. Moreover, some families of cells have been marked as corresponding to given fields by eye by a biologist, which give a parameter to compare to when finding the results.

4.1 Methodology and results

The fields we select are colored in red and yellow[Figure 5]. We select two sets of the cells in the same field and each set has five cells. We calculate the Hurst parameters of each pair and average them. The data size we used is around 50. From the result of Figure 6, in the field 21 the average Hurst parameters are 0.469

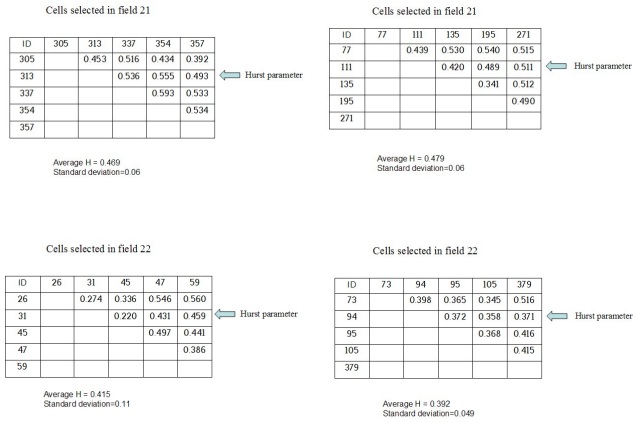

Figure 6: average Hurst parameter of two sets of cells in two fields

and 0.479, that is, they are very similar. Inside field 22 the values are 0.392 and 0.415, very similar between them, but different enough from the other field. This is a first indication that goes in the direction of our expectations. Then, we do more detailed work, calculating the Hurst parameter of a cell and its neighbors in different fields.. We used fields $1, 2, 6$ pre-selected by biologists in BioEmergences platform and calculate the Hurst parameters between the center cells and their neighbors. Since in this step, different neighbors have different lifetime steps, the

Timestep	101–150		151–200		201–250		251–300	
selection	H	N.of center cell	H	N.of center cell	H	N.of center cell	H	N.of center cell
2	0.3327	56	0.3232	49	0.2343	50	0.2596	30
6	0.3236	18	0.3215	21	0.3200	24	0.3183	24
1	0.5696	34	0.4403	44	0.3766	46	0.3351	24

Figure 7: Average Hurst parameter in three different fields in terms of different time steps

quantity of data is different for each pair of cells. For this, we use the classical average pondered by variance, given by:

$$H = \frac{\sum \frac{1}{\sigma_i^2} \hat{H}_i}{\sum \frac{1}{\sigma_i^2}}$$

and the fact that the algorithm gives a variance that is inversely proportional to the number of time steps that are used. This given in the end:

$$\hat{H} = \frac{\sum T_i H_i}{\sum T_i}$$

where T_i is the time steps we use when calculating H_i.

From the numerical results in Figure 7, we find H in fields $1, 2, 6$ are different. For field 1, it always has highest H that decreases in time. Field 2 has similar trend with field 1 in this period but H_2 is always less than H_1. The difference between them is around 0.13. In field 6, the Hurst parameter is stable, around 0.32. Thus we see that these measures we do are dependent on the moment of the lifetime of the embryo, and as such, this should be considered when doing the analysis.

Once more, we remark that the effect of hydrodynamical movement between two cells is always positive in the sense of H, so that if the dominant behavior of the movement between two cells is of this form, we should find values that are very close to 1. In order to test also this hypothesis, we calculate Hurst parameters of cells in the boundary of different fields and compare the difference between them [Figure 8].

What we find is that indeed, when we take two cells that are neighbors, the simple fact that they are in the same field or not change a lot, and cells in different fields always have Hurst parameters that are very high. This is the kind of reasoning that can allow us to find the boundary between two fields using Hurst parameters.

Cell in field 1 ID center 124569305		Cell in field 1 ID center 124571459	
H of cell and its neighbors in field 1	H of cell and its neighbors in field 6	H of cell and its neighbors in field 1	H of cell and its neighbors in field 6
0.5904	0.7871	0.4774	0.8864
0.5678	0.7830	0.4327	0.8673
0.41583	0.7781	0.5058	0.8579
	0.8042	0.6149	0.8960
	0.7639	0.4431	0.8803
		0.3921	
		0.3203	

Figure 8: Comparison with H of cells in same field and in different fields

5 Conclusion

In this last part or the report, we showed that the analysis of Hurst parameter can be used as one important measure to distinguish different morphogenetic fields. Of course, this is a first evidence of the efficiency of this method, and does not allow yet the effective segmentation of the embryo. For this goal, a further study about the establishment of thresholds for the distinction of cells in the same and in different fields. An important point about this methodology is that it is independent of the subtract where it acts. That means that it can be applied to different embryos without being changed. While we can expect that this same application will work for other zebra fishes, it is difficult to foresee the results for other animals.

A natural way to continue this work would be to obtain the whole map of the embryo, with the representation of Hurst parameters point by point. This would already give more precise results, as the quantity of cells would improve and the there would be the possibility of using larger ranges of neighbors. With this map, it would be possible to work together with biologists in order to find a way to get thresholds that are connected with their expectations.

References

[1] Peltier R F, Lévy-Véhel J. A new method for estimating the parameter of fractional Brownian motion[J]. *Rapport de recherche-institut national de recherche en informatique et en automatique,1994.*

[2] Coeurjolly J F. Estimating the parameters of a fractional Brownian motion by discrete variations of its sample paths[J]. *Statistical Inference for stochastic processes, 2001, 4(2): 199-227.*

[3] Jean-Francois C. Simulation and identification of the fractional Brownian motion: abibliographical and comparative study[J]. *Journal of statistical software, 2000, 5: 1-53.*

[4] Beran J. Statistics for long-memory processes[M]. *Chapman Hall/CRC, 1994.*

[5] Coeurjolly J F. Hurst exponent estimation of locally self-similar Gaussian processes using sample quantiles[J]. *The Annals of Statistics, 2008, 36(3): 1404-1434.*

[6] Achard S, Coeurjolly J F. Discrete variations of the fractional Brownian motion in the presence of outliers and an additive noise[J]. *Statistics Surveys, 2010, 4: 117-147.*

[7] Clegg R G. A practical guide to measuring the Hurst parameter[J]. *arXiv preprint math/0610756, 2006.*

[8] Dieker T. Simulation of fractional Brownian motion[J]. *MSc theses, University of Twente, Amsterdam, The Netherlands, 2004.*