Cell Trajectory Clustering: Towards the Automated Identification of Morphogenetic Fields in Animal Embryogenesis

Juan Raphael Diaz Simões[1,2], Paul Bourgine[1,3], Denis Grebenkov[2] and Nadine Peyriéras[1]
[1]*BioEmergences USR3695, CNRS, Université Paris-Saclay, 91198 Gif-sur-Yvette Cedex, France*
[2]*CNRS, Université Paris-Saclay, Route de Saclay, 92128 Palaiseau Cedex, France*
[3]*Complex Systems Institute Paris Île-de-France UPS3611, CNRS, 113 rue Nationale, 75013 Paris, France*

Keywords: Animal Embryogenesis, Cell Lineage, Clustering, Path Integrals.

Abstract: The recent availability of complete cell lineages from live imaging data opens the way to novel methodologies for the automated analysis of cell dynamics in animal embryogenesis. We propose a method for the calculation of measure-based dissimilarities between cells. These dissimilarity measures allow the use of clustering algorithms for the inference of time-persistent patterns. The method is applied to the digital cell lineages reconstructed from live zebrafish embryos imaged from 6 to 13 hours post fertilization. We show that the position and velocity of cells are sufficient to identify relevant morphological features including bilateral symmetry and coherent cell domains. The method is flexible enough to readily integrate larger sets of measures opening the way to the automated identification of morphogenetic fields.

1 INTRODUCTION

Embryogenesis involves the formation of boundaries between cell populations, which leads to the individuation of cell compartments (Fagotto, 2014). Cells gathered in the same compartment are expected to have behavioral similarities, depending on the animal species and the spatio temporal morphogenetic sequence. The formation of morphogenetic fields and compartments being a progressive phenomenon, similarities in cell behaviors should be present at early stages. Furthermore, behavioral similarities should be passed from mother to daughters along the cell lineage.

In this context, the proposed similarity (or dissimilarity) measure of cell behavior should have the following properties:

- It should be flexible enough to deal with different sets of *measures*, different animal species and different periods of the organism development.
- It should consider the cell as a structure with a *temporal* coherence.
- It should take into account the relative persistence of cell characteristics along their *lineage*.

We propose a method to measure dissimilarities between cells based on their behaviors along the cell lineage. The dissimilarity calculation method is inspired from path integrals in quantum mechanics and stochastic processes (Feynman, 1965) and is valid for measures taken at the level of single cells throughout the duration of their cell cycle.

This measure is further used for the *automatic clustering of cell trajectories*.

The method is applied to the analysis of digital cell lineage trees reconstructed from live zebrafish embryos imaged in 3D+time. We show that a minimal set of parameters can lead to clusters highlighting morphogenetic features from the organism bilateral symmetry to finer cellular domains consistent with the cell compartments that shape the presumptive organs.

2 ALGORITHM DESCRIPTION

The algorithm can be decomposed into the following steps:

1. Extracting measures from input data. In the case presented here, the data consisted in the cell lineage reconstructed from partial 3D+time imaging of developing zebrafish embryos (Faure et al., 2016). The digital cell lineage provides the cell positions in space and time. Velocities are calculated using a discrete derivative similar to low-pass filters (Holoborodko, 2008). This choice is justified by the fact that the noise on the trajectories is unknown;

Simões, J., Bourgine, P., Grebenkov, D. and Peyriéras, N.
Cell Trajectory Clustering: Towards the Automated Identification of Morphogenetic Fields in Animal Embryogenesis.
DOI: 10.5220/0006259407460752
In *Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2017)*, pages 746-752
ISBN: 978-989-758-222-6

2. Calculating the *genealogic dissimilarities* between cell trajectories. This step is described in detail in the following sections;
3. Using these dissimilarities as input for a clustering algorithm. We used *spectral clustering* (Ng et al., 2001) since it is easy to implement, efficient enough to allow the test of parameters and works with sparse dissimilarity matrices (von Luxburg, 2007).

In the following sections we present the steps for the calculation of genealogic dissimilarities, comment on some practical aspects of parameters and efficiency and discuss the results. The overall methodology is schematized in Fig. 1.

3 TRAJECTORIES, GENEALOGIES AND DISSIMILARITIES

The data sets used here are *partial*, since the microscope images only cover a limited region of the embryo. This leads to the following division of a cell lifetime. We define the *lifetime of a cell* as the interval of time between its birth at the time of it's mother's division or its entrance into the field of view and its division or exit from the field of view. From this definition mother cells and daughters are different entities, connected by the cell lineage.

In addition, cells can only be compared at the same point in time as their behavior changes during the embryogenesis. Our algorithm is made to overcome this difficulty and compare cell behaviors throughout an entire relevant developmental period by extending the cell lifetime period by that of its progeny through a probabilistic approach.

3.1 Defining Cell Rajectories

Given a totally ordered time set T, which may be discrete or continuous, and a state set S of measures, we define a *trajectory* to be a dependent tuple (I,x), where $I =]I_-, I_+]$ is an interval in T and $x : I \to S$ is a function into the state space. We call the space of trajectories in time T and state S by $\mathcal{T}(T,S)$.

3.2 An Algebra for Trajectories

The calculation of dissimilarities between trajectories involve the intersection and concatenation of trajectories, the fundamental operations of the method. However, because it's not true that every pair of cells share a point in time where they coexist, we will need to introduce the notion of undefined values to deal with these cases.

We define a *possibly undefined value of type* X to be an element of the set $\overline{X} = X \cup \{\mathbb{U}\}$ where the *undefined* value $\mathbb{U}$ has been added. Any function $f : X \to Y$ can be lifted to a function $\overline{f} : \overline{X} \to \overline{Y}$ by

$$\begin{aligned} \overline{f}(\mathbb{U}) &= \mathbb{U} \\ \overline{f}(x) &= f(x) \end{aligned}$$

We use $X = (I,x)$ and $Y = (J,y)$ as example trajectories. We define *trajectory concatenation* $\vee$ for trajectories satisfying $I_+ = J_-$ by

$$X \vee Y = (I \cup J, x \vee y)$$

where

$$[x \vee y](t) = \begin{cases} t \in I & \Rightarrow x(t) \\ t \in J & \Rightarrow y(t) \end{cases}$$

This operation can be lifted to $\overline{\mathcal{T}(T,S)}$, by making $\mathbb{U}$ a unity

$$X \overline{\vee} Y = \begin{cases} X = \mathbb{U} & \Rightarrow Y \\ Y = \mathbb{U} & \Rightarrow X \\ \text{otherwise} & \Rightarrow X \vee Y \end{cases}$$

We define moreover *trajectory intersection* $\wedge : \mathcal{T}(T,S) \times \mathcal{T}(T,S) \to \overline{\mathcal{T}(T,S^2)}$ by

$$X \wedge Y = \begin{cases} I \cap J \neq \emptyset & \Rightarrow (I \cap J, (x,y)|_{I \cap J}) \\ \text{otherwise} & \Rightarrow \mathbb{U} \end{cases}$$

An example is shown in Fig. 2. This operation can be lifted to $\overline{\mathcal{T}(T,S)}$ by defining $\mathbb{U}$ to be a zero

$$X \overline{\wedge} Y = \begin{cases} X = \mathbb{U} & \Rightarrow \mathbb{U} \\ Y = \mathbb{U} & \Rightarrow \mathbb{U} \\ \text{otherwise} & \Rightarrow X \wedge Y \end{cases}$$

With these definitions, it is not hard to prove the distributivity of intersection over concatenation

$$Z \overline{\wedge} (X \overline{\vee} Y) = (Z \overline{\wedge} X) \overline{\vee} (Z \overline{\wedge} Y) \tag{1}$$

In the following sections, we will always write $\wedge$ and $\vee$ instead of $\overline{\wedge}$ and $\overline{\vee}$ for easier reading, but all operations are defined with relation to the later.

3.3 Genealogic Trajectories

A *genealogic trajectory* is a cell trajectory that has been extended probabilistically by the trajectories of its progeny. The definition is recursive. For any trajectory X we define i_X as an equiprobable zero-one random variable. The genealogical trajectory $\widetilde{X}$ is a trajectory-valued random variable such that

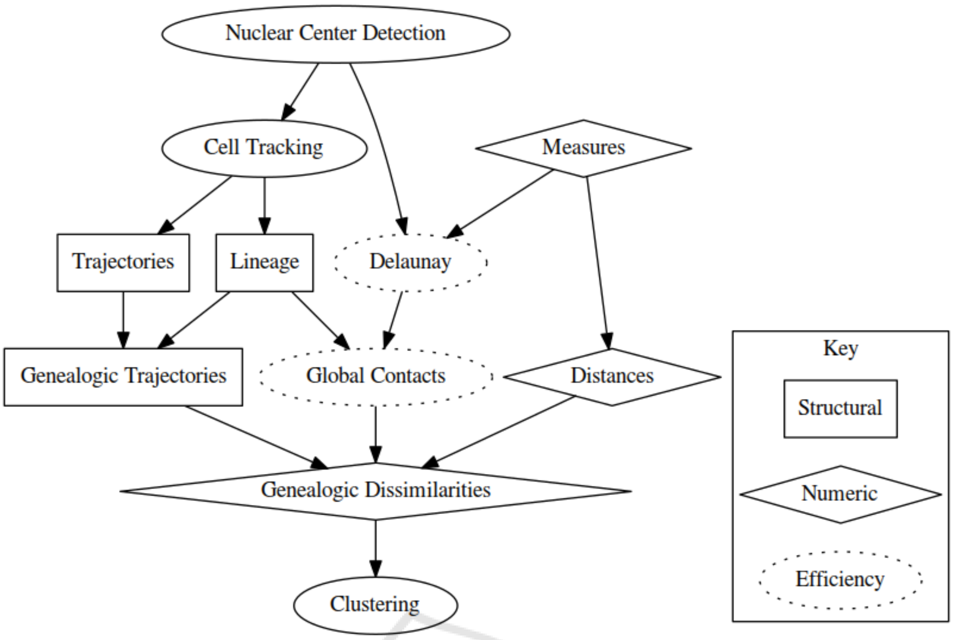

Figure 1: Scheme describing the steps of the algorithm. In square boxes are the steps necessary for the construction of the structural entities: the cell lineage and genealogical trajectories. In diamond-shaped boxes are the numerical entities, corresponding to measures, in this case position and velocity and the corresponding distances and dissimilarities. In dotted boxes are the entities necessary for the efficient calculation of the clustering, namely the definition of neighboring cells.

1. If X has no siblings, $\widetilde{X} = X$;
2. If X has siblings Y_0 and Y_1 then $\widetilde{X} = X \vee \widetilde{Y}_{i_X}$.

An example is shown in Fig. 2. Moreover, if $X \neq Y$ then i_X and i_Y assumed to be independent. This allows the simple calculation of expectations.

3.4 Dissimilarities and Weights

A dissimilarity in S is a symmetric positive function $d : S \times S \to \mathbb{R}$. Given the two trajectories $X = (I,x)$ and $Y = (J,y)$, $X \wedge Y = (K,z)$, a dissimilarity d in S and a measure μ in T, we define the *partial dissimilarity* $P_{\mu,d}$ between two trajectories as

$$P_{\mu,d}(X,Y) = \left(\mu(K), \int_K d \circ z \, d\mu\right)$$

The set $\mathbb{R}^2$ has a natural monoidal structure given by the sum of coordinates represented by $\oplus$. By giving the following monoidal structure to $\overline{\mathbb{R}}^2$

$$a \oplus b = \begin{cases} a = \mathbb{U} & \Rightarrow b \\ b = \mathbb{U} & \Rightarrow a \\ \text{otherwise} & \Rightarrow a \oplus b \end{cases}$$

We can prove that $P_{\mu,d}$ satisfies the following

$$P_{\mu,d}(X \vee Y) = P_{\mu,d}(X) \oplus P_{\mu,d}(Y) \qquad (2)$$

Finally, we define the *dissimilarity* $D_{\mu,d}$: $\mathcal{T}(T,S) \times \mathcal{T}(T,S) \to \overline{\mathbb{R}}$ as

$$D_{\mu,d}(X,Y) = \overline{\operatorname{div}}\, P_{\mu,d}(X \overline{\wedge} Y)$$

where $\operatorname{div}(m,s) = s/m$, considering $0/0 = \mathbb{U}$.

The *genealogical dissimilarity* between two trajectories is defined as

$$\widetilde{D}(X,Y) = \mathbb{E}[D(\widetilde{X},\widetilde{Y})] \qquad (3)$$

where $\mathbb{E}$ denotes expectation. This definition is analogous to path integrals in quantum mechanics.

Equation (2) gives a way to decompose genealogical dissimilarities into regular dissimilarities, meaning that genealogical dissimilarities can be calculated from the matrix of pairwise partial dissimilarities

$$P_{\mu,d}(Z \wedge (X \vee Y)) = P_{\mu,d}(Z \wedge X) \oplus P_{\mu,d}(Z \wedge Y)$$

Finally, we define the *similarity* $S_{\mu,d,\sigma} : \mathcal{T}(T,S) \times \mathcal{T}(T,S) \to \mathbb{R}$ as

$$S_{\mu,d,\sigma}(X,Y) = \exp\left[-\frac{\widetilde{D}_{\mu,d}(X,Y)^2}{2\sigma^2}\right] \qquad (4)$$

where σ is a scale factor.

3.5 Properties

Let us consider a simple lineage with a cell M, two daughters C_1 and C_2 and a cell N with no daughters.

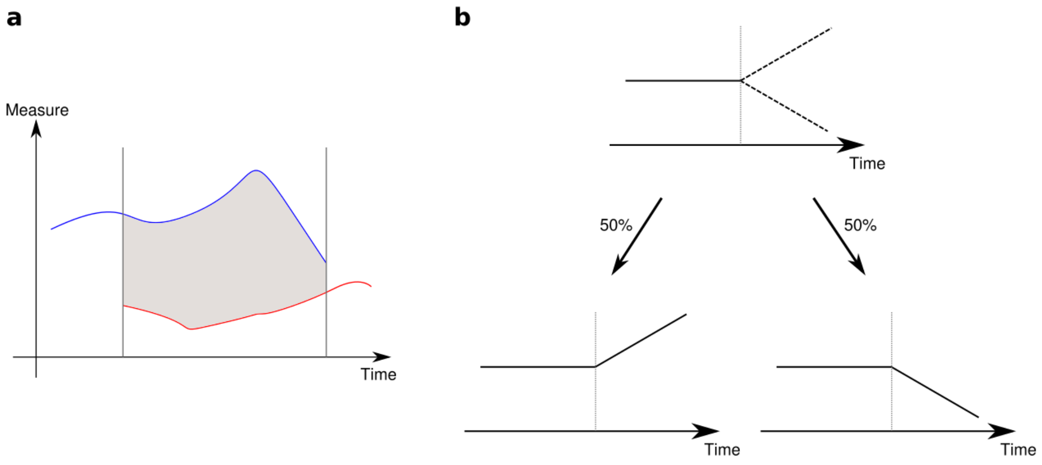

Figure 2: **a** Intersection of the trajectories of two cells (red and blue) represented in an abstract measure space. The starting point of each cell trajectory can be due to either cellular division of the mother or entrance of the cell into the imaging field of view. It may also result from artefacts of the tracking algorithm. Similarly, the ending point of each cell trajectory corresponds either its division, its exit out of the imaging field of view or to tracking errors. Only the difference of measures (grey) inside this interval are taken into account. **b** A schematic cell lineage with the lifetime if a mother cell (continuous line) and the lifetime of its two daughters (dashed lines). The probabilistic transformation of the mother history gives rise to two different paths with equal probabilities. This process is done recursively throughout consecutive cell divisions along the cell lineage.

If M and N do not coexist at any point in time, formula (1) states

$$\widetilde{D}_{\mu,d}(M,N) = \frac{\widetilde{D}_{\mu,d}(C_1,N) + \widetilde{D}_{\mu,d}(C_2,N)}{2}$$

In particular

$$\widetilde{D}_{\mu,d}(M,C_1) = \frac{\widetilde{D}_{\mu,d}(C_1,C_2)}{2} = \widetilde{D}_{\mu,d}(M,C_2) \qquad (5)$$

This equation shows that the mother is equidistant to its daughters and that if both daughters stay close to each other, their mother stays close to both of them, leading to a degree of coherence between their clustering allocations.

Another important property is given by the application of equation (2), adding a point in time for Z. This can be written as

$$P_{\mu,d}(X \wedge (Z \vee Z')) = P_{\mu,d}(X \wedge Z) \oplus P_{\mu,d}(X \wedge Z')$$

meaning that the partial dissimilarity can be processed incrementally in time.

3.6 Definition of Similarities

If the state space has the form $\prod_k S_k$, where each coordinate has a corresponding dissimilarity d_k, any term of the form

$$d_\lambda = \sum_k \lambda_k d_k$$

is also a dissimilarity on S, where every $\lambda_k > 0$. If s_k is the similarity associated to d_k then

$$s_\lambda = \sum_k \lambda_k s_k$$

is also a similarity and this is the definition used in this article. In the case of two dissimilarities that gives:

$$s_\lambda = \lambda \exp\left(-\frac{d_1^2}{2\sigma_1^2}\right) + (1-\lambda)\exp\left(-\frac{d_2^2}{2\sigma_1^2}\right)$$

In order to have a good equilibrium between similarities, we choose σ_k to be equal to the standard deviation of d_k.

4 COMPUTATIONAL EFFICIENCY

Following the regular clustering algorithm, we need to provide the whole similarity matrix, which gives a complexity of $O(n^2)$ on the number of trajectories. Moreover, because the matrix is dense, extracting a even few eigenpairs from it is very costly.

A common method of simplification is to use only the few largest similarities, making the matrix sparse and its largests eigenvectors fast to calculate. However, the whole matrix still has to be calculated.

We propose an alternative simplification by calculating the similarities between neighbor trajectories only. Two trajectories are considered neighbors if

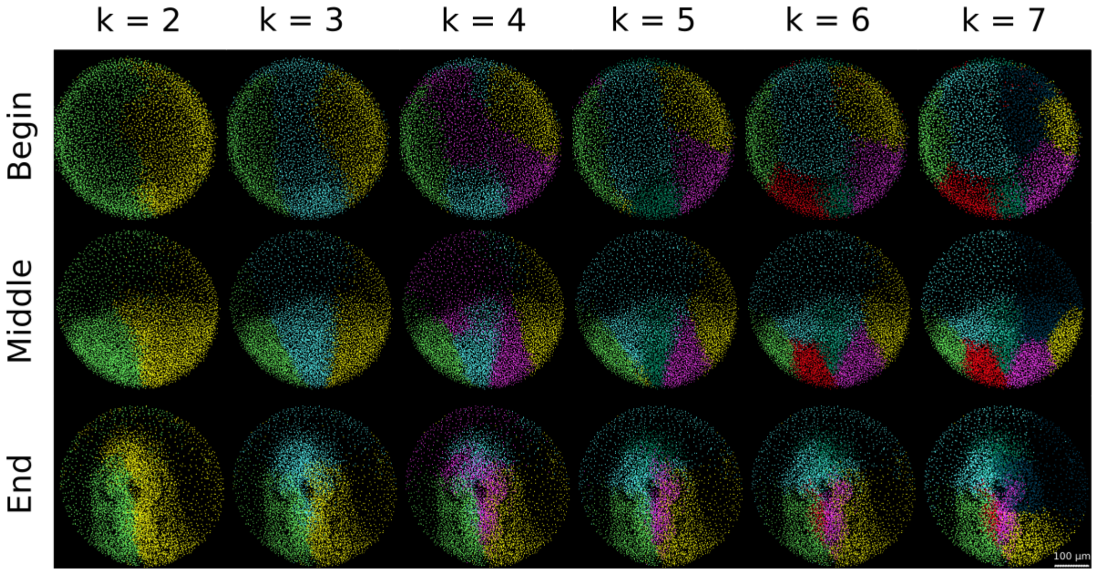

Figure 3: **Patterns defined by the clustering algorithm** Processing of the dataset 141108aF with $\lambda = 0.5$. 3D rendering of the embryo displayed with dots corresponding to detected nuclei. The different clusters are distinguished by their color chosen arbitrarily. The chosen number of clusters varies from 2 to 7 displayed in the different columns. The embryo is observed from the animal pole, anterior to the top at 6h54 (Begin - first row), 9h31 (Middle - second row) and 12h10 hpf (End - third row). Scale bar 100 μm.

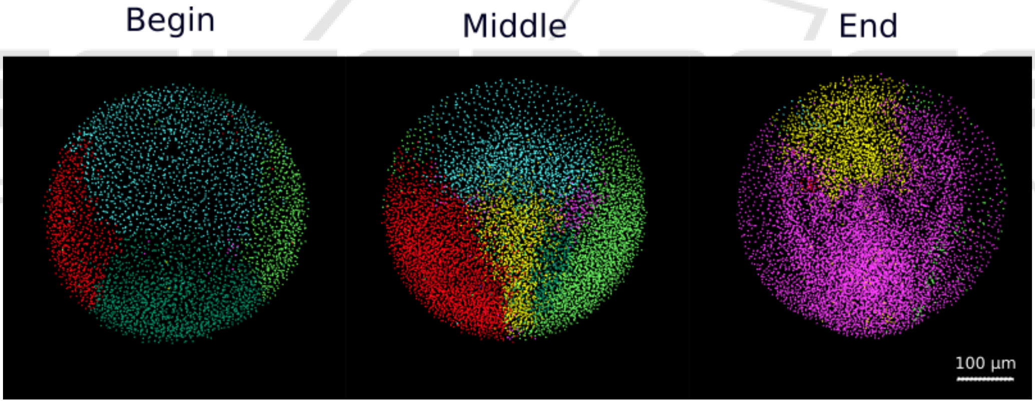

Figure 4: **Example of the artifact of temporal variability of the segmentation patterns** Processing of dataset 071226a with $\lambda = 0.5$ and $k = 6$. 3D rendering as in Figure 3. The embryo is observed from the animal pole, anterior to the top from 6h54 (begin), 9h31 (middle) and 12h10 hpf (end). Scale bar 100 μm. The flow of cells into and out of the imaged volume breaks the coherence of the cell lineage and consequently the spatial and temporal coherence of the patterns identified by our method.

they are neighbors one time step at least, in either position of velocity (or any other chosen euclidean measure). The concept of neighborhood is based on the Delaunay tesselation performed at each time step and for every measure.

A typical data set encompassing zebrafish embryonic development from 6hpf to 13 hpf imaged from the animal pole (Faure et al., 2016) has around 350 time steps, 6000 cells per time step and altogether 100000 cell trajectories. Given these numbers, the tradeoff between the complexity added by the tesselation at each time step and a global dense matrix is very positive. The original algorithm would process only very small data sets, while the improved version processes typical data sets within a few hours on a standard desktop computer.

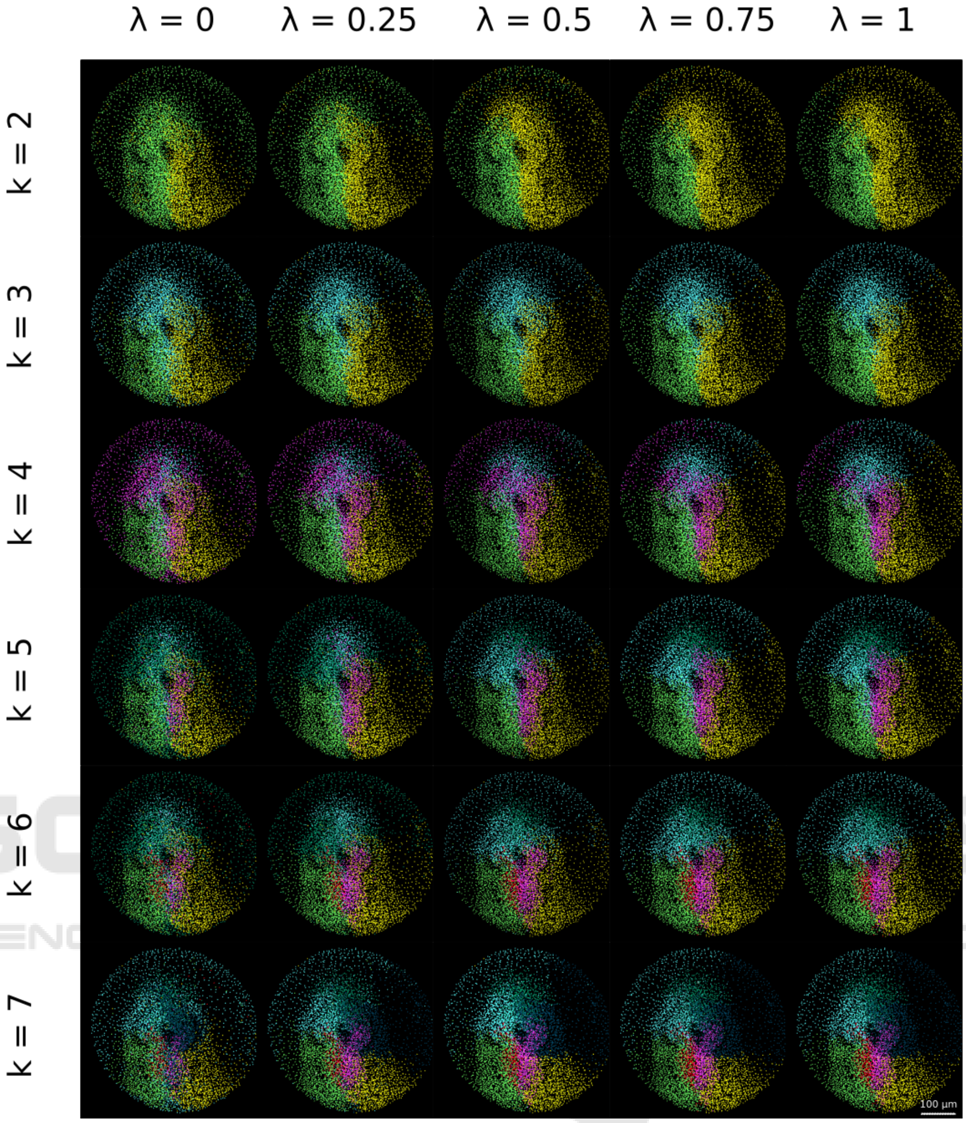

Figure 5: **Parameter space exploration for the specimen 141108aF**. The patterns obtained with the clustering algorithm when varying λ and k are displayed. Each row corresponds to the chosen number of clusters varying from 2 to 7. Different values of λ corresponding to the relative weight of cell position and velocity are displayed in columns, with the relative weight of position increasing from left to right.

5 RESULTS : APPLICATION TO ZEBRAFISH DIGITAL LINEAGE TREES

The relevance of the method to the analyzis of digital cell lineage trees has been assessed on two datasets corresponding to wild type zebrafish embryos identified as 141108aF and 071226a, developing over 8 hours encompassing gastrulation stages. These details have been described in detail in (Faure et al., 2016). The application of our algorithm to 141108aF (Fig. 3) shows that for a number of clusters equal to 2, the embryo is segmented through its bilateral symmetry plane. This appears as an interesting emerging feature, not being imposed at all by the method. Furthermore, when increasing the number of clusters, embryonic tissues are segmented in visually coherent domains.

The results of the algorithm are very sensitive to the persistence of cells inside the imaged volume, as shown by the results for 071226a (Fig. 4). As the imaged volume encompasses only the animal top of the embryo, there is an extensive flow of cells into and out of the imaged volume that creates a temporal segmentation that blurs the coherence the spatial organization that we expect to characterize. This limitation comes from the data, not from the algorithm, and would be solved if the data encompassed the whole embryo. Artifacts due to the incompleteness of the imaging data also compromise the possibility to quantify interindividual differences.

The algorithm can be tuned by exploring the parameter space defined by λ and *k* (Fig. 5), λ corresponding to the relative weight given to the different measures and *k* to the number of clusters. We observed, as expected, that privileging cell position over cell velocity led to more compact patterns and domains with sharper boundaries.

6 CONCLUSION

Our algorithm for measuring cell behavior similarity based on cells' trajectory clustering takes into account an arbitrary number of measures derived from digital cell lineage trees and translates them into coherent cell groups.

The major limit identified in the study come from the incompleteness of the imaging data. It should however be noted that taking into account an even larger number of cells would bring other difficulties.

The results that reveal coherent domains based on cell behavior similarity are meaningful for the biologist as they automatically reveal morphological landmarks. A next step will be to systematically compare the obtained patterns with patterns defined otherwise, such as gene expression patterns or fate maps. Our algorithm is expected to be a valuable addition to the growing toolbox for algorithmically augmented observation and the analysis of animal embryonic morphogenesis, opening new paths of research.

REFERENCES

Fagotto, F. (2014). The cellular basis of tissue separation. *Development 141, 3303-3318.*

Faure, E. et al. (2016). A workflow to process 3d+time microscopy images of developing organisms and reconstruct their cell lineage. *Nature Communications 2016 25 Feb; 7:8674.*

Feynman, R. P.; Hibbs, A. R. (1965). *Quantum Mechanics and Path Integrals*. New York: McGraw-Hill.

Holoborodko, P. (2008). Smooth noise robust differentiators. http://www.holoborodko.com/ pavel/numerical-methods/numerical-derivative/smooth-low-noise-differentiators/.

Ng, A. Y., Jordan, M. I., and Weiss, Y. (2001). On spectral clustering: Analysis and an algorithm. In *Advances In Neural Information Processing Systems*, pages 849–856. MIT Press.

von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing, 17 (4).*