Randomness and variability in animal embryogenesis, a multi-scale approach

Paul Villoutreix

UNIVERSITÉ PARIS DESCARTES

École doctorale Frontières du Vivant

Aléatoire et variabilité dans l'embryogenèse animale, Une approche multi-échelle

Par Paul Villoutreix

Thèse de doctorat de Biologie Mathématique

Dirigée par Giuseppe Longo et Nadine Peyriéras

Présentée et soutenue publiquement le 3 Juillet 2015

Devant un jury composé de :

Stéphane DOUADY	Rapporteur	CNRS, Paris VII
Sylvie MAZAN	Rapporteure	CNRS, UPMC
Franck VARENNE	Examinateur	CNRS, Université de Rouen
David MCCLAY	Examinateur	Duke University
Giuseppe LONGO	Directeur de thèse	CNRS, ENS
Nadine PEYRIÉRAS	Directrice de thèse	CNRS

PARIS DESCARTES UNIVERSITY
"Frontières du Vivant" PhD program

Randomness and variability in animal embryogenesis, A multi-scale approach

PhD Thesis in Mathematical Biology

PAUL VILLOUTREIX

Prepared under the joint supervision of
Giuseppe LONGO and Nadine PEYRIÉRAS

Public defense - July 3rd, 2015 - Examining committee

Stéphane DOUADY	Referee	CNRS, Paris VII
Sylvie MAZAN	Referee	CNRS, UPMC
Franck VARENNE	Member	CNRS, Université de Rouen
David MCCLAY	Member	Duke University
Giuseppe LONGO	Thesis advisor	CNRS, ENS
Nadine PEYRIÉRAS	Thesis advisor	CNRS

Abstract

We propose in this thesis to characterize variability quantitatively at various scales during embryogenesis. We use a combination of mathematical models and experimental results

In the first part, we use a small cohort of digital sea urchin embryos to construct a prototypical representation of the cell lineage, which relates individual cell features with embryo-level dynamics. This multi-level data-driven probabilistic model relies on symmetries of the embryo and known cell types, which provide a generic coarse-grained level of observation for distributions of individual cell features. The prototype is defined as the centroid of the cohort in the corresponding statistical manifold. Among several results, we show that intra-individual variability is involved in the reproducibility of the developmental process.

In the second part, we consider the mechanisms sources of variability during development and their relations to evolution. Building on experimental results showing variable phenotypic expression and incomplete penetrance in a zebrafish mutant line, we propose a clarification of the various levels of biological variability using a formal analogy with quantum mechanics mathematical framework. Surprisingly, we find a formal analogy between quantum entanglement and Mendel's idealized scheme of inheritance.

In the third part, we study biological organization and its relations to developmental paths. By adapting the tools of algebraic topology, we compute invariants of the network of cellular contacts extracted from confocal microscopy images of epithelia from different species and genetic backgrounds. In particular, we show the influence of individual histories on the spatial distribution of cells in epithelial tissues.

Résumé

Nous proposons dans cette thèse de caractériser quantitativement la variabilité à différentes échelles au cours de l'embryogenèse. Pour ce faire, nous utilisons une combinaison de modèles mathématiques et de résultats expérimentaux.

Dans la première partie, nous utilisons une petite cohorte d'oursins digitaux pour construire une représentation prototypique du lignage cellulaire, reliant les caractéristiques des cellules individuelles avec les dynamiques à l'échelle de l'embryon tout entier. Ce modèle probabiliste multi-niveau et empirique repose sur les symétries des embryons et sur les identités cellulaires; cela permet d'identifier un niveau de granularité générique pour observer les distributions de caractéristiques cellulaires individuelles. Le prototype est défini comme le barycentre de la cohorte dans la variété statistique correspondante. Parmi plusieurs résultats, nous montrons que la variabilité intra-individuelle est impliquée dans la reproductibilité du développement embryonnaire.

Dans la seconde partie, nous considérons les mécanismes sources de variabilité au cours du développement et leurs relations à l'évolution. En nous appuyant sur des résultats expérimentaux montrant une pénétrance incomplète et une expressivité variable de phénotype dans une lignée mutante du poisson zèbre, nous proposons une clarification des différents niveaux de variabilité biologique reposant sur une analogie formelle avec le cadre mathématique de la mécanique quantique. Nous trouvons notamment une analogie formelle entre l'intrication quantique et le schéma Mendélien de transmission héréditaire.

Dans la troisième partie, nous étudions l'organisation biologique et ses relations aux trajectoires développementales. En adaptant les outils de la topologie algébrique, nous caractérisons des invariants du réseaux de contacts cellulaires extrait d'images de microscopie confocale d'épithéliums de différentes espèces et de différents fonds génétiques. En particulier, nous montrons l'influence des histoires individuelles sur la distribution spatiales des cellules dans un tissu épithélial.

Remerciements

Il serait surprenant qu'une thèse sur l'aléatoire ne comporte pas une composante de hasard, celle-ci a bénéficié de nombreuses rencontres plus ou moins fortuites. Ce sont toutes ces personnes que je souhaite remercier ici.

Mes remerciements vont en premier lieu à mes directeurs de thèse Giuseppe Longo et Nadine Peyriéras. Je les remercie de m'avoir soutenu tout au long de ces années de thèse. L'originalité de leurs travaux, leur ouverture d'esprit, leur énergie ont nourri ma curiosité scientifique et m'ont emmené vers des domaines nouveaux pour moi en biologie, en mathématiques, en philosophie.. Je pense en particulier à toutes les rencontres et discussions passionnantes ayant eu lieu au cours des réunions CIM organisées par Giuseppe. Je pense aussi à l'exploration permanente de Nadine sa capacité à s'enthousiasmer et sa profonde réflexion. Je les remercie de m'avoir offert autant de liberté.

Je souhaite remercier plusieurs chercheurs qui ont nourri mon travail au fil des directions où mes recherches m'ont mené. Paul Bourgine pour ses intuitions, son enthousiasme et sa curiosité insatiable, mon travail sur les oursins lui doit beaucoup. Gunnar Carlsson pour sa disponibilité, sa vivacité et pour m'avoir ouvert de nombreuses portes en mathématiques. Monica Nicolau pour m'avoir permis de passer plusieurs mois à l'université de Stanford. René Doursat pour sa patience face à mes notations parfois hasardeuses! Ses idées en modélisation des systèmes vivants et sa grande clarté et ouverture d'esprit sont très appréciables.

Les réunions avec les membres de mon comité de thèse ont été très bénéfiques. Je remercie particulièrement Franck Varenne pour ses encouragements répétés, Philippe Herbomel pour son sens biologique, Michel Morange pour ses perspectives historique, Michel Bitbol pour m'avoir conforté sur la piste d'une "interprétation quantique" de la variabilité chez les squints.

Cette thèse n'aurait pu avoir eu lieu dans une autre école doctorale que Frontières du Vivant. Je remercie François Taddei d'avoir insufflé cette énergie interdiscplinaire à un nombre toujours plus en grand d'étudiants. Merci à Ariel Lindner, Claire Ribrault, David Tareste et Annemiek Cornelissen de m'avoir initié à la recherche au cours du Master AIV, et aussi Vincent Fleury et Annemiek pour leurs cours passionnés d'embryologie physique. Merci à Gilles Fleury de m'avoir soutenu à Supélec vers cette trajectoire particulière.

Je remercie vivement les membres du jury; Stéphane Douady et Sophie Mazan d'avoir accepté d'être rapporteurs et Franck Varenne et David McClay d'avoir accepté d'être ex-

aminateurs.

Une thèse se fait aussi en équipe, merci à Louise pour m'avoir initié à l'embryologie expérimentale, Thierry pour ses inventions perpétuelles, Dimitri pour son sens de la discussion et ses jumeaux, Matthieu pour ses loutres et ses topos, Julien De. pour ses embryons virtuels. Merci à tous les membres de la plateforme BioEmergences, Monique, Amparo, Yannick, Adeline B, Julien Du., Adeline R, Mathieu B., Clovis, Adil, Sylvia, Mark, Gaëlle, Barbara, Manu.

Merci également à l'Institut des Systèmes Complexes d'avoir financé et hébergé la majorité de cette thèse. La diversité des personnes que l'ont peut y croiser est très appréciable. Je remercie notamment Elisa pour sa solidarité en fin de thèse et son accent italien, Jean-Philippe pour son chic et son ouverture, Guillaume pour son flegme et ses conseils avisés, Romain pour TuxKart, Mathieu L. pour son sens de l'improvisation, Fabien pour ses piétons et bien sûr toutes les personnes qu'on peut y rencontrer et qui en font un merveilleux lieu de travail, David, Laurence, Catherine, Marlène, Maud, Julie, Pierre, Mazyiar, Alexandre, Jean-Baptiste, Samuel, Salma, Wandé, Tam Kien et tous les autres..

Je suis très heureux d'avoir pu participer aux réunions CIM régulièrement organisées par Giuseppe et je remercie tous ceux qui font vivre ce groupe de recherche. Merci notamment à Maël pour avoir ouvert de nombreuses voies de recherches passionnantes et pour être toujours prêt à les partager, merci à Nicole pour ses perspectives kantiennes, merci à Matteo pour l'organisation du vivant et des carnavals, merci à Angelo pour ses éclairages philosophiques, merci à Ana et Carlos d'être toujours attentifs à la portée théorique des concepts biologiques.

Finalement merci à mes amis qui m'ont beaucoup entouré. Merci en particulier aux amis des Marsouins, Jehanne, Clément, Ariane, Antoine F. et Antoine D., Aleksandra, Jean, Aude, Tatiana, c'était pour moi un rendez-vous indispensable! Merci à Peva de m'avoir mis entre les mains le livre de Bailly et Longo lorsque je cherchais ma voie, merci à Linda pour le soutien et les encouragements pendant de nombreuses années, et bien sûr merci à ma famille qui a toujours été très présente pour moi.

Contents

General introduction

In the introduction of the *Origin of species* [49], Charles Darwin refers to the "Mystery of mysteries" for the question of how species evolve and replace each other. Almost two centuries later, this question remains of great importance and even if some major advances have been made in the understanding of the mechanisms underlying evolution, the principle of *descent with modification*, which constitutes together with the principle of natural selection the grounds of Darwinian evolution, lacks a full comprehension. As a contribution to this problem, this work adresses the question of variation during embryogenesis and development.

Multicellular organisms are the result of a morphogenetic process involving multiple levels of organization, from molecules to cells to tissues to organs, interacting in a complex manner. This complexity is witnessed by causal processes that can be bottom-up, for example from the molecular level to the cellular level through gene expression, and top-down, for example from the tissue level to the cellular level through mechanical constraints. Moreover, an organism builds and maintains itself, it is self-organized. This idea of a self-organizing nature of living organisms can be traced back to Immanuel Kant's *Critique of Judgment* [116]. He considered organisms as entities where "every part is thought as owing its presence to the agency of all the remaining parts, and also as existing for the sake of the others and of the whole". Self-organization is at the very basis of complex systems science [117] and requires specific methodological approaches for its understanding.

The development of an organism is a dynamical phenomenon where any event occurring at a given scale at a certain time is involved in the subsequent stages of development at all level of organizations. Therefore any variation occurring in this process has consequences on other parts of the organisms at later stages. This integrated nature of the organism and its relations to variation had already been noticed by Charles Darwin who designated it as *correlated variations* in the fifth chapter of the *Origin of species*:

Correlated Variations - I mean by this expression that the whole organization is

> so tied together during its growth and development, that when slight variations in any one part occur, and are accumulated through natural selection, other parts become modified.

The question remains as to how these correlated variations occur and shape the exploration and generation of diversity.

Understanding the diversity of forms in the living begins with the question of the sources of morphological differences between individual organisms. Usually, differences between individual organisms are attributed to genetic differences and variation in environmental conditions. Recent experimental results suggest that other phenomena have to be accounted for when considering sources of diversity. One striking example is the work of Raj et al. in 2010 [181], where clonal organisms of a mutant strain of the worm *Caenorhabditis Elegans* are grown in homogeneous environment and result in variable phenotypes. In this example the phenotypic variation is ascribed to stochastic gene expression. The concept of stochastic gene expression [65] covers the processes involved in the variability of quantity of the protein expressed for a given gene. Many other mechanisms at various levels of organization are able to generate diversity; they will be described in more depth in the course of this dissertation. This raises the question of how to integrate this variety of mechanisms generating variability when considering organisms as a whole.

Recent progresses in biological imaging technology and other quantitative techniques have allowed to reconsider many biological phenomena. The observation of stochastic gene expression in single cells [65] is a major example of the breakthroughs made possible with technological developments. In the field of developmental biology, the development of *in toto* and *in vivo* microscopy technology such as 2-photon microscopy, or single planar illumination microscopy, have opened new perspectives on the study of embryogenesis. In particular the complete digital reconstruction of the cell lineage during the first few hours of the zebrafish development has been made possible by joint innovations in imaging techniques and image processing [163]. This kind of phenomenological reconstruction provides information on developing organisms which had never been observed in a quantitative way before, leading to a reinterpretation of many of the processes. In addition to technical challenges, the very large datasets generated require new analytic methodologies to extract significant information [149]. In the meantime, a reinterpretation of the relations between data and models is needed, new paradigms such as data-driven models and hypothesis-driven models have emerged during the last decade [119].

These large sources of quantitative data at various scale obtained in living multicel-

lular organisms shed a new light on processes occurring during development. Phenomena that have been mostly qualitatively and verbally described can now be characterized quantitatively in order to understand underlying principles [208]. However, we largely lack concepts, methods and tools to use this data for deciphering biological complexity. On a theoretical and epistemological level, biological organisms are multi-scale objects involving several levels of organization that are usually described with heterogeneous theoretical frameworks [33]. A central characteristic of biological objects is their historical nature, they are the result of both ontogenetic and phylogenetic trajectories, defining them as historical entities; they may not be reproducible identically. Indeed, phylogenetic trajectories are the result of an interaction and a co-constitution of organisms and their environments involving single events and small numbers [86], [137], similarly ontogenetic trajectories can be considered as a sequence of symmetry breakings involving contingent events [135], [15]. This historicity and variability of biological objects is at the center of Darwinian theory of evolution and should be at the ground of our understanding of biological organization. This is an epistemological specificity of biology contrasting with the *ahistoricity* of most physical objects. On this aspect the following quote of the physicist Max Delbrück is particularly illuminating:

> The complex accomplishment of any one living cell is part and parcel of the [fact] that any one cell represents more an historical than a physical event. These complex things do not rise every day by spontaneous generation from the non-living matter - if they did, they would really be reproducible and timeless phenomena, comparable to the crystallization of a solution, and would belong to the subject matter of physics proper. No, any living cell carries with it the experiences of a billion years of experimentation by its ancestors. You cannot expect to explain so wise an old bird in a few simple words [1]

Following this line of thought, we will consistently discuss the relevance of the mathematical concepts used and transferred from one discipline to the other by considering their epistemological justification.

Given the *systemic* nature of organisms and the *variety of sources* of variation, as well as the *historicity* of organisms, it is natural to ask for the characteristics of variability during embryogenesis. What makes individual singular? How variation in individuals influence variability at the population level? What are the relations between variability and probabil-

1. Max Delbrück - "A Physicist Looks at Biology", Address Delivered at the Thousandth Meeting of the Academy, 1949

ities? And variability and randomness? How to measure and quantify variation at several scales during development? Given the differences between individual specimen, can we define a normal prototypical development for a species?

We support the following thesis: variations at all scales during development shape the exploration of diversity of forms, specifically designed mathematical tools are required to characterize and quantify these variations.

Characterizing normality In the first part, we will consider the concept of normality and reproducibility of development. The question will be to measure how similar and how different are embryonic developments of the same species in normal conditions. To this aim, we will use a data set of **five digitally reconstructed sea urchin *Paracentrotus lividus* embryos** at the single cell resolution. The specimens of this small cohort are developing from the 32 cells stage (4 hours post fertilization) to hatching (around 500 cells, 10 hours post fertilization). Using the BioEmergences worlkflow [70] their complete 3D+time cell lineage and the shape of each cell was obtained from 2-photon microscopy acquisitions. Using this large data set, we will investigate the different levels of variability. This variability is first witnessed within an organism among the cells and underlies cell differentiation, it is the intra-individual variability. This variability is then witnessed between specimen among the cohort and is the result of individual specific histories, it is the inter-individual variability. While intra-individual variability can be well characterized by considering distributions of cell features, inter-individual variability requires to establish generic comparable features allowing to place individual specimen on the same footing without averaging out significant intra-individual variability.

Similar dynamical patterns are found at the level of the whole embryo and in each morphogenetic field. These patterns concern the evolution of the number of cells, the cell surface and volume and the number of neighbors. A preliminary linear spatio-temporal scaling is however necessary to make them fully comparable among specimen of the cohort. The value of the coefficients defining this scaling are a first step to characterize inter-individual variability.

Variability in the distribution of individual cell features and symmetries of the embryos prevent to identify unique cells from one organism to the other. To compare embryos and characterize intra-individual variability, we define a generic coarse-grained level of observation based on inherent symmetries and similar fates (Mesomere, Macromeres and Large and Small Micromeres). The corresponding group of cells are clustered according to this

identity and generation. They form the unite of our study of variability in the sea urchin development. Cells are considered exchangeable within these groups: the description of the distribution of cell features is not affected by permutation of the cells. Therefore we rely on the **de Finetti's theorem** to guarantee the use of empirical probability distribution as a good descriptor of individual cell feature distribution within a group of cells [8], [45].

Approximated parametric probability distributions within groups of cells and approximated independency of the cell features distributions between groups of cells is the basis for **a multi-level data-driven probabilistic model of the cell lineage** reproducing embryo-level dynamics for each specimen from measures of individual cell features. The same structure relating individual cell features and embryo level dynamics is found in each specimen. Parameters of the probability laws estimated empirically define uniquely individual specimen. This structure serves as the basis for a **prototypical model of development** among the cohort representing invariant features while preserving intra-individual variability [186], [213]. We use the framework of Information Geometry [10], [159] to define the prototype as the Kullback-Leibler centroid in the associated statistical manifold enabling to obtain **a unique set of parameters** representing the cohort.

In addition to representing quantitatively every measured cellular processes, this model addresses formally and quantitatively the question of regulation in development and the concept of morphogenetic field. Moreover it can be used to characterize a notion of structural stability and irreversibility. Eventually, this multi-level data-driven probabilistic model will be employed as a basis for an hypothesis driven biomechanical model using the Mecagen modeling platform. This model enables a spatial embedding of the prototypical cell lineage, the values of biomechanical parameters are obtained by parameter exploration strategies and matching with empirical data leading to a phenotypic phase diagram [55], [186].

Overall, this work provides a picture of development where the reproducibility of embryo level dynamics emerges from variability at the individual cell level. This picture contrasts with the traditional view of the development as a finely tuned process. Chapter 1 describes this work at the broadest level, its reading doesn't require a strong mathematical background. Chapter 2 provides a more mathematical description of the multi-level data-driven probabilistic model of the lineage and the construction of the prototype. Chapter 3 discusses some perspectives raised by the model such as a notion of irreversibility of desorganization implied by the structure found in the embryos and a formal characterization of robustness of development.

Characterizing diversity In the second part, we will study the emergence of diversity in evolution and how it is shaped by development. The question consists in how to relate variations at the individual specimen level to variation at the population level and how to characterize the influence of mutant development for diversification.

The variety of mechanisms at the origin of variation stimulate this question. In chapter 4, we try to characterize variability in biology with respect to formalization in mathematics and physics. The review of the main mechanisms sources of variation shows an heterogeneity of processes, from genetic mutations [139] to stochastic gene expression [65], through epigenetic changes [9]. The timescales at which they operate and their various mechanisms of inheritability are obstacles for their integration, although they all contribute to the generation of diversity. We then turn to mathematics and physics to explore the characteristics of the frameworks used for the **formalization of uncertainty**. Probability theory offers a framework to handle events in a context of uncertainty but doesn't provide any definition of randomness itself. This framework, following Kolmogorov's axiomatisation, rely on the boolean algebra of sets to construct the set of events (σ-algebra) which is a model that may show some limitations, for example in the case of quantum mechanics. Chaos theory and ergodic theory are two frameworks in classical physics which provide two different characterizations of randomness [14]. In both cases, it is an epistemic concept. Quantum mechanics on the other hand rely on an objective use of randomness, as an intrinsic component of the object under study. The Hilbert space structure used in quantum mechanics probabilities enables operations that were not possible in the Kolmogorovian framework such as a tensor product between space of possible corresponding to different observables [19]. Finally, we will argue that, given the high complexity of biological organisms and the heterogeneous nature of the mechanisms source of variability, an alternative approach consists in using models to explore the repertoire of possibles although they will always provide an incomplete description of the space of possible.

The experimental study of the *squint* mutant line of the zebrafish *Danio rerio* will then bring an example of **variable phenotypic expressivity** and **incomplete penetrance** [169]. The results of a quantitative assessement of the distribution of phenotypes in the progeny of homozygote mutants shows a discrete list of phenotypes that may be incomplete and in unpredictable proportions. This experimental study is then the basis for an analogy with the mathematical framework used in quantum mechanics, since the traditional Kolmogorovian framework shows some limitations when the description of the space of pos-

sible is incomplete. The analogy rely on the possibility to use the Hilbert space vector space structure and the possibility to perform tensor products between spaces of possible. By differentiating between uncertainty at the level of the observables, at the level of possible phenotypes and the probability of obtaining these observables, we will clarify some aspects of biological variation [211]. In particular, this framework can handle the emergence of new observable and the emergence of new phenotypic value for a given observable. We obtain a **formal analogy with entanglement** in Mendel's idealized scheme of inheritance that is interpreted as a trace of biological organization.

Finally, we will explore the concept of an **ontogenetic tree**, which is an attempt to organize divergence patterns between developments among mutants of the same species [101]. This question relates to the concept of developmental constraints or canalization, and is a first step toward an understanding of how variability in development shapes the space of possible forms [5]. We show with a data set containing a large number of description of mutant developments [29] that the zebrafish's pharyngula stage is actually the stage from which the highest number of mutants begin to diverge. However, possible biases in the data set are discussed.

Quantifying shape In the third part, we focus on epithelial organization and on the traces left by individual histories. The relative universality of this structure allow to study variability among several species and within mutant developments. Epithelial morphogenesis results from a sequence of events involving cell proliferation, cell movement, cell death and cell extrusion, leading consequently to a complex landscape. Using a data set of epithelial images, we will consider the quantitative characterization of the network of cellular contacts [67]. The network of cellular contacts gives a good estimation of epithelial organization, however traditional tools from complex networks theories show some limitations for its study since this network is highly constrainted by the underlying topology of the tissue.

A similar problem arise when trying to characterize the structure of the cosmic web. It is an historical structure shaped by random events at various scales which forms a complex landscape. Using a discrete analog of the approach developed in [200], [38] we propose to compute topological invariants of the network of cellular contacts by adapting the framework of persistent homology. To this aim, we begin by considering a discrete version of the level set functions on the network using the number of neighbors as the parameter. The number of neighbors is used as a measure of "density" in the network. Computation

of the sub and super level sets with varying threshold for the parameter number of neighbor generates two filtrations, i.e. two sequences of nested subspaces, from a network. These sequences unfold the structure of the network. For each value of the parameter, **Betti numbers** can be computed on the corresponding subspaces of the filtrations. These Betti numbers roughly measure the i-th dimensional holes in the considered space and are topological invariants. The sequence of Betti numbers values obtained when exploring the range of value of the parameter enable to compute a persistence diagram which is a **topological signature of the network**. This signature is automatically extracted from confocal images of epithelia from Drosophila and Chick embryos. It is used to compare and classify tissues.

To make sense of these topological signatures, we introduce a model of **random triangulated surface**. This model has the same number of neighbors distribution as the empirical network, nodes are linked randomly to form a triangulated surface. The topological signature is computed on this model of random triangulated surface and enable to estimate the distance of empirical networks to a random spatial distribution of cells. Significant deviation from the random model is obtained for the different cases studied; non randomness of the spatial distribution is a measure of the influence of the morphogenetic process for the construction of the network of cellular contacts. It shows that events occurring at the individual cell level have an influence on the global morphology of the tissue. The role of Myosin II, an element of the cellular cytoskeleton, is estimated by comparing the results of the method on knocked out mutants and wild type embryos. Overall we show that the level of organization of the cell cannot be uncoupled from the tissue organization. This study is reported in chapter 8 and in [209], [212].

Finally, we consider the dynamical aspects of the network of cellular organization by proposing some perspective on possible characteristics that could be computed on developing tissue. These dynamics involve the branching structure of the cell lineage as well as its spatial unfolding. They are described in chapter 9.

Part I

Characterizing normality

Introduction

In this part, we investigate the question of the reproducibility of development in normal conditions. This question is crucial for developmental biology. Embryogenesis has been believed to be highly reproducible and finely tuned leading to the metaphor of the "execution of a program" encoded in the genes [111], [76].

In the last decade, the reproducibility of the development of the *Drosophila* early development has been quantified at several level of observation and has been shown to be highly reproducible in particular at early stages. At the genetic level the quantification of the morphogen concentration profiles such as Bicoid indicates a reproducibility with 10% variation, interpreted as a precise control over absolute concentrations and responses to small concentration differences. At the cellular level, the cell membrane lengthening during cellularization of the *Drosophila* is a highly reproducible process ([130], [72], [59]) allowing to calibrate measurements very accurately. At a more macroscopic level the fly wing vein patterns are highly reproducible, the precision is in the range of a single cell width [1].

For the sea urchin, Eric Davidson proposed the idea of an "invariant cell lineage" [51]. In the seminal article by Sultan and Horvitz the lineage of the nematode sea elegant *Caernohabidtis elegans* has been shown to be reproducible at the single cell resolution [202]. However, until now, characterization of the development of the sea urchin at single cell resolution has not been performed. The work presented in this part concerns the quantitative study of sea urchin development at the single cell resolution.

Quantitative and integrative approaches in developmental biology

Quantitative approaches in developmental biology require to integrate dynamics occurring at several scales. Several steps are needed to understand the mechanisms occur-

ring at these various scales, from data acquisition, to reconstruction, to modeling. The BioEmergences platform [2] has been pioneer in this field. The approach conceptualized by Nadine Peyriéras and Paul Bourgine is summarized in the epistemological triangle as represented on figure 1.

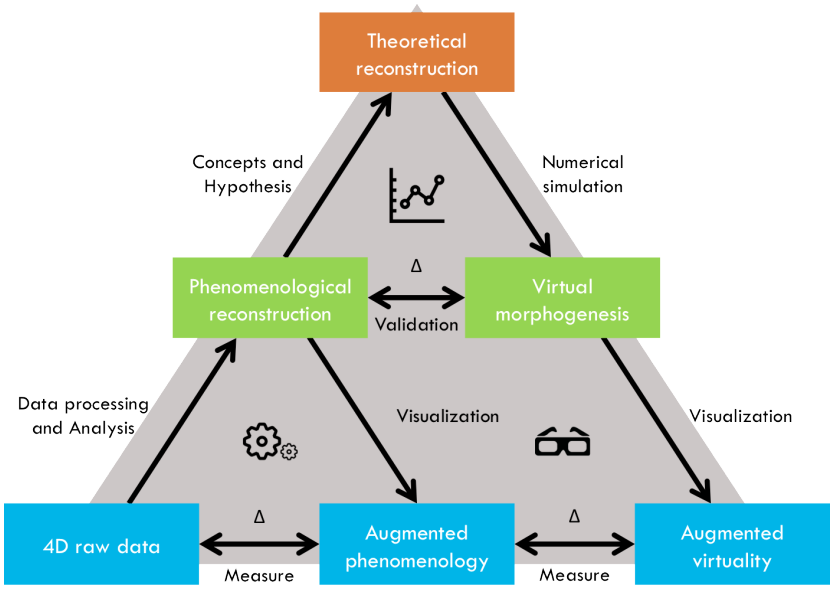


Figure 1: BioEmergences epistemological triangle showing the various steps required from data acquisition, to phenomenological reconstruction, to modeling - Image courtesy of the BioEmergences platform see footnote 2

The first step consists in producing an accurate *in toto* and *in vivo* image acquisition of the embryo development. This step involve the development of efficient imaging technologies. The most popular technologies are two-photon microscopy [57], [97] and light sheet microscopy [206], [166]. The concept of two photon microscopy consists in exciting fluorophores which can be of different kinds and are distributed at relevant places in the embryo using two photons reaching the desired energy level when they converge and superpose at the focal point. This technique enables deep imaging in living tissues. The concept of the light sheet microscopy consists in using a thin sheet of light instead of light focused on one point, this technique enables quicker image acquisition but require to rotate the sample. Fluorophores consist most of the time of fluorescent proteins which are translated from RNA injected at a precocious stage. These fluorophores commonnly fuse to the membrane of the cell or to the nucleus. These techniques enable to measure developing embryo without perturbing their normal development. The data sets obtained

2. `http://www.bioemergences.eu`

consist in 3-dimensional images at regular consecutive time steps [149], therefore 4D raw data as indicated on the bottom left of the epistemological triangle on figure 1.

The second step consists in extracting biologically relevant information from the 4D raw data sets. This step is done through image processing. In particular, the position of the cell nuclei at each time step is a relevant information to reconstruct the spatiotemporal cell lineage [16], [163]. In many studies the shape of individual cells carries useful information [185]. From this phenomenological reconstruction, useful features can be extracted, such as the position and velocity field of cells [147], other cinematic description of morphogenesis can be found in [132]. The accuracy of the phenomenological reconstruction can be assessed through visual inspection by experts using visual platform such as the MoveIt software [70].

Once the phenomenological reconstruction has been established for one or several embryos, it is useful to propose theoretical hypotheses to interpret the data. These hypotheses can be extracted from the data [109], [108], [28]. Or they can be brought from external knowledge, for example physical hypotheses or previous experimental results [55], [220], [219], [96], [18], [75], or information theoretic approaches such as the concept of complexity [79], [13].

The main question that remains is how to compare theoretical assumptions with empirical data, given the variability at all scales observed in developing embryos.

Multi-level approach for the study of the sea urchin early embryogenesis

The sea urchin *Paracentrotus lividus* is a model organism widely used in developmental biology. Sea urchin's embryos have many advantages that led to this status of model organism. The eggs are easily accessible, fertilization can be controlled, embryo are transparent and develop quickly, enabling their observation with microscopy techniques [69]. The use of the sea urchin for embryology can be traced back to the XIXth century, beginning in 1847 with three publications documenting fertilization, "Sur le développement des oursins" (On the development of sea urchins) by Adolphe Dufossé, "Auszug aus einem Berichte des Akameikers v. Baër, aus Triest" (Excerpt from a report by the University Graduate von Baër in Trieste), by Karl Ernst von Baër and "Observations sur le mécanisme et les phénomènes qui accompagnent la formation de l'embryon chez l'oursin comestible" (Observations on the mechanism and phenomena accompanying the formation of the

embryo of the edible sea urchin), by Alphonse Derbès [31]. These studies were followed by a famous work by Oskar Hertwig on sperm and egg pronuclear fusion in 1876. In 1891, Hans Driesch used the sea urchin to perform experiments on development, showing that a complete embryo could develop from extracted cells refuting preformation and mosaic theories.

During the XXth century the sea urchin has been the basis for major discoveries. Tim Hunt and collaborators discovered the role of cyclins in the sea urchin development, which are key regulators of the cell cycle [68]. The sea urchin development has also been used as a basis for the comprehensive study of the gene regulatory network [53]. The most recent development of this approach can be found on the biotapestry website [3].

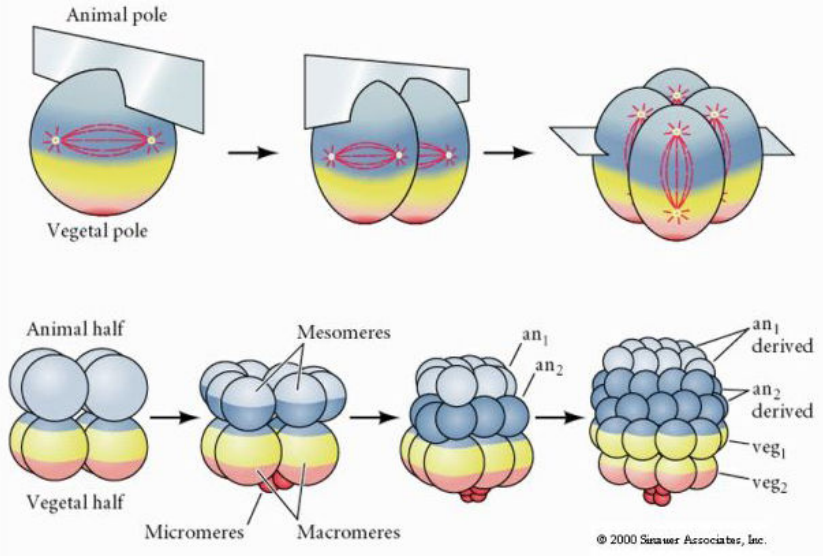


Figure 2: Diagram representing the first steps of the sea urchin cleavage patterns, from [83]

As any multicellular organism, the sea urchin develop from a single fertilized egg. It undergoes a radial holoblastic cleavage, meaning that all the cell divide in a stereotypic way, with division axis either parallel or at right angle with what will become the animal vegetal axis. The first and second cleavages are both meridional and perpendicular to each other, as shown on figure 2. They are followed by an equatorial cleavage perpendicular to them. The three first cleavages divide cells symmetrically, leading to a symmetrical 8-cell blastula. The fourth division round is different of the first three, the animal and the vegetal tier don't divide similarly. The cells of the animal tier divide meridionally into eight cells

3. http://www.biotapestry.org/

with similar volumes called the Mesomeres cells. The cells of the vegetal tier don't divide symmetrically, the division occur in the equatorial plan, producing four large cells called Macromeres, close to the Mesomeres, and four small cells called Micromeres at the vegetal pole. At this 16-cell stage where the morphological symmetry breakings occur, the cells are well identifiable. The next round of division is equatorial for the Mesomeres, forming two tiers of eight similar cells. The Macromeres divide meridionally, forming a tier of eight cells below the Mesomeres. The Micromeres divide somewhat later forming two sets of cells, four Large Micromeres and four Small Micromeres. The Small Micromeres divide once more, then cease dividing before the larval stage, see figure 3. When the embryo has attained the 32-cells stage, it has begun to form a blastocoel which is a proteinaceous fluid within the embryo, and during the next rounds of division which are less stereotypical, cells organize themselves as a single layered epithelium surrounding the blastocoel.

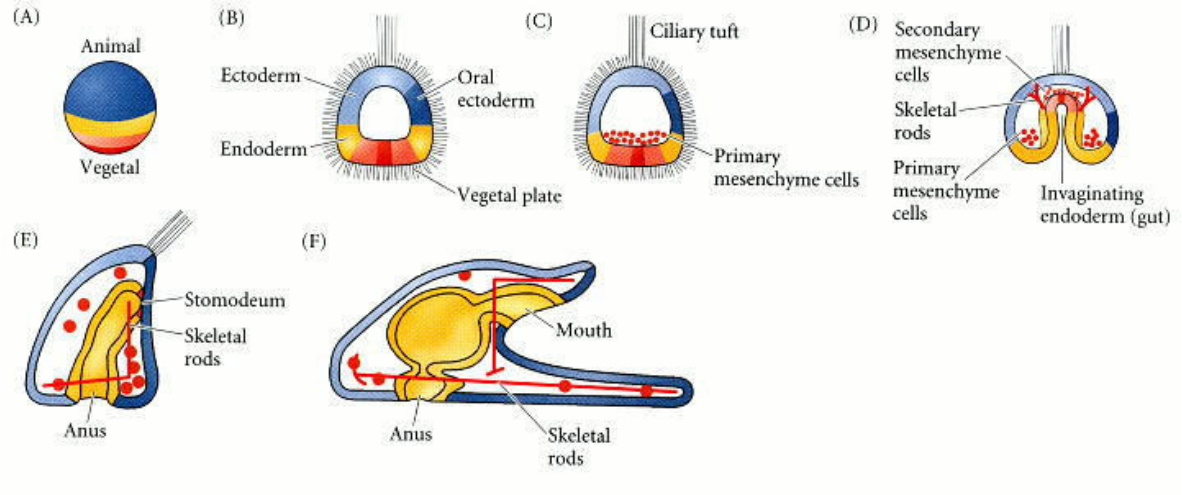


Figure 3: Normal sea urchin development, following the fate of the cellular layers of the blastula. (A) Fate map of the zygote. (B) Late blastula with ciliary tuft and flattened vegetal plate. (C) Blastula with primary mesenchyme. (D) Gastrula with secondary mesenchyme. (E) Prism-stage larva. (F) Pluteus larva. Fates of the zygote cytoplasm can be followed through the color pattern. (Courtesy of D. McClay.) - Image and caption from [83]

At the end of the cleavage period, which correspond to proliferation without cell death and without many cell movements within the blastula, the cells begin to undergo complex morphogenetic movements. Figure 3 shows a schematic drawing of this gastrulation movements which will later give rise to a free swimming pluteus larva.

The study of the sea urchin gene regulatory network (GRN) underlying development lead to the establishment of maps of interaction such as the one represented on figure 4 for

Mesomere cells during the period from 6 to 17 hours. Obtaining such a map require to test each gene individually. On the scheme, each gene is represented with its name (e.g. Nodal or Lefty), and the links between the genes represent interactions such as induction or inhibition. We can observe the modularity of this network which depend only on signals from Mesomere cells, except for some maternal inputs and the expression of the gene Wnt8 in the cells Veg2 (a subpopulation of the Macromeres lineage). The establishment of the gene regulatory network has been described in [53] and contains more than 40 genes. The dynamics in space and time of this gene regulatory network has been studied as a boolean computational model [171], it is argued in this article that the data underlying this gene regulatory network contains sufficient information to explain the complex developmental process of gene expression. The main limitation of this approach is that the spatial resolution is coarser than the individual cell level where interactions between genes take place. This limitation can be overcome with analysis at the single cell resolution such as the one presented in this part of the dissertation.

In the following we propose to investigate quantitatively the morphogenesis of the sea urchin *Paracentrotus lividus* from the 32-cells stage to more than 400 cells, by observing the phenomenology of individual cells, shape and proliferation dynamics.

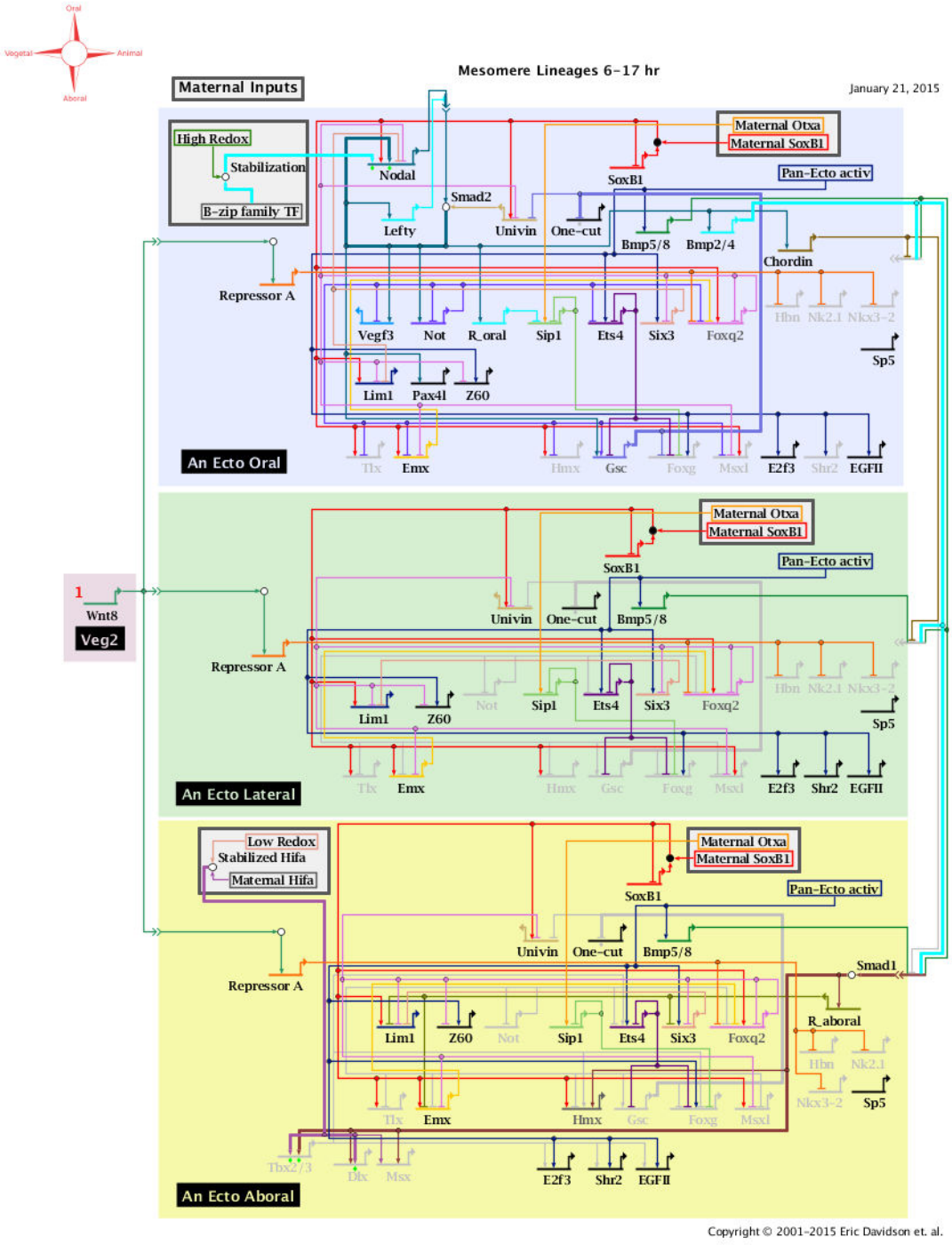


Figure 4: Diagram representing the gene regulatory network of the Mesomeres during 6 to 17 hours post fertilization in the sea urchin *Strongylocentrotus purpuratus* from Eric Davidson's website at Caltech `http://sugp.caltech.edu/endomes/` as of January 21, 2015

Overview of the part

The main object of this part of the dissertation is the comparative study of a small cohort of digitally reconstructed sea urchin embryos from *in toto* and *in vivo* 2-photon microscopy imaging. Using the BioEmergences workflow, the complete reconstruction of the cell lineage and individual cell shape is obtained for 5 wild type embryos. The question which is raised by this data set is how to characterize intra-individual variability which underlies cell differentiation and inter-individual variation associated to individual developmental histories. Given the multi-scale nature of the embryonic development, the answer to this question depends on the level of observation chosen. Indeed, features at the individual cell level can be very variable from one individual to the other, whereas embryo-level dynamics appear highly reproducible.

The original approach that we develop to integrate these various levels of observation is summarized on figure 5. Based on the image acquisition and the phenomenological reconstruction, a set of features is obtained for each individual cells. These features are the length of the cell cycle, the division, the mean volume, the mean surface area, the mother/daughter volume or surface area ratio. Since we have access to the complete lineage, it is possible to follow their evolution through the genealogy. At a more macroscopic level, we can look at embryo-level dynamics such as the evolution of the number of cells through time, the evolution of the total cell volume, the evolution of the total cell surface area.

To integrate the various levels of observation, we defined an intermediate coarse-grained level of observation between individual cell features and embryo-level dynamics by clustering individual cell features into identifiable groups of cells. These groups are defined by common cell identity (Mesomeres, Macromeres, Large and Small Micromeres) and common generation (number of cycles undergone since fertilization), hence identifiable in each specimen of the cohort. This generic coarse grained level of observation is the basis unite of our comparative study. It allows to compare individual characteristics without averaging out intra-individual variability.

The link between individual cell features is characterized with a data-driven multilevel probabilistic model relying on this intermediate level of observation. Individual cell features distribution can indeed be described and approximated through parametrized probability distributions. These probability distributions are then combined using the branching structure of the cell lineage. The parameters governing the probability distributions

provide a signature of each individual specimen in the cohort. They are distributed in a statistical manifold where inter-individual comparison can be performed.

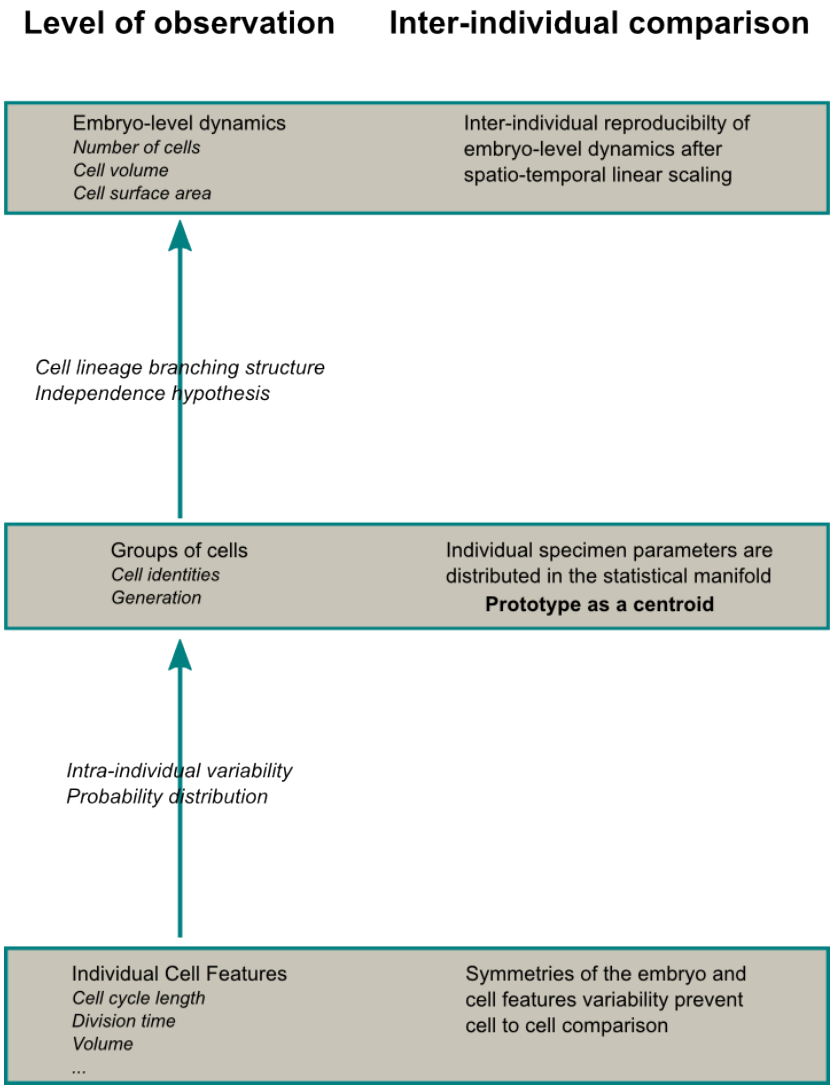


Figure 5: Scheme representing the proposed approach to quantify inter individual variability in the sea urchin development and produce a prototypic model.

After having characterized intra- and inter-individual variability in the development of the sea urchin, it is interesting to ask for the possibility of representing uniquely the development in a quantitative manner. This question is answered by establishing a prototypical representation of the cell lineage as the centroid of the cohort in the statistical manifold. This prototype can then be used as a basis for an hypothesis based modeling of morphogenesis. Indeed, this prototype enables to integrate quantitative parameters values in a biomechanical model, which can then be compared to empirical data on aspects

not belonging to the input of the model.

Chapter 1 describes the sea urchin's development in a cohort of digitally reconstructed specimens. The specimens have been observed under normal condition and correspond to a typical range of variability occurring in a normal development. This chapter represents a complete integrative approach for the study of development morphogenesis, as it combines empirical measures, digital reconstruction, data analysis. Establishment of a prototypical model for proliferation and cell volume dynamics enables to test biomechanical hypotheses through parameter exploration with a biomechanical model.

Chapter 2 describes the data-driven multi-level probabilistic model underlying the analysis presented in chapter 1 in depth. It requires some mathematical background. The contribution of this chapter is the establishment of the prototype

Chapter 3 discusses some perspectives of the multi-level probabilistic for our understanding of embryogenesis, such as the concept of irreversibilty of desorganisation during the considered period of development because of the lack of regulation of individual cell features.

Chapter 1

Predicting sea urchin's normal development from a small cohort of digital embryos

Abstract *The quantitative comparison of developing sea urchin embryos from a small cohort of digital specimens is the basis for the construction of a prototypic cell lineage tree, sufficient to predict the spatio temporal cell organization of a normal sea urchin blastula. This is achieved i) by finding the statistical models fitting best the phenotypic macroscopic phenotypic features, ii) and embed the corresponding artificial prototypic cell lineage in the 3D space via a biomechanical model. The resulting 3D model is made to systematically explore a space of parameters to fit the experimental data in order to test biological hypotheses.* [1]

1.1 Introduction

The question of finding the time and the locus for the apparition of differences between individuals has irrigated the science of embryology ([111]). Large genetic screens have sought to find genetic determinants of these differences ([161], [92]). The construction of the genetic regulatory network of the sea urchin should reveal the dynamics of

1. This chapter is an early version of a paper involving Barbara Rizzi, Louise Duloquin, Julien Delile, René Doursat, Paul Bourgine and Nadine Peyriéas - This study has been presented under various forms at the conferences "The developmental biology of the sea urchin" XXI and XXII, at Woods Hole, MA, USA (2012 - 2014) and at the Gordon Research Conference "Stochastic physics in biology" at Ventura Beach, CA, USA (2015)

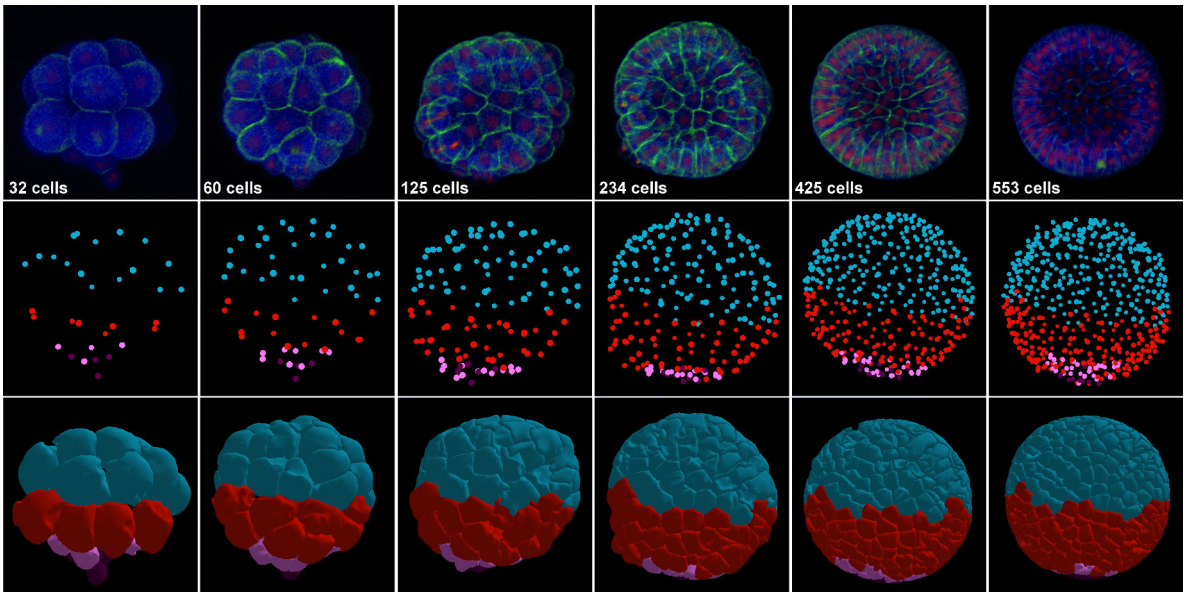
these determinants during development ([53]). However, recent results have shown that the linear relation between genetic regulatory network and phenotype can be complicated by stochastic ([181]) effects. On the other hand, physical determination and constraints at the scale of the tissue or of the whole embryo canalize the space of possible shapes ([189], [94], [74]). Relating these two approaches, genetic determinants and physical constraints at the scale of the tissue, requires to understand the relations between individual cell phenomenology and transformations at the whole embryo level. Phenomenological reconstruction of live embryo development ([163]) generates data that allow to investigate quantitatively such questions. We propose to use the full digital reconstruction of a small cohort of developing sea urchin to unfold the relations between cell, tissue and whole embryo dynamics.

The sea urchin has been studied as an animal model since over a century ([66], [144]). The study of early embryogenesis provides insight in differentiation of cells in the different layers ([53], [52]). The description of the morphogenetic changes associated to these differentiation processes are lacking a quantitative description. In particular the quantitative study of the sea urchin blastula development is a good model to tackle this problem because the morphogenetic changes are undergone smoothly by the embryo allowing live imaging over a long period of cleavage.

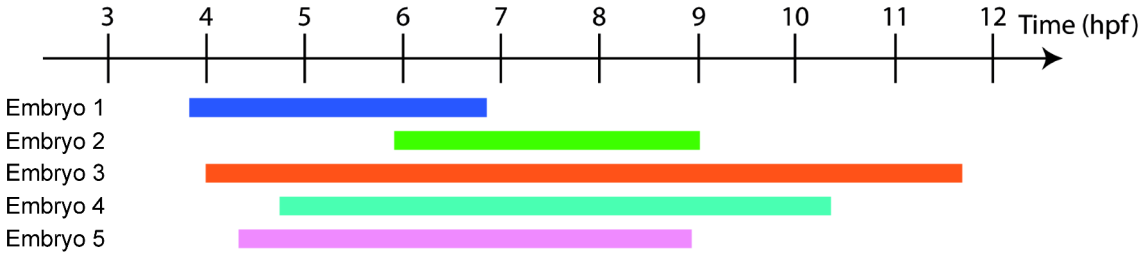
The full digital reconstruction of live specimen reveals simultaneously quantitative features at the individual cell level and at the scale of the whole embryo. The development is orchestrated by changes in size, shape, number, position and gene expression of cells. The question remains of how these quantities are related to each other and in which way do the macroscopic dynamics of the development emerge from the micro characteristics of the cells. Are the individual cell features aggregated together in a unique precise way that would create the precise patterns that we observe, suggesting a precise developmental mechanism regulating the development, or the global dynamics emerge from loose relations between the cells, favoring robust emerging process.

1.2 A small cohort of digital sea urchin throughout their cleavage period

The quantitative comparison of the cell lineage and cell behaviors was achieved through the full reconstruction of digital specimens from 2-photon microscopy imaging of live embryos. Nuclear and membrane staining obtained by 1-cell stage injection ([70]) of syn-



(a)



(b)

Figure 1.1: (a) Reconstruction of digital specimens from 3D+time in toto imaging. First line - Volume rendering of raw images from 2-photon laser scanning microscopy, H2B-mCherry and farnesylated eGFP staining. Second line - Nuclei detection and cell tracking. Coloured dots represent the cell positions, cell trajectories displayed as streamlines over the next 5 consecutive time steps. Third line - Surface rendering of segmented cell membranes. Color code: Mesomeres (blue), Macromeres (red), Large (pink) and Small (purple) Micromeres. (b) Temporal sequences covered by the imaging of the five different specimens analyzed in this study.

thetic mRNA encoding H2B-mcherry fusion protein and farnesylated eGFP respectively, was used to image embryos developing from the 32-cell stage until the hatching blastula stage (Figure 1.1 - 1st line). Image data sets were processed through the BioEmergences reconstruction workflow ([70]) to provide the complete cell lineage (Figure 1.1 - 2nd line) and the segmentation of cell shapes (Figure 1.1 - 3rd line). The visualization interface Mov-IT ([163]) was used to validate and correct the cell tracking and manually label cells according to their distribution in known populations ([11]).

1.3 Feature Extraction and Measuring

To define normal sea urchin's development we used 5 developing specimens imaged in similar experimental condition with the same set up. Acquisition lasted from 2 to 8 consecutive hours with a time resolution of 2 to 4 min, beginning 4 to 5 hours post fertilization (32 cells - Figure 1.1 (d)). Cell lineage combined with cell segmentation provides the life length and division time of cells, as well as the volume and surface area (Chapter 2 - section 2.3.1). Mesomeres, macromeres and large and small micromeres cell populations were marked at the 32-cell stage.

Morphological changes at the embryo-level are witnessed by the evolution of the number of cells, the cell volume and the cell surface area (Chapter 2 - Figure 2.4 A, D, G). These dynamics have similar patterns in each specimen of the cohort. However spatio temporal rescaling overcome a first level of interindividual variability, making these dynamics comparable from one embryo to the other, in whole embryo and in each morphogenetic field (Chapter 2 - Figure 2.4 B, E, H). It consists in an affine transformation of the time dependency (two parameters - Chapter 2 - Figure 2.4 C) and a linear transformation of the spatial dependency (one parameter - Chapter 2 - Figure 2.4 F).

To compare quantitatively cell features among specimens of the cohort, it is necessary to find generic coarse-grained levels of description because symmetries of the embryo prevent to identify individual cells (Chapter 2). Relying on exchangeability of the cell fate at this period for cells belonging to the same population (Mesomeres, Macromeres, Large micromeres, Small micromeres), cells were clustered in groups of cells sharing identity and generation. These groups of cells forms the basic unit for the comparison and modeling in this study of the sea urchin development.

1.4 Emergence of embryo-level dynamics from individual cell features

To relate the macroscopic dynamics with the individual cell features, we propose a data-driven multi-level probabilistic model of the cell lineage which rely on its branching nature. After each cell cycle, a cell divide into two. The features of the cells are chosen randomly using the corresponding probability distributions defined for each groups of cells independently of any genealogical relationships (Chapter 2 - section 2.5). The macroscopic dynamics observed in each embryo and within each morphogenetic fields are ac-

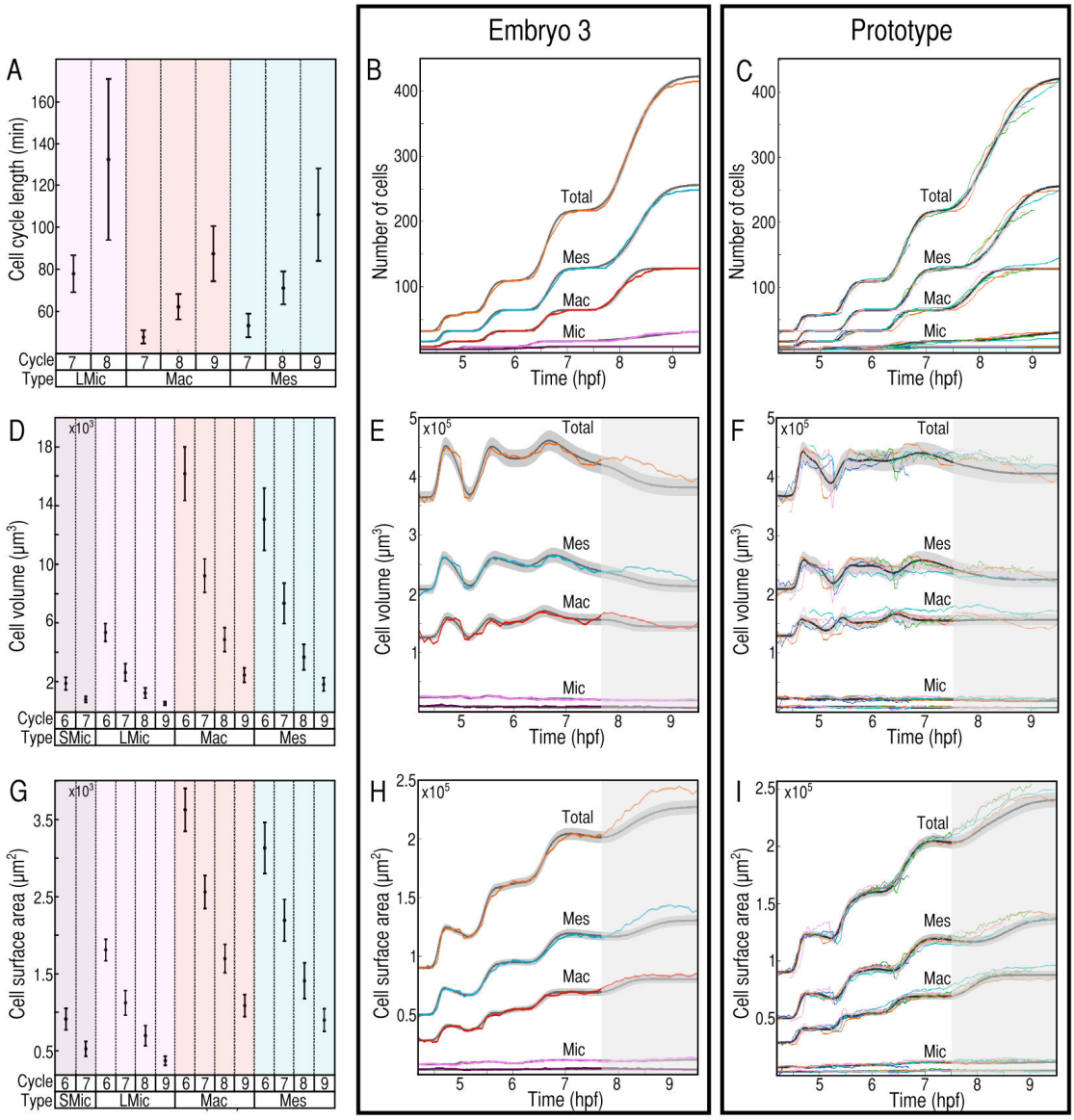


Figure 1.2: Probabilistic modeling. Each individual of the cohort is associated to one color. The black lines with grey intervals correspond to 300 realizations of the model, mean (black) and standard deviation (grey). The first column (A,D,G) shows the value of the parameters, cell cycle lenth, cell volume, cell surface area for the prototypical representation of the cohort. The second column (B, E, H) corresponds to the model for one individual of the cohort. The third column (C, F, I) corresponds to the prototypical model over the cohort. The first line (B,C) shows the evolution of the number of cells, relying on the distribution of life lengths in the different subpopulations (A). The second line (E,F) shows the evolution of the cellular volume, relying on the distribution of the volume (D). The third line (H,I) shows the evolution of the cellular surface area, relying on the distribution of the surface area in the different subpopulations (G).

curately reconstructed with empirical parameters for each group of cells (Figure 1.2 B, E, H). Variability in the division times results from the successive addition of variability in the life lengths, leading to a continuous desynchronization of the cell cycles (Chapter 2). The variation in the total cellular volume and surface area result from variable mean characteristics with invariant cell dynamics (Chapter 2). These results suggest that the high reproducibility of embryo-level dynamics emerges from individually loosely regulated cell features within a branching structure and population level characteristics.

Each specimen of the cohort is represented by a set of parameters sufficient to reproduce embryo-level dynamics. The cohort is represented as a set of points in the associated statistical manifold and the prototypical representation of the cohort is defined as the centroid of the specimens in this parameter space (Chapter 2 and [10], [159]). Prototypical statistics for the groups of cells are defined using this methodology (Figure 1.2 A, D, G). Intraindividual variability is represented by the prototypical standard deviation computed for each cell feature. A representation of the normal development of the sea urchin during cleavage is obtained by simulating prototypical embryo-level dynamics from the probabilistic model (Figure 1.2 C, F, I).

1.5 Spatial modeling

To understand the relations between the individual cell features and the shape of the embryo, the prototypical model of the cell lineage is embedded in space with a biomechanical model using the MecaGen modeling platform (Figure 1.3 A - [55], [56]). Each cell is represented by a single cylindrical particle oriented along the apico-basal axis of the epithelium (Figure 2.18). As cells are extremely small and sticky (low Reynolds number, [179]), inertia is neglected in favor of viscosity forces and the cell displacement is caused by their immediate mechanical interactions. Thus, the relation between displacement and the net force applied on cells by their neighbors is ruled by an overdamped equation of motion. At the bastula stage, the increasingly epithelial nature of the cells induces a decomposition of the force exerted between two neighbors into a set of tangential and normal components: the attraction-repulsion force maintains the integrity of the cell volume and controls the stiffness and the adhesion of interaction in the tangential direction, and the planarity conservation force maintains the planarity of the monolayered sea urchin epithelium (Figure 1.3 B,C). In between a pair of neighbor cells, the attraction coefficient varies depending whether the pair belongs to the same subpopulation (homotypic) or not

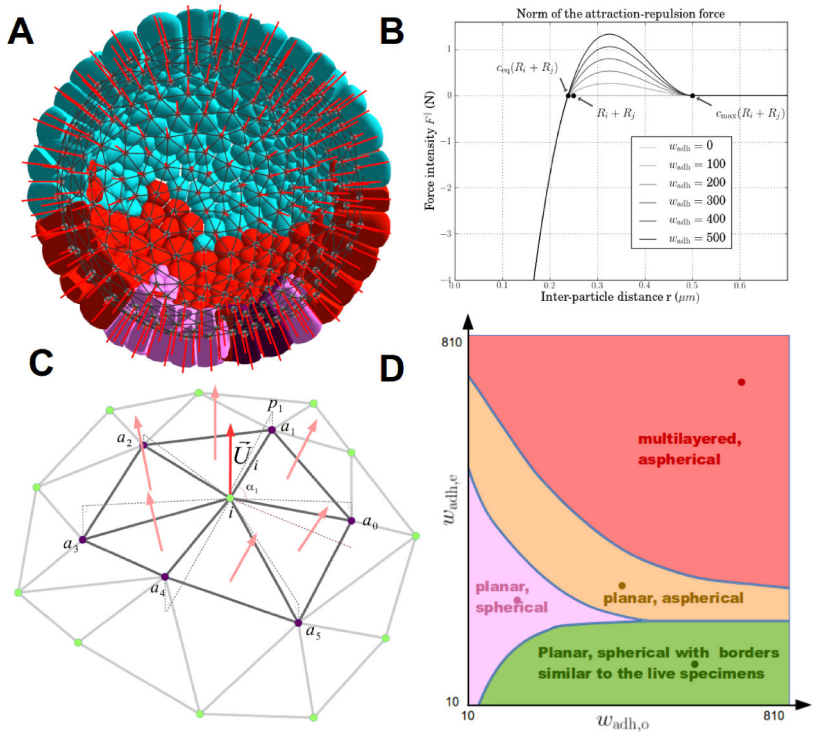


Figure 1.3: A. Each dot represents a cell center, the edges relating them are calculated from the two steps spatial neighborhood algorithm, via metric and topological criteria. The red axes represent the cell axis $\vec{U}_i$ oriented along the apico-basal direction. Membrane surfaces are calculated a posteriori for rendering purpose. Color indicates indentity similar to figure 1.1. B. Intensity of the attraction-repulsion force $\vec{F}^{\parallel}$ exerted between two neighbor cells as a function of their relative distance. This force displays a similar shape to interatomic potential derived forces like Morse or Lennard-Jones (Chapter 2 - Figure 2.6). The force well is modulated by varying the appropriate attraction coefficient w_{adh}. C. The cell axis $\vec{U}_i$ (red arrow) is calculated by averaging the 6 surrounding triangle outward normal vectors (light red arrows). D. Phenotypic phase diagram, axis correspond to $(w_{\text{adh,e}}, w_{\text{adh,o}})$ with $k_{\text{rig}} = 10000$ and $\alpha_{\text{gab}} = 1.0$. Four distinct phenotype domain are obtained when exploring the parameter space. Figure by Julien Delile

(heterotypic). Cell division timing follows the prototypical probabilistic model mentioned above (Chapter 2) and their orientation is performed in the tangential domain with an angle chosen randomly using a uniform distribution.

Parameter space exploration determines the parameter sets which govern realistic spatial enfolding of the sea urchin embryo development. Validation uses the sphericity of the global embryo shape, the maintenance of the monolayered epithelium and the similarity of the inter-subpopulation border shapes with those observed in the digital embryos (Chapter 2 section 2.7). The best fitting domain is obtained for low heterotypic adhesion, confirming that clear-cut border can be obtained without the need of biasing the division orientation (Figure 1.3 Db:Mean). Moreover, when the attraction-force intensity is larger than the planarity force intensity, the embryo epithelium agglomerate into 3D aggregate. This transition runs through different phenotypic states: from the highly spherical and planar embryos to the collapsed magma of cells, some starfish shaped embryos may appear in the parameter region where planarity is obtained even in the absence of sphericity (Figure 1.3 - D).

1.6 Discussion

In summary, we demonstrate that, despite interindividual variability during cleavage period, it is possible to uncover underlying invariant structures driving macroscopic dynamics and morphological changes. Their reproducibility emerge from loosely regulated cell features along the branching cell lineage. In particular we show that the desynchronization of cell division increases continuously during cleavage, and not as successive period of synchronicity, metasynchronicity and asynchronicity as previously stated ([83]). And if a pseudo gradient from vegetal to animal has been suspected ([60]), our study suggests that it results from the characteristics of cell life length variability and does not necessarily need a material support.

By finding a relevant generic coarse-grained level of observation to compare specimen we overcome the problem of intraindividual variability and symmetries that arise from multiscale observations. Modeling of the cell lineage and its 3D biomechanical embedding assure the sufficiency of the characteristics defined at the level of groups of cells to describe accurately the development of one specimen of the cohort, thus suggesting a regulation at the level of populations of cells and not at the level of individual cells. This period of development may not need to require a fine tuned genetic regulation. This sparse

description of the development for each specimen lead to the modeling of a prototypical embryonic development which is the centroid of the cohort in the space of models. This prototypical representation defines the normal development with a generality level guaranteed by the number of specimen in the cohort. This framework may form a basis to bridge the gap between experimental biology and theoretical biology.

Chapter 2

Variability in the sea urchin development: A multi-level data driven probabilistic model

Abstract *This chapter describes a data-driven multi-level probabilistic underlying the comparative study of a small cohort of digital sea urchin embryos presented in chapter 1. To relate individual cell features with embryo-level dynamics, it is necessary to define an intermediate generic coarse-grained level of observation. This level enables to characterize probability laws that are the basis of a comprehensive probabilistic model. Corresponding parameters are distributed in a statistical manifold where each embryo is identified by a small set of points. A prototype is obtained by computing the Kullback-Leibler centroid among specimen of the cohort. This prototype serves as a basis for a spatial embedding through biomechanical modeling with the MecaGen platform.* [1]

2.1 Introduction

The development of the sea urchin blastula from 32 to 540 cells happens through cell proliferation with no cell death. Early embryonic territories and cell morphology allow to categorize cells according to known cell types, namely Mesomeres, Macromeres, Large Micromeres, and Small Micromeres [50] as represented on figure 2.2. Symmetry by rotation along the animal vegetal axis as well as cellular variability prevent to identify individual

1. The establishment of the multi-level data-driven probabilistic model has highly benefited from advices by Paul Bourgine whom we warmly acknowledge

cells from one specimen to the other.

Using digital reconstructions of *in toto* and *in vivo* developing sea urchin embryo, we propose a data-driven multi-level probabilistic model that relates individual cell features with embryo-level dynamics. These embryo level dynamics are the number of cells, the cell volume and the cell surface area. The high reproducibility of these dynamics is observed after performing a linear spatio-temporal rescaling.

We propose to model the cell lineage tree as a binary branching process where the division probability of a cell depends on its age. Each cell of the lineage lives for a random time before giving rise to two new cells. The same process applies to these two new cells. This approach can be compared to a binary Bellman-Harris process, yet identical probability among generations will not be assumed [122]. Once the cell genealogy has been established, the dynamics of the cellular volume and surface is investigated along the cell lineage. An intermediate level of description is defined by clustering cells by common cell type and cell cycle, allowing interindividual comparison. Life lengths and division times in these groups of cells are found to be well described with gaussian probability distributions, mean volume and surface area and mother/daughter ratio for the volume and the surface area are found to be well described with log-normal distributions. The same multi-level probabilistic model relates distribution of individual cell features in these groups of cells with embryo-level in each specimen of the cohort. The geometry of the space of probability distributions allow to assess the accuracy of our model which is found to be high compared to inter-individual variability. Moreover, this model provides a description of the propagation of intra-individual variability in the cell lineage. Finally, based on this multi-level probabilistic model, we define a unique prototypical representation of this period of development by aggregating individual specimen statistics as Kullback-Leibler centroids in the relevant morphospace. This prototypical representation is used to generate artificial cell lineages which will be the basis of a biomechanical model. Parameters of the biomechanical model are fitted to the empirical data by comparing the simulations with the reconstruction.

2.2 Image acquisition and digital reconstruction

This section describes the protocol to obtain images of developing sea urchin embryos, and the image processing algorithms used to reconstruct the cell lineage and the cell shape from these images. This section describes Louise Duloquin's work for the image acquisi-

tion and Barbara Rizzi's work for the image processing [70].

2.2.1 Image acquisition

To obtain microscopy images of developing sea urchin embryo, it was first necessary to inject oocytes from *Paracentrotus lividus* with 150 μg/ml H2B-mcherry and 150 μg/ml eGFP-ras synthetic mRNAs [148]. Embryos were either maintained between slide and coverslip covered with protamine or embedded in 0.25% low-melting-point agarose sea water at the center of a 3 cm Petri dish. Image acquisition was performed with 2-photon microscopy with simultaneous excitation at two different wavelengths (1030 nm and 980 nm). Acquisition of 5 developing specimen is obtained while maintaining similar experimental conditions. They lasted from 2 to 8 hours with a time resolution of 2 to 4 min, beginning 4 to 5 hours post fertilization.

2.2.2 Image processing

Nucleus center detection The detection of the position of cell nuclei relies on the intensity of the signal emitted by the fluorescent histone fusion protein H2B-mcherry. A difference of gaussian algorithm is applied on the images, resulting in a band pass filter sharpening cell nuclei signal, which enables center detection by maxima identification on the image. Value of the parameters were manually chosen with the MoveIt software superimposing raw data with the detected cell positions.

Cell tracking The cell tracking consists in the complete spatio-temporal genealogy of the cells, it is obtained by linking cell nuclei from one time frame to the other when they represent the same cell, or two daughters after mitosis. The tracking is obtained in five steps. The first one consists in linking nearest nuclei when this relation is reciprocal (it is not necessarily the case when there is a mitosis). The second step accounts for mitoses by connecting nuclei without predecessor with the nearest nuclei in the previous frame. The third step refine the tracking obtained by assuming that the cells have small displacement and hence don't change of neighborhood quickly. An energy function is minimized by simulated annealing. The fourth step consists in considering cells that do not live long as false positive nuclear centers. The fifth step consists in visual inspection of the lineage obtained with the MoveIt software which superimposes raw data with the reconstructed lineage.

Cell segmentation The cell segmentation consists in the reconstruction of individual cell shapes. The fluorescent signal associated to the cell membrane is obtained thanks to the GFP-Ras fluorescent protein. The cell membrane shape are reconstructed using a generalized version of subjective surfaces method [222]. This method enable to reconstruct missing boundaries, relevant in the context of the images presenting a low signal-to-noise ratio and incomplete membrane contours. This method is performed on the images that are first smoothed with a geodesic mean curvature flow filter [126]. Numerical implementation of both filtering and segmentation algorithms has been performed on finite difference schemes [164, 194]. Further details about the employed methods can be found in [126] and [222].

Using these image acquisition and image processing techniques, we obtained a small cohort of 5 digital embryos with completely validated reconstructed cell lineages and reconstructed individual cell shape.

2.3 Multi-level measures and rescaling

The digital reconstruction of a small cohort of developing embryos was analyzed at different levels of observation, from individual cell features to embryo-level dynamics via relevant coarse-grained cell groups. We observed similar patterns of evolution in the measured quantities, such as cell volumes and numbers of neighbors, across all specimens of the cohort. Interindividual variability was captured by a linear temporal rescaling of the cell number charts, and a spatial rescaling of the geometric charts. Spatial symmetries in the embryos defined groups of exchangeable cells that constitute the basis of the probabilistic model.

2.3.1 Individual cell features

The digital reconstruction of entire sea urchin specimens allowed us to extract a variety of features at the level of individual cells as presented. For any cell i, we defined the following quantities (Fig. 2.1):

- its *life length*: x_i, corresponding to the time between two consecutive divisions
- a *mitosis time*: m_i, denoting the moment at which the cell i divides into two daughter cells (and ceases to exist as such)
- its *current volume*: $v_i(t)$, with the *average volume* over its life length (using discrete

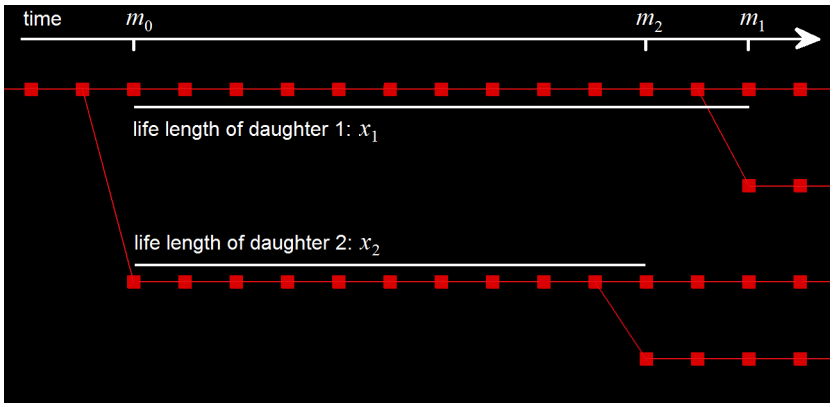


Figure 2.1: Sample of the cell lineage indicating how life lengths are calculated.

time steps):

$$\overline{v}_i = \frac{1}{x_i}\int_{m_i-x_i}^{m_i} v_i(t)dt \simeq \frac{1}{x_i}\sum_{t=m_i-x_i}^{m_i-1} v_i(t) \tag{2.1}$$

– its *current surface area*: $s_i(t)$, with the *average surface area* over its life length:

$$\overline{s}_i \simeq \frac{1}{x_i}\sum_{t=m_i-x_i}^{m_i-1} s_i(t) \tag{2.2}$$

– a *generation number*: n_i, denoting the number of past divisions on cell i's lineage branch, which includes itself and all its ancestors since fertilization
– a *cell type*: k_i, taking one of four possible values: 1 for mesomeres (Mes), 2 for macromeres (Mac), 3 for large micromeres (LMic), and 4 for small micromeres (SMic) [50] (Fig. 2.2)
– its *current degree*: $d_i(t)$, equal to the number of neighbors of cell i at time t, assuming that the spatial distribution of cells can be represented by an undirected graph of cellular contacts, in which cells correspond to nodes and cell-cell interactions to edges [80, 67] (Fig. 2.3).

Each cell i also has a mother identified by j (a shortcut notation for the function $j(i)$ mapping any cell index to its mother's index), which is used in the definition of two additional features:

– a daughter/mother *average volume ratio*: $a_i = \overline{v}_i/\overline{v}_j$
– a daughter/mother *average surface ratio*: $b_i = \overline{s}_i/\overline{s}_j$

and three relationships:

– the mitosis time of a cell is equal to the sum of its life length and the mitosis time of

its mother: $m_i = m_j + x_i$
- the generation is incremented by one at each division: $n_i = n_j + 1$
- starting from the 32-cell stage, each cell type is conserved across divisions throughout the period considered here (development of the blastula): $k_i = k_j$.

Morphogenetic fields and global state variables

The *cell lineage*, denoted by $\mathscr{L}$, is the set of all cells that the embryo contained during a given period of time. From there, various subsets were defined:
- the *current embryo*, the set of all cells alive at time t: $\mathscr{L}(t) = \{i \in \mathscr{L} \mid m_i - x_i \leq t < m_i\}$
- *morphogenetic fields*, the sets of cells of a given type k across all generations: $\mathscr{L}^k = \{i \in \mathscr{L} \mid k_i = k\}$
- *current fields*, the morphogenetic fields at time t: $\mathscr{L}^k(t) = \{i \in \mathscr{L}(t) \mid k_i = k\}$.

At the level of the entire embryo, and at each time step t, the global measures were:
- the *number of cells*, cardinality of the current embryo: $N(t) = |\mathscr{L}(t)|$
- the *total cellular volume*, sum of all individual cell volumes: $W(t) = \sum_{i \in \mathscr{L}(t)} v_i(t)$
- the *total cellular surface area*, sum of individual cell surface areas: $Z(t) = \sum_{i \in \mathscr{L}(t)} s_i(t)$
- the *total number of contacts*, sum of the local degrees: $C(t) = \sum_{i \in \mathscr{L}(t)} d_i(t)$

These quantities were also calculated inside each current morphogenetic field k, replacing $\mathscr{L}(t)$ by $\mathscr{L}^k(t)$ in the above definitions and using the notations $N^k(t)$, $W^k(t)$, $Z^k(t)$, and $C^k(t)$.

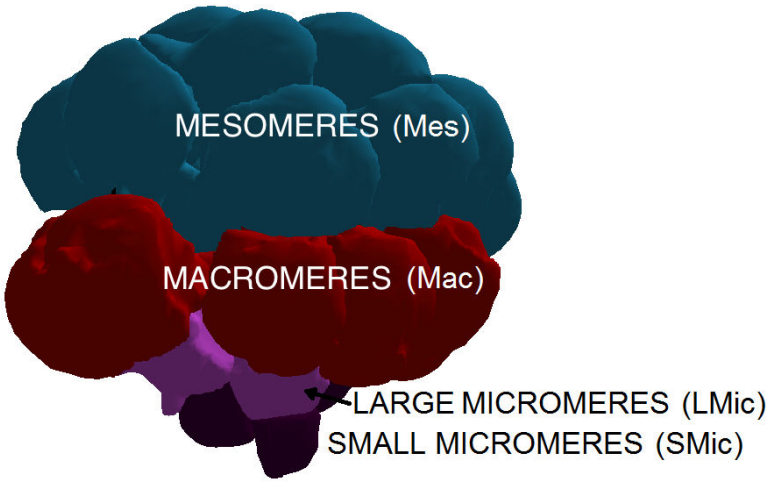


Figure 2.2: The four cell populations identifiable at the 32-cell stage: 16 mesomeres, 8 macromeres, 4 large micromeres, and 4 small micromeres.

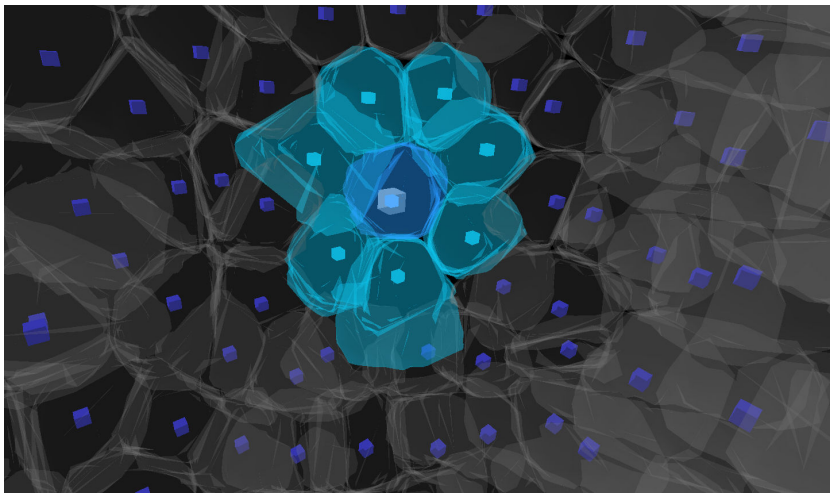


Figure 2.3: Example of a cell with seven neighbors (visualized with the Mov-IT software).

Normalized embryo dynamics

The number of cells, total cellular volume, and total cellular surface area were measured empirically in each embryo of the cohort $\mathscr{E}$, indexed by $e \in [1,\ 5]$ and denoted $N_e(t)$, $N_e^k(t)$, and so on. The evolution over time of these variables was plotted concurrently (Fig. 2.4A,D,G). We observed that the patterns of evolution across the various specimens were similar but did not overlap: some were globally faster, some slower; some were larger, other smaller. To filter out this variability and facilitate interindividual comparison, we applied temporal and spatial rescaling functions.

Temporal rescaling For each observed cell number $N_e(t)$, we considered its inverse function $t_e(N)$ equal to the time at which embryo e contained a given number N of cells (symmetric of Fig. 2.4A). Since $N_e(t)$ is a monotonically increasing function (no cell death during the period considered), so is $t_e(N)$. However, because of the relative synchrony among cell cycles, N_e can remain constant for several time steps (plateaus in Fig. 2.4A). Therefore, to obtain a single-valued function, we defined $t_e(N) = \min\{t \mid N_e(t) = N\}$ over the size interval of e, $[N_{e,\min},\ N_{e,\max}]$ (if N was not part of the observed discrete values, $t_e(N)$ was interpolated). Then, to minimize the discrepancy between the observed time of an embryo, $t_e(N)$, and the mean time of the five specimens, $\langle t \rangle(N)$, an affine transform was

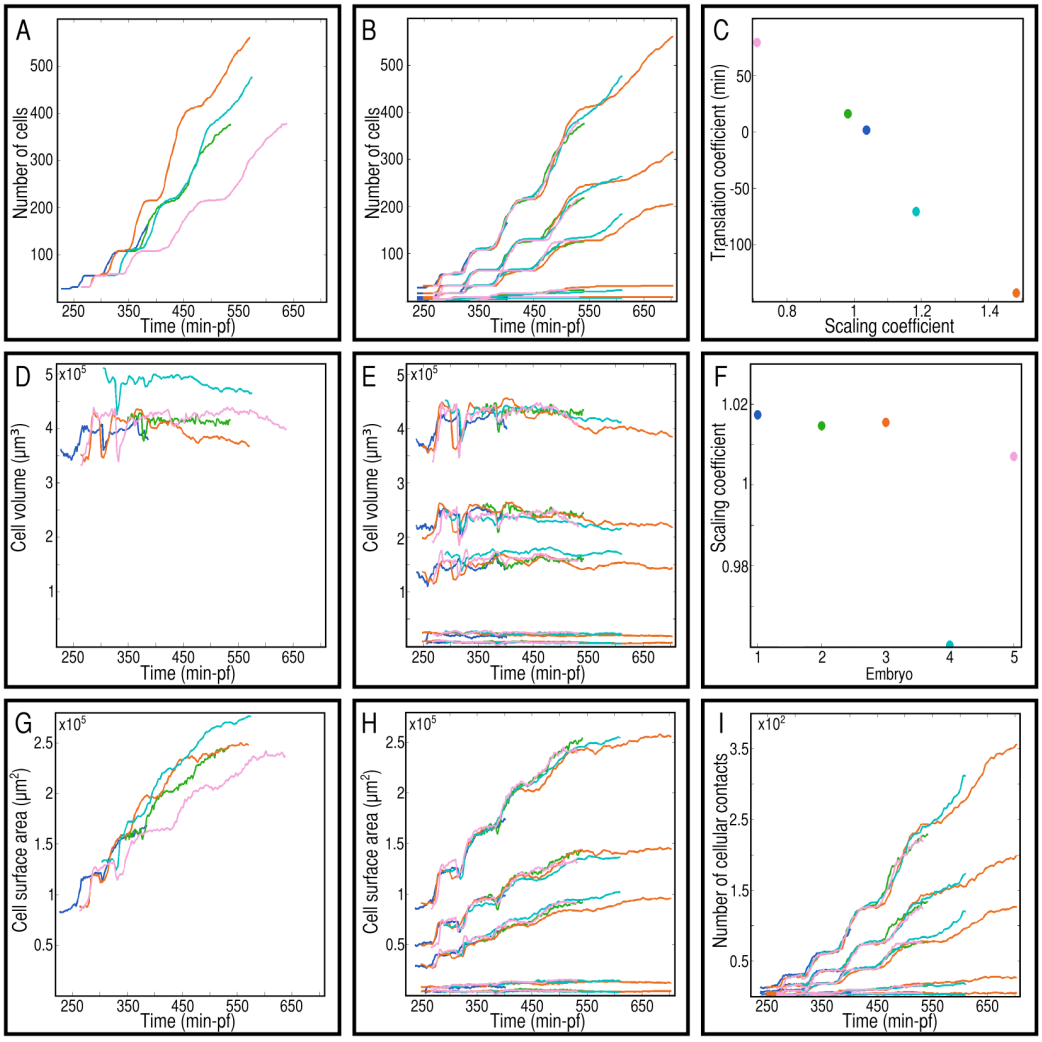


Figure 2.4: Embryo-level dynamics before and after temporal and spatial rescaling. Starting from the 32-cell stage, the global measures of five specimens $e = 1,...,5$ are plotted over their respective imaging periods. A) Number of cells $N_e(t)$ in each embryo. Original periods in minutes post-fertilization (mpf): 225-396 mpf (blue), 345-507 mpf (green), 260-588 mpf (orange), 300-575 mpf (cyan), and 260-638 mpf (pink). B) Temporally rescaled number of cells in each embryo, $N_e(\alpha_e t + \beta_e)$ (top curves), and each morphogenetic field, $N_e^k(\alpha_e t + \beta_e)$ (four groups of lower curves, from top to bottom: Mes, Mac, LMic, SMic). The rescaled periods are: 240-398 mpf (blue), 357-539 mpf (green), 246-702 mpf (orange), 291-619 mpf (cyan), and 265-531 mpf (pink). C) Affine transform parameters (α_e, β_e) used in the previous top curves. D) Total cellular volumes $W_e(t)$. E) Temporally and spatially rescaled total cellular volumes, $(\gamma_e)^3 W_e(\alpha_e t + \beta_e)$ (top), and field cellular volumes $(\gamma_e)^3 W_e^k(\alpha_e t + \beta_e)$ (same order as B), using the same parameters (α_e, β_e) as before. F) Coefficients γ_e used in the previous top curves. G) Total cellular surface areas $Z_e(t)$. H) Temporally and spatially rescaled total cellular surface areas, $(\gamma_e)^2 Z_e(\alpha_e t + \beta_e)$ (top), and field cellular surface areas $(\gamma_e)^2 Z_e^k(\alpha_e t + \beta_e)$ (same order as B), using the same parameters (α_e, β_e) and γ_e as before. I) Temporally rescaled total numbers of contacts $C_e(\alpha_e t + \beta_e)$ and field numbers of contacts $C_e^k(\alpha_e t + \beta_e)$.

performed on the time axis using a cost function:

$$F_e(\alpha,\beta)=\sum_{N=N_{e,\min}}^{N_{e,\max}}\left(\alpha\, t_e(N)+\beta-\langle t\rangle(N)\right)^2 \quad \text{with} \quad \langle t\rangle(N)=\frac{1}{|\mathscr{E}_N|}\sum_{e\in\mathscr{E}_N} t_e(N), \tag{2.3}$$

where $\mathscr{E}_N$ is the sublist of embryos e such that $N\in[N_{e,\min},\ N_{e,\max}]$ (including by interpolation). For each embryo, the parameters adopted for the transform satisfied $(\alpha_e,\beta_e)=\arg\min F_e(\alpha,\beta)$. Finally, taking the inverse again, we obtained five rescaled total numbers $N_e(\alpha_e t+\beta_e)$ and four groups of five rescaled field numbers $N_e^k(\alpha_e t+\beta_e)$, which are plotted in Fig. 2.4B. The five parameter pairs (α_e,β_e) are shown in Fig. 2.4C.

Spatial rescaling Based on the same temporal rescaling, a linear transform was also performed along the spatial dimensions to minimize the discrepancy between the temporally rescaled cellular volume of an embryo, $W_e(\alpha_e t+\beta_e)$ and its mean value over all specimens:

$$G_e(\gamma)=\sum_{t=(t_{e,\min}-\beta_e)/\alpha_e}^{(t_{e,\max}-\beta_e)/\alpha_e}\left(\gamma^3 W_e(\alpha_e t+\beta_e)-\langle W\rangle(t)\right)^2 \quad \text{with} \quad \langle W\rangle(t)=\frac{1}{|\mathscr{E}_t|}\sum_{e\in\mathscr{E}_t} W_e(\alpha_e t+\beta_e), \tag{2.4}$$

where $\mathscr{E}_t$ is the sublist of embryos e such that $(\alpha_e t+\beta_e)\in[t_{e,\min},\ t_{e,\max}]$ (including by interpolation). For each embryo, $\gamma_e=\arg\min G_e(\gamma)$ was then defined as the optimal coefficient. The original volumes $W_e(t)$ are plotted in Fig. 2.4D. The five rescaled total volumes $(\gamma_e)^3 W_e(\alpha_e t+\beta_e)$ and four groups of five rescaled field volumes $(\gamma_e)^3 W_e^k(\alpha_e t+\beta_e)$ each are plotted in Fig. 2.4E, while coefficients γ_e are shown in Fig. 2.4F. Finally, we used the same coefficients squared for the rescaled cellular surface areas, $(\gamma_e)^2 Z_e(\alpha_e t+\beta_e)$ and $(\gamma_e)^2 Z_e^k(\alpha_e t+\beta_e)$ (Fig. 2.4H; the original curves $Z_e(t)$ are shown in Fig. 2.4G).

Neighborhoods The total numbers of contacts in each embryo and each morphogenetic field did not require an extra spatial transform to highlight interindividual similarities. After carrying over the temporal rescaling parameters found above, the curves $C_e(\alpha_e t+\beta_e)$ and $C_e^k(\alpha_e t+\beta_e)$ already presented a high degree of overlap among embryos $e\in[1,\ 5]$ (Fig. 2.4I). This is consistent with the fact that neighborhood relationships are topological features, thus independent from metric deformations.

2.3.2 Intermediate cell groups

Symmetries in the embryo, such as the rotational symmetry around the animal-vegetal (AV) axis, prevent the identification and matching of individual cells from one specimen to another. Unique identification of cells based on their morphological characteristics cannot be done without ambiguity either. To overcome these issues, a generic *coarse-grained level* of observation was needed. We chose here to define new subsets of $\mathscr{L}$ (adding to the list of Section 2.3.1):

– *generational cell groups*, or simply "cell groups", the subsets of generation-n cells inside morphogenetic fields k: $\mathscr{L}^{n,k} = \{i \in \mathscr{L} \mid n_i = n \,\&\, k_i = k\}$.

Note that $\mathscr{L}^{n,k}$ is not the same as $\mathscr{L}^k(t)$: the latter is a snapshot of morphogenetic field k at time t and therefore may contain a mix of generations if some cells have divided more frequently than others. Conversely, cells in the former group may have to be taken at different times. These sets offer two useful viewpoints on the embryo, one focusing on its global state, the other on cell statistics. In any case, the greater n and t, the more distant these two sets are likely to be.

At this intermediate scale, it became possible to identify and map cell groups $\mathscr{L}_e^{n,k}$ across embryos. At the lower level, cells belonging to the same group were expected to display similar individual features. At the sixth generation, each embryo e contained a total of 32 cells composed of 16 Mes, 8 Mac, 4 LMic, and 4 SMic cells, which can be written: $|\mathscr{L}_e^{6,1}| = 16$, $|\mathscr{L}_e^{6,2}| = 8$, $|\mathscr{L}_e^{6,3}| = |\mathscr{L}_e^{6,4}| = 4$ for all $e \in [1,\ 5]$. Since each cell gives rise to two daughter cells at each cycle and there is no cell death, the number of cells of a given type continued doubling, i.e. $|\mathscr{L}_e^{n,k}| = 2^{n-6}|\mathscr{L}_e^{6,k}|$ for $n \geq 6$ during the period of interest.

The six cell features considered here are the life length x_i, mitosis time m_i, average cell volume $\overline{v}_i$, average cell surface area $\overline{s}_i$, and the daughter/mother ratios of average volume a_i and surface area b_i. The dispersion of these features observed in each group $\mathscr{L}_e^{n,k}$ of the empirical data is represented by *distributions*, modeled by sequences of random variables. For example, $(x_1, x_2, ..., x_m)$ denotes the sequence of the values that the first feature, life length, takes in the m cells of a given group $\mathscr{L}_e^{n,k}$. Each x_i is the realization of a random variable X_i pertaining to cell i. Cells being indistinguishable within the same group, however, their order in the sequence is irrelevant. This property is referred to as "exchangeability" of the sequence of random variables and is formalized as follows: for any permutation π of the indices $[1,\ m]$, the joint probability distribution of these variables is the same:

$$P(X_{\pi(1)}, X_{\pi(2)}, ..., X_{\pi(m)}) = P(X_1, X_2, ..., X_m). \tag{2.5}$$

This property leads to de Finetti's theorem, which can be summarized by saying that an infinite sequence of exchangeable random variables is a "mixture" of independent and identically distributed (i.i.d.) sequences [8, 45]. As for a finite exchangeable sequence, it can also be approximated by a mixture of i.i.d. sequences [58]. One consequence of this theorem is that the following empirical measure P_{m,X_g} constitutes a summary statistic for the probability density P of the X sequence:

$$P_{m,X_g}(I) = \frac{1}{m}\sum_{i=1}^{m}\delta_{X_i}(I), \text{ with } \delta_x(I) = 1 \text{ if } x \in I \text{ and } \delta_x(I) = 0 \text{ if } x \notin I, \tag{2.6}$$

where I is an interval of values or union of intervals (i.e. an element of the "σ-algebra" on $\mathbb{R}$) and δ_x is the Dirac measure. If m could increase to ∞, P_{m,X_g} would converge asymptotically to the underlying theoretical probability distribution P of the sequence of exchangeable random variables.

The motivation of the probabilistic model presented here and in the next section was to provide an idealized representation of the dynamics of multicellular development relating the distributions observed in each cell group with the global dynamics of the cell lineage. To simplify notations, let us define an embryo-specific cell group index $g = (e, n, k)$ so that $\mathscr{L}_e^{n,k}$ can be equivalently written $\mathscr{L}_g$. With this, our goal was the following:

- link the distribution in each group $\mathscr{L}_g$ of life lengths: $X_g \sim \{x_i \mid i \in \mathscr{L}_g\}$ and the distribution of mitosis times: $M_g \sim \{m_i \mid i \in \mathscr{L}_g\}$ to the evolution of the total number of cells in the embryo $N_e(t)$
- link the distribution in each group $\mathscr{L}_g$ of average volumes $\overline{V}_g$ and the distribution of daughter/mother volume ratios A_g to the evolution of the total cellular volume $W_e(t)$
- link the distribution in each group $\mathscr{L}_g$ of average surface areas $\overline{S}_g$ and the distribution of daughter/mother surface ratios B_g to the evolution of the total cellular surface area $Z_e(t)$.

2.4 Observation and approximation of multi-level statistics

From the empirical distributions of individual cell features, we derived parametrized representations approximating their probability distributions in each cell group. Combining cell features along the cell lineage required considering simple models of inheritance. As explained below, we calculated correlations based on an assumption of linearity in in-

tercellular dependencies and, obtaining low values, concluded that the individual features of a cell were largely independent from its mother's or sister's features. This independence was then taken as a founding hypothesis of our model.

2.4.1 Estimation of cell feature distributions in cell groups

To interpret the empirical feature distributions, capture their evolution and predict the developmental dynamics, it was best to describe them in terms of parametric models. First, however, because of the finite and relatively short duration of the period of observation, some groups $\mathscr{L}_g$ were incompletely represented in the dataset. Therefore, a distribution of cell features in a group was considered to be significant only if the number of observed cells constituted at least 95% of the full cardinality $|\mathscr{L}_g|$ calculated above. Moreover, certain features, such as the volume v_i and surface area s_i, required the observation of their evolution during whole cell cycles. Incomplete cell cycles were taken into account only if the mean period of observation of individual cell dynamics in a group was at least 20 min. Based on these criteria, we retained 252 distributions of individual cell features across all cell groups $\mathscr{L}_g$ (corresponding to a rough average of 4×3 significant distribution-generation pairs in each one of the 5×4 embryo-cell type fields).

Graphical assessment of these histograms led us to categorize them into two major types: *normal distributions* (i.e. Gaussian curves) for life lengths x_i and mitosis times m_i; and *log-normal distributions* for average cell volumes $\overline{v}_i$, average cell surface areas $\overline{s}_i$, daughter/mother volume ratios a_i and surface area ratios b_i (where a random variable X is said to be log-normally distributed if the random variable $Y = \log X$ is normally distributed).

We performed a chi-squared goodness-of-fit test on each distribution to validate our assessment. This test evaluates the proximity of the empirical frequency with the theoretical one, i.e. whether the random variables are i.i.d. under a normal distribution or a log-normal distribution. To this aim, we first calculated the *empirical mean and standard deviation* of each random variable in each cell group $\mathscr{L}_g$ using a classical maximum likelihood estimation [40]. For example, in the case of life lengths X_g these two quantities

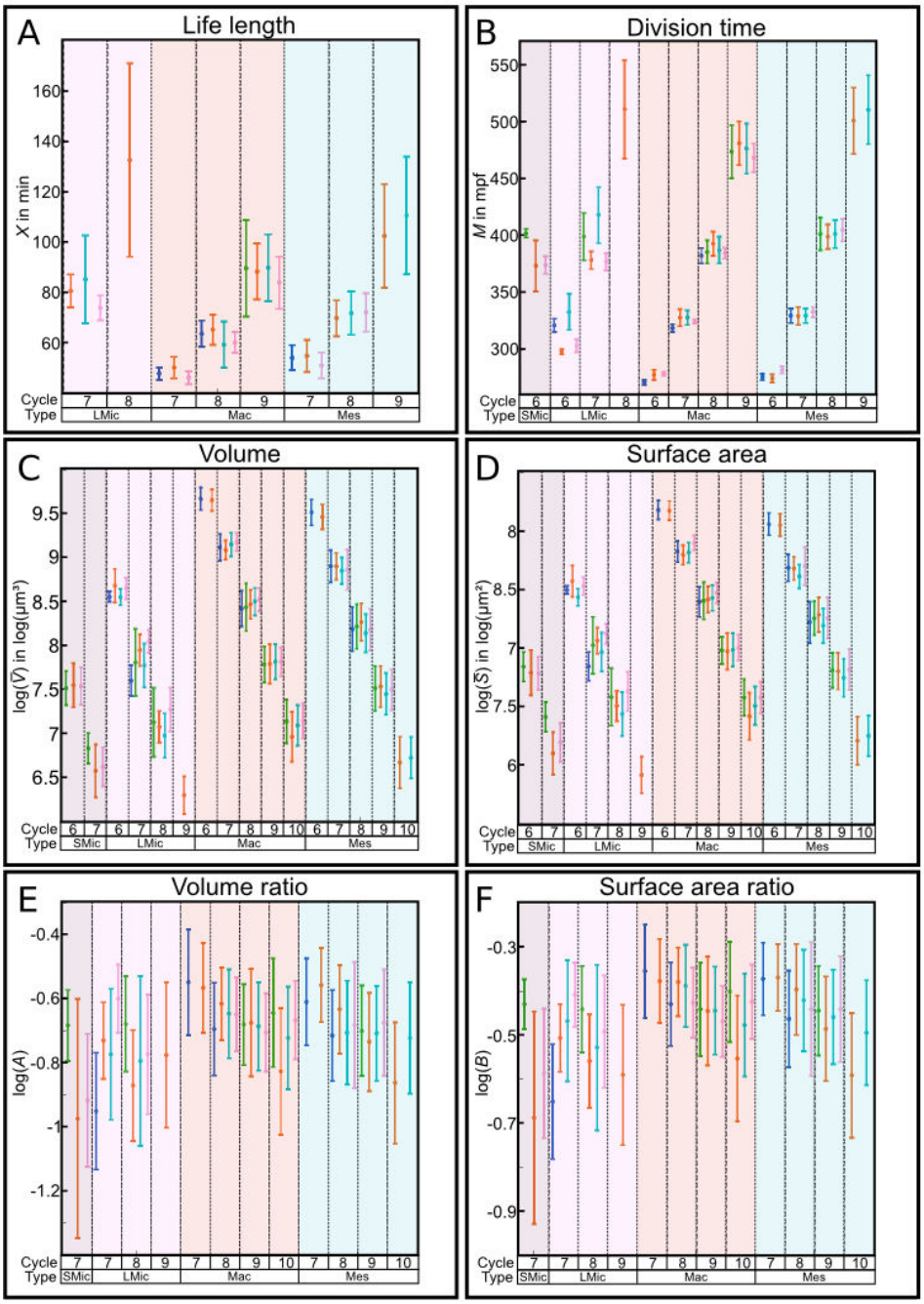


Figure 2.5: Mean and standard deviation bars representing the normal and log-normal approximations of the cell feature distributions in various cell groups $\mathscr{L}_g$. For each feature (one per frame), the cell groups $g = (e, n, k)$ cover all four cell types k (except life length), one or several generations $n \in [6,\ 10]$ per type, and most of the five embryos e per generation-type combination (using the same colors as Fig. 2.4).

are

$$\mu_{X_g} \approx \widetilde{\mu}_{X_g} = \frac{1}{N_g} \sum_{i \in \mathscr{L}_g} x_i \tag{2.7}$$

$$\sigma_{X_g} \approx \widetilde{\sigma}_{X_g} = \left(\frac{1}{N_g - 1} \sum_{i \in \mathscr{L}_g} (x_i - \widetilde{\mu}_{X_g})^2 \right)^{1/2} \tag{2.8}$$

where $N_g = |\mathscr{L}_g|$. The same formulas were applied to variables M_g, $\log \overline{V}_g$, $\log \overline{S}_g$, $\log A_g$, or $\log B_g$, for each one of the 252 available distributions, yielding parameters $(\widetilde{\mu}_{M_g}, \widetilde{\sigma}_{M_g}), ..., (\widetilde{\mu}_{B_g}, \widetilde{\sigma}_{B_g})$. All 252 (μ, σ) pairs are plotted in Fig. 2.5.

Then, based on the empirical mean and standard deviation, our categorization hypothesis was challenged in each distribution by calculating a p-value corresponding to the probability to find a good fit with a normal or log-normal curve. Taking again X_g as an example, the distribution of life lengths $\{x_1, x_2, ..., x_m\}$ over the m cells of group $\mathscr{L}_g$ was categorized into a fixed number l of discrete bins, $(\hat{x}_1, \hat{x}_2, ..., \hat{x}_l)$, where $\hat{x}_1 = \min\{x_i\}$ and $\hat{x}_{h+1} - \hat{x}_h = \Delta\hat{x} = (\max\{x_i\} - \min\{x_i\})/l$ for all $h \in [1,\ l-1]$. In this histogram, the *observed* distribution corresponds to the number of values falling in each bin and was given by the empirical measure $P_{m,X_g}(I_h)$ (equation 2.6), where $I_h = [\hat{x}_h,\ \hat{x}_h + \Delta\hat{x})$ and I_l is closed, while the *expected* distribution was given by $P_{m,\widetilde{X}_g}(I_h)$, where $\widetilde{X}_g$ is the theoretical normal distribution calculated from $(\widetilde{\mu}_{X_g}, \widetilde{\sigma}_{X_g})$. The chi-squared statistic computes the discrepancy between both distributions as follows:

$$\chi^2_g = \sum_{h=1}^{l} \frac{(P_{m,X_g}(I_h) - P_{m,\widetilde{X}_g}(I_h))^2}{P_{m,\widetilde{X}_g}(I_h)}. \tag{2.9}$$

This statistic was assumed to follow a chi-squared distribution with $\kappa = l - 3$ degrees of freedom [145] (the number of frequencies, l, reduced by the number of parameters of the fitted distribution, μ and σ, and the constraint $\sum_h P_{m,X_g}(I_h) = 1$). This gave the p-value, which is the probability of observing a test statistic *at least* as extreme, i.e. 1 minus the cumulative distribution function of χ^2_g for κ degrees of freedom. Finally, our Gaussian-fit model was deemed adequate for a given feature in a group $\mathscr{L}_g$ if its p-value was greater than 0.05 (Fig. 2.6).

Note that the chi-squared test of goodness-of-fit produces inaccurate results if the size of the sample is too small—i.e., by convention, less than 5 elements in a bin [145]. Reducing the number of frequencies l increases the number of elements in each bin, but also decreases the number of degrees of freedom κ, which must remain greater than 3 for the

test to be applicable here. Based on these constraints, 128 distributions out of 252 had to be excluded from the evaluation.

For most of the remaining 124 distributions, the validity of our assumptions about their type (normal or log-normal) could be retained. Only 20 of them had a p-value under 0.05, among which 12 were under 0.01 (Fig. 2.7). The complete ranges of p-values obtained were the following: [3.7e-8, 0.42] for life lengths x_i, [3.6e-6, 0.46] for mitosis times m_i, [8.5e-3, 0.89] for average cell volumes $\overline{v}_i$, [5e-3, 0.96] for average cell surface areas $\overline{s}_i$, [0.06, 0.84] for average volume ratios a_i, and [1.3e-3, 0.76] for average surface area ratios b_i. The histogram of all 124 p-values is plotted in Fig. 2.8.

Ideally, the worst-fit cases could be used to subdivide cell groups $\mathscr{L}_g$ into smaller classes (e.g., oral and aboral cells) reflecting the multimodal aspect of certain distributions, or to adopt different parametric models (e.g., exponential instead of Gaussian curves). Given the small size of the cohort, however, and the non-reproducibility of these outlier distributions in various specimens, it was difficult to rely on such deviations to define new cell types or new models. Moreover, biological relevance of the inferred cell clusters would have to be validated by correlation with genetic expression in the form of a developmental "atlas" [41], which is out of the scope of this study. This is why we preferred keeping the simplicity of normal and log-normal approximations as they captured the essential characteristics of statistical distributions (their mean and variance) even in marginal cases. For now, we left out any potential additional information for the sake of understandability and interpretability of the whole dataset. The description of distribution shapes can be refined in future work by using a larger cohort of specimens.

2.4.2 Cell volume and surface area dynamics

The volume and surface area of a cell are not constant during its life length because of various contractions and dilations. To highlight the fluctuations of the cellular geometry around its average, we define two other cell quantities:

- the *normalized volume*: $\omega_i(u) = v_i(t)/\overline{v}_i$
- the *normalized surface*: $\phi_i(u) = s_i(t)/\overline{s}_i$

where $t = m_j + ux_i$ (m_j is the mother's mitosis time) and $u \in [0,\ 1]$ represents the cell's normalized age, which can be equivalently defined by $t = m_i - (1-u)x_i$. Then, the microdynamics of cell geometry can be described by taking the mean of these quantities in each

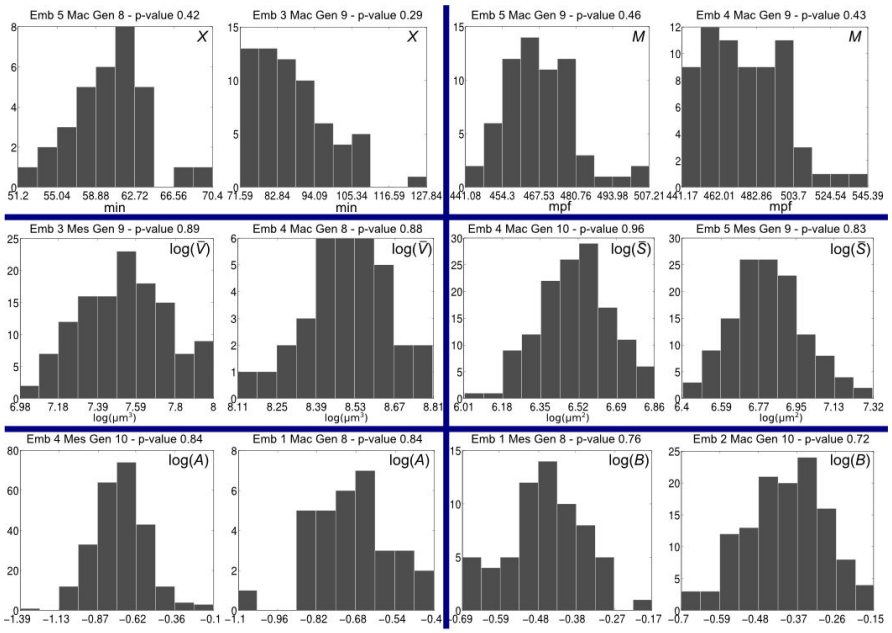


Figure 2.6: The 12 cell feature distributions (two per variable type) with the highest p-value, i.e. the best cases of fit by normal or log-normal curves with respect to the chi-squared test. Left to right, top to bottom: the two X_g and two M_g distributions closest to normal curves; the two $\overline{V}_g$, two $\overline{S}_g$, two A_g and two B_g distributions closest to log-normal curves. Titles indicate the embryo e, cell type k, generation n and p-value.

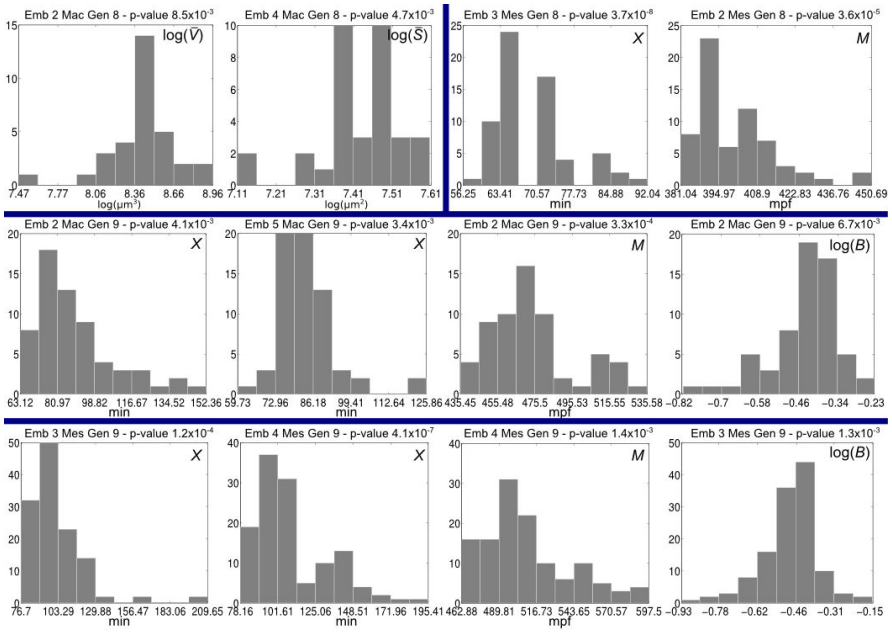


Figure 2.7: The 12 worst-fit cases (p-value < 0.01) of cell feature distributions, where the assumption of normal or log-normal curves could not be retained with respect to the chi-squared goodness-of-fit test. They fall into four n, k categories: 8, Mac; 8, Mes; 9, Mac; and 9, Mes. Most of them concern X and M.

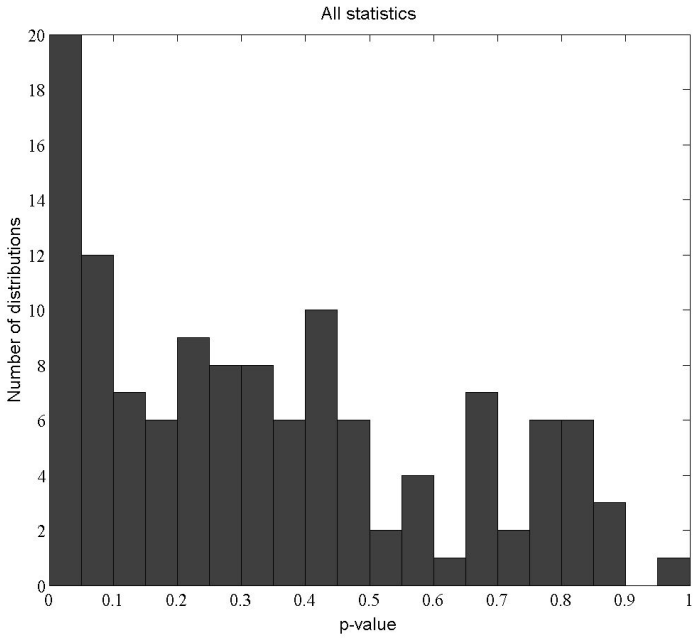


Figure 2.8: Histogram of all p-values obtained by chi-squared goodness-of-fit test on 124 distributions of cell features. The six variables considered here were X_g, M_g, $\log \overline{V}_g$, $\log A_g$, $\log \overline{S}_g$ and $\log B_g$ across the five embryos e, four cell types k, and all generations n within the period during which the cell feature was observable.

group $\mathscr{L}_g$ (Fig. 2.9,2.10):

$$\widetilde{\omega}_g(u) = \frac{1}{N_g} \sum_{i \in \mathscr{L}_g} \omega_i(u) \quad \text{and} \quad \widetilde{\phi}_g(u) = \frac{1}{N_g} \sum_{i \in \mathscr{L}_g} \phi_i(u), \tag{2.10}$$

making the assumption that the functions ω_g and ϕ_g are essentially deterministic, i.e. the variations in volume and surface *normalized in time and amplitude* are the same for all cells $i \in \mathscr{L}_g$ (hence the notations $\widetilde{\omega}_g, \widetilde{\phi}_g$ instead of $\widetilde{\mu}_{\Omega_g}, \widetilde{\mu}_{\Phi_g}$). Their particularity compared to the other six empirical means calculated previously (equation 2.7) is their dependency on time (normalized to align the signals).

2.4.3 Independence along the lineage

Having defined parametric representations $(\widetilde{\mu}, \widetilde{\sigma})$ for each cell feature distribution within each cell group $\mathscr{L}_g$, we wanted to investigate the *relationships* among these variables. Since cells are genealogically related, it was natural to hypothesize some degree of cor-

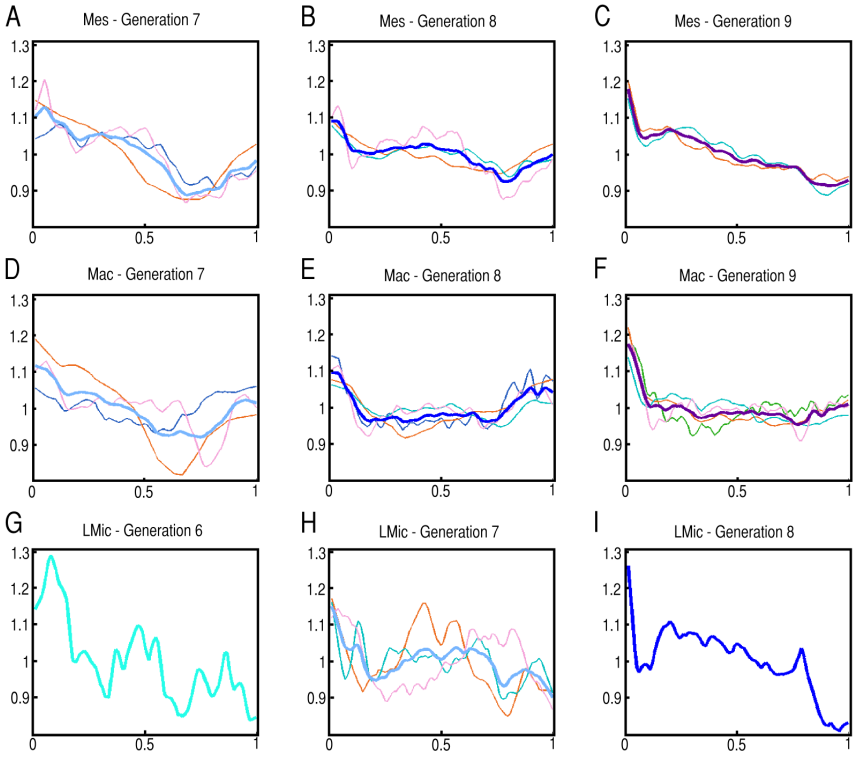


Figure 2.9: Estimated cell volume microdynamics $\widetilde{\omega}_g(u)$ in various groups $\mathscr{L}_g$ of types $k = 1,2,3$ and generations $n = 6,7,8,9$. Thin lines correspond to individual embryos e of the cohort $\mathscr{E}$ (one or several per group). Thick lines represent cohort averages $\langle \widetilde{\omega}_g(u) \rangle_{e \in \mathscr{E}}$.

relation between cells belonging to the same descent. With the goal to find a reduced number of parameters describing the phenomenology of the process, we assumed *linear dependencies* between the cell distributions of daughters and mothers, and among sisters. This is written $Y = \lambda + \mu X$, where the scalar parameters λ and μ can be estimated by linear regression based on a set of N realizations of the random variables X and Y: $\{(x_i, y_i)\}_{i=1,\dots,N}$. The empirical optimal values $\widetilde{\lambda}$ and $\widetilde{\mu}$ are the ones that minimize the *residual sum of squares* over the samples:

$$SS_{\text{res}} = \sum_{i=1}^{N} (y_i - (\widetilde{\lambda} + \widetilde{\mu} x_i))^2 = \sum_{i=1}^{N} \epsilon_i^2, \tag{2.11}$$

where ϵ_i denotes the residual error terms such that $y_i = \widetilde{\lambda} + \widetilde{\mu} x_i + \epsilon_i$ and is assumed to be normally distributed around zero. Then, to characterize the accuracy of this linear estimation, we used the *coefficient of determination*, R^2, which measures the percentage of

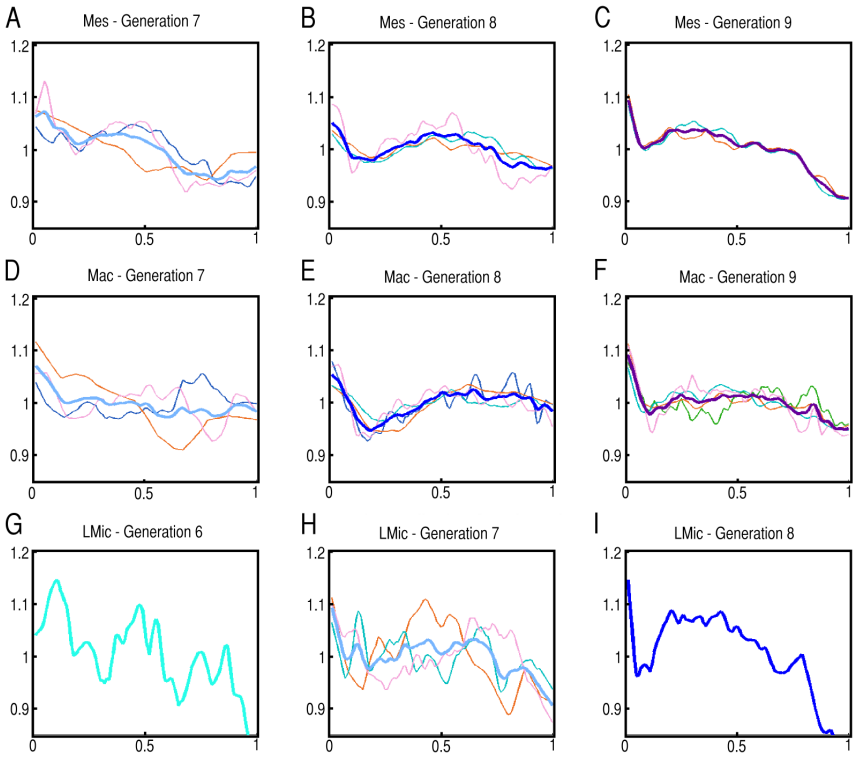


Figure 2.10: Estimated cell surface microdynamics $\widetilde{\phi}_g(u)$ in various groups $\mathscr{L}_g$ of types $k = 1,2,3$ and generations $n = 6,7,8,9$. Thin lines correspond to individual embryos e of the cohort $\mathscr{E}$ (one or several per group). Thick lines represent cohort averages $\langle \widetilde{\phi}_g(u) \rangle_{e \in \mathscr{E}}$.

variation of one variable explained linearly by the other [183, 40]. If we define the *total sum of squares* by

$$SS_{\text{tot}} = \sum_{i=1}^{N} (y_i - \widetilde{y})^2 \quad \text{with} \quad \widetilde{y} = \frac{1}{N} \sum_{i=1}^{N} y_i, \tag{2.12}$$

where $\widetilde{y}$ is the empirical mean of Y, then the percentage of variation that remains unexplained linearly is given by the ratio $SS_{\text{res}}/SS_{\text{tot}}$, and the coefficient of determination is equal to its complement:

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}. \tag{2.13}$$

For a "simple" linear regression (i.e. one with a single explanatory variable), it can be shown that the coefficient of determination is equal to the square of the estimated Pearson's coefficient of correlation:

$$R^2 = \widetilde{\rho}^2_{X,Y}, \text{ where } \rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\mu_{XY} - \mu_X \mu_Y}{\sigma_X \sigma_Y}. \tag{2.14}$$

The greater R^2, the more X and Y tend to be linearly dependent, with a binary decision threshold generally set at 0.6. Note that R^2 does not measure nonlinear dependencies, thus can remain low even if X and Y are strongly related through quadratic or logarithmic curves, for example.

Here, this coefficient was applied to six pairs of features calculated among the 252 distributions deemed sufficiently represented in the dataset (Section 2.4.1, Fig. 2.5). In the following, for a group index $g = (e, n, k)$ we use the shorthand notation $g' = (e, n-1, k)$ to refer to the previous-generation group in the same embryo.

Daugther's life length vs. mother's mitosis time As above, we assumed a simple model of linear dependency between the life length x_i of a cell $i \in \mathscr{L}_g$ and the mitosis time m_j of its mother $j \in \mathscr{L}_{g'}$: $x_i = \lambda + \mu m_j$. The coefficients of determination $R^2 = \widetilde{\rho}^2_{X_g, M_{g'}}$ were computed between the distributions $X_g \sim \{x_i \mid i \in \mathscr{L}_g\}$ and $M_{g'} \sim \{m_j \mid j \in \mathscr{L}_{g'}\}$ at several generations n. For this calculation, we used the 23 groups $\mathscr{L}_g$ of Fig. 2.5A in which all cell cycles were completely known, i.e. had observed start and end mitosis times, and their corresponding 23 groups $\mathscr{L}_{g'}$. In the end, the values of R^2 that we obtained fell in the range [3e-3, 0.51], which indicated only weak or inexistant linear dependencies.

Sisters' life lengths For each pair of sister cells $i_1, i_2 \in \mathscr{L}_g$, ordered such that $x_{i_1} > x_{i_2}$, we also assumed $x_{i_1} = \lambda + \mu x_{i_2}$, where the longer life length was included in a set $X_{g,1}$ and the shorter one in another set $X_{g,2}$. The coefficient of determination $R^2 = \widetilde{\rho}^2_{X_{g,1}, X_{g,2}}$ was then computed in the same 23 groups and the values obtained belonged to the interval [0.095, 0.77], revealing some linear dependency among sisters' life lengths within certain groups. More precisely, R^2 coefficients scored above 0.6 in the following seven groups $g = (e, n, k)$: (4, 7, Mac), (5, 7, Mac), (3, 8, Mac), (4, 8, Mac), (1, 7, Mes), (3, 7, Mes) and (3, 9, Mes). In the other 16 groups, there was no clear linear dependency.

Sisters' volume ratios Concerning the volume and surface area, several models of dependency can be explored. Usually, it is assumed that the volume of a mother is conserved in its two daughters through mitosis, meaning that the average volumes should verify: $\overline{v}_{i_1} + \overline{v}_{i_2} = \overline{v}_j$. Dividing by $\overline{v}_j$, this is equivalent to a linear relationship $a_{i_1} = 1 - a_{i_2}$ between the average daugther/mother volume ratios of the sister cells. To verify if there is such a dependency, we computed the coefficient of determination $R^2 = \widetilde{\rho}^2_{A_{g,1}, A_{g,2}}$ as above between pairs of sister cells $i_1, i_2 \in \mathscr{L}_g$ ordered such that $a_{i_1} > a_{i_2}$. Over the 38 cell groups $\mathscr{L}_g$ of Fig. 2.5C in which all volumes could be confidently measured, the coefficients of

determination belonged to [1.4e-3, 0.88], with 34 values below 0.61 and the other four values in [0.75, 0.88] corresponding to cell groups (3, 7, SMic), (3, 7, Mac), (4, 8, Mac) and (1, 7, Mes). However, the $\widetilde{\lambda}$ and $\widetilde{\mu}$ parameters for these groups were found in $[-0.45, 0.12]$ and [0.87, 3.02] respectively, i.e. far from 1 and -1. Altogether, these results largely invalidated the hypothesis of volume conservation across mitosis.

Daughter's volume ratio vs. mother's volume It was also legitimate to ask whether the average daughter/mother volume ratio a_i of a cell $i \in \mathscr{L}_g$ and the mean volume $\overline{v}_j$ of its mother $j \in \mathscr{L}_{g'}$ were correlated. However, the coefficients of determination R^2 computed in the same 38 groups as above ranged in [1.9e-4, 0.43], indicating no linear relationship.

Sisters' surface ratios; Daughter's surface ratio vs. mother's surface Finally, the same tests were performed for surface area features, yielding similar negative results. All relationships are summarized in Table 2.1 and the histogram of all the coefficients of determination evaluated is shown on Fig. 2.11.

Overall, given the weak values of R^2 obtained for the above six relationships of individual features in the different cell groups (R^2 was greater than 0.6 in only 20 cases out of 198 investigated pairs of distributions), we adopted the viewpoint that the various random variables were independent between and within cell groups. Although some of these correlation values may be ascribed to underlying deterministic factors responsible for cell-to-cell variability [193], our goal in the present work is to find parameters that can summarize relationships between random variables, hence the poor statistical significance of the linear regression test led us to a global assumption of independence. The following section examines the relations between probability laws in the different cell groups and how they are combined in the cell lineage.

2.5 Multi-level probabilistic model

In this section we use the parametrized description of the distributions of individual cell features obtained for each cell group in the preceding section and their combination in the cell lineage. The individual cell features considered are the life length, the mitosis time, the volume and the surface area. The resulting probabilistic model predicts accurately the final distribution of individual cell features comparatively to inter-individual dif-

daughter $i \in \mathscr{L}_g$	mother $j \in \mathscr{L}_{g'}$	sister $i_1 \in \mathscr{L}_g$	sister $i_2 \in \mathscr{L}_g$	$R^2_{\min}$	$R^2_{\max}$	#groups g such that $R^2 > 0.6$
x_i	m_j			3e-3	0.51	0 of 23
		x_{i_1}	$> x_{i_2}$	0.095	0.77	7 of 23
		a_{i_1}	$> a_{i_2}$	1.4e-3	0.88	0 of 38
a_i	$\overline{v}_j$			1.9e-4	0.43	5 of 38
		b_{i_1}	$> b_{i_2}$	0.02	0.81	0 of 38
b_i	$\overline{s}_j$			1.4e-5	0.39	8 of 38

Table 2.1: Summary of the linear regression test for six pairs of features across 198 distributions in 61 groups.

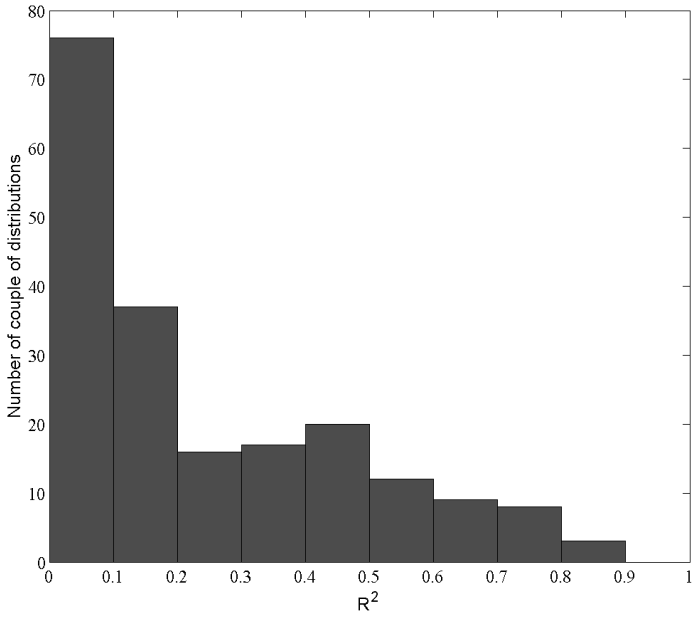


Figure 2.11: Histogram of the coefficient of determination $R^2 = \widetilde{\rho}^2_{X,Y}$ for the various pairs of distributions X and Y displayed in Table 2.1. Only 20 values rank above 0.6.

ferences among the cohort. It reproduces embryo-level dynamics within each specimen of the cohort (number of cells, total cellular volume, total cellular surface area). Moreover it provides a model of how intra-individual variability propagates across the cell lineage.

Proliferation dynamics

To model the dynamics and evolution of the cell lineage, the relations between the moment of division and the life length of each cell in the cell lineage have to be established. $X_{n,k}$ is the random variable describing the cell cycle length (time between two consecutive mitoses) within the group $\mathscr{L}^{n,k}$, and $M_{n,k}$ is the random variable describing the mitosis time in $\mathscr{L}^{n,k}$. Based on these notations, we propose the following model, derived from the analysis of the data:

$$M_{n,k} = X_{n,k} + M_{n-1,k} \tag{2.15}$$

$$M_{n,k} \sim N(\mu_M^{n,k}, \sigma_M^{n,k}) \tag{2.16}$$

$$X_{n,k} \sim N(\mu_X^{n,k}, \sigma_X^{n,k}) \tag{2.17}$$

$$P(X_{n,k} \mid M_{n-1,k}) = P(X_{n,k}) \tag{2.18}$$

where N represents a normal distribution of mean μ and variance σ^2 [40]. Equation 2.18 shows that $X_{n,k}$ and $M_{n-1,k}$ are assumed to be independent from each other, P represents the probability law governing $X_{n,k}$.

This model is instantiated independently in each branch of the cell lineage with the following relation among mitosis times:

$$m_i = m_j + x_i \tag{2.19}$$

where $i \in \mathscr{L}^{n,k}$ and $j \in \mathscr{L}^{n-1,k}$ are the indices of a cell and its mother, m_i is a realization of $M_{n,k}$ and represents the mitosis time of cell i, x_i is a realization of $X_{n,k}$ and represents the life length of cell i, m_j is a realization of $M_{n-1,k}$ and represents the mitosis time of cell j.

We can relate the parameters governing the different distributions:

$$M_{n,k} \sim N\left(\mu_M^{n,k}, \sigma_M^{n,k}\right) \tag{2.20}$$

$$\mu_M^{n,k} = \sum_{r=n^{0,k}+1}^{n} \mu_X^{r,k} + \mu_M^{n^{0,k},k} \tag{2.21}$$

$$(\sigma_M^{n,k})^2 = \sum_{r=n^{0,k}+1}^{n} (\sigma_X^{r,k})^2 + (\sigma_M^{n^{0,k},k})^2 \tag{2.22}$$

where $n^{0,k}$ indicates the initial cycle at which the cells of type k are found. The variability

of mitosis times, represented by the variance of the Gaussian distribution is the sum of the variability of cycle lengths in each cell group along the lineage tree. Similarly, the mean mitosis time is the sum of mean cell cycle lengths over the preceding cell cycles in the lineage.

Proof. Let's first establish a recurrence relation between the random variables:

$$\begin{aligned} M_{n,k} &= X_{n,k} + M_{n-1,k} \\ &= X_{n,k} + X_{n-1,k} + M_{n-2,k} \\ &= X_{n,k} + X_{n-1,k} + \ldots + X_{n^{0,k}+1,k} + M_{n^{0,k},k} \end{aligned}$$

The observation window of the developing embryo in our data is finite. The observation begins at the 32 cells stage at the earliest, i.e. with cells found at the 6th cell cycle in each of the four subpopulations. The observation ends at the 408 cell stage at the latest, i.e. with cells found at the 10th cell cycle for the Mesomeres and the Macromeres, at the 9th cell cycle for the Large Micromeres and at the 7th cell cycle for the Small Micromeres. Since the recurrence cannot be traced back to the origin, we have to provide a boundary condition for the initial distribution of mitosis time, $M_{n^{0,k},k}$.

The first distribution of mitosis time $M_{n^{0,k},k}$ is a Gaussian distribution:

$$M_{n^{0,k},k} \sim N(\mu_M^{n^{0,k},k}, \sigma_M^{n^{0,k},k}) \tag{2.23}$$

where $n^{0,k}$ indicates the initial cycle at which the cells of type k are found.

The probability distribution of mitosis time, $M_{n,k}$, is obtained recursively from the parameters of the cycle length distribution and the fact that the density of two independent random variables is the convolution of their densities. Since all the random variables are independent, we obtain the following relations between the probability densities:

$$f_{M_{n,k}} = f_{X_{n,k}} * f_{M_{n-1,k}} \tag{2.24}$$

$$f_{M_{n,k}} = f_{X_{n,k}} * f_{X_{n-1,k}} * f_{M_{n-2,k}} \tag{2.25}$$

$$f_{M_{n,k}} = f_{X_{n,k}} * f_{X_{n-1,k}} * \ldots * f_{X_{n^{0,k}+1,k}} * f_{M_{n^{0,k},k}} \tag{2.26}$$

where $*$ denotes the convolution product and f_{RV} denotes the probability density of a random variable RV.

In the case of Gaussian distributions, the convolution of two distributions is another

Gaussian whose mean and variance are the sums of their respective components. [32]. The probability density of a Gaussian distribution is $f(x \mid \mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}.\exp(-\frac{(x-\mu)^2}{2.\sigma^2})$. We obtain for two Gaussian distributions with parameters (μ_1,σ_1 and μ_2,σ_2):

$$f(x \mid \mu_1,\sigma_1) * f(x \mid \mu_2,\sigma_2) = f(x \mid \mu_1+\mu_2, \sqrt{\sigma_1^2+\sigma_2^2}) \tag{2.27}$$

Each distribution of the cell cycle lengths are Gaussian distributions, $\forall n,k, f_{X_{n,k}} = f(x \mid \mu_X^{n,k}, \sigma_X^{n,k})$ and $f_{M_{n^{0,k},k}} = f(x \mid \mu_M^{n^{0,k},k}, \sigma_M^{n^{0,k},k})$. Consequently, combining equation 2.26 with equation 2.27, we obtain:

$$f_{M_{n,k}}(x) = f(x \mid \sum_{r=n^{0,k}+1}^{n} \mu_X^{r,k} + \mu_M^{n^{0,k},k}, \sqrt{\sum_{r=n^{0,k}+1}^{n} (\sigma_X^{r,k})^2 + (\sigma_M^{n^{0,k},k})^2}) \tag{2.28}$$

□

This result guarantees that the knowledge of the distributions of cell cycle lengths in each cell group is sufficient to characterize the distributions of mitosis time for each of them.

Cell volume dynamics

The cell volume varies throughout cell proliferation. The variation of volume is decomposed into a stochastic and a deterministic component. The stochastic component concerns the evolution of the mean volume of the cells along the cell lineage and the deterministic component serves at modulating the cell volume around its mean by contraction and expansion throughout the cell cycle.

To model the first component, we use a random variable $\overline{V}_{n,k}$ describing the mean cell volume in the group $\mathscr{L}^{n,k}$, and another, $A_{n,k}$, describing the daughter/mother mean volume ratio in the same group. We propose the following model that relate the mean volume and the daughter/mother ratio:

$$\overline{V}_{n,k} = A_{n,k}\overline{V}_{n-1,k} \tag{2.29}$$

$$\overline{V}_{n,k} \sim lnN(\mu_V^{n,k}, \sigma_V^{n,k}) \tag{2.30}$$

$$A_{n,k} \sim lnN(\mu_A^{n,k}, \sigma_A^{n,k}) \tag{2.31}$$

$$P(A_{n,k} \mid \overline{V}_{n-1,k}) = P(A_{n,k}) \tag{2.32}$$

where lnN represents the log-normal distribution with parameters μ and σ, $A_{n,k}$ being log-normally distributed is equivalent to $\log(A_{n,k})$ being normally distributed with mean μ and variance σ^2 [131, 40]. Equation 2.32 shows that the random variables $A_{n,k}$ and $\overline{V}_{n-1,k}$ are assumed to be independent, P is the probability law governing $A_{n,k}$.

This model is instantiated in each branch of the cell lineage with the following relation, for a cell $i \in \mathscr{L}^{n,k}$ with mother $j \in \mathscr{L}^{n-1,k}$:

$$\overline{v}_i = a_i \overline{v}_j \tag{2.33}$$

where a_i represents the ratio of a cell mean volume and its mother mean volume, it is a realization of the random variable $A_{n,k}$. $\overline{v}_j$ and $\overline{v}_i$ are realizations of $\overline{V}_{n,k}$ and $\overline{V}_{n-1,k}$ and represent the mean volume of cell i and j respectively.

The volume of a cell is not constantly equal to its mean value during the cell cycle. The deterministic function $\omega^{n,k}$ to which we refer as the volume micro dynamic model the variation of volume which occurs between two consecutive mitoses. This function is defined for each cell group $\mathscr{L}^{n,k}$ and represents the expansion and contraction that a cell volume undergoes during its cell cycle. Its domain is the interval $[0,1]$ which is the percentage of elapsed cell cycle (the function is equal to 0 outside this interval) and its range is the interval $[0.8,1.3]$ and it is a normalized function: $\int_0^1 \omega^{n,k} = 1$. Fig. 2.9 shows the estimated shape of $\omega^{n,k}$ for the different cell groups.

Using this function $\omega^{n,k}$ and keeping the same notations as in the equation 2.33, the volume v_i is obtained at each time step t for a cell $i \in \mathscr{L}^{n,k}$ with mother $j \in \mathscr{L}^{n-1,k}$:

$$v_i(t) = \overline{v}_i \omega^{n,k}(u_i(t)) \tag{2.34}$$

$$= a_i \overline{v}_j \omega^{n,k}(u_i(t)) \tag{2.35}$$

where the function u transports the time t to values in the interval [0,1] which is the domain of $\omega^{n,k}$ and describe the percentage of elapsed life length. u is defined for cell i as $u_i(t) = \frac{t-m_j}{m_i-m_j}$ if $m_j \leq t \leq m_i$ and $u_i(t) = 0$ otherwise.

The parameters of the various probability laws can be related in the following way:

$$\overline{V}_{n,k} \sim lnN\left(\mu_V^{n,k}, \sigma_V^{n,k}\right) \tag{2.36}$$

$$\mu_V^{n,k} = \sum_{r=n^{0,k}+1}^{n} \mu_A^{r,k} + \mu_V^{n^{0,k},k} \tag{2.37}$$

$$(\sigma_V^{n,k})^2 = \sum_{r=n^{0,k}+1}^{n} (\sigma_A^{r,k})^2 + (\sigma_V^{n^{0,k},k})^2 \tag{2.38}$$

where $n^{0,k}$ indicates the initial cycle at which the cells of type k are found. As for the mitosis times, the variability of the volume $(\sigma_V^{n,k})^2$ is the sum of the variabilities in the daughter/mother coefficients, $(\sigma_A^{r,k})^2$, along the lineage tree with the initial variability $(\sigma_V^{n^{0,k},k})^2$.

Proof. Using the initial distribution and the parameters of the daughter/mother mean volume ratio $\mu_A^{n,k}, \sigma_A^{n,k}$ at each subsequent cell cycle, the mean volume distribution is calculated with a recurrence relation and the property that the product of two independent random variables log-normally distributed is log-normally distributed.

We have the recurrence:

$$\overline{V}_{n,k} = A_{n,k}\overline{V}_{n-1,k} \tag{2.39}$$

$$= A_{n,k}A_{n-1,k}\overline{V}_{n-2,k} \tag{2.40}$$

$$= A_{n,k}A_{n-1,k}...A_{n^{0,k}+1,k}\overline{V}_{n^{0,k},k} \tag{2.41}$$

we can take the logarithm:

$$\log(\overline{V}_{n,k}) = \log(A_{n,k}) + \log(\overline{V}_{n-1,k}) \tag{2.42}$$

$$= \log(A_{n,k}) + \log(A_{n-1,k}) + \log(\overline{V}_{n-2,k}) \tag{2.43}$$

$$= \log(A_{n,k}) + \log(A_{n-1,k}) + ... + \log(A_{n^{0,k}+1,k}) + \log(\overline{V}_{n^{0,k},k}) \tag{2.44}$$

The last equation can be re-written:

$$\log(\overline{V}_{n,k}) = \sum_{r=n^{0,k}+1}^{n} \log(A_{r,k}) + \log(\overline{V}_{n^{0,k},k}) \tag{2.45}$$

where $\forall n,k, \log(A_{n,k}) \sim N(\mu_A^{n,k}, \sigma_A^{n,k})$ and $\forall n,k, \log(\overline{V}_{n,k}) \sim N(\mu_V^{n,k}, \sigma_V^{n,k})$.

All the random variables are independent. As already stated in equation 2.27, the sum of two independent random variables governed by a Gaussian distribution is a Gaussian

Mean volume		Mean surface area
$\overline{V}_{n,k} \sim lnN(\mu_V^{n,k}, \sigma_V^{n,k})$	$\longleftrightarrow$	$\overline{S}_{n,k} \sim lnN(\mu_S^{n,k}, \sigma_S^{n,k})$
Mother/daughter volume ratio		Mother/daughter surface ratio
$A_{n,k} \sim lnN(\mu_A^{n,k}, \sigma_A^{n,k})$	$\longleftrightarrow$	$B_{n,k} \sim lnN(\mu_B^{n,k}, \sigma_B^{n,k})$
Volume micro dynamic		Surface area micro dynamic
$\omega^{n,k}$	$\longleftrightarrow$	$\phi^{n,k}$

Table 2.2: Identification between variables describing cell volume and cell surface area in any cell group $\mathscr{L}^{n,k}$

distribution whose parameters are the sum of the parameters of the initial distributions. Therefore we obtain:

$$\log(\overline{V}_{n,k}) \sim N(\mu_V^{n,k}, \sigma_V^{n,k}) \tag{2.46}$$

$$i.e. \quad \overline{V}_{n,k} \sim lnN(\mu_V^{n,k}, \sigma_V^{n,k}) \tag{2.47}$$

$$\mu_V^{n,k} = \sum_{r=n^{0,k}+1}^{n} \mu_A^{r,k} + (\mu_V^{n^{0,k},k}) \tag{2.48}$$

$$(\sigma_V^{n,k})^2 = \sum_{r=n^{0,k}+1}^{n} (\sigma_A^{r,k})^2 + (\sigma_V^{n^{0,k},k})^2 \tag{2.49}$$

□

Cell surface area dynamics

Cell surface areas and cell volumes are modeled in the same way, identifications between the quantities involved in their model are summarized in the table 2.2:

Model summary

The description of the model is summarized in tables 3.1 and 3.2. Table 3.1 describes the relations between the random variables at the level of the cell groups and table 3.2 describe how the model is instantiated in each branch of the cell lineage tree.

A cell lineage is computed with this model by beginning with the 32 cells stage (16 Mes, 8 Mac, 4 LMic, and 4 SMic), using the fact that each cell divide into two and the relations described in table 3.2 to generate the microscopic features. Relying on the cell lineage we can compute macroscopic dynamics in each morphogenetic field, namely the number of

cells $N^k(t)$, the cellular volume $W^k(t)$, and the cellular surface area $Z^k(t)$, as described in section 2.3.1. These macroscopic features are themselves random variables. However, we will not give an analytical expression of their probability distribution since they are difficultly tractable and will be simulated, leading to a numerical representation of their properties.

Parameters are estimated from digital embryos as described in section 2.4.1. We recall that it requires a first spatial and temporal rescaling over the cohort to overcome a first level of inter individual variability (section 2.3.1).

Cell group	$\mathscr{L}^{n,k}$
Cardinality	$\lvert\mathscr{L}^{n,k}\rvert = 2 * \lvert\mathscr{L}^{n-1,k}\rvert = 2^{n-n^{0,k}} \lvert\mathscr{L}^{n^{0,k},k}\rvert$
Division time	$M_{n,k} = X_{n,k} + M_{n-1,k} = \left(\sum_{r=n^{0,k}+1}^{n} X_{r,k}\right) + M_{n^{0,k}}$ $M_{n,k} \sim N(\mu_M^{n,k}, \sigma_M^{n,k})$ $X_{n,k} \sim N(\mu_X^{n,k}, \sigma_X^{n,k})$ $P(X_{n,k} \mid M_{n-1,k}) = P(X_{n,k})$ $\mu_M^{n,k} = \sum_{r=n^{0,k}+1}^{n} \mu_X^{r,k} + \mu_M^{n^{0,k},k}$ $(\sigma_M^{n,k})^2 = \sum_{r=n^{0,k}+1}^{n} (\sigma_X^{r,k})^2 + (\sigma_M^{n^{0,k},k})^2$
Mean volume	$\overline{V}_{n,k} = A_{n,k}\overline{V}_{n-1,k} = \left(\prod_{r=n^{0,k}+1}^{n} A_{r,k}\right)\overline{V}_{n^{0,k},k}$ $\overline{V}_{n,k} \sim lnN(\mu_V^{n,k}, \sigma_V^{n,k})$ $A_{n,k} \sim lnN(\mu_A^{n,k}, \sigma_A^{n,k})$ $P(A_{n,k} \mid \overline{V}_{n,k}) = P(A_{n,k})$ $\mu_V^{n,k} = \sum_{r=n^{0,k}+1}^{n} \mu_A^{r,k} + \mu_V^{n^{0,k},k}$ $(\sigma_V^{n,k})^2 = \sum_{r=n^{0,k}+1}^{n} (\sigma_A^{r,k})^2 + (\sigma_V^{n^{0,k},k})^2$
Mean surface area	$\overline{S}_{n,k} = B_{n,k}\overline{S}^{n-1,k} = \left(\prod_{r=n^{0,k}+1}^{n} B_{r,k}\right)\overline{S}_{n^{0,k},k}$ $\overline{S}_{n,k} \sim lnN(\mu_S^{n,k}, \sigma_S^{n,k})$ $B_{n,k} \sim lnN(\mu_B^{n,k}, \sigma_B^{n,k})$ $P(B_{n,k} \mid \overline{S}^{n,k}) = P(B_{n,k})$ $\mu_S^{n,k} = \sum_{r=n^{0,k}+1}^{n} \mu_B^{r,k} + \mu_S^{n^{0,k},k}$ $(\sigma_S^{n,k})^2 = \sum_{r=n^{0,k}+1}^{n} (\sigma_B^{r,k})^2 + (\sigma_S^{n^{0,k},k})^2$

Table 2.3: Summary of the relations between the different random variables in the model - the definitions of the different notations can be found in the previous sections 2.5, 2.5, 2.5

Cells	$j \in \mathscr{L}^{n-1,k}$ and $i \in \mathscr{L}^{n,k}$ i is one of the two daughters of j
Division time	$m_i = x_i + m_j$ x_i **is a realization of** $X_{n,k}$ m_j **is a realization of** $M_{n-1,k}$
Volume	$v_i(t) = a_i.\overline{v}_j.\omega^{n,k}(u_i(t))$ a_i **is a realization of** $A_{n,k}$ $\overline{v}_j$ **is a realization of** $\overline{V}_{n-1,k}$ $\omega^{n,k}$ is the deterministic volume micro dynamic $u_i(t)$ is the percentage of elapsed life length of cell i
Surface area	$s_i(t) = b_i.\overline{s}_j.\phi^{n,k}(u_i(t))$ b_i **is a realization of** $B_{n,k}$ $\overline{s}_j$ **is a realization of** $\overline{S}^{n-1,k}$ $\phi^{n,k}$ is the deterministic surface area micro dynamic $u_i(t)$ is the percentage of elapsed life length of cell i

Table 2.4: Instantiation of the model independently in any branch of the cell lineage. The stochastic components are highlighted in bold- the definitions of the different notations can be found in the previous sections 2.5, 2.5, 2.5

Model evaluation

To assess the significance of our model, we propose to measure the distance between the empirical distribution $(M_{n,k}, \overline{V}_{n,k}, \overline{S}^{n,k})$ with the distribution predicted by the model.

For a cell cycle n, the parameters $\mu^{n,k}$ and $\sigma^{n,k}$ of the distribution of individual cell features can be predicted from the values of the parameters governing the distributions of previous cell cycles with the following relationships.

For the moment of division:

$$M_{n,k} \sim N\left(\mu_M^{n,k}, \sigma_M^{n,k}\right) \tag{2.50}$$

$$\mu_M^{n,k} = \sum_{r=n^{0,k}+1}^{n} \mu_X^{r,k} + \mu_M^{n^{0,k},k} \tag{2.51}$$

$$(\sigma_M^{n,k})^2 = \sum_{r=n^{0,k}+1}^{n} (\sigma_X^{r,k})^2 + (\sigma_M^{n^{0,k},k})^2 \tag{2.52}$$

for the volume:

$$\overline{V}_{n,k} \sim lnN\left(\mu_V^{n,k}, \sigma_V^{n,k}\right) \tag{2.53}$$

$$\mu_V^{n,k} = \sum_{r=n^{0,k}+1}^{n} \mu_A^{r,k} + \mu_V^{n^{0,k},k} \tag{2.54}$$

$$(\sigma_V^{n,k})^2 = \sum_{r=n^{0,k}+1}^{n} (\sigma_A^{r,k})^2 + (\sigma_V^{n^{0,k},k})^2 \tag{2.55}$$

and for the surface area:

$$\overline{S}_{n,k} \sim lnN\left(\mu_S^{n,k}, \sigma_S^{n,k}\right) \tag{2.56}$$

$$\mu_S^{n,k} = \sum_{r=n^{0,k}+1}^{n} \mu_B^{r,k} + \mu_S^{n^{0,k},k} \tag{2.57}$$

$$(\sigma_S^{n,k})^2 = \sum_{r=n^{0,k}+1}^{n} (\sigma_B^{r,k})^2 + (\sigma_S^{n^{0,k},k})^2 \tag{2.58}$$

Since we measure both the statistics of $M_{n,k}$, $\overline{V}_{n,k}$, $\overline{S}^{n,k}$ and $X_{n,k}$, $A_{n,k}$, $B_{n,k}$ for all cell groups $\mathscr{L}^{n,k}$. We can compare the empirical statistics of $M_{n,k}$, $\overline{V}_{n,k}$, $\overline{S}^{n,k}$ and the values of the parameters predicted by the model.

A good measure of distance between probability distributions is the symmetrized Kullback-Leibler divergence [10, 180, 159]. The Kullback-Leibler divergence between two probabil-

ity distributions, having p and q as probability densities, is defined as:

$$KL(p||q) = \int_x p(x) \log(\frac{p(x)}{q(x)}).dx \tag{2.59}$$

The Kullback-Leibler divergence between two Gaussian distributions $p(x|\mu_p, \sigma_p^2)$ and $p(x|\mu_q, \sigma_q^2)$ is:

$$KL(p(x|\mu_p, \sigma_p^2)||p(x|\mu_q, \sigma_q^2)) = \frac{1}{2}.(2.\log\frac{\sigma_q}{\sigma_p} + \frac{\sigma_p^2}{\sigma_q^2} + \frac{(\mu_q - \mu_p)^2}{\sigma_q^2} - 1) \tag{2.60}$$

and can be symmetrized into:

$$D(p(x|\mu_p, \sigma_p^2), p(x|\mu_q, \sigma_q^2)) = \frac{1}{2}.KL(p(x|\mu_p, \sigma_p^2)||p(x|\mu_q, \sigma_q^2)) + \frac{1}{2}.KL(p(x|\mu_q, \sigma_q^2)||p(x|\mu_p, \sigma_p^2)) \tag{2.61}$$

$$= \frac{1}{4}.(\frac{\sigma_q^2}{\sigma_p^2} + \frac{\sigma_p^2}{\sigma_q^2} + (\mu_q - \mu_p)^2(\frac{1}{\sigma_p^2} + \frac{1}{\sigma_q^2}) - 2) \tag{2.62}$$

For each specimen of the cohort we obtain the comparison between the empirical measures of coefficients $M_{n^{\text{fin},k},k}$, $\overline{V}_{n^{\text{fin},k},k}$, $\overline{S}_{n^{\text{fin},k},k}$, where $n^{\text{fin},k}$ represents the last cell cycle observed in the window of observation (we write n^{fin} when the cell type is obvious). We can estimate the distance between the model and the data. In the following, the exponent emp stands for measured statistics and the exponent mod corresponds to the parameters predicted by the model.

For the mitosis times:

$$D(M_{n,k}^{\text{emp}}, M_{n,k}^{\text{mod}}) = \frac{1}{4}.\left(\frac{(\sigma_{M,\text{mod}}^{n,k})^2}{(\sigma_{M,\text{emp}}^{n,k})^2} + \frac{(\sigma_{M,\text{emp}}^{n,k})^2}{(\sigma_{M,\text{mod}}^{n,k})^2} + (\mu_{M,\text{mod}}^{n,k} - \mu_{M,\text{emp}}^{n,k})^2(\frac{1}{(\sigma_{M,\text{emp}}^{n,k})^2} + \frac{1}{(\sigma_{M,\text{mod}}^{n,k})^2}) - 2\right) \tag{2.63}$$

where $\mu_{M,\text{mod}}^{n,k} = \sum_{r=n^{0,k}+1}^{n} \mu_{X,\text{emp}}^{r,k} + \mu_{M,\text{emp}}^{n^{0,k},k}$ and $(\sigma_{M,\text{mod}}^{n,k})^2 = \sum_{r=n^{0,k}+1}^{n} (\sigma_{X,\text{emp}}^{r,k})^2 + (\sigma_{M,\text{emp}}^{n^{0,k},k})^2$.

For the volume:

$$D(\overline{V}_{n,k}^{\text{emp}}, \overline{V}_{n,k}^{\text{mod}}) = \frac{1}{4}.\left(\frac{(\sigma_{V,\text{mod}}^{n,k})^2}{(\sigma_{V,\text{emp}}^{n,k})^2} + \frac{(\sigma_{V,\text{emp}}^{n,k})^2}{(\sigma_{V,\text{mod}}^{n,k})^2} + (\mu_{V,\text{mod}}^{n,k} - \mu_{V,\text{emp}}^{n,k})^2(\frac{1}{(\sigma_{V,\text{emp}}^{n,k})^2} + \frac{1}{(\sigma_{V,\text{mod}}^{n,k})^2}) - 2\right) \tag{2.64}$$

where $\mu_{V,\text{mod}}^{n,k} = \sum_{r=n^{0,k}+1}^{n} \mu_{A,\text{emp}}^{r,k} + \mu_{V,\text{emp}}^{n^{0,k},k}$ and $(\sigma_{V,\text{mod}}^{n,k})^2 = \sum_{r=n^{0,k}+1}^{n} (\sigma_{A,\text{emp}}^{r,k})^2 + (\sigma_{V,\text{emp}}^{n^{0,k},k})^2$.

And for the surface area:

$$D(\overline{S}^{\text{emp}}_{n,k}, \overline{S}^{\text{mod}}_{n,k}) = \frac{1}{4}.\left(\frac{(\sigma^{n,k}_{S,\text{mod}})^2}{(\sigma^{n,k}_{S,\text{emp}})^2} + \frac{(\sigma^{n,k}_{S,\text{emp}})^2}{(\sigma^{n,k}_{S,\text{mod}})^2} + (\mu^{n,k}_{S,\text{mod}} - \mu^{n,k}_{S,\text{emp}})^2 (\frac{1}{(\sigma^{n,k}_{S,\text{emp}})^2} + \frac{1}{(\sigma^{n,k}_{S,\text{mod}})^2}) - 2 \right) \tag{2.65}$$

where $\mu^{n,k}_{S,\text{mod}} = \sum_{r=n^{0,k}+1}^{n} \mu^{r,k}_{B,\text{emp}} + \mu^{n^{0,k},k}_{S,\text{emp}}$ and $(\sigma^{n,k}_{S,\text{mod}})^2 = \sum_{r=n^{0,k}+1}^{n} (\sigma^{r,k}_{B,\text{emp}})^2 + (\sigma^{n^{0,k},k}_{S,\text{emp}})^2$.

The measure of distance in itself is not easily interpretable because we don't have a benchmark for it. To assess relevant orders of magnitude, we compute the same distance between the measures obtained in the various embryo of the cohort. For a specimen e of the cohort, the distance between the model and the empirical distribution is computed for the last observable cell cycle $n^{\text{fin},k}$. This quantity is normalized with the averaged inter individual difference. We evaluate the normalized difference between the model and the data for an embryo e with the function ev defined, for the mitosis times, as:

$$\text{ev}(M,k) = \frac{D(M^{\text{emp}}_{n^{\text{fin}},k,e}, M^{\text{mod}}_{n^{\text{fin}},k,e})}{\frac{1}{|\mathscr{E}|-1} \sum_{l\in\mathscr{E}, l\neq e} D(M^{\text{emp}}_{n^{\text{fin}},k,e}, M^{\text{emp}}_{n^{\text{fin}},k,l})} \tag{2.66}$$

for the volumes:

$$\text{ev}(\overline{V},k) = \frac{D(\overline{V}^{\text{emp}}_{n^{\text{fin}},k,e}, \overline{V}^{\text{mod}}_{n^{\text{fin}},k,e})}{\frac{1}{|\mathscr{E}|-1} \sum_{l\in\mathscr{E}, l\neq e} D(\overline{V}^{\text{emp}}_{n^{\text{fin}},k,e}, \overline{V}^{\text{emp}}_{n^{\text{fin}},k,l})} \tag{2.67}$$

and for the surface areas:

$$\text{ev}(\overline{S},k) = \frac{D(\overline{S}^{\text{emp}}_{n^{\text{fin}},k,e}, \overline{S}^{\text{mod}}_{n^{\text{fin}},k,e})}{\frac{1}{|\mathscr{E}|-1} \sum_{l\in\mathscr{E}, l\neq e} D(\overline{S}^{\text{emp}}_{n^{\text{fin}},k,e}, \overline{S}^{\text{emp}}_{n^{\text{fin}},k,l})} \tag{2.68}$$

These measure of differences between the model and the data are computed for each morphogenetic field k, in each embryo such that $n^{\text{fin},k} > n^{0,k}$. They are computed for division times, volumes and surface areas. The results are grouped in the histogram on Fig. 2.12. 36 predicted sets of parameters among 48 (75%) show a distance significantly smaller than the reference distance (ev < 0.5), 10 predicted sets of parameters are in the same order of magnitude than the reference distance ($0.5 \leq \text{ev} < 1.5$), and 2 predicted sets of parameters differ significantly from the empirical data; they correspond to the statistics of volume and surface in the embryo 5 for the Mesomere ($k = 1$). As a conclusion, except for these two "outlying" distributions, the model predicts accurately the dynamics of the

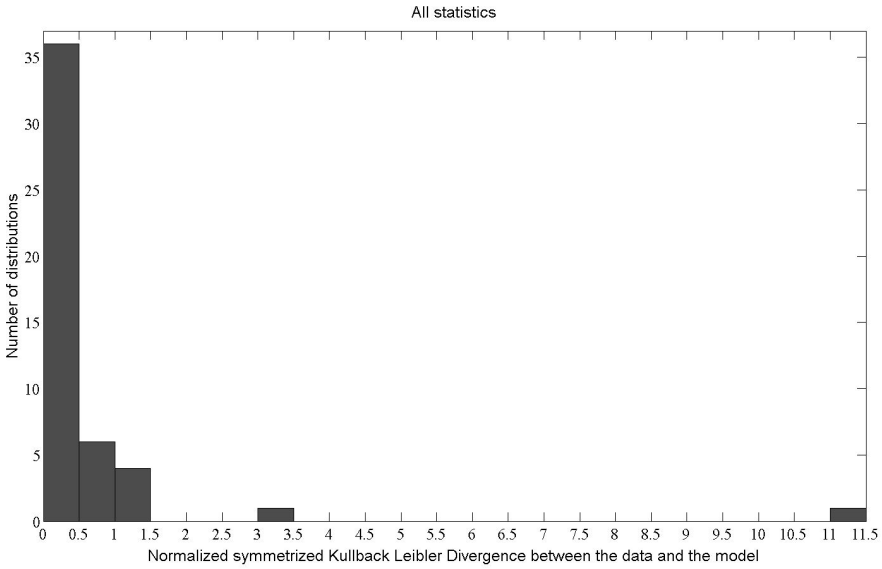


Figure 2.12: Histogram of the score of the model for all morphogenetic field k such that $n^{\text{fin},k} > n^{0,k}$ in each embryo for division times, volumes and surface areas, using the evaluation functions defined in equations 2.66, 2.67, 2.68

sea urchin development.

Simulation of artificial cell lineages with volume and surface area dynamics

300 realizations of the cell lineage were simulated in order to generate significant model output. Matlab random number generator was used to simulate the realizations of random variables. Macroscopic features were computed and compared to the empirical value obtained for each specimen, as represented on figures 2.13 and 2.14.

Each embryo e of the cohort $\mathscr{E}$ has a specific window of observation. This window determines the observed of cell groups $\mathscr{L}_e^{n,k}$. $n^{0,k}$ denotes the initial cell cycle where cells of type k are observable in a significant number and $n^{fin,k}$ the last cell cycle where cells of type k are observable in a significant number (we write n_0 and n_{fin} when the cell type k is obvious), as defined in the section about parameter estimation. The algorithm used to generate the cell lineage is a straightforward implementation of the model described in the section 2.5.

The finite range of observation of the cell lineage prevents to have access to the complete cell cycle for the first and last generation, thus the volume ($\omega^{n,k}$) and surface area ($\phi^{n,k}$) micro dynamics cannot be estimated for these cycles. The shaded area on the right

of the plot corresponding to the cell volume and the cell surface area represents the incomplete cell cycles on figures 2.13 and 2.14.

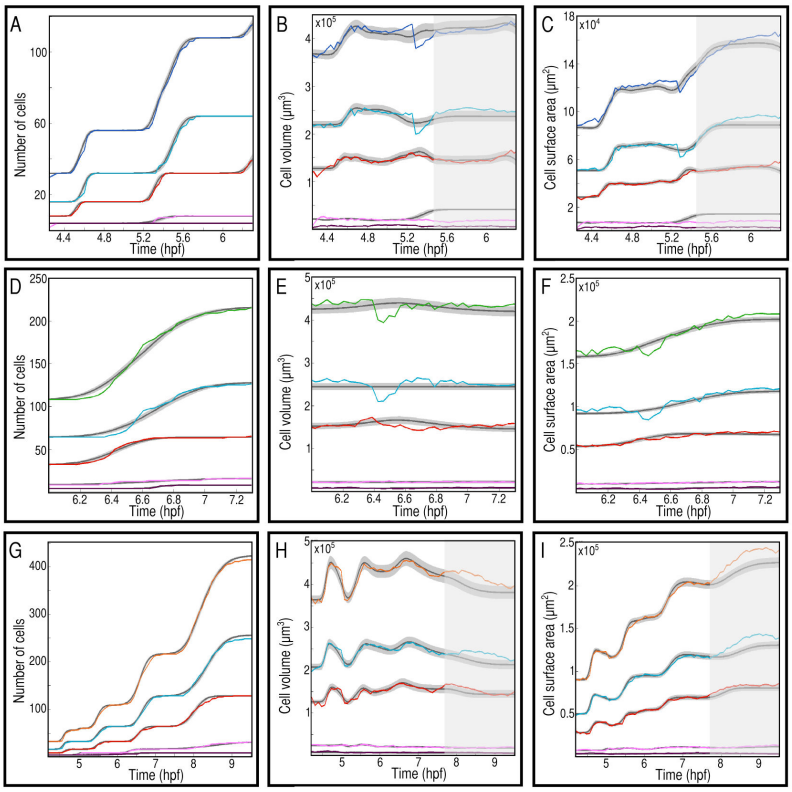


Figure 2.13: Probabilistic modeling for each embryo of the cohort & A) D) G) in color number of cells in embryo 1,2,3 - in black and grey mean and standard deviation of 300 realizations of the model B) E) H) in color cellular volume in embryo 1,2,3 - in black and grey mean and standard deviation of 300 realizations of the model C) F) I) in color cellular surface in embryo 1,2,3 - in black and grey mean and standard deviation of 300 realizations of the model - The shaded area on the right of B, C, H and I refers to incomplete cell cycles

2.5.1 Prototype

The probabilistic model provides an invariant structure relating individual cell features with embryo level dynamics by using an intermediate coarse-grained level (cell groups clustered by common cell type and generation). The same structure appears in each specimen of the cohort. Sufficient statistics are identified for each cell group, providing a "signature" identifying each specimen of the cohort. To get a unique representation of the sea

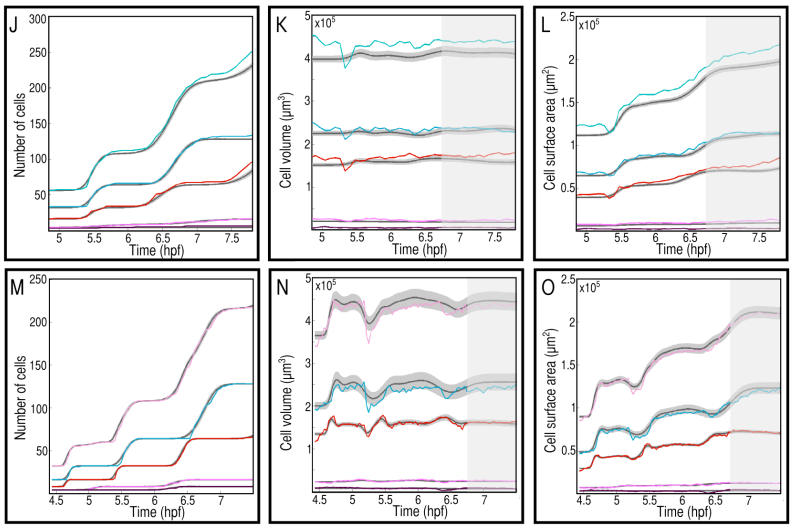


Figure 2.14: Fig. 2.13 continued J) M) in color number of cells in embryo 4,5 - in black and grey mean and standard deviation of 300 realizations of the model K) N) in color cellular volume in embryo 4,5 - in black and grey mean and standard deviation of 300 realizations of the model L) O) in color cellular surface in embryo 4,5 - in black and grey mean and standard deviation of 300 realizations of the model - The shaded area on the right of K, L, N and O refers to incomplete cell cycles

urchin blastula normal development, measures of each specimen have to be aggregated. These aggregated measures associated to the identified invariant structure are meant to provide a prototypical representation of the cohort. The aggregation procedure is based on the **geometrical concept of centroid** of a set of points which generalizes the notion of average in higher dimensional spaces. The mathematical framework of information geometry is relevant here to deal with spaces of probability distributions needing to be endowed with a relevant notion of distance [10, 180]. Bregman divergences generalize euclidean distance and unify it to the statistical Kullback-Leibler divergence. In the following we use the Kullback-Leibler divergence as the relevant notion of distance to compare probability distribution.

Kullback-Leibler centroid

We give in this section the main formulas needed to compute the centroid of the cohort of sea urchin. More refined mathematical developments can be found in [159].

The Kullback-Leibler centroid is defined as follow:

$$c = \arg\min_{c \in S} \frac{1}{n}. \sum_{i=1}^{n} \frac{KL(p_i||c) + KL(c||p_i)}{2} \tag{2.69}$$

Therefore combining equation 2.60 and 2.71, we obtain:

$$c = \arg\min_{c \in S} \frac{1}{n} \frac{1}{4} \sum_{i=1}^{n} \left((2.\log \frac{\sigma_c}{\sigma_i} + \frac{\sigma_i^2}{\sigma_c^2} + \frac{(\mu_c - \mu_i)^2}{\sigma_c^2} - 1) + (2.\log \frac{\sigma_i}{\sigma_c} + \frac{\sigma_c^2}{\sigma_i^2} + \frac{(\mu_i - \mu_c)^2}{\sigma_i^2} - 1) \right)$$

which simplifies into

$$c = \arg\min_{c \in S} \frac{1}{n} \frac{1}{4} \sum_{i=1}^{n} \left(\frac{\sigma_i^2}{\sigma_c^2} + \frac{\sigma_c^2}{\sigma_i^2} + (\mu_i - \mu_c)^2 . (\frac{1}{\sigma_c^2} + \frac{1}{\sigma_i^2}) - 2 \right) \tag{2.70}$$

Prototype as a centroid in the relevant morphospace

The probabilistic model provides an invariant structure relating individual cell features with embryo level dynamics by using an intermediate coarse-grained level (cell groups clustered by common cell type and generation). The same structure appears in each specimen of the cohort. Sufficient statistics are identified for each cell group, providing a "signature" identifying each specimen of the cohort. To get a unique representation of the sea urchin blastula normal development, measures of each specimen have to be aggregated.

These aggregated measures associated to the identified invariant structure are meant to provide a prototypical representation of the cohort. The aggregation procedure is based on the geometrical concept of centroid of a set of points which generalizes the notion of average in higher dimensional spaces. The mathematical framework of information geometry is relevant here to deal with spaces of probability distributions needing to be endowed with a relevant notion of distance [10, 180]. Bregman divergences generalize euclidean distance and unify it to the statistical Kullback-Leibler divergence. In the following we use the Kullback-Leibler divergence as the relevant notion of distance to compare probability distribution.

Kullback-Leibler centroid

We give in this section the main formulas needed to compute the centroid of the cohort of sea urchin. More refined mathematical developments can be found in [159].

The Kullback-Leibler centroid is defined as follow:

$$c = \arg\min_{c \in S} \frac{1}{n}. \sum_{i=1}^{n} \frac{KL(p_i||c) + KL(c||p_i)}{2} \tag{2.71}$$

Therefore combining equation 2.60 and 2.71, we obtain:

$$c = \arg\min_{c \in S} \frac{1}{n} \frac{1}{4} \sum_{i=1}^{n} \left((2.\log \frac{\sigma_c}{\sigma_i} + \frac{\sigma_i^2}{\sigma_c^2} + \frac{(\mu_c - \mu_i)^2}{\sigma_c^2} - 1) + (2.\log \frac{\sigma_i}{\sigma_c} + \frac{\sigma_c^2}{\sigma_i^2} + \frac{(\mu_i - \mu_c)^2}{\sigma_i^2} - 1) \right) \tag{2.72}$$

which simplifies into

$$c = \arg\min_{c \in S} \frac{1}{n} \frac{1}{4} \sum_{i=1}^{n} \left(\frac{\sigma_i^2}{\sigma_c^2} + \frac{\sigma_c^2}{\sigma_i^2} + (\mu_i - \mu_c)^2 . (\frac{1}{\sigma_c^2} + \frac{1}{\sigma_i^2}) - 2 \right) \tag{2.73}$$

Prototype as a centroid in the relevant morphospace

The prototypical representation of the sea urchin blastula normal development is modeled with the same structure as for each individual specimen. This structure is summarized in section 2.5 and relates individual cell features with global dynamics through statistics at the level of groups of cell. Using this multi-level probabilistic model, the prototype is defined as a centroid in the relevant morphospace (the idea is schematized on Fig. 2.15). Because the statistics at the level of cell groups are assumed to be either normal or log-normally distributed, the relevant morphospace consists in the statistical manifold

S defined for each random variable by the coordinates system (μ, σ^2).

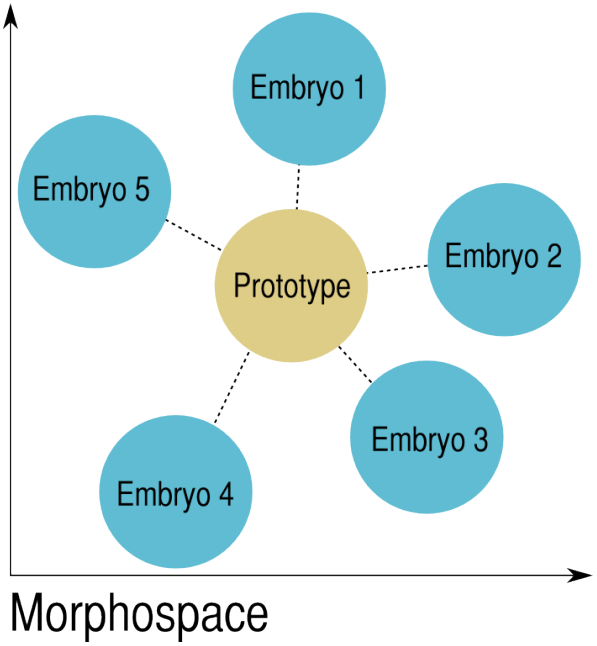


Figure 2.15: Schematic representation of the concept of prototype in the relevant morphospace - the morphospace considered here is the statistical manifold S and the prototype is the centroid of the cohort computed as a minimization problem described in equation 2.74

The parameters of the distributions of individual cell features in each cell groups are estimated as the centroid of the individual distributions found for each specimen of the cohort.

Thus the parameters of the prototype (μ^p, σ^p) are obtained by solving the optimization problem described in equation 2.73 over the cohort $\{(\mu_e, \sigma_e) \mid e \in \mathscr{E}\}$.

$$(\mu^p, \sigma^p) = \arg \min_{(\mu,\sigma)\in\mathbb{R}_+^2} \frac{1}{n}\frac{1}{4}\sum_{e\in\mathscr{E}}\left(\frac{\sigma_e^2}{\sigma^2} + \frac{\sigma^2}{\sigma_e^2} + (\mu_e - \mu)^2.(\frac{1}{\sigma^2} + \frac{1}{\sigma_e^2}) - 2\right) \tag{2.74}$$

To actually compute the coefficients used in the simulations, this minimization problem is solved numerically with the estimated empirical parameters for each cell group $\mathscr{L}_g$. The results of the computation are shown on Fig. 2.16. The distributions of life length X and mitosis time M within cell groups $\mathscr{L}_g$ are assumed to be normal distributions, the prototypical coefficients are computed directly from the equation 2.74. The distributions of volume V, surface area S, daughter/mother mean volume ratio A and mother/daughter

mean surface area ratio B are assumed to be log-normal distributions, thus the prototypical coefficients are computed from the logarithm of these random variables which are normally distributed.

Finally, the prototypical volume and surface area microdynamics $\omega(t)$ and $\phi(t)$ which are deterministic functions are obtained for each cell group $\mathscr{L}_g$ by a simple average over the cohort. They are represented in bold on the figures 2.9 and 2.10.

Simulation

300 realizations of prototypical cell lineages are produced and the macroscopic features are compared with the empirical values as shown on Fig. 2.17.

Artificial cell lineages decorated with cellular volume and cellular surface can be provided for the prototypical model in the same way as for each embryo. We used as input coefficients: $\widetilde{\mu}_{X,p}, \widetilde{\sigma}_{X,p}$, $\widetilde{\mu}_{A,p}, \widetilde{\sigma}_{A,p}$, $\widetilde{\mu}_{\omega}$, $\widetilde{\mu}_{B,p}, \widetilde{\sigma}_{B,p}$, $\widetilde{\mu}_{\phi}$, $n_0 = 6$ (the initial stage is the 32 cell stage) and $\widetilde{\mu}^{n_0,k}_{M,p}, \widetilde{\sigma}^{n_0,k}_{M,p}$, $\widetilde{\mu}^{n_0,k}_{V,p}, \widetilde{\sigma}^{n_0,k}_{V,p}, \widetilde{\mu}^{n_0,k}_{S,p}, \widetilde{\sigma}^{n_0,k}_{S,p}$. The simulation ends at the 408 cell stage with $n^{fin,1} = n^{fin,2} = 10$ for the Mesomeres and the Macromeres, $n^{fin,3} = 9$ for the Large Micromeres and $n^{fin,4} = 7$ for the Small Micromeres. Section 2.5.1 defines the notation and the procedure of estimation of their value.

2.6 Biomechanical model description

The prototypical model of the cell lineage, with cell volume and cell surface area dynamics, is used as a basis for a biomechanical modeling of the sea urchin early embryogenesis. This model is obtained with MecaGen modeling platform corresponding to Julien Delile's work under the supervision of René Doursat and Nadine Peyriéras ([55], [56]).

This model describes an organism as a set of interacting particles, representing the cells. Cells are small and "sticky", which is a main property of their physical behaviors. The consequence is the disappearance of inertial forces [179] and low value of Reynolds number. In this situation, applied forces produce a velocity and not an acceleration. The displacement is proportional to the instantaneous force and inversely proportional to a damping coefficient λ. For a multicellular system, the motion of each cell i with an interacting neighborhood $\mathscr{N}_i$ of cells j is governed by the equation:

$$\lambda_i \vec{v}_i = \sum_{j \in \mathscr{N}_i} \vec{F}_{ij} \tag{2.75}$$

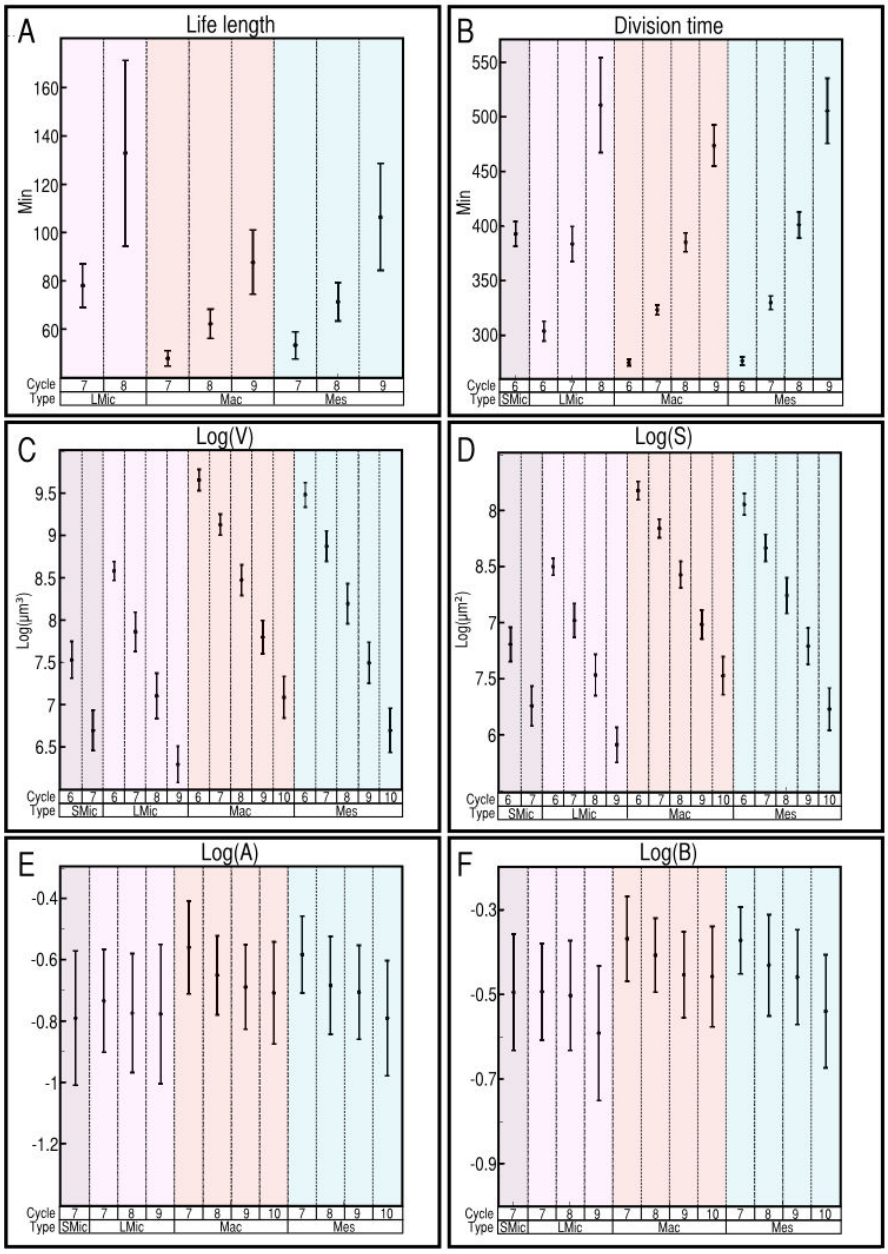


Figure 2.16: Prototypical coefficients for each group of cell $\mathscr{L}^{n,k}$ A) Mean and standard deviation of the life lengths B) Mean and standard deviation of the division times C) Mean and standard deviation of the log of the volume D) Mean and standard deviation of the log of the surface area E) Mean and standard deviation of the log of the daughter/mother volume ratio F) Mean and standard deviation of the log of the daughter/mother mean surface ratio

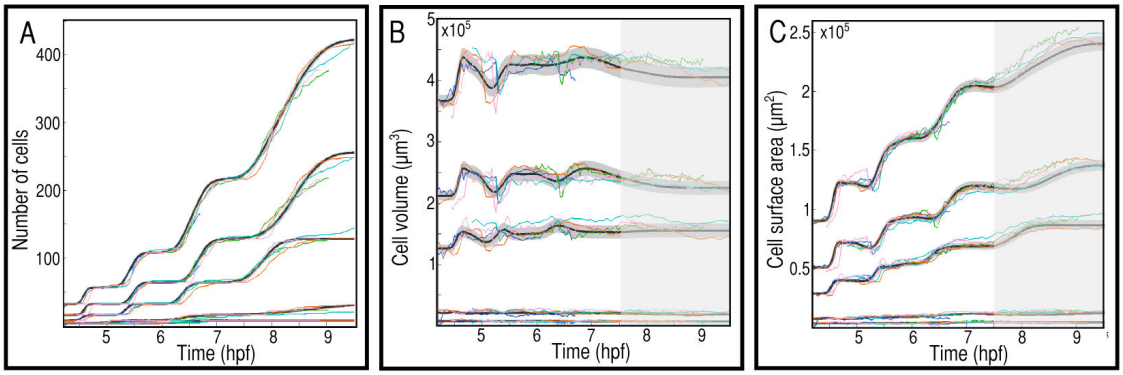


Figure 2.17: In each graph, one color corresponds to one embryo and the 300 realizations of the prototypical probabilistic model are represented with mean (black) and standard deviation (grey). All curves are represented as a function of developmental time A) number of cells B) cellular volume C) cellular surface - for B and C, shared areas on the right correspond to the end of the window of observation and therefore to incomplete cell cycles

where $\vec{F}_{ij}$ is the interaction force exerted by cell j over cell i.

Each particle correspond to a cell, each of them having a certain shape. This shape is modeled as a cylinder as shown on figure 2.18. Cell shape is defined by a lateral radius R, a half-height $R^{\perp}$ and a normal vector $\vec{U}$. The cylinder axis $\vec{U}$ is oriented orthogonally to the epithelial surface. Throughout this study, we assume that the depth of the tissue remain constant and identical for each cell as the embryo develops itself, thus $R_i^{\perp} = R_0^{\perp}$ for each cell. And therefore, the relation between the tangential radius of each cell i and its current volume $V_i(t)$ reads:

$$R_i(t) = \sqrt{\frac{V_i(t)}{2\pi R_0^{\perp}}} \tag{2.76}$$

The motion of each cell is determined by its interactions with its neighborhood. The establishment of this neighborhood is performed in two steps, first a metrical assessment of the neighborhood and then a topological criteria to filter results of the first step. This two-step procedure, which is described in ([55], [56]) enable to compute for each cell i its interacting neighborhood $\mathcal{N}_i$.

Once the topological list of neighbors is determined, the surface orientation of the monolayered sea urchin can be expressed at each cell location by averaging the outward normal vectors of the n triangles formed by n topological neighbors (Fig 2.19).

The interaction force $\vec{F}_{ij}$ exerted by cell j over cell i leads the swarm of cells toward an

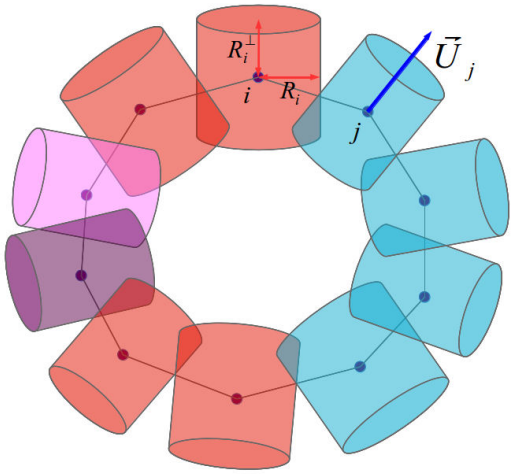


Figure 2.18: **Cells are represented by cylindrical particle.** Each cell is defined by a lateral radius R, a half-height $R^{\perp}$ and a normal vector $\vec{U}$. - Figure by Julien Delile

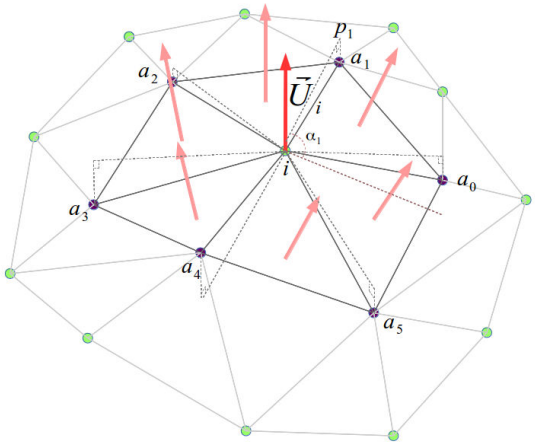


Figure 2.19: **Cell axis determination procedure.** The cell axis $\vec{U}_i$ (red arrow) is calculated by averaging the 6 surrounding triangle outward normal vectors (light red arrows). In order to define the surrounding triangles, each neighbor position is first projected on the plane orthogonal to $\vec{U}_i$ (here, only p_1 the projection of the neighbor a_1 is displayed). Then neighbors are sorted according to the angle α_j formed by the vector (i, p_j) and an arbitrary vector belonging to the orthogonal plane (the light red and dashed line). Figure by Julien Delile

equilibrium state, it is the sum of two components:

- an *attraction-repulsion* interaction force $\vec{F}^{\parallel}_{ij}$, which maintains the integrity of the cell's volume and controls both the stiffness and the adhesion of the interaction in the direction of the surface plane; this force can be modulated via the stiffness and adhesion coefficients, and
- a *planarity conservation* interaction force $\vec{F}^{\perp}_{ij}$, which maintains the planarity of a monolayered epithelium; this force can be modulated via a planar rigidity coefficient.

Moreover, the damping coefficient λ_i, which plays here a role somewhat equivalent to the mass m_i in Newton's Second Law, is proportional to the surface of the cell: $\lambda_i = \lambda_0 S_i$ with $S_i = 2\pi R_i(R_i + 2R_i^{\perp})$ as the cell shape is considered cylindrical. The equation of motion used throughout the study will be:

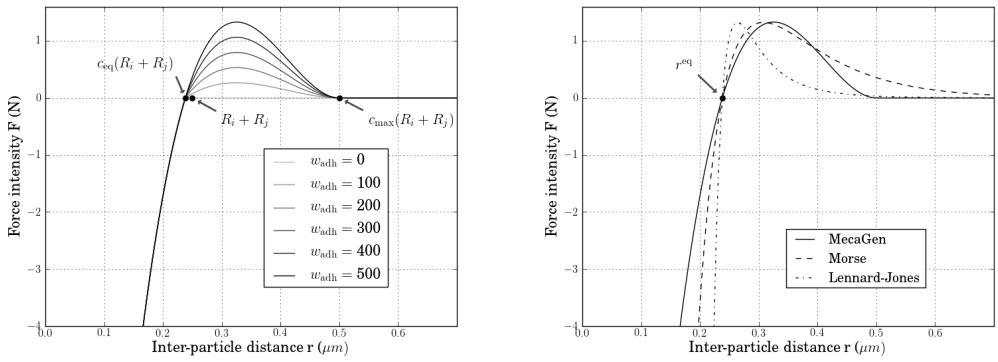
$$\vec{v}_i = \frac{1}{\lambda_0 S_i} \sum_{j \in \mathcal{N}_i} \left(\vec{F}^{\parallel}_{ij} + \vec{F}^{\perp}_{ij} \right) \tag{2.77}$$

The *attraction-repulsion* interaction force $\vec{F}^{\parallel}_{ij}$ is a spring-like force derived from an elastic potential. Its expression is given by the following system of equations

$$\vec{F}^{\parallel}_{ij} = \begin{cases} -w_{\text{rep}}(r_{ij} - r_{ij}^{\max})^2 (r_{ij} - r_{ij}^{\text{eq}}).\vec{u}_{ij} & \text{if } r_{ij} < r_{ij}^{\text{eq}} \\ -w_{\text{adh}}(r_{ij} - r_{ij}^{\max})^2 (r_{ij} - r_{ij}^{\text{eq}}).\vec{u}_{ij} & \text{if } r_{ij} \geq r_{ij}^{\text{eq}} \text{ and } r_{ij} < r_{ij}^{\max} \\ \vec{0} & \text{if } r_{ij} \geq r_{ij}^{\max} \end{cases} \tag{2.78}$$

where w_{rep} and w_{adh} are a repulsion and an adhesion coefficient, r_{ij}^{eq} is an equilibrium distance between two cells corresponding to the densest packing in the 2D plane, $r_{ij}^{\max}$ is the maximal distance under which cells can adhere, finally r_{ij} corresponds to the distance between the two cells. The magnitude of this force is plotted on figure 2.20 (a) with varying values of w_{adh}. Alternative possible models are shown on figure 2.20 (b).

The *planarity conservation* interaction force ensures that the spatial configuration of the swarm of cells remain planar during the simulation. Between two neighboring cells, i and j, this planar rigidity coefficient is proportional to the planar rigidity coefficient k_{rig}, and to the dot product between the neighborhood vector $(\vec{X}_j - \vec{X}_j)$ ($\vec{X}_j$ and $\vec{X}_j$ are the positions of i and j) and the normalized sum of both neighbor normal vectors $\vec{n}_{ij} = \frac{\vec{U}_i + \vec{U}_j}{|\vec{U}_i + \vec{U}_j|}$. The force is oriented in the direction of $\vec{n}_{ij}$ and the force complete expression reads:



(a) (b)

Figure 2.20: (a)Plot of the attraction-repulsion force $\vec{F}^{\parallel}$ modulated with the variable values of w_{adh} (b) Comparison of the attraction-repulsion force used here with alternative classical attraction-repulsion forces. The equilibrium distance is $r^{\mathrm{eq}} = r_{ij}^{\mathrm{eq}} = c_{\mathrm{eq}}(R_i + R_j$ with $R_i = 0.1$ and $R_j = 0.15$. The solid line is the attraction/repulsion potential $\vec{F}_{ij}^{\parallel}$ with $w_{\mathrm{adh}} = w_{\mathrm{rep}} = 500$. The dashed line is a force derived from the Morse potential, its equation reads: $2Dke^{-k(r-r^{\mathrm{eq}})} - 2Dke^{-2k(r-r^{\mathrm{eq}})}$ with $D = 0.265$ and $k = 10$. The dot-dashed line is force derived from the Lennard-Jones potential, its equation reads: $-\frac{24\epsilon}{\sigma}(2(\frac{r}{\sigma})^{-13} - (\frac{r}{\sigma})^{-7})$ with $\epsilon = 0.117$ and $\sigma = 2^{-1/6} r^{\mathrm{eq}}$. The force parameters were selected in order that the equilibrium distance and the maximum value of the three curves match approximately on the plot. Figure by Julien Delile

$$\vec{F}^{\perp}_{ij} = k_{\text{rig}} \left((\vec{X}_j - \vec{X}_i).\vec{n}_{ij} \right) \vec{n}_{ij}$$

The notations are described on figure 2.21.

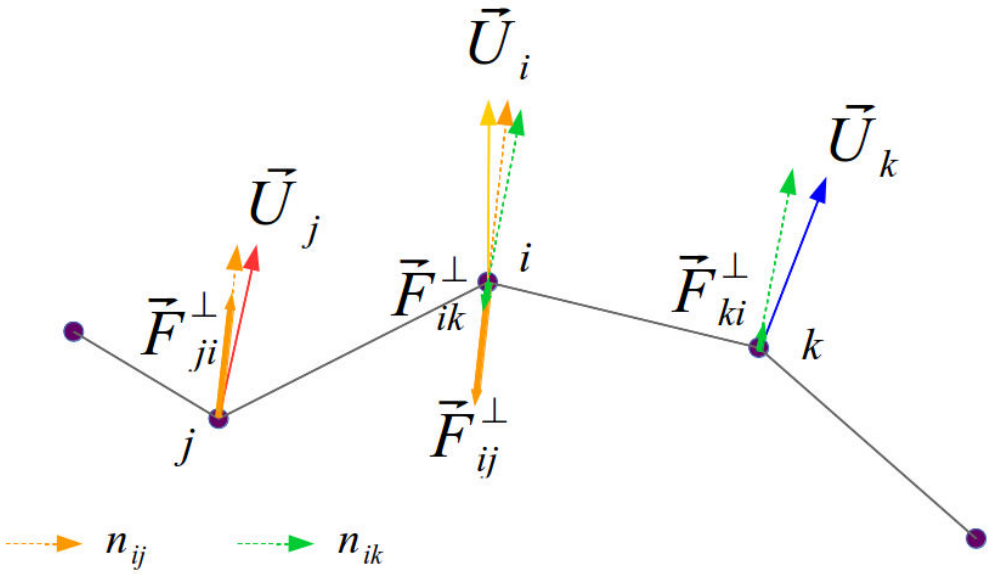


Figure 2.21: **Schematic of the planar rigidity force between neighbor epithelial cells.** Here, cells i, j and k are represented in 2D, with their respective normal axes $\vec{U}_i$ (yellow arrow), $\vec{U}_j$ (red arrow) and $\vec{U}_k$ (blue arrow). The planar rigidity forces exerted between neighbor cells i and j, and between i and k are aligned colinearly with the shared normal vectors $\vec{n}_{ij}$ (orange vector), and $\vec{n}_{ik}$ respectively. Forces are displayed by the thicker arrows. The intensity of the force is lesser between i and k because the relative position vector $\vec{u}_{ik}$ and $\vec{n}_{ik}$ are nearly orthogonal. As a result of these forces, neighbor cells are always attracted back to the lateral domain of each cell and thus maintain the planarity of the tissue. Figure by Julien Delile

Mitoses are triggered when the cell cycle is complete as computed with the prototypical probabilistic model. When that event occurs, a division axis $\vec{M}_{\text{m}}$ is randomly selected in the local tangential plane of the tissue.

Similarly, cell volumes are computed with the prototypical probabilistic model. As mentionned in section before, the apico-basal radius of a daughter cell remains equal to the one of its mother $R^{\perp}_{\text{m}} = R^{\perp}_{\text{d1}} = R^{\perp}_{\text{d2}} = R^{\perp}_{0}$. The daughter cell lateral radii are deduced from their respective volume.

2.7 Comparison to experimental data

This section describes the metric used to compare the simulation of the biomechanical model embedding the prototypical probabilistic model in space. The metric enables to compare the simulation to the empirical data-set with varying values of the free parameters such as the adhesion coefficient w_{adh}. This approach validates the hypotheses underlying the model.

Three metrics based on three different aspects of the topology of the embryo are designed (D_n for the number of neighbors, D_c for the number of neighbors variation, and D_b for the inter-population borders). In addition, we also calculate four objective functions characterizing some properties of simulated embryos only (C_s evaluates the cohesion of the tissue, S_s and N_s the sphericity of the simulated embryo shapes, and P_s the planarity of the epithelium). The evaluation of these comparisons is performed in section 2.7.4.

In the following, we will associate the letter s with the spatial simulations and e with the digital reconstructions.

2.7.1 Metrics

Comparison protocol To compare a spatially simulated embryo with the cohort, we used the three topological features: 1) the distribution of the number of neighbors 2) the distribution of the rate of neighborhood change and 3) the distribution of the neighbors shared at the border.

Topological features The topology of an embryo is well described by the network of cell-cell contacts which can be used as an estimator of the shape of the embryo. At time t, the vertices of the network are the nuclei of the cells, denoted by $\mathscr{L}(t)$ (with the same notation as in section 2.3.1). The edges of the network are the junctions between cells, denoted by $\mathscr{E}(t)$.

The neighborhood of a given cell i at time t is $\mathscr{N}_i^t = \left\{ j \in \mathscr{L}(t) : (i,j) \in \mathscr{E}^t \right\}$. The number of neighbors (degree) d_i^t of this cell can be calculated with the cardinality of its neighborhood:

$$d_i^t = |\mathscr{N}_i^t|$$

The network is evolving through time. To get an idea of the dynamics that it undergoes, we computed the rate of change of the neighborhood per unit of time, c_i^t. We compared

the neighborhood at time t and $t+dt$ and made the sum of the number of lost cells and gained cells. More formally it can be described as:

$$c_i^t = \frac{1}{dt}(\mathcal{N}_i{}^t \Delta \mathcal{N}_i{}^{t-dt}) = \frac{1}{dt}(\mathcal{N}_i{}^t \cup \mathcal{N}_i{}^{t-dt}) \setminus (\mathcal{N}_i{}^t \cap \mathcal{N}_i{}^{t-dt})$$

were Δ represents the symmetrical difference between two sets.

The embryos are composed of different subpopulations k that have already been presented above, in section 2.3.1. The topological network of cellular contacts allows to define a notion of boundary between them, it will be the set of cells of a given type that share a contact with the cells of another type. Formally, the border of the cell of type k_1 with the cell of type k_2 is: $\mathscr{B}^t_{k_1 \to k_2} = \{i \in \mathscr{L}^{k_1}(t) : \exists j \in \mathcal{N}_i^t$, such that $j \in \mathscr{L}^{k_2}(t)\}$. With this definition of the border, the number of contacts of a given cell i (type k_1) shared at the border $\mathscr{B}^t_{k_1 \to k_2}$ can be written:

$$b_i^t(k_1 \to k_2) = |\{j \in \mathcal{N}_i^t : \text{type}(j) = k_2\}|$$

Distributions of topological features in the groups of cells We would like to compare embryos based on these topological features. We have shown that it is impossible to compare developing embryos at the level of the individual cell because of an inherent variability in the cellular behaviors. We chose to compare these features at the level of the subpopulations of cells to be coherent with the models whose parameters are defined at this resolution. We therefore focused our attention on the distribution of topological features in the groups of cells.

At a given time t, for a subpopulation k, we can define the distribution of the number of neighbors using its normalized histogram $h_{d,k}(t)$. Since the number of neighbors of a cell is a discrete quantity, we defined naturally the bins as being the number of neighbors. Namely for a bin $n \in \mathbb{N}$, we compute the frequencies of occurrence as:

$$p^t_{d,k}(n) = \frac{1}{|\mathscr{L}^k(t)|}.|\left\{i \in \mathscr{L}^k(t) : d_i(t) = n\right\}|$$

and the complete histogram is the vector, $h_{d,k}(t) = (p^t_{d,k}(1), p^t_{d,k}(2), ..., p^t_{d,k}(N))$, N being the maximum number of neighbors observed. The histogram is normalized: $\Sigma_{k=1}^N p^t_{d,k}(k) = 1$, it can be viewed as a discrete probability density.

In the same manner we computed a normalized histogram for the rate of change of the neighborhood per unit of time $h^t_{c,g}$. To link the bins with interpretable biological

quantity, we considered as a base unit one change in the neighborhood in 4 minutes (equal to rate of change of 0.25 per minute). The histogram $h_{c,k}(t)$ is defined on the set $X = ([0, 0.25],]0.25, 0.5], ...,]C - 0.25, C])$, with C being the maximum value observed for the rate of change in the neighborhood. The frequencies of occurrence for a bin $x \in X$ can be calculated with the following formula:

$$p^t_{c,k}(n) = \frac{1}{|\mathscr{L}^k(t)| * 0.25}.|\left\{i \in \mathscr{L}^k(t) : c^t_i \in x\right\}|$$

the complete histogram being the vector $h_{c,k}(t) = (p^t_{c,g}(0), p^t_{c,g}(1), ..., p^t_{c,g}(C))$, with $\Sigma^C_{k=0} p^t_{c,g}(k) * 0.25 = 1$. The histogram can be considered as a discretized continuous probability density.

For the boundaries, we considered all the possible cases of contact between cells. There are four cell populations, SMic, LMic, Mac and Mes, which gives us 12 boundaries (considered asymmetrically): (SMic → LMic, SMic → Mac, ..., Mac → Mes, Mes → Mac). For each of them (symbolized as $g \to l$), we computed the normalized histogram of the shared number of contacts at the border $\mathscr{B}^t_{g\to l}$, $h^t_{b,g\to l}$. As for the number of neighbors, natural bins are formed by the integers. The frequency of occurrence of a bin $n \in \mathbb{N}$ is :

$$p^t_{b,g\to l}(n) = \frac{1}{|\mathscr{B}^t_{g\to l}|}.|\left\{i \in \mathscr{B}^t_{g\to l} : b^t_{i,g\to l} = n\right\}|$$

the complete histogram is the vector $h_{b,g\to l}(t) = (p^t_{b,g\to l}(1), p^t_{b,g\to l}(2), ..., p^t_{b,g\to l}(S))$, with S with the maximum number of neighbors shared at a border. The histogram is normalized $\Sigma^S_{k=1} p^t_{b,g\to l}(k) = 1$ and can be considered as a density distribution.

Measures of similarity To measure the similarity of two histograms we chose to compute the Matusita distance which is related to the cosine of the angle between the two histograms and varies between 0 (identical) to 1 (opposite). First we define the Bhattacharyya coefficient (Bh) between two histograms p and q having the same set of bins X.

$$\text{Bh}(p, q) = \Sigma_{x\in X} \sqrt{p(x)}\sqrt{q(x)} = cos(\theta)$$

where θ is the angle between the vectors $(\sqrt{p(1)}, \sqrt{p(2)}, ..., \sqrt{p(N)})$ and $(\sqrt{q(1)}, \sqrt{q(2)}, ..., \sqrt{q(N)})$, equal to 1 when the distributions are identical. This coefficient is defined for two continuous density distribution as $Bh(p, q) = \int \sqrt{p(x)}\sqrt{q(x)}dx$. The Matusita distance is defined as $\text{Mat}(p, q) = \int (\sqrt{p(x)} - \sqrt{q(x)})^2 dx$, the two quantities are related by the relationship: $\text{Mat}(p, q) = 2 * (1 - Bh(p, q))$. In the following, we will use the discretized version of this

distance defined for discrete histograms $(p(1), p(2), ..., p(N))$, $(q(1), q(2), ..., q(N))$:

$$d(p,q) = \frac{1}{2}\Sigma_{x\in X}((\sqrt{p(k)} - \sqrt{q(k)})^2 * \text{db})$$

where db is the size of the bin x. This distance is equal to zero if the two distributions are identical and equal to one when they are opposite, it is applicable to any kind of distribution and even if any bin is empty (which is a reason of preferring this measure to the chi-squared statistics).

Histogram resampling to calculate similarities As stated previously the first step of the comparison between the embryo and the spatial simulation consists in comparing the distribution of the features at different time steps. An issue is that the digital embryos have all their own time sampling, ranging from 2min per time step to 5min. To overcome this problem, we decided to use a common time sampling for all the embryos and for the spatial simulation. We used a time interval ranging from 240 min post fertilization to 702 min, with a sample every 3 min. We interpolated the value of the histograms used for the comparison at each time point of this time interval.

Given a simulated embryo s:
for *each specimen* e_i *of the cohort* **do**
  for *each subpopulation (k), or ordered couple of subpopulation* $(g \rightarrow l)$ **do**
    for *each time t of the common time interval* **do**
      compute the three distances
      $d(h^t_{d,k,e_i}, h^t_{d,k,s})$, $d(h^t_{c,k,e_i}, h^t_{c,k,s})$,$d(h^t_{b,g\rightarrow l,e_i}, h^t_{b,g\rightarrow l,s})$;
    end
    average over the entire time interval for each distance: $\overline{d}_{d,k,e}$, $\overline{d}_{c,k,e}$, $\overline{d}_{b,g\rightarrow l,e}$;
  end
  average over all the subpopulations: $\overline{\overline{d}}_{n,e}$,$\overline{\overline{d}}_{c,e}$,$\overline{\overline{d}}_{b,e}$;
end
average over the cohort: $D_n = \overline{\overline{\overline{d}}}_n$, $D_c = \overline{\overline{\overline{d}}}_c$, $D_b = \overline{\overline{\overline{d}}}_b$;

Algorithm 1: Comparison protocol for the three topological metrics

With this procedure, we obtain a measure of similarity for each of the three topological feature that was used to define a protocol of comparison as shown on the Algorithm 1.

2.7.2 Objective functions

In addition to the three metrics introduced above, we measure intrinsic features of the simulated embryos via objective functions. These functions are defined in the following for a given time step. They are subsequently averaged over the entire time interval.

Tissue cohesion objective function C_s The tissue cohesion objective function, which represents the spacing between neighbor cells. Neighbor cells i, j tend to maintain their effective distance r_{ij} close to their equilibrium distance r_{ij}^{eq} but they may detach from each other depending on their biomechanical properties. The definition of C_s reads:

$$C_s = \frac{1}{|\mathscr{E}|} \sum_{i \in \mathscr{L}} \sum_{j \in \mathscr{N}_i} \frac{\left| r_{ij} - r_{ij}^{\text{eq}} \right|}{r_{ij}^{\text{eq}}}$$

where $\mathscr{E}$ is the of edges in the network of cellular contacts, $\mathscr{L}$ the set of cells and for a cell i, $\mathscr{N}_i$ is its neighborhood.

Sphericity objective function S_s and N_s An important criteria to evaluate the model is the shape of the simulated embryos. We designed two functions, S_s and N_s for that purpose. S_s is the ratio between standard deviation and the average of the set of distances between the cell positions X_i and the embryo center $\overline{X}$. The rationale is that this function has a low value for a small standard deviation and a large average. Its value is null in the case of a spherical embryo.

$$S_s = \frac{\sigma}{\mu} \quad \text{where } \mu = \frac{1}{|\mathscr{L}|} \sum_{i \in \mathscr{L}} \left| X_i - \overline{X} \right| \quad \text{and } \sigma = \sqrt{\frac{1}{|\mathscr{L}|} \sum_{i \in \mathscr{L}} (\left| X_i - \overline{X} \right| - \mu)^2}$$

The second sphericity objective function N_s is similar to S_s and its rationale is that in the case of a spherical embryo, the cell axes $\vec{U}_i$ are equal to their normalized relative position vector $\vec{X}_{i,\text{center}} = \frac{X_i - \overline{X}}{\left| X_i - \overline{X} \right|}$. The set of dot products between $\vec{U}_i$ and $\vec{X}_{i,\text{center}}$ should have a minimal standard deviation and an average value close to 1. The equation of the function N_s reads:

$$N_s = \frac{\sigma}{\mu} \quad \text{where } \mu = \frac{1}{|\mathscr{L}|} \sum_{i \in \mathscr{L}} \frac{1 + \vec{U}_i \cdot \vec{X}_{i,\text{center}}}{2} \quad \text{and } \sigma = \sqrt{\frac{1}{|\mathscr{L}|} \sum_{i \in \mathscr{L}} (\frac{1 + \vec{U}_i \cdot \vec{X}_{i,\text{center}}}{2} - \mu)^2}$$

Planarity objective function P_s For each cell, the planarity objective function P_s aims at characterizing the unicity of the epithelium layer by measuring the ratio of cells belonging to the surrounding apico-basal neighborhood. We split the set of neighbors $\mathcal{N}_i$ into two complementing sets, one covering the lateral domain $\mathcal{N}_i^{\parallel}$ and the other the apical domain $\mathcal{N}_i^{\perp}$. They are defined as the following:

$$\mathcal{N}_i^{\parallel} = \left\{ j \in \mathcal{N}_i : \frac{\vec{X}_j - \vec{X}_i}{\left\| \vec{X}_j - \vec{X}_i \right\|} \cdot \vec{U}_i < \eta \right\}$$

$$\mathcal{N}_i^{\perp} = \left\{ j \in \mathcal{N}_i : \frac{\vec{X}_j - \vec{X}_i}{\left\| \vec{X}_j - \vec{X}_i \right\|} \cdot \vec{U}_i \geq \eta \right\}$$

A threshold value η controls the repartition between the two sets. Here, we set $\eta = cos(\frac{\pi}{4})$. The planarity objective function is written:

$$P_s = \frac{1}{|\mathcal{L}|} \sum_{i \in \mathcal{L}} \frac{|\mathcal{N}_i^{\perp}|}{|\mathcal{N}_i|}$$

In the case of a mono-layered epithelium P_s value is 0 and its value increases as the cells agglomerate in a 3D tissue.

2.7.3 Initial State

For the spatial simulation, an initial spatial state of the cells was needed in order to begin the simulation at the 32 cells stage. We choose to use the measured initial state of one of the five embryos of the cohort. This solution, although not optimal, was more satisfying than an artificial 32 cells stage. We wanted to avoid a spatial averaging that would have been without biological meaning.

2.7.4 Validation - Parameter space

The purpose of this exploration study is to determine the parameter sets which are responsible for realistic spatial enfolding of the sea urchin embryo development. The criteria of validation are the sphericity of the global embryo shape, the maintenance of the monolayered epithelium and the similarity with the empirical degree, rate of change of the neighborhood and the inter-subpopulation border shapes. This parameter space of the simulation is 4D, comprising the planar rigidity coefficient k_{rig}, the gabriel criteria coefficient α_{gab} and two specific adhesion coefficients controlling the attractive part of the

relaxation force $\vec{F}_{ij}^{\parallel}$: $w_{\text{adh,e}}$ between pairs of heterotypic cells and $w_{\text{adh,o}}$ between pairs of homotypic cells. The other free parameters of the model are set at constant values (Table 2.5). The timestep Δt between two simulated time step is 6 seconds.

Note concerning the force amplitude coefficients $w_{\text{adh,e}}$, $w_{\text{adh,o}}$, w_{rep} and k_{rig} : these coefficients are all divided by the damping coefficent λ_0 in the master equation of motion. Simulations are strictly equivalent if the ratio force amplitude coefficient over damping remains constant.

Table 2.5: Range and cardinalities of the parameters explored in this study.

	Min.	Max.	Cardinality
$w_{\text{adh,e}}$	10	1000	20
$w_{\text{adh,o}}$	10	1000	20
w_{rep}	100	100	1
k_{rig}	6000	16000	11
α_{gab}	0.9	1.3	0.02
λ_0	3000	3000	1
c_{max}	2	2	1

Results of the exploration of the parameter space are shown on figure 1.3 D.

Chapter 3

Perspectives and open problems raised by the probabilistic model of development

Abstract *In this chapter we explore perspectives raised by the data-driven multi-level probabilistic model developed in the previous chapter. We begin by characterizing the algebraic structure underlying the relations between random variables, leading to identyifing a Monoid structure. This structure can form the basis of a dynamical system representation of the model enabling to derive some properties such as stability of the dynamics or law of evolution. We also use this model to study its relation to previous theoretical proposition in developmental biology.*

The data-driven multi-level probabilistic model obtained in the previous chapter can be used to discuss several aspects of developmental processes. Among these aspects, we will explore the ideas of robustness, developmental irreversibility, chreodes and epigenetic landscape. The main goal of this chapter is to identify some interesting perspectives.

We will begin by characterizing this idea of developmental irreversibility that arise from the structure of the model.

Let's first recall the main characteristics of the multi-level probabilistic model obtained in the previous section operating at the level of the group of cells $\mathscr{L}^{n,k}$ and their combination through the lineage :

Cell group	$\mathscr{L}^{n,k}$
Cardinality	$\lvert\mathscr{L}^{n,k}\rvert = 2 * \lvert\mathscr{L}^{n-1,k}\rvert = 2^{n-n^{0,k}} \lvert\mathscr{L}^{n^{0,k},k}\rvert$
Division time	$M_{n,k} = X_{n,k} + M_{n-1,k} = \left(\sum_{r=n^{0,k}+1}^{n} X_{r,k}\right) + M_{n^{0,k}}$ $M_{n,k} \sim N(\mu_M^{n,k}, \sigma_M^{n,k})$ $X_{n,k} \sim N(\mu_X^{n,k}, \sigma_X^{n,k})$ $P(X_{n,k} \mid M_{n-1,k}) = P(X_{n,k})$ $\mu_M^{n,k} = \sum_{r=n^{0,k}+1}^{n} \mu_X^{r,k} + \mu_M^{n^{0,k},k}$ $(\sigma_M^{n,k})^2 = \sum_{r=n^{0,k}+1}^{n} (\sigma_X^{r,k})^2 + (\sigma_M^{n^{0,k},k})^2$
Mean volume	$\overline{V}_{n,k} = A_{n,k}\overline{V}_{n-1,k} = \left(\prod_{r=n^{0,k}+1}^{n} A_{r,k}\right)\overline{V}_{n^{0,k},k}$ $\overline{V}_{n,k} \sim lnN(\mu_V^{n,k}, \sigma_V^{n,k})$ $A_{n,k} \sim lnN(\mu_A^{n,k}, \sigma_A^{n,k})$ $P(A_{n,k} \mid \overline{V}_{n,k}) = P(A_{n,k})$ $\mu_V^{n,k} = \sum_{r=n^{0,k}+1}^{n} \mu_A^{r,k} + \mu_V^{n^{0,k},k}$ $(\sigma_V^{n,k})^2 = \sum_{r=n^{0,k}+1}^{n} (\sigma_A^{r,k})^2 + (\sigma_V^{n^{0,k},k})^2$
Mean surface area	$\overline{S}_{n,k} = B_{n,k}\overline{S}^{n-1,k} = \left(\prod_{r=n^{0,k}+1}^{n} B_{r,k}\right)\overline{S}_{n^{0,k},k}$ $\overline{S}_{n,k} \sim lnN(\mu_S^{n,k}, \sigma_S^{n,k})$ $B_{n,k} \sim lnN(\mu_B^{n,k}, \sigma_B^{n,k})$ $P(B_{n,k} \mid \overline{S}^{n,k}) = P(B_{n,k})$ $\mu_S^{n,k} = \sum_{r=n^{0,k}+1}^{n} \mu_B^{r,k} + \mu_S^{n^{0,k},k}$ $(\sigma_S^{n,k})^2 = \sum_{r=n^{0,k}+1}^{n} (\sigma_B^{r,k})^2 + (\sigma_S^{n^{0,k},k})^2$

Table 3.1: Summary of the relations between the different random variables in the model - the definitions of the different notations can be found in the previous sections 2.5, 2.5, 2.5

Cells	$j \in \mathscr{L}^{n-1,k}$ and $i \in \mathscr{L}^{n,k}$ i is one of the two daughters of j
Division time	$m_i = x_i + m_j$ x_i **is a realization of** $X_{n,k}$ m_j **is a realization of** $M_{n-1,k}$
Volume	$v_i(t) = a_i.\overline{v}_j.\omega^{n,k}(u_i(t))$ a_i **is a realization of** $A_{n,k}$ $\overline{v}_j$ **is a realization of** $\overline{V}_{n-1,k}$ $\omega^{n,k}$ is the deterministic volume micro dynamic $u_i(t)$ is the percentage of elapsed life length of cell i
Surface area	$s_i(t) = b_i.\overline{s}_j.\phi^{n,k}(u_i(t))$ b_i **is a realization of** $B_{n,k}$ $\overline{s}_j$ **is a realization of** $\overline{S}^{n-1,k}$ $\phi^{n,k}$ is the deterministic surface area micro dynamic $u_i(t)$ is the percentage of elapsed life length of cell i

Table 3.2: Instantiation of the model independently in any branch of the cell lineage. The stochastic components are highlighted in bold- the definitions of the different notations can be found in the previous sections 2.5, 2.5, 2.5

3.1 The probabilistic model implies a monoid structure

3.1.1 Monoid Structure

It is interesting to look at the algebraic structure underlying the successive addition, or multiplication, of independent variables, either for the relation: $M_{n,k} = X_{n,k} + M_{n-1,k}$ or for the relations $\overline{V}_{n,k} = A_{n,k}.\overline{V}_{n-1,k}$ and $\overline{S}_{n,k} = B_{n,k}.\overline{S}_{n-1,k}$, which are equivalently described as addition of independent variables as $\ln \overline{V}_{n,k} = \ln A_{n,k} + \ln \overline{V}_{n-1,k}$ and $\ln \overline{S}_{n,k} = \ln B_{n,k} + \ln \overline{S}_{n-1,k}$. The basic property of a sum of independent random variables is that their probability laws are convolved. This property is particularly useful in the case of gaussian distributions because the convolution of two gaussian distributions is a gaussian distribution. Since all the probability considered here are gaussian distributiion (for the ln version of the volume and surface area), we obtain a monoid structure for the random variables equipped with the addition.

Sum of independent gaussian random variable It can be shown that the set of gaussian distributions with the convolution product forms a Monoid. The neutral element is $\mathcal{N}(0,0)$, the convolution of a gaussian distribution gives a gaussian distribution (stability). If gaussian distribution has a strictly positive standard deviation, then it cannot be inversed. The evolution of the distribution of mitosis can only be toward more desynchronization and thus disorganization. The development can be seen as the monoid action on the group of cells. In terms of symmetry, it can be said that the symmetry identified with the exchangeability principle are conserved by the monoid action.

Proof. Let's denote $\mathcal{S}$ the set of gaussian distribution and $*$ the convolution product. $(\mathcal{S}, *)$ is a monoid if the operation $*$ is stable, associate and has an identity element.

stability For two gaussian distributions with parameters $((\mu_1,\sigma_1),(\mu_2,\sigma_2))$, we have $f(x|\mu_1,\sigma_1) * f(x|\mu_2,\sigma_2) = f(x|\mu_1+\mu_2,\sqrt{\sigma_1^2+\sigma_2^2})$. The convolution product of two gaussian distributions is a gaussian distribution.

$$\forall (x,y) \in \mathcal{S}^2, x * y \in \mathcal{S}$$

commutativity For two gaussian distributions with parameters $((\mu_1,\sigma_1),(\mu_2,\sigma_2))$, we have

$$\begin{aligned} f(x|\mu_1,\sigma_1) * f(x|\mu_2,\sigma_2) &= f(x|\mu_1+\mu_2,\sqrt{\sigma_1^2+\sigma_2^2}) \\ &= f(x|\mu_2+\mu_1,\sqrt{\sigma_2^2+\sigma_1^2}) \\ &= f(x|\mu_2,\sigma_2) * f(x|\mu_1,\sigma_1) \end{aligned}$$

Therefore, $\forall x, y \in \mathscr{S}^2, x * y = y * x$.

associativity For three gaussian distributions with parameters $((\mu_1,\sigma_1),(\mu_2,\sigma_2),(\mu_3,\sigma_3))$, we have:

$$\begin{aligned} (f(x|\mu_1,\sigma_1) * f(x|\mu_2,\sigma_2)) * f(x|\mu_3,\sigma_3) &= f(x|\mu_1+\mu_2,\sqrt{\sigma_1^2+\sigma_2^2}) * f(x|\mu_3,\sigma_3) \\ &= f(x|(\mu_1+\mu_2)+\mu_3,\sqrt{(\sigma_1^2+\sigma_2^2)+\sigma_3^2}) \\ &= f(x|\mu_1+(\mu_2+\mu_3),\sqrt{\sigma_1^2+(\sigma_2^2+\sigma_3^2)}) \\ &= f(x|\mu_1,\sigma_1) * f(x|\mu_2+\mu_3,\sqrt{\sigma_2^2+\sigma_3^2}) \\ &= f(x|\mu_1,\sigma_1) * (f(x|\mu_2,\sigma_2) * f(x|\mu_3,\sigma_3)) \end{aligned}$$

The convolution product is an associative operation on the set of gaussian distribution $\mathscr{S}$.

identity element An element $e \in \mathscr{S}$ is an identity element if $\forall x \in \mathscr{S}, e * x = x * e = x$. $*$ is commutative, thus we need to find e with parameters $(\mu_e,\sigma_e) \in \mathbb{R}^2$ such that $\forall x \in \mathscr{S}$ with parameters $(\mu,\sigma) \in \mathbb{R}^2$, $f(x|\mu_e,\sigma_e) * f(x|\mu,\sigma) = f(x|\mu_e+\mu,\sqrt{\sigma_e^2+\sigma^2}) = f(x|\mu,\sigma)$. The solution is a gaussian distribution with parameters $\mu_e = 0, \sigma_e = 0$. This element is the Dirac delta function and is denoted by δ.

Given the properties of the convolution product on the set of gaussian distributions, $(\mathscr{S},*)$ is said to form a monoid. □

A corollary of this result is that the tendency towards more desynchronization is irreversible. Indeed, the successive convolutions of gaussian distributions necessary lead to a gaussian distribution with a greater variance. This operation is not invertible (except for gaussian distribution with null variance).

In biological terms, it means that the rounds of division are less and less synchronized throughout development. Similarly the dispersion of the volume and surface area

increases with time, even if their mean values are decreasing. This irreversibility seems to underlie the process of cell differentiation. As written in Davidson *et al.* [53]

> the fundamental feature of developmental transcriptional systems in higher (bilaterian) animals is that it always moves inexorably forward, never reversing direction

Here, it is the dispersion of the morphological features that goes inexorably toward being more spread.

However, this result rely on the fact that the random variables defined in each groups of cells are supposed independent of each other. It is possible to conceive mechanisms that would enable cells to control dispersion while proliferating. Regulation of the cell cycle and other morphological features are required to avoid going inexorably toward more disorganization.

3.1.2 Formalization as a dynamical system

This underlying algebraic structure as a monoid allows to describe the development as a dynamical system. To define this system (following the definition on wikipedia) it is necessary to have a tuple (T, M, ψ) where T is a monoid, M is a set and ψ is a function:

$$\psi : U \subset T \times M \to M$$

and ψ has the properties of a flow:

- $I(x) = \{t \in T : (t, x) \in U\}$
- $\psi(0, x) = x$
- $\psi(t, \psi(s, x)) = \psi(t + s, x)$

The structure of monoid here is given by the set of gaussian distribution together with the convolution product and the set of M correspond to the set of groups of cells embedded in the manifold defined by the coordinate system (μ, σ), which is the phase space of the system. The function ψ is the evolution function, and an orbit/trajectory of the system is the set $\{\psi(t, x) : t \in I(x)\}$ and correspond to the development of one specimen in the cohort.

However, the function ψ is unknown and can only be obtained empirically for now, i.e. interpolated from the sequence of points for each developmental trajectories.

With a dynamical system formalism and a phase space, we can define a notion of structural stability and a notion of vector field in the phase space. The vector field in the phase

space can be interpolated between the measured points and from that we can expect to compute Lyapounov experiments or topological entropy that gives a measure of the complexity of development. The structural stability could be investigated by characterizing singularity points of the vector field and the periodic trajectories.

Fokker-Planck Equation One possibility for the analytic characterization of the function ψ is to characterize the evolution of probability distribution through time by way of an equation. One candidate equation could be the Fokker-Planck equation, indeed

> Brownian motion of a particle is described by a stochastic differential equation $dX_t = \mu dt + \sigma dW_t$, where the X_t are particle positions in $\mathbb{R}^n$. μ is the drift velocity, σ is an $n \times n$ matrix and dW_t represents an n-dimensional normal Wiener process. The Fokker–Planck equation (also called forward Kolmogorov equation) describes the temporal evolution of the probability density $p(X_t)$:
>
> $$\frac{\partial p}{\partial t} = -\nabla.(\mu.p) + \nabla.(D\nabla p), \text{ where } D = \frac{1}{2}\sigma\sigma^T$$
>
> If μ and D are constant, the Fokker–Planck equation reduces to a drift-diffusion equation that can be solved analytically. **The fundamental solutions are Gaussian distributions which drift and widen with time.** [1]

The fundamental solutions of this equation can be fitted to the empirical gaussian distributions obtained in the sea urchin embryo. The use of this equation on our system could be interpreted as the random diffusion of the cells in a morphological space with a drift generated by cell differentiation.

3.1.3 Fluctuation theory and robustness

It could be interesting to characterize the robustness of the model by studying its behavior when submitted to small fluctutations. We know that after three cell cycles, the probability density of a random variable f_{123} such as the division time is the result of the convolution of probability densities of random variables such as the life length (f_1, f_2, f_3). We have the equality

$$f_{123} = f_1 * f_2 * f_3$$

1. "Brownian Motion in 2D and the Fokker-Planck Equation" from the Wolfram Demonstrations Project - http://demonstrations.wolfram.com/BrownianMotionIn2DAndTheFokkerPlanckEquation/

We would like to relate small deviation on the density distribution Δf_1, Δf_2, Δf_3 with deviation on the final probability density Δf_{123}. The result should indicate how the perturbations at different stages of development affect the final result. It is likely that earlier fluctuations are more amplified than late ones.

More formally, we would like to show a relation such that $\Delta f_{123} = \Delta f_1 * f_2 * f_3 + f_1 * \Delta f_2 * f_3 + f_1 * f_2 * \Delta f_3 + 2.(\Delta f_1 * \Delta f_2 * f_3 + \Delta f_1 * f_2 * \Delta f_3 + f_1 * \Delta f_2 * \Delta f_3) + \Delta f_1 * \Delta f_2 * \Delta f_3$ can be simplified into $\Delta f_{123} = \Delta f_1 * f_2 * f_3 + f_1 * \Delta f_2 * f_3 + f_1 * f_2 * \Delta f_3$ by considering only small Δ.

However, to define these fluctuations, we would be interested to keep the same metric that we have used to compare embryos and to define the prototype as a centroid. These equalities would require to combine Kullback Leibler divergence (relative entropy) with a convolution.

On this point, we can quote Oliver Johnson ([114] p.33):

> Although we would like to prove results concerning the behaviour of relative entropy on convolution, it proves difficult to do so directly, [...] Specifically the logarithm term in the definition of entropy behaves in a way that is hard to control directly on convolution.

Therefore, even if the metric that we have used to define the prototypical probabilistic model is highly efficient to compute a prototype in the statistical manifold it is less suited for a perturbation theory in developmental biology.

3.2 Parameters evolution

3.2.1 Waddington's epigenetic landscape

The multi-level probabilistic model obtained for each specimen of the cohort of sea urchin can be fruitfully brought closer to the old idea of an epigenetic landscape during development. Indeed, the highest unknown that stays after having formalized this model is the following question "Why does the probability laws have such parameters values?". The life length increases at each generation, the mean volume is divided, but not exactly in half.

One possibility is that the value of these parameters is the result of dynamics of the underlying genetic network that governs these characteristics and which may imply a certain level of variability as well. In that case, the concept an epigenetic landscape is a relevant

illustration of this idea, see figures 3.1 and 3.2. In our model, at each cell cycles, the valleys widen as the result of the increase in variability (the standard deviations are added). This increase in variability may well result in differentiation of cells into subpopulations. The idea of the cells being pushed forward into the valleys is also relevant since the model present an additive behavior with no regulation.

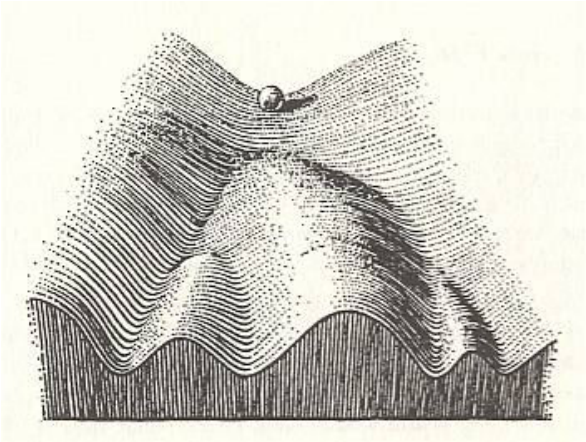


Figure 3.1: Epigenetic landscape I. "The path followed by the ball, as if rolls towards the spectator, corresponds to the developmental history of a particular part of the egg." (Waddington CW. The Strategy of the Genes. London, Allen and Unwin (1957), [215])

3.2.2 Kupiec's ontophylogenesis

Finally this work can be brought closer to Kupiec theory of ontophylogenesis. Jean-Jacques Kupiec claimed that the processes occurring during ontogenesis, i.e. during development, could be considered as probabilistic ([128],[127]). The reproducibility of the development coming from a principle analog to natural selection within the organism. The results of our study don't disprove this approach since individual cell features can indeed be modeled with probability distribution and the parameters values of these probability distributions may well arise from the interaction between the cells and their environment.

Overall, this probabilistic perspective on development sheds new light on the robustness of development. The high reproducibility of development seems to emerge from highly variable individual cell features. This perspective is at the opposite of the idea of the execution of a finely tuned developmental program. Therefore, we may need a new theoretical framework to understand development.

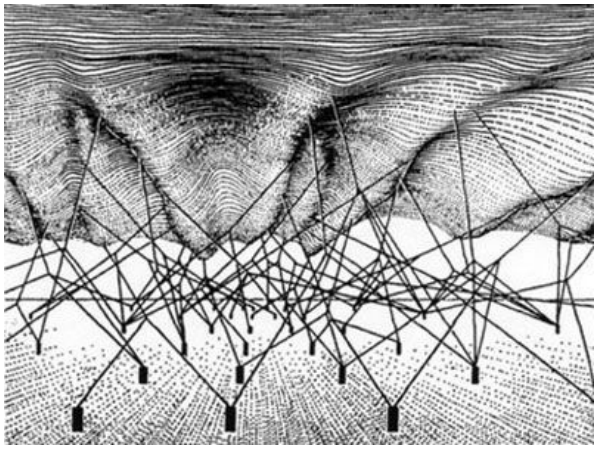


Figure 3.2: Epigenetic landscape II. “The complex system of interactions underlying the epigenetic landscape. The pegs in the ground represent genes, the strings leading from the chemical tendencies that the genes produce. The modeling of the epigenetic landscape, which slopes down from above one’s head towards the distance, is controlled by the pull of these numerous guy-ropes which are ultimately anchored to the genes.” (Waddington CW. The Strategy of the Genes. London, Allen and Unwin (1957), [215])

Conclusion

To conclude on this part we have shown the first example of a complete integrative approach for the study of multi-scale dynamics during the sea urchin development. This work aimed at characterizing intra- and inter-individual variability at the level of the individual cell. By studying the dynamics occurring in a cohort of 5 digital embryos, we showed that an intermediate level of observation was required between reproducible embryo-level dynamics and variable individual cell features. This coarse-grained level of observation served as a basis for the establishment of a data-driven multi-level probabilistic model. This model was found to have an invariant structure among individuals of the cohort. Taking advantage of the branching structure of the cell lineage, the value of parameters governing probability in the model were sufficient to identify uniquely specimen and reproduce embryo-level dynamics. The invariance of the structure relating individual cell features and embryo-level dynamics enabled to define a prototypical representation of the cell lineage associated to individual cell features.

This data-driven prototypic representation of the cell lineage was then used as a basis for a biomechanical modeling using the MecaGen platform. This model allowed to infer the value of adherence parameters by systematic parameter exploration and fit to empirical data. Because of its data-driven nature, the prototype was able to free the biomechanical model of a certain number of parameters governing the proliferation rates or the shape of cells. The remaining free parameters were associated to the hypotheses underlying the biomechanical model.

Finally, this data-driven multi-level probabilistic model sheds new light and opens some perspective on theoretical conception of the development. Indeed, the traditional view of development as a finely tuned process is weakened by the observed variability at the individual cell level. The good adequacy of the empirical data with the probabilistic description enables to suggest some theoretical hypothesis, such as the development could rely on this variability at the individual cell level. This could be interpreted as an

extension of the evolutionary process within the organism. Although empirical results are also compatible with a more traditional epigenetic interpretation such as the one developed by Waddington by means of an epigenetic landscape. To sum up the content of this part, there is an intrinsic variability at the individual cell level underlying cell differentiation and this variability is involved in the reproducibility of embryo-level dynamics.

Part II

Characterizing diversity

Introduction

We propose to study in this part of the dissertation the theoretical status of biological variability and the relations of variability during development with darwinian evolution as well as its relation to randomness.

In the first chapter, we introduce the question of the relations between randomness, variability and diversity in biology. We review the mechanisms at the origin of variation, how uncertainty is modeled in mathematics and physics and how hypothesis based model can help to explore the range of possible at the price of approximations on other levels of organization.

In the second chapter of this part, we present the experimental results obtained when studying **variable phenotypic expression and incomplete penetrance** in the *squint*^{cz35} zebrafish mutant line. The results of this experiment show that the list of possible phenotypes is discrete, these phenotype are obtained in unpredictable proportions and the variations in the list of possible phenotypes may be incomplete even if the parents are homozygote mutants. These experimental results suggest a need for a clarification of the different levels of variability in biology, this is the subject of the following chapter.

In the third chapter of this part we study the relations between variability during development and evolution. **Using a formal analogy with quantum mechanics, we propose a clarification of the various levels of variability**; probability of obtaining a given phenotype, uncertainty on the set of possible phenotypes for a given observable and unpredictability at the level of the observable itself. Uncertainty on the set of observable is a much stronger form of uncertainty than a probability of obtaining a given phenotype, it is specific to the historical nature of biological objects and prevent to consider the space of possible *a priori*. Surprisingly, we find a formal analogy between quantum entanglement and Mendel's idealized scheme of inheritance which we relate to biological organization.

In the fourth chapter of this part, we consider the relationships between variations in individual development and observed diversity of phenotypes. Given the path-dependency

and historical nature of development we propose to gather individual developments, normal and pathological, into an **ontogenetic tree**. This ontogenetic tree structure enables to define a developmental proximity between phenotypes revealing the influence of developmental stages. By considering a large number of empirical descriptions of zebrafish developments we show that the pharyngula stage has the highest number of diverging developmental paths, suggesting an empirical basis for its status of phylotypic stage. However, the data set used may include possible biases.

Chapter 4

Sources of biological diversity and randomness

Abstract *In this chapter, we review the various concepts of variability developed in biology, from gene mutation to stochastic gene expression and epigenetic effects. We then turn to mathematical and physical theories of randomness and models of uncertainty. Probability theory provides a framework to handle unpredictable events, but doesn't provide a definition for randomness. This theory is extended to capture specific aspects of quantum mechanics. Characterizations of randomness are provided for chaotic systems and in ergodic theory. To overcome the complexity of biological organisms, one possibility consists in using modeling approaches describing some mechanisms and explore regions of the space of possible. However, the multi-scale nature of dynamics occurring in organisms make those models necessarily incomplete. We propose to use multi-scale prototypical model to ground modeling approaches.*

This chapter is an extended and translated version of [210]. The idea of this chapter is to review the various mechanisms that can be sources of variability in biology, from gene mutation, to epigenetic and stochastic effects. Once their main characteristics have been exhibited, we then turn to mathematics and physics to see how unpredictable phenomena have been formalized. However, we show that the different frameworks are heterogeneous. We then explore the possibility to use models in biology to describe space of possible associated to some mechanisms. While reducing the generality of the statements, these models may integrate several aspects of biological variability. However, they are necessarily incomplete regarding the multi-scale and historical nature of organisms. These obstacles may be overcome by the use of multi-scale prototype of biological organ-

isms.

4.1 Sources of variability in biology

In this section we review the main mechanisms sources of variability. We begin with the most established results before considering more recent ones on stochasticity in gene expression.

4.1.1 Gene mutations

Permanent alteration of the DNA sequence constitutes the most famous source of variation in biology. Some mutations correspond to a punctual alteration of the nucleotide sequence: they can lead to changes in the encoded protein by the mutated gene, or to a complete absence of its expression. Other mutations correspond to a more significant modification of the DNA, as a chromosome gain for example (long chain of nucleotides). In any cases, if the mutations are not silent, they alter the considered organism. For example, with the zebrafish *Danio rerio*, if the gene encoding the Oep protein, receptor of the Nodal signaling pathway, the development fails: fishes become cyclops [169].

It is usually assumed that these mutations happen spontaneously, independently of their fitness in a given environment. The question of the influence of the environment on the variation of organism's traits is associated to a debate between Jean-Baptiste de Lamarck and Charles Darwin theories on evolution. For the first one, whose theory is the "transformisme Lamarckien", variations in the environment induce variations in organisms and these variations are oriented, with the internal efforts of the organism, toward the goal of being adaptive. For example, the fact that food is more abundant in height would have led giraffes to increase the length of their neck and their front legs while trying to reach it [1]. For Darwin, and in contrast to Lamarck, variations occur independently of their fitness in the environment. Living beings are continuously varying. Less adapted individuals will be disadvantaged with respect to others and thus reproducing in smaller

1. « Relativement aux habitudes, il est curieux d'en observer le produit dans la forme particulière et la taille de la girafe (camelo-pardalis) : on sait que cet animal, le plus grand des mammifères, habite l'intérieur de l'Afrique, et qu'il vit dans des lieux où la terre, presque toujours aride et sans herbage, l'oblige de brouter le feuillage des arbres, et de s'efforcer continuellement d'y atteindre. Il est résulté de cette habitude, soutenue, depuis longtemps, dans tous les individus de sa race, que ses jambes de devant sont devenues plus longues que celles de derrière, et que son col s'est tellement allongé, que la girafe, sans se dresser sur les jambes de derrière, élève sa tête et atteint à six mètres de hauteur (près de vingt pieds). » [129], p. 256.

proportions. A clear separation between variability and fitness can be found in Darwin's theory.

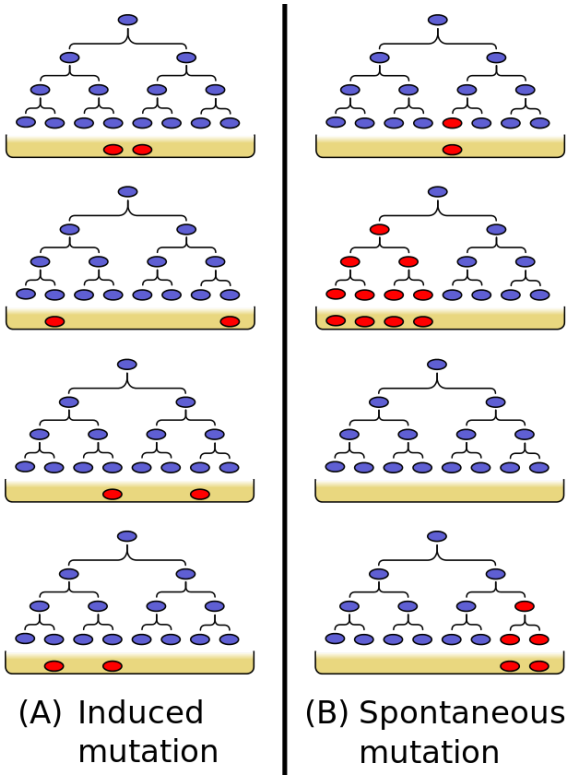


Figure 4.1: Illustration of Luria Delbrück fluctuation test. Schematic trees represent cell proliferation in an environment containing a virus. Red colonies survive the virus attack. Mutations induced by the presence of virus, case (A), are present in similar proportions in every tree. Spontaneous mutations enabling survival to virus attack, case (B), appear randomly in the trees resulting in highly fluctuating size of colonies - Adapted from wikipedia

In the beginning of the 1940s, before the discovery of the DNA structure by Watson and Crick, it had been observed that bacteria had the capacity to adapt and survive an attack of bacterial virus. In 1943, Luria and Delbrück published a famous article demonstrating that the capacity of *Escherichia coli* to acquire immunity appears with time, independently of the presence of the virus [139]. This result was proved by measuring fluctuations on the size of colonies surviving virus attacks and comparing these fluctuations to a theoretical model. The theoretical model enables to differentiate between mutations induced by the virus presence and spontaneous mutations. See figure 4.1 for a schematic explanation of the model. Acquisition of immunity have been associated to genetic mutations. By proving that this acquisition occurs spontaneously, and thus independently of their fitness,

their non Lamarckian nature was validated. Therefore, this result contributed to give to genetic mutation the status of main source of diversity for the living [153], [151].

It has been shown since, that the mutation rate of some bacteria could be increased in response to variations in the environment [24]. Some mechanisms have a direct action on the rate at which the genome varies in time. This is the case of the SOS system in *Escherichia coli*, or as a response to the stress induced by changes in the temperature of the environment or changes in the amount of nutrient. These examples show that internal variations depend on environmental variations. The environment has a direct influence on the mutation rate. Mutations are however not directed towards a better fitness: this phenomenon does not correspond to a movement back toward Lamarckism. The environment can induce a more or less important capacity to vary, without reconsidering the idea that these variations are independent of their fitness.

Genetic mutations are an important source of variation in biology. They form the basis of many mathematical evolutionary theories, particularly in population genetics [140]. In these theoretical frameworks, mutations are supposed to be spontaneous, independent of each other and independent of their fitness. However, these assumptions carry at least two presuppositions that can be discussed. The first one corresponds to the assumption of a direct (linear) relation between a gene and a phenotypic trait; the second one corresponds to the hypothesis of an independent variation between genes or phenotypic traits within an organism and with respect to the environment, this is an assumption on the modularity of variation [2]. The organization of living beings, the coupling between several levels of organization, the fact that the whole and the parts form a complex network of relations where retroactions are numerous, and where relations with the environment modify the internal space, are many reasons to suppose on one side that the linear relation between a gene or a set of genes and a phenotypic trait is not necessarily unequivocal, and on the other side that variation cannot be split up among the different parts of an organism. [3]

2. Some models take into account interactions between genes in the formation of phenotypes with the concept of epistasis, complicating the picture [174]. Conversely, some models take into account the effect of mutations on several phenotypes [201]

3. Darwin's definition of correlated variations is very meaningful here; "Correlated Variations - I mean by this expression that the whole organization is so tied together during its growth and development, that when slight variations in any one part occur, and are accumulated through natural selection, other parts become modified." [49] - chapter 5

4.1.2 Epigenetic and stochasticity

In parallel to the study of genetic mutations, works have been conducted on other mechanisms sources of variation. It is the case for the phenomena grouped together under the term "epigenetic", although associated concepts have evolved with time. More recently study of stochasticity in single cell reveal new mechanisms. We will see here what are the differences that these studies have contributed to bring to the established conceptions of biological variations.

It is difficult to give a unique definition of what is epigenetic because the phenomena described with this word are numerous and its definition has evolved with scientific discoveries. Historically, this term has been introduced by Conrad Waddington in the 1940s to name all the mechanisms involved in the process of expressing the genotype into the phenotype. As from the 1960s, following the discovery of the molecular mechanisms of the lac operon by Jacob and Monod [112], the role previously attributed to epigenetic is captured by genetic through gene regulatory networks and the metaphor of the genetic program. To transform a static code stored in the DNA into a phenotype, it wasn't necessary to invoke an external mechanism anymore; genes could act directly on each other during the execution of the genetic program [152]. Starting from there, epigenetic has become the study of the modulation of the genetic expression by way of chromatin modification (DNA and its protein skeleton). One of the main mechanism is histone methylation. An history of the concept of epigenetic can be found in [110]. The first study on histone methylation is [9]. Histone methylation is a process leading to a change in chromatin spatial organization. Opening or closing of chromatin will allow or not transcription of certain genes. These variations in the access to DNA don't constitute a definite modification of the transcription mechanism, but still have a certain stability in time. One of the main characteristic of the epigenetic modifications of chromatin is their transmission across generation, through mitoses and in certain conditions through meiosis, i.e. within an organism or during sexual reproduction. This research area is very active and has been joined by the study of various mechanisms that share an effect on the modulation of genetic expression. Those mechanisms are for example the inactivation of the X chromosome in mammalian development [162], the process by which one of the two copy of the X chromosome is inactivated in female XX mammals during early embryogenesis; or the role of the maternal proteome in the gamete, i.e. the set of all the proteins already present in the female gamete before fertilization; or study on prion, a protein whose spatial folding is transmissible to other prion protein [146].

Epigenetic modification modify the gene expression level whereas DNA mutations alter the structure of the proteins and hence the qualitative effect resulting from the gene expression. It can however be noticed that quantitative modification can lead to important qualitative effect. For example, during embryonic development, a change in the timing of expression period of a gene can imply an important effect on the final phenotype of an organism [4] [154]. These effects are called developmental heterochronies and the observation of the developmental sequence at every level of organization is a widely used tool to compare embryogenesis among species. Conceptually, epigenetic modifications are close to DNA mutations, although there is a slight difference with respect to heredity, since the first ones can be reversed, or more precisely are less stable in time, whereas the second ones are considered as permanent [182].

Besides, recent results have demonstrated the existence of stochastic phenomena in gene expression. In particular, the work of Michael Elowitz's team has shown with time-lapse visualization of the expression of two fluorescent probes with two different colors associated to the same regulatory sequence, that the level of protein expression is variable in time within a cell and occur in a stochastic manner as shown on figure 4.2. Using these two different fluorescent probes, we can differentiate between the component of variability associated to the expression of each coding region depending on the same regulatory sequence, it is the so-called intrinsic noise. The common component of variability is associated to environmental variation, the so-called extrinsic noise [65]. This stochastic gene expression is called *noise* because it can be associated to random fluctuations but doesn't change qualitatively gene expression. One of the reason for the success of this approach comes from the mathematics used to describe this experiment. The variability observed can be characterized with the Fano Factor η^2_{tot} defined as the ratio of the variance of the measure over the mean: $\eta^2_{tot} \equiv \frac{\sigma^2_x}{\langle x \rangle^2}$. η_{tot} is the noise in gene expression. It can be shown that in the dual reporter experiment, the noise can be decomposed into an intrinsic part associated to the decorrelated variations of the two genes within a cell, i.e. an intrinsic source of noise within the process, η^2_{int}, and into an extrinsic part associated to the correlated variations of the two genes within a cell, i.e. a common external influence on the process, η^2_{ext}. The relation reads $\eta^2_{tot} = \eta^2_{int} + \eta^2_{ext}$ ([203], [65], [168]). These relations assume a static environment and have to be changed to account for a fluctuating environment [98].

4. These temporal variations in the sequence of events during development can also take place at other, more macroscopic, levels. It is the case of the maxillary bone between various groups of salamanders which is morphologically highly different because of variations in the growth periods and morphogenesis periods. These results are presented in [6]

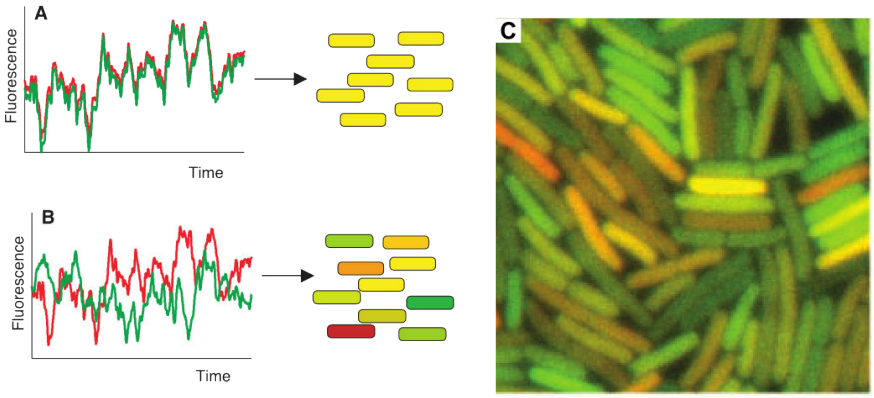


Figure 4.2: This figure is adapted from Elowitz *et al.* 2002 paper on stochastic gene expression [65]. A and B show the basic concept of the experiment: two genes having the same regulatory sequence express two different flurorescent protein. In the case A, both genes have the same expression pattern, in the case B, the two genes have differing gene expression patterns. Figure C shows the result of an experimental measurement for this experiment in a strain of *Escherichia coli* (RP22). The distribution of colors from green to red proves the presence of stochastic gene expression in individual cells

Historically, previous studies had already pointed out variability in gene expression. This the case of the 1957 article "Enzyme induction as a all-or-none phenomenon" ([160]). However, they weren't able to measure gene expression in time to prove a form of stochasticity in the process.

Another work reports an experiment combining stochasticity with inheritance. The article shows that cells can inherit a stochastic switch. A population of yeast *Saccharomyces cerevisiae* change their phenotype in a stochastic way between two semi stable epigenetic states. This stochastic switch is correlated among related individual cells, indicating an inheritance of epigenetic determinants of stochasticity [118].

It can be shown more generally that phenotypic variability in a clonal population of cells can reflect noise in the transcriptome underlying the choice of fate between several metastable states of the gene regulatory network [42], [138].

When considering multicellular organisms, this noise may have dramatic effects on development. This is the case of a mutant line of the nematode *Caenorhabditis elegans* whose capacity to form intestinal cell precursor is highly variable for genetically identical individual in the same environmental conditions. This variability can be explained by the stochasticity of the gene expression and the amplification of small variation by a bimodal gene regulatory network [181]. A bimodal gene regulatory network is a network that

has two stable states. The bimodal gene regulatory network is very sensitive to fluctuations in gene expression. The small number of molecules and the possibility of degrading RNA molecules enable these small variations. Amplified by highly non-linear responses of some gene regulatory network, these small variations, considered as noise, can have dramatic consequences on development and thus on organisms phenotypic traits and fitness [5].

In addition to these stochastic phenomena during gene expression, we can consider the stochastic events occurring during cell division. These events contribute to differentiate cells from each other. The random distribution of the proteome during mitoses is the most studied phenomenon [105]. In the same manner than for stochastic gene expression, if the number of molecules is small and the gene regulatory network is non linear, the effects associated to variations related to cell division can be highly amplified and have consequences on the whole organism.

The set of epigenetic phenomena, the stochastic effects during gene expression or during cell division constitute an important source of variability for the living, different from classical genetic mutations. Their status is also important to understand evolution, even if they are less permanent than mutations [182]. In particular these variations will have a higher effect if they occur early in development for multicellular organisms.

It is possible to consider with Jean-Jacques Kupiec that organisms are built around this intrinsic variation and a principle of natural selection acting within the organism. This is an extension of darwinian evolution to the formation of the individual in an "ontophylogenetic" process [127], [128]. It seems necessary to take into account the specific development of every organism as an essential component of biological variation in order to understand the exploration of possible forms during evolution. The specific development of each organism will be the sequence of singular events which has come along the constitution of an organism [135].

By looking at the mechanisms sources of variability, we notice two important aspects, on one side these mechanisms involve the systemic structure of organisms, in the sense that several processes coming from different mechanisms are mixed in an organism, and on the other side they are involved in their ontogenesis, meaning that the sequence of events occurring during development will be decisive for the constitution of an organism. The processes described are heterogeneous. Can they all be assembled and qualified as random? How should this randomness modeled? Can it be associated to forms that have

5. For a review on epigenetic phenomena during development, see [103]

already been defined in physics? And more generally, what are the impacts of these aspects of variability on models of morphogenesis ?

4.2 Randomness and its formalisms in mathematics and physics

4.2.1 Probability theory

The main problem when we try to model situations involving random events is the problem of prediction. We would like to determine in advance the set of all possible results for a given experiment and the way these possible results can occur, the frequency at which they appear for example. In mathematics, the idea of randomness first appeared with the concept of probability. The probability concept is rooted in the work of Blaise Pascal and was also developed in the "Logique de Port-Royal" [6] in the middle of the 17th century. Pascal's wager is one of the first example of a mathematical reasoning in a context of uncertainty. In this wager, he computes the gain of believing in god, given a probability that god exists, in order to justify to believe in god even if its existence cannot be proved (the whole wage can be found in *Pensées* part III, §233). Mathematization of this concept has taken some time before being considered as a true branch of mathematics. By enunciating a small set of axioms in 1933, Kolmogorov contributed to set probability theory on a firm ground [124]. See [197] for an historical perspective of the mathematical context surrounding the enunciation of the axioms, for an historical account of the emergence of probability see [91]. Axioms of probability theory give constraints on the way to compute probabilities and hence on the tools enabling to measure randomness. They however avoid to define randomness as such. Probabilities always have to be interpreted. Probabilities are sometimes considered as deriving from imperfect knowledge of the experimenter relative to his object of study; they are considered as subjective. Probabilities are some other time considered as an objective property of the object; they are considered as objective. Additional epistemological interpretations of probabilities can be found between these two extreme positions, see for example [84] for a philosophical review of the various meanings of probability. Probabilities don't tell us anything about randomness in itself since the mathematical theory is compatible with these various interpretations.

Kolmogorov's six axioms of probability are enunciated in the following way ([124],

6. The Port-Royal Logic is a very influential book, first published anonymously in 1662 in Paris, it is acknowledged to Antoine Arnauld and Pierre Nicole. It has been a reference in language theory and logic until the 19th century

[197]). Let E denote a set, the set of elementary events, and $\mathscr{F}$ denote a set of subsets of E, which are the random events:

1. $\mathscr{F}$ is a field of sets (it's an algebra over E, i.e. it is closed under the intersection and union of pairs of sets and under complements of individual sets)
2. $\mathscr{F}$ contains E
3. To each set A from $\mathscr{F}$ is assigned a nonnegative real number $P(A)$. This number $P(A)$ is called the probability of the event A
4. The probability of the E is $P(E) = 1$
5. If A and B are disjoint, then $P(A \cup B) = P(A) + P(B)$
6. If $A_1 \supseteq A_2 ...$ is a decreasing sequence of events from $\mathscr{F}$ with $\cap_{n=1}^{\infty} A_n = \emptyset$, then $\lim_{n\to\infty} P(A_n) = 0$

The sixth axiom, also called *axiom of continuity*, is equivalent to countable additivity of P when the first five axioms are given. It means that, for any infinite sequence of disjoints sets $A_1, A_2, ...$ in $\mathscr{F}$, P satisfies the equality:

$$P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$$

Concerning this sixth axiom, it is interesting to read its interpretation by Kolmogorov (quotation translated in [197], [124]):

> Since the new axiom is essential only for infinite fields of probability, it is hardly possible to explain its empirical meaning ... In describing any actual observable random process, we can obtain only finite fields of probability. Infinite fields of probability occur only as idealized models of real random processes. *This understood, we limit ourselves arbitrarily to model that satisfy Axiom VI.* So far this limitation has been found expedient in the most diverse investigations

We understand here that this asymptotic axiom which is needed for the probability calculus has no empirical counterpart. It is an infinity, useful to construct the space of possible of probability theory. It may be interesting to compare it to other forms of infinity used to construct spaces of possible in mathematics and physics [134].

Another way formalization of probability theory states that $(\Omega, \mathscr{F}, P)$ is a measure space with $P(\Omega) = 1$. Formalizing probability theory as a special case of measure theory is the major contribution made by Kolmogorov. Ω is the sample space, $\mathscr{F}$ is the space of events

and P is the probability measure. $\mathscr{F}$ is a σ-algebra; i.e. closed under complement, union of countably many sets and intersection of countably many sets. This is the model of events at play. These rules are those of boolean logic on sets. This logic will constrain the way phenomena can be represented through these probabilities. In particular, it is worth noting that the closure by complement require to have a space of possibles *a priori* that gives the complementary of any events. Moreover, the closure by union of countably many sets requires to assume a modularity of the modeled events, that is to say that any two events can be associated in any order without having any influence on each other. For more detail and interpretations, see [93].

To illustrate these notions, we can describe the example of dice throwing. The set of elementary equally likely events correspond to each face of a dice $\Omega = \{1,2,3,4,5,6\}$. If we consider the experiment consisting in throwing two times consecutively the dice, then we may want to compute the probability that the dice have the same value. Let's denote this event by A in $\mathscr{F}$. All 36 possible sequences of two throws of dice correspond to $\{\{1,1\},\{1,2\},\{1,3\},\{1,4\},\{1,5\},\{1,6\},\{2,1\},\{2,2\}...\}$, they are all the possible variants that are *a priori* possible as outcomes of the experimental conditions. All 6 sequences of identical consecutive throws corresponding to $\{\{1,1\},\{2,2\},\{3,3\},\{4,4\},\{5,5\},\{6,6\}\}$ are included in the set A, therefore, in each of these 6 cases, event A has taken place. If all the sequences of two consecutive throws are considered equally likely, then the probability of A is $P(A) = \frac{6}{36} = \frac{1}{6}$.

Studying the axioms shows clearly that probability theory is a framework to handle random events, but doesn't give any definition of randomness itself. The specific characteristics of the space of possible, of the probability measure, of the space of events, all depend on the modeled situation.

4.2.2 Randomness in algorithmic theories

We have to turn our attention to algorithmic theories of information to find attempts at defining randomness in itself. In the 1960s, the problem of characterizing an infinite random sequence of symbols was open. Kolmogorov proposed a concept of incompressibility for finite sequences of symbols; incompressible sequences are those which can not be produced by a shorter program that the size of the sequence itself using a universal Turing Machine. In 1966, the swedish mathematician Martin Löf proposed a definition of infinite random sequence of symbols by considering all the sequences that possess all conceivable statistical properties of randomness using algorithmic theory to formally define the

notion of a test of randomness [142]. It is ultimately this definition, which is considered the best definition of randomness in mathematics. It has been shown that under certain conditions given by Chaitin, infinite random sequences are exactly those which have all finite incompressible initial segments. This definition of randomness is widely accepted as the fundamental characterization of the notion of a random sequence. It should be noted however, that this definition is based on sets of infinite sequences and statistical tests which are not computable, leaving therefore a large gap between this definition and empirical results.

For example, the infinite binary sequence 0101010101010101... is not random since it can be compressed as a repetition of 01. On the hand, we can consider a Chaitin Omega number which is defined as the halting probability of a universal self-delimiting Turing machine. It is a computable enumerable and (algorithmically) random number. The first 64 bits of a Chaitin Omega have been computed by Calude, Dinneen and Shu in [34]:

0000001000000100000110001000011010001111110010111011101000010000.

Some results bridge the gap between these abstract characterizations of randomness and physical systems in asymptotic cases [78]. We will now focus on physical systems presenting unpredictable behaviors.

4.2.3 Randomness in dynamical systems and ergodic theory

Randomness in dynamical systems The idea of randomness in physics cannot be approached without considering mathematical determinism as it appears in the study of dynamical systems and chaos theory which is its unpredictable counterpart. To understand randomness in dynamical systems, we have to go back to the XIXth century. Pierre Simon Laplace published *A Philosophical Essay on Probabilities* within which he developed his views on determinism. In his view, the world is ruled by a set of causes and effects relationships and only an incomplete knowledge about the state of the world at a given moment could be an obstacle to a perfect knowledge of these relationships and hence to a prediction of its futures states. This perspective is clearly expressed in the following famous sentence where he develops the idea of a demon with an infinite knowledge about the world:

> We may regard the present state of the universe as the effect of its past and the cause of its future. An intellect which at a certain moment would know all forces that set nature in motion, and all positions of all items of which nature is composed, if this intellect were also vast enough to submit these data to analysis, it

> would embrace in a single formula the movements of the greatest bodies of the universe and those of the tiniest atom; for such an intellect nothing would be uncertain and the future just like the past would be present before its eyes.

This view is mathematically supported by the Cauchy-Lipschitz theorem, (also known as the Picard–Lindelöf theorem). For a differential equation

$$\frac{\partial y(t)}{\partial t} = f(t, y(t))$$

with initial condition $y(t_0) = y_0$, if the function f is regular enough (lipschitz continuous) with respect to y and continuous in t, then for some $\epsilon > 0$, there is a unique solution on the interval $[t_0 - \epsilon, t_0 + \epsilon]$. The proof for this theorem involve the use of a fixed point result.

However the works of Henri Poincaré in the late nineteenth century on the stability of the solar system and the three-body problem have weakened this position and laid the foundation to what would become chaos theory. In a seminal article Poincaré shows that in some dynamical system, ruled by a set of non linear differential equation, a small uncertainty on the initial conditions can lead to very large fluctuations on futur behavior [178]. This contradicts Laplace's view since for the latter, an approximate knowledge on the initial conditions should result in an approximation of the same magnitude for the knowledge on the evolution of the system. This means that perturbations could not have an important effect on trajectories. Laplace, like his contemporaries, was confident in linearization methods of differential equations [141]. Poincaré's result is therefore very important because it shows that even if the phenomena we observe are very well defined, it may be impossible to predict the evolution of the system. Perturbations that may seem irrelevant can actually impact the evolution of the system highly significantly. On this point, it is interesting to quote Poincaré from his book Science et méthode, 1908 [177]:

> A very small cause which escapes our notice determines a considerable effect that we cannot fail to see, and then we say that that effect is due to chance. If we knew exactly the laws of nature and the situation of the universe at the initial moment, we could predict exactly the situation of that same universe at a succeeding moment. But, even if it were the case that the natural laws had no longer any for us, we could still only know the initial situation *approximately*. If that enabled us to predict the succeeding situation *with the same approximation*, that is all that we require, and we should say that the phenomenon had been predicted, that it is governed by laws. But it is not always so; it may hap-

> pen that small differences in the initial conditions produce very great ones in the final phenomena. A small error in the former will produce an enormous error in the latter. Prediction becomes impossible and we have the fortuitous phenomenon.

Unpredictability resulting from sensitivity to initial conditions can be viewed as randomness. Even if we knew perfectly the laws of evolution of a system, our inability to know perfectly its state leads to its unpredictability. It is an epistemic kind of randomness stemming from the interstice between the theory and the physical object. For this reason the status of measure becomes dominating, it determines our access to the objects and thus our ability to know the object [14].

A chaotic behavior can be quantified with Lyapunov exponents. They express mathematically the sensitivity to initial conditions and the rate at which two close solutions of dynamical system diverge. Given two trajectories in a phase space, their initial separation is given by $|\delta Z_0|$. The Lyapounov exponent λ is defined as $|\delta Z(t)| \approx e^{\lambda t}|\delta Z_0|$ where $|\delta Z(t)|$ is the difference between the two solutions at time t.

Ergodic theory When considering gas, another sort of problem for the prediction of the system's evolution appears. Following the work of Ludwig Boltzmann, gaz are composed of a high number of atoms and molecules, each of them having its own dynamic. Therefore this high number of elements makes it impossible to follow each particle individually. It is thus assumed that each individual particle has a random motion and collide with other particles in a stochastic way. In order to derive the macroscopic behavior from the behavior of each individual particle, one of the necessary condition was the *ergodic hypothesis* [77]. This assumption is required to prove the equipartition of energy in the kinetic theory of gases. This assumption says that the average of a process over time of a system and the average over the statistical ensemble are the same [20]. It means that each element of the system goes everywhere in the space of possible in a uniform fashion. This statement of uniformity of individual behavior is a form of randomness and involve an asymptotic statement which require the use of infinity. This randomness results from the too high number of elements in interaction.

These two forms of randomness found in dynamical systems theory and in ergodic theory are somehow opposed, they correspond respectively to a too tiny cause or too many causes. However, they both relate to an epistemic notions. They are defined in the frame of classical physics where trajectories of individual particles are assumed. Quantum physics

has a radically different approach of the problem of randomness.

4.2.4 Randomness in quantum mechanics - Quantum mechanics as a generalized probability theory

Quantum physics studies phenomena at the scale of atoms. The formalization of quantum mechanics, which happened in the beginning of the XXth century, is an epistemological revolution. The concept of trajectory itself has been changed compared to classical physics. Schrödinger equation which describes the evolution of a quantum system, no longer governs the path of a body but the trajectory of a probability amplitude (a normalized vector state in a Hilbert space). When a quantum system is measured, the result of the measure is related to the kind of measure applied and the results are predicted in probability. A quantum state is described as a linear combination of possible measure results. Each of the coefficients of this combination gives the probability to obtain the associated possible result according to the Born rule.

It is worth noting that the classical probability structure relying on Kolmogorov's axioms (as presented above) has been replaced by a structure using a normalized state vector in a Hilbert space. The σ-algebra has been replaced by an orthoalgebra within which various experimental contexts, sometimes incompatible, can be combined. It becomes however difficult to use the word object, and it is more common to refer to a process inseparable of the experimental conditions which are also the very conditions of its existence [22]. The traditional configuration where an observer observe a system from the outside is blurred, this situation is part of the formalism. In this field, the theory is build around a probability amplitude. The concept of randomness is at the center, although it seems to be more related to our relations with the object of the study rather than to the properties of the object as such. In any case, the form of randomness used in quantum mechanics can be considered as objective, since the theory cannot overcome it. The violation of Bell inequalities and their experimental measurement proves that no theory with local hidden variables is compatible with the quantum mechanics framework. This theory has a high predictive power. There is however no definition for the randomness in itself or characterization for the mechanisms sources of randomness.

After reviewing these model of randomness in physics and mathematics, we understand that this notion covers very rich and varied fields. Formalisms used can be very different and are not necessarily compatible. To summarize, the main differences rely on the choice of the space of possibles to describe the results of an experiment, the capacity to

separate the object from the context of the experiment, the experimental reproducibility and finally the asymptotic character of the studied system. Non predictability is the underlying concept, the inability to predict uniquely the result of an experiment. Probabilities provide a calculus framework on the results of an experiment with uncertain results. The choice of the rules used in the computations depend on the model of the experiment and thus on the concept of randomness at play.

As presented in the first part of this chapter, the sources of variability in biology are numerous and of various nature. We need a notion of variability, of non predictability and thus of randomness that could place these effects on the same footing. The various time scales at play, the diversity of mechanisms at play, the systemic organization of organisms coupling these different phenomena, makes it difficult to reduce biological variability to one or the other form of randomness met in physics. It is for example possible to have a combination of chaotic and quantum effects [33]. The question remains as to how these couplings between all these phenomena are an obstacle to models of biological variability. Is it possible to decompose in a meaningful way these phenomena or to state the adequate simplifying hypotheses in order to obtain a model of variability? Can we isolate some morphogenetic processes and model them independently to capture one or the other aspect of variability?

4.3 Variability and models in biology

4.3.1 Models and simulation as tools for exploring some dynamics of the living

Here and in the following, we will consider that a simulation is the computation of a specific realization of a model, whereas the model itself is a set of assumptions that aims to describe a target phenomenon - with a lower degree of generality than a theory.

One of the first work describing at the same time a model and a simulation of a morphogenetic process in the living is due to Alan Turing. In his 1952 article « The chemical basis of morphogenesis » [207], he explores the possibilities of a mathematical model of morphogenesis. He proposes a system of differential equations which describe the spatial distribution of chemical elements in an idealized biological tissue through a reaction diffusion mechanism. This non-linear system of equation generates stationary spatial patterns from an initially homogeneous situations (symmetrical with respect to the op-

eration of interchanging the cells [207] p.42) and with an initial local perturbation. Although it is possible to solve this problem analytically and thus to exhibit explicitly some solutions with simplifying hypotheses, Alan Turing recommend in his article the use of computer simulations[7]. These simulations solve numerically the system of differential equations. They exhibit some valid solutions and enable us thus to construct a representation of some solutions. They allow avoiding simplifying assumptions, opening the way to more complicated cases than those presented in the article, for example by integrating mechanical constraints between cells in addition to chemical reactions. This article shows how a mathematical model which is "a simplification and an idealization" sheds light on some mechanisms of the living by describing fundamental characteristics. This model of reaction-diffusion is still at the center of numerous contemporary research [156]. The use of computer simulation enables to describe explicitly solution of models by sacrificing mathematical generality in favor of the intelligibility of specific solutions. Computer simulation is a tool for exploring mathematical models.

This practice carries with it a set of constraints that should be taken care of for the study of biological variation. Indeed, all computer simulations rely on a Turing machine. A Turing machine is a discrete state machine, based on a Laplacian paradigm of identical iteration. As shown by Giuseppe Longo, the perfect iteration that underlies all computer sciences introduces an inevitable difference between a simulation and some dynamics occuring in the physico-chemical continuum like chaotic systems [133]. Turing's 1952 work on morphogenesis, subsequent to his invention of the so-called Turing machine, is a work on the non-linear dynamics defined in the mathematical continuum. These dynamics are sensitive to initial conditions, he uses the words "catastrophic growth" ([207] p. 64). Turing needed the mathematical continuum to make assumptions on morphogenetic mechanisms in the living. This mathematical continuum is then approximated to be simulated on a computer. More precisely, if one considers a chaotic system, the study of its behavior is limited by the ability to reproduce the initial conditions whose access is limited through a necessarily finite measure. Simulation of this kind of system would be identically reproducible if the same parameters are set as inputs. These discussions on the profoundly reproducible status of computer simulation highlight the difference in the epistemological status between the model which can be developed in the mathematical continuum and the simulation which happen in the discrete. They should however not hide the ex-

7. "It might possible be possible, however, to treat a few particular cases in detail with the aid of a digital computer. This method has the advantage that it is not so necessary to make simplifying assumptions as it is when doing a more theoretical type of analysis [207] p.72

ploratory power of computer simulation.

Many examples of computer simulations allow to convince oneself of the relevance of some mechanisms in the morphogenesis of the living. Refinements of Turing's reaction-diffusion system describe well the formation of some patterns in the development of the zebrafish [157], [125]. Cellular Potts model is another approach which tries to describe relations between cells in terms of adhesion within a tissue by using method from statistical physics [88]. By generalizing the Ising model, Potts model describes some cellular displacements. The Ising model is one the simplest model of statistical physics presenting a phase transition. Moreover, the deformation of a 2 dimensional tissue in a 3 dimensional structure has been modeled with a vertex model during *Drosophila* morphogenesis, the tension differences between cells is used to faithfully reproduce empirical measures [165]. At a more molecular scale, the gene regulatory network in the development of the sea urchin has been modeled with dynamical boolean network [171] (it is a network where the nodes can take values from a discrete set of values depending on their interactions). This model enable to reproduce the main steps of the dynamic of the gene regulatory network as it is observed. To take into account epigenetic effects, it is possible to introduce probabilistic interactions between genes [198]. Finally, recent attempts try to model the coupling between gene regulatory network and mechanical interactions between cells during early zebrafish embryogenesis, explaining some phases of development [8].

The question then is to disturbe the mechanisms reproducing empirical phenomena in order to test the predictive power of the model. Given the high complexity of organism, it is to be expected that the mechanisms observed and modeled have a narrow range of applicability and that the mechanisms change highly when considering adjacent situation.

These different models aiming at describing embryogenesis offer sets of reasonable assumptions for some specific phenomena at a given level of organization at a given time of embryogenesis. They allow to describe and explore the possible results of these morphogenetic processes. They give an idea of what may be the repertoire of possible forms determined by the mechanisms. Yet, the question of knowing what is the scope of these models is still open. To what extend can we draw general conclusions from the study of these mechanisms considered separately? What is the legitimacy of the assumptions made in the models? Are these approaches accounting for the aspects of variability described previously?

8. Thèse Julien Delile : « From Cell Behavior to Tissue Deformation: Computational Modeling and Simulation of Early Animal Embryogenesis », under the supervision of Nadine Peyriéras and René Doursat

4.3.2 Living organisms are organized objects involving different levels of organization with heterogeneous dynamics

These various mechanisms are not sufficient to describe the complexity of an organism if they don't allow to integrate the various levels of organization. The experimental results presented in the first section of this chapter show how the coupling between different levels of organization can have an effect on biological variability. It should hence be taken into account in models of variability. Most of the models in developmental biology assume a separability between the phenomena occurring at various scales concerning measurement and modeling. For example Eric Davidson's work consists in modeling the gene regulatory network in the sea urchin embryo [9] from experimental results. This model is simulated with a boolean network. It may show some limitation as it cuts the embryo in parts of tissues that are coarser than the level of the cell and in time period of several hours, much longer than a cell cycle. The problem is that genetic interactions occur within a cell where genetic expression takes place. The simplifying hypothesis which consists in abstracting the level of organization of the unique cell may lead to wrong interpretations of the dynamics of gene expression. On the opposite, one can consider work coupling the dynamics of genetic expression at the cellular level with physical processes at play at the tissue level in the *Drosophila* during mesoderm invagination [216]. The problem which may arise here is the role of previous developmental steps in the phenomenon. For example early events in development may have a later effect on the tissue organization. The simplifying assumption here consists in focusing on a short time period without considering earlier steps in study of causal mechanisms. These two approaches may show some limitations in explaining embryogenesis while considering organisms as a whole.

It is possible to consider an alternative approach in some multi-scale approach such as the phenomenological reconstruction of a complete organism [163] as shown in the first part of this dissertation. This work consists in reconstructing in an automated way the dynamics occurring at different scales. From 3D+time image acquisition, raw data are processed in order to automatically detect the position of the nuclei of each cells in time as well as the shape of the cellular membrane. From these data, it is possible to obtain the spatio-temporal cell lineage. This approach enable the access to the dynamics of each individual cells, as well as the dynamics of the tissue and the whole embryo. The obtained data set is a digital embryo. This reconstruction of a process involving contingent events

9. http://sugp.caltech.edu/endomes/

at various levels enable to construct a duplicate of the empirical space [208] were the phenomena with heterogeneous dynamics are combined, and can be studied as a whole. The process of automated reconstruction introduces some assumptions on the nature of the observed phenomena it is the "thickness" of the digital object. By offering an access to the living, which is digital, quantitative, multi-scale, taking into account the temporal dimension of the embryonic development, this phenomenology reduces the distance between the simulations and the empirical results. It reduces this distance in two ways. First, because it allows to confront quantitatively the simulation and experimental data, which is rarely the case in developmental biology. Secondly, because the underlying assumptions required for the automatic processing of the data are assumptions on the observed objects, and are thus close to a model. The aim differs however from an explicative model since the idea is to reproduce accurately the empirical data and not to test the existence of mechanisms allowing to explain them.

Thanks to this kind of approaches, the question of variability can be asked in a multi-scale and temporal manner. In particular, the phenomenology of variability at various levels of organization can be studied. Variability can be transmitted from one level of organization to the other. Similarly, its transmission through time within the embryo is an interesting question, for example along the cell lineage. By studying the phenomenological reconstruction of five sea urchin embryos during the first hours of development, we could quantify this variability at different scales, from the individual scale to the groups of cells sharing the same type and generation to the whole individual. (see chapter 1 and 2 of this dissertation). We show that the inter individual variation is high for the proliferation dynamics and cellular volume if only the cellular level is considered, although it decreases when considering a coarse-grained level, cells clustered by cell type and generation. Similarly, we show that the variation observed at a given generation is only weakly transmitted from generation to generation at the level of groups of cells. These groups of cells are defined with morphological criteria identifiable for each specimen and seem to be a relavant level of observation to describe embryonic development. Using these two results, the dynamics at the level of the group of cells can be averaged over the cohort into a prototypical representation of development. This prototypic description takes into account intrinsic variability of the groups of cells through the use of a probability distribution and allow to describe the behavior at the scale of the whole embryo by combining the behavior of each of the groups. This is a multi-scale quantitative prototypic representation of embryogenesis. By shedding light on the relations between different levels of organization and by

identifying invariant features among individuals, this kind of object seems to be a good candidate to be at the basis of a multi-scale model of morphogenesis with a certain level of generality and leading thus to a theoretical level of explanation.

The prototypical reconstruction of the sea urchin is a very simple case, on a model organism which is expected to be highly reproducible. The question of integrating the variability stemming from the different levels of organization is open in most of the cases.

4.4 Conclusion

The two main obstacles for modeling possible forms in the living are on one side the systems aspect, which couples various levels of organization and heterogeneous dynamics within a single organism and on the other side the temporal aspect, the fact that all the steps of the ontogenetic trajectory constituted of variable phenomenon at all scales has an impact on the organism. Biological objects are the result of an evolutive history, phylogenetic and of an individual trajectory, constituted of singular events, epigenetic and stochastic. We have shown that the sources of this variability are multiple, precluding *a priori* to reduce these phenomena to one or the other form of randomness already studied in physics. Moreover, probabilities only provide a framework for calculus to describe phenomena and don't say anything on the phenomena themselves. They are useful once the model of variability has been established.

Biological objects are organized, multi-scale, the result of a complexe morphogenetic process. Thus, models focusing on one level of organization and including too strong hypotheses on other levels of organization may not have a sufficient generalizing power in other experimental frameworks and may neglect important phenomena. Phenomenological reconstruction such as presented in the first part of this dissertation allow to access multi-scale dynamics with heterogeneous dynamics. Using these multi-scale dynamics, variability can be measured at all levels of organization. It is possible to identify the relevant levels of organization to describe a model of variability. The multi-scale prototypical representation of the embryonic development of the sea urchin embryo is an example showing that it is possible to study variability in an integrative way, overcoming some limitations described in this article. Models relying on this kind of approach could acquire sufficient generality to reduce the distance between empirical and theoretical biology.

Chapter 5

Variable phenotypic expressivity and incomplete penetrance of the zebrafish mutant line squint^{cz35}

Abstract *This chapter studies a phenomenon of incomplete penetrance in the squint zebrafish mutant line. By measuring the distribution of phenotypes (interocular distance) in progeny of identified pairs of homozygote parents, it is shown that the list of possible phenotypes is discrete with intermediate phenotypes and in unpredictable proportions. The statistical distributions put aside factors related to genetic backgrounds or environment. These complex inheritance patterns suggest a variety of possible relationships between early molecular events and late phenotypes.*

5.1 Introduction

This chapter explores a phenomenon of incomplete penetrance and variable phenotype expression in a zebrafish mutant line. The experimental results and their interpretation presented here also serve as a basis for the next chapter.

The notion of incomplete penetrance describes the situation when individuals carrying identical mutants alleles show either mutant or wild-type phenotypes as it has been observed in the nematode *Caenorhabditis elegans* [102] by Horvitz and Sulston in 1980, and more recently in 2010 by Raj *et al.* [181] . The introduction of this concept can be traced back to Timoféeff-Ressovsky [204] and Romaschoff [187] who noticed it independently in 1925. The notion of variable (phenotypic) expressivity describes the idea that

individuals with identical mutant alleles have variable phenotypes, not necessarily wild type, although it has been found associated to incomplete penetrance[102]. Incomplete penetrance and variable expressivity can sometimes be explained by variation in environmental conditions [218], variation in the genetic background [36] but also by stochastic effects in gene expression [181]. In normal conditions this variability in gene expression may be buffered to ensure proper development. We explore here quantitatively a zebrafish mutant line presenting incomplete penetrance and variable phenotypic expressivity. After having described the phenomenology of the phenotypic variation, we will point out methodological problems raised when trying to explain this observed variability with underlying molecular variability.

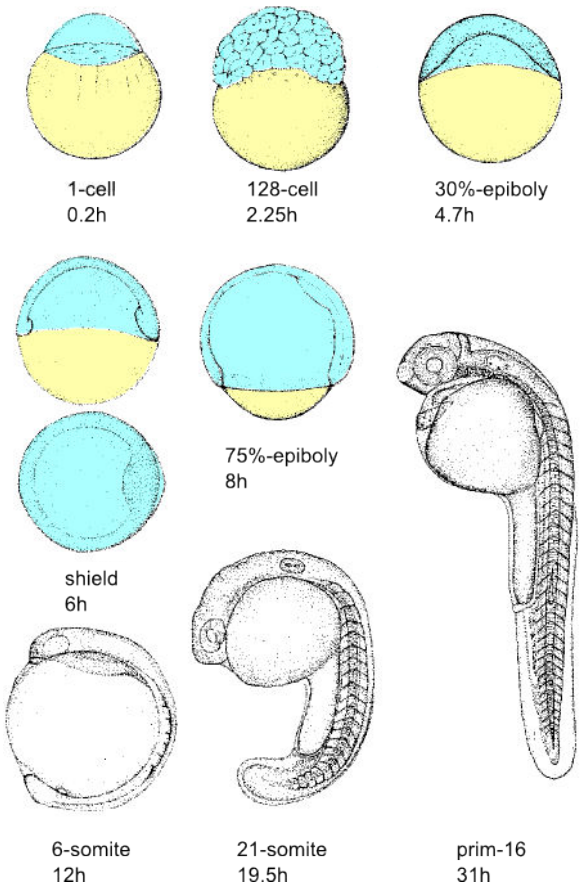


Figure 5.1: Main stages of Zebrafish development. The names of the stages and the corresponding developmental times are indicated below the figures. For the first stages, cells are represented in blue and the yolk sac in yellow. After 12h of development the distinction between the yolk sac and the embryo is clear. Adapted from [122]

The zebrafish is a vertebrate model organism appreciated for its high manipulability

in laboratories, its rapid development and its transparency during early embryogenesis. It is also known for its regenerative abilities. The main stages of its development are shown on figure 5.1. After fertilization, the egg proliferate on the yolk sack before undergoing complex morphogenetic movements such as epiboly, and convergence extension. At 30 hours post fertilization (hpf) the embryo shows an almost final morphology, in particular its head is well formed, and its two eyes are separated by the forebrain. One of the main signaling pathway involved in this morphogenesis is called Nodal. Defects in Nodal related genes such as *squint*, *oep* or *cyclop* lead to dramatic phenotypic defects in the midline such as cyclopia and in the establishment of the embryonic axes [169] [44]. Similar defects are found in human embryos; defects of the midline division of the developing forebrain into cerebral hemisphere with concomitant facial midline defects are known as holoprosencephaly (HPE) [173]. Examples of the consequences of these developmental defects in humans and mice are shown on figure 5.2.

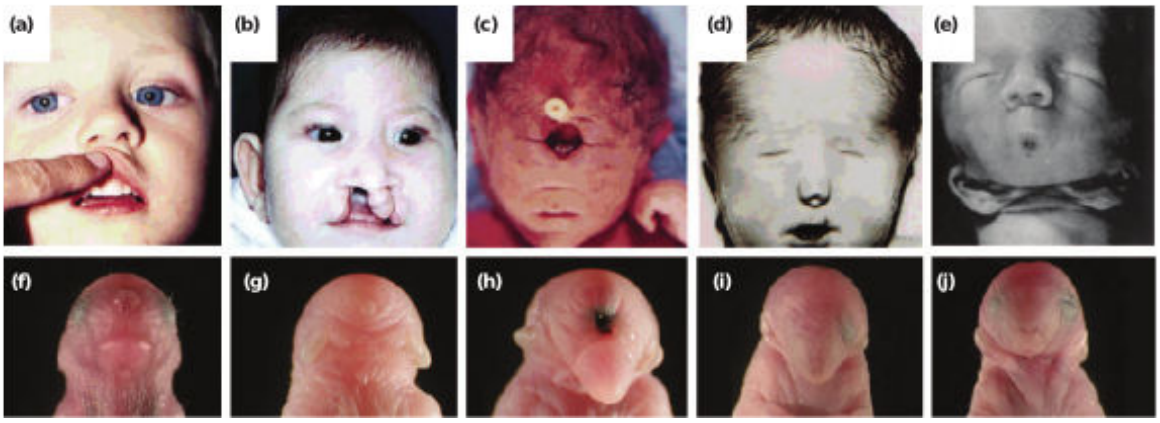
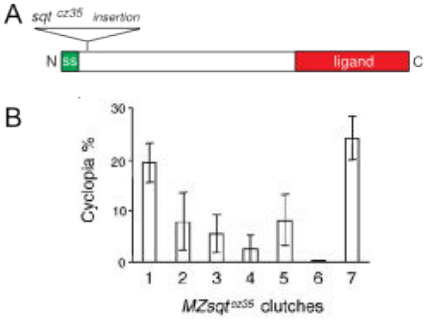


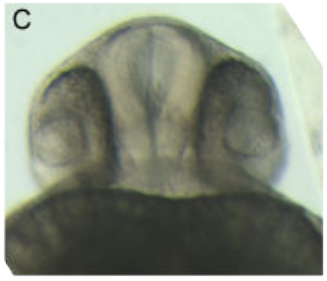
Figure 5.2: Spectrum of craniofacial phenotypes in humans (a-e) and Twsg1-/- mice (f-j). (a) Single central incisor; (b) microcephaly, midface hypoplasia with bilateral cleft lip and palate; (c) cyclopia with proboscis above the fused eye; and (d) hypotelorism and a single nostril. (e) Agnathia with downward displacement of the ears and microstomia. (f) Wild type; (g) severe anterior truncation; (h) cyclopia with proboscis; (i) single nostril with agnathia; and (j) agnathia. Figure and caption from [173]

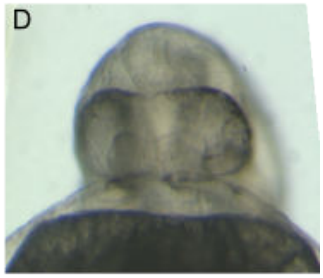
The sqt^{cz35} zebrafish mutant line presents a phenomenon of incomplete and variable phenotypic expressivity. Among other phenotypes, homologous *squint* mutants (*Sqt* -/-) show cyclopia as shown on figure 5.3 D. Some mutants escape the defectuous phenotype and become viable and breeding adults as shown on figure 5.3 C. When two homozygote mutants are crossed, the proportion of embryo expressing the cyclop mutant phenotype in the progeny is highly variable, from 0% to 24 %, statistics obtained in [169] are shown on

figure 5.3 B. The viability of some of the *squint* mutants may be interpreted as a recovery since an earlier phenotype, a delayed formation of the dorsal organizer shows complete penetrance. This incomplete penetrance may result from an otherwise buffered variability in the nodal signaling pathway that is uncovered by the absence of the *squint* protein. To establish a model for the inheritance patterns of variability in phenotypic expressivity of the *squint* mutants, we characterized the phenotypes of the progeny of 10 identified pairs of couples of homozygote mutants during 8 consecutive weeks. This experiment allowed us to study the variations in the distribution of phenotypes through time and among couples.

Previous study on *squint* penetrance shows that the incidence of cyclopia among progeny is dependent on the genetic background of the parents, statistics obtained in [169] are shown on figure 5.3 B. Although crossing of all combination of 8 male and female heterozygote mutants shows that no simple scheme of inheritance can explain the proportion of cyclops in the progeny, statistics obtained in [169] are shown on figure 5.3 E.







E

Transmission of cyclopia susceptibility from sibling sqt$^{cz35/cz35}$ fish, as scored by cyclopia penetrance among progeny

	Female 1	Female 2	Female 3	Female 4	Female 5	Female 6	Female 7	Female 8
Male 1	Med (8.5%)			Med (5.6%)				Med (3.2%)
Male 2		Low (0%)					Low (0.6%)	
Male 3			High (21%)					
Male 4	High (20%)			Med (3.4%)				Low (0%)
Male 5	Med (10%)			Low (1.5%)	Med (4.3%)			Med (3.7%)
Male 6			Low (0%)			Low (0%)		
Male 7		Low (0%)					High (24%)	
Male 8	Med (5.2%)			Low (0%)	Med (2.4%)			Med (2%)

Figure 5.3: A. Figure showing the insert of 1.9kb in *squint* gene (mutant *squint*^{cz35}) [71] B. Incidence of cyclopia in crossing of identified couple of mutants, from [169] C. Wild type phenotype in a homozygote *squint*^{cz35} mutant D. Cyclopia phenotype in a homozygote *squint*^{cz35} mutant E. Proportion of cyclopia in random crossings of homozygous mutants rule out simple models of inheritance, results from [169]

5.2 Materials and methods

Mutant line *squint*^{cz35} - 10 pairs identified by PCR The *squint*^{cz35} mutation consists in an insert of 1.9 kb in the *squint* gene which is a ligand of the Nodal signaling pathway leading to a truncated expression of the *squint* gene. To perform our experiment we identified 10 pairs of *squint* mutant sqt^{cz35} by fin-clipping male and female *squint* adults and genotyping by PCR with specific primers [71]: reverse primers detecting the presence (5'-ATATAAAATCAGTACAACCGCCCG-3') or absence (5'-GCCAGCTGCTCGCATTTTATTCC-3') of the insertion were used with a forward primer (5'-GAGCTTTATTTCAATAACTGCGTG-3') present in wild-type and sqt^{cz35} alleles. See figure 5.3 A.

Fixation at 30 hpf Controlled fertilization was performed by using aquariums where pairs of fishes were maintained isolated. The night before the egg laying a transparent wall separates male and female, in the morning the wall is removed and freshly layed eggs are controlled every 15 minutes. Fertilized eggs were then stored in an incubator at 28°C. Environmental conditions were similar during the 8 weeks of the experiment. After 30 hours post fertilization (30hpf), embryos were fixed with paraformaldehyde (PFA) to stop the development and maintain the morphology comparable within the assemblies of embryos.

Photographs and measures - Fiji Once the embryos were fixed, they could be photographed to provide a quantitative measure of their morphology. They were dechorionated to be better manipulated and have a better resolution. They were positioned with the head in the good orientation, by using rails in agarose. Photographs were taken with a binocular microscope in bright light. Measures of interocular distance have been manually extracted with FIJI software as shown on figure 5.4. A total of 928 measures is obtained.

5.3 Results

Protocol - Number of embryos in each egg clutch The experimental protocol is summarized on the figure 5.4. It consisted in controlling the spawning of identified pairs (male and female breeding couple) of homozygote squint mutants during 8 consecutive weeks. Pictures of each embryo of the progeny were then taken at 30 hours post fertilization. They were used to measure the interocular distance.

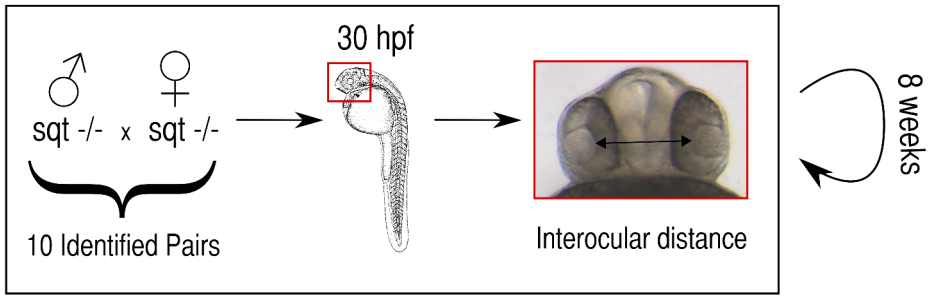


Figure 5.4: Protocol. Every week during 8 weeks, all embryo of the progeny of identified couples of homozygote squint mutants were photographed at 30 hpf in controlled fertilized condition.

The number of embryo in the progeny of each pairs was very variable through time although the conditions were maintained identical from one embryo to the other and through time. Table 5.1 shows the distribution of the number of embryos through time for the different pairs considered. We see that pair 1 and pair 5 generated embryos on the most regular basis, they were breeding 6 times among the 8 weeks of the experiment. Pair 3 and pair 4 generated embryos half of the time of the experiment, i.e. 4 times. Pair 7 and pair 8 generated embryos two times with a low number of embryos (max 16 embryos). Pair 6 and pair 9 generated embryos only once and it was less than 8 embryos. Pair 2 and pair 10 were unable to reproduce during the experiment.

***Squint* penetrance** The measures performed on the embryos enable to compute the number of cyclops during each week of the experiment, they are reported in the following table 5.2 and correspond to the penetrance of the *squint* phenotype:

The values of penetrance shown in the table 5.2 are very variable among the different couples (from 0% to 75%) and throughout the experiment for an identified pair, for example for couple 5 from 3.3% to 32.4%. In contrast to the results obtained in [169], these results suggest that the variations in the penetrance rate of cyclopia in Sqt-/- embryos may be determined by other factors than the genetic background.

Couple 5 We first consider a single pair throughout the experiment. The pair number 5 has the highest number of embryo (283) during the 8 weeks of the experiment providing therefore the most significant results. The distribution of interocular distance are reported on figure 5.5, five histograms are shown corresponding to weeks 1, 3, 4, 6, 8. The experiment performed during week 2 could not be exploited for measurement of the interocular

Week	27 Jan. 1	02 Feb. 2	09 Feb. 3	21 Feb. 4	28 Feb. 5	06 March 6	13 March 7	20 March 8
Couple 1	5	-	-	55	56	44	33	42
Couple 2	-	-	-	-	-	-	-	-
Couple 3	53	-	-	7	-	-	64	47
Couple 4	-	-	-	-	81	18	31	118
Couple 5	40	28	58	37	-	87	-	61
Couple 6	-	-	-	-	-	-	-	4
Couple 7	-	-	-	-	2	-	-	8
Couple 8	-	-	-	1	16	-	-	-
Couple 9	-	-	-	-	8	-	-	-
Couple 10	-	-	-	-	-	-	-	-

Table 5.1: Number of eggs for each week of the experiment. Each line corresponds to an identified *sqt*-/- pair

distance.

When considering week 1, i.e. the first histogram of figure 5.5, we can observe a proportion of 17.5% of cyclops embryos. In addition, we observe intermediate phenotypes between the wild type phenotype and the cyclop phenotypes. These intermediate phenotypes are characterized by inter ocular distances that are intermediate between 0 and 140 μm. Two intermediate phenotypic classes are identified during week 1, they correspond to interocular distances equal to 45 μm and 90 μm. The phenotypic class corresponding to the lowest interocular distance is also represented in the progeny obtained in week 3 and 8. The other phenotypic class corresponding to interocular distance around 140 μm is represented in all of the considered histograms. Altogether, we see that the phenotypes are not continuously distributed along the axis corresponding to interocular distances.

This description of the phenotypic classes rely on manual classification. It could be interesting in the future to use automatic clustering algorithm on the phenotypic value to classify mutants. However, in this case, given the high disproportion between embryos corresponding to intermediate values of the phenotype (very few embryo each week, less than 10) and embryo corresponding to wild type phenotypes (up to 80 % of the embryos),

Week	1	2	3	4	5	6	7	8
Couple 1	0	-	-	10.9%	12.5%	13.6%	15.1%	17%
Couple 2	-	-	-	-	-	-	-	-
Couple 3	5.7%	-	-	28.6%	-	-	1.6%	4.3%
Couple 4	-	-	-	-	6.2 %	0%	6.4%	1.7%
Couple 5	17.5%	-	22.4%	32.4%	-	31.03%	-	3.3%
Couple 6	-	-	-	-	-	-	-	75%
Couple 7	-	-	-	-	50%	-	-	37.5%
Couple 8	-	-	-	0%	23.5%	-	-	-
Couple 9	-	-	-	-	0%	-	-	-
Couple 10	-	-	-	-	-	-	-	-

Table 5.2: Proportion of cyclops for each week of the experiment in % of the number of embryo generated during the considered week. Each line corresponds to an identified *sqt*-/- pair. Photorgraphs obtained during week 2 could not be exploited

lead to inaccurate results.

Proportions when merging the measures of all weeks The histograms are represented on figure 5.6, each histogram corresponds to the merging of the phenotypic distributions in the progeny of one identified couple. We see that the shape of the distribution are highly variable from one couple to the other. This result seems to rule out a dependency of the phenotype distribution only on the gene *squint*. Since all the couples are heterozygote mutants, such a unique dependency would be witnessed by a similar phenotypic distribution. For each of the couples, the phenotypes are not continuously distributed from cyclopy to two well formed eyes. The phenotypic classes are similar from one couple to the other, although some intermediate phenotypic classes are not represented in couple 7, 8 and 9. The proportions are very variable from one couple to the other.

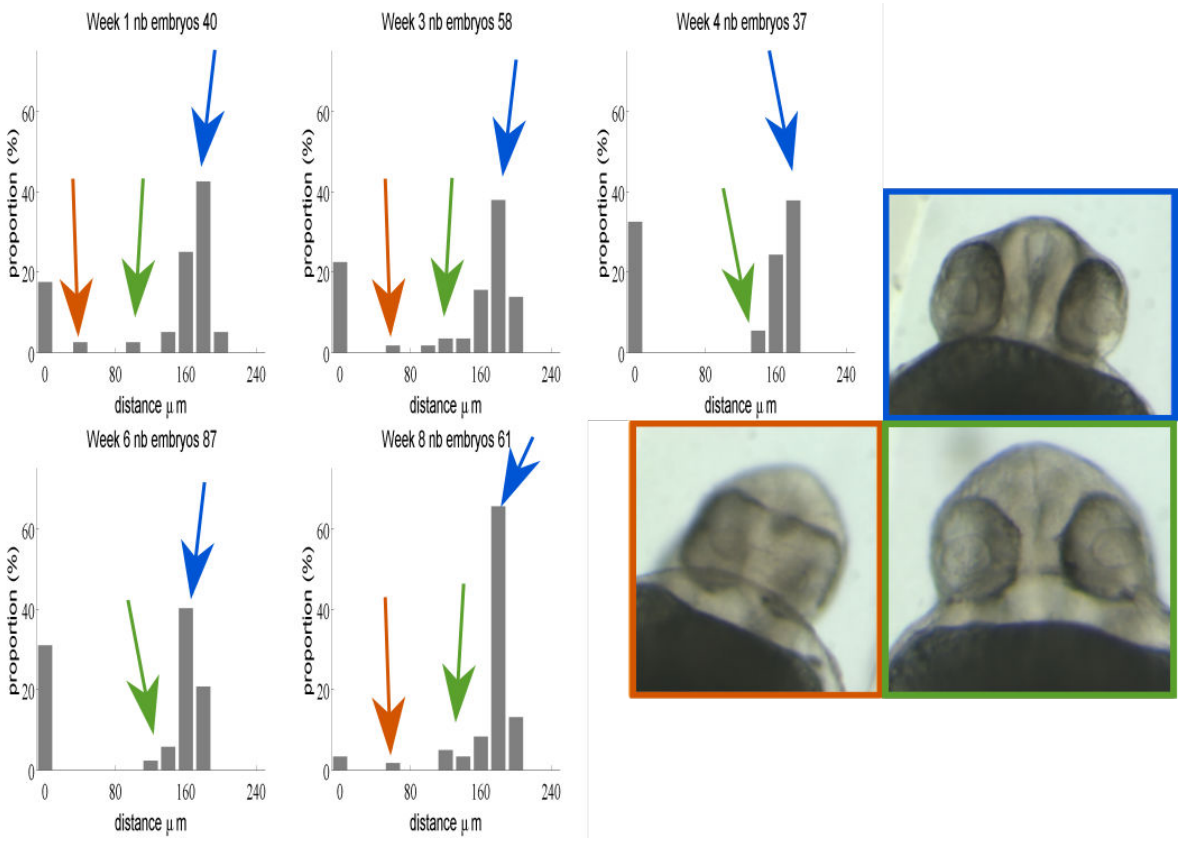


Figure 5.5: Proportion of Cyclops and intermediate phenotypes for the pair number 5. Histograms correspond to week 1, 3, 4, 6, 8 of the experiment and show the distribution of interocular distance (id). The proportion of Cyclops varies from 3.3% to 32.4%. Mutants with wild type phenotypes are observed (blue id $\geq 140\mu m$). Two intermediate phenotypes can be observed (orange $0 < \text{id} < 80\mu m$ and green $80\mu m \leq \text{id} < 140\mu m$).

5.4 Discussion

Overall, this experiment on the distribution of phenotypes in the progeny of homozygote *squint* mutant shows variable phenotypic expression in unpredictable proportions and with a possible incomplete list of accessible phenotypes. This last point is underlied by the fact that the list of intermediate phenotypes is not the same among the various couples. Therefore it can be supposed that other couples of mutants present other intermediate phenotypes.

The observed variation in the phenotypes could result from a cryptic variation of β-catenin, involved in the nodal signalling pathway, revealed by the absence of squint. This maternal protein has a redundant role compared to the protein squint [195].

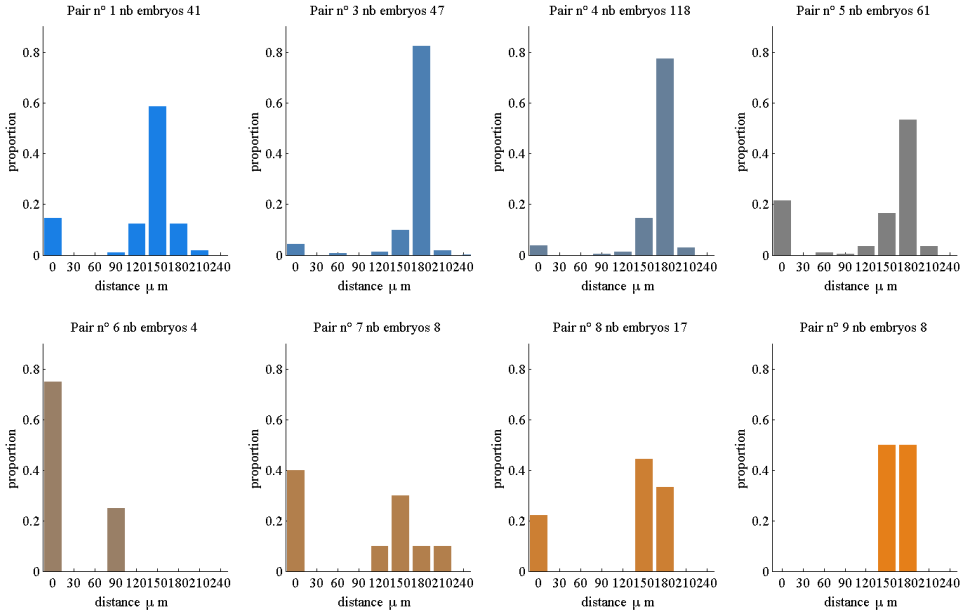


Figure 5.6: Each histogram represents the distribution of every egg clutches merged. Results are shown for each 8 pairs having been able to have a progeny (pairs: 1, 3, 4, 5, 6, 7, 8, 9).

To find correlation, and possible causal links, between early phenotypes such as the spatial distribution of β-catenin and late phenotypes such as the interocular distance requires to measure the spatial distribution of β-catenin. This measure should lead to various phenotypic classes. The question that will arise then is how to relate these statistical distributions?

Figure 5.7 shows the hypothetical causal links between early molecular events, early phenotypes, and late phenotypes. The case (a) corresponds to a direct link between an early molecular event and a late phenotype, for example the absence of *oep* lead to cyclopia with complete penetrance. The case (b) corresponds to situations when the symmetry breaking hasn't occurred yet, for example in a bimodal gene network two metastable states are possibles. The case (c) corresponds to convergence during development, when for example two embryos with different phenotypes converge to same phenotype. The case (d) corresponds to the case when an early phenotype is observed but doesn't have a late counter part, because this causal link has been observed only once and only at the molecular level for example. The case (e) is the same idea as (d) but when only the late phenotype has been observed. Finally it can be assumed that some causal links may oc-

cur but have not been observed yet.

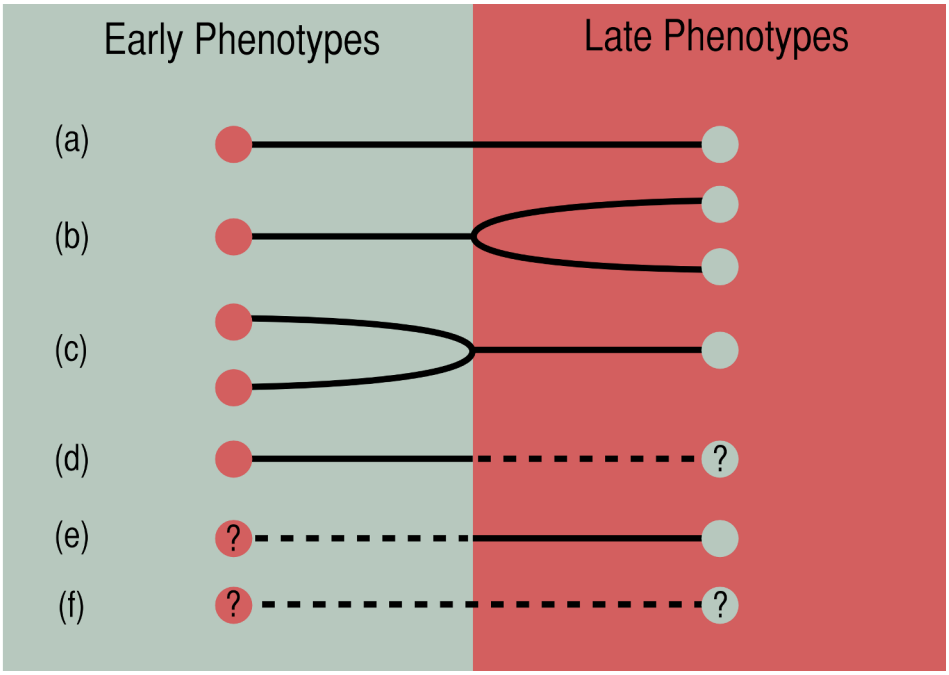


Figure 5.7: Possible causal chains between late phenotypes (eg. interocular distance) and early phenotypes (eg. spatial distribution of β-catenin). Each experiment trying to link early phenotypes with late ones should deal with these causal pathways a) simplest causal chain where an early phenotype can be linked uniquely to the late phenotype b) symmetry breakings occur lately in development c) two early different phenotypes are buffered into one late phenotype d) the late phenotype counterpart of an early phenotype has not been observed e) the early counterpart of a late phenotype has not been observed f) early and late phenotypes have not been observed but are possible

The presence of symmetry breakings (b), convergence (c), non-exhaustivity of the measurement (d,e) and possible new causal relationships are important methodological obstacles for statistical inference. What does these methodological obstacles imply for biology? Can it be assumed that this situation is the norm in biology instead of a special case of inheritance and determination? Thus, the choice of the space of possibles is a central problem, it is the central problem of the next chapter.

Chapter 6

Biological diversity and quantum mechanics formalism

Abstract *We present in this article a mathematical approach to account for diversity in biology using the framework of quantum mechanics. To construct our characterization, we rely on two principles. The first one is that the list of observables in biology is not predictable and their possible value are uncertain. The second one is that relations between various events during establishment of a phenotype are related in a complex way. These two statements justify the successful use of a quantum like framework to account for statistics. An illustration is given with the study of the squint mutant line of the zebrafish. A formal analog of entanglement is identified for Mendel's idealized scheme of inheritance. Finally, clarification of the various levels of biological variation is enabled with this formal analogy.*

6.1 Introduction

The two principles by which Darwin proposed to explain the diversity of biological forms are descent with modification and natural selection [49]. The sources of variation underlying descent with modification are multiple, from combinatorial genetic lottery, to mutations of the DNA, to epigenetic effects such as DNA methylation or histone modification, to stochastic effects during gene expression, [104], [181]. However, the theoretical status of variability is still unclear.

Following a line of thoughts initiated by Longo, Bailly and Montévil [15], [136], [135] individual organisms can be considered as resulting from a specific onto- and phylogenetic

trajectory involving contingent events. They are thus historical entities. This perspective has strong implications for the theoretical status of biological objects. In particular, the space of possible is not completely predictable, radically new observables are likely to emerge in the course of evolution, for example a new organ or a new function for an already existing organ [137].

In this article, we propose to analyze this unpredictability in the light of the quantum mechanics formalism in order to characterize the various levels of uncertainty met in biology. Previous approaches have brought closer biology and quantum mechanics. Schrödinger in his book "What is life" [196], Rosen in a paper discussed quantum information as an approach for describing genetics [188]. More recently, Khrennikov and colleagues have extended the use of quantum mechanics formalism to biological situations [17], with a fine analysis of epigenetic evolution [12]. The mathematical structure of quantum mechanics contains several epistemological characteristics of quantum mechanics, such as the dependency of the empirical results to the experimental context, and more particularly the measure instruments [23], [21]. Other disciplines sharing similar epistemological constraints may benefit from the use of the quantum formalism. The emerging field of quantum interaction provides examples of successful applications in cognitive sciences, decision theory, computer sciences.

6.2 Variability in biology, emergence of new phenotypes

Any biological inquiry begins by studying a particular observable. An observable corresponds to a measurable characteristic that can provide knowledge about an organism. It can be a phenotypic trait, a genotype, a function. Drawing a direct analogy with quantum mechanics, a biological observable will be represented as an hermitian operator in a hilbert space. The eigenvectors of the operator being the possible values of the observable. For example, the observable can correspond to the “wrinkled” or “smooth” character of a pea as in Mendel’s experiments. The observable can also be the presence or not of eyes, such as in the species of fish *Astyanax mexicanus* which comes in at least two forms, with or without eyes (for the blind cave form) [113].

Following the classical framework of quantum mechanics [47], in a population where only two possible phenotypes coexist: phenotype 1 and 2 can be symbolized by the two vectors $|p_1\rangle$ and $|p_2\rangle$, the observable as A. A population of similar organisms in homogeneous environment is described by a normalized state vector $\psi = a_1.|p_1\rangle + a_2.|p_2\rangle$ with

parameters a_1 and a_2 in $\mathbb{R}$ such that $a_1^2 + a_2^2 = 1$. Following the Born rule, the a_i coefficient corresponds to a probability amplitude and a_i^2 is the probability to measure the value $|p_i\rangle$ of the observable A. In practice this probability is obtained empirically from the observed frequency f_i. It captures a first level of uncertainty about the result of biological experiments. The use of a probability amplitude leads to an additional degree of freedom in the choice of the sign: $a_i = -+\sqrt{f_i}$. We notice here that we restrict our use of probability amplitude to real value and not complex values as in quantum mechanics to avoid unnecessary degrees of freedom. The choice of the observable depends on the observer, its possible values and their empirical frequencies depend on the experimental system, this is the main difference between A, $|p_i\rangle$ and a_i.

For example, in Mendel's experiment, the observable A corresponds to the outer appearance of peas, its possible values are "wrinkled" (symbolized by the vector $|w\rangle$) or "smooth" (symbolized by the vector $|s\rangle$) and their probability f_i can be measured empirically, leading to the probability amplitude $a_i = -+\sqrt{f_i}$ and the descriptor of a population of peas as a normalized state vector $\psi = a_1 |w\rangle + a_2 |s\rangle$. Similarly, the fishes of the species *Astyanax mexicanus* can be characterized with an observable A detecting the presence of the eyes, its possible values are with eyes ($|\text{eyes}\rangle$) or eyeless ($|\text{eyeless}\rangle$) and as in the previous example, the distribution of phenotypes in a population of similar organisms can be described with the normalized state vector $\psi = a_1.|\text{eyes}\rangle + a_2.|\text{eyeless}\rangle$. If a finer approach is necessary, the observable can correspond to size of the eye instead of a binary detection of presence.

To summarize, we can formalize the distribution of phenotypes in a population of similar organisms in the following way:

- an observable A which is associated to a measurement procedure
- a set of possible phenotypes, $\{|p_i\rangle\}_i$, corresponding to the possible values of the observable *actually* observed, they form the basis of eigenvectors of A
- a set of probability amplitude, $\{a_i\}_i$, associated to the empirical frequency $f_i = a_i^2$ of each phenotype $|p_i\rangle$. To reduce the number of degree of freedom, we limit them to real values.
- a state vector $\psi = \sum_i a_i |p_i\rangle$ describing the distribution of phenotypes in the observed population

By increasing the size of the population of similar organisms under study it is possible to obtain new unexpected result for the observable, because the exploration of diversity is still at work (even in homogeneous environment). This is where historicity of

individual organisms plays a major role. Indeed, an ontogenetic trajectory is a sequence of highly contingent events, for example stochastic gene expression, asymmetric cell division, epigenetic effects. These contingent events are constitutive of the final phenotype of an organism and thus play a role in the diversity observed. The list of possible phenotypes may not be completely predictable because of yet unobserved contingent event. If a new unexpected phenotype is observed, it is necessary to update the space of possible. Say, we know that the space of possible H is well described with the list of possible phenotypes $(|p_1\rangle, |p_2\rangle, ..., |p_n\rangle)$ for observable A. The new unexpected phenotype will be denoted $|p_{n+1}\rangle$. Then the new space of possible H_{new} will be defined by the basis $(|p_1\rangle, |p_2\rangle, ..., |p_n\rangle) \times (|p_{n+1}\rangle) = (|p_1\rangle, |p_2\rangle, ..., |p_n\rangle, |p_{n+1}\rangle)$ and the corresponding observable A will add $|p_{n+1}\rangle$ to the set of its eigenvectors. While not changing fundamentally the observable, this approach allows to capture a second level of uncertainty on the set of possible values of an observable. The dimension of the space of possible related to the observable is increased in a linear way. The fact that the space of possible is defined as a vector space allows to increment its set of possible values with a cartesian product. The use of a normalized state vector allows to avoid renormalization techniques when considering previous experimental results. Using this approach shows that the prediction of the possible results of a biological experiment is related to the knowledge of the observer and the context of previously observed phenotypes. We presuppose the unability to predict completely the list of possible phenotypes for a given observable.

It is possible, moreover, that during the course of evolution, a radically new observable emerges which doesn't fit in the space of possible of a previously defined observable, or, stated otherwise, that cannot be described with the same act of measuring. This is Goldsmith's hypothesis of *hopeful monster*. This level of uncertainty is captured by defining a new observable B. It is a much more stronger form of uncertainty than a probability since it requires to define a new space of possibles.

There is therefore a strong difference between an unpredictability in terms of probability of having a given phenotype, at the level of the set possible phenotypes for a given observable and the unpredictability of the observable itself. To illustrate these ideas, we present unpublished experimental data.

6.2.1 *Squint* experiment - an incomplete list of phenotypes

The *squint*^{cz35} mutant line of the zebrafish presents a phenomenon of incomplete penetrance and variable phenotypic expressivity ([170] and chapter 5 of this dissertation).

Crossing homozygote mutants lead to progeny with various phenotypes, from complete cyclopia to two well formed eyes in unpredictable proportions. To quantify the variation in phenotypes we conceived a protocol characterizing embryos in the most reproducible way. Figure 5.4 presents the design of the experiment, 10 couples of identified homozygote mutant (Sqt $-/-$) were crossed during 8 consecutive weeks and progeny grown in controlled conditions. Each week, pictures of each embryo of the progeny were taken at 30 hours post fertilization. These pictures were used to measure an interocular distance.

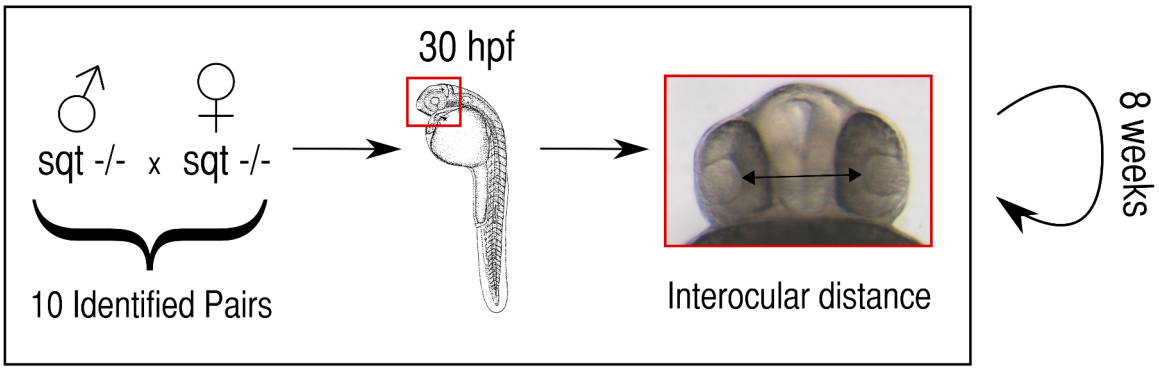


Figure 6.1: Protocol. Every week during 8 weeks, progeny of homozygote squint mutants (Sqt $-/-$) couples were systematically photographed at 30 hours post fertilization (hpf) in controlled fertilized condition to measure interocular distance

The number of embryo in the progeny of each pair was very variable through time even if the conditions were kept identical from one pair of embryo to the other. Table 5.1 shows the number of embryos through time for the different pairs considered. Notably, we can see that couple 2 and 10 were unable to produce embryo. Couple 6, 7, 8, 9 produced very few embryos during the course of the experiment (a total of 4, 10, 17, 8 respectively). Couple 5 provided embryos on a regular basis.

Histograms corresponding to distributions of measures during 5 weeks are shown on figure 6.2 for couple 5 (pictures of week 2 were unfortunately unusable, reducing the number of samples). The first observation is that the proportion of cyclops is variable through time (embryo with null interocular distance) from 2% to 32% of the progeny. The second observation is that the distribution of phenotypes doesn't form a continuum between cyclops and two eyed phenotypes. We then observe that embryos with non-zero interocular distance can be classified according to a reduced set of phenotypic classes. In the course of the experiment, 3 phenotypic classes are identified, associated to the colors orange, green and blue on figure 6.2. These classes are associated to specific morphologies as shown on the photographs presented on the figure 6.2, and to specific values of inte-

	27 Jan.	02 Feb.	09 Feb.	21 Feb.	28 Feb.	06 March	13 March	20 March
Couple 1	5	-	-	55	56	44	33	42
Couple 2	-	-	-	-	-	-	-	-
Couple 3	53	-	-	7	-	-	64	47
Couple 4	-	-	-	-	81	18	31	118
Couple 5	40	28	58	37	-	87	-	61
Couple 6	-	-	-	-	-	-	-	4
Couple 7	-	-	-	-	2	-	-	8
Couple 8	-	-	-	1	16	-	-	-
Couple 9	-	-	-	-	8	-	-	-
Couple 10	-	-	-	-	-	-	-	-

Table 6.1: Number of fertilized eggs during the 8 week of the experiment for each couple of identified homozygote *squint* mutant

rocular distance id which are identified manually. Cyclops corresponds to id = 0, orange phenotypes to $0 < \text{id} < 80\mu m$, green phenotypes to $80\mu m \leq \text{id} < 140\mu m$ and blue phenotypes to $\text{id} \geq 140\mu m$ We obtain the following empirical frequencies and state vectors:

- Week 1 - cyclops : 17,5%, orange : 2,5%, green : 5%, blue : 75%
  $\psi_1 = 0,42.ket\text{cyclops} + 0,16.ket\text{orange} + 0,2.ket\text{green} + 0,87.ket\text{blue}$
- Week 3 - cyclops : 22,414%, orange : 1,7241 %, green : 8,6207%, blue : 67,241%
  $\psi_3 = 0,47.ket\text{cyclops} + 0,13.ket\text{orange} + 0,29.ket\text{green} + 0,82.ket\text{blue}$
- Week 4 - cyclops : 32,432 %, orange : 0 %, green : 2,7027%, blue : 64,865%
  $\psi_4 = 0.57.ket\text{cyclops} + 0.ket\text{orange} + 0.16.ket\text{green} + 0.80.ket\text{blue}$
- Week 6 - cyclops : 31,034%, orange : 0%, green : 2,2989%, blue : 66,667%
  $\psi_6 = 0,56.ket\text{cyclops} + 0.ket\text{orange} + 0,15.ket\text{green} + 0,82.ket\text{blue}$
- Week 8 - cyclops : 3,2787 %, orange : 1,6393 %, green : 4,918%, blue : 90,164%
  $\psi_8 = 0,18.ket\text{cyclops} + 0,13.ket\text{orange} + 0,22.ket\text{green} + 0,95.ket\text{blue}$

During the first week of observation, the 4 phenotypic classes were already observed. However, if the measures had begun at week 4, the phenotype *orange* wouldn't have been observed until week 8. The space of possible phenotypes would then have been the Hilbert space of dimension 3, $H = (ket\text{cyclops}, ket\text{green}, ket\text{blue})$ for week 4, 5, 6, 7. At week 8 the

new phenotype *orange* would have emerged leading to increment the space of possible with the new phenotype: $H_{new} = H \times ket\text{orange} = (ket\text{cyclops}, ket\text{green}, ket\text{blue}, ket\text{orange})$.

This example shows that it is possible to observe the emergence of new phenotypes for a given observable, here the interocular distance. It is not impossible that by continuing this experiment several more weeks, other new unexpected phenotypes would emerge. The structure of the space of possible as a hilbert space is necessary to perform a cartesian product with the new direction defined by a new phenotype. The use of state vector allow to define probabilities in a space whose dimensions can be incremented while keeping its normalization.

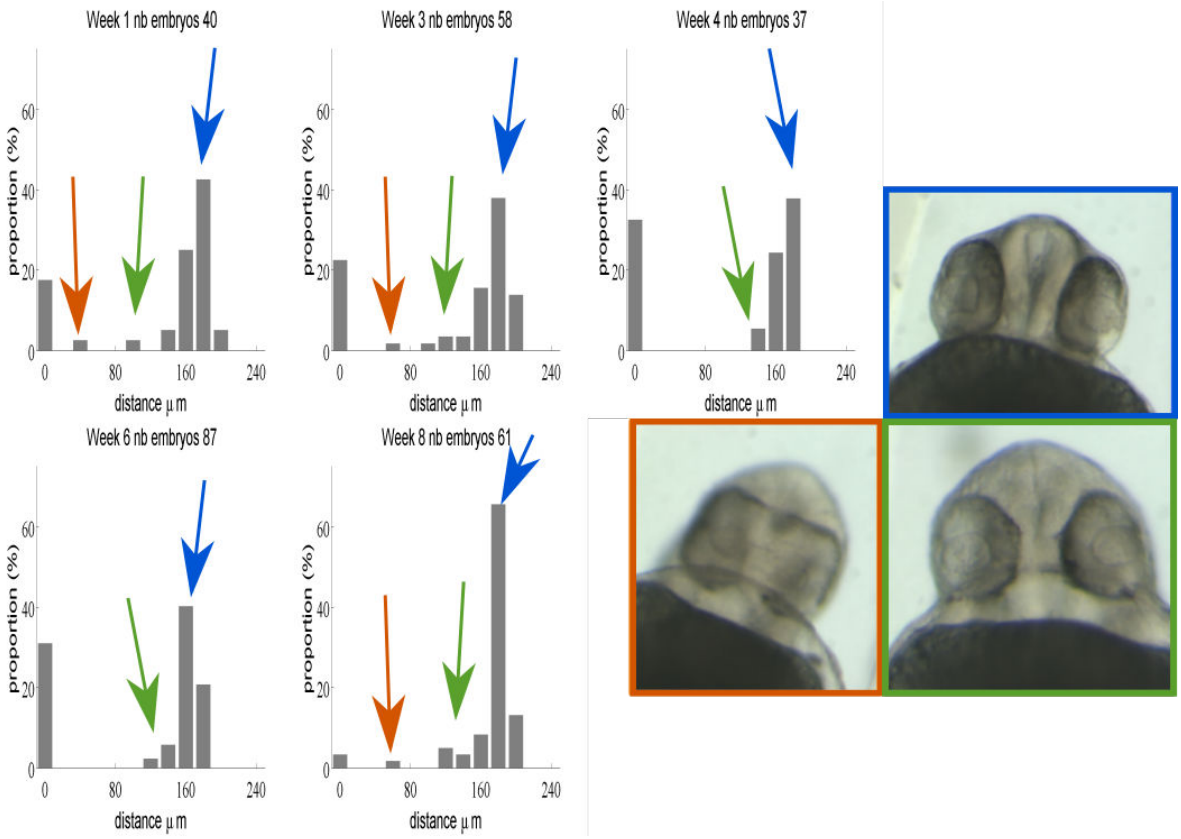


Figure 6.2: Distributions of interocular distance (id) among progeny for the pair number 5 during 5 weeks. The proportion of cyclops varies from 3.3% to 32% (id = $0\mu m$). Mutants with wild type phenotypes are observed (blue id $\geq 140\mu m$). Two intermediate phenotypes can be observed (orange $0 <$ id $< 80\mu m$ and green $80\mu m \leq$ id $< 140\mu m$).

This example, is a good example of an experiment where, for a given observable, the list of possible phenotypes is not completely predictable and where the probabilities for each possible phenotypes seem unpredictable. Finding causal links for the constitution

of the various phenotypes could allow to predict with certainty the value of the observable. However, these causal links can only be obtained by correlating the values of several observables.

6.3 Correlating several observables

Given populations of similar organisms, we may be interested in correlating various observables A, B. For example correlating phenotypes with genotypes or phenotypes with various environments. Most of the time, causal links are difficult to identify. Organisms involve multiple levels of organization where the causal relations can be top-down and bottom-up, for example for the molecular level to the cellular level through gene expression, or from the tissue level to the cellular level through mechanical constraints. In the work of Raj et al. [181], the phenotype under study is the intestinal cell fate during embryogenesis of the mutant nematode *Caenorhabditis elegans*, and it is correlated to stochastic gene expression in an underlying bimodal gene regulatory network.

If considered independently, two observables gives two state vectors $|\psi\rangle_A$ and $|\psi\rangle_B$, each of them being a distribution of phenotypes in a population of similar organisms in homogeneous environment. They are linear combination of their eigenvectors. These eigenvectors define two spaces of possible, H_A and H_B having for basis $(|a_1\rangle, |a_2\rangle, ..., |a_n\rangle)$ and$(|b_1\rangle, |b_2\rangle, ..., |b_m\rangle)$ respectively.

If considered jointly the space of possible becomes the tensor product of the two individual spaces of possible $H_A \otimes H_B$ since all possible combinations of observables values can be expected. The result of an experiment correlating two observables can be modeled as a state vector $|\psi\rangle$ in the product space. It can be written as a combination of the elements of the basis of $H_A \otimes H_B$.

$$|\psi\rangle = \sum_{i,j} c_{i,j}.|a_i\rangle \otimes |b_j\rangle,\ (i,j) \in \{1,...,n\} \times \{1,...,m\} \tag{6.1}$$

The size of the new space of possible is the product of the size of the two spaces associated to each observables. The presence of non-linear causal chains, possible bifurcations along ontogenetic paths, complex interactions between levels of observations, implies that the structure of the tensor space may be non trivial.

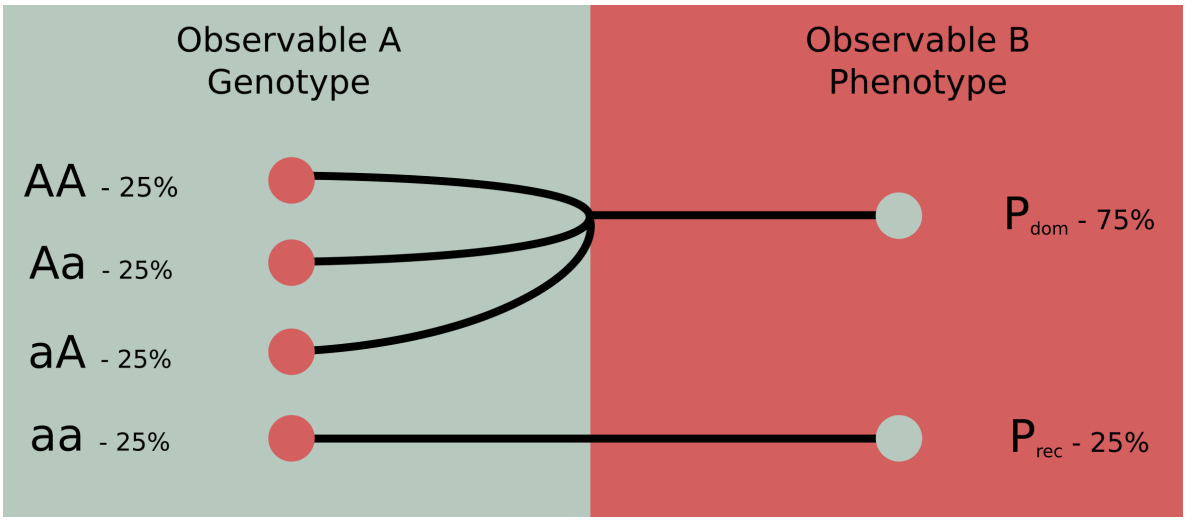


Figure 6.3: Scheme representing Mendel's model of inheritance in the first generation from the hybrids. On the left, the distribution of inheritance factors in the population with their relative frequencies (genotype). On the right, the distribution of phenotypes with their relative frequencies.

6.3.1 Mendel's model of inheritance: a formal analog of entanglement

In this section, we use an idealization of Mendel's experimental results on transmission of hereditary character to show the necessity of a tensor product for the space of possible of two correlated observables. Mendel's experiment consisted in constructing pure lines of peas, where crossing two individuals with a given phenotypic traits led to individual with the same phenotypic traits [150]. These pure lines were then hybridized, giving rise to a first generation of hybrids showing only one of the two phenotypes. The first generation from hybrid was then presenting the two ancestors phenotypes in proportion 3 to 1. To understand these distributions of phenotypes, Mendel proposed a model of inheritance involving hereditary factors and a phenomenon of recessivity and dominance. This model is summarized on figure 6.3 for the first generation from the hybrids.

The first observable is the genotype, with the hereditary factors A and a. In the first generation of hybrids possible genotypes are: $\{|AA\rangle, |aA\rangle, |Aa\rangle, |aa\rangle\}$. The second observable is the phenotype ("wrinkled" or "smooth"), the set possible phenotypes is $\{|P_{dom}\rangle, |P_{rec}\rangle\}$, dominant and recessive phenotype respectively. The state vector describing the distribution of phenotypes $|\psi\rangle$ can be written, according to the idealized experimental results:

$$|\psi\rangle = \sqrt{0.25}.|AA\rangle\otimes|P_{dom}\rangle + \sqrt{0.25}.|Aa\rangle\otimes|P_{dom}\rangle + \sqrt{0.25}.|aA\rangle\otimes|P_{dom}\rangle + \sqrt{0.25}.|aa\rangle\otimes|P_{rec}\rangle \tag{6.2}$$

If the two spaces of possible were separable, we should be able to write the state vector as

$$|\psi\rangle = (a.|AA\rangle + b.|aA\rangle + c.|Aa\rangle + d.|aa\rangle)\otimes\left(e.|P_{dom}\rangle + f.|P_{rec}\rangle\right) \tag{6.3}$$

However this equality leads to the following system of equations

$$\begin{cases} a.e = \sqrt{0.25} \\ a.f = 0 \\ b.e = \sqrt{0.25} \\ b.f = 0 \\ c.e = \sqrt{0.25} \\ c.f = 0 \\ d.e = 0 \\ d.f = \sqrt{0.25} \end{cases}$$

which has no solution.

Therefore we have

$$|\psi\rangle \neq (a.|AA\rangle + b.|aA\rangle + c.|Aa\rangle + d.|aa\rangle)\otimes\left(e.|P_{dom}\rangle + f.|P_{rec}\rangle\right) \tag{6.4}$$

$\forall(a,b,c,d,e,f) \in \mathbb{R}$

Equation 6.4 shows the non separability of state vector on the spaces corresponding each observable's subspaces. This is a formal analog of quantum entanglement [64]. This phenomenon stems from biological organization, the fact that multiple levels of organization are related to each other. Indeed, mixes of $|P_{dom}\rangle$ and $|P_{rec}\rangle$ do not exist at the same time in a single organism.

6.4 Discussion

We have shown in this article that biological variation can be handled to a certain extend with the mathematical framework of quantum mechanics. The distinction has been made between the unpredictability of the list of possible phenotypes corresponding to a well defined observable and unpredictability of possible observables. Variations in the choice of phenotypes is described with a probabilistic framework. This quantum probabilistic framework uses the structure of the hilbert space to be easily updated when new possible phenotypes are observed with a simple cartesian product between the actual space of possible and the direction defined by the new phenotype, a non trivial operation in Kolmogorov's simplex framework. The possibility to perform tensor product between spaces of possibles of several observables enables to consider complex causal relationships in a sound way.

To summarize, three levels of unpredictability are defined. The first one corresponds to the incompleteness of the set of observables, which is closely related to historicity of evolution and to the knowledge of the observer. The second one is related to the possible value of the observable, it involves the historicity of each developmental path and can be described with classical and quantum probabilities, as exemplified with the *squint* experiment. The third one is related to the complex correlations that relates various observables and is formalized with a tensor product, it is related to the organization of biological systems as shown with Mendel's idealized scheme of inheritance.

With no assumption on the mechanisms underlying biological variation, this approach throws light on the theoretical status of randomness in biology. The strongest form of randomness and its specificity with respect to physics is the unpredictability at the level of the set of observables [35], [33], [95]. It involves the observer as situated in space and time because the definition of the set observables is indexed on the knowledge of the observer, relying on retrospective analysis of the measure and an impossibility to predict future observable.

Chapter 7

Evolution and development: toward an ontogenetic tree

Abstract *This chapter investigates the relations between individual developments, normal and pathological, and the space of possible forms. The ontogenetic tree structure is proposed to organize observations on mutant developmental paths. It serve as a basis to define a developmental distance between developments. This developmental distance can be compared to a phenotypic distance to quantify the path dependency of a phenotype. Finally, the measure of the rate of diversification in the zebrafish development supports the description of the pharyngula stage as a phylotypic stage.*

7.1 Introduction

In this chapter, we propose some perspectives on the relationships between development and evolution. The interaction between development and evolution is a major question of evolutionary developmental biology ([87], [6], [5], [199], [217]). Originally, Darwin's theory of evolution uses as a fundamental principle the notion of descent with modification but without characterizing the underlying mechanisms sources of variation [49]. On the other hand Mendel's quantitative work on inheritance proposed a model explaining the distribution of phenotypes in controlled population through the transmission of inheritance factors[150].

Later during the XXth century, the proponents of the "Modern Synthesis" provided a synthesis between these two theories [106]. Associated to this theory comes the idea that genotype and phenotype should be clearly separated, and that the random variations oc-

cur at the level of the genotype. These approaches of population genetics lead to systems of equations such as $P = G + E$, a phenotypic trait P is the sum of genetic G and environmental E effects and thus $\mathrm{Var}(P) = \mathrm{Var}(G) + \mathrm{Var}(E) + 2.\mathrm{Cov}(G,E)$ where $\mathrm{Var}(P)$ represents the variance of a phenotypic trait P, $\mathrm{Var}(G)$ represents the genetic variance, $\mathrm{Var}(E)$ the variance in environmental conditions and $\mathrm{Cov}(G,E)$ the covariance between variations in the genotype and the environmental conditions [120]. As explicitly represented in this equation, this theory completely abstracts out the concept of development, and considers a direct mapping between the genotype and the phenotype, with selection acting at the level of the phenotype and variations occurring at the level of the genotype. However, these kinds of models have often been criticized for reducing organisms to their genes in a so-called "beanbag genetics" [143].

Taking the development into account as a phenomenon extended in time and occurring at different scales when considering evolutionary theories requires to introduce other concepts. The first problem consists in being able to compare developing organisms leading to phenotypes that are not necessarily comparable, for example between different species. Before the advent of developmental genetics, the concept of heterochrony during development has been proposed as a good operational tool to compare development in close or distant species ([87], [6]). Heterochrony, which has first been introduced by Ernst Haeckel, characterizes the changes in the timing of events during development. After having identified homologous parts in various organisms, it is possible to classify the transformation between developments in several classes, Acceleration, Neoteny, Hypermorphosis, Progenesis, Post displacement and Pre displacement depending of the changes in the control parameters which are the growth rate, the shape changes and the time of onset and offset of development of a given shape. These concepts have been successfully used to compare and understand the differences in the morphology of three species of salamanders.

Embryonic developments among animals of the vertebrate phylum has been shown to go through a very similar stage after presenting great variations in the early stages. This is the so-called phylotypic stage, leading to a developmental hourglass model. First proposed by Von Baer in 1828 [214], this model is considered as showing developmental constraints in development. The phylotypic stage has been supported by recent molecular evidence [115], [107].

Many morphological approaches have been put aside for many years after the development of developmental genetics in the 1980s. In particular, homeotic genes have been

shown to be involved in the establishment of the body plan in many organisms [161] , for example Hox genes in the *Drosophila* [184]. After these discoveries, the new field of evolutionary developmental biology was concerned with explaining the evolution of these genes which regulate development in many organisms.

Opposite to this last perspective is the "physicalist" approach developed by Vincent Fleury who explains the body plan only with physical constraints on the tissue generated by the early embryo, considered as a "tissue flow" [73]. This approach provides a very interesting perspective in terms of constraints undergone by the tissue which is a physical material. These physical constraints are relatively universal because they apply to any tissue having the same boundary conditions. However this approach only characterizes one level of organization, the tissular one, and neglect the behavior of individual cells.

The key concepts in the approaches bringing together development and evolution are the notions of canalization, plasticity, developmental stability, developmental constraints, evolvability and how development shapes evolution and how evolution constrain development [54], [30], [123]. Cryptic developmental variation can also be considered as an explanation for buffered phenotypic variability [61].

In the following, we would like to consider the developmental paths and the distribution of symmetry breaking as a way to characterize diversification. Following suggestions made in the frame of the so called dynamical structuralism, we propose here to link a historical perspective of morphogenesis with a comparison of resulting phenotypes using the zebrafish as a model organism[217]. First we will explore a concept proposed to correlate phenotype and genotype while including development in the *Drosophila* embryogenesis, under the name ontogenetic tree. We propose then to use this concept to link a "developmental proximity" with a "phenotypic proximity" in order to highlight structural constraints in developing zebrafish.

7.2 Reconstructing the ontogenetic tree of the *Danio rerio* embryogenesis

7.2.1 The concept of an ontogenetic tree

In a series of papers, Ho and Saunders develop the idea of an ontogenetic tree as a basis for a rational approach of taxonomy [101], [99]. They refer explicitly to the *Romantische Naturphilosophie* in Germany after the seminal work of Goethe where he introduced

the notion of morphology and the theory of transformation from an *Urpflanze* in 1790 [85]. This relates also to the debate between Cuvier and Geoffroy Saint-Hilaire in 1830 on the existence of a unique body plan for all animals [176], which later led to comparative anatomy. Of course, these conceptions have to be considered in their historical context and should not be transposed as such for contemporary science. On the other hand, in Darwin's evolutionary theory, published a few years later, the theory of forms and transformation is not central anymore, the main driving process comes from natural selection and thus adaptation. Morphologies diverge from each other among the genus, taxa, species, as the result of random variations. Although very interesting, these historical considerations belong to history of science and epistemology and should be discussed elsewhere.

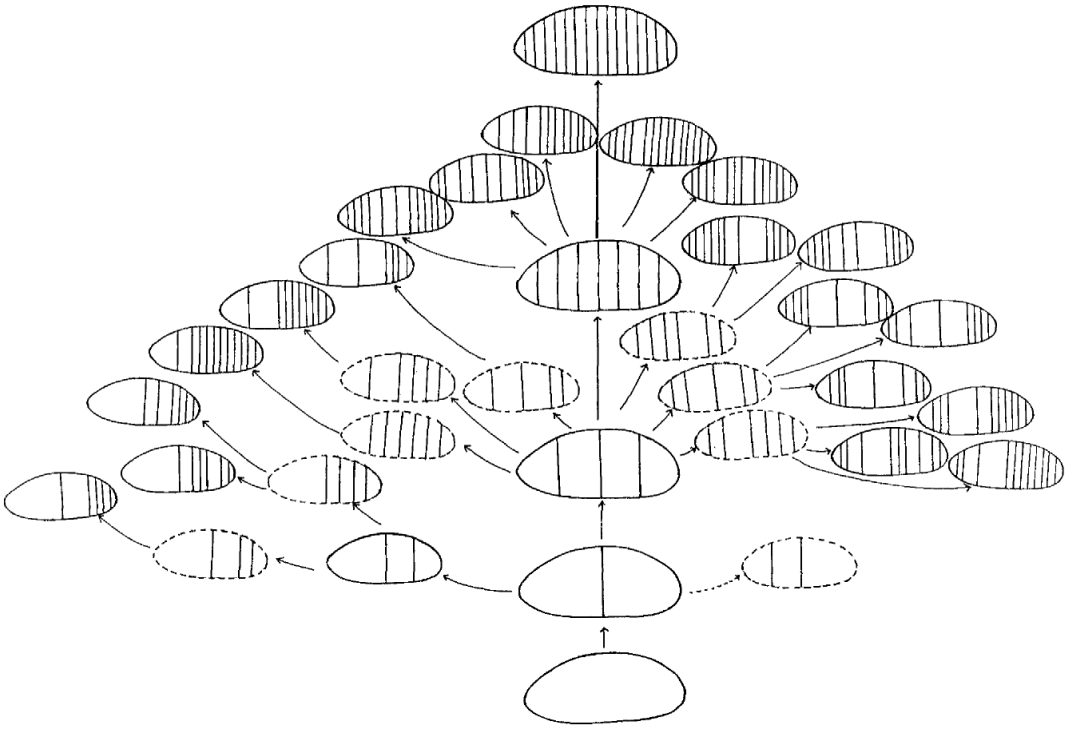


Figure 7.1: The ontogenetic tree of the observed morphologies in *Drosophila* with hypothetical intermediate phenotypes (dotted outline) based on developmental rules. Figure adapted from [101]

In their articles, Ho and Saunders refer to a set of experimental results resulting from genetic perturbations of the *Drosophila* embryos. These perturbations affect the segmentation process. What they argue is that these perturbations lead to phenotypic changes that cannot be completely arbitrary and necessarily depend on the developmental path.

Therefore, the final phenotypes are constrained by the intermediate stages through which the embryos are passing during development. The genetic perturbations generate successive bifurcations from the normal development. All of these developments are gathered together within an ontogenetic tree. Each bifurcation will generate a new branch in the tree. Figure 7.1 shows an illustration of the concept of ontogenetic tree for the *Drosophila melanogaster* in [101] based on data published in [100].

The approach and the structure of an ontogenetic tree are very appealing concepts, although the underlying assumptions for the process of segmentation may not be relevant anymore in regard of contemporary results. The structure of a tree can be formalized mathematically. The distance between resulting phenotype can be characterized by considering the length of the common developmental path, which constrain highly the subsequent development.

7.2.2 Formalization of the tree

Developmental processes are usually described with successive developmental stages corresponding to the setting up of the different parts of the organisms. Because any variation in the development will affect all the subsequent stages of the process, we can describe hierarchically the possible paths that stem from each mutational event. The pooling of different development paths that arise from different mutational profiles results in a hierarchical tree where each bifurcation corresponds to a symmetry breaking. The symmetry breakings that we consider here consist, for any part of the organism, in 1) change in timing, 2) abnormal development, or 3) absence of development. We will call this tree an ontogenetic tree in reference to previous work done on the drosophila segmentation [101] presented in the previous subsection. Thus, we can compute a "developmental distance", two organisms will be developmentally close if they share long branches, i.e. if they share a long common history.

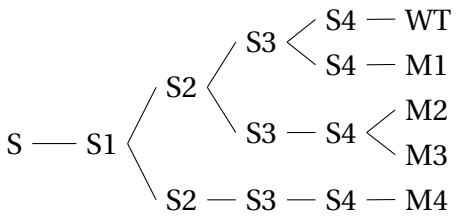
As a toy example, let's consider the following set of developmental sequences and symmetry breaking presented in table 7.1. Suppose that the development of the wild type (WT) is composed of four developmental stages $\{S1, S2, S3, S4\}$. Then, the developmental sequence associated to the wild type is $\{0,0,0,0\}$. A symmetry breaking in a mutant development will be denoted by a 1 in the developmental sequence at the corresponding developmental stage. For example the mutant M1 undergoes a symmetry breaking compared to the wild type at the third stage, therefore its developmental sequence is denoted as $\{0,0,1,0\}$. If several mutant developments are considered, they can be ordered with a

lexicographical order as shown in the table 7.1.

	S1	S2	S3	S4
WT	0	0	0	0
M1	0	0	1	0
M2	0	1	0	0
M3	0	1	0	1
M4	1	0	0	0

Table 7.1: Toy model of developmental sequences covering five developmental stages. From top to bottom, the sequences are ordered with a lexicographical order

The order obtained enables to gather the developmental sequences on a tree. With the toy model presented on table 7.1, the following ontogenetic tree is obtained



There is a first symmetry breaking at stage S1, separating the developments of (WT, M1, M2, M3) on one side and (M4) on the other side. Then, a second symmetry breaking is observed at stage S2 separating the developments of (WT, M1) on side and (M2, M3) on the other side. In the branch gathering (WT, M1), a symmetry breaking is observed at stage S3 separating the two developmental paths. In the branch gathering (M2, M3) the two developmental paths diverge at the stage S4.

If we define a developmental proximity d as the length of the common developmental path, we obtain the following inequalities:

$$d(M2,M3) \leq d(WT,M1) \leq d(WT,M2) \leq d(WT,M4)$$

The values of these developmental proximities can then be compared to a phenotypic proximity between the resulting embryos, e.g. based on the morphological similarity or on the functional similarity.

7.2.3 Observing the phylotypic stage

To explore these ideas further, we used the Zebrafish Model Organism Database (zfin.org) that gather current works on the zebrafish [29]. This database relies on a temporal ontology that describes the main developmental stages and a morphological ontology that describes the main parts of the organisms. 7317 mutations have been registered, with the description of their development using this ontology. We transformed each development description in a binary sequence that summarizes the normal stages and the symmetry breakings as defined in the previous section. After having ordered them using a lexicographical order we were able to map them on the same tree as exemplified on figure 1.

The first observation that can be made is that the distribution of developmental paths is spread over most of the developmental stages. At first sight, density of branches among the various developmental stages seems uniformly distributed. Moreover, we have plotted in different colors the mutant developmental paths affecting the eyes (yellow) and the developmental paths described without indication of an effect on the eye. The two colors seem to be well distributed along the developmental stages, indicating that there are no preferential stages for effects on the eyes.

Are the mutant developmental paths distributed uniformly along the ontogenetic tree? When and where do the higher number of development paths arise? One way to compute answers to these questions is by considering for each developmental stage the number of developmental path which undergo a first symmetry breaking at this stage, i.e. diverging from the wild type development. For example, with the toy data set presented on the table 7.1, there is one developmental path diverging from the wild type at stage S1 (M4), two at stage S2 (M2, M3) and one at stage S3 (M1).

Figure 7.3 shows the result of the computation of the number of diverging developmental paths for each stage of the zebrafish embryogenesis. This number can be interpreted as a diversity potential. We can see that except for the early embryogenesis the diversity potential increases with the succession of the developmental periods until the pharyngula period before decreasing with the hatching period. The pharyngula period corresponds to the period of organogenesis. This result is coherent with the fact that this period has been considered as the phylotypic stage, the main axis are already established and the variation can therefore take place on less critical parts of the organism.

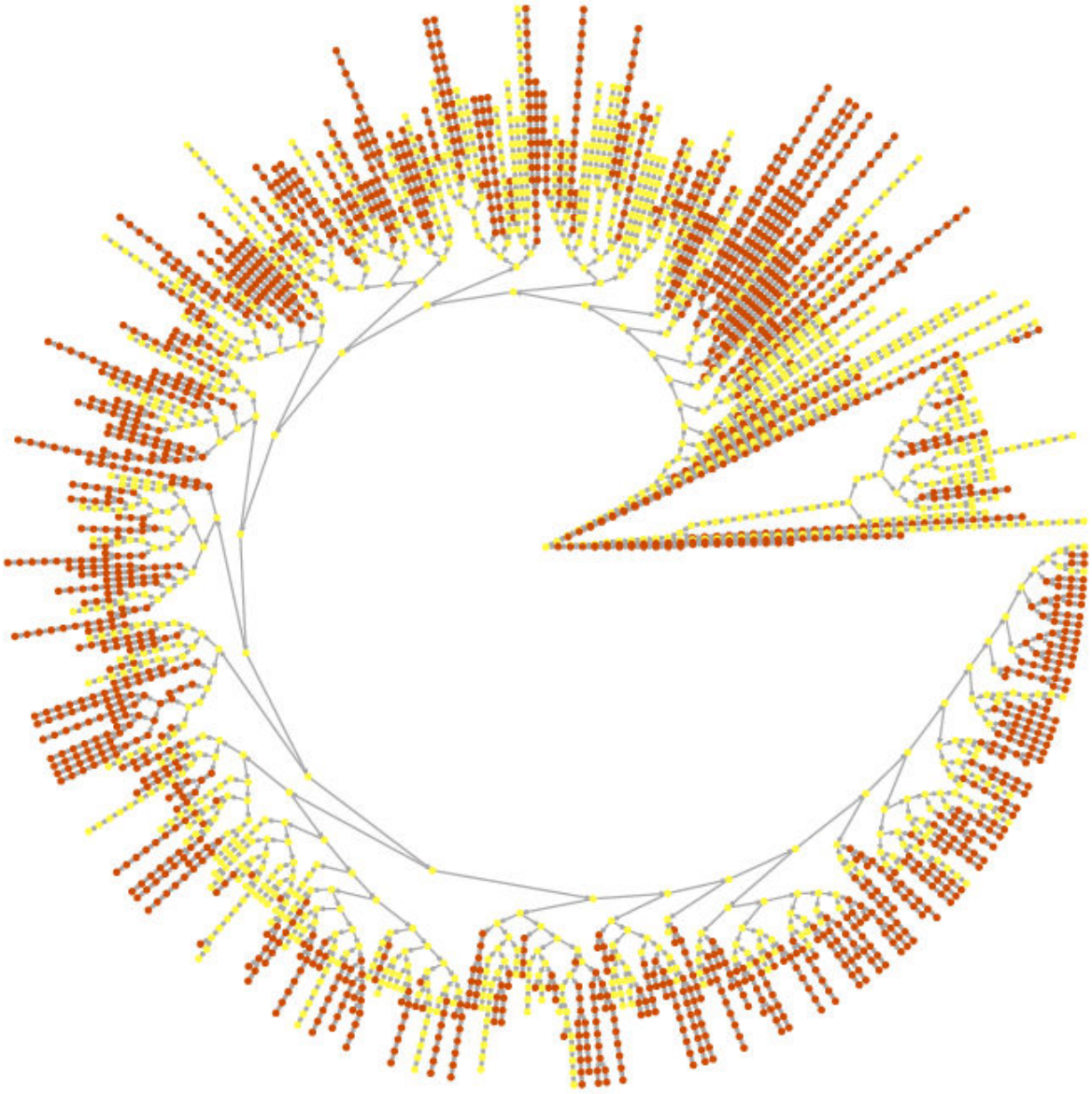


Figure 7.2: Reconstruction of the zebrafish ontogenetic tree using zfin.org database. Each leaf of this tree corresponds to a set of mutant phenotypes that share a similar history, each node corresponds to a developmental stage, each branching corresponds to a symmetry breaking. Nodes are colored in yellow if they correspond to phenotypes having a default on the eye and in red if not

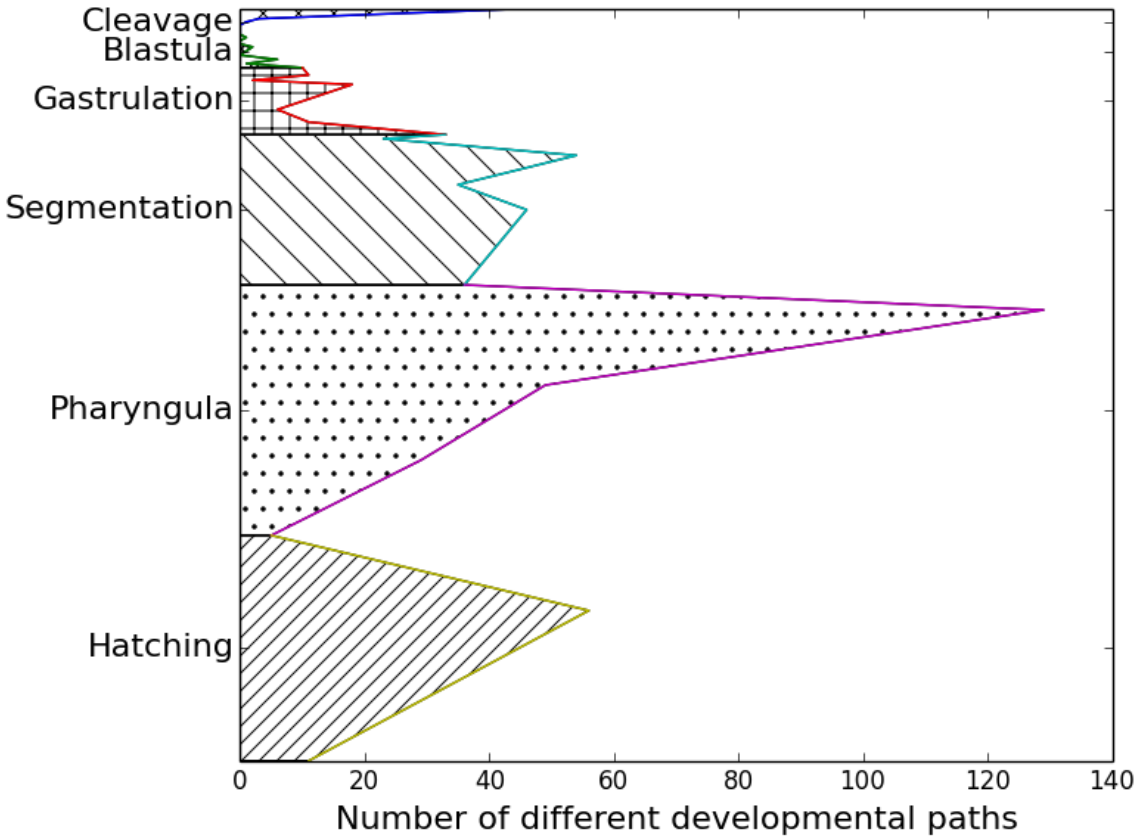


Figure 7.3: Diversity potential per developmental stage and period evaluated by counting the number of diverging branches from the wild type development

7.3 Discussion and conclusion

The main flaws of this study reside in the nature of the dataset which can be biased by 1) the chosen structures and phenotypes for the ontology used to describe the development (epistemic flaw) 2) the description biases related to the teams making the experiments (epistemic flaw) 3) the mutagenesis methodology (sampling methodology) .

It could be interesting to provide a null model of tree, to show that the diversity potential is really significantly higher for the pharyngula period compared to randomly distributed changes along the developmental paths.

Overall, we see that the concept of an ontogenetic tree can be useful to compare a large number of developmental sequences. The divergence pattern between developmental sequences can explain some of the morphological similarity between species, since divergence is more likely after critical stages of development having been gone through. Therefore the early stages of embryogenesis canalize the possible forms of final phenotype. Variation seems to proceed mostly on already established organization. Variation cannot be considered as leading to any arbitrary possible phenotype. This view relates to von Baer conception of diversification by successive specialization. Understanding the development in different species, could explain and help classify the morphological differences observed in later stages.

Conclusion

The main idea unifying the studies presented in this part of the dissertation is to understand the relation between variations in individual development and generation of diversity. The multi-scale nature of organisms, i.e. their *organization* and the path dependency of development, i.e. their *historicity*, require to give to variability a specific role.

The review of the main mechanisms at the origin of diversity in Chapter 4 points out that variability can emerge at multiple scales in an organism, from gene mutation to stochastic effects and epigenetic events. The heterogeneous nature of these mechanisms make them difficult to integrate in a common framework. However they are relevant to understand diversity since they all have an effect on resulting phenotypes upon which natural selection act.

Turning to mathematics and physics doesn't simplify the picture. Indeed, the formalisms developed rely on different hypotheses, for example in chaotic systems theory and ergodic theory, the characterizations of randomness are different. The mathematics of probability theory provide a framework enabling calculus on uncertain events without defining any notion of randomness. The assumptions underlying the axioms, however, involve some modeling hypotheses on the described situation. These assumptions may not be always the most suited as witnessed by the probability theory developed in quantum mechanics, generalizing Kolmogorovian probability theory which is the most common framework used in classical situations.

Given the heterogeneous nature of mechanisms at the origin of diversity, and the heterogenous nature of mathematical frameworks used to model situation of uncertainty, it is argued that, as an alternative to a theoretical integration of sources of diversity, it is possible to use hypothesis based modeling and simulations to explore the range of possibles allowed by a set of identified mechanisms. However, the simplifying assumptions necessary to establish such models reduce the generality of their results.

After these general considerations on variability in the living, we turn our attention, in

Chapter 5, to a specific experimental system which shows surprisingly high levels of variability. Crossing homozygote zebrafish mutants lead to a phenomenon of variable phenotypic expressivity and incomplete penetrance, in the *squint* mutant line. The mutation affect the *Nodal signaling* pathway which is involved in the establishment of the main body axes of the zebrafish. A quantitative description of the distribution of phenotypes (interocular distance) in the progeny of identified homozygote mutant parents leads to the following properties of variability: the discrete list of possible phenotypes is explored in unpredictable proportions, this list is possibly incomplete. Trying to find determinants to this variability involve complex causal relationships.

Theoretical implications of the results of the *squint* experiment are investigated in Chapter 6. In this chapter, we propose an analogy with quantum mechanics to clarify the different levels of variability encountered in biology. Indeed, variability leads, for a given observable, such as the interocular distance in the *squint* experiment, to the observation of phenotypes in variable proportions. These proportions can be described in terms of probability. However, this list of possible phenotypes may be extended with the observation of an emerging new phenotype. In that case, the space of possible associated to the observable needs to updated, this can be performed with a cartesian product between the previous space of possible and the new direction defined by the new phenotype. This operation requires a vector space structure for the set of probability distributions. Finally, in the course of evolution a radically new phenotype can emerge, requiring new observables for its description. In that case, the new observables define new spaces of possibles. This level of unpredictability is much higher than the one described by probability theory in a pre-given set of possible phenotypes. It is specific of the historical nature of biological objects. After having clarified these levels of uncertainty formally with a mathematical framework analog to the framework developed in quantum mechanics, it is possible to describe a way to infer causal links or correlation between observable values. Using Mendel's scheme of inheritance as an example, we show that the dominance-recessivity relationships imply a phenomenon formally analog to quantum entanglement. This effect can be explained by the fact, that given biological organization, two phenotypes cannot coexist in the same organism.

The path dependency of developmental processes implies that the relations between genotype and phenotype are not linear. The *squint* experiment shows that the relations can be highly complex. However, it is possible to try to measure and quantify the influence of path dependency on resulting phenotypes. This is the object of Chapter 7. In this chap-

ter, we propose a formal structure, an "ontogenetic tree", to gather mutant developments and characterize their diverging patterns in relation to each other. Using this tree structure, we can compute a developmental distance between phenotypes by measuring the length of their common development. This developmental distance has to be compared to a phenotypic distance in order to highlight the effect of the different developmental stages on the resulting phenotype. As an example we use a data set of numerous descriptions of zebrafish mutant developments. By computing the ontogenetic tree, we show that the pharyngula period of development generates the highest number of diverging developmental paths, canalizing therefore subsequent degrees of freedom for development. These results should however be manipulated carefully since multiple biases can exist in the data set.

Altogether, we show that variability at all scales has an effect on diversity. The integration of this variability in a unique framework is an open and difficult problem which is not reducible to already existing mathematical framework because of the *organization* and *historicity* of biological objects. It is a central problem of biology that should be considered in any study.

Part III

Quantifying biological shapes

Introduction

In this part, we consider the question of shape. We propose to study epithelial organization in developing embryos. Epithelial tissues constitute one of the simplest structure found in development. The question is to find invariants in this structure among several species and genetic backgrounds and way to quantify variability in a generic manner on this relatively universal structure.

One approach for the study of epithelial organization considers the network of cellular contacts from static images. This mathematical structure enables to characterize the spatial distribution of cells. Previous studies have used the tools of complex networks studies, however these approaches underestimate the role of underlying topological constraints of the tissue which reduce the power of these measures. Developmental histories leading to epithelial organization prevent to find simple symmetrical structures, epithelia are indeed the result of cell proliferation, cell motility and cell extrusion producing complex structures.

Departing from an analogy with the structure of the cosmic web which is highly historical in nature, we propose to define topological invariants based on the tools of persistent homology. Using this approach we were able to compare and classify a wide range of tissues and get back to species and genetic background classification.

To assess the significance of this approach, we introduce a model of random triangulated surface. This model constrains local characteristics: the degree distribution in the network is set. The choice of neighbors in the network is free and random, the only condition being that the resulting network can be embed in a surface. This approach uncovers the higher-order spatial constraints in the distribution of the cells and shows the existence of patterns involving groups of cells in significant proportions. These constraints can be interpreted as *the mark of individual developmental histories*. Relying on a data set consisting in confocal images of epithelia from Drosophila and Chick embryos, we perform a comparative study in Chapter 8.

Chapter 9 proposes some perspectives on a dynamical characterization of the network of cellular contacts. In particular we argue that time should not be considered as a similar dimension as space dimensions since its evolution changes the nature of the network. Finally we propose a measure that combine the study of genealogical relations between cells and their spatial organization at the same time. This measure should be able to quantify the unfolding of shape in space and time.

Chapter 8

Using persistent homology to quantify tissue shape and organization

Abstract *Epithelial tissues are simple cellular structures found in developing embryo in different species. The organization of these tissues has been studied using local properties of the network of cellular contacts. However, accounting for global and spatial properties requires extending these approaches. Using persistent homology on the network of cellular contacts reveals global topological characteristics. To assess the significance of these characteristics, we provide a model of random triangulated surfaces with arbitrary degree distribution. These oriented surfaces are obtained by randomly gluing oriented polygons; this process results in planar graphs with appropriate degree distribution. We explore the topological characteristics of these surfaces and compare them to a set of empirical data. Differences between the null model and the data provide insights for the understanding of underlying biological processes and equip us with a notion of "level of randomness". This notion can be used to evaluate the contribution of genetic factors such as the presence of Myosin II.* [1]

8.1 Introduction

Embryonic morphogenesis is a complex process involving regulations at all scales. Epithelia are simple tissular structures found in a wide range of embryo. An epithelium is a

1. The study presented in this chapter has greatly benefited from the supervision of Gunnar Carlsson during an extended stay at Stanford University Mathematics Department (Feb.- Jul. 2014). This work has been presented at the conference "Algebraic Topology - Methods, Computation and Science 6" at Vancouver, Canada (05/26-30/2014)

monolayer of densely packed cells. The spatial distribution of cells forms a complex structure shaped by cell proliferation, cell motility and cell extrusion throughout development [90].

Resulting spatial structure and individual cell shapes within the tissue depend on this sequence of events. Cell shapes in 2D can be modeled as polygons. This model allow to relate cell proliferation to distribution of cell shapes in epithelia [80]. A generalization of this approach considers the whole network of cellular contacts [67]. The network of cellular contacts is an abstract representation of the epithelial organization. Each cell is a node and the contacts between cells are represented as edges. Similar patterns of cellular contacts has been studied in crack patterns [27], [26], or in foam dynamics [121].

The authors of [67] import the tools of complex networks for the study of the network of cellular contacts. This approach is pursued in a subsequent study [191]. We take in this chapter a different approach to study the network. Using invariants defined in algebraic topology, we propose to quantify high-order structures in the network of cellular contacts. We develop an algorithm of discrete pattern recognition to describe structure involving small groups of cells in the network of cellular contacts. This approach gives us access to intrinsic and tissue-level characteristics of the network of cellular contacts.

8.2 Global characterization of epithelial tissues

8.2.1 Network of cellular connectivity

The structure of an epithelium can be described with the network of connectivity between the cells: two cells are connected if their membranes are in contact. This network can be formalized as a couple (V,E) where $V = \{v_1, v_2, ...v_n\}$ is the set of vertices in the network corresponding to each cell in the tissue and $E = \{\{v_i, v_j\} \in V^2\}$ is the set of unoriented edges corresponding to each cellular contact in the tissue [7]. Figure 8.1 shows a representation of this network in a sample of tissue. This network is obtained by segmenting the cells in the image. Although it is represented in space, it is a purely combinatorial structure. In figure 8.1, the nodes positions correspond to mass center of the associated cells. It differs from a voronoi tessellation of the space based on the cell centers because regions defined by cell membranes may not be convex.

In a network, the degree of a node corresponds to the number of edges coming to this node, it corresponds to the number of neighbors of a cell. The frequency distribution of n-

sided cells in a tissue as described in [80] is exactly the degree distribution in the network of cellular connectivity. Figure 8.2 shows this distribution in different type of epithelia. We can see that the shape of this distribution varies greatly from one type of tissue to the other, although they are all centered on 6-sided cells.

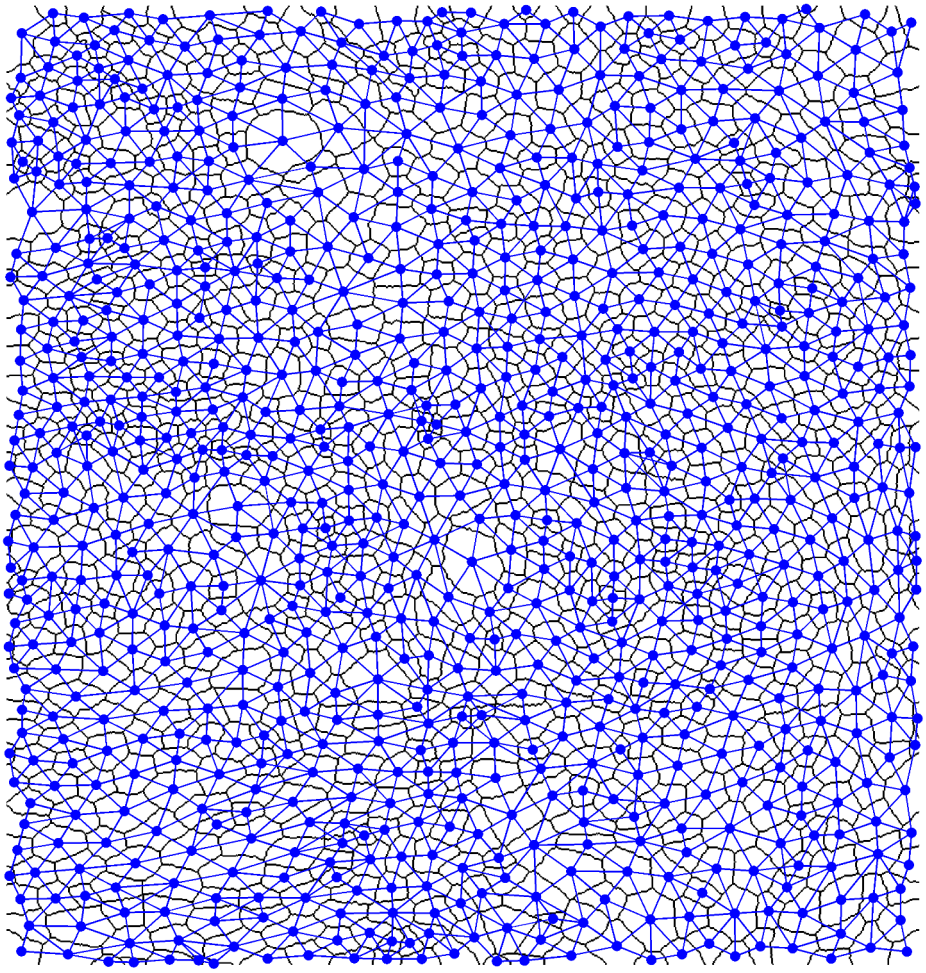
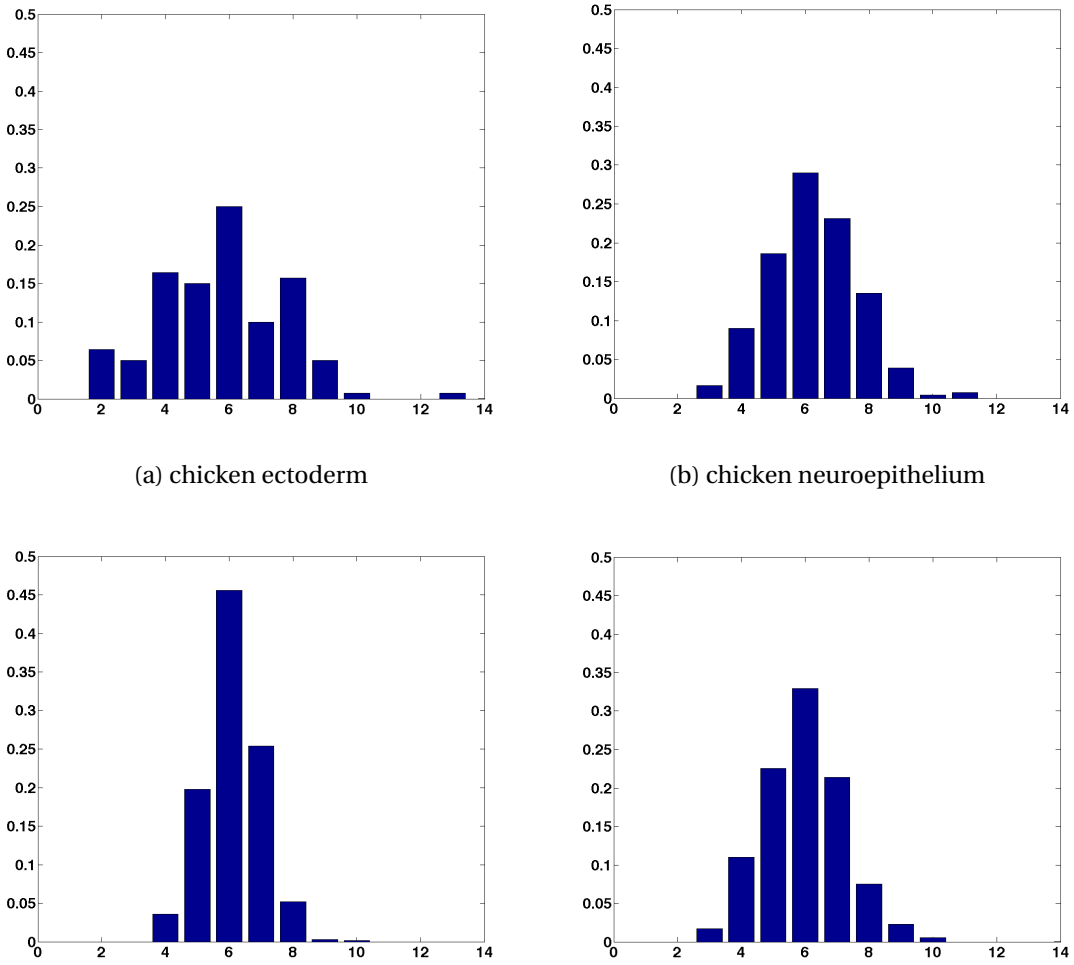


Figure 8.1: Sample of epithelium from the Drosophila Wing Prepupa where the cells membranes are represented in black and the network of cellular connectivity is visualized in blue. The vertices of the network are placed on the mass centers of the cells, although the network of cellular connectivity is a purely combinatorial object

The study of the network of cellular connectivity has been recently introduced in the



(a) chicken ectoderm

(b) chicken neuroepithelium

(c) drosophila wing prepupa

(d) drosophila mutant wing prepupa

Figure 8.2: Degree distribution computed from the network of cellular connectivity in various tissues. Number of neighbors varies around the mean value of 6. The proportion is indicated as an empirical frequency: number of nodes with a given number of neighbor divided by the total number of nodes.

paper [67]. Authors of this paper use three measures to characterize this network: the degree, the clustering coefficient and the average degree of neighbors. The degree is the number of neighbors of a cell. The clustering coefficient is a measure of local packing which is defined for a cell as the ratio between the number of observed interconnections in its neighborhood and the number of interconnections in a complete graph having the same number of nodes (i.e. where all the nodes are interconnected). The average degree of neighbors is the averaged degree over the neighbors of the considered node.

8.2.2 Complex networks approach shows some limitations

The degree, the clustering coefficient and the average degree of neighbors are measures of network developed in the context of complex systems approaches [48]. If complex network approaches are relevant for the study of sociological networks, they may show some limitations for the study of epithelia which are highly constrained by the underlying topology. Network of cellular contacts in epithelia can be assimilated to a triangulated planar graph as a first approximation. A planar graph has the property to be embeddable in a plane without any edge crossing. A triangulated planar graph has for dual network, a network where each node has degree 3.

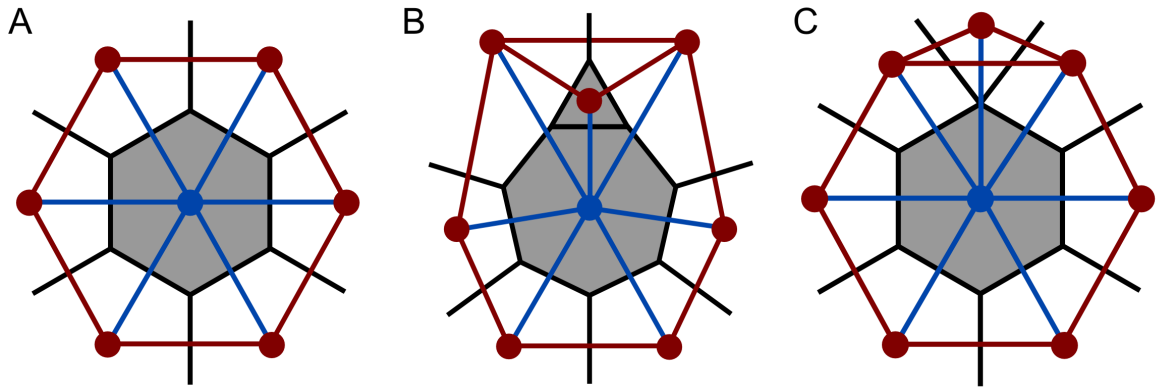


Figure 8.3: Schematic representation of the network of cellular contacts. k blue edges connect a node to its neighbors. Red edges connect neighbors to each other. The clustering coefficient is defined for a cell as the ratio between the number of observed interconnections in its direct neighborhood and the number of interconnections in a complete graph having the same number of nodes (i.e. where all the nodes are interconnected). A) Most common case, clustering coefficient $C = 2/(k-1)$ B) Case with a 3-sided cell, $C = 2.(k+1)/(k.(k-1))$ C) Case with a 4-sided junction, $C = 2.(k+1)/(k.(k-1))$

It is easy to show that in a planar triangulation, the clustering coefficient of a node

having k neighbors is equal to $\frac{2}{k-1}$. Therefore, the measure of the clustering coefficient doesn't provide more information than the number of neighbors. The average degree of neighbors is a measure the connectivity of neighbors and gives information about wider context in the network. The accuracy of this last measure is nevertheless dependent on the dispersion of the degree in the network, and may not provide better information than the degree distribution itself for networks with low degree variation. These classical measures of complex networks have been introduced in the context of communications or sociological networks where the underlying space has a complex topology [7]. They have been very successful to describe Internet topology or networks of actors. The limitations presented here derive from the global topology of epithelia. It is thus necessary to introduce new measure that would be able to provide global information about the network of cellular contacts, and which are not reducible to the number of neighbors.

Alternative approaches can be found in the study of topological properties of networks. There is a large mathematical literature that describe the topology of network ([89]), yet the traditional characteristics like the euler characteristics or the genus may not be accurate to distinguish between different tissues since most of the time epithelia are planar graphs (or close to it). They are contractible, that means that they have the homotopy type of a point. Some differences may appear due to transient processes such as rosette formation that induces 4-sided or 5-sided junctions and appear transiently during the process of cell intercalation [25], [205]. They are however described as unstable features of the networks and may not form the basis of stable signature of a tissue topology. The following will present more sophisticated tools aimed at finding robust signatures of the topology of the networks and unfolding the richness of their organization.

8.2.3 Persistent homology

A useful way to approach the topology of a complex landscape is to use the framework of persistent homology ([38], [37], [62]). This method has been applied successfully in cosmology to characterize the filamentary structure of the cosmic web ([200]) among many examples. The idea is to define a sequence of nested subspaces of the investigated space indexed by a parameter and to observe how the topology of these spaces change with the parameter. These changes are then summarized by a barcode or a persistence diagram, which gives a topological signature of the space. This method refine the description of the space by providing a multi scale representation, unfolding its organization. Moreover, stability results show that this description is robust [43].

In the case of the network of cellular connectivity, we can use the number of neighbors as a proxy for the density of the space. For each network (V,E), two filtrations are defined. A filtration is a nested sequence of spaces. The filtrations considered here are indexed with a parameter k (natural integer) corresponding to values of degree. They are defined as a discrete analog of the sub- and super level sets functions. We first define two sequences of nested networks, $\underline{Sub} = \{Sub(k)\}_{k\in\mathbb{N}}$ and $\underline{Sup} = \{Sup(k)\}_{k\in\mathbb{N}}$:

$$Sub(k) = (\{v_i \in V : d_i \leq k\}, \{e = \{v_i, v_j\} \in E : d_i \leq k \;\&\; d_j \leq k\}) \tag{8.1}$$

$$Sup(k) = (\{v_i \in V : d_i \geq k\}, \{e = \{v_i, v_j\} \in E : d_i \geq k \;\&\; d_j \geq k\}) \tag{8.2}$$

$Sub(k)$ and $Sup(k)$ are subnetworks of (V,E) such that each node and each edge connect nodes having at most (at least respectively) k neighbors in (V,E). We have therefore $Sub(1) \subset Sub(2) \subset ... \subset Sub(k_{max})$, and $Sup(k_{max}) \subset Sup(k_{max}-1) \subset ... \subset Sup(1)$, with k_{max} being the highest degree measured in (V,E). These subnetworks are the analogs of sub and super level sets in the continuous case. For each value of the parameter k, clique complexes $Sub^*(k)$ and $Sup^*(k)$ are constructed from the subnetwork $Sub(k)$ and $Sup(k)$. Clique complexes contain simplices for each clique of the network as shown on figure 8.4. A clique is a complete graph, i.e. a set of nodes with an edge for each pair of node. For example, if a set of three nodes are completely connected, the clique complex of this graph is a triangle, if four nodes are completely connected, the clique complex is a tetrahedron. The sets of clique complexes $\underline{Sub^*} = \{Sub^*(k)\}_{k\in\mathbb{N}}$ and $\underline{Sup^*} = \{Sup^*(k)\}_{k\in\mathbb{N}}$ form natural filtrations as they are constructed on top of the filtrations $\underline{Sub}$ and $\underline{Sup}$. Figure 8.5 gives an example of the filtration.

The idea of the persistent homology algorithm is to describe the topology of the spaces contained in the filtrations $\underline{Sub^*}$ and $\underline{Sup^*}$ for each value of the parameter k. Persistent features will correspond to intrinsic features of the object, whereas features associated to noise will be unstable and won't last. The topological features considered are the Betti numbers, which measure the i-dimensional "holes" of a space. For example, for a torus, as represented on figure 8.6, the number of 0-dimensional holes (β_0) is 1, it is the number of connected components. The number of 1-dimensional holes (β_1), i.e. "circles", is 2, the one that encircles the void in the middle of the torus and the one that encircles the torus. The number of 2-dimensional holes (β_2), i.e. "voids", is 1, the one inside the torus.

Computing these topological invariants for each subspaces of the filtrations results in a multi-scale object called a barcode that summarizes the evolution of these features

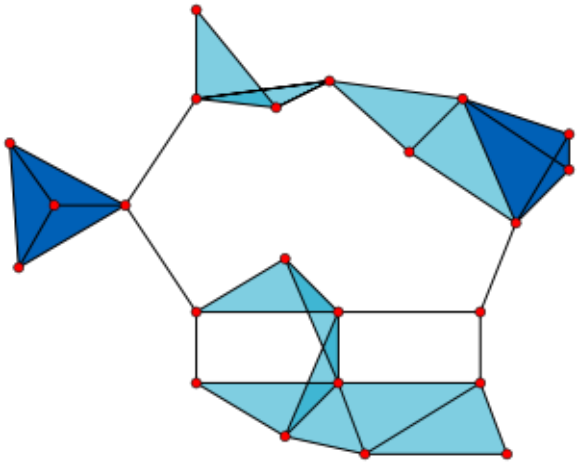
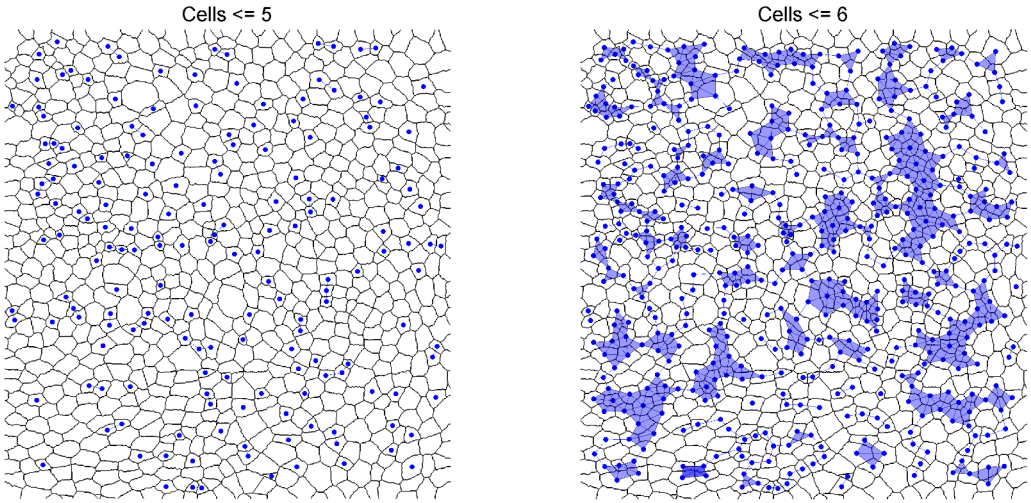


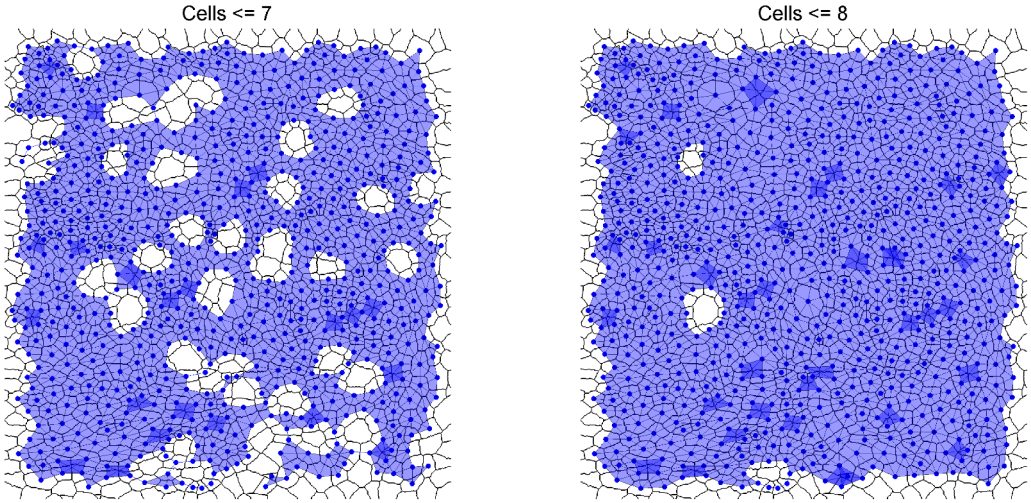
Figure 8.4: Example of a Clique complex on top of a network. Nodes are in red, edges in black, triangles in light blue when three nodes are connected, tetrahedron in deep blue when four nodes are connected - From wikipedia

throughout the value of the parameter indexing the sequence. Bars in the barcode correspond to generator of i-dimensional holes. A new bar appears when a new i-dimensional hole appears in the filtration and the bar disappears when the hole disappears. Long bars can be interpreted as robust features features of the space whereas small bars can be induced by local small variations. Implementation of the algorithm is described in [223]. We used the javaplex library [2] to compute a barcode for each epithelium. Figure 8.7 shows barcodes for the sample of tissue represented on figure 8.1. The first line of a barcode shows the evolution of the number of connected components which is visualized through the number of bars. The second line shows the evolution of cycles in the networks and the way they are filled. The barcodes corresponding to sub and super level sets in this example are not identical, both approaches provide useful information, and they will be considered jointly in the following.

The filtration represented on figure 8.5 can be directly related to the barcode represented on figure 8.7 (a). When the parameter is equal to $d = 5$, we see that sub complexes consist only in unrelated connected components, the barcode has bars only for Betti 0 which counts the number of connected components. When the parameter is increased to $d = 6$ we see still a significant number of connected components and the birth of holes. This is visible on the first line of the barcode where the number of bars for Betti 0 is constant and new bars appears for Betti 1. For $d = 7$, the simplicial complex become a single connected component where most of the nodes are connected, it contains holes. The number of bars for Betti 0 falls to 1 the single connected component. The numbers of bars



(a) (b)



(c) (d)

Figure 8.5: Sequence of simplicial complexes $\{Sub^*(k)\}_k$ for the same sample of tissue represented in figure 8.1 (from Drosophila Wing Prepupa), with $k = 5$, $k = 6$, $k = 7$ and $k = 8$. Cell membranes are represented in black and simplicial complexes in blue. Nodes are positioned on the mass center of cell surface area, although simplicial complexes are a purely combinatorial structure.

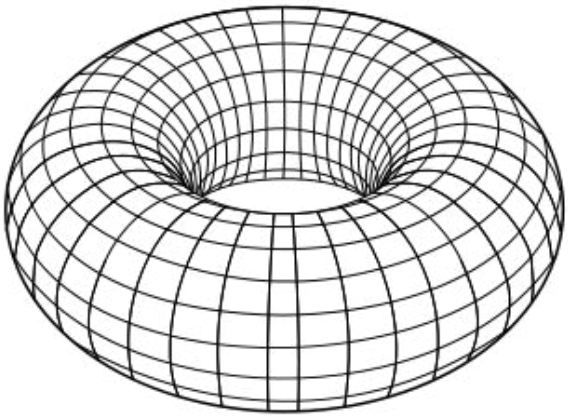


Figure 8.6: Example of a torus, its Betti numbers have for values $\beta_0 = 1$, $\beta_0 = 2$, $\beta_0 = 1$ - Source wikimedia

for Betti 1 is maintained, although discontinuously indicating change of the position of the holes. For $d = 8$, we still have a single connected component, where the number of holes is highly reduced. The number of bars for Betti 0 stays at 1, the single connected component. The number of bars for Betti 1 falls with the decrease of holes number.

Barcodes computed over the filtrations $\underline{Sub^*}$ and $\underline{Sup^*}$ defined above offer a multi scale representation of the epithelial topology. They carry useful information that can be used to investigate the global organization of the network of cellular connectivity. These barcodes provide a qualitative information that can be compared from one sample to the other because they represent similar features in the networks. Their differences reflect changes in the organization. To make these comparisons possible on a broad range of tissue samples it is however necessary to introduce quantitative features associated to each barcodes.

8.2.4 Quantitative comparison by computing features on top of barcodes

In order to compute quantitative features based on the barcodes which are the objects summarizing the evolution of the topological features of the spaces along the filtration, we used a methodology proposed in [3]. In this article the authors suggest to use the coordinates of the bars of a bacode as the domain of functions that will return quantitative value characterizing the barcodes. These functions have to satisfy a certain number of criteria that makes them sufficiently generic over different types of barcodes. These criteria

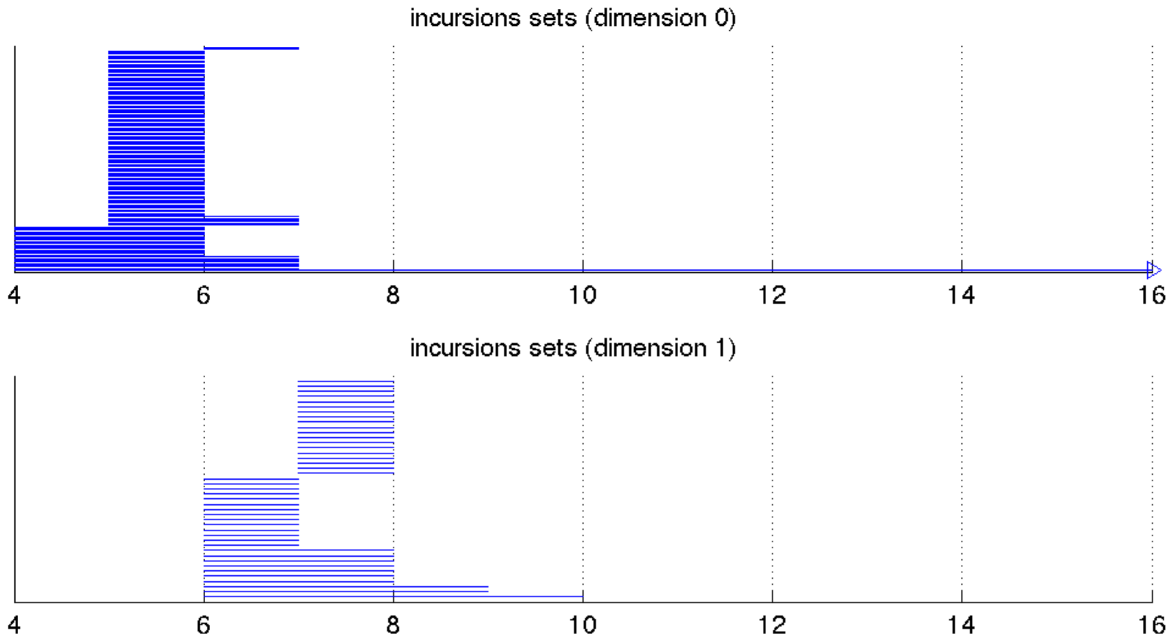


Figure 8.7: Barcodes for $\underline{Sub^*}$ for the sample of tissue represented in figures 8.1 and 8.5. Each individual blue line correspond to a generator of the i-dimensional homology. Dimension 0 shows the number of connected component. Dimension 1 shows the number of holes. In each graph abscissa correspond to the value of the parameter in the filtration, here the degree of a node in the network

are sufficient to define functions in terms of tropical polynomials that will be used in the following.

For each Betti number, corresponding barcodes can be written as sets of couples $\{(x_i, y_i)\}_i$, with $(y_i > x_i)$, each of them representing a bar. Only bars of finite length will be considered. We denote by n the number of bars. We define the functions that are relevant to our analysis in the following way:

$$\frac{1}{n}\sum_i (y_i - x_i) \tag{8.3}$$

$$\frac{1}{n}\sum_i (y_i - x_i)^2 \tag{8.4}$$

$$\frac{1}{n}\sum_i (y_i + x_i)(y_i - x_i) \tag{8.5}$$

$$\frac{1}{n}\sum_i (y_i + x_i)^2(y_i - x_i) \tag{8.6}$$

$$\tag{8.7}$$

Each of these functions returns a unique real value associated to one of the Betti number. They highlight different aspects of the barcodes: the lengths of the bars for the first two, with more weight on longer bars for the second one, and the relative shift in the barcode for the last two of them, with more weight accorded to this shift for the fourth one. They are divided by n, the number of bars, to avoid biases introduced by variable of the samples.

Thanks to these functions computed over the barcodes, we obtain four features for each Betti numbers in each filtration of each sample of tissue. Betti 0 and Betti 1 are considered only, Betti 2 have only infinite bars or no bars. In summary, 16 features are obtained for each sample of tissues: 4 features x 2 barcodes x 2 filtrations. These features can be used to compare a broad range of epithelial tissues.

8.2.5 Classification of tissues

Using the data set provided with the article [67], we used these vectors of features as a basis for a statistical analysis over different type of epithelia. The data set consists of sample of tissues extracted from live embryo and imaged with confocal microscopy. 9 to 15 samples have been imaged in each of the different following situations: drosophila Wing

Prepupa, drosophila mutant Wing Prepupa (reduced level of Myosin II expressed), chicken Neuroepithelium, chicken embryonic Ectoderm. For each sample, the network of cellular connectivity has been extracted. Each sample of tissue is plotted with respect to the characteristics of its degree distribution on figure 8.8. We can see that the different tissue types are clustered according to their degree distributions. Even if they have a similar general organization as an epithelium, the networks present some specificities depending on the tissue type.

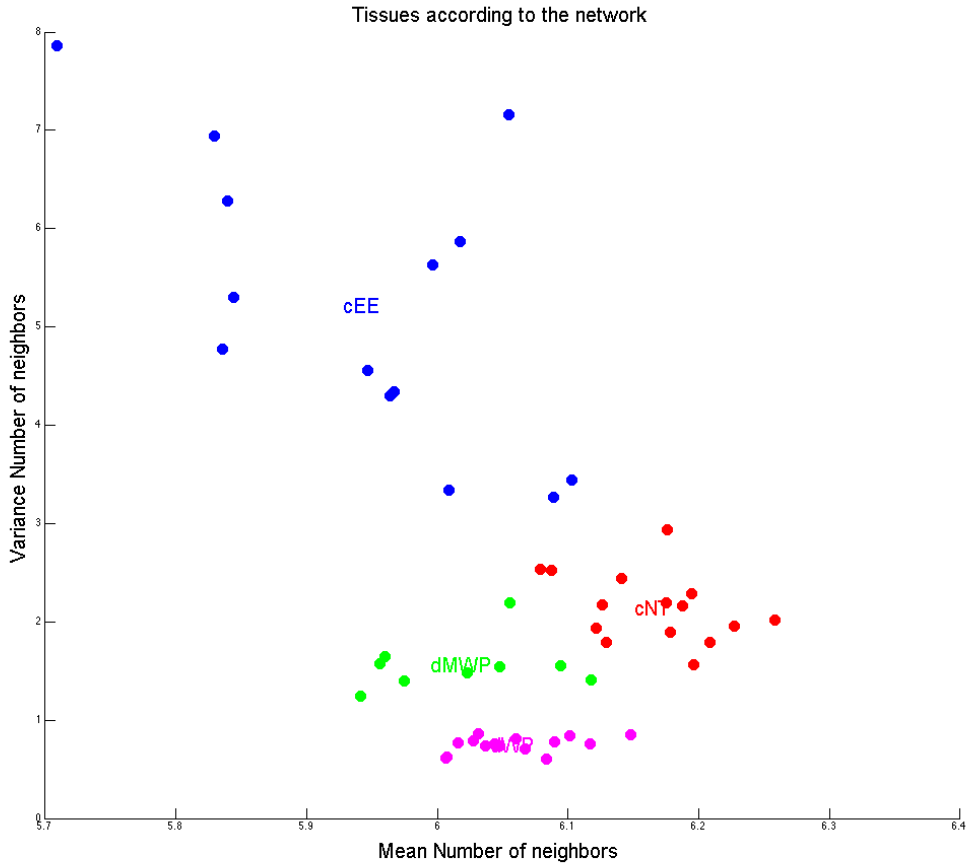


Figure 8.8: The 52 samples of epithelium are well separated using the main characteristics of the degree distribution. Each point corresponds to one sample of tissue and is plotted in the space defined by the mean number of neighbors versus the standard deviation. Each color corresponds to one type of epithelium - Magenta: Drosophila Wing Prepupa (dWP) - Green: Drosophila Mutant Wing Prepupa (dMWP) - Red: Chick Neuroepithelium (cNT) - Blue: Chick Ectoderm (cEE)

Persistent homology has been computed on the filtrations $\underline{Sub^*}$ and $\underline{Sup^*}$, and the

vector of 16 features has been computed. These samples are represented in a 16-dimensional space. To visualize their distribution, they have been projected on the plane defined by the two first principal components of the principal component analysis as represented in figure 8.9. This figure shows that points are clustered according to tissue type indicating a robust topological signature generated by this methodology. The spread of points is not the same for the different tissue types. It is the highest for samples from chicken Ectoderm, whereas it is the lowest for sample from the drosophila's Wing Prepupa.

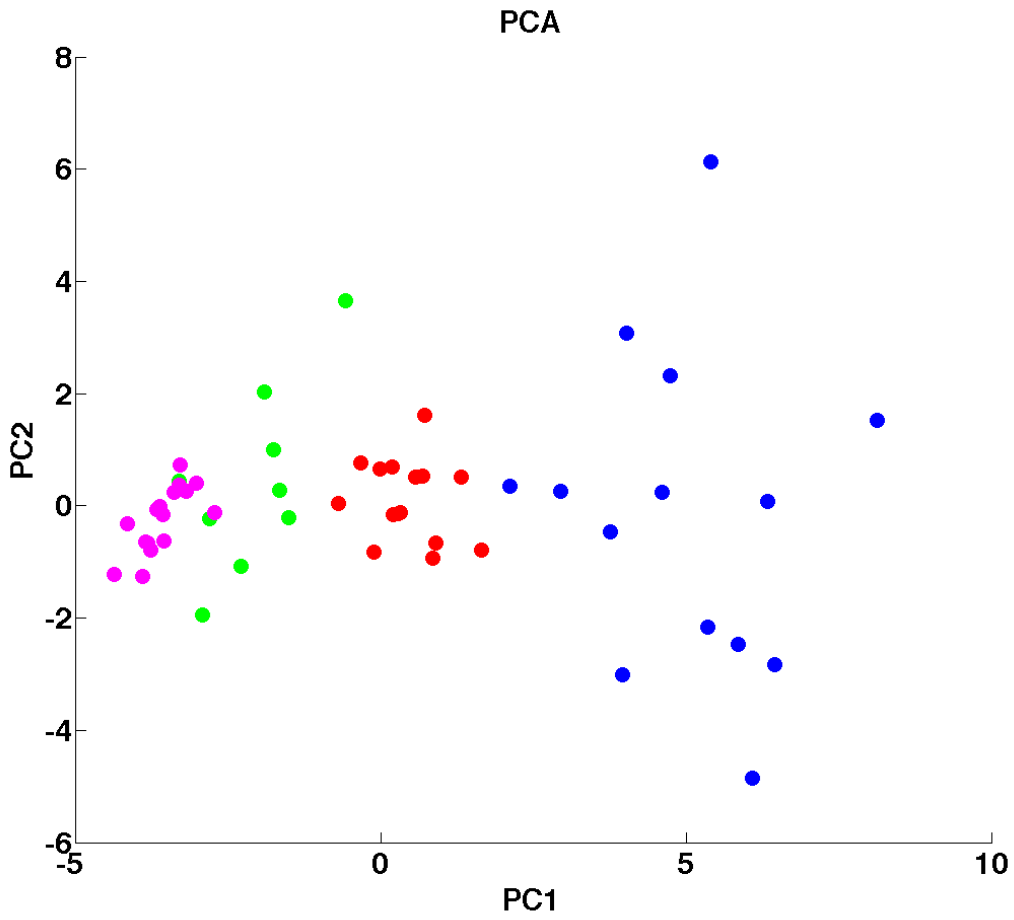


Figure 8.9: Principal component analysis performed on the quantitative features, defined in section 8.2.4, associated to each sample of tissue - Magenta: Drosophila Wing Prepupa - Green: Drosophila Mutant Wing Prepupa - Red: Chick Neuroepithelium - Blue: Chick Ectoderm

Given the apparent separability of the clusters, it is expected that more sophisticated statistical analysis and machine learning tools would capture precisely the value of the

features obtained for each type of tissue. These observations prove that images of epithelia can be compared and classified using persistent homology. Moreover, we will see below that computing persistent homology enables to characterize some aspects of the organization that are missing when considering the degree distribution alone.

8.2.6 Summary

The automatic pipeline that has been developed in this section is described by the following steps (see table 8.1):

sample of tissue (membranes)

↓

cell segmentation

↓

network of cellular contacts

(purely combinatorial object)

↓

filtration on the sub and super level sets of the number of neighbors

↓

persistent homology algorithm

↓

barcodes

↓

quantitative features

↓

comparison and classification

Table 8.1: Steps of the automatic pipeline enabling comparison and classification of epithelial tissues

8.3 Random surfaces with arbitrary degree distribution to model tissue topology

In the previous section we provided a method to compare samples of tissues taking into account their global organization. However, to gain a biological understanding of this organization, the differences observed in the features values need to be interpreted. We propose in this section a model of random graphs that tries to make sense of the observed differences. The question that will support these investigations is the question of how local relations between cells affect the global organization.

8.3.1 Use of a null model

One way of answering the question of how local interactions between cells influence the global organization of the tissue is to use a null model that will match some properties of the network while keeping other properties free. Since the question is to understand how relations between cells affect the global topology of the network, the properties that will be set are the local properties of cells, thus the degree. The choice of the type of connections between vertices will then affect the global organization of the network. Comparing an empirical network with its random counterpart having the same degree distribution will highlight the role of the connections between nodes on topological features.

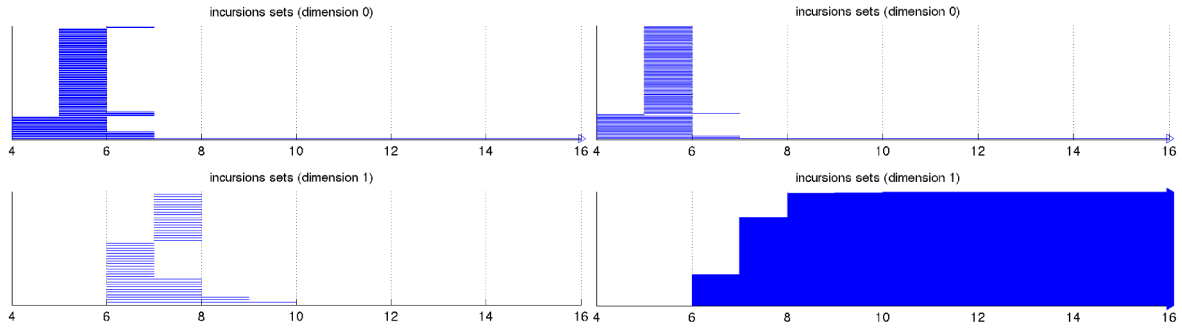
8.3.2 Topological hypotheses are necessary

The simplest model of random graph with arbitrary degree distribution is the Newman Strogatz model [158]. In this model each node has a degree randomly drawn from the degree distribution. The nodes are then randomly connected to each other. With this construction process, the degree distribution is prescribed in the network and nothing is assumed about the relations between the nodes.

We generated random networks with empirical degree distribution for each sample of tissue. The previous section described a way to quantify global topology of any combinatorial network, thus random networks can be studied in the same way. Figure 8.10 shows the comparison of the barcodes between an actual sample of tissue (on the left) and the random network having the same degree distribution (on the right). The comparison of these two figures shows a clear difference in the barcodes of these two networks.

This difference mostly rely on the barcodes obtain for Betti 1 (cycles). The barcodes corresponding to Betti 0, describing the simply connected components, are very similar in the two situations. First order topology is captured by the degree distribution. Thus, the barcode of Betti 0 seem to be in that case another representation for the degree distribution which is independent of the way nodes are related to each other. Higher order structures are much more different as highlighted by the differences in Betti 1. High order structures arise from the interconnection between nodes.

The topology corresponding to higher order structures seems to have a very specific organization in the actual sample of tissue compared to the random network. Connecting nodes randomly induces a complex underlying topology which is very far from the highly constrained network observed in biological tissues, which can be embeddable in the plane. This property canalizes these higher order structures. Our approach using persistent homology is proved to be relevant here, since the difference between the random network and the empirical data wouldn't have been noticeable based only on a comparison of the degree distribution.



(a) (b)

Figure 8.10: Comparison of the barcodes for an actual sample of tissue (a) and its random counterpart (b) using the Newman Strogatz model of random networks

To understand the relations between cells in epithelial networks of cellular contacts and refine our understanding of epithelial organization, more explicit constraints have to be imposed in the null model.

8.3.3 Randomly gluing polygons

One way of describing epithelial topology, already proposed in [80], is to approximate cells with polygons and consider epithelia as tessellation of the plane. Using this idea,

a random epithelium can be modeled as a random tessellation that would maintain and control the distribution of polygons. Random orientable surfaces can be obtained by randomly gluing triangles [175]. The resulting triangulation forms a planar graph embeddable in the sphere. However these kinds of model don't control the degree distribution. We generalize this model by randomly gluing polygons (they inherit the characteristics of the random triangulated surfaces since the dual graph is a random triangulation), while keeping the orientation.

The steps of construction of the random surface obtained by gluing polygons are described on the figure 8.11. The idea is to begin first with a randomly chosen polygon, then to fill its empty sides with randomly drawn polygons (while preserving the orientation and the degree distribution), and to reiterate with its neighbors. Whenever possible empty sides are connected with already existing empty sides (while preserving the orientation). The orientability of the whole surface is guaranteed by constraining the orientation when adding any new polygon to the construction.

This model of random network imposes in addition to a prescribed degree distribution that the dual graph has degree three. It is the easiest way to construct a graph that is embeddable in the plane.

Any degree distribution can be provided as an input, we show several examples on figure 8.12.

Figure 8.13 shows the barcode for a sample of tissues and its random counter part obtained by randomly gluing polygons. Barcodes for Betti 0 are similar, which is analogous to what has been obtained in the case of the Newman Strogatz random network. This similarity is explained by the fact that the number of simply connected components is mostly dependent on the shape of the degree distribution. Barcodes for Betti 1 are very similar between the data and the model, this is a clue toward the fact that this model is able to capture the features present in the global organization of the tissue. These results show that the random model captures well the global organization of the network by imposing to the graph to be embeddable in the plane, confirming that this is a major constraint on this type of networks.

8.3.4 Topological characteristics of random surfaces

Before going into a systematic comparison of the data and the model, the stability and robustness of the model needs to be assessed. To evaluate the stability of the model, it is relevant to study its dependency to input parameters. The parameters are the number

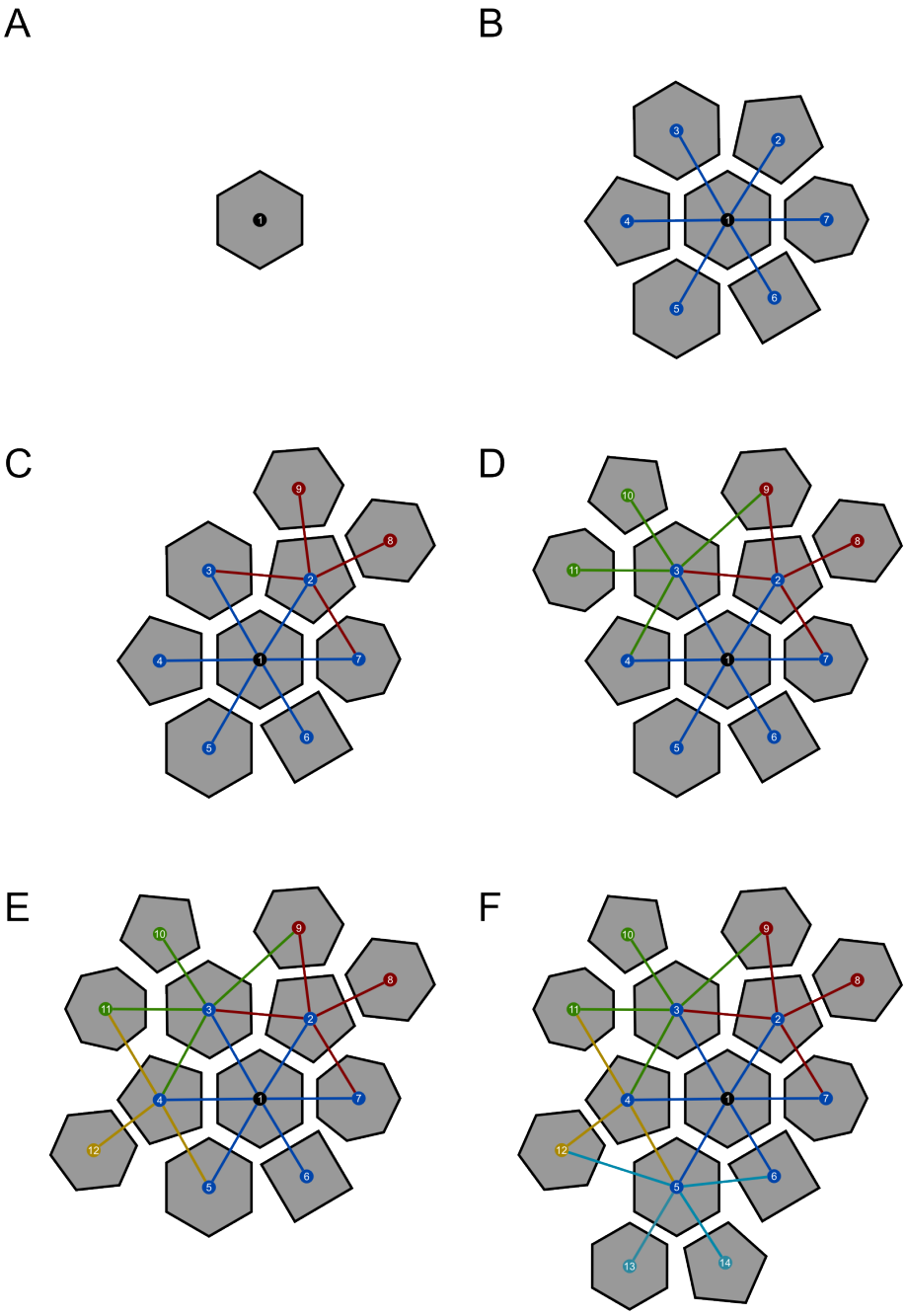
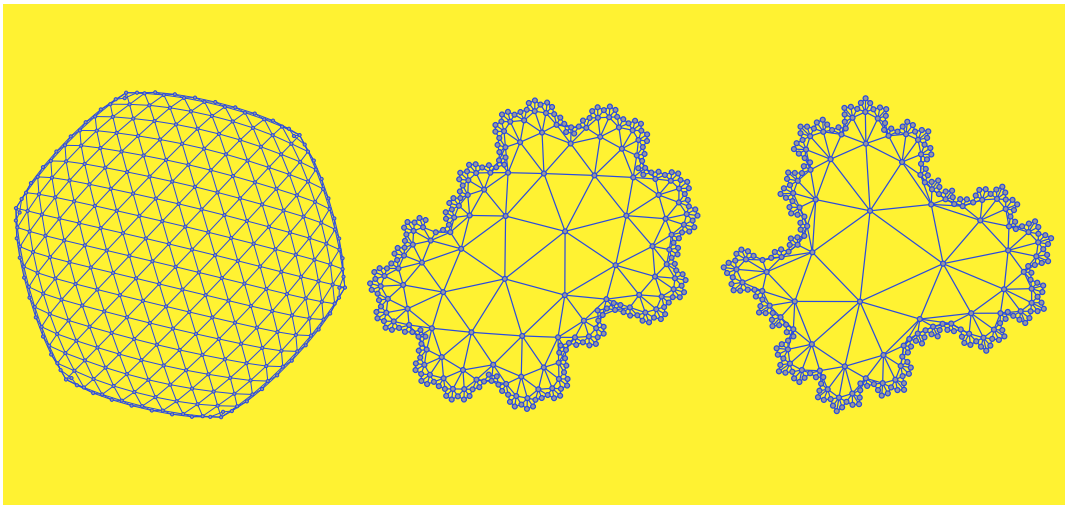
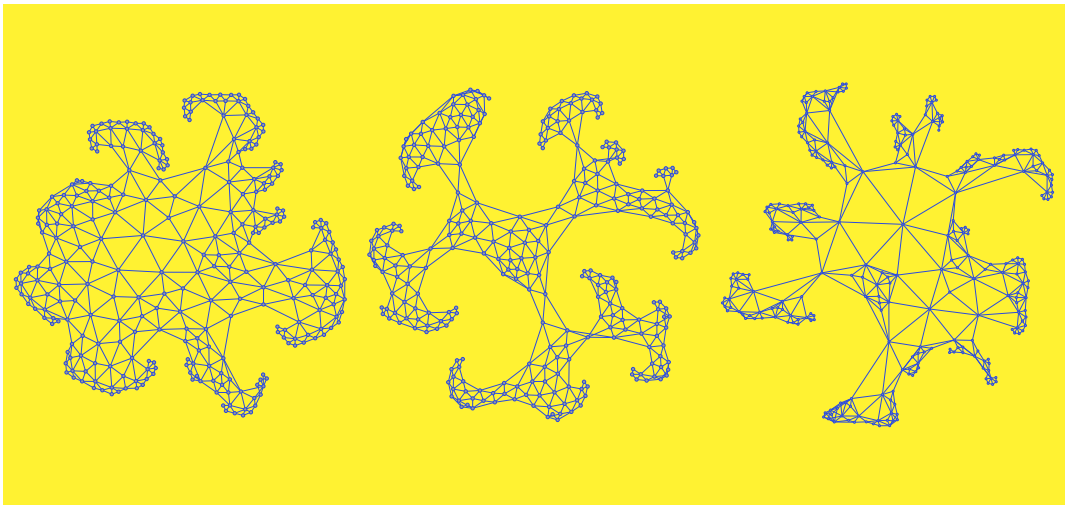


Figure 8.11: Steps of construction of the random triangulated surface with appropriate degree distribution. Every polygons are drawn from the same degree distribution. The order of gluing is indicated by the number on the polygons.

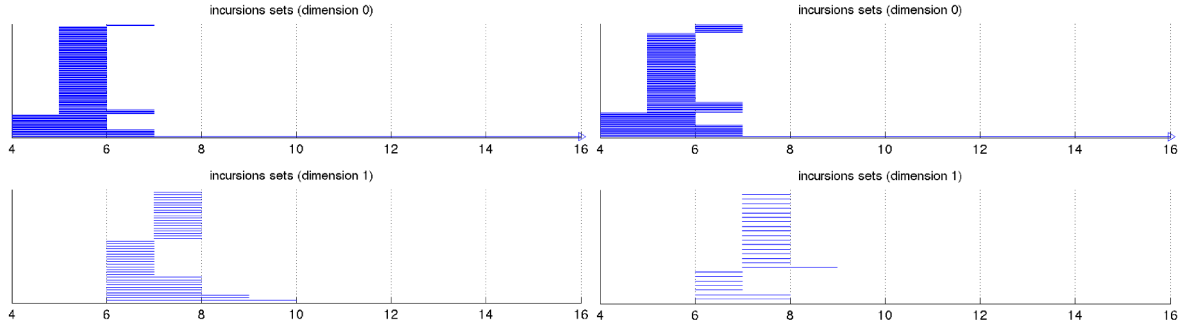


(a) (b) (c)



(d) (e) (f)

Figure 8.12: Examples of random networks obtained by gluing polygons, when the distribution is a) only 6-sided polygons b) only 7-sided polygons c) only 8-sided polygons d) 5 and 7 sided polygons equiprobable e) 5 and 6 sided polygons equiprobable f) 4 and 8 sided polygons equiprobable. The spatial representation has been obtained with the gephi software, it corresponds roughly to a minimization of repulsion forces between the nodes.



(a) (b)

Figure 8.13: Comparison of the barcodes for an actual sample of tissue (a) and its random counterpart (b) generated as a random triangulated surface with the same degree distribution

of nodes and the shape of the degree distribution used to construct the model. A way to evaluate the stability of the model is to compute its topological features as defined in the first section, equations 8.4, 8.5, 8.6, 8.7, and look at their evolutions while changing the parameter values.

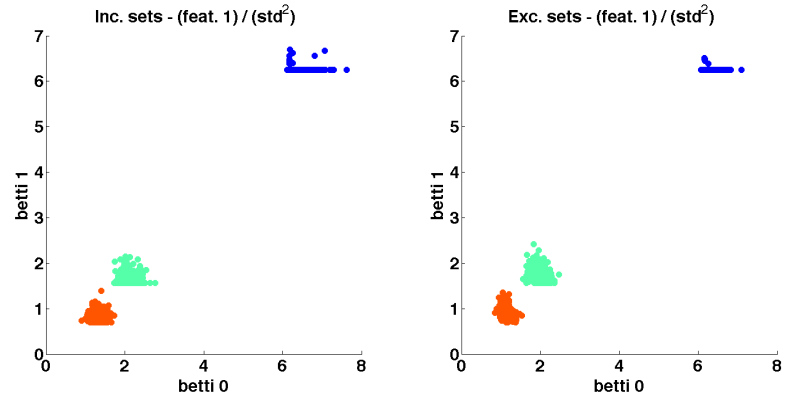
To assess empirically the topological characteristics of the model of random surfaces, we performed a systematic exploration of parameters. Degree distributions were chosen as normal distribution with mean taking values in the set (5,6,7,8) and a standard deviation taking values in the set (0.4,0.8,1.2), the number of polygons varied from 100 to 100 000. The results show a scaling of the features value with the parameters of the degree distribution, independently of the number of polygons. Figures 8.14 and 8.15 show the evolution of the features values. The high degree of clustering of the features values for the simulations proves the stability of the features values, and their independence to the number of nodes.

8.3.5 Comparison of the null model and the data for each of the features

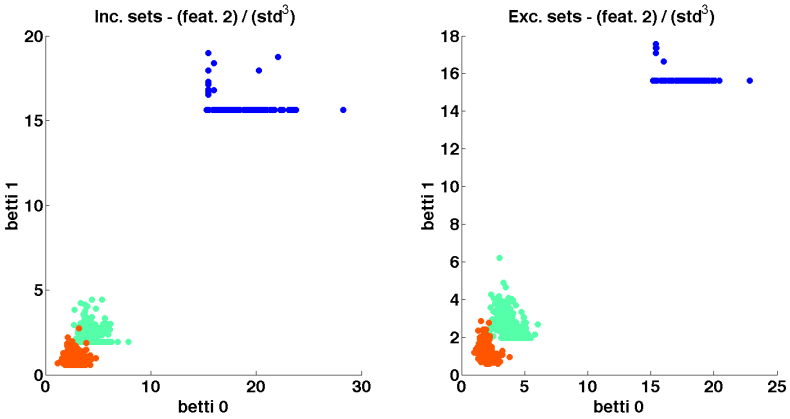
For each sample we computed the corresponding random network by using the empirical degree distribution and the same number of nodes. 5 realizations of the random model have been simulated for each sample of tissue.

For each of the features

We found that the data and the model have similar features values for the features 1,2 and that they vary more for features 3 and 4. These features have been defined in the first

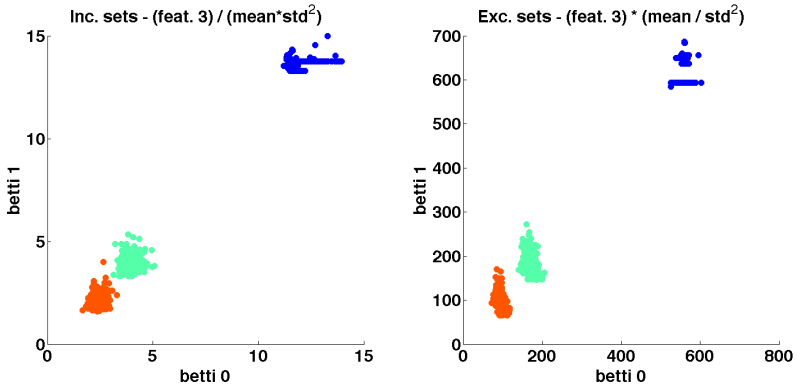


(a)

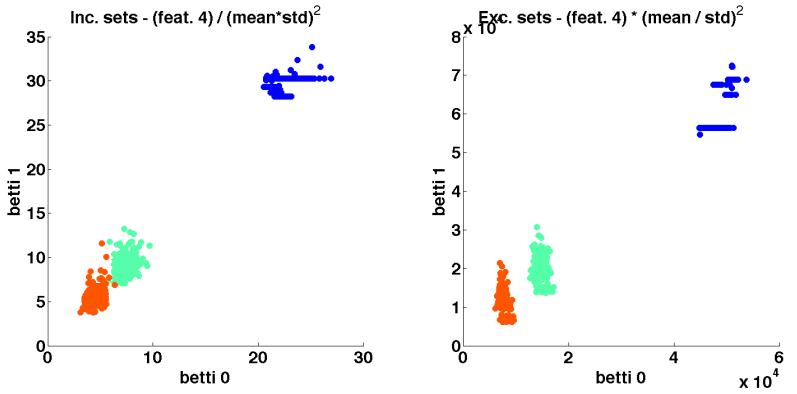


(b)

Figure 8.14: Scaling of the topological features with the model's parameters and independence of the topological features' values with respect to the number of polygons in the model. Degree distribution were chosen as normal distribution with mean (μ) taking values in the set (5,6,7,8) and a standard deviation taking values in the set (0.4,0.8,1.2), the number of polygons varied from 100 to 100 000. Points are colored according to standard deviation value. Graphs (a) correspond to the evolution of feat $1/\sigma^2$. Feat 1 has been defined according to equation 8.4. Left graph correspond to $\underline{Sub^*}$ and right graph to $\underline{Sup^*}$. Graphs (b) correspond to the evolution of feat $2/\sigma^3$. Feat 2 has been defined according to equation 8.5. Left graph correspond to $\underline{Sub^*}$ and right graph to $\underline{Sup^*}$.



(a)



(b)

Figure 8.15: Scaling of the topological features with the model's parameters and independence of the topological features' values with respect to the number of polygons in the model. Degree distribution were chosen as normal distribution with mean (μ) taking values in the set (5,6,7,8) and a standard deviation (σ) taking values in the set (0.4,0.8,1.2), the number of polygons varied from 100 to 100 000. Points are colored according to standard deviation value. Graph (a)-left corresponds to the evolution of feat $3/\mu\sigma^2$ for $\underline{Sub^*}$. Feat 3 has been defined according to equation 8.6. Graph (a)-right corresponds to the evolution of feat $3\mu/\sigma^2$ for $\underline{Sup^*}$. Graph (b)-left corresponds to the evolution of feat $4/(\mu\sigma)^2$ for $\underline{Sub^*}$. Feat 4 has been defined according to equation 8.7. Graph (b)-right corresponds to the evolution of feat $4(\mu/\sigma)^2$ for $\underline{Sup^*}$.

section, equations 8.4, 8.5, 8.6, 8.7. They correspond to functions computed on top of the barcodes.

Distance between the null model and the data - "randomness level"

For each epithelium a degree distribution is computed using the network of cellular contacts, 5 random triangulated surfaces are generated according to this degree distribution. The 16 features described in section 8.2.5 are computed for the random surfaces. Distribution of random surfaces and the empirical samples are compared in the 16-dimensional space with a measure of proximity d.

For principal component i of the PCA defined in section 8.2.5

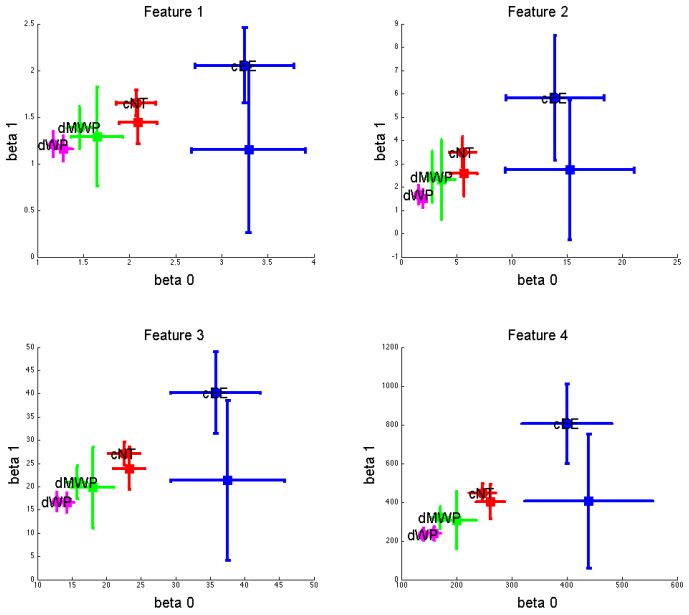
$$d_i = e_i . \frac{\mu_i^{emp} - \mu_i^{rand}}{\sigma_i^{emp}} \tag{8.8}$$

where e_i is the eigenvalue associated to principal component i, μ_i^{emp} is the mean value of empirical samples associated to one of the tissue type (chicken ectoderm, chicken neuroepithelium, drosophila wing preppie, drosophila mutant wing preppie) along i and σ_i^{emp} the standard deviation (over the set of similar tissue type), μ_i^{rand} is the mean value of the null models corresponding to the same empirical samples along i.

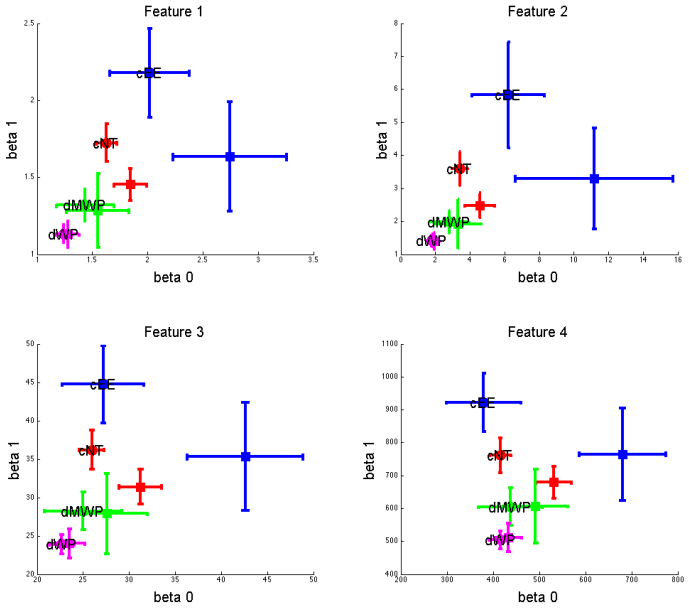
The measure of proximity d is defined as the sum of d_i over all of the 16 dimensions.

$$d = \sum_{i=1}^{16} d_i \tag{8.9}$$

Figure 8.17 shows the results. Chick Neuroepithelium, in red, is the closest of the four epithelial types examined here to the null model. It means that in this kind of epithelial organization do not present patterns at the scales of the groups of cells, as much as the other types of epithelia do. The Drosophila Wing Prepupa is the most distant from the null model, it means that this tissue has the most constrained organization. Drosophila Mutant Wing Prepupa is as expected closer to the null model than the wild type. Indeed, Myosin II which is an element of the cytoskeleton is involved in epithelial organization. However, surprisingly, this distance is higher than for the Chick Neuroepithelium. That means that other factors than Myosin II contribute to epithelial organization.



(a) *Sub**



(b) *Sup**

Figure 8.16: Computation of the features for the data and the model for sub- (a) and super- (b) level sets. The results for the empirical data are plotted with the name superimposed and a rounded label. The results for the null model are plotted with squared label. Magenta: Drosophila Wing Prepupa - Green: Drosophila Mutant Wing Prepupa - Red: Chick Neuroepithelium - Blue: Chick Ectoderm

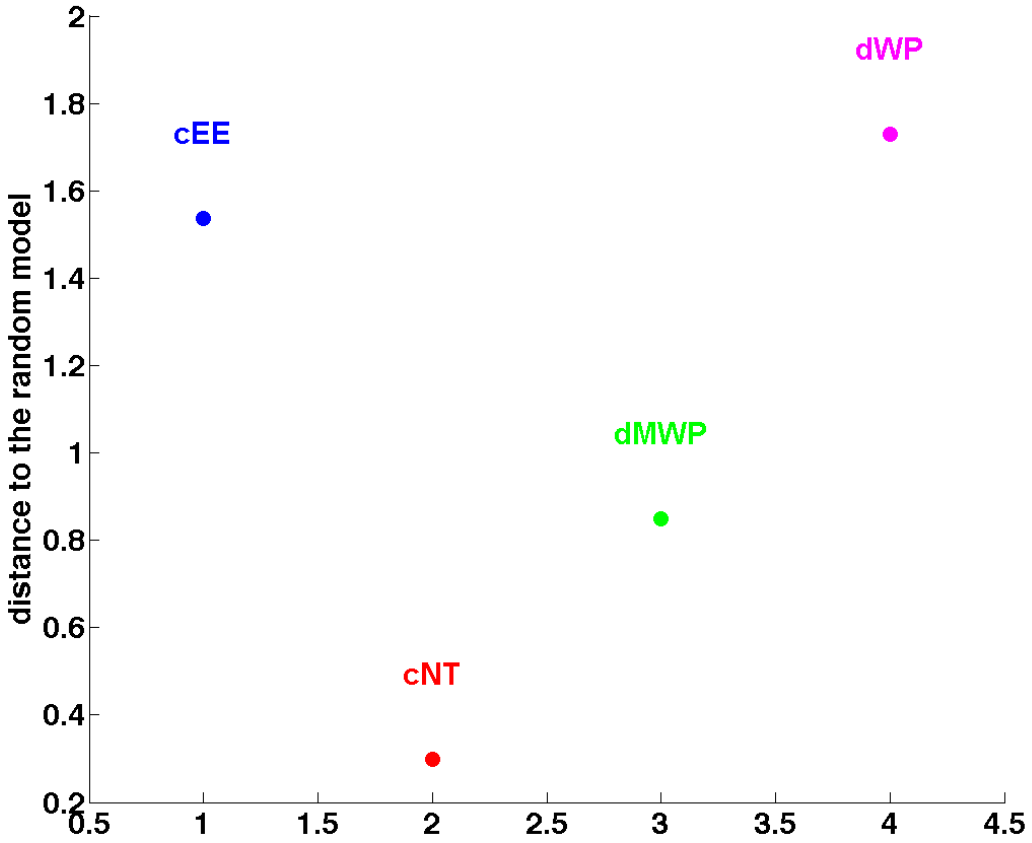


Figure 8.17: Topological distance to random null model as defined in 8.9. Magenta: Drosophila Wing Prepupa - Green: Drosophila Mutant Wing Prepupa - Red: Chick Neuroepithelium - Blue: Chick Ectoderm

8.4 Discussion and Conclusion

The use of persistent homology introduced in this article provides a new tool to quantify patterns in networks of cellular contacts. A set of features characterizing the topology of the network at several scales can be obtained automatically from raw data using exclusively topological information. Epithelia of different types have reproducible features which can be used to classify and quantify them. Although those features are less optimal than the degree distribution alone to discriminate different tissue types, they are useful for characterizing patterns at the level of groups of cells. These patterns at the level of groups of cells (neighborhood) are inadequately described with traditional features used in complex networks analysis because they are highly constrainted by the (quasi)-triangulated nature of the network of cellular contacts. Using a null model of random triangulated surface with arbitrary degree distribution we quantify the presence of high-order patterns in the network. The significance of their presence varies between the various epithelial types and can be related to the specifities of the underlying biological processes.

Using this methodology, we quantified the contribution of Myosin II and cellular cytoskeleton to the formation of spatial patterns at the level of groups of cells. Indeed, mutant specimens present a higher proximity to null model and thus less significant presence of ordered pattern than the wild type. However, we see that chicken neuroepithelium (cNT) is even closer to the null model. The role of Myosin II is therefore not completely determinant for the presence of spatial patterns and should be combined to other processes to explain the structure of epithelia. The different nature of epithelia, squamous, tubular, may also be relevant to understand the specificities of the spatial distribution of cells. Finally the sequence of events, such as cell proliferation, cell motility, and cell extrusion, may be the most significant contributions to the morphogenesis of epithelia. To test and quantify the contribution of the various biological processes, it seems necessary to systematically perform perturbation experiments leading to a quantification of ordered patterns in the resulting networks of cellular contacts.

In order to increase the applicability of the proposed method, the approach could be generalized to characterize patterns and organization in 3D structures within epithelia without limits *a priori*. Dynamical evolution of epithelia may be a more difficult problem since the number and structure of the simplicial complex associated to the network of cellular contacts may vary from one time step to the other with cell motility and cell proliferation. Finally developing the null model into a more analytical object could lead to the

characterization of statistical distribution of spatial networks, and for example the computation of p-value for the ordered nature of the organization of the network of cellular contacts.

Chapter 9

Tissue shape dynamics: cell proliferation and cell displacements

Abstract *This chapter proposes some perspectives for the characterization, quantification and comparison of developing tissues as evolving networks. It is shown that time should not be considered as simple dimension since the nature of the network is changed through its flow. An adaptation of the persistent algorithm is proposed using genealogy and relatedness as a parameter. Its use should highlight the intertwining between cell proliferation and cell displacement.*

9.1 Introduction

In the previous chapter, we have proposed an approach characterizing the network of cellular contacts from static images. This approach was able to uncover specific higher-order constraints in the spatial distribution of cells that we ascribe to developmental histories. However, to confirm these assumptions and describe morphogenesis, it would be necessary to characterize dynamically the evolution of these networks. We propose in this chapter some perspectives in this direction.

A previous study which has gained much attention characterizes the dynamic nature of epithelial organization by means of cell division [80], [81]. A simple model describes the evolution of the cellular contacts as shown on figure 9.1. This model leads to the establishment of a markov chain whose evolution converges toward a distribution of polygons (degree distribution) found in various epithelia (Drosophila, Hydra and Xenopus).

However the distribution of polygons found in [80] is not as universal as expected and

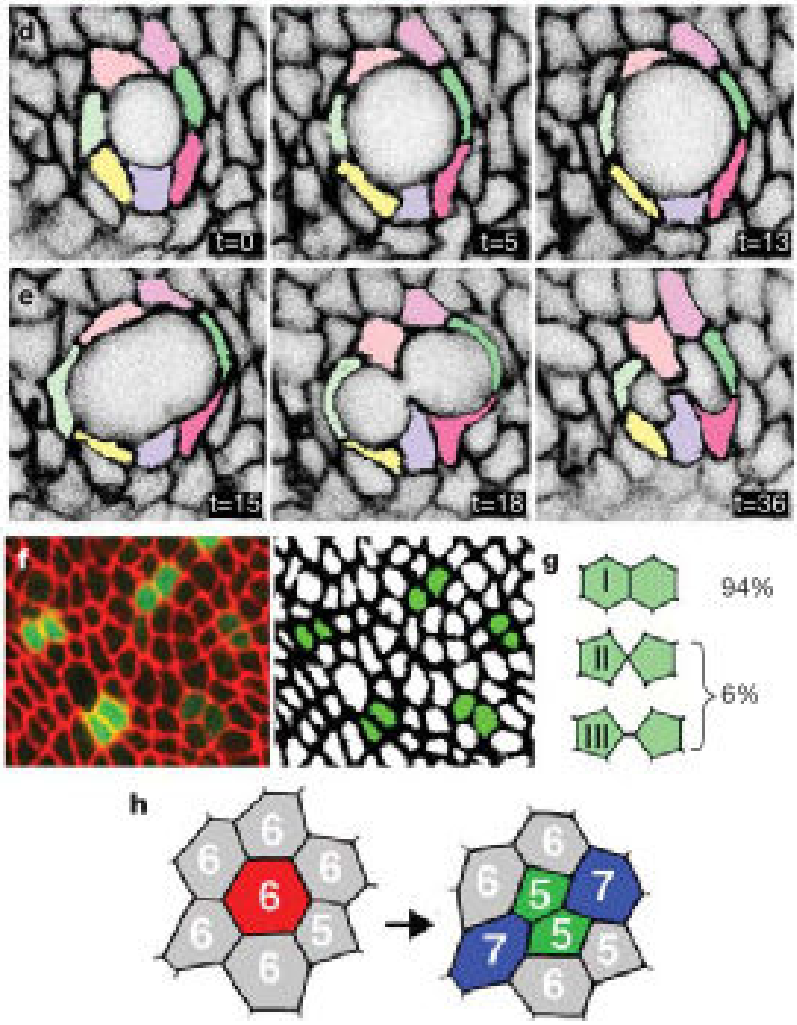


Figure 9.1: Empirical evidence underlying the establishment of the model of cell division in an epithelial tissue in [80] "d, Dilation of the junctional lattice permits rounding of a seven-sided mitotic cell during stages corresponding to prophase–metaphase. Owing to compression and stretching of the pseudocoloured neighbours, no cell-neighbour exchanges occur (n = 18 dilating cells). Units of t are in minutes. e, During stages corresponding to anaphase through cytokinesis, local topology (connectivity between cells) remains unchanged; the mitotic cell approaches abscission surrounded by the same cohort of seven neighbours (n = 23 cytokinetic cells). f, Two-cell clones marked by heritable expression of GFP (green) imaged at the level of the septate junctions stained with anti-Dlg (red). g, In approximately 94% of cell divisions, cytokinesis resolves with formation of a new cell interface, resulting in the type I conformation of mitotic siblings. h, Summary diagram of topology changes during cell division."

since the publication of this article, several studies have proposed alternative approaches taking into account other mechanisms such as the mechanical constraints on cleavage

patterns, or cell displacements to explain shifts from this theoretical distribution [192], [4], [167], [221], [82], [219], [190].

The original model proposed by Gibson *et al.* [80] rely on the geometric properties of the cells in an epithelium. It postulates that the cells don't move and that the orientation of division is chosen randomly. These two hypotheses seem realistic and form the basis of the model, they can be modulated depending on the modeling context, for example if the polarity of the cell division is highly constrained or if the cells are highly mobile. In addition to this two hypotheses, the model use a *mean field* assumption. Under this assumption, it is postulated that the cells divide synchronously and that each cell gain one new contact from a neighboring dividing cell in average. This last assumption seem to be the most difficult, in particular in development, because epithelial organization is far from an equilibrium or asymptotic state. As shown in the previous chapter, patterns involving groups of cells are present in significant proportions in epithelia and the underlying topology constrains the network. Neighborhoods cannot be considered uniformly with a mean field assumption, they are biased by these patterns.

As shown in the previous chapter, when using a measure of organization taking into account higher order spatial structures, we find that the developmental sequence of events has had an influence on the resulting shape of the tissue. We will present in the next sections a measure considering at the same time the spatial organization and the history of the tissue after a short review of already existing measures.

9.2 Time evolving networks

9.2.1 Time evolution of static measurements

The most straightforward approach consists in observing the dynamical evolution of measures defined on static networks. For example, we can look at the average degree in an evolving network. The average and standard deviation of the degree distribution obtained from the network of cellular contacts in one of the digitally reconstructed sea urchin presented in Chapter 1 of this dissertation is shown on figure 9.2. This measure shows a tendency towards 6 neighbors in average and an increasing standard deviation.

Although very useful, this kind of approach can only describe the system at a macroscopic level since the detail of individual cell trajectories is not taken into account. Relating this measure with individual cell dynamics require a model, such as [80].

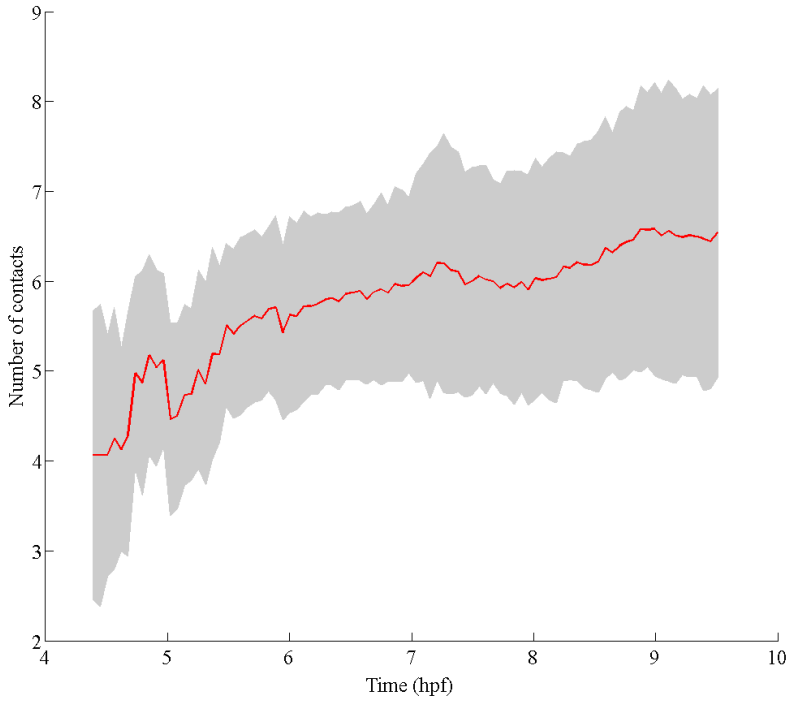


Figure 9.2: Evolution of the mean number of neighbors (red) and standard deviation (grey) through time in the sea urchin development in one of the 5 specimens studied in chapter 1

Moreover, we have seen in the previous chapter that the degree distribution can be misleading and doesn't describe the spatial distribution of cells with respect to each other. Similar to the evolution of the degree distribution in time, we can measure the evolution of the topological signatures of epithelial organization obtained in the previous chapter. At a given time step, the algorithm defined in the previous chapter can be applied to the network of cellular contacts. The result is a barcode, which is a mathematical object describing the evolution of topological features regarding varying values of a parameter. A barcode contains bars representing generators of i-dimensional "holes". The value of the parameter where these bars appear and disappear are the birth and death parameter value. A barcode can be represented as a persistence diagram where the coordinate axis are birth and death of these bars. A generalization of the barcode for time varying system is the concept of Vineyard as represented on figure 9.3 adapted from [155], [46]. The points in the persistence diagram are extended to "spaghetti" or "vine" according to a temporal dimension. This construction has been successfully used for the study of evolving set of sensors. The main limitation of this approach is that vineyards require a continuity of the set in

time, meaning that mitoses or cell extrusion should be excluded in our case. Moreover, this object is not completely stable and computable.

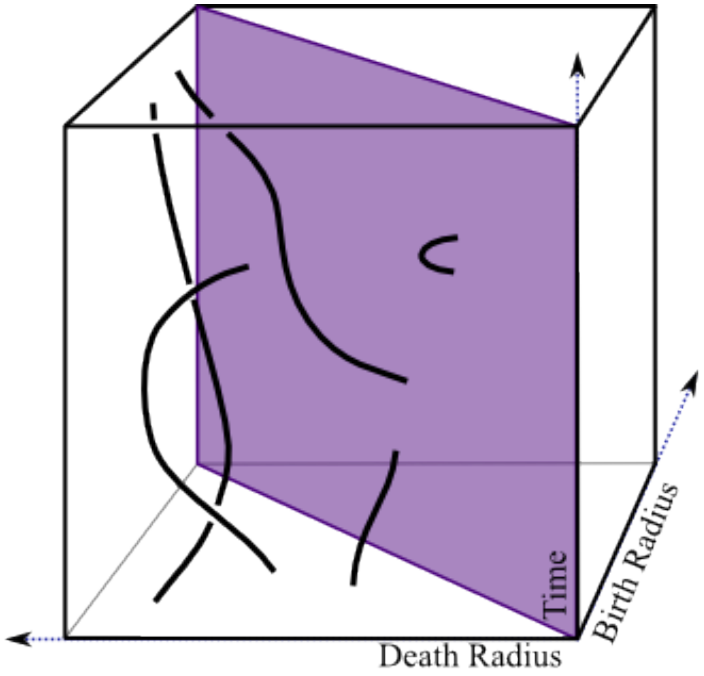


Figure 9.3: Vineyard as mathematical characterization of an evolving topological space. The persistence diagram is updated at each time step - Image adapted from [155]

We see that extension of measure of static networks in time is appealing but not completely efficient to solve our problem.

9.2.2 Looking at spatiotemporal networks

Another line of thoughts consists in considering time evolution as another parameter of the space under consideration.

In algebraic topology, at least two possibilities have been developed. The first one is multidimensional persistence. This is the generalization of persistent homology to two or n- parameters [39]. The main problem that appears with this approach is the fact that it has been shown that there is no simple summary like the barcode for 1-dimensional persistence [223]. Moreover, as for the vineyards, the multi-dimensional approach assume an identical underlying topological space in each direction. Yet, in our case, cells are proliferating, thus changing the nature of the underlying topological space.

The second possibility stems from a collaboration between the mathematician Edelsbrunner and the biologist Heisenberg [63]. They propose the concept of the Medusa to characterize the phenomenon of cell sorting during development, with an approach considering space and time as two similar dimensions. Cell sorting is a phenomenon where two types of cells that are initial mixed segregate into two spatially distinct populations. This mathematical construction seem to be relevant for a problem where the parts are moving in time. However, they don't take the proliferation dynamics into consideration.

9.3 Using genealogy as a parameter - historical dependency of shape

As already suggested in [172] and in the previous chapter, the combination of cell proliferation and cell displacement produces the specific organization of the network of cellular contact. A good measure of this process should integrate these two phenomena. A solution can be found by considering the relatedness between cells as the criteria to construct the sequence of nested spaces in the filtration.

Given a network (V,E) of cellular contacts, we can define relations among cells. Let's define a function R that returns the degree of relatedness of two cells. If two cells $(i,j) \in V^2$ are sisters, i.e. they share the same mother cell, we will write $R(i,j) = 1$. If the two cells $(i,j) \in V^2$ are cousins, i.e. they share the same great mother cell, we will write $R(i,j) = 2$. We can notice that $R(i,j) = 1 \Rightarrow R(i,j) = 2$, but the opposite is not necessarily true. Similarly, if two cells $(i,j) \in V^2$ share the same great great mother, we will write $R(i,j) = 3$, with $R(i,j) = 1 \Rightarrow R(i,j) = 2 \Rightarrow R(i,j) = 3$. R is defined in the same way for higher degree of relatedness. If no ancestor is known for a couple of cells $(i,j) \in V^2$, we write $R(i,j) = 0$. An illustration of this function is shown on figure 9.4.

In a similar way as in Chapter 8, we can use the sub level sets of this function R to define a filtration $\underline{\text{Fil}} = \{\text{Fil}(k)\}_{k\in\mathbb{N}}$ on the network of cellular contacts:

$$\text{Fil}(k) = (\{v_i \in V\}, \{e = \{v_i, v_j\} \in E : R(v_i, v_j) \leq k\}) \quad \forall k \in \mathbb{N} \tag{9.1}$$

This filtration contains for each value of the parameter k, all the vertices of the network. The edges are included in the filtration with a condition on relatedness. The higher degree of relatedness is considered, the more complete is the filtration. By taking the clique complexes $\text{Fil}^*(k)$ on this filtration for each value of k as described in Chapter 8 section 8.2.3,

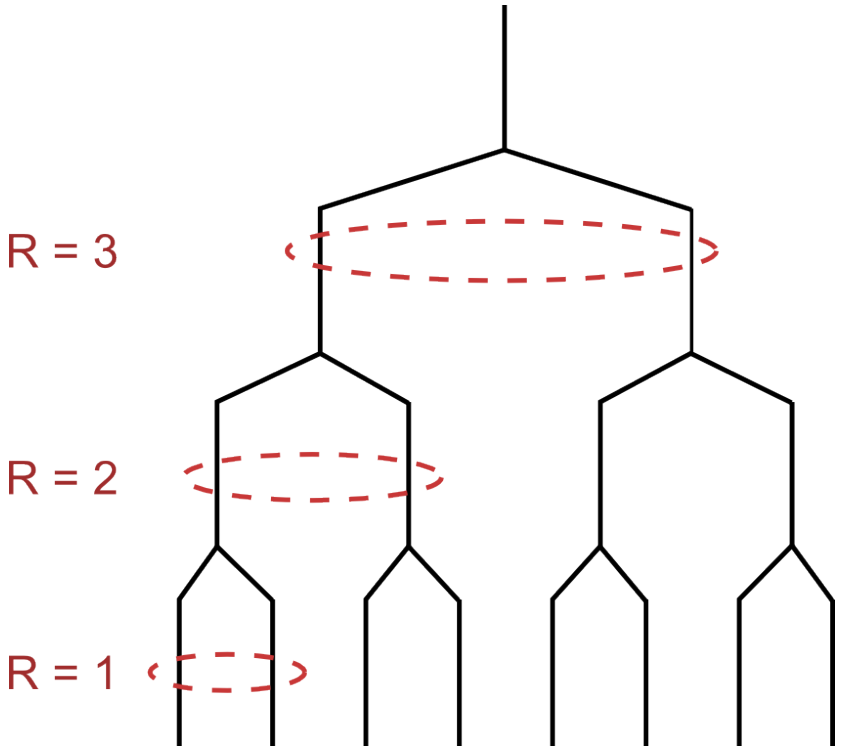


Figure 9.4: Visualization of the function R that returns the degree of relatedness of two cells. A cell lineage is represented in black. When considering cells at the bottom of the tree, any couple of cells in the branches below the red circles have relatedness indicated by the value of R

we obtain the filtration $\underline{\text{Fil}}^*$ on which we can apply persistent homology. In that case, persistent homology algorithm will measure the mixing of cells through displacements and proliferation.

For example if cells only divide with no displacement in a tissue, then the algorithm will detect a decreasing number of connected components and no holes. Whereas, if the cells are highly mixed, the number of connected components will decrease less rapidly and the number of holes will increase.

This measure is a first strong step toward solving the problem of the quantification of the intertwining between cell proliferation and cell displacement in an evolving network of cellular contact.

9.4 Conclusion

The perspectives presented in this chapter show that the problem of understanding the spatial organization of epithelia in the light of cell displacement and cell proliferation is interesting and require to be studied. The simple extension of static characterization of networks in a dynamical framework is not that simple. Several problems may arise from the use of another dimension when considering topological characterization. The proliferation undergone by the cells raises numerous problems as the topological nature of the tissue is changed in time. It is a difficult problem to match two simplicial complexes, it is much harder when the number of elements vary in time.

However, we present in the last section an appealing perspective that would allow to combine cell lineage and cell spatial organization. This measure which integrate cell proliferation and cell movements should help to understand the unfolding of tissue shape during morphogenesis.

Conclusion

The main goal of this part was to uncover the relations between spatial organization of tissue and development in epithelia, which is one of the most simple tissular structure. This approach which tries to characterize invariant features among embryo of different species and between mutant and normal development is intended to reveal some universal pattern in the shape of embryo. Shape in a developing organism is dependent on the specific sequence of events occurring during its developmental trajectory.

The first problem raised by such aims is the question of finding a generic way of describing tissues. The network of cellular contacts is a good candidate. It is composed by nodes corresponding to the cells and by edges corresponding to their contacts. It is a purely combinatorial structure and is thus not affected by questions of size. Moreover, it allows to consider the level of individual cells in relation to the whole tissue.

Descriptive tools developed in previous studies investigating this question show some limitations because they neglect the high constraints of the underlying tissue topology. We proposed to reveal underlying topological invariants using the mathematical framework of persistent homology. This mathematical framework has been proved to be relevant in the study of the structure of the cosmic web which is a historical structure in the sense that its establishment involve random events at multiple scales and cannot be reduced to a small number of parameters.

To apply the mathematical framework of persistent homology on images of epithelia, we defined for each sample of tissue two filtrations using the number of neighbors as a proxy for density in the tissue. These filtrations, which are nested sequences of simplicial complexes, are the basis for the establishment of a multi-scale topological signature of the tissue. This signature can then be used to compare and classify tissue between samples from *Drosophila* and *Chick* embryos. Signature of common tissue types are clustered, suggesting the possibility of an automatic classification of embryo images.

The significance of these measures is assessed by comparing the topological signa-

ture of a null model of random surface with arbitrary degree distribution. This model constrain local features, degree distribution, while letting the connection between nodes unconstrained. The construction process ensure that the resulting network can be embeddable in the plane, a realistic topological constraint. By comparing empirical samples with their random counterparts, randoms network with similar degree distribution, we show that the different types of tissues present different "degrees of randomness" in their spatial organization. Low "degree of randomness" reflects the significant presence of patterns involving groups of cells. These differences between tissues reflect the type of epithelium, squamous or tubular, and the presence of molecular determinant of shape such as Myosin II. We conjecture that these patterns of groups of cells are the traces of the sequence of developmental events.

Measures combining spatial organization and developmental history of a tissue are discussed in Chapter 9. It is shown that the dynamical study of measures defined on static networks doesn't seem sufficiently informative to understand the effect of individual cell histories without the use of models. Looking at the evolving network as a spatio-temporal structure is an appealing direction. However, the proliferative nature of the cells change the nature of the topological space under investigation in time, leading to difficult mathematical formalization. Finally, we propose a way to combine genealogical information with spatial organization by adapting the persistent homology approach defined in Chapter 8. This approach which uses the cell lineage should reveal the interaction between cell displacement and cell proliferation as the origin of tissular organization.

Altogether, we see that, even for a simple structure as an epithelium, the multi-scale nature of biological objects raises challenges that require the design of specific mathematical approaches. These mathematical approaches should tackle the question of *organization* and *developmental history* of organisms.

General conclusion

The approach developed throughout this work is a combination of theoretical, mathematical and experimental considerations concerning the measure and interpretation of variability in animal embryogenesis. This very general subject has led to a certain number of more specific studies, the main contributions are the following:

- ⋆ In Part I, we considered the question of the reproducibility of development. The study of intra- and inter-individual variability in a small cohort of digitally reconstructed sea urchins has led to the establishment of **multi-level data-driven probabilistic model** relating variable cell features with reproducible embryo-level dynamics. This model forms the basis for a **prototypical representation of the sea urchin development** as the centroid of the cohort, based on empirical parameters estimation. Surprisingly, variability at the individual cell level plays a significant role in the establishment of reproducible embryo-level dynamics, weakening the traditional view of the development as a finely-tuned process at the individual cell level.
- ⋆ In Part II, we considered the question of the diversity of phenotypes in the living, or how variations in individual developments affect the space of possible. The main contribution consists in using **a formal analogy with the mathematical framework of quantum mechanics to clarify the various levels of randomness in biology**. This approach enables to account for the high variability observed in the zebrafish *squint* mutant line and classical Mendelian scheme of inheritance. The latter is shown to be **formally analog to quantum entanglement**. We also discuss the relations between variability in biology and randomness in physical and mathematical theories.
- ⋆ In Part III, we study epithelial structure using the network of cellular contacts. By using **topological invariants and a generalized model of random network**, we show the presence of patterns involving groups of cells. We assume that **these patterns result from the sequence of events involved in the morphogenesis of tissues**. We propose a measure to characterize spatial organization which incorporates lineage

relationships between cells.

****

This work is part of a larger movement introducing quantitative approaches and computation in biology. The wealth of data generated by new imaging technologies and ever increasing computational power has opened the way for new perspectives on the understanding of biological problems such as embryonic development or evolution. The quantitative turn undergone by biology in the last decade demands a reinterpretation of many biological concepts. These challenges are associated to developments and extensions of mathematical approaches which are required to handle and make sense of these, sometimes massive, data sets. However, these mathematical developments cannot be performed by themselves, they require an underlying theoretical and epistemological basis to keep them interpretable. These observations form the general framework within which this work has taken place.

The main concepts around which this thesis has evolved are the notions of *organization* and of *historicity* of organisms. These notions are closely related to *variability*, the central theme of this thesis. Indeed, organization is the result of an evolution, whose main ingredient is variation, on the other hand, historicity stems from this intrinsic variability which is ubiquitous in biology. The concept of organization expresses itself in the multi-scale nature of organisms involving dynamics that can be heterogeneous in nature. The concept of historicity expresses itself in development by the path-dependency of the process, meaning that the specific sequence of events has an effect on the final phenotype. Historicity in evolution is associated to the contingency of events resulting from variability at all scales in biological organisms. In this context, we have tried in the course of this dissertation to characterize mathematically some notions of variability at different scales, thus in relation to the concept of organization, and along the different stages of development in animal embryogenesis, thus in relation to historicity of development.

More precisely, in the comparative study of a small cohort of digitally reconstructed sea urchin embryos developed in chapters 1 and 2 we are confronted with these two notions of *organization* and *historicity*. Embryo-level developmental dynamics present similar patterns from one specimen to the other, whereas individual cell features are variable and spatial symmetries of the embryo prevent to compare individual cell between specimen of the cohort. This multi-scale nature of the development is captured by defining groups

of cells as an intermediate level of observation which takes advantage of these symmetries. The historicity of the developmental process is witnessed by the accumulation of variability and consequently the increase of the spread of distributions of individual cell features through time. This process underlies cell differentiation (chapter 3). These interesting relations between individual cell features and embryo-level dynamics shed new light on the sources of reproducibility. Where a finely-tuned process at the cellular level would have been expected, we find an averaged process over variable cell features.

The variety of mechanisms sources of variability in individual cells and of diversity in population of organisms is reviewed in chapter 4. This multiplicity of phenomena is the main obstacle for a proper characterization of biological variability. Their heterogeneity prevent their multi-scale integration. Indeed, we show that several concepts exist in mathematics and physics for the formalization of the idea of *randomness* which underlies frameworks modeling situations of uncertainty. These frameworks are not compatible as such. Two common properties are the need for an *a priori* space of possible and the use of an infinity to define this space of possible, or fluctuations in the case of chaotic systems. To avoid having to integrate these various forms of randomness, an alternative can be to use relevant simplifying assumptions on other levels of organization or on past history. The price being a reduced generality of the results.

An experiment presenting complex patterns of variability is presented in chapter 5. This experiment involves the *squint* mutant line of the zebrafish. The results of this experiment show in the progeny of homozygote couples of mutants a phenomenon of variable phenotypic expressivity and incomplete penetrance; from complete cyclopia to two well formed eyes. We obtain embryos with phenotypes from a list of phenotypes in unpredictable proportions, without being able to ensure the completeness of the list. These results suggest complex scheme of inheritance of maternal factors and/or complex causal processes during development, supported by the various possible sources of variability described in chapter 4.

The variability observed and measured in chapter 4 motivates the reflections developed in chapter 5. These reflections rely on an analogy with the quantum mechanics mathematical framework as an attempt to account for the various levels of variability observed in the living. The interest of this analogy is to be able to consider sets of probability distributions as vector spaces. The problem consists in clarifying the difference between variable phenotypes in a fixed list (which can be described with classical probability), new emerging phenotypes to be added to the list of already known phenotypes (which is best

described as a cartesian product) and the emergence of a new observable which require a new space of possible. These emergences of new phenotypes or new observables are derived from the historical nature of onto- and phylogenetic trajectories. In addition, this framework can account for complex relations between observables, with a tensor product coupling the space of possible corresponding to each of these observables. This property is used to account for Mendelian scheme of inheritance and is shown to be formally analog to quantum entanglement. We interpret it as a consequence of the organization of biological organisms preventing the coexistence of two phenotypes in the same individual.

To study the relations between variations in individual developments and patterns of diversification, we consider in chapter 6 the concept of an ontogenetic tree which enables to define a notion of developmental proximity. This approach is intended to compare a phenotypical similarity with the developmental one in order to quantify the influence of path dependency. Several assumptions are made concerning the canalization on possible phenotypes as a result of the constraints imposed by the already undergone developmental stages before diverging, this approach is underlied by an empirical data set describing numerous mutant zebrafish developments.

This second part on diversity and diversification involves many different situations. The organized and historical aspects of development drive and canalize diversification, while being at the same time the result of this evolutionary history. And if biological variability cannot be reduced to randomness as defined in physics and mathematics, this dependence to an evolutionary history may give it its specific status. Biological variability implies a stronger form of randomness, at the level of the observable and thus of the space of possible itself.

Finally in chapter 7 and 8 we turned to a more specific problem, the characterization of epithelial organization. Although epithelia are simple cellular structures compared to complete organisms, their presence in many species during embryonic development gives to this problem a high level of generality. Epithelial organization is formalized as a network of cellular contacts whose underlying topological invariants are found to enable comparison and classification over a wide range of specimens. The framework of persistent homology is used to define these topological invariants. By comparing these tissue shapes to a null random model, we found that they contain pattern involving small groups of cells in significant number. We interpret these patterns as resulting from the sequence of events involved in the morphogenesis of these tissues. To characterize the relation between ep-

ithelial shape and the developmental sequence we propose a methodology based on the cell lineage and cell network of cellular contacts. Overall, we show in this last part that organization and historicity are intertwined in morphogenesis.

Perspectives raised by this dissertation are hopefully numerous. We have made some suggestions in the course of the manuscript, particularly in chapters 3 and 9. However, here are some perspectives

The data-driven multi-level prototypical probabilistic model obtained in chapter 2 can serve as a basis for a dynamical system description of development, a law of evolution needs to be found to relate the value of the parameters. The approach is general enough to be expected to be found applicable in the development of other species where there is relative synchrony of cell cycles using the same mathematical framework. Moreover, the approach consisting in defining a relevant coarse-grained level of observation preventing averaging out significant information and being generic enough to enable comparison between specimens is likely to be transferable in other developing systems. As an extension of the model it could be interesting to combine the analysis of the cell lineage with a spatial analysis of the distribution of cells.

The probabilistic approach developed in the first part raises numerous questions on the reproducibility and robustness of development. An interesting line of research could be to perform experimental perturbations of the system to see if the mathematical structure uncovered stays the same. For example by modulating the length of the cell cycles.

In the second part, we have considered the diversity of the living. The main perspective here would be to be able to integrate the variability at all scales in a *theory of organisms*. Such a theory would enable to describe more formally this notion of biological randomness at the level of the space of possible. This would lead to better understanding of the concept of prediction in biology

In the third part, we have considered epithelial organization, and its relations to cell proliferation and displacement. It could be interesting to perform perturbations on developing systems, e.g. on cell proliferation or cell motility, to quantify their influence on the resulting shape. Another perspective consists in generalizing the approach to more complex tissues, for example during epithelial folding to describe the transition from two to three-dimension, by taking into account the curvature for example. This approach can also be generalized to three-dimensional situations.

Overall, we anticipate that the field of embryology will greatly benefit from an understanding of variation at every scales. This understanding requires quantitative and math-

ematical approaches to be grounded on stable settings. It will hopefully enable to link deeply development and evolution, while giving major roles to the notions of organism and of historicity of living objects.

Placing *variation* at the foundation of his theory of evolution is one of the major paradigm shifts operated by Charles Darwin. Our results suggest that we may need the same kind of revolution for developmental biology.

Bibliography

[1] Laurent Abouchar, Mariela D Petkova, Cynthia R Steinhardt, and Thomas Gregor. Precision and reproducibility of macroscopic developmental patterns. *arXiv preprint arXiv:1309.6273*, 2013.

[2] Henry Adams and Andrew Tausz. Javaplex tutorial, 2011.

[3] Aaron Adcock, Erik Carlsson, and Gunnar Carlsson. The ring of algebraic functions on persistence bar codes. 2012.

[4] Tinri Aegerter-Wilmsen, Alister C Smith, Alix J Christen, Christof M Aegerter, Ernst Hafen, and Konrad Basler. Exploring the effects of mechanical feedback on epithelial topology. *Development*, 137(3):499–506, 2010.

[5] Pere Alberch. Ontogenesis and morphological diversification. *American Zoologist*, 20(4):653–667, 1980.

[6] Pere Alberch, Stephen Jay Gould, George F Oster, and David B Wake. Size and shape in ontogeny and phylogeny. *Paleobiology*, pages 296–317, 1979.

[7] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002.

[8] David J Aldous. *Exchangeability and related topics*. Springer, 1985.

[9] VG Allfrey, R Faulkner, and AE Mirsky. Acetylation and methylation of histones and their possible role in the regulation of rna synthesis. *Proceedings of the National Academy of Sciences of the United States of America*, 51(5):786, 1964.

[10] Shun-ichi Amari and Hiroshi Nagaoka. *Methods of information geometry*, volume 191. American Mathematical Soc., 2007.

[11] Lynne M Angerer and Robert C Angerer. Animal–vegetal axis patterning mechanisms in the early sea urchin embryo. *Developmental biology*, 218(1):1–12, 2000.

[12] Masanari Asano, Irina Basieva, Andrei Khrennikov, Masanori Ohya, Yoshiharu Tanaka, and Ichiro Yamato. A model of epigenetic evolution based on theory of open quantum systems. *Systems and synthetic biology*, 7(4):161–173, 2013.

[13] Ricardo BR Azevedo, Rolf Lohaus, Volker Braun, Markus Gumbel, Muralikrishna Umamaheshwar, Paul-Michael Agapow, Wouter Houthoofd, Ute Platzer, Gaëtan Borgonie, Hans-Peter Meinzer, et al. The simplicity of metazoan cell lineages. *Nature*, 433(7022):152–156, 2005.

[14] Francis Bailly and Giuseppe Longo. Randomness and determinism in the interplay between the continuum and the discrete. *Mathematical Structures in Computer Science*, 17(02):289–305, 2007.

[15] Francis Bailly and Giuseppe Longo. *Mathematics and the natural sciences: the physical singularity of life*. Imperial College Press, 2011.

[16] Zhirong Bao, John I Murray, Thomas Boyle, Siew Loon Ooi, Matthew J Sandel, and Robert H Waterston. Automated cell lineage tracing in caenorhabditis elegans. *Proceedings of the National Academy of Sciences of the United States of America*, 103(8):2707–2712, 2006.

[17] Irina Basieva, Andrei Khrennikov, Masanori Ohya, and Ichiro Yamato. Quantum-like interference effect in gene expression: glucose-lactose destructive interference. *Systems and synthetic biology*, 5(1-2):59–68, 2011.

[18] Martin Behrndt, Guillaume Salbreux, Pedro Campinho, Robert Hauschild, Felix Oswald, Julia Roensch, Stephan W Grill, and Carl-Philipp Heisenberg. Forces driving epithelial spreading in zebrafish gastrulation. *Science*, 338(6104):257–260, 2012.

[19] Ingemar Bengtsson and Karol Zyczkowski. *Geometry of quantum states: an introduction to quantum entanglement*. Cambridge University Press, 2006.

[20] George D Birkhoff. Proof of the ergodic theorem. *Proceedings of the National Academy of Sciences*, 17(12):656–660, 1931.

[21] Michel Bitbol. *Mécanique quantique: une introduction philosophique*. Editions Flammarion, 1996.

[22] Michel Bitbol. La mécanique quantique comme théorie des probabilités généralisées. In Etienne Klein & Yves Sacquin, editor, *Prévision et Probabilité dans les Sciences*. Éditions Frontières, Paris, 1998.

[23] Michel Bitbol. *Théorie quantique et sciences humaines*. CNRS, 2009.

[24] Ivana Bjedov, Olivier Tenaillon, Benedicte Gerard, Valeria Souza, Erick Denamur, Miroslav Radman, François Taddei, and Ivan Matic. Stress-induced mutagenesis in bacteria. *Science*, 300(5624):1404–1409, 2003.

[25] J Todd Blankenship, Stephanie T Backovic, Justina SP Sanny, Ori Weitz, and Jennifer A Zallen. Multicellular rosette formation links planar cell polarity to tissue morphogenesis. *Developmental cell*, 11(4):459–470, 2006.

[26] S Bohn, S Douady, and Y Couder. Four sided domains in hierarchical space dividing patterns. *Physical review letters*, 94(5):054503, 2005.

[27] S Bohn, L Pauchard, and Y Couder. Hierarchical crack pattern as formed by successive domain divisions. *Physical Review E*, 71(4):046214, 2005.

[28] Floris Bosveld, Isabelle Bonnet, Boris Guirao, Sham Tlili, Zhimin Wang, Ambre Petitalot, Raphaël Marchand, Pierre-Luc Bardet, Philippe Marcq, François Graner, et al. Mechanical control of morphogenesis by fat/dachsous/four-jointed planar cell polarity pathway. *Science*, 336(6082):724–727, 2012.

[29] Yvonne Bradford, Tom Conlin, Nathan Dunn, David Fashena, Ken Frazer, Douglas G Howe, Jonathan Knight, Prita Mani, Ryan Martin, Sierra AT Moxon, et al. Zfin: enhancements and updates to the zebrafish model organism database. *Nucleic acids research*, 39(suppl 1):D822–D829, 2011.

[30] Ingo Brigandt. From developmental constraint to evolvability: how concepts figure in explanation and disciplinary identity. In *Conceptual Change in Biology*, pages 305–325. Springer, 2015.

[31] Elissa Briggs and Gary M Wessel. In the beginning... animal fertilization and sea urchin development. *Developmental biology*, 300(1):15–26, 2006.

[32] PA Bromiley. Products and convolutions of gaussian distributions. *Medical School, Univ. Manchester, Manchester, UK, Tech. Rep*, 3:2003, 2003.

[33] Marcello Buiatti and Giuseppe Longo. Randomness and multilevel interactions in biology. *Theory in Biosciences*, 132(3):139–158, 2013.

[34] Cristian S Calude, Michael J Dinneen, Chi-Kou Shu, et al. Computing a glimpse of randomness. *Experimental Mathematics*, 11(3):361–370, 2002.

[35] CS Calude and G Longo. Classical, quantum and biological randomness as relative incomputability. Technical report, Department of Computer Science, The University of Auckland, New Zealand, 2014.

[36] Örjan Carlborg and Chris S Haley. Epistasis: too often neglected in complex trait studies? *Nature Reviews Genetics*, 5(8):618–625, 2004.

[37] Gunnar Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, 2009.

[38] Gunnar Carlsson. Topological pattern recognition for point cloud data. *Acta Numerica*, 23:289–368, 5 2014.

[39] Gunnar Carlsson and Afra Zomorodian. The theory of multidimensional persistence. *Discrete & Computational Geometry*, 42(1):71–93, 2009.

[40] George Casella and Roger L Berger. *Statistical inference*, volume 70. Duxbury Press Belmont, CA, 1990.

[41] Carlos Castro-González, Miguel A Luengo-Oroz, Louise Duloquin, Thierry Savy, Barbara Rizzi, Sophie Desnoulez, René Doursat, Yannick L Kergosien, María J Ledesma-Carbayo, Paul Bourgine, et al. A digital framework to build, visualize and analyze a gene expression atlas with cellular resolution in zebrafish early embryogenesis. *PLoS computational biology*, 10(6):e1003670, 2014.

[42] Hannah H Chang, Martin Hemberg, Mauricio Barahona, Donald E Ingber, and Sui Huang. Transcriptome-wide noise controls lineage choice in mammalian progenitor cells. *Nature*, 453(7194):544–547, 2008.

[43] Frédéric Chazal, David Cohen-Steiner, Leonidas J Guibas, Facundo Mémoli, and Steve Y Oudot. Gromov-hausdorff stable signatures for shapes using persistence. In *Computer Graphics Forum*, volume 28, pages 1393–1403. Wiley Online Library, 2009.

[44] Yu Chen and Alexander F Schier. The zebrafish nodal signal squint functions as a morphogen. *Nature*, 411(6837):607–610, 2001.

[45] Yuan Shih Chow and Henry Teicher. *Probability theory: independence, interchangeability, martingales*. Springer, 2003.

[46] David Cohen-Steiner, Herbert Edelsbrunner, and Dmitriy Morozov. Vines and vineyards by updating persistence in linear time. In *Proceedings of the twenty-second annual symposium on Computational geometry*, pages 119–126. ACM, 2006.

[47] Claude Cohen-Tannoudji, Bernard Diu, and Frank Laloë. *Quantum Mechanics. Vol. I & II*. Wiley, New-York, 1991.

[48] L da F Costa, Francisco A Rodrigues, Gonzalo Travieso, and Paulino Ribeiro Villas Boas. Characterization of complex networks: A survey of measurements. *Advances in Physics*, 56(1):167–242, 2007.

[49] Charles Darwin. *On the origins of species by means of natural selection*. John Murray, 1859.

[50] E H Davidson, R a Cameron, and a Ransick. Specification of cell fate in the sea urchin embryo: summary and some proposed mechanisms. *Development*, 125(17):3269–90, September 1998.

[51] Eric H Davidson. Lineage-specific gene expression and the regulative capacities of the sea urchin embryo: a proposed mechanism. *Development*, 105(3):421–445, 1989.

[52] Eric H Davidson. *Gene activity in early development*. Elsevier, 2012.

[53] Eric H Davidson, Jonathan P Rast, Paola Oliveri, Andrew Ransick, Cristina Calestani, Chiou-Hwa Yuh, Takuya Minokawa, Gabriele Amore, Veronica Hinman, Cesar Arenas-Mena, et al. A genomic regulatory network for development. *Science*, 295(5560):1669–1678, 2002.

[54] Vincent Debat and Patrice David. Mapping phenotypes: canalization, plasticity and developmental stability. *Trends in Ecology & Evolution*, 16(10):555–561, 2001.

[55] Julien Delile, René Doursat, and Nadine Peyriéras. Computational modeling and simulation of animal early embryogenesis with the mecagen platform. In A Kriete and R Eils, editors, *Computational Systems Biology, 2nd edition.*, pages 359–405. Academic Press, Elsevier, 2013.

[56] Julien Delile, Herrmann Matthieu, Nadine Peyriéras, and René Doursat. Mecagen: A cell-based computational model of embryogenesis coupling mechanical behavior and gene regulation. *submitted*, 2015.

[57] Winfried Denk, James H Strickler, and Watt W Webb. Two-photon laser scanning fluorescence microscopy. *Science*, 248(4951):73–76, 1990.

[58] Persi Diaconis and David Freedman. Finite exchangeable sequences. *The Annals of Probability*, pages 745–764, 1980.

[59] Julien O Dubuis, Reba Samanta, and Thomas Gregor. Accurate measurements of dynamics and reproducibility in small genetic networks. *Molecular systems biology*, 9(1), 2013.

[60] Rosalie E Langelan Duncan and Arthur H Whiteley. The echinoid mitotic gradient: effect of cell size on the micromere cleavage cycle. *Molecular reproduction and development*, 78(10-11):868–878, 2011.

[61] Fabien Duveau and Marie-Anne Félix. Role of pleiotropy in the evolution of a cryptic developmental variation in caenorhabditis elegans. *PLoS biology*, 10(1):e1001230, 2012.

[62] Herbert Edelsbrunner and John Harer. *Computational topology: an introduction*. American Mathematical Soc., 2010.

[63] Herbert Edelsbrunner, Carl-Philipp Heisenberg, Michael Kerber, and Gabriel Krens. The medusa of spatial sorting: topological construction. *arXiv preprint arXiv:1207.6474*, 2012.

[64] Albert Einstein, Boris Podolsky, and Nathan Rosen. Can quantum-mechanical description of physical reality be considered complete? *Physical review*, 47(10):777, 1935.

[65] Michael B Elowitz, Arnold J Levine, Eric D Siggia, and Peter S Swain. Stochastic gene expression in a single cell. *Science*, 297(5584):1183–1186, 2002.

[66] Susan G Ernst. A century of sea urchin development. *American zoologist*, 37(3):250–259, 1997.

[67] Luis M Escudero, Luciano da F Costa, Anna Kicheva, James Briscoe, Matthew Freeman, and M Madan Babu. Epithelial organisation revealed by a network of cellular contacts. *Nature communications*, 2:526, 2011.

[68] Tom Evans, Eric T Rosenthal, Jim Youngblom, Dan Distel, and Tim Hunt. Cyclin: a protein specified by maternal mrna in sea urchin eggs that is destroyed at each cleavage division. *Cell*, 33(2):389–396, 1983.

[69] Dimitri Fabrèges. Phenotypic variation and resilience in sea urchin morphogenesis. In Edgar Raymond Banks, editor, *Sea Urchins: Habitat, Embryonic Development and Importance in the Environment*. Nova Science Publishers Inc, 2014.

[70] Emmanuel Faure*, Thierry Savy*, Barbara Rizzi*, Camilo Melani*, Mariana Remesikova*, Robert Spir*, Olga Drblikova*, Robert Cunderlik*, Gaëlle Recher*, Benoît Lombardot*, Dimitri Fabrèges*, Mark Hammons, Louise Duloquin, Ingrid Colin, Jozef Kollar, Sophie Desnoulez, Pierre Affaticati, Benoît Maury, Adeline Boyreau, Jean-Yves Nief, Pascal Calvat, Philippe Vernier, Monique Frain, Georges Lutfall, Yannick Kergosien, Pierre Suret, René Doursat, Alessandro Sarti, Karol Mikula, Nadine

Peyriéras, and Paul Bourgine. An algorithmic workflow for the automated processing of 3d+time microscopy images of developing organisms and the reconstruction of their cell lineage. *in revision*, 2015.

[71] Benjamin Feldman, Michael A Gates, Elizabeth S Egan, Scott T Dougan, Gabriela Rennebeck, Howard I Sirotkin, Alexander F Schier, and William S Talbot. Zebrafish organizer development and germ-layer formation require nodal-related signals. *Nature*, 395(6698):181–185, 1998.

[72] Lauren Figard, Heng Xu, Hernan G Garcia, Ido Golding, and Anna Marie Sokac. The plasma membrane flattens out to fuel cell-surface growth during drosophila cellularization. *Developmental cell*, 27(6):648–655, 2013.

[73] Vincent Fleury. Clarifying tetrapod embryogenesis, a physicist's point of view. *The European Physical Journal Applied Physics*, 45(03):30101, 2009.

[74] Vincent Fleury. Clarifying tetrapod embryogenesis by a dorso-ventral analysis of the tissue flows during early stages of chicken development. *Biosystems*, 109(3):460–474, 2012.

[75] Gabor Forgacs and Stuart A Newman. *Biological physics of the developing embryo*. Cambridge University Press, 2005.

[76] Evelyn Fox Keller. *Le siècle du gène*. Gallimard, 2003.

[77] Roman Frigg, Joseph Berkovitz, and Fred Kronz. The ergodic hierarchy. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Summer 2014 edition, 2014.

[78] Stefano Galatolo, Mathieu Hoyrup, and Cristóbal Rojas. Effective symbolic dynamics, random points, statistical behavior, complexity and entropy. *Information and Computation*, 208(1):23–41, 2010.

[79] Nicholas Geard, Seth Bullock, Rolf Lohaus, Ricardo BR Azevedo, and Janet Wiles. Developmental motifs reveal complex structure in cell lineages. *Complexity*, 16(4):48–57, 2011.

[80] Matthew C Gibson, Ankit B Patel, Radhika Nagpal, and Norbert Perrimon. The emergence of geometric order in proliferating metazoan epithelia. *Nature*, 442(7106):1038–1041, 2006.

[81] William T Gibson and Matthew C Gibson. Cell topology, geometry, and morphogenesis in proliferating epithelia. *Current topics in developmental biology*, 89:87–114, 2009.

[82] William T Gibson, James H Veldhuis, Boris Rubinstein, Heather N Cartwright, Norbert Perrimon, G Wayne Brodland, Radhika Nagpal, and Matthew C Gibson. Control of the mitotic cleavage plane by local epithelial topology. *Cell*, 144(3):427–438, 2011.

[83] SF. Gilbert. *Developmental Biology. 6th edition*. Sinauer Associates, 2000.

[84] Donald Gillies. *Philosophical theories of probability*. Routledge, 2000.

[85] Johann Wolfgang von Goethe. *Versuch die Metamorphose der Pflanzen zu erklären*. Gotha, Ettingersche Buchhandlung, 1790.

[86] S.J. Gould. *Wonderful Life: The Burgess Shale and the Nature of History*. W.W. Norton, 1989.

[87] Stephen Jay Gould. *Ontogeny and phylogeny*. Harvard University Press, 1977.

[88] François Graner and James A Glazier. Simulation of biological cell sorting using a two-dimensional extended potts model. *Physical review letters*, 69(13):2013, 1992.

[89] Jonathan L Gross and Thomas W Tucker. *Topological graph theory*. Courier Dover Publications, 2001.

[90] Charlène Guillot and Thomas Lecuit. Mechanics of epithelial tissue homeostasis and morphogenesis. *Science*, 340(6137):1185–1189, 2013.

[91] Ian Hacking. *L'émergence de la probabilité*. Seuil, 2002.

[92] Pascal Haffter, Michael Granato, Michael Brand, Mary C Mullins, Matthias Hammerschmidt, Donald A Kane, Jörg Odenthal, FJ Van Eeden, Yun-Jin Jiang, Carl-Philipp Heisenberg, et al. The identification of genes with unique and essential functions in the development of the zebrafish, danio rerio. *Development*, 123(1):1–36, 1996.

[93] Jean Harthong. *Probabilités & statistiques. De l'intuition aux applications*. Diderot éditeur, collection Arts et Sciences, 1996.

[94] Bing He, Konstantin Doubrovinski, Oleg Polyakov, and Eric Wieschaus. Apical constriction drives tissue-scale hydrodynamic flow to mediate cell elongation. *Nature*, 508(7496):392–396, 2014.

[95] Thomas Heams. Randomness in biology. *Mathematical Structures in Computer Science*, 24(03):e240308, 2014.

[96] Carl-Philipp Heisenberg and Yohanns Bellaïche. Forces in tissue morphogenesis and patterning. *Cell*, 153(5):948–962, 2013.

[97] Fritjof Helmchen and Winfried Denk. Deep tissue two-photon microscopy. *Nature methods*, 2(12):932–940, 2005.

[98] Andreas Hilfinger and Johan Paulsson. Separating intrinsic from extrinsic fluctuations in dynamic biological systems. *Proceedings of the National Academy of Sciences*, 108(29):12167–12172, 2011.

[99] Mae-Wan Ho. An exercise in rational taxonomy. *Journal of theoretical biology*, 147(1):43–57, 1990.

[100] Mae-Wan Ho, Alistair Matheson, Peter T Saunders, Brian C Goodwin, and Anna Smallcombe. Ether-induced segmentation disturbances in drosophila melanogaster. *Roux's archives of developmental biology*, 196(8):511–521, 1987.

[101] Mae-Wan Ho and Peter T Saunders. Rational taxonomy and the natural system. *Acta Biotheoretica*, 41(4):289–304, 1993.

[102] H Robert Horvitz and John E Sulston. Isolation and genetic characterization of cell-lineage mutants of the nematode caenorhabditis elegans. *Genetics*, 96(2):435–454, 1980.

[103] Sui Huang. Non-genetic heterogeneity of cells in development: more than just noise. *Development*, 136(23):3853–3862, 2009.

[104] Sui Huang. The molecular and mathematical basis of waddington's epigenetic landscape: A framework for post-darwinian biology? *BioEssays*, 34(2):149–157, 2012.

[105] Dann Huh and Johan Paulsson. Random partitioning of molecules at cell division. *Proceedings of the National Academy of Sciences*, 108(36):15004–15009, 2011.

[106] Julian Huxley et al. Evolution. the modern synthesis. *Evolution. The Modern Synthesis.*, 1942.

[107] Naoki Irie and Shigeru Kuratani. Comparative transcriptome analysis reveals vertebrate phylotypic period during organogenesis. *Nature communications*, 2:248, 2011.

[108] S Ishihara, K Sugimura, SJ Cox, I Bonnet, Y Bellaïche, and F Graner. Comparative study of non-invasive force and stress inference methods in tissue. *The European Physical Journal E*, 36(4):1–13, 2013.

[109] Shuji Ishihara and Kaoru Sugimura. Bayesian inference of force dynamics during morphogenesis. *Journal of theoretical biology*, 313:201–211, 2012.

[110] Eva Jablonka and Marion J Lamb. The changing concept of epigenetics. *Annals of the New York Academy of Sciences*, 981(1):82–96, 2002.

[111] François Jacob. *La logique du vivant: une histoire de l'hérédité*. Gallimard, 1987.

[112] François Jacob and Jacques Monod. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of molecular biology*, 3(3):318–356, 1961.

[113] William R Jeffery. Cavefish as a model system in evolutionary developmental biology. *Developmental biology*, 231(1):1–12, 2001.

[114] Oliver Johnson. *Information theory and the central limit theorem*, volume 8. World Scientific, 2004.

[115] Alex T Kalinka, Karolina M Varga, Dave T Gerrard, Stephan Preibisch, David L Corcoran, Julia Jarrells, Uwe Ohler, Casey M Bergman, and Pavel Tomancak. Gene expression divergence recapitulates the developmental hourglass model. *Nature*, 468(7325):811–814, 2010.

[116] Immanuel Kant. *Critique of judgment*. 1790.

[117] Stuart A. Kauffman. *The origins of order: Self-organization and selection in evolution*. Oxford university press, 1993.

[118] Benjamin B Kaufmann, Qiong Yang, Jerome T Mettetal, and Alexander van Oudenaarden. Heritable stochastic switching revealed by single-cell genealogy. *PLoS biology*, 5(9):e239, 2007.

[119] Douglas B Kell and Stephen G Oliver. Here is the evidence, now what is the hypothesis? the complementary roles of inductive and hypothesis-driven science in the post-genomic era. *Bioessays*, 26(1):99–105, 2004.

[120] Oscar Kempthorne. An introduction to genetic statistics. 1957.

[121] N Kern, D Weaire, A Martin, S Hutzler, and SJ Cox. Two-dimensional viscous froth model for foam dynamics. *Physical Review E*, 70(4):041411, 2004.

[122] Marek Kimmel and David E. Axelrod. *Branching Processes in Biology*, volume 19 of *Interdisciplinary Applied Mathematics*. Springer New York, New York, NY, 2002.

[123] Christian Peter Klingenberg. Evolution and development of shape: integrating quantitative approaches. *Nature Reviews Genetics*, 11(9):623–635, 2010.

[124] Andrei Kolmogorov. *Grundbegriffe der Wahrscheilichkeitsrechnung*. Springer, 1933.

[125] Shigeru Kondo and Takashi Miura. Reaction-diffusion model as a framework for understanding biological pattern formation. *Science*, 329(5999):1616–1620, 2010.

[126] Z. Krivá, K. Mikula, N. Peyriéras, B. Rizzi, A. Sarti, and O. Stašová. 3d early embryogenesis image filtering by nonlinear partial differential equations. *Medical Image Analysis*, 14(4):510 – 526, 2010.

[127] Jean-Jacques Kupiec. *L'origine des individus*. Fayard, 2008.

[128] Jean-Jacques Kupiec and Pierre Sonigo. *Ni Dieu ni gène, pour une autre théorie de l'hérédité*. Seuil, 2000.

[129] Jean-Baptiste Lamarck. *Philosophie Zoologique [1809]*. Edition Flammarion, 1994.

[130] Bomyi Lim, Núria Samper, Hang Lu, Christine Rushlow, Gerardo Jiménez, and Stanislav Y Shvartsman. Kinetics of gene derepression by erk signaling. *Proceedings of the National Academy of Sciences*, 110(25):10330–10335, 2013.

[131] Eckhard Limpert, Werner A Stahel, and Markus Abbt. Log-normal distributions across the sciences: Keys and clues. *BioScience*, 51(5):341–352, 2001.

[132] Bentoît Lombardot. *Description intégrative des cinématiques de la structure embryonnaire: morphogenèse précoce du poisson zébré*. PhD thesis, Ecole polytechnique, 2010.

[133] Giuseppe Longo. Laplace, turing et la géométrie impossible du "jeu de l'imitation" aléas, déterminisme et programmes dans le test de turing. *Intellectica*, 35(2):131–161, 2002.

[134] Giuseppe Longo. Mathematical infinity "in prospettiva" and spaces of possibilities. *Visible, a Semiotics Journal*, 9, 2011.

[135] Giuseppe Longo and Maël Montévil. From physics to biology by extending criticality and symmetry breakings. *Progress in biophysics and molecular biology*, 2:106, 2011.

[136] Giuseppe Longo and Maël Montévil. *Perspectives on Organisms: Biological Time, Symmetries and Singularities*. Springer Science & Business Media, 2013.

[137] Giuseppe Longo, Maël Montévil, and Stuart Kauffman. No entailing laws, but enablement in the evolution of the biosphere. In *Proceedings of the 14th annual conference companion on Genetic and evolutionary computation*, pages 1379–1392. ACM, 2012.

[138] Richard Losick and Claude Desplan. Stochasticity and cell fate. *Science*, 320(5872):65–68, 2008.

[139] Salvador E Luria and Max Delbrück. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics*, 28(6):491, 1943.

[140] Michael Lynch and John S Conery. The origins of genome complexity. *science*, 302(5649):1401–1404, 2003.

[141] Angelo Marinucci. Tra ordine e caos: metodi e linguaggi tra fisica, matematica e filosofia. Aracne, 2011.

[142] Per Martin-Löf. The definition of random sequences. *Information and control*, 9(6):602–619, 1966.

[143] Ernst Mayr et al. Animal species and evolution. *Animal species and their evolution.*, 1963.

[144] David R McClay. Evolutionary crossroads in developmental biology: sea urchins. *Development*, 138(13):2639–2648, 2011.

[145] John H McDonald. *Handbook of biological statistics*, volume 2. Sparky House Publishing Baltimore, MD, 2009.

[146] Michael P McKinley, David C Bolton, and Stanley B Prusiner. A protease-resistant protein is a structural component of the scrapie prion. *Cell*, 35(1):57–62, 1983.

[147] Amy McMahon, Willy Supatto, Scott E Fraser, and Angelike Stathopoulos. Dynamic analyses of drosophila gastrulation provide insights into collective cell migration. *Science*, 322(5907):1546–1550, 2008.

[148] Andrew P McMahon, Constantin N Flytzanis, Barbara R Hough-Evans, Karen S Katula, Roy J Britten, and Eric H Davidson. Introduction of cloned dna into sea urchin egg cytoplasm: replication and persistence during embryogenesis. *Developmental biology*, 108(2):420–430, 1985.

[149] Sean G Megason and Scott E Fraser. Imaging in systems biology. *Cell*, 130(5):784–795, 2007.

[150] Gregor Mendel. *Versuche über Pflanzenhybriden*, volume 44. 1866.

[151] Francesca Merlin. Le "hasard évolutionnaire" de toute mutation génétique, ou la vision consensuelle de la synthèse moderne. *Bulletin d'Histoire et d'épistémologie des sciences de la vie*, 18(1):79–108, 2011.

[152] Jacques Monod. *Le hasard et la nécessité*. Seuil, Paris, 1970.

[153] Michel Morange. *Histoire de la biologie moléculaire*. La Découverte, 2013.

[154] Eric G Moss. Heterochronic genes and the nature of developmental time. *Current Biology*, 17(11):R425–R434, 2007.

[155] Elizabeth Munch. *Applications of Persistent Homology to Time Varying Systems*. Duke University, 2013.

[156] James D Murray. Mathematical biology i: An introduction, vol. 17 of interdisciplinary applied mathematics, 2002.

[157] Akiko Nakamasu, Go Takahashi, Akio Kanbe, and Shigeru Kondo. Interactions between zebrafish pigment cells responsible for the generation of turing patterns. *Proceedings of the National Academy of Sciences*, 106(21):8429–8434, 2009.

[158] Mark EJ Newman, Steven H Strogatz, and Duncan J Watts. Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, 64(2):026118, 2001.

[159] Frank Nielsen and Richard Nock. Sided and symmetrized bregman centroids. *Information Theory, IEEE Transactions on*, 55(6):2882–2904, 2009.

[160] Aaron Novick and Milton Weiner. Enzyme induction as an all-or-none phenomenon. *Proceedings of the National Academy of Sciences of the United States of America*, 43(7):553, 1957.

[161] Christiane Nüsslein-Volhard and Eric Wieschaus. Mutations affecting segment number and polarity in drosophila. *Nature*, 287(5785):795–801, 1980.

[162] Ikuhiro Okamoto, Arie P Otte, C David Allis, Danny Reinberg, and Edith Heard. Epigenetic dynamics of imprinted x inactivation during early mouse development. *Science*, 303(5658):644–649, 2004.

[163] Nicolas Olivier, Miguel A Luengo-Oroz, Louise Duloquin, Emmanuel Faure, Thierry Savy, Israël Veilleux, Xavier Solinas, Delphine Débarre, Paul Bourgine, Andrés Santos, Nadine Peyriéras, and Emmanuel Beaurepaire. Cell lineage reconstruction of early zebrafish embryos using label-free nonlinear microscopy. *Science*, 329(5994):967–971, 2010.

[164] Stanley Osher and James A Sethian. Fronts propagating with curvature-dependent speed: algorithms based on hamilton-jacobi formulations. *Journal of computational physics*, 79(1):12–49, 1988.

[165] Miriam Osterfield, XinXin Du, Trudi Schüpbach, Eric Wieschaus, and Stanislav Y Shvartsman. Three-dimensional epithelial morphogenesis in the developing *Drosophila* egg. *Developmental cell*, 24(4):400–410, 2013.

[166] Periklis Pantazis and Willy Supatto. Advances in whole-embryo imaging: a quantitative transition is underway. *Nature Reviews Molecular Cell Biology*, 15(5):327–339, 2014.

[167] Ankit B Patel, William T Gibson, Matthew C Gibson, and Radhika Nagpal. Modeling and inferring cleavage patterns in proliferating epithelia. *PLoS computational biology*, 5(6):e1000412, 2009.

[168] Johan Paulsson. Summing up the noise in gene networks. *Nature*, 427(6973):415–418, 2004.

[169] Wuhong Pei and Benjamin Feldman. Identification of common and unique modifiers of zebrafish midline bifurcation and cyclopia. *Developmental biology*, 326(1):201–211, 2009.

[170] Wuhong Pei, P Huw Williams, Matthew D Clark, Derek L Stemple, and Benjamin Feldman. Environmental and genetic modifiers of squint penetrance during zebrafish embryogenesis. *Developmental biology*, 308(2):368–378, 2007.

[171] Isabelle S Peter, Emmanuel Faure, and Eric H Davidson. Predictive computation of genomic logic processing functions in embryonic development. *Proceedings of the National Academy of Sciences*, 109(41):16434–16442, 2012.

[172] Anne-Cécile Petit, Emilie Legué, and Jean-François Nicolas. Methods in clonal analysis and applications. *Reproduction Nutrition Development*, 45(3):321–339, 2005.

[173] Anna Petryk, Daniel Graf, and Ralph Marcucio. Holoprosencephaly: signaling interactions between the brain and the face, the environment and the genes, and the phenotypic variability in animal models and humans. *Wiley Interdisciplinary Reviews: Developmental Biology*, 4(1):17–32, 2015.

[174] Patrick C Phillips. Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics*, 9(11):855–867, 2008.

[175] Nicholas Pippenger and Kristin Schleich. Topological characteristics of random triangulated surfaces. *Random Structures & Algorithms*, 28(3):247–288, 2006.

[176] Jean Piveteau. Le débat entre cuvier et geoffroy saint-hilaire sur l'unité de plan et de composition. *Revue d'histoire des sciences et de leurs applications*, 3(4):343 – 363, 1950.

[177] H. Poincaré. *Science and Method*. Dover Publications, 2013.

[178] Henri Poincaré. Sur le problème des trois corps et les équations de la dynamique. *Acta mathematica*, 13(1):A3–A270, 1890.

[179] RM Purcell. Life at low Reynolds number. *American Journal of Physics*, 1977.

[180] C Radhakrishna Rao. Information and accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, 37(3):81–91, 1945.

[181] Arjun Raj, Scott A Rifkin, Erik Andersen, and Alexander van Oudenaarden. Variability in gene expression underlies incomplete penetrance. *Nature*, 463(7283):913–918, 2010.

[182] Oliver J Rando and Kevin J Verstrepen. Timescales of genetic and epigenetic inheritance. *Cell*, 128(4):655–668, 2007.

[183] C Radhakrishna Rao. *Linear statistical inference and its applications*, volume 22. John Wiley & Sons, 2009.

[184] Michael Regulski, Katherine Harding, Richard Kostriken, Francois Karch, Michael Levine, and William McGinnis. Homeo box genes of the antennapedia and bithorax complexes of drosophila. *Cell*, 43(1):71–80, 1985.

[185] Barbara Rizzi and Nadine Peyriéras. Towards 3d in silico modeling of the sea urchin embryonic development. *Journal of chemical biology*, 7(1):17–28, 2014.

[186] Barbara Rizzi*, Paul Villoutreix, Julien Delile*, Louise Duloquin*, Thierry Savy, Matthieu Herrmann, Emmanuel Faure, Dimitri Fabrèges, Yannick Kegosien, Paul Bourgine, René Doursat, and Nadine Peyriéras. Predicting sea urchin's normal development from a small cohort of digital embryos. *In preparation*, 2015.

[187] DD Romaschoff. Über die variabilität in der manifestierung eines erblichen merkmales (abdomen abnormalis) bei drosophila funebris f. *Journal für Psychologie und Neurologie*, 31:323–325, 1925.

[188] Robert Rosen. A quantum-theoretic approach to genetic problems. *The bulletin of mathematical biophysics*, 22(3):227–255, 1960.

[189] Wilhelm Roux. *Über die Entwicklungsmechanik der Organismen.* 1890.

[190] Patrik Sahlin and Henrik Jönsson. A modeling study on how cell division affects properties of epithelial tissues under isotropic growth. *PloS one*, 5(7):e11750, 2010.

[191] Daniel Sánchez-Gutiérrez, Aurora Sáez, Alberto Pascual, and Luis M Escudero. Topological progression in proliferating epithelia is driven by a unique variation in polygon distribution. *PloS one*, 8(11):e79227, 2013.

[192] Sebastian A Sandersius, Manli Chuai, Cornelis J Weijer, and Timothy J Newman. Correlating cell behavior with tissue topology in embryonic epithelia. *PloS one*, 6(4):e18081, 2011.

[193] Oded Sandler, Sivan Pearl Mizrahi, Noga Weiss, Oded Agam, Itamar Simon, and Nathalie Q Balaban. Lineage correlations of single cell division time as a probe of cell-cycle dynamics. *Nature*, 519(7544):468–471, 2015.

[194] Alessandro Sarti, Ravi Malladi, and James A. Sethian. Subjective surfaces: A method for completing missing boundaries. *Proceedings of the National Academy of Sciences*, 97(12):6258–6263, 2000.

[195] Stephan Schneider, Herbert Steinbeisser, Rachel M Warga, and Peter Hausen. β-catenin translocation into nuclei demarcates the dorsalizing centers in frog and fish embryos. *Mechanisms of development*, 57(2):191–198, 1996.

[196] Erwin Schrödinger. *What is life*. Cambridge University Press, 1944.

[197] Glenn Shafer and Vladimir Vovk. The sources of kolmogorov's" grundbegriffe". *Statistical Science*, pages 70–98, 2006.

[198] Ilya Shmulevich, Edward R Dougherty, Seungchan Kim, and Wei Zhang. Probabilistic boolean networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics*, 18(2):261–274, 2002.

[199] J Maynard Smith, Richard Burian, Stuart Kauffman, Pere Alberch, John Campbell, Brian Goodwin, Russell Lande, David Raup, and Lewis Wolpert. Developmental constraints and evolution: a perspective from the mountain lake conference on development and evolution. *Quarterly Review of Biology*, pages 265–287, 1985.

[200] Thierry Sousbie. The persistent cosmic web and its filamentary structure–i. theory and implementation. *Monthly Notices of the Royal Astronomical Society*, 414(1):350–383, 2011.

[201] Frank W Stearns. One hundred years of pleiotropy: a retrospective. *Genetics*, 186(3):767–773, 2010.

[202] John E Sulston and H Robert Horvitz. Post-embryonic cell lineages of the nematode, caenorhabditis elegans. *Developmental biology*, 56(1):110–156, 1977.

[203] Peter S Swain, Michael B Elowitz, and Eric D Siggia. Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proceedings of the National Academy of Sciences*, 99(20):12795–12800, 2002.

[204] NW Timoféeff-Ressovsky. Über den einfluss des genotypus auf das phänotypen auftreten eines einzelnes gens. *Journal für Psychologie und Neurologie*, 31:305–310, 1925.

[205] Georgios Trichas, Aaron M Smith, Natalia White, Vivienne Wilkins, Tomoko Watanabe, Abigail Moore, Bradley Joyce, Jacintha Sugnaseelan, Tristan A Rodriguez, David Kay, et al. Multi-cellular rosettes in the mouse visceral endoderm facilitate the ordered migration of anterior visceral endoderm cells. *PLoS biology*, 10(2):e1001256, 2012.

[206] Thai V Truong, Willy Supatto, David S Koos, John M Choi, and Scott E Fraser. Deep and fast live imaging with two-photon scanned light-sheet microscopy. *Nature methods*, 8(9):757–760, 2011.

[207] Alan Mathison Turing. The chemical basis of morphogenesis. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 237(641):37–72, 1952.

[208] Franck Varenne. La reconstruction phénoménologique par simulation : vers une épaisseur du simulat. In D. Parrochia & V. Tirloni, editor, *Formes, systèmes et milieux techniques*, pages 107–123. Edition Jacques André, Lyon, 2012.

[209] Paul Villoutreix. Random triangulated surfaces with arbitrary degree distribution reveal embryonic epithelial organization. In *Algebraic Topology - Methods, Computation and Science 6 (Vancouver)*, 2014.

[210] Paul Villoutreix. Vers une modélisation multi-échelle de la variabilité biologique? In Franck Varenne, Marc Silberstein, Sébastien Dutreuil, and Philippe Huneman, editors, *Modéliser & Simuler. Epistémologie et pratique de la modélisation et de la simulation, tome 2*, pages 643 – 664. Editions matériologiques, 2014.

[211] Paul Villoutreix. Biological diversity and quantum formalism: Entanglement of genotype-phenotype relations. *In preparation*, 2015.

[212] Paul Villoutreix, Gunnar Carlsson, and Nadine Peyriéras. Topological analysis of epithelia uncovers high-order structures in spatial distribution of cells. *In preparation*, 2015.

[213] Paul Villoutreix, Barbara Rizzi, Julien Delile, Louise Duloquin, Emmanuel Faure, Thierry Savy, Paul Bourgine, and Nadine Peyriéras. A probabilistic modeling of the cell lineage highlights interindividual variability in paracentrotus lividus early development. In *The Developmental Biology of the Sea Urchin XXII*, 2014.

[214] Carl-Ernst von Baer. *Über Entwicklungsgeschichte der Thiere. Beobachtung und Reflexion.-Königsberg, Gebrüder Bornträger 1828-1837*, volume 1. Gebrüder Bornträger, 1828.

[215] Conrad Hal Waddington. *The strategy of the genes*. Allen and Unwin, 1957.

[216] Yu-Chiun Wang, Zia Khan, Matthias Kaschube, and Eric F Wieschaus. Differential positioning of adherens junctions is associated with initiation of epithelial folding. *Nature*, 484(7394):390–393, 2012.

[217] Gerry Webster and Brian C Goodwin. The origin of species: a structuralist approach. *Journal of Social and Biological Structures*, 5(1):15–47, 1982.

[218] Mary Jane West-Eberhard. Phenotypic plasticity and the origins of diversity. *Annual review of Ecology and Systematics*, pages 249–278, 1989.

[219] Fengzhu Xiong, Wenzhe Ma, Tom W Hiscock, Kishore R Mosaliganti, Andrea R Tentner, Kenneth A Brakke, Nicolas Rannou, Arnaud Gelas, Lydie Souhait, Ian A Swinburne, et al. Interplay of cell shape and division orientation promotes robust morphogenesis of developing epithelia. *Cell*, 159(2):415–427, 2014.

[220] Fengzhu Xiong, Andrea R Tentner, Peng Huang, Arnaud Gelas, Kishore R Mosaliganti, Lydie Souhait, Nicolas Rannou, Ian A Swinburne, Nikolaus D Obholzer, Paul D Cowgill, et al. Specified neural progenitors sort to form sharp domains after noisy shh signaling. *Cell*, 153(3):550–561, 2013.

[221] Jennifer A Zallen. Planar polarity and tissue morphogenesis. *Cell*, 129(6):1051–1063, 2007.

[222] C. Zanella, M. Campana, B. Rizzi, C. Melani, G. Sanguinetti, P. Bourgine, K. Mikula, N. Peyriéras, and A. Sarti. Cells segmentation from 3-d confocal images of early zebrafish embryogenesis. *Image Processing, IEEE Transactions on*, 19(3):770–781, March 2010.

[223] Afra Zomorodian and Gunnar Carlsson. Computing persistent homology. *Discrete & Computational Geometry*, 33(2):249–274, 2005.