CONTENTS

I SPATIAL ANALYSIS OF COMPLEX BIOLOGICAL TISSUES FROM SINGLE CELL GENE EXPRESSION DATA 1

1 INTRODUCTION 3

2 CAPTURING GENE EXPRESSION IN *platynereis dumerilii*'S BRAIN 5
- 2.1 Platynereis dumerilii, an ideal organism of brain development studies 5
  - 2.1.1 General description 5
  - 2.1.2 Larval development 6
- 2.2 Gene expression in Platynereis' developing brain 8
  - 2.2.1 Platynereis' nervous development until 48hpf 8
  - 2.2.2 Spatial organization of complex biological tissues like the brain 9
  - 2.2.3 Generalities about gene expression and development 9
- 2.3 Capturing gene expression in the laboratory 11
  - 2.3.1 In-situ hybridization assays 11
  - 2.3.2 Building a image library of gene expression for Platynereis 11
  - 2.3.3 RNA sequencing 13

3 FROM TISSUE TO SINGLE CELL TRANSCRIPTOMICS, A PARADIGM SHIFT 15
- 3.1 Spatially referenced single cell-like in-situ hybridization data 15
- 3.2 Singe cell RNA sequencing, building a map of the full transcriptome 15
- 3.3 About the quantitative trait of single cell expression data 16
- 3.4 Binarizing gene expression datasets 16
- 3.5 Preliminary results on mapping single cell RNA-seq data in from Platynereis' brain 16

4 HIDDEN MARKOV RANDOM FIELD BASED CLUSTERING FOR SINGLE CELL GENE EXPRESSION DATA 17
- 4.1 Markov Random Field prior distribution 17
- 4.2 Hidden Markov model 18
- 4.3 Parameter estimation using the EM algorithm 19
- 4.4 Mean field approximations 20
- 4.5 Maximization 21
- 4.6 Estimating K 21

II APPENDIX 23
A APPENDIX 25

BIBLIOGRAPHY 27

LIST OF FIGURES

Figure 1 *Platynereis dumerillii*'s larva and adult forms. 5

Figure 2 *Platynereis dumerillii*'s larvea development at 48hpf or late trochophore. Striped in red is indicated the area which forms the developing brain of the larvae. 7

Figure 3 *Platynereis dumerillii*'s stereotypical and synchronous development. In green and red are two different *P. dumerillii* individuals' with the same gene expression being highlighted. They show extremely similar patterns of development. 7

Figure 4 Fluorescent in-situ hybridization assays to create a 169 genes catalogue of gene expression in the brain of *P. dumerillii*. From the live tissue cut into thin fixed layers, every slice is stained with a reference gene and a gene of interest that will reveal the areas of expression under fluorescent microscopy. The process repeated 169 times for key genes in *P. dumerillii* development has been generated by [31] 12

LIST OF TABLES

LISTINGS

ACRONYMS

Part I

SPATIAL ANALYSIS OF COMPLEX BIOLOGICAL TISSUES FROM SINGLE CELL GENE EXPRESSION DATA

1 INTRODUCTION

This is where the introduction goes

2

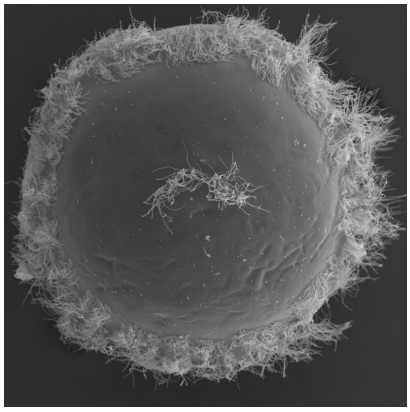
CAPTURING GENE EXPRESSION IN *PLATYNEREIS DUMERILLII*'S BRAIN

2.1 PLATYNEREIS DUMERILII, AN IDEAL ORGANISM OF BRAIN DEVELOPMENT STUDIES

2.1.1 *General description*

P. dumerillii is a marine annelid of the class Polychaeta, it has been established as one of the main marine animal models in the fields of evolutionary, developmental and neurobiological biology as well as ecology and toxicology [15, 29, 13, 6, 8, 9]. As a member of the bilateria *P. dumerillii* has a defined bilateral symmetry.

P. dumerilliipopulates shallow (no more than 3m) hard ocean floors around the world. It is commonly found in the Mediterranean sea, the north Atlantic coast of Europe as well as in the shallow seas surrounding Sri Lanka, Java and the Philippines. Eggs, embryos and larvae are roughly 160 µm while the adults can measure up to 6cm in length.



(a) Larval form of *P. dumerillii*. Image: MPI for Developmental Biology.

(b) Adult *P. dumerillii*. Image: Arendt group, EMBL

Figure 1: *Platynereis dumerillii*'s larva and adult forms.

They are several reasons why *P. dumerillii* has been chosen as a model by numerous laboratories. In terms of evolution *P. dumerillii* shows several interesting characteristics. It belongs to the lophotrochozoan taxon of the bilaterian animals as opposed to most of the well established models animal which either belong to the ecdysozoans (*Caenorhabditis elegans*, *Drosophila melanogaster*) or the deuterostomes (mouse, human). Lophotrochozoans being extremely under represented, *P. dumerillii* as a model organism is essential to comparative approach

on bilaterian biology.

P. dumerillii also shows an exceptionally slow evolutionary lineage. It has even been described as a "living fossil" for that reason [9]. This means that the ancestral developmental characteristics of *P. dumerillii* are at an image of the common past of all bilaterians. To illustrate this fact an interesting example described in [5, 30] is the conserved molecular topography of the genes responsible for the development of the central nervous system between *P. dumerillii* and all vertebrates. This slow evolutionary rate confers *P. dumerillii* the advantage of being a link between fast evolving models like *drosophila* and vertebrates.

In terms of practicality, *P. dumerillii* can easily be kept and bred in captivity producing offspring throughout the year [8]. The behavioural characteristics of *P. dumerillii* mating ritual have been well studied. The "nuptial dance" happens on the water surface, male and female releasing the sperm and eggs synchronously, respectively. This activity is synchronized by pheromones released into the water [35]. Over 2000 individuals can be produced within a single batch. Every new individual will undergo embryonic then larval development before reaching *P. dumerillii*'s adult form.

2.1.2 *Larval development*

Similarly to the other polychaetes, the larval development of *P. dumerillii* can be decomposed into three main anatomical stages: the trochophore, the metotrochophore and the nectochaete. The trochophore is spherical and moves via a equatorial belt of ciliated cells as well as an apical organ possessing a ciliary tuft [25, 22] as seen on figure 1a and schematically on figure 2. the metotrochophore stage is characterized by the development of a slightly elongated segmented trunk compared to that of the trochophore [12]. The next stage is the nectochaete larvae that resembles the adult (figure 1b) in most of the traits especially with parapodial appendages used for swimming and crawling [12]. This traditional subdivision has been applied to *P. dumerillii* [14].

Aside from this purely anatomical subdivision, an additional staging systems exists and has become the norm for current studies. The development is measured in *hours post fertilization* (hpf) at 18°C.

A key factor making *P. dumerillii* such an interesting model to work with is the fact that after fertilization, the $\approx$ 2000 larva will start developing at the exact same time, in a synchronous fashion. Furthermore, the larval development of *P. dumerillii* follows a very stereotypical pattern with very little variation from one individual to the other

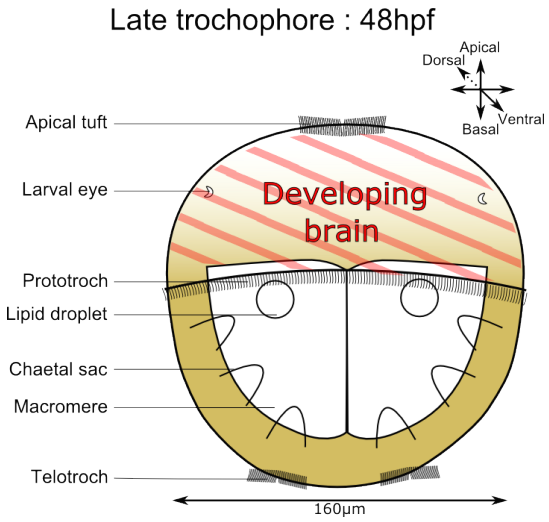


Figure 2: *Platynereis dumerillii*'s larvea development at 48hpf or late trochophore. Striped in red is indicated the area which forms the developing brain of the larvae.

and even between batches provided the temperature is kept constant [8, 6]. An example showing the similarity between individuals during development can be seen on figure 3. this is a very important feature as it allows biologists to repeat experiments on several individuals at a very close developmental stage even if they are from different batches.

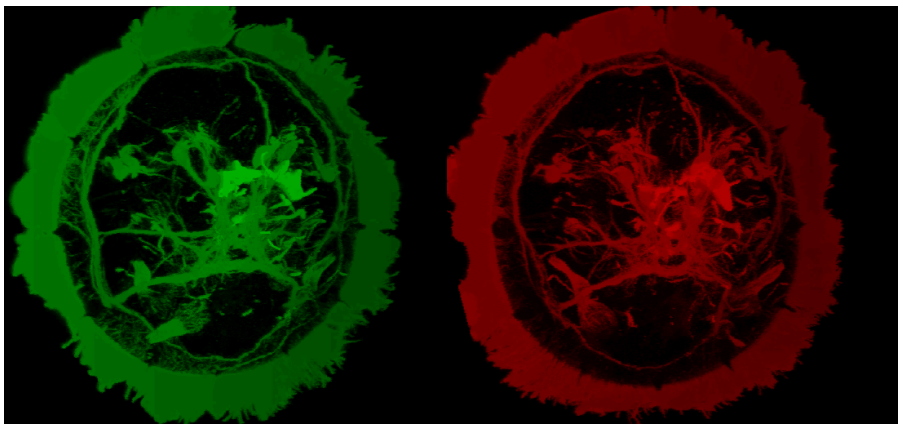


Figure 3: *Platynereis dumerillii*'s stereotypical and synchronous development. In green and red are two different *P. dumerillii* individuals' with the same gene expression being highlighted. They show extremely similar patterns of development.

Describing the entire development of *P. dumerillii* does not fall within the scope of this thesis. Indeed, we will only be interested in the brain of *P. dumerillii*'s larvae at 48hpf. Therefore, it is important to have an anatomical idea of what the brain looks like at this time

in development and what inherent characteristics will be the most interesting to investigate.

2.2 GENE EXPRESSION IN PLATYNEREIS' DEVELOPING BRAIN

2.2.1 *Platynereis' nervous development until 48hpf*

The main purpose of this thesis is not to fully understand the patterns of development in *P. dumerillii*'s larval brain. Therefore we will only give a brief summary of what the main component of the brain are at 48hpf, the time point we will be interested in in the next chapters. *P. dumerillii*'s larval brain development is detailed in [9].
From the early trochophore (24-26hpf) neural system development starts taking place. The apical ganglion forms at the apical tuft. It contains one sertonergic cell and a few neurons that link to the nerve of the ciliary band of the larva called the prototroch (see figure 2 for better understanding). This allows the first movements of the larve thanks to the cialited cells of the prototroch.
The mid-trochophore (26-40 hpf) sees the formation of the first cerebral commissure, it is a band of nerves interconnecting the ventral nerve cord and the brain. This trait is a typical feature of annelids brains. During this phase the apical ganglion becomes bigger with three more serotonergic cells.
The late trochophore (40-48hpf)sees the formation of the second commissure in the ventral nerve cord. It is at the end of this stage that the brain starts to become more complexity with a notable increase in the number of neutrites.
The data we will use in the rest of this thesis will not encapsulated the whole larvae, just the brain (see figure 2) thus excluding the ventral part of the nervous system. The best studied areas of the brain are the larval eyes, the developing adult eyes, the apical organ on the dorsal side. On the ventral side are located the mushroom bodies a pair of structures that are known to play a role in olfactory learning and memory in insects and annelids [31]. A schematic representation of those areas is shown on figure FIGURE brain areas.

Even at a very early stage in a relatively simple organism, the brain quickly becomes a complex tissue. Cell types diverge and functional areas are formed. Before trying to understand more about *P. dumerillii*'s brain organization, it is interesting to ask the more general question about how complex tissues such as the brain are defined spatially.

2.2.2 *Spatial organization of complex biological tissues like the brain*

This section is not intended to demonstrate a specificity of the *P. dumerillii*'s brain, it is meant to ask some of the fundamental questions that intrinsically motivate the work presented in the rest of this thesis. Complex tissues, the obvious example of which is the brain, could be viewed as an interconnected mosaic of cells having different functions, working together to achieve the global function of the organ.

If we look closely at this mosaic of cells, the spatial organisation of this mosaic is not random. Cells that serve the same function will often be close from each other, thus defining functional tissues. However, the spatial coherency of those tissues is not necessarily always the same. Some cell types could be formed of cells scattered inside another more spatially coherent tissue. To illustrate that fact, an interesting example is the difference between the spatial coherency of cells forming the neuronal tissue in the brain and cells forming a well defined region in the brain like the mushroom bodies. When asking the question, is it likely that this cell is fully surrounded by the same cell type, the extensions created by the axons of neurones will decrease this probability. Indeed, axons will grow through other types of tissues to reach their destination, making the overall spatial coherency of "neuronal" tissue smaller than very well spatially defined tissues.

When trying to analyse the full structure of the brain with an automated method, keeping in mind that fact could prove important to improve the results. This fact and its consequences on the work presented in this method are further discussed in section cite section spatial clustering.

So far, we have only regarded organs and cell types as regarding their anatomical traits. But as mentioned in the introduction 1 the functional heterogeneity of complex tissues goes further that simple anatomical traits. We need to work on traits that fundamentally represent how cells are functioning.

2.2.3 *Generalities about gene expression and development*

When speaking about developmental biology it it should be noted that the term “cell" will be referring to eukaryotic cells and more specifically those of multicellular organisms. Every cell in a complex organism possesses the same genome, that is, the sum of all the genetic information contained in the cell (nucleus and other compartments). This fundamental homogeneity is in plain contradiction with the heterogeneity observed anatomically. If every cell has the exact

same DNA, where does the great variability between cell types come from (what makes a neurone become a neurone and not a pancreatic cell). Answering this sort of questions defines the field of developmental biology.

The short and rather complete answer to any developmental biology question actually is: same genome but different pattern of gene expression. As indeed gene expression is the central, most important, most studied cellular activity. Gene expression even is general common denominator of life as large parts of the mechanisms making up gene expression are actually shared by every living creature known to man.

Of course to understand what gene expression is, we must first define what genes are. The precise definition of a gene is still controversial. The concept of a "factor that conveys traits from parents to offspring" was laid by Gregor Mendel in 1866 [19] when the accepted theory at the time was based on blending inheritance where the traits of the parents appeared mixed in the offspring following a continuous gradient. The most recent published definition of a gene followed the publication of the ENCODE project [7]. It states that a gene is "A gene is a union of genomic sequences encoding a coherent set of potentially overlapping functional products."

Gene expression is the way cells express their genes. Expression of a gene is the process of transcribing the DNA of that particular gene. The product of gene expression is RNA molecules and there are several ways to look at gene expression. In a cell or tissue, at a given time point we can to choose to look whether a gene is expressed or not (binary expression) or how much a certain gene is expressed (quantitative expression).

Most RNA molecules are translated into proteins that can have very different purposes some will directly serve in the cellular life as functional/structural agents (elements of the ATP synthase for example) others will have a regulatory effect on gene expression. In other terms the expression gene a, coding for protein A might activate, accelerate, inactivate or decelerate expression of gene b and potentially others. This outlines the complex interdependent regulatory system that is gene expression. For precise examples gene regulation see [11, 27, 10, 1].

Add figure for gene expression Camille

During development mechanisms exist that allow gene expression to become differential as the divisions occur. This is how the asymmetrical axis (dorso-ventral, and basal-apical) of the body are defined.

The main mechanism involves chemical gradients. The first of these gradient has to come from the original cell which musty contain some asymmetrically distributed chemical so that the first divisions lead to non identical cells. In the case of *Platynereis dumerillii*, the body axis are defined between 2hpf and 7hpf [9].

As described, gene expression is the key factor during tissue development. The ability to study gene expression patterns has revolutionized the fields of developmental biology. Technological innovation has been the main driving factor of this revolution. In the next section we will present two methods to capture gene expression.

2.3 CAPTURING GENE EXPRESSION IN THE LABORATORY

2.3.1 *In-situ hybridization assays*

In-situ hybridization (ISH) is an experimental technique where the practitioner is able to determine in which cells of the tissue under study a particular RNA is expressed. As opposed to Southern blotting, ISH assays not only allow to know whether a gene is expressed or not, but also where in the tissue it is expressed. First proposed in 1969 by Pardue [23] and John [16] independently, in-situ hybridization (ISH) used radioactive tritium labelled probes on a photographic emulsion to reveal on which chromosome particular genomic components were located. With the development of fluorescent labelling techniques [17, 24] allowing for faster, more sensitive and of course safer hybridization assays [28] compared to radioactive probes, Fluorescent in-situ hybridization (FiSH) quickly became the standard technique to study gene expression in the spatial context of the biological tissue. Importantly, using multiple fluorescent probes of different colours allowed the simultaneous localization of several RNA fragments in the tissues [21].

2.3.2 *Building a image library of gene expression for Platynereis*

During his PhD, Raju Tomer and colleagues cite thesis member of the Detlev Arendt lab in EMBL, used Fluorescent in-situ hybridization to create an image library of gene expression in the brain of *P. dumerillii*. He was able to record gene expression in the full brain at 48hpf for 169 genes. In practice each individual larvae was dissected to isolate the brain, which was then cut into thin slices and fixed. Each individual slice was then stained with two different fluorescent probes corresponding to two messenger RNAs (RNAm). One of the gene is considered a reference, as it is always hybridized in all the assays (the main reference gene used was Emx) along an other gene of interest,

see figure 4.

As mentioned previously, the larval development of *P. dumerillii* is highly similar in every individual larvae. In the case of this study requiring a lot of different assays conducted each time a on different animal, the stereotypical development of *P. dumerillii* has proven essential. Indeed, having the same reference localized in all the assays has allowed Tomer to align all the other gene expression patterns onto this scaffold.

The result is an image library of 169 gene expression pattern in the full brain of *P. dumerillii* with a exploitable spatial reference that allows for a very precise mapping.

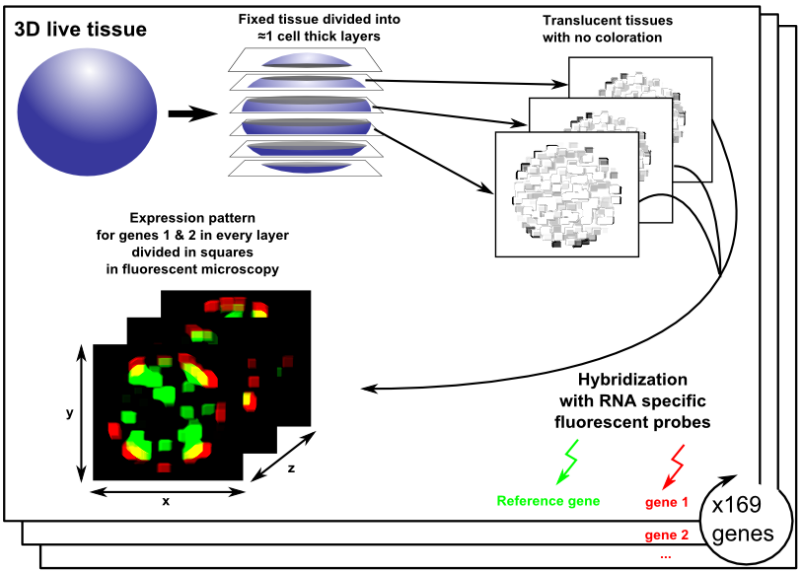


Figure 4: Fluorescent in-situ hybridization assays to create a 169 genes catalogue of gene expression in the brain of *P. dumerillii*. From the live tissue cut into thin fixed layers, every slice is stained with a reference gene and a gene of interest that will reveal the areas of expression under fluorescent microscopy. The process repeated 169 times for key genes in *P. dumerillii* development has been generated by [31]

However useful and practical fluorescent in-situ hybridization may be, such assays are limited in terms the quantity of gene one is able to study. Indeed, each individual larvae only provides the expression of two genes, one being the reference. Crucial developments in sequencing technologies have brought a way to study the expression of the whole transciptome landscape in a single assay, RNA sequencing.

2.3.3 *RNA sequencing*

Whole Transcriptome Shotgun Sequencing (WTSS) also called RNA sequencing (RNA-seq) [20, 32] has developed alongside Next Generation Sequencing (NGS) technique used to retrieve the genome (DNA). Instead, when preparing the starting material, only the RNAs are selected using magnetic beads and a using the polyA tail on the 3' side of messenger RNAs (RNAm). - FIG 3 : about RNA sequencing for tissues - Explanations about the technique - Obtaining the full transcriptome at once - Necessity of having the genome to map to or a list of known genes (primR) - Discuss the starting RNA quantity - Discuss the fact that gene expression is averaged over the tissue losing spatial information.

3

FROM TISSUE TO SINGLE CELL TRANSCRIPTOMICS, A PARADIGM SHIFT

3.1 SPATIALLY REFERENCED SINGLE CELL-LIKE IN-SITU HYBRIDIZATION DATA

Dividing images into "cells"

- Images with a good enough resolution can determine expression at the single cell level. - But every cell is different in terms of shape and size => need for a cell model - in-situ hybridization keeps the spatial information of every cell - Present possibilities for cell model with membrane markers etc... but to start with, a simple mode is possible => next paragraph

A simple cell model, the "cube" data

- FIG 4 : From images to luminiscence cube data - Present the cube cell model, and its assumptions - cells have roughly the same size - cells are roughly cubical - Present the choice of size for the cubes (3um or 6 um) - This model introduces errors (cells divided or several cells in one cube, empty cubes between cells) - => When working on this data we will need methods that are able to smooth those mistakes over.

3.2 SINGE CELL RNA SEQUENCING, BUILDING A MAP OF THE FULL TRANSCRIPTOME

Sequencing single cell RNA contents

- Same as tissue sequencing but with a lot less starting material - Present the main techniques used (see with Luis) : Microfluidics and others - We obtain the full transcriptome of every cell sequenced

Mapping back gene expression to a spatial reference

- Single cell RNA-seq at the moment does not allow to track cell localization - Need to map the transcriptome back to a spatial reference - Use in-situ hybridization results as reference

3.3 ABOUT THE QUANTITATIVE TRAIT OF SINGLE CELL EXPRESSION DATA

Light contamination in in-situ hybridization data

- FIG 5 : show light intensity across one slice - Explain problem of scale and light contamination

Technical noise in single cell RNA-seq data

- FIG 6 : show "typical" correlation plot from single cell RNA-seq with the noise increasing when reducing starting material - Both methods are currently unreliable quantitatively => need to binarize

3.4 BINARIZING GENE EXPRESSION DATASETS

Binarizing in-situ hybrdization datasets

- With biological knowledge and a limited number of genes - Possibility to compare spatially the resulting binary expression patterns to microscope data and adjust for each gene the threshold manually

Binarizing whole transcriptomes

- Manual curation no longer possible - Thresholding ideally with density peaks - Problems that may occur and possible solutions (figure?)

3.5 PRELIMINARY RESULTS ON MAPPING SINGLE CELL RNA-SEQ DATA IN FROM PLATYNEREIS' BRAIN

Single cell RNA-seq in Platynereis' brain

- Present the data (number of cell) - Present the method used (to dissolve the brain, to capture the cells, to sequence the cells

Mapping back RNA-seq data back to PrimR in-situ hybridization assays

- Select the overlapping genes - Present mapping method (Nuno's pipeline) - Present simple mapping technique and why it is not satisfactory - Present John's method - FIG 6: find a nice way to show a few good examples of mapping

4

HIDDEN MARKOV RANDOM FIELD BASED CLUSTERING FOR SINGLE CELL GENE EXPRESSION DATA

4.1 MARKOV RANDOM FIELD PRIOR DISTRIBUTION

Let S be a finite set of sites, each of which represents one “cube" of data (see the results section for a detailed description of the data). Given the coordinates of each site, we were able to define a neighbourhood system on S using a first order neighbourhood system, i.e the 6 closest sites. S and its neighbourhood system can be viewed as a connecting graph G. Let C be the set of cliques of G. C is therefore the set of all sites that are all neighbours from one another.

Let a Random Field Z be defined as a set of random variables $Z = \{Z_i, \forall i \in S\}$ each Zi taking its value in $[1, K]$. For every site $i \in S$, let $N(i)$ represent the set of its neighbours and $\mathbf{z}_{S-\{i\}}$ a realization of the field restricted to $S - \{i\} = \{j \in S, j \neq i\}$. Z is a Markov Random Field if and only if it verifies the Markov property at every site :

$$\forall i \in S, P_G(z_i \mid z_{S-\{i\}}) = P_G(z_i \mid z_j, j \in N(i)) \tag{1}$$

Equation (1) states that the realization of the field at any site $i \in S, z_i$ can be fully determined using only the state of its neighbours $N(i)$. In other words each “cube" is only dependent upon its neighbours. The Hammersley-Clifford theorem state that if Z is a Markov Random Field, the join distribution of the field follows a Gibbs distribution so that :

$$P_G(\mathbf{z}) = \frac{e^{-H(\mathbf{z})}}{\sum_{\mathbf{z}'} e^{-H(\mathbf{z}')}} \tag{2}$$

$H(\mathbf{z})$ is called the Energy function and is summed over the cliques of the graph C. Considering that we are working with an order one neighbouring graph, C is the set of all the couples of sites (i, j) that are neighbours. We chose to consider H as a function of vector $\text{ } = (\beta_1, \ldots, \beta_K)$ containing K parameters, one per cluster and $v_{i,j}$ a potential function set to 1 in our method.

$$H(\mathbf{z}) = -\sum_{i \in S} \beta_{z_i} \sum_{\substack{i,j \\ neighbours}} v_{i,j} \times_{[z_i = z_j]} \tag{3}$$

The denominator in (2) where $\mathbf{z}'$ represents all the possible realizations of the field is a normalizing constant that we will refer to as

$W(\zeta)$.
This model is closely related to a K-color Potts model [33] although instead of a single parameter β for the entire model, we assign one β per cluster. Equation (3) is a decreasing function of every component of ζ and of the number of neighbouring "cubes" in the field having the same class. This Energy thus favours spatially regular partitions and a higher value of β_h, with $1 \leqslant h \leqslant K$ will amplify the smoothing effect, or coherence over cluster h. We chose to use one spatial smoothness parameter per cluster because of the nature of the data we are dealing with. Indeed, in a biological context, it is expected that some tissues will be more spatially coherent than others.

From this prior distribution we have K unknown parameters $\zeta = (\beta_1, \ldots, \beta_K)$ to be estimated by the model. It is important to note at this point that $W(\zeta)$ is summed over all possible realizations of the field Z, this is an exponentially complex sum as the cardinality of S rises. Therefore the computation of the normalizing factor becomes intractable very quickly. To address this problem, we are going to need to make some approximations in order to compute this quantity (see Mean Field Approximations).

We have described the prior distribution of a Markov Random Field representing our partition, we now need to describe the relationship between Z and the data.

4.2 HIDDEN MARKOV MODEL

As Z is unknown a priori and represents the partition, let Y be a set of random variables representing the observations (the in-situ hybridization data). We have to assume conditional independence of the observations given the partition Z so that, with f_{z_i} the density function relative to cluster $z_i, i \in S$:

$$p(\mathbf{y} \mid \mathbf{z}; \Theta) = \prod_{i \in S} p(y_i \mid z_i; \Theta) \tag{4}$$

$$= \prod_{i \in S} f_{z_i}(y_i \mid z_i; \Theta) \tag{5}$$

We define one unknown parameter per cluster: $\Theta = (\eta_1, \ldots, \eta_K)$. It is interesting to note that this part of the model is equivalent to an independent mixture model [18]. Indeed, hidden Markov models can be viewed as independent mixture models where Z is a set of independent, identically distributed random variables, which happens when $\beta = 0$.

Because the 169 genes chosen by Tomer et al. [31] are key genes involved in the early development of Platynereis' brain, the assumption

of conditional independence given the realization of the field seems acceptable. The validity of this hypothesis may however be argued for future RNA-seq datasets representing entire transcriptomes (see discussion).
Given a particular cluster $h \in [1, K]$ and M the set of considered genes, we assume that each gene $m \in M$ follows a Bernoulli distribution with parameter $\theta_{h,m}$. We then have one unknown Bernoulli parameter per gene per cluster so that :

$$\begin{aligned}\Theta &= (\pi_1, \ldots, \pi_K) \\ &= \begin{pmatrix} \theta_{1,1} & \ldots & \theta_{1,K} \\ \vdots & \ddots & \vdots \\ \theta_{M,1} & \ldots & \theta_{M,K} \end{pmatrix}\end{aligned}$$

The conditional density function $f_i, i \in S$ can be expressed as :

$$f_i(y_i \mid z_i; \Theta) = f_i(y_i \mid z_i; \pi_{z_i}) = \prod_{m \in M} \theta_{z_i,m}^{y_{i,m}} \times (1 - \theta_{z_i,m}^{1-y_{i,m}}) \qquad (6)$$

Looking at both fields Z and $Y \mid Z$ together, the complete likelihood of the model is expressed as :

$$P_G(\mathbf{y}, \mathbf{z} \mid \Theta, \beta) = f(\mathbf{y} \mid \mathbf{z}, \Theta) P_G(\mathbf{z} \mid \beta) = \frac{exp\{-H(\mathbf{z} \mid \beta) + \sum_{i \in S} \log f_i(y_i \mid z_i, \theta_z i)\}}{\sum_{z'} e^{-H(\mathbf{z})}} \qquad (7)$$

Because equation (7) is a Gibbs distribution, using the Hammersley-Clifford theorem we can conclude that the conditional field Y given $Z = \mathbf{z}$ is another a Markov Random Field with the Energy function

$$H(\mathbf{z} \mid \mathbf{y}, \beta, \Theta) = H(\mathbf{z} \mid \beta) - \sum_{i \in S} \log f_i(y_i \mid z_i, \Theta)$$

In our case, the goal is to recover the unknown realization of $Z : \mathbf{z}$. To this end we need to maximize the values of all the parameters of the model $\Phi = (\Theta, \beta)$. We will also need to determine the unknown value K. This will be determined a posteriori by computing the BIC over the model full likelihood [26].

4.3 PARAMETER ESTIMATION USING THE EM ALGORITHM

The EM principle can be applied to estimate the parameters $\Phi = (\Theta, \beta)$ of the hidden MRF model. After initializing the clusters $\mathbf{z}$, we

choose Φ^{l+1} at iteration $(l+1)$ in order to maximize the model's expectation:

$$\begin{aligned} Q(\Phi \mid \Phi^l) &= \sum_{\mathbf{z}} p(\mathbf{z} \mid \mathbf{y}; \Phi^l) \log p(\mathbf{y}, \mathbf{z}; \Phi) \\ &= \underbrace{\sum_{\mathbf{z}} p(\mathbf{z} \mid \mathbf{y}; \Phi^l) \log p(\mathbf{y} \mid \mathbf{z}; \Theta)}_{R_y(\Theta \mid \Phi^l)} + \underbrace{\sum_{\mathbf{z}} p(\mathbf{z} \mid \mathbf{y}; \Phi^l) \log p(\mathbf{z} \mid \zeta)}_{R_z(\zeta \mid \Phi^l)} \end{aligned} \tag{8}$$

The decomposition in (8) allows us to consider separately the maximization of $R_y(\Theta \mid \Phi^l)$ and $R_z(\zeta \mid \Phi^l)$:

$$\Theta^{l+1} = \arg\max_{\Theta} R_y(\Theta \mid \Phi^l)$$

$$\zeta^{l+1} = \arg\max_{\zeta} R_z(\zeta \mid \Phi^l)$$

We estimate $Q(\Theta \mid \Phi^l)$ in the E step by further developing it using equation (5):

$$\begin{aligned} Q(\Theta \mid \Phi^l) = R_y(\Theta \mid \Phi^l) &= \sum_{\mathbf{z}} p(\mathbf{z} \mid \mathbf{y}; \Phi^l) \sum_{i \in S} \log f_{z_i}(y_i; \Theta) \\ &= \sum_{i \in S} \sum_{h=1}^{K} [\log f_h(y_i; \Theta)] \; p(Z_i = h \mid \mathbf{y}; \Phi^l) \end{aligned}$$

Therefore, at each iteration we need to compute in the E step the following quantity :

$$t_{ih}^{m+1} = p(Z_i = h \mid \mathbf{y}; \Phi^l)$$

Computing this conditional probability is problematic because of the dependence between neighbouring "cubes", and an exact value cannot be obtained without considerable computing resources. As mentioned previously, we also need to approximate the normalizing constant $W(\zeta)$. Approximation methods include Besag's pseudo-likelihood [2] to compute $W(\zeta)$, and simulating the posterior distribution of Z given $\mathbf{y}$ with the parameters at iteration (l), with a Gibbs sampler to estimate t_{ih}^{m+1} [3].

However, another method exists, the mean field approximation originally proposed in the field of statistical mechanics. Since then, it has been used in a variety of fields including computer vision [34] and more recently to approximate the distribution of both $W(\zeta)$ (with a single β) and t_{ih}^{m+1} [36]. We present here the extension of this method to a model with $\zeta = (\beta_1, \ldots, \beta_K)$.

4.4 MEAN FIELD APPROXIMATIONS

The idea behind this approximation is to compute intractable quantities at any point $i \in S$ by setting the values of all the other sites in

the field to their mean values. As seen in equation (1) in the case of a MRF, this is equivalent to fixing the values of $N(i)$ only.

When computing t_{ih}^{m+1}, the mean fields approximation yields the following fixed point equation [4] for $i \in S$ and $1 \leqslant h \leqslant K$:

$$t_{ih}^{m+1} \approx \frac{f_h(y_i; \nu_h^m) \exp\{\beta_h^m \sum_{j \in N(i)} t_{jh}^{m+1}\}}{\sum_{u=1}^{K} f_u(y_i; \nu_u^m) \exp\{\beta_u^m \sum_{j \in N(i)} t_{ju}^{m+1}\}} \tag{9}$$

For the normalizing constant $W(\zeta)$, if we apply the mean-field approximation, using equation (3), we can write :

$$W(\zeta) = \sum_{\mathbf{z}'} \exp(-H(\mathbf{z}')) \approx \sum_{i \in S} \sum_{\mathbf{z}_i} \exp(-H(\mathbf{z}_i)) = \sum_{i \in S} \sum_{\mathbf{z}_i} \exp(\beta_{z_i} \sum_{N(i)} [z_i = z_j])$$

With this new set of equations, we are now able to estimate all quantities needed in the E step to maximize the model's expectation.

4.5 MAXIMIZATION

After the E step, the maximizing Φ is relatively straight forward. For Θ, once the the $t_{ih}^m = p(Z_i = h \mid \mathbf{y}; \Phi^l)$ have been computed during the E-step, we use those probabilities to assign each cell to its cluster at step l. Once the new partition is created, the maximization of Θ can be computed iteratively for cluster $h \in [1, K]$ and gene $m \in M$ with $Expr_{h,m}$ the number of cells expressing gene m in cluster h and Num_h the total number of cells in cluster h.

$$\theta_{m,h}^{l+1} = \arg\max_{\Theta} R_y(\Theta \mid \Phi^l) = \frac{Expr_{h,m}}{Num_h}$$

We then need to maximize ζ^{l+1}, to this end, we use a gradient ascent algorithm for each $\beta_h^{l+1}, h \in [1, K]$ on the function $R_z(\zeta \mid \Phi^l)$. This process is done iteratively. (CITE LAMIAE)

We are now able to compute a partition over K clusters by applying the previously described EM algorithm. However, we still need to find a way of choosing K.

4.6 ESTIMATING K

Without any prior knowledge, choosing the right number of clusters K is challenging. We decided to use an a posteriori method relying on the final log Likelihood of the model derived from equation (7):

$$\log L(\Phi) = \log P_G(\mathbf{y}, \mathbf{z} \mid \Theta, \zeta)$$

Because $\mathrm{Log}\ L(\Phi)$ monotonically increases with the number of parameters of the model, the BIC approach penalizes the addition of new parameters to the model. Let P be the total number of parameters in the model and N the cardinality of S, the BIC is expressed as:

$$-2 \log L(\Phi) + P \log N$$

By computing the final likelihood for a large range of possible K values, the minimal resulting BIC will be chosen as the optimal number of classes, $\hat{K}$ for our dataset. This approach is not ideal (see Discussion) but yields good results when applied to simulated data (see Results).

Part II

APPENDIX

A

APPENDIX

BIBLIOGRAPHY

[1] James E Balmer and Rune Blomhoff. Gene expression regulation by retinoic acid. *Journal of lipid research*, 43(11):1773–1808, 2002.

[2] Julian Besag. Statistical analysis of non-lattice data. *The statistician*, pages 179–195, 1975.

[3] Bernard Chalmond. An iterative gibbsian technique for reconstruction of< i> m</i>-ary images. *Pattern recognition*, 22(6):747–761, 1989.

[4] M. Dang and G. Govaert. Spatial fuzzy clustering using EM and markov random fields. *International Journal of System Research and Information Science*, 8(4):183–202, 1998.

[5] Alexandru S Denes, Gáspár Jékely, Patrick RH Steinmetz, Florian Raible, Heidi Snyman, Benjamin Prud'homme, David EK Ferrier, Guillaume Balavoine, and Detlev Arendt. Molecular architecture of annelid nerve cord supports common origin of nervous system centralization in bilateria. *Cell*, 129(2):277–288, 2007.

[6] Adriaan WC Dorresteijn. Quantitative analysis of cellular differentiation during early embryogenesis ofplatynereis dumerilii. *Roux's archives of developmental biology*, 199(1):14–30, 1990.

[7] EA Feingold, PJ Good, MS Guyer, S Kamholz, L Liefer, K Wetterstrand, FS Collins, TR Gingeras, D Kampa, EA Sekinger, et al. The encode (encyclopedia of dna elements) project. *Science*, 306 (5696):636–640, 2004.

[8] Albrecht Fischer and Adriaan Dorresteijn. The polychaete platynereis dumerilii (annelida): a laboratory animal with spiralian cleavage, lifelong segment proliferation and a mixed benthic/pelagic life cycle. *Bioessays*, 26(3):314–325, 2004.

[9] Antje Fischer, Thorsten Henrich, and Detlev Arendt. The normal development of platynereis dumerilii (nereididae, annelida). *Frontiers in zoology*, 7(1):31, 2010.

[10] Clay Fuqua, Matthew R Parsek, and E Peter Greenberg. Regulation of gene expression by cell-to-cell communication: acyl-homoserine lactone quorum sensing. *Annual review of genetics*, 35(1):439–468, 2001.

[11] Manfred Gossen and Hermann Bujard. Tight control of gene expression in mammalian cells by tetracycline-responsive promoters. *Proceedings of the National Academy of Sciences*, 89(12): 5547–5551, 1992.

[12] Valentin Häcker. *Die pelagischen Polychaeten-und Achaetenlarven der Plankton-expedition...* Lipsius & Tischer, 1898.

[13] Jörg D Hardege. Nereidid polychaetes as model organisms for marine chemical ecology. *Hydrobiologia*, 402:145–161, 1999.

[14] Carl Hauenschild, G Czihak, A Fischer, and R Siewing. *Platynereis dumerilii: mikroskopische Anatomie, Fortpflanzung, Entwicklung*. Fischer, 1969.

[15] Thomas H Hutchinson, Awadhesh N Jha, and David R Dixon. The polychaete platynereis dumerilii (audouin and milne-edwards): a new species for assessing the hazardous potential of chemicals in the marine environment. *Ecotoxicology and environmental safety*, 31(3):271–281, 1995.

[16] HA John, ML Birnstiel, and KW Jones. Rna-dna hybrids at the cytological level. *Nature*, 223(5206):582, 1969.

[17] JE Landegent, De Wal, N Jansen In, RA Baan, JHJ Hoeijmakers, and M Van Der Ploeg. 2-acetylaminofluorene-modified probes for the indirect hybridocytochemical detection of specific nucleic acid sequences. *Experimental cell research*, 153(1):61–72, 1984.

[18] Geoffrey McLachlan and David Peel. *Finite mixture models*. Wiley. com, 2004.

[19] G Mendel. Versuche ber pflanzen-hybriden. verh. *Naturforsch. Ver. Brnn*, 4:347, 1866.

[20] Ryan D Morin, Matthew Bainbridge, Anthony Fejes, Martin Hirst, Martin Krzywinski, Trevor J Pugh, Helen McDonald, Richard Varhol, Steven JM Jones, and Marco A Marra. Profiling the hela s3 transcriptome using randomly primed cdna and massively parallel short-read sequencing. *Biotechniques*, 45(1):81, 2008.

[21] PM Nederlof, D Robinson, R Abuknesha, J Wiegant, AHN Hopman, HJ Tanke, and AK Raap. Three-color fluorescence in situ hybridization for the simultaneous detection of multiple nucleic acid sequences. *Cytometry*, 10(1):20–27, 1989.

[22] Claus Nielsen. Trochophora larvae: Cell-lineages, ciliary bands, and body regions. 1. annelida and mollusca. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 302(1): 35–68, 2004.

[23] Mary Lou Pardue and Joseph G Gall. Molecular hybridization of radioactive dna to the dna of cytological preparations. *Proceedings of the National Academy of Sciences*, 64(2):600–604, 1969.

[24] D Pinkel, J Landegent, C Collins, J Fuscoe, R Segraves, J Lucas, and J Gray. Fluorescence in situ hybridization with human chromosome-specific libraries: detection of trisomy 21 and translocations of chromosome 4. *Proceedings of the National Academy of Sciences*, 85(23):9138–9142, 1988.

[25] Greg W Rouse. Trochophore concepts: ciliary bands and the evolution of larvae in spiralian metazoa. *Biological Journal of the Linnean Society*, 66(4):411–464, 1999.

[26] Gideon Schwarz. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

[27] Kazuo Shinozaki, Kazuko Yamaguchi-Shinozaki, and Motoaki Seki. Regulatory network of gene expression in the drought and cold stress responses. *Current opinion in plant biology*, 6(5):410–417, 2003.

[28] RR SWIGER and JD TUCKER. Fluorescence in situ hybridization: A brief review. *Environmental and molecular mutagenesis*, 27 (4):245–254, 1996.

[29] Kristin Tessmar-Raible and Detlev Arendt. Emerging systems: between vertebrates and arthropods, the lophotrochozoa. *Current opinion in genetics & development*, 13(4):331–340, 2003.

[30] Kristin Tessmar-Raible, Florian Raible, Foteini Christodoulou, Keren Guy, Martina Rembold, Harald Hausen, and Detlev Arendt. Conserved sensory-neurosecretory cell types in annelid and fish forebrain: insights into hypothalamus evolution. *Cell*, 129(7):1389–1400, 2007.

[31] Raju Tomer, Alexandru S Denes, Kristin Tessmar-Raible, and Detlev Arendt. Profiling by image registration reveals common origin of annelid mushroom bodies and vertebrate pallium. *Cell*, 142(5):800–809, 2010.

[32] Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 2009.

[33] Fa-Yueh Wu. The potts model. *Reviews of modern physics*, 54(1): 235, 1982.

[34] Alan L Yuille. Generalized deformable models, statistical physics, and matching problems. *Neural Computation*, 2(1):1–24, 1990.

[35] Erich Zeeck, Tilman Harder, and Manfred Beckmann. Uric acid: the sperm-release pheromone of the marine polychaete platynereis dumerilii. *Journal of Chemical Ecology*, 24(1):13–22, 1998.

[36] Jun Zhang. The mean field theory in em procedures for markov random fields. *Signal Processing, IEEE Transactions on*, 40(10): 2570–2583, 1992.

COLOPHON

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede. The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*". `classicthesis` is available for both LaTeX and LyX:

http://code.google.com/p/classicthesis/

Happy users of `classicthesis` usually send a real postcard to the author, a collection of postcards received so far is featured here:

http://postcards.miede.de/

DECLARATION

This thesis:

- is my own work and contains nothing which is the outcome of work done in collaboration with others, except where specied in the text;
- is not substantially the same as any that I have submitted for a degree or diploma or other qualication at any other university; and
- does not exceed the prescribed limit of 60,000 words.

Cambridge, 2014

Jean-Baptiste Olivier
Georges Pettit