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

Clustering and visualizing functionnal tissues in *P. dumerillii*

JEAN-BAPTISTE OLIVIER GEORGES PETTIT

UNIVERSITY OF CAMBRIDGE

2014 – version 0.9

Ohana means family.
Family means nobody gets left behind, or forgotten.

— Lilo & Stitch

Dedicated to the loving memory of Rudolf Miede.
1939 – 2005

ABSTRACT

This is wehere the abstarct will go...

PUBLICATIONS

Some ideas and figures have appeared previously in the following publications:

Put your publications from the thesis here. The packages multibib or bibtopic etc. can be used to handle multiple different bibliographies in your document.

Le temps ne fais rien a l'affaire...

— George Brassens

ACKNOWLEDGMENTS

My thanks will go there

CONTENTS

LIST OF FIGURES

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

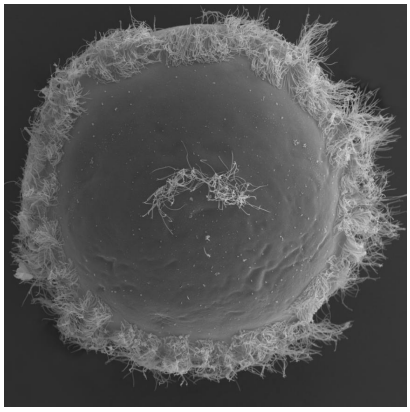
CAPTURING GENE EXPRESSION IN *PLATYNEREIS DUMERILLII*'S BRAIN

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 [11, 16, 9, 5, 6, 7]. As a member of the bilateria *P. dumerillii* has a defined bilateral symmetry.

*P. dumerillii*populates 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 [7]. 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 [4, 17] 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 [6]. 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 [21]. 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 as seen on figure 1a [14, 13]. the metotrochophore stage is characterized by the development of a slightly elongated segmented trunk compared to that of the trochophore [8]. 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 [8]. This traditional subdivision has been applied to *P. dumerillii* [10].

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 [6]. An example showing the similarity between individuals during de-

velopment can be seen on figure 2.

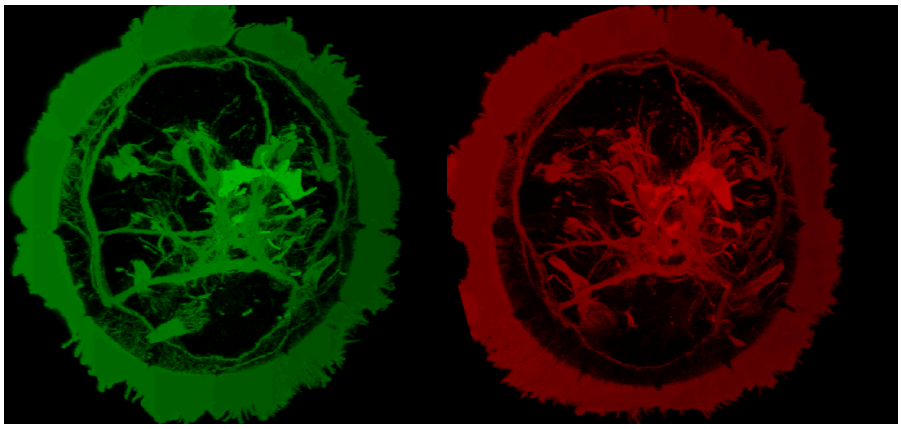


Figure 2: *Platynereis dumerillii*'s stereotypical and synchronous development. In green and red are two different *P. dumerillii* individuals' with the same gene expression being highlighted. They show extremely similar patterns of development.

2.2 GENE EXPRESSION IN PLATYNEREIS' DEVELOPING BRAIN

2.2.1 *Mechanisms of gene expression*

- Generalities about gene expression - Genes, transcription factors, all cells have the same genome expression differs

2.2.2 *Platynereis' brain development*

- General development pattern (symmetries, eyes, mushroom) - LOOK for a paper summing that up - State of the brain at 48hpf because we'll use it later

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

- Some tissues will be spatially well defined others will be scattered - Example with the pancreas ? - Introducing the idea of the spacial coherency heterogeneity

2.3 CAPTURING GENE EXPRESSION IN THE LABORATORY

2.3.1 *In-situ hybridization assays*

- FIG 2 : In situ hybridization principles - Explanations about the technique

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

- Stereotypical development allows one gene to be considered as reference - Different individuals are "replicates" - Mapping to a scaffold created by the reference gene

2.3.3 *RNA sequencing*

- 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 [19] 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: $\Theta = (\alpha_1, \ldots, \alpha_K)$. It is interesting to note that this part of the model is equivalent to an independent mixture model [12]. 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. [18] 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 :

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

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

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

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 [15].

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 [1] 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} [2].

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 [20] and more recently to approximate the distribution of both $W(\zeta)$ (with a single β) and t_{ih}^{m+1} [22]. 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 [3] 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] Julian Besag. Statistical analysis of non-lattice data. *The statistician*, pages 179–195, 1975.

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

[3] 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.

[4] 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.

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

[6] 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.

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

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

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

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

[11] 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.

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

[13] 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.

[14] 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.

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

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

[17] 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.

[18] 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.

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

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

[21] 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.

[22] 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