LU-TP 22-20
March 2022

Modelling of gene regulatory and spatial mechanisms which pattern the early human neural tube

Fredrik Bertilsson

Department of Astronomy and Theoretical Physics,
Lund University

Master thesis supervised by Victor Olariu

LUND
UNIVERSITY

Abstract

Tissue patterning during development is thought to be driven largely by gene regulatory responses to spatial gradients of molecular signals, known as morphogens. In this thesis, early patterning of the human neural tube is studied using two different gene regulatory network models. The models portray patterning in the rostral-caudal direction as a response to gradients of WNT and FGF signalling. The first model is based on a tri-stable network topology which describes patterning into forebrain, midbrain, and hindbrain regions; the second model describes the formation and maintenance of the isthhmic organizer, and also subsequent midbrain formation. The two models were optimized against time-resolved bulk and single-cell RNA sequencing data from human embryonic stem cells undergoing 14 days of patterning in an *in vitro* microfluidics-based setup. In the setup, a GSK3 inhibitor is used to mimic the patterning effects of the *in vivo* early neural tube WNT signalling gradient. The resulting RNA data are in the form of time series over multiple GSK3 inhibitor concentrations. Clustering of the RNA data reveals that it is meaningful to group genes into the forebrain, midbrain, and hindbrain categories of the tri-stable model. It also reveals additional gene co-expression patterns within these categories, which is useful for further modelling work. The optimized network models are studied in reaction-diffusion simulations of WNT and FGF morphogen driven patterning in both a realistic 3D neural tube geometry, and in a 3D model of the tissue from the microfluidics setup. It is found that patterning of the tri-stable model is significantly influenced by WNT production, decay, and diffusion mechanisms, and that the WNT secretion locations of an earlier spatial 3D tube model, used as a foundation for this work, should be modified in order to produce a WNT gradient which is more similar to the *in vivo* WNT gradient shape. The isthmic organizer network model is found to accurately describe isthmic organizer formation at the intersection of regions with high OTX2 and GBX2 gene expression. Furthermore, the model isthmic organizer, once formed, maintains itself in the absence of an external WNT signalling gradient. WNT and FGF diffusion from the model isthmic organizer is found to induce midbrain identity adjacent to the organizer in approximately the right position. Overall, both models capture several key patterning features, and they will serve as useful stepping stones towards even more refined neural patterning models.

Populärvetenskaplig beskrivning

Hur en enda befruktad äggcell kan utvecklas till ett igenkännligt foster över loppet av några månaders tid är ett fascinerande mysterium. Vi vet nog alla att celler kan dela sig, och att det är på så vis som våra kroppar växer under fosterstadiet. Men en cell som delar sig kommer ge upphov till två *identiska* dotterceller. Hur är det då möjligt att en äggcell kan ge upphov till alla de helt olika sorters celler som finns i ett foster redan någon månad in i dess utveckling? Hjärnceller ser ju helt annorlunda ut än exempelvis blodceller, och de har helt skilda ”arbetsuppgifter” i kroppen. Alltså måste det vara något mer än celldelning i görningen för att ett foster ska kunna utvecklas på rätt sätt.

En stor del av svaret på utvecklingens gåta finns i våra gener. Många vet att generna innehåller all information om vad cellerna ska göra, men det är nog inte lika många som vet att inte alla gener i en cell är ”påslagna” samtidigt. Lite förenklat kan man säga att det är vilka gener som är av och påslagna som avgör vad cellen gör för något – därför har varje sorts cell sitt egna ”fingeravtryck” av vilka gener som är av och på. En hjärncell har alltså helt andra gener som är påslagna än blodceller, och det är på grund av detta som de skiljer sig så mycket åt.

Intressant nog finns det gener vars viktigaste uppgift tycks vara att slå av eller på andra gener. Man kan då tänka sig att vissa gener är viktigare än andra: om en viss viktig gen exempelvis slås på av en annan, så kanske det får så stora konsekvenser för cellen att den blir till en helt annan sorts cell. Det är precis det som sker vid åtskilliga tillfällen i ett fosters utveckling: vid vissa tidpunkter byter några av fostrets celler identitet på grund av hur vissa viktiga gener påverkar varandra. För att förstå vad som händer kan man rita ett sorts släktträd för cellerna: högst upp har vi den befruktade äggcellen, som tidigt delas upp i olika grupper (”trädgrenar”) av celler baserat på vad som händer med deras gener. Varje cellgrupp kan i sin tur utvecklas till alltmer specifika celltyper, i ett eller flera steg. Den här processen kallas för *differentiering*.

Men det räcker inte med differentiering för att ett foster ska utvecklas rätt. Vi vill ju inte bara ha en massa olika sorters celler i en enda röra: hur skulle det se ut? Nej, varje celltyp måste dessutom hamna på rätt plats i kroppen, vilket kallas för *mönsterbildning*. Detta kan exempelvis ske genom att celler som befinner sig nära varandra skickar olika sorters signaler till varandra.

Differentiering och mönsterbildning är väldigt komplexa processer. För att förstå dem bättre så har biologer på senare år börjat ta alltmer hjälp av fysiker, som är duktiga på att *modellera* krångliga saker – det vill säga, de kan förenkla något krångligt så att det går att förstå, förhoppningsvis utan att tappa alltför mycket av de intressanta detaljerna. De förenklade modellerna kan sedan användas i till exempel datorsimuleringar. Detta arbete går ut på att göra just sådana simuleringar, utgående från modeller av mönsterbildning och differentiering i ett visst stadium av den tidiga mänskliga hjärnans utveckling. Tanken är att testa hur väl modellerna beskriver resultatet av nya experiment, och att i ljuset av dessa experiment försöka göra modellerna mer exakta och användbara. Förhoppningen är att detta i längden kan hjälpa oss att bättre förstå vilka gener som är viktiga för differentiering i just det här stadiet av hjärnans utveckling, och hur celler kommunicerar med varandra för att den tidiga hjärnans mönsterbildning ska ske på rätt sätt.

Contents

List of acronyms

FB	Forebrain
MB	Midbrain
HB	Hindbrain
SC	Spinal cord
hESC	Embryonic stem cell
iPSC	Induced pluripotent stem cell
WNT	Wingless-Int1
MHB	Midbrain-hindbrain boundary
IsO	Isthmic organizer
SHH	Sonic hedgehog
DV	Dorsal-ventral
RC	Rostral-caudal
GSK3	Glycogen synthase kinase 3
MiSTR	Microfluidic-controlled stem cell regionalization
sc-RNAseq	Single-cell RNA sequencing
ODE	Ordinary differential equation
DNA	Deoxyribonucleic acid
mRNA	Messenger ribonucleic acid
ESC	Embryonic stem cell
TF	Transcription factor
GRN	Gene regulatory network
S-A	Shea-Ackers
PCA	Principal component analysis
SNN	Shared nearest neighbours
UMAP	Uniform manifold approximation and projection
DE	Differential evolution
L-BFGS-B	Limited memory Broyden–Fletcher–Goldfarb–Shanno box constrained

1 Introduction

Many of the processes which underlie the development of organisms from embryo to adult are currently not well understood. The general picture is clear: stem cells in the embryo proliferate via cell division, and over time the stem cells start differentiating into the full variety of cell types found in the adult body. However, the mechanisms responsible for the cellular changes that occur during particular stages of development are often unknown. In particular, it is not always clear how stem cells mature into different cell types depending on where in a developing embryo they are located. This kind of spatially distinct differentiation is known as *patterning*.

A particular type of cell is characterized by which of its genes are expressed and which are not. Any viable patterning mechanism must thus somehow regulate gene expression. Moreover, it must do so in a spatially distinct manner. One mechanism capable of this is the action of morphogens. A morphogen is a signalling molecule which is secreted from specific types of source cells, which are found at *organizers* - morphogen source locations. The morphogen is then transported inter-cellularly, often via diffusion [1], which leads to a concentration gradient. Cells close to the morphogen source are then heavily affected by it, and cells far away less so. Morphogens have been shown to play key roles in several important *in vivo* developmental patterning events [2], [3]. Figure 1.1 below illustrates how a morphogen gradient can split a tissue into regions with distinct cell types.

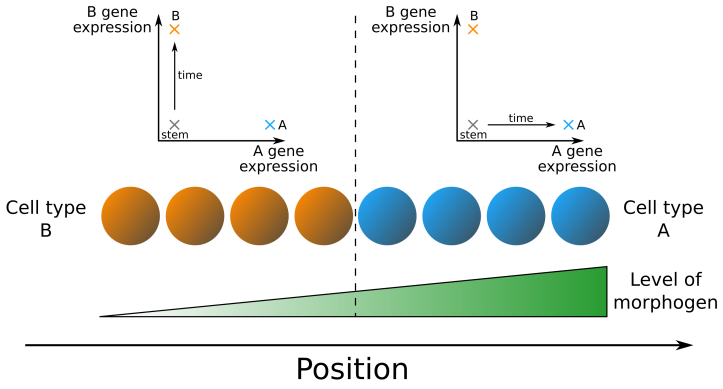


Figure 1.1: Conceptual sketch which illustrates morphogen-driven patterning of stem cells into distinct cell types A (blue) and B (orange) over time. The morphogen source is located to the right. Stem cells differentiate into cell type A where they are exposed to high levels of morphogen - this activates genes characteristic of cell type A. Cell type B is induced at positions where morphogen levels are low. The morphogen gradient thus establishes a boundary between different kinds of tissue.

One example of developing tissue which undergoes morphogen-driven patterning is the human central nervous system. It is comprised of the brain and the spinal cord, and they develop from a common progenitor tissue called the *neural tube*. The subdivision into brain and spinal cord regions is one of the earliest neural tube patterning events, together with the subdivision of the brain into forebrain (FB), midbrain (MB), and hindbrain (HB) regions (figure 1.2, on the following page) - in human, these events occur a few weeks into development [4].

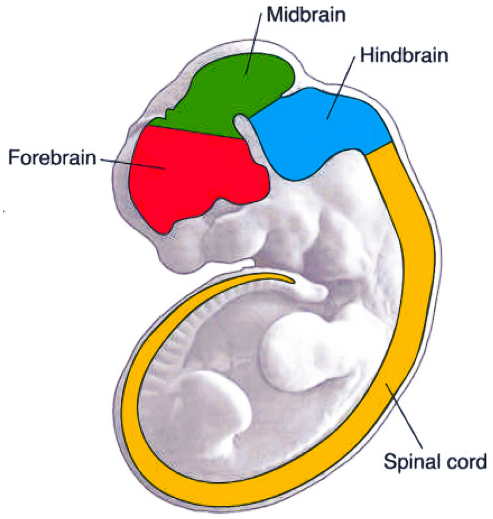


Figure 1.2: A simplified picture of a human fetus a few weeks into development, with the neural tube highlighted in four different colours. Each colour signifies a region of the tube which contains a distinct cell type - forebrain (FB), midbrain (MB), hindbrain (HB), and spinal cord (SC). These regions arise due to a spatial variation in cell differentiation fates - patterning. Figure obtained from [4].

Neural tube patterning has lately attracted attention due to the emergence of promising stem cell-based therapies for Parkinson's disease [5]. This disease is characterized by the degradation of dopaminergic neurons. The idea behind the emerging therapies is to replace a patient's own degraded dopaminergic neurons with new, fully functioning ones. Dopaminergic neurons can be obtained using directed *in vitro* differentiation of human embryonic stem cells (hESC:s) [6]. Instead of starting from hESC:s, an alternative approach is to sample some of the patient's own fully differentiated cells and reprogram them into stem cells, then known as induced pluripotent stem cells (iPSC:s) [7]. With either approach, one must understand what kinds of extra-cellular conditions are required to drive differentiation in the dopaminergic neuron direction. This is the connection to neural tube patterning - *in vivo* progenitors of dopaminergic neurons are produced in the midbrain region of the neural tube. A better understanding of what leads to the FB / MB / HB sub-division of the developing brain is thus highly relevant for designing more efficient *in vitro* dopaminergic neuron differentiation protocols.

Morphogens are known to be involved in neural tube patterning [8]. However, it is less clear which genes the morphogens regulate, and how. Furthermore, morphogen target genes may also regulate each other's expression in complex ways. To better understand these regulatory mechanisms, FB / MB / HB patterning has been studied in *in vitro* experiments, using substances which emulate the action of the *in vivo* morphogen [9]. These experiments have resulted in gene expression data sets for the genes known to be involved in the studied patterning event. In turn, these data sets have been useful for inference of likely regulatory connections [10]. This has made it easier to set up realistic computational models of neural tube patterning, as exemplified by the work of Brambach

et al. (2021) [11]. However, until recently, no time-resolved gene expression data sets of *in vitro* FB / MB / HB patterning were available. This changed in 2020 with the publication of Rifes et al. (2020) [12]. Time-resolved data makes it possible to conduct a more detailed co-expression analysis, and also to perform more sophisticated reality checks of computational models.

The goal of this thesis work is to refine and extend already existing models of human neural tube patterning. The core of the work is two different gene regulatory network models. The models are used to simulate patterning of the brain region in the neural tube (figure 1.2), mainly along the rostral - caudal (head-tail) tube axis. The first model is a slightly revised version of the model in [10] - it is now assessed in more detail, in light of the new experimental data from Rifes et al. (2020) [12]. The second model is more novel and aims to describe the induction of the MB / HB boundary, also known as the *isthmic organizer*, and the ensuing patterning of the MB region. Both models, once optimized towards experimental data, are used in 3D lattice models to simulate patterning in relevant geometries. This makes it possible to study how morphogen diffusion, production, and decay mechanisms influence the patterning.

2 Background

2.1 Neural tube patterning

The first patterning event of a human embryo is the process of *gastrulation*, where embryonic stem cells differentiate into three basic cell lineages: the germ layers *ectoderm, mesoderm* and *endoderm* (for more details, see appendix A.1). Ectodermal progenitors then form a slab of tissue called the *neural plate*, which subsequently proceeds to fold in on itself to form the neural tube [4]. The initial cylindroid morphology of the tube is then altered into the curved, more irregular shape that was shown in figure 1.2.

Morphogen-driven neural patterning starts already at the neural plate stage. The main morphogens responsible for it are a set of closely related signalling proteins, which are encoded by genes from the Wingless-Int1 (WNT) family. The subdivision into forebrain (FB), midbrain (MB), and hindbrain (HB) regions along the rostral-caudal (RC) axis is largely driven by a concentration gradient of these WNT morphogens, such that the caudal (HB) end experiences higher levels of WNT signals than the rostral (FB) end (figure 2.1, below). By yet unknown mechanisms, this concentration gradient is formed via signalling from a tissue known as the paraxial mesoderm, which lies adjacent to (paraxial with) the neural plate / tube along the rostral-caudal axis [13]. Hence, the paraxial mesoderm acts as an organizer. Neural tube cells which are more caudally positioned than HB develop into the spinal cord (SC). It is becoming increasingly clear that SC cell identity is established in a process separate from the FB/MB/HB sub-division [14] - hence, patterning of the SC is not an area of focus in this thesis.

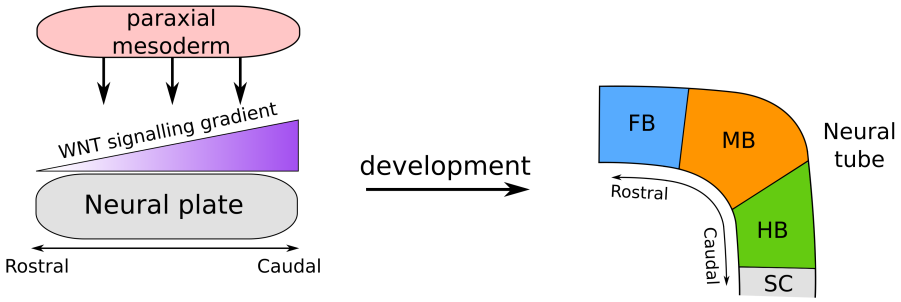


Figure 2.1: Simplified sketch which shows key events in early neural patterning. Paraxial mesoderm tissue induces a WNT signalling gradient in the neural plate, which is sustained for some time while the neural plate folds in on itself and forms the neural tube. The WNT gradient results in rostral-caudal patterning of the tube into forebrain, midbrain, and hindbrain.

Both WNT and other morphogens continue to play a role in neural tube patterning after the initial mesodermal WNT gradient has established the rostral-caudal axis. As the tube continues to develop, *secondary* organizers start to appear through various mechanisms. Morphogens secreted from them further refines the tube pattern. One of the most significant secondary organizers is the *isthmic organizer* (IsO), which forms at the midbrain-hindbrain boundary (MHB). Since the function of the IsO and the mechanisms which leads to its induction are the subject of the second model in this thesis, the IsO is

discussed more thoroughly in connection with that model in section 4. Other organizer examples are the *roof plate*, which secretes WNT, and the *floor plate*, which secretes the morphogen Sonic Hedgehog (SHH). SHH patterns the tube along the dorsal-ventral (DV, back-to-belly) tube axis.[1] Figure 2.2 shows the approximate locations of these secondary organizers in the tube, along with the main morphogens they produce.

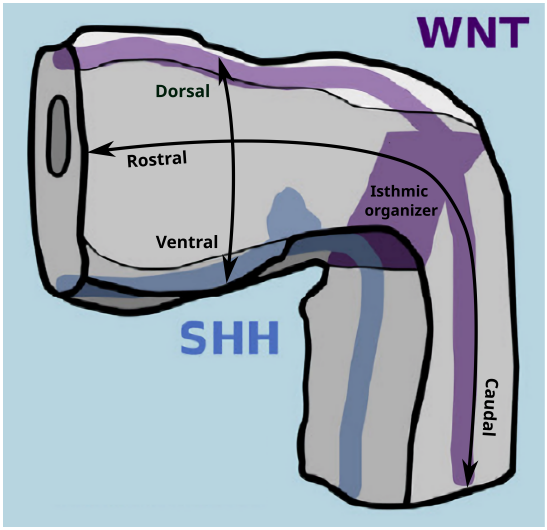


Figure 2.2: Patterning axes of the neural tube, together with notable WNT (purple) and SHH (blue) secondary organizer locations. Figure obtained from [11], modified by addition of patterning axes.

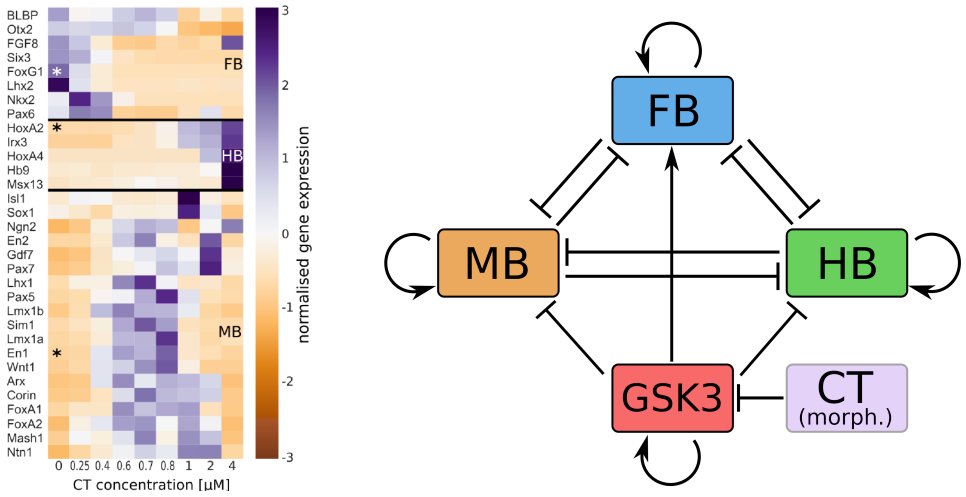
Patterning along both the rostral-caudal (RC) and DV axes have been modelled in the context of gene regulatory networks affected by spatial gradients of morphogen signals. As this thesis focuses on patterning along the RC axis, key results of Brambach et al. (2021) [11] pertaining to RC patterning are reviewed in the next subsection.

2.2 Previous modelling of rostral-caudal patterning

Up until recently, there was no regulatory network model which comprehensively described the full sub-division into the three regions FB, MB, and HB. In Brambach et al. (2021) [11], such a model was formulated. It is based on experimental gene expression data from Kirkeby et al. (2012) [9], where human neural tube progenitor cells were cultivated *in vitro* in the presence of different levels of a particular glycogen synthase kinase 3 (GSK3) inhibitor known as CT. GSK3 inhibition emulates the effects of WNT morphogen signalling (see appendix A.2). In the data it was observed that low CT levels induced genes characteristic of FB cell identity, and that higher CT levels induced MB / HB identity genes. Brambach et al. (2021) [11] formalized this observation by hierarchically clustering the data, thereby demonstrating that it appears meaningful to organize genes in three clusters, corresponding to genes whose levels of expression are elevated in FB, MB, and HB cells. Based on the clustering results, a regulatory network model of

[1] SHH does not play a role in the models in this thesis, but it is mentioned for completeness since it plays an important role in a DV patterning model considered in Brambach et al. (2021) [11].

patterning into FB, MB, and HB regions was formulated. The gene expression data from Kirkeby et al. (2012) [9] is shown in figure 2.3(a), and the regulatory network model is shown in figure 2.3(b). The network notation is explained in appendix B.1.



(a) (b)

Figure 2.3: (a) Gene expression levels from Kirkeby et al. (2012) [9] for different CT levels. Note the clear trend of how the expression profile becomes increasingly caudal with increasing levels of CT - i.e. with increasing levels of emulated WNT signalling. (b) The tri-stable regulatory network constructed by Brambach et al. (2021) [11] based on the data in (a). CT, which emulates the morphogen action of WNT, inhibits GSK3, which in turn induces FB and inhibits MB and HB - hence, high levels of CT / WNT will induce a MB / HB fate. Sub-figure (a) was obtained from [11].

The FB, MB, and HB sub-network topology is a generalized bi-stable switch network (see appendix E.2). It is henceforth referred to as the tri-stable network, as there are three distinct patterning regions which are represented as nodes in the network. Since these FB, MB, and HB nodes represent regions with a particular expression profile, and not actual genes, the network does not necessarily indicate how FB, MB, and HB genes mechanistically regulate each other - it merely intends to capture the overall variation of expression profiles with CT levels.

Brambach et al. (2021) [11] performed simulations of the tri-stable network model dynamics, and optimized the model against individual genes in the experimental data (figure 2.3(a)) as FB, MB, and HB network nodes. Simulations with the optimized model were generally in good agreement with data. However, it was difficult to evaluate the fit quality, as data were only available for a single time point. As was discussed in the introduction to this thesis, time-resolved data from a similar experiment would allow an assessment of how accurately the network can capture experimentally observed patterning dynamics. That would in turn both aid in uncovering more details of the underlying regulatory mechanisms, and help constrain which FB / MB / HB genes are most reasonable to use as FB / MB / HB nodes in the network. Furthermore, data from additional, earlier, time points would be helpful for correct initialization of the model. Fortunately, these kinds of data are now available from the work of Rifes et al. (2021) [12].

2.3 New experimental patterning data permits more sophisticated modelling

The research group which supplied the data used in Brambach et al. (2021) [11] have continued to study neural patterning. In Rifes et al. (2020) [12], a microfluidics system termed "microfluidic-controlled stem cell regionalization" (MiSTR) was used to set up a linear concentration gradient of the GSK3 inhibitor CT across a strip of hESC tissue. Thereby, patterning due to the *in vivo* mesodermal WNT gradient (figure 2.1) was emulated *in vitro*: spatially continuous rostral-caudal patterning of the hESC tissue was achieved over time during the MiSTR experiment. The experiment was run for 14 days, and the CT gradient was removed at day 9. Bulk and single-cell RNA sequencing (sc-RNAseq) gene expression data which was collected at multiple time points revealed a clear spatial trend in which genes were expressed, consistent with FB / MB / HB patterning as described in section 2.1. Strikingly, the induced pattern remained at day 14, five days after the removal of the CT gradient. Figure 2.4 below outlines the time course of the MiSTR experiment in (a), and the data collection protocol in (b).

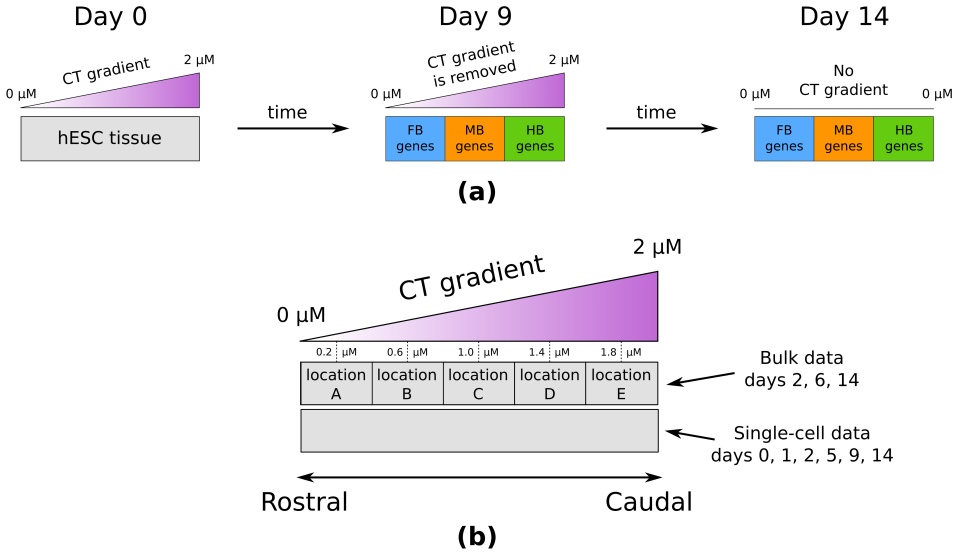
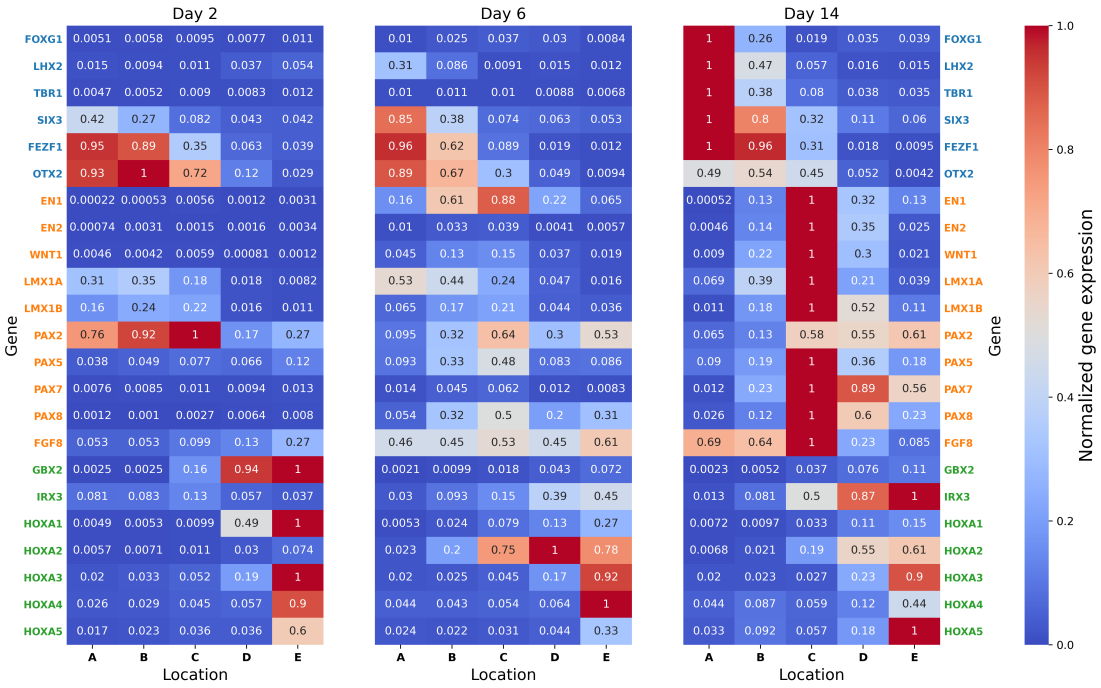
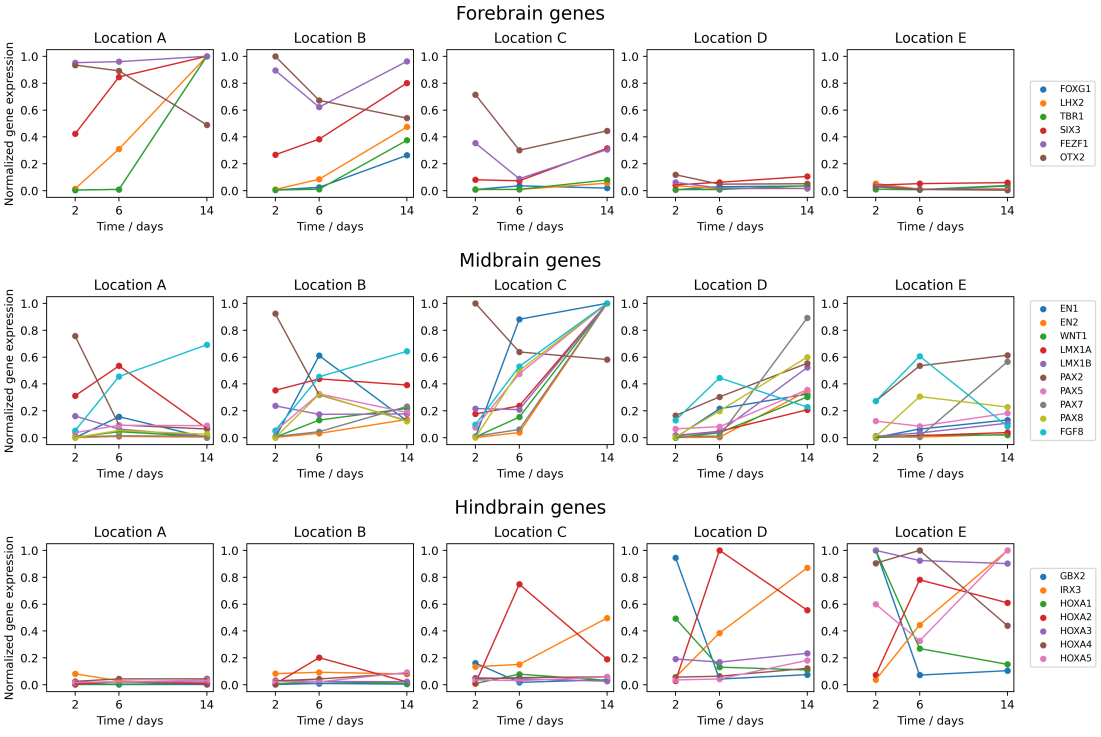


Figure 2.4: The MiSTR experiment in Rifes et al. (2020) [12]. (a): Time course of the experiment. The gradient of the GSK3 inhibitor CT was present for the first 9 days, which induced rostral-caudal patterning of the hESC tissue. After removal of the CT gradient at day 9, the induced pattern persisted. (b) Outline of the MiSTR data collection protocol. The patterned tissue was sub-divided into an upper and a lower half along the CT gradient. From one of the tissue halves, gene expression was measured for all individual cells using sc-RNAseq. The other tissue half was sub-divided into five spatial regions A (most rostral) to E (most caudal), each corresponding to a particular average level of CT. Average (bulk) gene expression levels were measured in each of these regions.

Both bulk and sc-RNAseq data from the MiSTR experiment were used for model optimization in this thesis. Figure 2.5 on the following page shows the bulk data obtained from the different MiSTR locations A-E (figure 2.4(b)). Data are shown as both heat maps and as time series. The data have been processed and normalized: see appendix C for a detailed description of these processing steps.



(a)



(b)

Figure 2.5: Normalized and averaged bulk gene expression data for all MiSTR experiment locations across days 2, 6, and 14 of patterning. (a): Heat map of gene expression distribution across MiSTR locations A-E for each time point. Gene names are coloured like in figure 2.4, such that FB, MB, and HB marker genes are shown as blue, orange, and green, respectively. (b): Time series of FB, MB, and HB marker genes in each MiSTR location.

Based on figure 2.5, gene expression patterns are in agreement with the FB, MB, and HB gene categories: FB genes are mostly expressed in the rostral-most locations, HB genes in the caudal-most locations, and so on. Hence, this new data set is consistent with the data from Brambach et al. (2021) [11] in figure 2.3(a). To obtain a more quantitative measure of similarity between expression patterns of different genes across locations, the bulk and sc-RNAseq data sets were clustered using a bottom-up, hierarchical clustering algorithm (see appendix F).

3 Model 1: rostral-caudal FB-MB-HB-patterning

The first model considered in this thesis is the tri-stable rostral-caudal patterning model from Brambach et al. (2021) [11], which was described in section 2.2. The purposes behind investigating it further are to

- Optimize the model against the new time-resolved gene expression data sets from Rifes et al. (2021) [12], described in section 2.3, and assess the fit quality for different sets of (FB, MB, HB)-genes.
- Perform spatial simulations of how the optimized model patterns in a reaction-diffusion model of WNT morphogen driven patterning in a realistic 3D neural tube geometry, given well-known WNT secretion locations. This was done in Brambach et al. (2021) [12] as well, and it would be interesting to see whether or not simulations of the same model, but optimized against different (and better) data, reproduces their results. Furthermore, many aspects regarding how WNT production, decay, and diffusion mechanisms influence the patterning were left unexplored in Brambach et al. (2021) [11], so they are explored in this thesis.

3.1 Model equations and optimization

The tri-stable regulatory network topology in figure 2.3 was used to write down ordinary differential equations (ODE:s) which describe the evolution of FB, MB, HB, GSK3, and CT node expression levels over time t. Employing a standard Shea-Ackers formalism (see appendix B.2), the tri-stable network ODE:s were written as in equations (3.1) on top of the following page. Expression levels are written as "[expression level]". Appendix B.2 describes the physical interpretation of the different types of parameters b, c, δ, and n. Note that CT levels were assumed to be steady-state, as was the case for the first 9 days of the MiSTR experiment. Also in accordance with experiments, the CT gradient was fully removed (i.e. put = 0) for $t >$ day 9.

By numerically solving eq. (3.1) and comparing the simulated levels with data, the tri-stable model was optimized. As the FB, MB, and HB nodes represent brain regions which are characterized by high expression levels of *several* genes, a specific FB/MB/HB gene combination had to be chosen during a particular optimization run. Since the data

$$
\begin{aligned}
\frac{d[\mathrm{FB}]}{dt} &= \frac{b_1 + c_1[\mathrm{GSK3}]^{n_1} + c_2[\mathrm{FB}]^{n_2}}{1 + c_1[\mathrm{GSK3}]^{n_1} + c_2[\mathrm{FB}]^{n_2} + c_3[\mathrm{MB}]^{n_3} + c_4[\mathrm{HB}]^{n_4}} - \delta_1[\mathrm{FB}], \\
\frac{d[\mathrm{MB}]}{dt} &= \frac{b_2 + c_6[\mathrm{MB}]^{n_6}}{1 + c_5[\mathrm{GSK3}]^{n_5} + c_6[\mathrm{MB}]^{n_6} + c_7[\mathrm{FB}]^{n_7} + c_8[\mathrm{HB}]^{n_8}} - \delta_2[\mathrm{MB}], \\
\frac{d[\mathrm{HB}]}{dt} &= \frac{b_3 + c_{10}[\mathrm{HB}]^{n_{10}}}{1 + c_9[\mathrm{GSK3}]^{n_9} + c_{10}[\mathrm{HB}]^{n_{10}} + c_{11}[\mathrm{FB}]^{n_{11}} + c_{12}[\mathrm{MB}]^{n_{12}}} - \delta_3[\mathrm{HB}], \\
\frac{d[\mathrm{GSK3}]}{dt} &= \frac{c_{13}[\mathrm{GSK3}]^{n_{13}}}{1 + c_{13}[\mathrm{GSK3}]^{n_{13}} + c_{14}[\mathrm{CT}]^{n_{14}}} - \delta_4[\mathrm{GSK3}], \\
\frac{d[\mathrm{CT}]}{dt} &= 0.
\end{aligned}
\tag{3.1}
$$

sets described in section 2.3 contain 6 FB, 10 MB, and 7 HB genes, optimization would ideally be performed for all $6 \cdot 10 \cdot 7 = 420$ FB/MB/HB combinations. However, optimization was only performed for a subset of combinations due to time constraints: given the challenging nature of the optimization task at hand, a single run took several hours to converge.

Optimization was formulated as a search for an optimal set of model parameters $\mathbf{p}$ which would result in simulations agreeing as well as possible with data. This was achieved by minimizing a particular cost function $C(\mathbf{p})$:

$$
C(\mathbf{p}) = \sum_g \sum_l \sum_t \left(s_{gl}(\mathbf{p}, t) - d_{gl}(t)\right)^2 + \sum_g \sum_l \alpha \left(\left. \frac{\partial s_{gl}(\mathbf{p}, t)}{\partial t} \right|_{t=\text{day } 14} \right)^2 . \tag{3.2}
$$

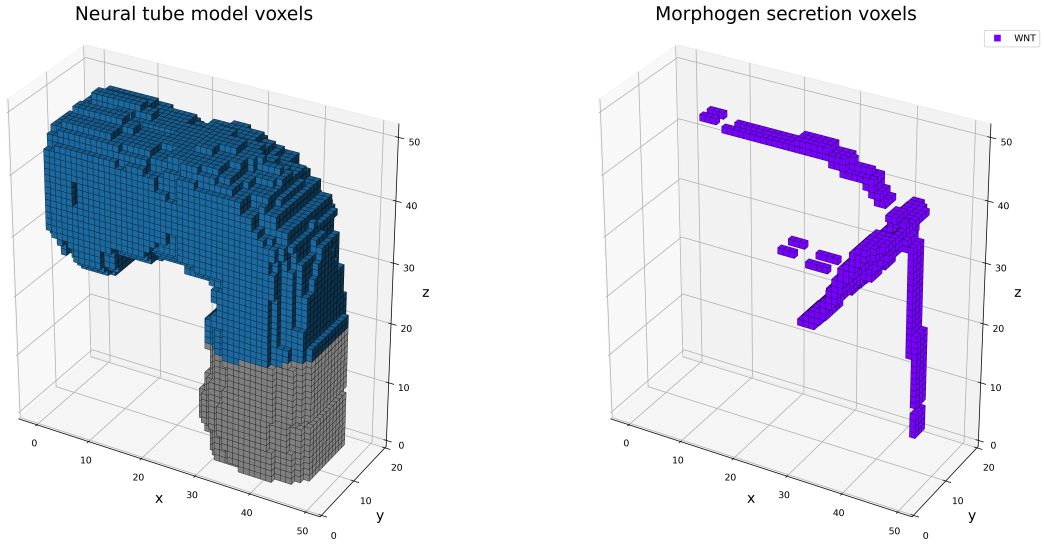
The cost function in eq. (3.2) consists of two types of terms. The first (left-most) terms are standard least-squares terms, used for minimizing differences between simulated levels $s_{gl}(\mathbf{p}, t)$ and data $d_{gl}(t)$ for each of the FB, MB, and HB genes g in the model. The sums over t and l run over all times t at which data were available, and over all experimental conditions, i.e. locations l (or, in other words, average CT levels at each location). The second (right-most) terms are steady-state enforcing terms, which were included to ensure that physically unrealistic simulation outcomes[2] were not considered optimal. The constant α is a hyperparameter which was manually tuned. A detailed description of the full optimization routine used to minimize $C(\mathbf{p})$ is provided in Appendix D.

3.2 Spatial simulations

Patterning of the tri-stable model as response to a WNT concentration gradient was simulated in a curved 3D voxel geometry, which represents the neural tube. The simulation geometry is shown in figure 3.1 (a) below. Figure 3.1(b) shows the sub-set of voxels in the simulation geometry which acted as WNT morphogen secretion sites in the simula-

[2] i.e. good fits to data, but blow-up of $s(\mathbf{p}, t)$) to very high levels after day 14.

tions. Note the similarity with figure 2.2 - the chosen secretion voxels represent the most significant *in vivo* WNT-secreting secondary organizer regions. The tube geometry and secretion voxel distribution were both obtained from Brambach et al. (2021) [11].



(a) (b)

Figure 3.1: Geometry of the 3D neural tube geometry used in patterning simulations. (a) All simulation voxels, with grey voxels representing spinal cord tissue - patterning was not simulated there. (b) Voxels which acted as WNT morphogen secretion sources.

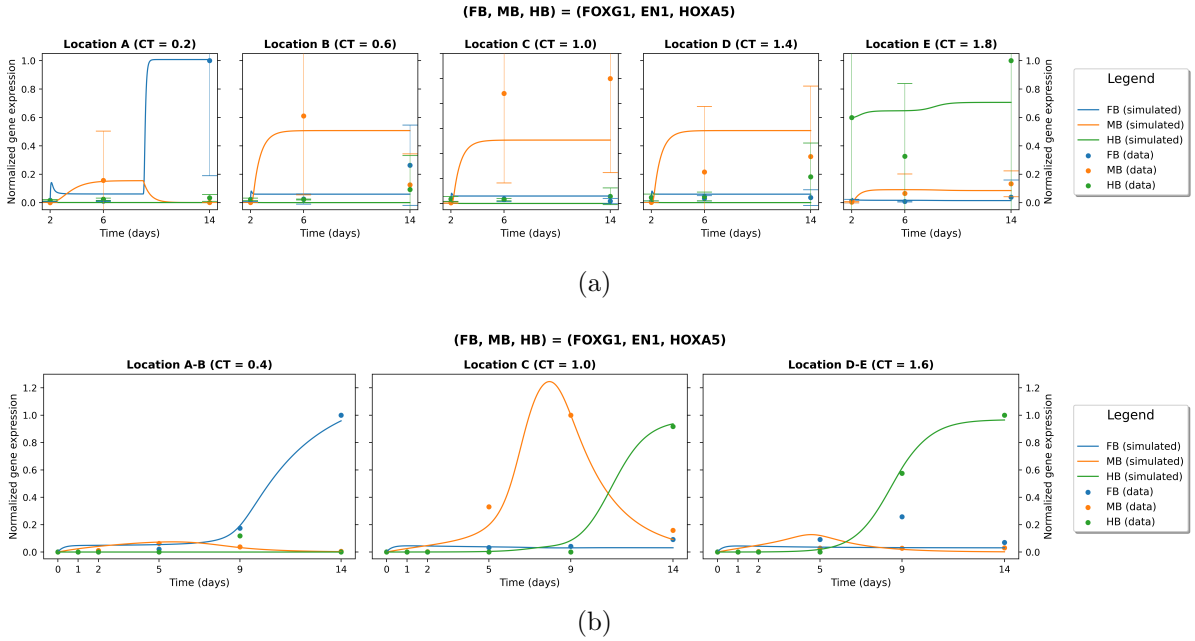
Patterning was formulated as a finite-difference reaction-diffusion problem on a lattice, see appendix E.1. The optimized tri-stable model equations (3.1) were solved in each lattice point, with the CT node level replaced by the local level of WNT. Hence, the WNT morphogen was assumed to affect the tri-stable network in exactly the same way as CT did during model optimization. A more detailed description of the 3D patterning simulation scheme is given in appendix E.3. Patterning was simulated with the tri-stable network optimized against both bulk and sc-RNAseq data. In the former case, data at day 2 were used as model initial conditions; in the latter, day 0 data were used as initialization.

3.3 Results and discussion

3.3.1 Model optimization

Good quality fits of the tri-stable model simulation outcomes were obtained for several different combinations of FB, MB, and HB genes, with the model optimized against both bulk and sc-RNAseq data. Figures 3.2 and 3.3 on the following page show fits to bulk (a) and sc-RNSAseq (b) data for (FB, MB, HB) = (FOXG1, EN1, HOXA5) and (OTX2, EN1, GBX2), respectively. The first gene combination is shown because it is a typical example of a combination exhibiting a clear spatial tri-stable pattern, similarly

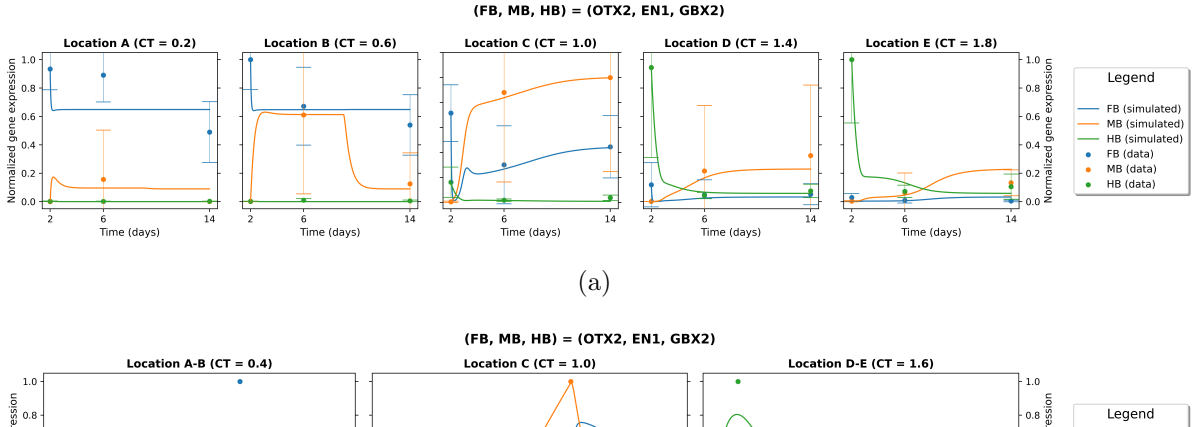
to (FOXG1, EN1, HOXA2) in Brambach et al. (2021) [11]. The other combination is shown because OTX2 and GBX2 are well-known early rostral / caudal markers.



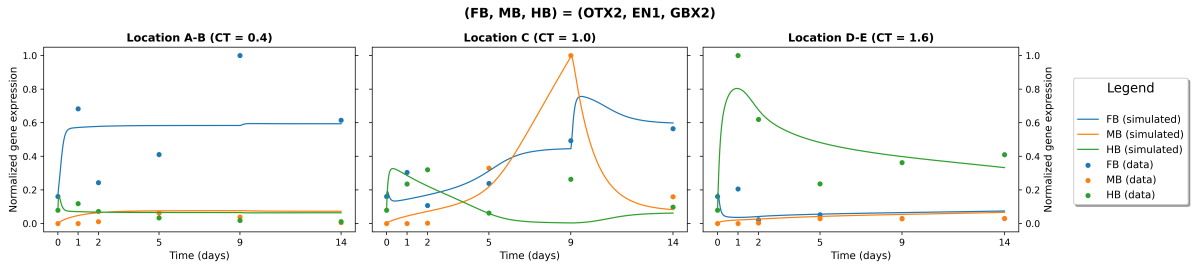
(a)

(b)

Figure 3.2: Simulation of the tri-stable network model dynamics when it is optimized against FOXG1, EN1, HOXA5, (a) bulk data and (b) sc-RNAseq data.



(a)



(b)

Figure 3.3: Simulation of the tri-stable network model dynamics when it is optimized against OTX2, EN1, GBX2, (a) bulk data and (b) sc-RNAseq data.

In figure 3.2(a), it is clear that the simulations agree well with the experimentally ob-

served tri-stable pattern at day 14. Importantly, this happens despite removal of the CT gradient at day 9, as in the MiSTR experiments. In location A, CT removal even seems to be what induces the correct final pattern. In the other locations, the pattern is barely affected by the day 9 CT transient. This kind of irreversibility, where a pattern is sustained even in absence of the signal which induced it, can be explained by the toggle switches in the tri-stable network topology (see appendix E.2, and also e.g. [15]). It is worth noting that the MB region is quite broad, as it spans three locations B-D.

For the corresponding simulations optimized against sc-RNAseq data in figure 3.2(b), the final pattern does not appear to be tri-stable, only bi-stable between FB and HB: MB starts losing its region C dominance around day 8. This does not seem to be caused by the CT removal at day 9, but rather by an asymmetry in the strength of the MB-HB-switch: MB is strongly repressed when HB increases, but HB is not when MB increases.

As a final note for (FOXG1, EN1, HOXA5), both simulation outcomes agree well with data, but they are also very different due to significant differences in the underlying data. The same can be said for the (OTX2, EN1, GBX2) simulations in figure 3.3 - both agree well with data, but they differ due to dissimilarities between bulk and sc-RNAseq data.

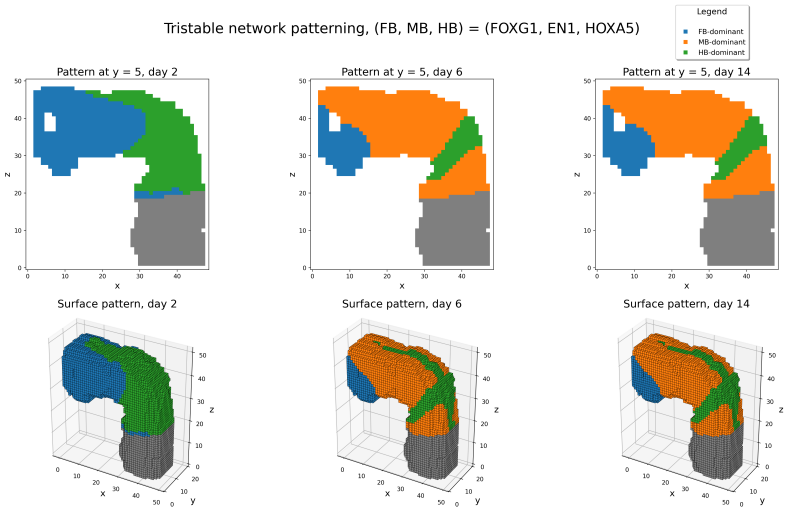
Since the agreement of simulations with data is good in figure 3.2 and 3.3, it seems like the tri-stable model indeed does describe the MiSTR data well, despite a large heterogeneity in the data for different FB/MB/HB genes and between bulk / sc-RNAseq data sets. This indicates a certain flexibility of the model - parameters of equations (3.1) can be tuned such that model simulation outcomes agree well with very different data sets (i.e. different choices of FB/MB/HB genes, or whether optimization is made against bulk or sc-RNAseq data). The model is therefore a correct but rather rough description of gene expression dynamics in rostral-caudal patterning, and does not e.g. reveal which genes are more or less important for the correct patterning to take place. Furthermore, some genes within a particular FB/MB/HB category are expressed very early during the patterning, and other genes within the same category are expressed very late. A further refined FB/MB/HB patterning model would take this into account, for example via a network consisting of different gene nodes being active for early and late patterning times.

3.3.2 Spatial simulations

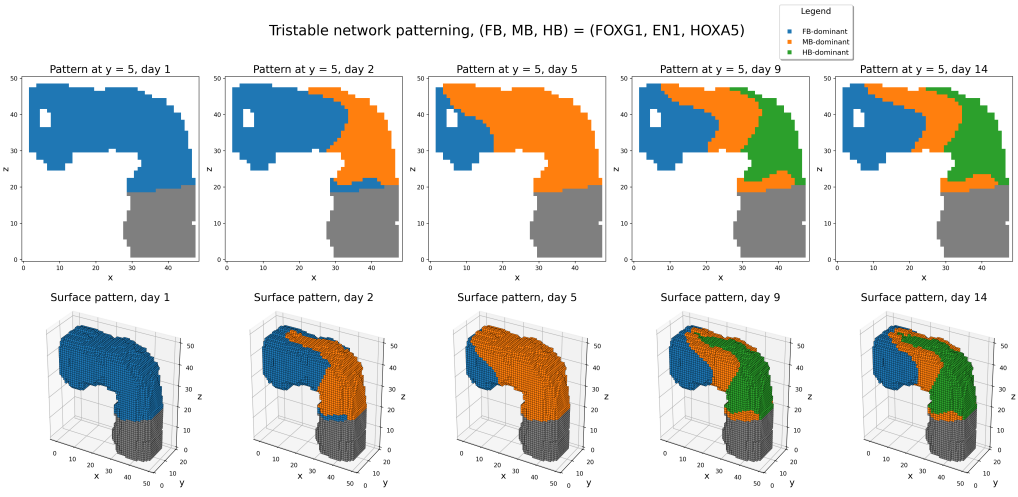
At first, simulations were made using the same WNT diffusion, production, and decay dynamics as in Brambach et al. (2021) [11]. This ensured that any differences in the simulated patterning would originate from differences in the model optimization, and not from differences in the morphogen implementation. Hence, the voxel spacing was set to 20 μm and the diffusion constant of WNT to $D_{\mathrm{WNT}} = 150.7\ \mu\mathrm{m}^2/\mathrm{s}$. Furthermore, WNT levels were assumed to have reached a steady-state level $[\mathrm{WNT}]^* = 2.0$ in morphogen secretion voxels, and to decay at a rate $\delta_{\mathrm{WNT}} = 0.01\ \mathrm{s}^{-1}$ elsewhere. WNT reaction terms R (see appendix E.1) were thus for a given voxel site s and time point n written as

$$R_s^{(n)} = \begin{cases} [\mathrm{WNT}]^* & \text{in secretion voxels,} \\ -\delta_{\mathrm{WNT}}[\mathrm{WNT}]_s^{(n)} & \text{outside secretion voxels.} \end{cases} \tag{3.3}$$

Figure 3.4 below shows the result of patterning simulations made with the tri-stable model optimized to FOXG1, EN1, and HOXA5 (as in figure 3.2), with morphogen / voxel conditions as described above. Each voxel is coloured according to the dominant (i.e. most expressed) gene at its location.



(a)



(b)

Figure 3.4: Tri-stable model patterning simulations optimized against (a) bulk and (b) sc-RNAseq data, with morphogen dynamics as in Brambach et al. (2021) [11]. The pattern order in the rostral-caudal direction at the final time point is the same for both simulations. Furthermore, they both exhibit the same kind of pattern reversal, where MB dominates not only rostrally but also caudally to the HB region.

Starting from the top left (most rostral) side of the tube in figure 3.4, the correct FB $\rightarrow$ MB $\rightarrow$ HB pattern shows up in the rostral-caudal direction in both (a) and (b). The relative sizes of patterning regions differ significantly between (a) and (b), though: in

particular, the MB region is very wide when simulating with the bulk data-optimized model, consistent with figure 3.2(a). Interestingly, the sc-RNAseq data-optimized model exhibits MB dominance at day 14 in a significant portion of the tube, despite this not being the case at day 14 in location C in figure 3.2(b).

In both figure 3.4(a) and (b) it is clear that the developing pattern exhibits an approximate reflection symmetry around the location where the isthmic organizer (IsO) secretion voxels are located: there, the pattern order reverses and varies as HB $\rightarrow$ MB when moving further in the caudal direction, instead of remaining HB. This can be easily understood from figure 3.1(b): the WNT morphogen secretion region distribution exhibits this kind of approximate symmetry. With isotropic WNT diffusion, it follows that the WNT gradient and thereby also the patterning process will exhibit a similar symmetry. This stems from that the WNT secretion distribution from Brambach et al. (2021) [11] portrays rostral-caudal pattering due to *secondary* WNT organizers, and not due to the mesodermal WNT gradient [13] that was emulated *in vitro* by Rifes et al. (2020) [12]. In essence, the unexpected pattern in figure 3.4 results from a biologically unrealistic WNT gradient shape - for the tri-stable model to pattern correctly, it is necessary to have a WNT gradient which increases *monotonously* in the rostral-caudal direction.

Regardless of the above, a few interesting insights may be gained by comparing the patterning results in figure 3.4 with those of Brambach et al. (2021) [11]. Particularly in figure 3.4(b), the pattern exhibits the same kind of dorsal-ventral *stacking* of the FB-MB-HB pattern. This can be attributed to the roofplate secretion regions. These regions help setting up a dorsal-ventral component of the WNT gradient, thereby producing some FB-MB-HB-patterning in this direction, referred to as stacking. This patterning characteristic can, however, be altered by modifying the WNT production dynamics. To demonstrate this, patterning was simulated using a modified form of the reaction terms in eq. (3.3):

$$R_l^{(n)} = \begin{cases} p_{\mathrm{WNT}} - \delta_{\mathrm{WNT}}[\mathrm{WNT}]_s^{(n)} & \text{in secretion voxels,} \\ -\delta_{\mathrm{WNT}}[\mathrm{WNT}]_s^{(n)} & \text{outside secretion voxels.} \end{cases} \tag{3.4}$$

The behaviour outside WNT secretion regions is the same as in eq. (3.3), but within secretion regions there is now a dynamic production of WNT at a constant rate p_{M} balanced by decay, instead of an artificially fixed constant level $[\mathrm{WNT}]^*$. Figure 3.5 on the following page shows simulated patterning using this modified form of the reaction terms, with the tri-stable model optimized against the same sc-RNAseq data as in figure 3.4(b). The values of $p_{\mathrm{WNT}} = 0.05\ \mathrm{s}^{-1}$ and $\delta_{\mathrm{WNT}} = 0.004\ \mathrm{s}^{-1}$ were chosen such that the three brain regions are similar in size to those in figure 3.4(b).

The overall FB-MB-HB pattern in figure 3.5 is very similar to that in figure 3.4(b), but there is much less stacking. The amount of stacking is minimized as a consequence of that the dorsal-ventral WNT gradient from the roof plate is less strong than before. This is because a dynamic production of WNT leads to the emergence of a bunching effect,

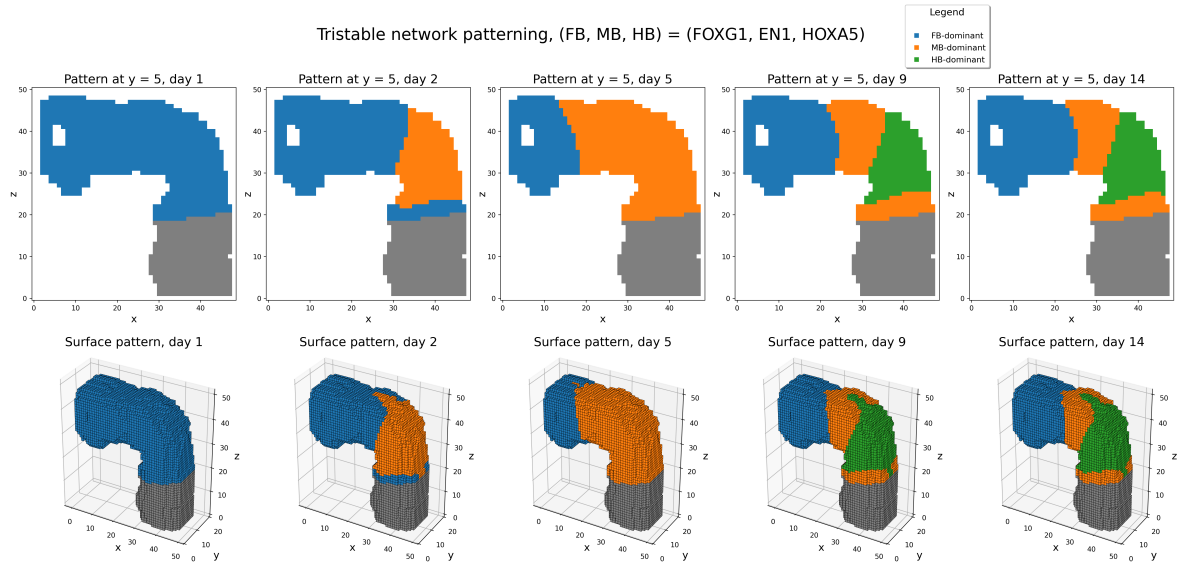


Figure 3.5: Tri-stable patterning simulations with the model optimized against sc-RNAseq data. WNT production is dynamic, with $p_{\mathrm{WNT}} = 0.05\ \mathrm{s}^{-1}$ and $\delta_{\mathrm{WNT}} = 0.004\ \mathrm{s}^{-1}$. This morphogen production / decay mechanism results in patterns which are nearly invariant in the dorsoventral direction (compare with figure 3.4(b)).

where the eventual steady-state WNT level will be higher in and around a large collection of secretion voxels, like the IsO, than in and around a smaller one, like the roof plate. Hence, differences in how morphogen production / decay is implemented can result in very different patterning results for the exact same gene regulatory model.

Differences in the WNT diffusion implementation can also affect the patterning. If the diffusion of WNT from a localized source can be characterized by a single diffusion constant D_{WNT}, the WNT gradient will decay exponentially in space [16]. However, in a recent paper on WNT diffusion in zebrafish [17], arguments are presented for that WNT actually diffuses via two different modes in the brain - one fast ($D_{\mathrm{WNT_{fast}}} \approx 60\ \mu\mathrm{m}^2/\mathrm{s}$) and one slow ($D_{\mathrm{WNT_{slow}}} \approx 6\ \mu\mathrm{m}^2/\mathrm{s}$).

If it is assumed that the WNT signalling which drives rostral-caudal patterning does not distinguish between WNT molecules carried via two distinct diffusion modes, it should then be the superposition $[\mathrm{WNT_{tot.}}] = [\mathrm{WNT_{fast}}] + [\mathrm{WNT_{slow}}]$ which acts as regulatory input to the patterning network. For the right production and decay rates, this can result in a steep gradient close to a WNT source (for example, the IsO), and a flatter gradient further away. This might influence the relative sizes of the FB / MB / HB regions, as is illustrated by a conceptual sketch in figure 3.6 on the following page.

Another patterning simulation with two diffusion modes, using the values of $D_{\mathrm{WNT_{fast}}}$ and $D_{\mathrm{WNT_{slow}}}$ from above, corroborates the picture provided by figure 3.6. The results of these simulations are shown in figure 3.7, on the following page.

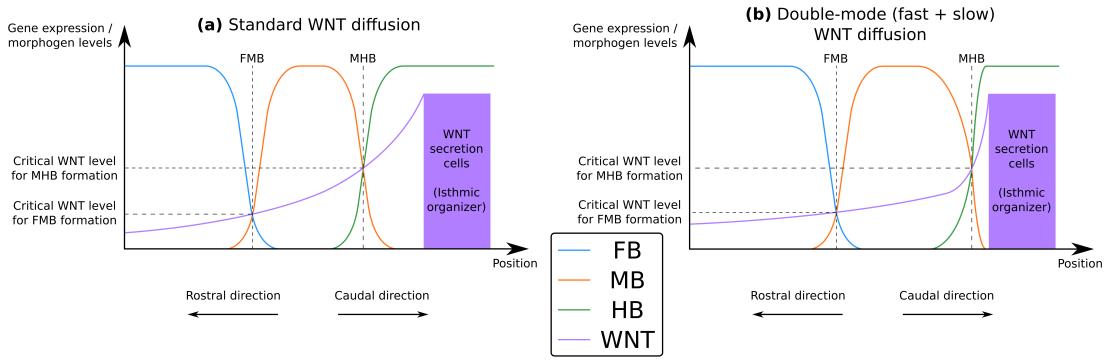


Figure 3.6: A combination of fast and slow WNT diffusion (b) can result in a WNT gradient shape which is very different from the standard gradient shape due to a single diffusion mode (a). Such differences in the WNT gradient can significantly influence the relative sizes of patterning regions.

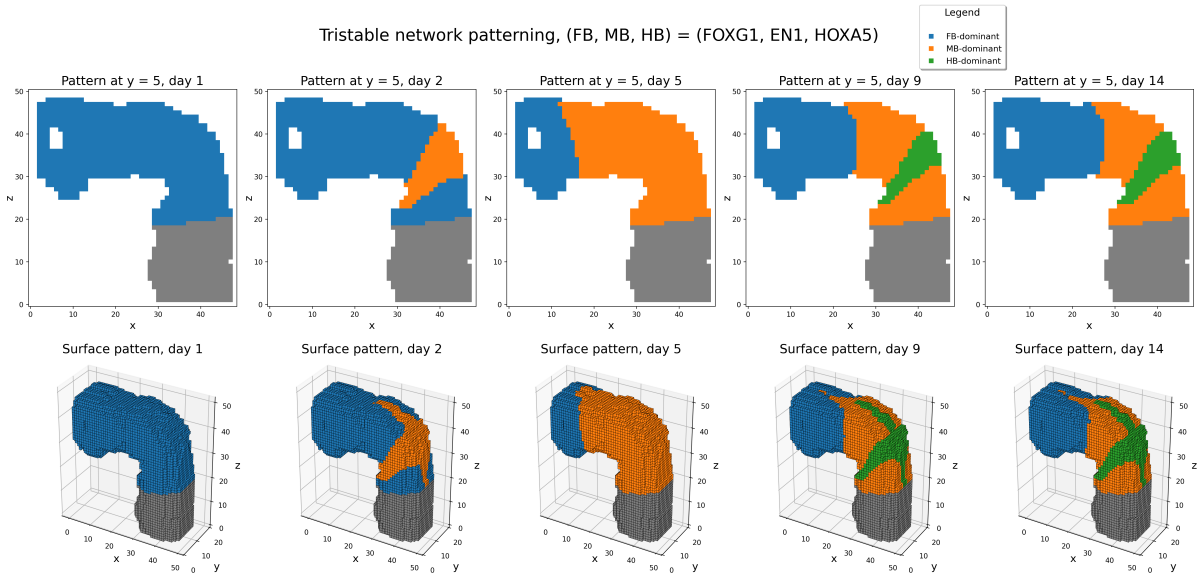


Figure 3.7: Tri-stable patterning simulations with the model optimized against sc-RNAseq data. WNT production is dynamic and happens via two separate diffusion modes, with $p_{\mathrm{WNT_{slow}}} = 0.0075\ \mathrm{s}^{-1}$, $p_{\mathrm{WNT_{fast}}} = 0.007\ \mathrm{s}^{-1}$, $\delta_{\mathrm{WNT_{slow}}} = 0.007\ \mathrm{s}^{-1}$, and $\delta_{\mathrm{WNT_{fast}}} = 0.0007\ \mathrm{s}^{-1}$. Note that the MB-HB boundary is placed close to the IsO, but the FB-MB boundary position remains similar to in figure 3.5.

To summarize, simulations with the tri-stable model optimized against time-resolved data from Rifes et al. (2020) [12] produces spatial patterns similar to those in Brambach et al. (2021) [11]. Furthermore, the simulated pattern is significantly affected by how WNT production, decay, and diffusion is implemented - patterning does not only depend on the patterning network model. Most importantly, rostral-caudal patterning into FB, MB, and HB regions requires a WNT gradient which increases monotonously in the rostral-caudal tube direction. Hence, in a more sophisticated spatial model, the Isthmic organizer is not the main *cause* of rostral-caudal patterning, but is instead *formed as a consequence of* rostral-caudal patterning. The second model in this thesis aims to capture how this occurs.

4 Model 2: isthmic organizer induction, maintenance, and patterning

4.1 Background, network model, and optimization

Instead of considering rostral-caudal patterning as the result of a tri-stable switch type network between coarsely defined FB, MB, and HB regions like in model 1, an alternative approach is to set up a patterning network model based on both established knowledge about how relevant neural tube genes interact, and about which genes are known to be highly expressed at both early and late stages of patterning. Using such an approach, a network model intended to capture the dynamics of isthmic organizer (IsO) formation and maintenance was constructed.

The main driver behind the construction of the IsO model described in this section is the observation that the neural tube does not seem to pattern directly into the three FB / MB / HB regions - instead, it is first patterned into only two regions: rostral and caudal, likely as a response to the previously mentioned gradient of WNT signalling from the paraxial mesoderm [13] (see figure 2.1). The early rostral and caudal regions are characterized by high expression levels of OTX2 and GBX2, respectively. These genes are known to directly antagonize each other in a bi-stable switch motif [18]. Importantly, the IsO is subsequently formed at the *intersection* of rostral and caudal regions, and midbrain (MB) region development is a result of subsequent IsO signalling activity, mediated by WNT1 and FGF8 proteins acting as morphogens [19]. Figure 4.1 on the following page illustrates this chain of events.

The constructed network model is shown in figure 4.2, on the following page. It is based on the picture of IsO development in figure 4.1, but also on previous modelling work on IsO patterning in Hock et al. (2012) [20]. Importantly, the CT network node represents the external WNT gradient in figure 4.1 - this is the paraxial mesoderm WNT gradient, which was emulated by the action of CT in the MiSTR experiments (see section 2.3).

The crucial network feature which determines the location of IsO formation is how the CT node affects the OTX2-GBX2 switch. This is because stronger network connections from CT to OTX2 and / or GBX2 will push the location where OTX2/GBX2 dominance flips (i.e. the dotted line in figure 4.1(a)) in the rostral direction, and vice versa. See appendix E.2 for a more detailed description of this effect. There is indeed experimental support for that mesodermal WNT signalling represses OTX2 and activates GBX2, which makes this positioning effect of the network model biologically grounded [21].

The two network genes WNT1 and FGF8, characteristic of the IsO, are repressed by GBX2 and OTX2, respectively. This helps restrict the spatial extent of the IsO. The midbrain node MB is induced by WNT and FGF morphogens which are secreted from the IsO. MB genes which are induced by WNT / FGF from the IsO are various EN and

PAX genes [20].

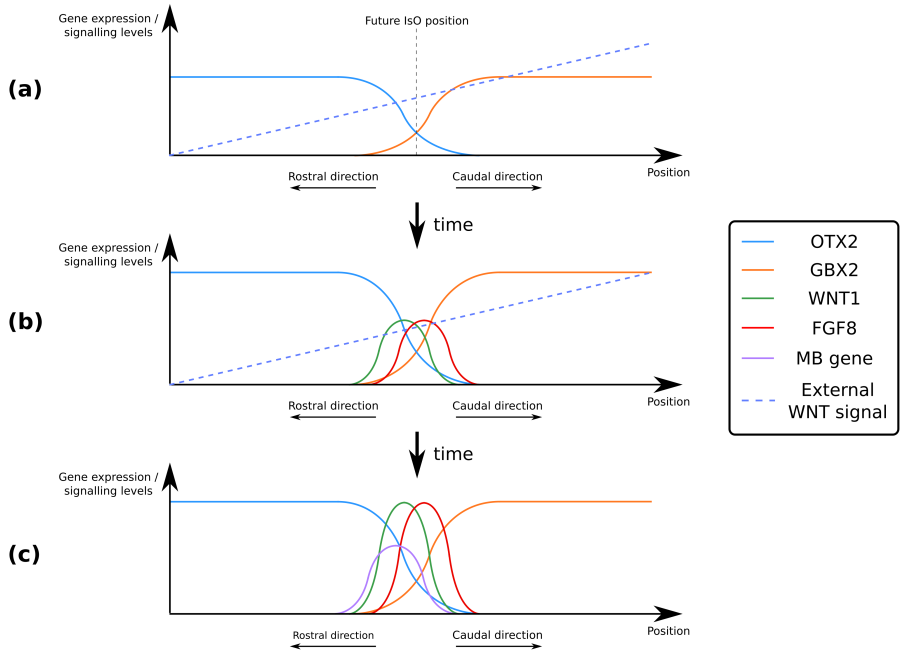


Figure 4.1: A simple sketch of IsO development over time. (a): the OTX2-GBX2 switch splits rostral and caudal regions as a response to an external WNT signalling gradient. (b) IsO genes, i.e. WNT1 and FGF8, start being expressed at the rostral-caudal intersection. (c): MB genes start being expressed at and around the IsO as a response to WNT and FGF signalling. The external WNT gradient may now no longer be needed to maintain the IsO - it is maintained via feedback from MB genes and from itself.

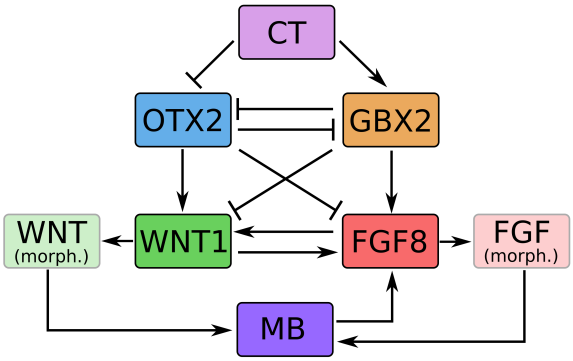


Figure 4.2: The constructed IsO development network model. CT emulates the initial WNT signalling gradient from the paraxial mesoderm. The node text (morph.) indicates morphogen or morphogen-like action of the WNT1 and FGF8 proteins.

With the network fully defined, its Shea-Ackers rate equations could be written down (see appendix B.2). However, a simplification of the network in figure 4.2 was first made in order to make model optimization against data tractable: the morphogen actions of WNT1 and FGF8 on MB were approximated by the direct actions WNT1 $\rightarrow$ MB and FGF8 $\rightarrow$ MB. With this, and the OTX2, GBX2, WNT1, and FGF8 nodes abbreviated

by their first letters, the network Shea-Ackers equations that were used during model optimization are:

$$\begin{aligned}
\frac{d[\mathrm{O}]}{dt} &= \frac{b_1}{1 + c_1[\mathrm{CT}]^{n_1} + c_2[\mathrm{G}]^{n_2}} - \delta_1[\mathrm{O}] \\
\frac{d[\mathrm{G}]}{dt} &= \frac{b_2 + c_3[\mathrm{CT}]^{n_3}}{1 + c_3[\mathrm{CT}]^{n_3} + c_4[\mathrm{O}]^{n_4}} - \delta_2[\mathrm{G}] \\
\frac{d[\mathrm{W}]}{dt} &= \frac{b_3 + c_5[\mathrm{O}]^{n_5} + c_6[\mathrm{F}]^{n_6}}{1 + c_5[\mathrm{O}]^{n_5} + c_6[\mathrm{F}]^{n_6} + c_7[\mathrm{G}]^{n_7}} - \delta_3[\mathrm{W}] \\
\frac{d[\mathrm{F}]}{dt} &= \frac{b_4 + c_8[\mathrm{G}]^{n_8} + c_9[\mathrm{W}]^{n_9} + c_{10}[\mathrm{MB}]^{n_{10}}}{1 + c_8[\mathrm{G}]^{n_8} + c_9[\mathrm{W}]^{n_9} + c_{10}[\mathrm{MB}]^{n_{10}} + c_{11}[\mathrm{O}]^{n_{11}}} - \delta_4[\mathrm{F}] \\
\frac{d[\mathrm{MB}]}{dt} &= \frac{b_5 + c_{12}[\mathrm{W}]^{n_{12}} + c_{13}[\mathrm{F}]^{n_{13}}}{1 + c_{12}[\mathrm{W}]^{n_{12}} + c_{13}[\mathrm{F}]^{n_{13}}} - \delta_5[\mathrm{MB}] \\
\frac{d[\mathrm{CT}]}{dt} &= 0.
\end{aligned} \tag{4.1}$$

Using eq. (4.1), optimization of the IsO model against MiSTR data was performed identically to how it was performed for model 1 (see section 3.1). EN and PAX genes were used as MB node during optimization.

4.2 Spatial simulations

Patterning of the optimized IsO model was simulated like for model 1 (see section 3.2). However, during simulations in the curved tube geometry from figure 3.1(a), the WNT source voxels from figure 3.1(b) were not used - one purpose of the IsO model is to predict where the IsO will form, so no *ad hoc* assumptions about its location are made. An artificial steady-state CT gradient was put in place along the rostral-caudal tube axis during the first 9 patterning days (and was removed after day 9) to replicate the MiSTR experiment conditions as closely as possible in a curved geometry.

In addition to simulations in the curved tube, patterning was also simulated in a simple rectangular 3D geometry of approximately the same dimensions as the MiSTR tissue. The simpler geometry was deemed useful for investigating the behaviour of a novel network, since it eliminates the potentially confounding effects of the curved neural tube geometry. The CT gradient was identical to the one used during the MiSTR experiments. The simple simulation geometry is shown in figure 4.3 on the following page, together with the steady-state CT gradient.

For simulations in the simple geometry, direct actions WNT1 $\rightarrow$ MB and FGF8 $\rightarrow$ MB were assumed, like during model optimization. Simulations in the curved tube geometry were made using both direct action, and with the full WNT1 and FGF8 morphogen effects from figure 4.2 re-introduced into the model. The purpose of re-introducing the morphogen effects was to investigate how this affected the spatial extent of the MB region.

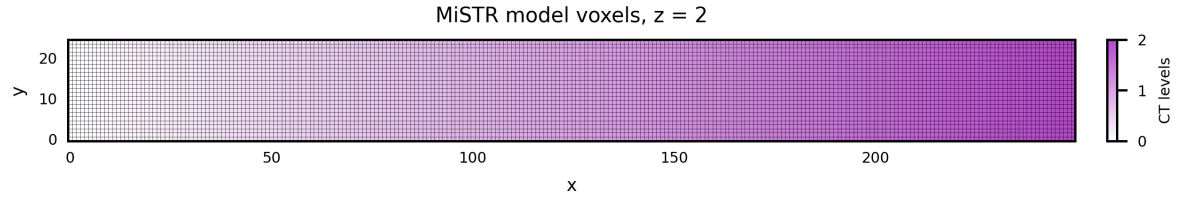


Figure 4.3: Geometry of a 2D slice of the 3D MiSTR voxel model, with the CT gradient along the x axis. The extent of the model along the z axis is 3 voxels.

Morphogens were assumed to diffuse with $D_{\mathrm{WNT/FGF}} = 150\ \mu\mathrm{m}^2/\mathrm{s}$. Their production and decay were assumed to follow equation (3.4), with the modification that production rates were not constant, but instead proportional to WNT1 / FGF8 gene expression levels. This ensured that a net production of WNT and FGF morphogens occurred only if their respective genes were highly expressed, so that only the IsO and no other parts of the tube produced morphogen signals.

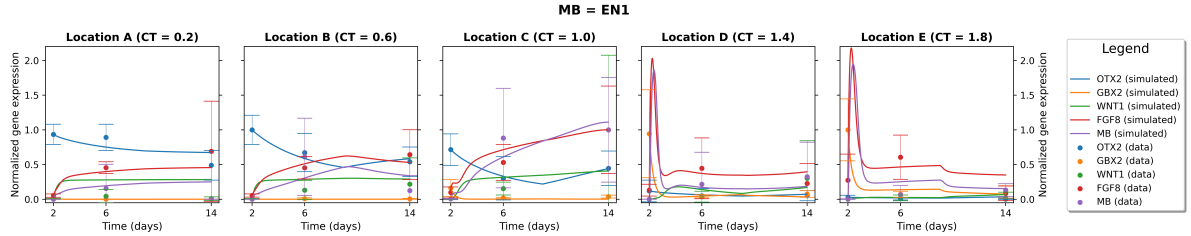
4.3 Results and discussion

4.3.1 Model optimization

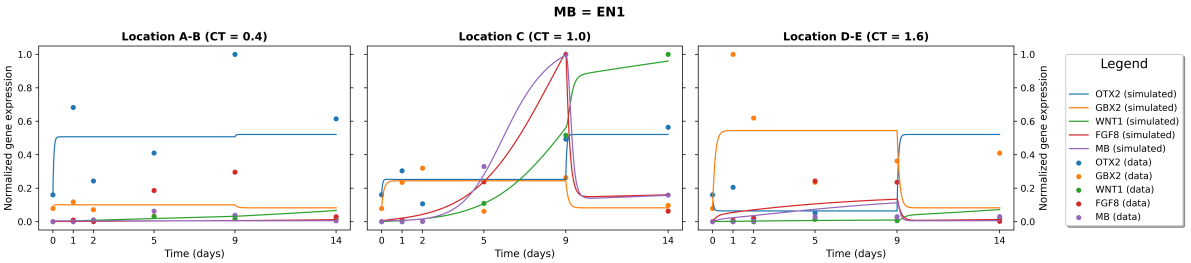
Figures 4.4 and 4.5 on the following page show simulations with the model optimized to MB = EN1 and MB = PAX5, respectively. When optimized to both bulk and sc-RNAseq data, IsO model simulation outcomes were in good agreement with data. In particular, good agreement was achieved in location C, where IsO and MB genes are highly expressed in the MiSTR data - this arguably means that location C is the most important location for the model to get right. It is worth noting that IsO / MB genes are always highly expressed at day 9, and never die down to very low levels at day 14. This suggests that the model IsO is sustained even in absence of the CT signal which helped induce it. The model also seems to capture that OTX2 and GBX2 are highly expressed already during the first few days of patterning, and that OTX2 (GBX2) is predominantly expressed in the most rostral (caudal) locations. It should be said that the OTX2-GBX2 switch flipped due to the removal of CT at day 9 for some of the simulations, though. However, this does not seem to have a very strong effect on the IsO and MB genes.

4.3.2 Spatial simulations

Figure 4.6 on page 23 shows the simulated patterning process of the IsO model optimized against MB = EN1 bulk data, and against MB = PAX5 sc-RNAseq data in the MiSTR tissue geometry (figure 4.3). In both patterning simulations, sharp domains where WNT1, FGF8, and MB are simultaneously co-expressed arise over time, despite the very different initializations of the two simulations. These domains arise at the intersection of regions with strong OTX2 and GBX2 expression, correctly positioned halfway along the CT gradient in the x-direction. This is in agreement with the picture of IsO induction that

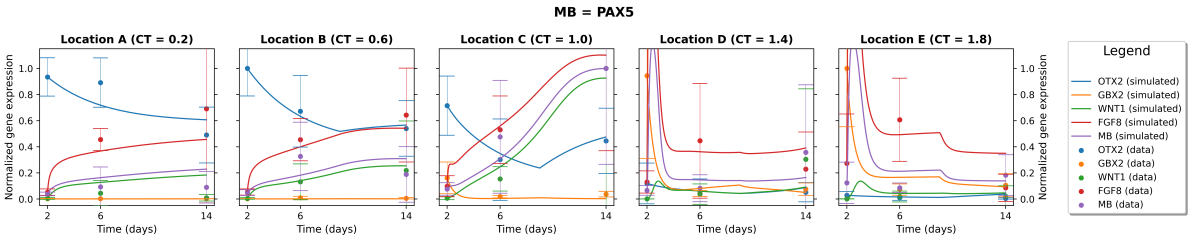


(a)

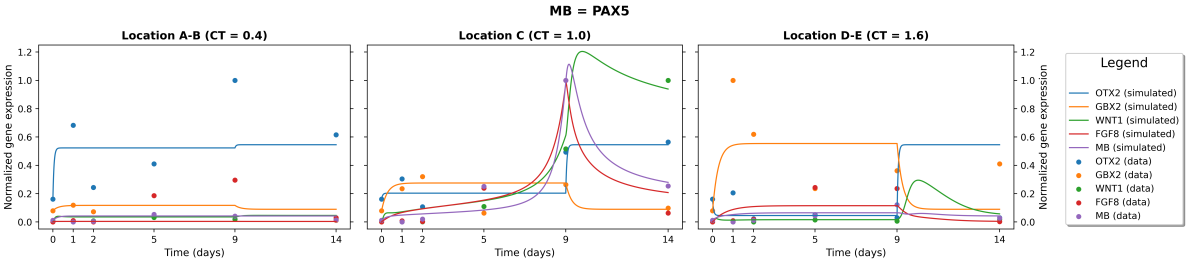


(b)

Figure 4.4: IsO network optimized to MB = EN1, (a) bulk data and (b) sc-RNAseq data.



(a)

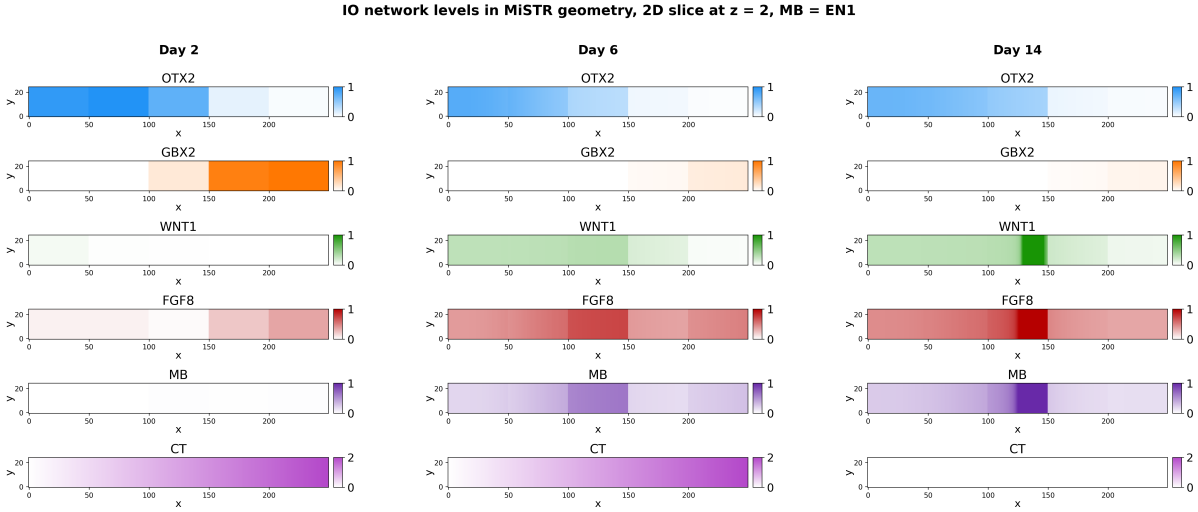


(b)

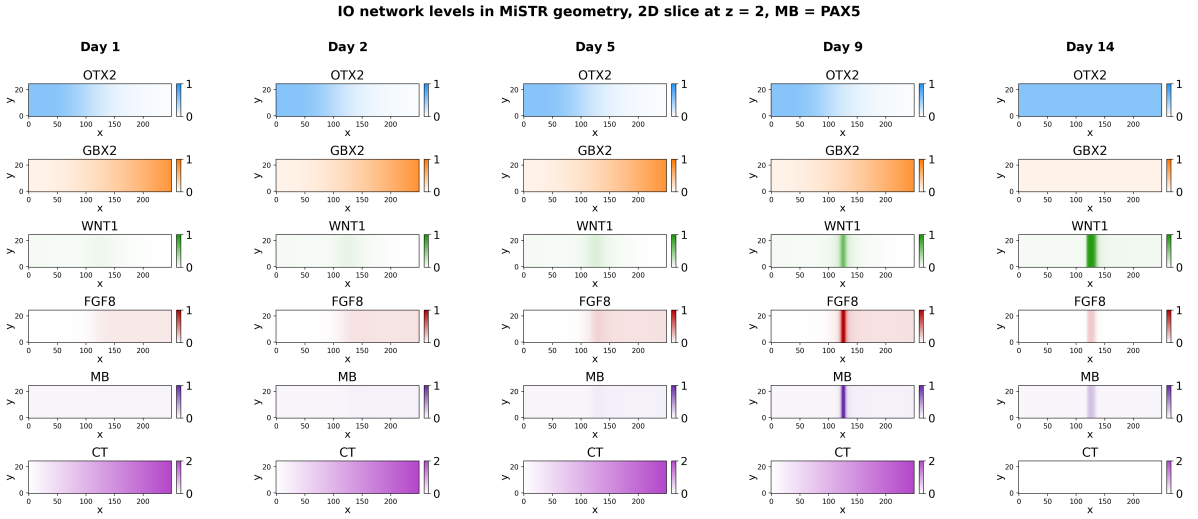
Figure 4.5: IsO network optimized to MB = PAX5, (a) bulk data and (b) sc-RNAseq data.

was shown in figure 4.1. The IsO domain is wider for the bulk data simulation, which seems to be a consequence of the piecewise uniform initialization at day 2.

Importantly, the IsO and MB domains survive even in the absence of the CT gradient, which is removed at day 9. This indicates that the IsO-MB genes maintain themselves



(a)



(b)

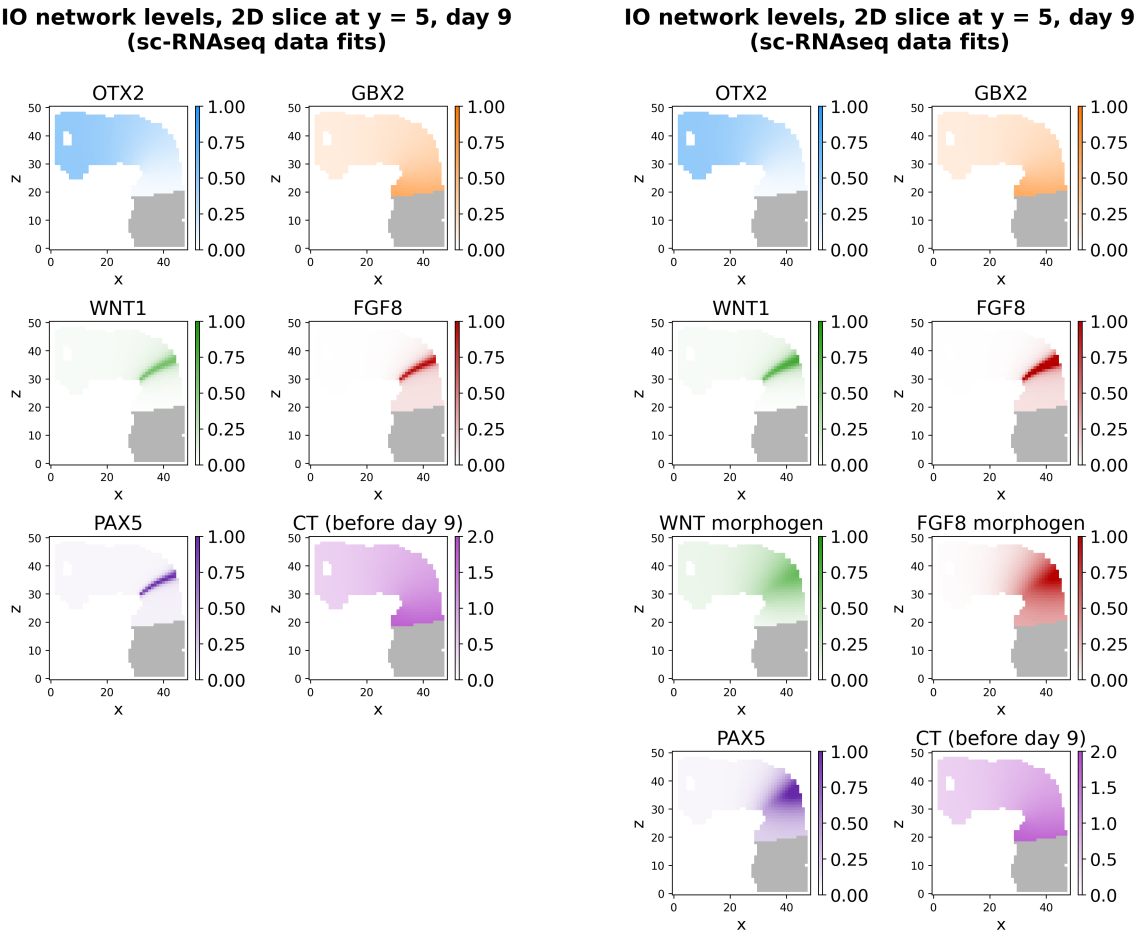
Figure 4.6: IsO model patterning in the MiSTR tissue geometry. The model is optimized to (a) bulk data and (b) sc-RNAseq data. The IsO domains where WNT1, FGF8, and MB are all highly expressed, i.e. at the tissue centre in both (a) and (b), remain after CT is removed at day 9.

once they are established, which can be considered a central feature of an organizer region in development. Additionally, the IO domain is still clearly present at day 14 in the sc-RNAseq fits, despite the flip of the OTX2-GBX2 switch in the caudal region of the tissue at day 9. More specifically, the sharp WNT1, FGF8, and MB peaks remain in the tissue centre, despite that the OTX2-GBX2 switch flip at day 9 eventually leads to that both OTX2 and GBX2 have completely homogeneous expression profiles across the tissue. Hence, the spatial OTX2 and GBX2 expression profiles are important for initially positioning the IsO and MB domains, but changes to those profiles do not seem to strongly affect the extent of the IsO and MB domains once these domains are established.

The key mechanism which facilitates both accurate initial positioning as well as maintenance of the IsO despite the CT loss at day 9 is the mutual WNT1-FGF8 activation, and possibly also the FGF8 self-activation via the loop FGF8 → MB → FGF8. To understand why the positive feedback between WNT1 and FGF8 is important, consider the model network in figure 4.2, but *without* such feedback. Given the expected early expression profiles of OTX2 and GBX2 as response to the CT gradient (compare with appendix E.2), it then follows that WNT1 would be most strongly expressed at the rostral end of the tissue, and FGF8 at the caudal end. Thus, no localized IsO domain would be formed at the centre. If, on the other hand, we have an already correctly positioned IsO, the lack of WNT1-FGF8 feedback would lead to a widening of the IsO domain when CT is removed (as long as the CT removal noticeably affects the OTX2 and GBX2 profiles, like in figure 4.4(b)). As an example, the WNT1 profile would widen caudally if GBX2 levels dropped globally at day 9, like they do in figure 4.4(b). *With* the WNT1-FGF8 feedback, this does not happen - high expression of WNT1 in a particular position is dependent on high expression of FGF8 in that position, and vice versa, and thereby it follows that the feedback prevents IsO widening. The combined WNT1-FGF8-MB sub-network thus behaves like a single self-activating network node which is repressed by both OTX2 and GBX2, reminiscent of the MB node in the tri-stable model.

It should be noted that the domains of high MB expression in figure 4.6 mostly overlap in space with the domains of high IsO expression. However, the MB domain should extend rostrally from the IsO, since the IsO acts as the MB-HB boundary *in vivo*. The MB region being too narrow likely stems from that WNT and FGF morphogen actions were replaced with direct actions WNT1 → MB and FGF8 → MB in the optimized model.

To investigate whether morphogen-mediated MB induction can broaden the MB region, morphogen action was re-introduced into the model as was described in section 4.2. Figure 4.7 on the following page shows the simulated IsO model pattern during day 9 in the 3D tube geometry, with direct WNT1 / FGF8 action in (a) and with morphogen WNT1 / FGF8 action in (b). In both (a) and (b), the IsO appears in the correct position (compare with figure 2.2). From (b) it is clear that WNT1 and FGF8 morphogen diffusion indeed can result in the widening of the MB (i.e. PAX5) domain, without simultaneously resulting in a significant widening of the IsO domain beyond its extent in (a). This is because PAX5 expression in (b) follows the morphogen distributions rather than the morphogen *gene* distributions as in (a): the former are wider than the latter. The widening does not happen only rostrally of the IsO, though, which would be a result even more in accordance with biological reality.



(a) (b)

Figure 4.7: IsO network patterning in the 3D tube model, as response to a CT gradient which increases monotonously in the caudal tube direction. The IsO is located where the tube has maximum curvature, i.e. where both WNT1, FGF8, and PAX5 are highly expressed. (a) direct actions by WNT1 and FGF8 induce PAX5. (b) Diffusing WNT1 and FGF8 morphogens induce PAX5. Here, $p_{\mathrm{WNT1}} = p_{\mathrm{FGF8}} = 0.0225$ s^{-1}, and $\delta_{\mathrm{WNT1}} = \delta_{\mathrm{FGF8}} = 0.01$ s^{-1}. Note the wider PAX5 region in (b) compared to in (a).

To summarize, the proposed IsO network model exhibits several important characteristics of a useful model of IsO establishment and maintenance: simulations are in reasonably good agreement with data, the IsO emerges at the right position, and the IsO maintains itself in the absence of external CT (i.e. WNT) signalling. The location of the MB domain is roughly correct, and diffusible WNT1 and FGF8 morphogens can act as a MB widening mechanism while simultaneously not widening the IsO itself. Further investigations will reveal whether the IsO model is indeed capable of restricting the MB region to lie only rostral to the IsO, as it should, or whether (and if so, how) the model needs to be modifed to achieve this.

5 Summary and outlook

Two different regulatory network models of early rostral-caudal neural tube patterning have been investigated. They were first optimized against the time-resolved bulk and sc-RNAseq gene expression data from the MiSTR experiment in Rifes et al. (2020) [12], and then used in spatial 3D simulations of WNT morphogen driven patterning. The first model was the tri-stable network from Brambach et al. (2021) [11] which describes patterning into FB, MB, and HB regions. The second model captures how the isthmic organizer (IsO) forms in a specific position along the rostral-caudal axis of the neural tube, and how it patterns the adjacent MB region via WNT and FGF morphogens.

For the tri-stable FB-MB-HB model, simulations of gene expression dynamics were in good agreement with data when the model was optimized against bulk and sc-RNAseq data for several different combinations of FB, MB, and HB genes. However, the expression dynamics of the optimized model was strongly dependent on which FB/MB/HB genes it was optimized against. This points towards a significant data heterogeneity between different genes within a given FB, MB, or HB category: they share the same spatial expression pattern, but the timing and dynamics of this pattern differs. This heterogeneity is clearly showcased by figure 2.5, and by the clustering results in appendix F. Hence, the tri-stable model can be said to accurately capture FB-MB-HB patterning on a coarse-grained level which leaves the internal regulatory dynamics of the FB, MB, and HB regions out of the picture.

Spatial patterning simulations of the optimized tri-stable model in the 3D tube geometry (figure 3.1) reproduced key results from similar simulations with the same regulatory model in Brambach et al. (2021), even though the model is now optimized against a different data set. It was found that dorsoventral stacking of the rostral-caudal pattern could be nearly eliminated by using a different implementation of morphogen production. In addition to this, a change in the WNT diffusion implementation led to a change in the resulting WNT gradient, such that the MB-HB boundary position was affected but the FB-MB boundary position was not. These results demonstrate that morphogen-driven patterning does not only depend on the patterning regulatory network as such, but also on the mechanisms behind how the morphogen gradient is established. Finally, the 3D simulations revealed a reflection symmetry of the pattern around the IsO. This symmetry arose from an arrangement of WNT secretion voxels which resulted in a WNT signalling gradient that emanated primarily from the tube centre. Patterning simulations using an alternative WNT gradient shape, more congruent with the early mesodermal WNT gradient, would likely remove the reflection symmetry from the pattern. Such simulations were not pursued in this work - instead, the reflection symmetry fuelled the thinking which led to the development of the second model describing IsO formation.

Simulations with the optimized IsO model were generally in good agreement with both bulk and sc-RNAseq data. Spatial simulations resulted in the formation of a correctly

positioned IsO-like domain in the MiSTR tissue geometry. Moreover, this domain sustained itself via positive feedback from itself and from the MB gene, even after loss of the external CT gradient. This can be considered an essential feature of a successful model of an organizer region - the organizer should remain after removal of the signal which helped to induce it. Finally, simulations in the 3D tube model revealed that WNT1 and FGF8 morphogen signalling from the IsO can act to broaden the MB region. As the MB broadening should be confined to the rostral side of the IsO, an interesting direction for future work would be to pursue a better understanding of the MB broadening mechanism.

There are many possible approaches for extending the existing network models. For example, the IsO network does not yet describe the patterning of the FB and HB regions. Perhaps OTX2 / GBX2 activate other FB / HB genes, who in turn regulate additional FB / HB genes that are expressed at even later stages of patterning? A more comprehensive literature study of the relevant genes could reveal this kind of interesting regulatory relationships to include in an extended IsO network model. For example, a bi-stable switch between SIX3 and IRX3 seem to be responsible for splitting the FB region into its two major later subdivisions: the diencephalon and the telencephalon [22]. In a similar way, it might be possible to extend the tri-stable model by e.g. introducing sub-nodes within the main FB, MB, and HB nodes such that a distinction is made between early and late patterning genes. Another approach for extending the existing models would be to figure out their most significant regulatory connections - a good starting point for this approach would be to conduct parameter sensitivity analyses.

The spatial aspects of the models are also worth investigating in more detail. For instance, a better understanding of the *in vivo* transport, production, and decay mechanisms of WNT and FGF morphogens would reduce the number of possibly unfounded assumptions about their signalling. It would also be interesting to explore the dorsoventral direction further - the focus of this thesis was constrained to rostral-caudal patterning.

Finally, it should be said that the 3D model assumes a fixed size and shape of the neural tube. This is not the case *in vivo*: the tube grows and undergoes substantial conformational changes during the first few weeks of development. Hence, its patterning must be robust to both scaling and alterations of the tube shape over time. To test this, one could imagine e.g. using an atlas of human development [23] to fix the tube shape at a few distinct times, and interpolate its shape in between those times. This would allow one to formulate the tube patterning problem in a dynamically morphing geometry, and solve it using more sophisticated finite element-based methods. That kind of model would also be useful for investigating how patterning is influenced by e.g. mechanical forces between and within cells [24]. It could be relevant to consider such a multi-level model in future work: even though the kind of gene regulation and morphogen based models explored in this thesis shed much light on neural tube patterning mechanisms, the inclusion of additional model components may lead to even more interesting insights.

Acknowledgements

First and foremost, I would like to thank Victor Olariu for providing the opportunity for me to work on the neural patterning project. I have greatly appreciated his approach of granting me a high degree of freedom to steer the project in whichever direction I found to be the most interesting and worthwhile. Simultaneously, he has always been there to provide on-point feedback on my approaches, and I thank him for bearing with my occasionally relentless barrages of lengthy Slack messages. It was also very nice to continue developing the 3D simulations as a summer intern: these months laid the groundwork for many of the later results of the thesis.

I also want to thank Agnete Kirkeby and her group for a very interesting discussion of the modelling work in early November: this meeting led to the inclusion of an important detail without which the results of the thesis would have been less interesting for sure. In addition, I thank Pedro Rifes and Gaurav Singh Rathore for providing access to the bulk and sc-RNAseq data, respectively.

Much of my work on the thesis was done from home, but I did manage to get a seat at the department in the end. Thank you Marc Pielies Avellí and Yukun Cheng for being very pleasant office mates, and I thank both you two, Valentin Nguyen-Flor, and Emil Andersson for interesting discussions about the project and related topics.

Last, but not least, thank you Natalie for all your love and support during this year (and for being a medical student, now medical intern, who often gives valuable input about my at times all too physics-y and simplified ideas about biological processes).

References

[1] K. Stapornwongkul and J-P. Vincent. Generation of extracellular morphogen gradients: the case for diffusion. *Nature Reviews Genetics*, 22:1–19, 03 2021. doi:10.1038/s41576-021-00342-y.

[2] F. Lopes et al. Spatial bistability generates hunchback expression sharpness in the drosophila embryo. *PLoS computational biology*, 4:e1000184, 10 2008. doi:10.1371/journal.pcbi.1000184.

[3] F. Ulloa and E. Martí. Wnt won the war: Antagonistic role of wnt over shh controls dorso-ventral patterning of the vertebrate neural tube. *Developmental dynamics : an official publication of the American Association of Anatomists*, 239:69–76, 11 2009. doi:10.1002/dvdy.22058.

[4] D.H. Sanes, T.A. Reh, and W.A. Harris. *Development of the Nervous System,* 4th edition. Academic Press, 2019.

[5] T. Kim, S.Y. Koo, and L. Studer. Pluripotent stem cell therapies for parkinson disease: Present challenges and future opportunities. *Frontiers in Cell and Developmental Biology*, 8:729, 08 2020. doi:10.3389/fcell.2020.00729.

[6] R. Gonzalez et al. Deriving dopaminergic neurons for clinical use. a practical approach. *Scientific reports*, 3:1463, 03 2013. doi:10.1038/srep01463.

[7] S. Mahajani et al. Homogenous generation of dopaminergic neurons from multiple hipsc lines by transient expression of transcription factors. *Cell Death & Disease*, 10, 11 2019. doi:10.1038/s41419-019-2133-9.

[8] J. Briscoe. Morphogens, modeling and patterning the neural tube: An interview with james briscoe. *BMC biology*, 13:5, 12 2015. doi:10.1186/s12915-014-0105-1.

[9] A. Kirkeby et al. Generation of regionally specified neural progenitors and functional neurons from human embryonic stem cells under defined conditions. *Cell reports*, 1:703–14, 06 2012. doi:10.1016/j.celrep.2012.04.009.

[10] M. Brambach. From naive pluripotent stem cells to neuron progenitors: A study of the central gene regulatory networks in human, 2018. Student Paper.

[11] M. Brambach et al. Neural tube patterning: From a minimal model for rostrocaudal patterning toward an integrated 3d model. *iScience*, 24:102559, 05 2021. doi:10.1016/j.isci.2021.102559.

[12] P. Rifes et al. Modeling neural tube development by differentiation of human embryonic stem cells in a microfluidic wnt gradient. *Nature Biotechnology*, 38:1–9, 11 2020. doi:10.1038/s41587-020-0525-0.

[13] U. Nordström et al. Progressive induction of caudal neural character by graded wnt signaling. *Nature Neuroscience*, 5:525–532, 01 2002. doi:10.1038/nn0602-854.

[14] M. Gouti, V. Metzis, and J. Briscoe. The route to spinal cord cell types: A tale of signals and switches. *Trends in genetics : TIG*, 31, 03 2015. doi:10.1016/j.tig.2015.03.001.

[15] J. Ferrell and W. Xiong. Bistability in cell signaling: How to make continuous processes discontinuous, and reversible processes irreversible. *Chaos (Woodbury, N.Y.)*, 11:227–236, 04 2001. doi:10.1063/1.1349894.

[16] U. Alon. *An Introduction to Systems Biology: Design Principles of Biological Circuits,* 2nd edition. Chapman & Hall/CRC Mathematical and Computational Biology. Taylor & Francis, 2020.

[17] S. Veerapathiran et al. Wnt3 distribution in the zebrafish brain is determined by expression, diffusion and multiple molecular interactions. *eLife*, 9, 11 2020. doi:10.7554/eLife.59489.

[18] F. Inoue et al. Gbx2 directly restricts otx2 expression to forebrain and midbrain, competing with class iii pou factors. *Molecular and cellular biology*, 32:2618–27, 05 2012. doi:10.1128/MCB.00083-12.

[19] C. Chi et al. The isthmic organizer signal fgf8 is required for cell survival in the prospective midbrain and cerebellum. *Development (Cambridge, England)*, 130:2633–44, 07 2003. doi:10.1242/dev.00487.

[20] S. Hock et al. Sharpening of expression domains induced by transcription and microrna regulation within a spatio-temporal model of mid-hindbrain boundary formation. *BMC systems biology*, 7:48, 06 2013. doi:10.1186/1752-0509-7-48.

[21] Y. Elkouby et al. Mesodermal wnt signaling organizes the neural plate via meis3. *Development (Cambridge, England)*, 137:1531–41, 03 2010. doi:10.1242/dev.044750.

[22] D. Brafman and K. Willert. Wnt/ beta-catenin signaling during early vertebrate neural development: Wnt signaling in neural development. *Developmental Neurobiology*, 77, 08 2017. doi:10.1002/dneu.22517.

[23] B. de Bakker et al. An interactive three-dimensional digital atlas and quantitative database of human development. *Science*, 354:aag0053–aag0053, 11 2016. doi:10.1126/science.aag0053.

[24] F. Abdel et al. Actuation enhances patterning in human neural tube organoids. *Nature Communications*, 12, 05 2021. doi:10.1038/s41467-021-22952-0.

[25] R. Phillips, J. Kondev, J. Theriot, and N. Orme. *Physical Biology of the Cell,* 2nd edition. Garland Science, 2013.

[26] B. Emerson. Specificity of gene regulation. *Cell*, 109:267–70, 06 2002. doi:10.1016/S0092-8674(02)00740-7.

[27] M. Uhlen et al. Proteomics. tissue-based map of the human proteome. *Science (New York, N.Y.)*, 347:1260419, 01 2015. doi:10.1126/science.1260419.

[28] K. Takahashi and S. Yamanaka. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, 126:663–76, 09 2006. doi:10.1016/j.cell.2006.07.024.

[29] C. Kaebisch et al. The role of purinergic receptors in stem cell differentiation. *Computational and Structural Biotechnology Journal*, 13, 11 2014. doi:10.1016/j.csbj.2014.11.003.

[30] L. Wolpert. Positional information and the spatial pattern of cellular differentiation. *Journal of theoretical biology*, 25:1–47, 11 1969. doi:10.1016/S0022-5193(69)80016-0.

[31] S. Pope and R. Medzhitov. Emerging principles of gene expression programs and their regulation. *Molecular Cell*, 71:389–397, 08 2018. doi:10.1016/j.molcel.2018.07.017.

[32] M. Levine and E. Davidson. Gene regulatory networks for development. *Proceedings of the National Academy of Sciences of the United States of America*, 102, 04 2005. doi:10.1016/B978-012088563-3/50022-5.

[33] M. Babu et al. Structure and evolution of transcriptional regulatory networks. *Current opinion in structural biology*, 14:283–91, 07 2004. doi:10.1016/j.sbi.2004.05.004.

[34] W.H. Press et al. *Numerical Recipes: The Art of Scientific Computing,* 3rd edition. Cambridge University Press, 2007.

[35] M.A. Shea and G.K. Ackers. The or control system of bacteriophage lambda. a physical-chemical model for gene regulation. *Journal of molecular biology*, 181:211–30, 02 1985.

[36] D. Gillespie. Exact stochastic simulation of coupled chemical-reactions. *J. of Physical Chemistry*, 81:2340–2361, 12 1977. doi:10.1021/j100540a008.

[37] Satija Lab. Seurat - r toolkit for single-cell genomics. https://satijalab.org/seurat/. [accessed 25 nov 2021].

[38] https://satijalab.org/seurat/reference/findclusters. [acessed 25 nov 2021].

[39] L. McInnes et al. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3:861, 09 2018. doi:10.21105/joss.00861.

[40] Private communication with Pedro Rifes, of Rifes et al (2020).

[41] https://docs.scipy.org/doc/scipy/reference/generated/scipy.integrate.odeint.html. [acessed 9 dec 2021].

[42] https://numba.pydata.org/. [acessed 9 dec 2021].

[43] A. Ghaffarizadeh, N. Flann, and G. Podgorski. Multistable switches and their role in cellular differentiation networks. *BMC bioinformatics*, 15:S7, 05 2014. doi:10.1186/1471-2105-15-S7-S7.

[44] V. Chickarmane et al. Transcriptional dynamics of the embryonic stem cell switch. *PLoS computational biology*, 2:e123, 10 2006. doi:10.1371/journal.pcbi.0020123.

Appendices

A Biological background

A.1 Fundamentals of development

A fully developed multi-cellular organism can contain a wide range of different cell types, which all have unique visual appearances and functional roles despite containing essentially the same genetic material in the form of deoxyribonucleic acid (DNA). The central dogma of molecular biology, illustrated in figure A.1, states that the information encoded in a gene's DNA is transcribed into messenger-ribonucleic acid (mRNA), which in turn undergoes translation - this produces a particular protein. Each protein performs one or several specific functions, and together all produced proteins can be said to orchestrate much of the organism's total behaviour [25]. From this chain of reasoning it should thus follow that cells identical in genomic composition will also necessarily be identical in form and function, but clearly that is *not* true. Hence, there must be something beyond the organismic genome as such which determines the identity of a particular type of cell.

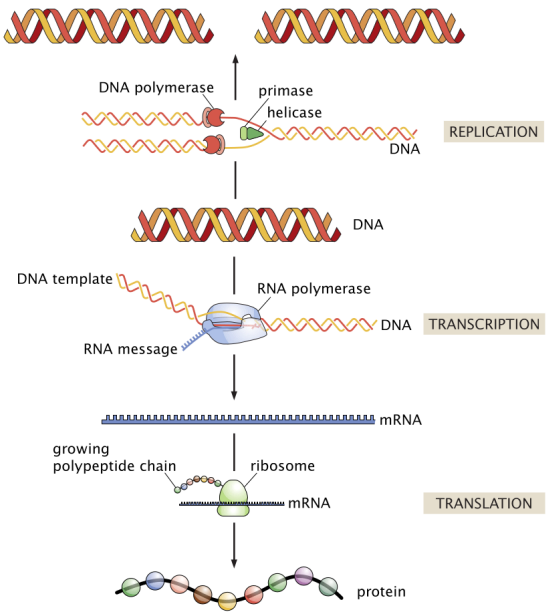


Figure A.1: The central dogma of molecular biology. Information stored in DNA as a gene - a unique sequence of base pairs - is transcribed by the enzyme RNA polymerase, which produces strings of mRNA. The mRNA is then translated into a unique chain of amino acids in a ribosome, and the gene's protein is thereby synthesized. Proteins can carry out a wide range of tasks, both inside and outside of a cell. Figure obtained from [25].

The apparent conundrum above is resolved with the insight that not all genes are actively coding for proteins (or being *expressed*) at a given moment in time - rates of transcription (and also translation) can vary greatly over time and across a genome due to a wide range of regulatory mechanisms [26]. Hence, protein levels can vary greatly between cells, which

explains differences in their form and function [27]. It is thus reasonable to identify a particular type of cell by its current state of gene expression and regulation. Given this fact, a natural question to ask is then how the cells in an embryo, who initially reside in only a few distinct gene expression states, develop into the multitude of different cell types which comprise the fully developed organism.

A developing embryo initially consists of so-called embryonic stem cells (ESC:s). They are *pluripotent*, meaning that they can develop into any kind of cell in the organism. This process of commitment towards more specific cell types is known as *differentiation*. It occurs in a directed and tree-like manner, such that a cell which has committed towards a somewhat more specific fate is restricted to develop into the corresponding sub-fates: it has become a *multipotent progenitor* rather than a pluripotent ESC. After one or several branching points in the differentiation tree, a progenitor will have committed to a final cellular state - this may be a skin cell, a neuron, or any other kind of highly specialized cell. Interestingly, it is also possible to reverse the differentiation process - to revert a fully committed cell back into the ESC state [28]. Such a reverted / reprogrammed cell is known as an induced pluripotent stem cell (iPSC). Figure A.2 illustrates some of the main aspects of a stem cell's possible differentiation paths, in what is known as a lineage tree.

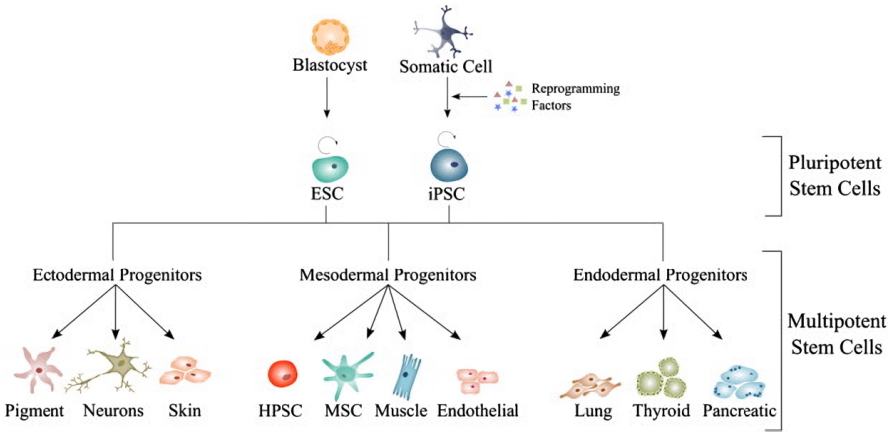


Figure A.2: A simplified lineage tree which shows possible paths of stem cell differentiation. Both embryonic and induced stem cells can commit to any of the main types of multipotent progenitor cells, who in turn can commit to increasingly specific cell fates. Figure obtained from [29].

To understand an organism's development from embryo to adult, one must be able to explain what makes some ESC:s differentiate into one particular lineage, and others into other lineages. In addition to this, one must also understand the spatio-temporal mechanisms which ensure that differently differentiated cells arrange themselves into a spatial distribution which corresponds to the organism's body plan - this is the *patterning* process [30], which is studied for the neural tube in this thesis. In a complete picture of development, patterning must be understood to occur in the context of cell proliferation, i.e. tissue growth.

A.2 WNT signalling

The WNT signalling pathways are highly conserved across phyla, and play important roles in a variety of developmental and other processes. For specifying cell fate in a neural tube development context, so-called canonical WNT signalling is thought to play the greatest role. It transcriptionally regulates specific target genes, whereas the non-canonical pathways regulate e.g. cytoskeleton structure and calcium levels within cells via other signalling cascades [22].

Figure A.3 below illustrates canonical WNT signalling by showing the intra-cellular responses to both an absence and a presence of extra-cellular WNT, which originates from cells that produce and secrete WNT proteins. In the absence of extra-cellular WNT, a complex of the proteins GSK3, Axin, and APC is present within cells. This complex binds to β-catenin, which is also present within cells, and thereby mediates the destruction of β-catenin via proteasomal degradation. This prevents β-catenin from reaching the nucleus, where it would have otherwise transcriptionally activated several target genes that are important for specifying neural cell fates. If, on the other hand, extracellular WNT is present, the destruction complex is destroyed. This allows β-catenin to reach the nucleus and activate its target genes. The destruction of the complex is mediated by the binding of extra-cellular WNT to the Frizzled (Fzd) and Lrp5/6 trans-membrane receptors. Via the Dishevelled (Dvl) protein, translocation of Axin and GSK3 to the surface membrane is initiated, whereupon the destruction complex activity is disrupted.

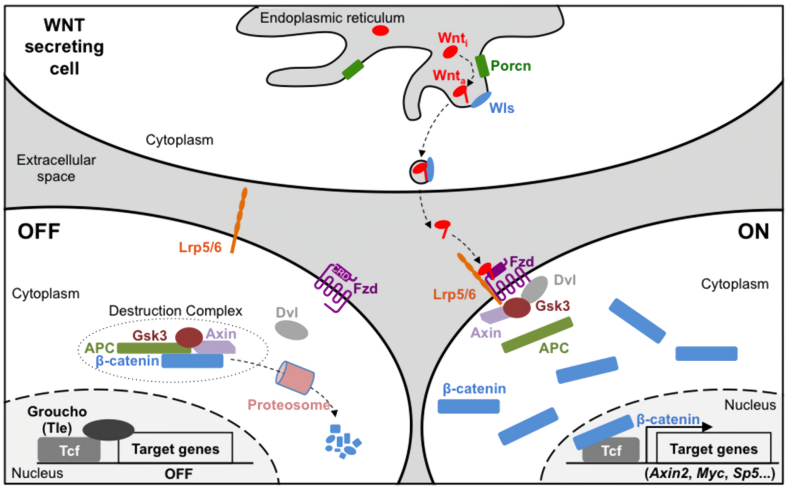


Figure A.3: WNT secretion, transport, and canonical signalling. Secreted WNT proteins are transported inter-cellularly, and bind to Frizzled (Fzd) receptors on other cells (ON). This initiates the destruction of a destruction complex, which, in the absence of WNT signalling (OFF), would have prevented β-catenin from reaching the nucleus to activate specific target genes. Figure from [22].

From figure A.3, it is also clear why the inhibition of GSK3 can be used to emulate the ON-state of canonical WNT signalling: as GSK3 is required for the destruction complex to function, inhibiting GSK3 allows β-catenin to activate its target genes for the same reason as when extra-cellular WNT is bound to its target receptors at the cell membrane.

B Modelling gene transcription and its regulation

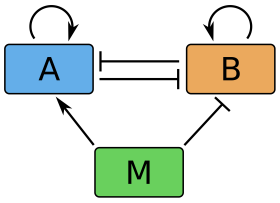
B.1 Fundamentals of gene expression and regulation

Let G be the entire protein-coding part of an organism's genome, i.e. the set of all its N genes g. Further, let $[\mathrm{G}](t) = ([\mathrm{g}_1](t), [\mathrm{g}_2](t), ..., [\mathrm{g}_N](t))$ be a vector containing the corresponding levels of gene expression at a given point in time t. This notation captures that gene expression is both time-dependent and may vary between genes. The variation of $[\mathrm{G}](t)$ is largely due to *regulation* of gene expression. There are many ways in which a given gene g_i can be regulated. Regulation can, for example, occur on the transcriptional level, via epigenetic-level modifications of DNA, or on the level of mRNA translation [31]. Throughout this work it will be assumed that all regulation is transcriptional, since it is well-known that transcriptional regulation plays a major role in many developmental processes [32]. In this case, it is natural to identify $[\mathrm{G}](t)$ as concentrations, or levels, of mRNA. For the sake of readability, the explicit time-dependence of [G] will, from this point forward, not be written out explicitly unless necessary.

Transcriptional regulation involves genes regulating each other. Some genes are translated to proteins known as *transcription factors* (TF:s), who act as regulatory input for transcription. Such input can be broadly categorized as *activating* or *repressing*. A TF which acts as an activator for the regulation of a gene g_i increases the rate of transcription $\frac{d[\mathrm{g}_i]}{dt}$ above a basal rate, for example by making it easier for RNA polymerase to bind to the relevant DNA segment. On the other hand, a repressor TF lowers the rate of transcription below a basal rate, for instance by directly blocking access for RNA polymerase to the gene's promoter region by binding to a segment of DNA adjacent to it.

From the above it follows that we can write down a *regulation matrix* $\mathbf{R}$, whose elements R_{ij} describe the type of regulatory action by gene g_i on gene g_j in G. In such a formulation one could, for instance, represent activation by $R_{ij} = +1$ and repression by $R_{ij} = -1$. If there is no regulatory effect by g_i on g_j, one can put $R_{ij} = 0$.

It is known that far from all genes act as TF:s, and those who do often only regulate the transcription of one or a few genes [33]. Hence, we have that $\mathbf{R}$ for any particular organism would be *sparse* - many of its entries would be zero. If one can then establish that only a few genes seem to be involved in a certain developmental process, one can limit oneself to only consider the sub-matrix **R'** which includes only the genes relevant for that particular process. **R'** determines the connectivity of a *gene regulatory network* (GRN), which consists of genes as nodes and regulatory interactions as directed edges. From this point forward, all further mentions of G refer to the set of genes in such a regulatory network, and not to the full genome. Activating GRN interactions are commonly represented by arrows, and repressing interactions are shown as blunt arrows. Figure B.1 on the following page depicts an example GRN and its corresponding regulation matrix.



(a)

$$\mathbf{R'} = \begin{bmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 1 & -1 & 0 \end{bmatrix}$$

(b)

Figure B.1: An example of a gene regulatory network (a) with genes A, B, and M, and their corresponding regulation matrix (b), with rows / columns in that order.

Neither **R'** nor a graphical depiction of a regulatory network like above embodies the fact that regulatory strengths and other important mechanistic details vary between genes - they only describe the network topology and types of regulatory connections.

B.2 A continuous and deterministic framework: the Shea-Ackers formalism

Given a GRN and expression levels $[\mathrm{G}](t_0) = [\mathrm{G}_0]$ of its constituent genes for some initial time t_0, we are interested in simulating $[\mathrm{G}](t)$ for times $t > t_0$. With the assumption of continuous time and levels, this problem is suitably viewed as that of solving a set of ordinary differential equations (ODE:s). Namely, we wish to solve

$$\frac{d[\mathrm{G}](t)}{dt} = f([\mathrm{G}](t)) \qquad \text{subject to} \quad [\mathrm{G}](t_0) = [\mathrm{G}]_0, \tag{B.1}$$

where f is some function of the current state of the network. This kind of ODE:s can be solved numerically using standard methods like the Euler or Runge-Kutta methods, which is particularly useful if f is complicated [34]. To solve eq. (B.1), we need to establish a form of f which is an accurate model of the network dynamics. It is assumed that mRNA can be produced via transcription, and decays at a rate proportional to [G]: this allows us to write f as the difference between a production term P and a decay term. For a given gene g_i, f will then be on the following form:

$$f([\mathrm{g}_i]) = P_i - \delta_i[\mathrm{g}_i], \tag{B.2}$$

where δ_i is the exponential decay constant of mRNA i. We now only need the form of P - this is the part of the model most laden with assumptions. Historically, mRNA transcription has been modelled as a catalysis-like process, where n TF:s bind to DNA to form an intermediate complex DNA_{bound} from which a protein X is (eventually) produced:

$$n \cdot \mathrm{TF} + \mathrm{DNA} \rightleftharpoons \mathrm{DNA}_{\text{bound}} \rightarrow \mathrm{X} + \mathrm{DNA}_{\text{bound}}. \tag{B.3}$$

If $n > 1$ we have *cooperativity*, i.e. that multiple TF:s of the same kind must jointly bind to DNA to have an effect on the rate of transcription.

Given the picture of mRNA transcription provided by eq. (B.3), there are multiple ways to find a production function P which faithfully captures how the rate of transcription is regulated by TF:s binding to / unbinding from DNA. In this thesis, the Shea-Ackers (S-A) formalism [35] is used. It is based on a statistical mechanics picture of gene regulation: P for a specific gene g_i is written as a fraction, which in a partition function-like manner yields the transcription rate from the probability of the gene's DNA being in a state of active transcription. For a single gene with basal transcription rate b_i being regulated by N_R and N_A *independently* acting repressor genes R and activator genes A, respectively, $P_{i,\,\mathrm{S-A}}$ is obtained as

$$P_{i,\,\mathrm{S-A}} = \frac{b_i + \sum_{j=1}^{N_A} c_j \left[\ \mathrm{A}_j\right]^{n_j}}{1 + \sum_{j=1}^{N_A} c_j \left[\ \mathrm{A}_j\right]^{n_j} + \sum_{k=1}^{N_R} c'_k \left[\mathrm{R}_k\right]^{n'_k}}, \tag{B.4}$$

where c_j/c'_k and n_j/n'_k represent regulatory strengths and cooperativities, respectively. If different TF:s do not act independently of each other, the form of $P_{i,\,\mathrm{S-A}}$ will differ somewhat from eq. (B.4). However, unless there is data indicating a specific logic for the regulatory process one wants to model, it may be a good idea to initially assume independent TF action. That assumption is made for the regulatory processes considered in this thesis. Furthermore, it is assumed that all protein levels (and thereby also the TF levels in eq. (B.4)) directly follow the levels of their corresponding mRNA.

As an example of applying eq. (B.4), the S-A ODE:s for the toy network in figure B.1 would be written as in eq. (B.5). [M] is assumed to have reached steady-state, since it plays the role of a morphogen which patterns the A-B sub-network (see appendix E.2).

$$\begin{aligned}
\frac{d[\mathrm{A}]}{dt} &= \frac{b_1 + c_1[\mathrm{A}]^{n_1} + c_2[\mathrm{M}]^{n_2}}{1 + c_1[\mathrm{A}]^{n_1} + c_2[\mathrm{M}]^{n_2} + c_3[\mathrm{B}]^{n_3}} - \delta_1[\mathrm{A}], \\
\frac{d[\mathrm{B}]}{dt} &= \frac{b_2 + c_4[\mathrm{B}]^{n_4}}{1 + c_4[\mathrm{B}]^{n_4} + c_5[\mathrm{M}]^{n_5} + c_6[\mathrm{A}]^{n_6}} - \delta_2[\mathrm{B}], \\
\frac{d[\mathrm{M}]}{dt} &= 0.
\end{aligned} \tag{B.5}$$

Solving the S-A ODE:s for a particular regulatory network results in gene expression levels $[\mathrm{g}_i](t)$ for all genes in the network. The level of each gene is bounded in the interval $[0, \mathrm{g}_{i,\mathrm{max}}]$, where $\mathrm{g}_{i,\mathrm{max}}$ is the maximum expression level of gene i. The total $[\mathrm{G}](t)$ can then be thought of as a "state-vector" which evolves over time and traces out a trajectory through the state-space of the network. The dynamics of such a trajectory depends on the network's connectivity architecture and initial conditions, and on the ODE model parameters. Some combinations of network architectures / initial conditions / parameters may lead to particular points in the state space functioning as attractors, or stable states, which the network can settle into after some time. In the context of a gene regulatory

network which controls cell differentiation / patterning, it then makes sense to identify such a stable state as a particular differentiation fate.

It should be said that an ODE-based approach assumes a completely deterministic time-evolution of the gene expression levels. Especially if there are few copies of each gene, this assumption may break down - stochastic fluctuations then become important. In that case, one has to resort to stochastic models such as making use of the Gillespie algorithm [36]. However, in this thesis it is assumed that stochastic dynamics are not relevant for the processes to be modelled - hence, only fully deterministic models are considered.

C Processing of experimental data

The supplied bulk and single-cell gene expression data had to be further processed in order to be useful as time series to optimize models against. These processing steps are detailed below. The actual data is displayed in section 2.3.

C.1 Bulk data

Gene expression levels E obtained from the bulk data measurements for cells from a particular MiSTR experiment can be represented as a 3-index array of gene expression levels, E_{gtl}, where indices g, t, and l represent a particular gene, time point, and one of the locations A-E, respectively. Data for 36 separate MiSTR experiments were supplied for a selection of genes relevant for early neural tube development.

The mean expression levels $\bar{E}$ across all $N = 36$ experiments were calculated via an element-wise average of the corresponding E:s, so that $\bar{E} = \frac{1}{N}\sum_{n=1}^{N} E^{(n)}$. For the purpose of fitting a regulatory network to the expression levels, $\bar{E}$ was normalized across time t and space l, separately for each gene, resulting in $\bar{E}^{(\text{norm.})}$. The normalization was made by dividing all expression levels for a particular gene by the maximum level of that gene across all times and locations. This conveniently transforms all expression levels to lie within the range $[0, 1]$, thereby making it possible to easily compare the spatio-temporal behaviour of different genes in the same plot window or heat map (figure 2.5).

Standard deviations σ_{gtl} of gene expression levels were calculated in the same manner as the mean expression levels, i.e. element-wise over the N MiSTR data arrays. Normalized standard deviations $\sigma^{(\text{norm.})}$ were then calculated as $\sigma^{(\text{norm.})} = \sigma\bar{E}/\bar{E}^{(\text{norm.})}$ Hence, if $\sigma_{gtl}^{(\text{norm.})} = 1$, the standard deviation equals the expression level. Standard deviations are shown as a heat map in figure C.1 on the following page.

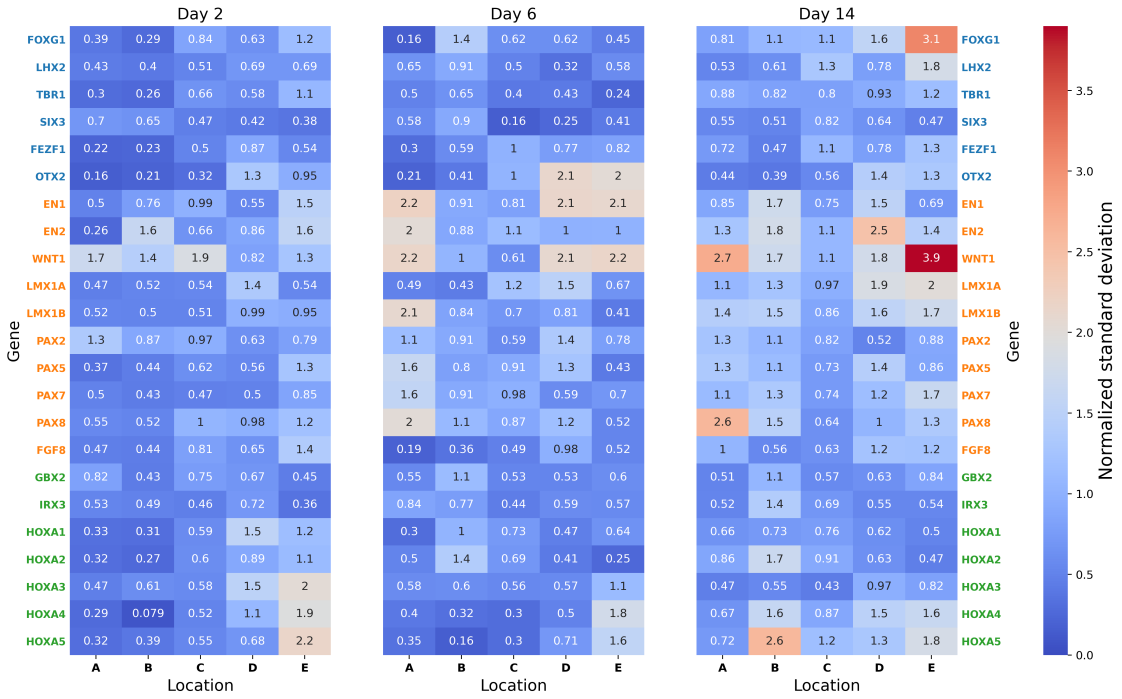


Figure C.1: Heat map showing normalized standard deviations of bulk data gene expression levels. Gene names are coloured according to whether the genes are characteristic of forebrain (blue), midbrain (orange), and hindbrain (green).

C.2 Single-cell data

For each time point, the sc-RNAseq data sets were supplied as so-called count matrices C_{gc} with genes g as rows and cells c as columns, with a count of n signifying n detections of a particular gene's mRNA in a cell during the single-cell sequencing. The count matrices had been pre-processed in the programming language R, using the package "Seurat" [37], to account for confounding factors such as batch effects and cell-cycle heterogeneity. They had also been subjected to principal component analysis (PCA) and subsequent clustering via a combination of a shared nearest neighbours (SNN) and a Louvain algorithm [38]. Only the 2000 most variable genes for each time point were selected to be clustered, as this helps in eliminating spurious clusters [37].

The goal of further processing of the count matrices was to construct time series. To be useful for correlation clustering and parameter fitting, they should have a structure similar to bulk data time series. A requirement for this is to have gene expression levels for multiple distinct CT values, or locations. This may be achieved by sorting cells into distinct groups, signifying locations, based on their position along the CT gradient in the MiSTR setup. If each cell would have been tagged with its position before being subjected to sc-RNAseq, this kind of grouping could have been performed as a trivial sorting and subsequent binning of columns in C_{gc} into the desired number of locations. However, no such spatial metadata was available. Hence, the spatial origin of a particular

cell in C_{gc} had to be inferred from its gene expression profile for known marker genes. The spatial grouping of cells was performed in R: the procedure is outlined as follows. First, the 2000-dimensional data had to be dimensionally reduced to permit meaningful visualization of clusters. This was performed using the algorithm "Uniform manifold approximation and projection" (UMAP), since it has been shown to preserve both local and global structure better than other popular dimensionality reduction techniques like e.g. t-SNE [39]. Using UMAP, the dimensionality was reduced to 2. Figure C.2 below shows the cell clusters in 2D UMAP space, in this case for the day 9 data set.

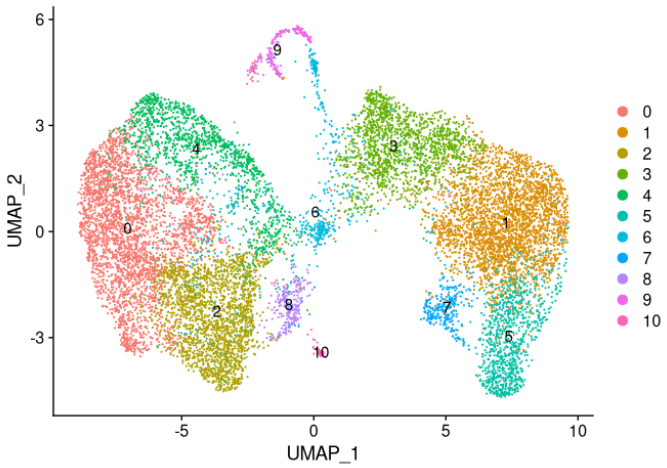
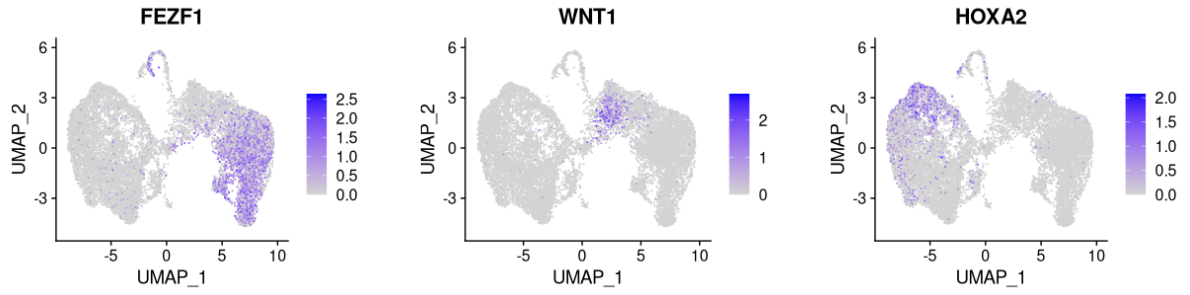


Figure C.2: 2D UMAP plot of the day 9 scRNAseq data set, with annotated clusters.

Since FB, MB, and HB regions of the neural tube are known to be characterized by high expression levels of specific marker genes at particular times, it is possible to use this knowledge to assign FB, MB, or HB identities to at least some of the clusters in figure C.2. An example is shown in figure C.3 below, once again for the day 9 data set: here, the distribution of expression levels is plotted for FEZF1 (FB marker), WNT1 (MB marker), and HOXA2 (HB marker).



(a) (b) (c)

Figure C.3: Distribution of selected marker gene expression levels in the 2D UMAP space from figure C.2 for the day 9 sc-RNAseq data set. (a)-(c) show key marker genes for FB, MB, and HB, respectively. Note the spatial separation between groups of cells exhibiting high expression of different marker genes.

It is clear that the rostral-caudal axis goes approximately from right to left in this UMAP visualization of the data set. Therefore, e.g. cluster 3 in figure C.2 consists predominantly of midbrain cells. Hence, those cells are likely to come from somewhere in the centre of the MiSTR tissue. By this kind of reasoning, and by using appropriate marker genes for each time point [40], three spatial regions per time point were identified. These were assumed to correspond to locations A-B, C, and D-E of the bulk data, and thus to CT levels of 0.6, 1.0, and 1.4, respectively. The MiSTR tissue was essentially spatially uniform at day 0 - the MiSTR cells are initially hESC:s. Hence, it was not possible to group the day 0 data into separate locations.

The expression levels of the cells in each location were then averaged separately for each gene. The levels obtained from this averaging were assumed to represent the mean gene expression levels for the corresponding locations. After this, a genome-wide normalized time series was created in the same way as for the bulk data, as was described in section C.1.

D Optimizing regulatory models against experimental data

The particular optimization approach used in this thesis can be justified as follows. Both of the considered regulatory model equations have a large parameter set $\mathbf{p}$ (35 and 36 parameters for model 1 and 2, respectively). This implies a high-dimensional cost function landscape $C(\mathbf{p})$, which immediately rules out brute-force optimization approaches[3]. Furthermore, the available gene expression time series to be optimized against are quite sparse in time. Fitting a complex model to only a few (as opposed to many) time points means that there can be a high degree of freedom in the kinds of network dynamics which results in reasonably small values of $C(\mathbf{p})$. Hence, C may have a large number of local minima. A purely local, gradient-based optimization algorithm would then get stuck in the closest local minimum, but more lower lying / optimal local minima of C may exist. To overcome this difficulty and find a $\mathbf{p}$ which is as optimal as possible, an approach which combines global and local optimization was used. This approach is described below.

The time-evolution of the Shea-Ackers equations of a particular regulatory model was optimized against experimental data in a two-step process, where a global optimizer (Differential Evolution, DE) was first used in a broad search for promising local minima in the cost function. Parameters corresponding to the bottom of the found minimum were then determined via a final, local, optimizer (L-BFGS-B), which used the winning parameters of the DE routine as starting parameters. The cost function $C(\mathbf{p})$ which was minimized can be found in section 3.1. The flowchart in figure D.1 outlines the Python-based optimization routine.

[3] even if each parameter axis is coarsely discretized by an array of length ~ 10, a model with e.g. 35 parameters requires $\sim 10^{35}$ evaluations of $C(\mathbf{p})$ for an exhaustive search - this is clearly intractable.

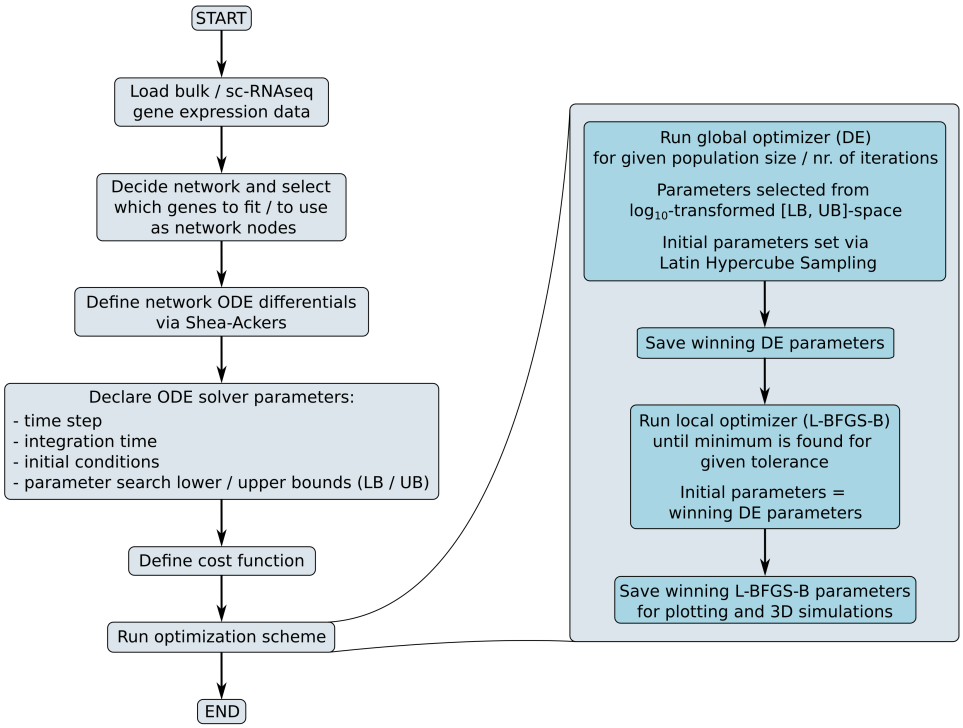


Figure D.1: Flow chart which describes the parameter optimization routine. Importantly, optimization is performed as a two-step process: a global optimizer first scans a wide region of the parameter space to find promising minima, and the bottom of the most promising minimum is then determined via a local optimizer.

ODE:s were solved with time in units of hours, using a total integration time of 12 days (288 h) and 14 days (336 h) for bulk and sc-RNAseq data, respectively. Experimentally measured expression levels at the first time point where data were available were used as initial conditions. The ODE:s were solved using SciPy:s "odeint" routine [41], which was chosen over a simple Runge-Kutta 4 routine due to speed considerations. This mostly stems from that the "odeint"-routine automatically detects ODE stiffness and switches between two solvers depending on if the problem is detected as stiff or not. Importantly, the Python library "numba" was also used to speed up the ODE solving by another factor of ≈ 10 [42]. The time step was 1/5000:th of the total integration time. Parameter bounds were, after trial and error, chosen as follows (written as [lower bounds, upper bounds]):

$$c_i\text{: } [10^{-3}, 10^{4}] \qquad n_i\text{: } [1, 4] \qquad \delta_i\text{: } [10^{-2}, 10^{0}] \qquad b_i\text{: } [10^{-5}, 10^{1}]$$

Parameter values were uniformly sampled in $\log_{10}$-spaces of the upper and lower bounds above. Initial DE parameters were chosen via so-called Latin hypercube sampling to ensure an even coverage of the parameter space. DE was run using a population size ranging from 100 to 300, with the total number of DE iterations ranging from 1000 to 2500. L-BFGS-B was run until $C(\mathbf{p})$ did not change by more than 10^{-4} between iterations.

E Spatial models of patterning: reaction-diffusion

E.1 Patterning on a 1D lattice

Patterning is in this thesis modelled to take place on a lattice, where each lattice site s represents a slab of tissue (in other words, a particular number of cells) to be patterned. For simplicity, consider a one-dimensional (1D) N_S-point lattice with the lattice sites located at x-ccoordinates $(x_1, x_2, \ ... \ , x_s, \ ... \ , x_{N_S-1}, x_{N_S})$. In each lattice point s, we wish to solve the gene regulatory ODE:s which govern the patterning. To do so for any time t, we need to simultaneously know the spatio-temporal distribution of the morphogen $[\mathrm{M}](x,t)$ over the lattice. If morphogens are transported inter-cellularly via diffusion-like processes, it is reasonable to assume that the spreading of $[\mathrm{M}](x,t)$ in time and space is, in the continuum limit, well described by the *diffusion equation*:

$$\frac{\partial[\mathrm{M}](x,t)}{\partial t} = D_{\mathrm{M}} \frac{\partial^2[\mathrm{M}](x,t)}{\partial x^2}, \tag{E.1}$$

where D_{M} is a diffusion constant with dimension [length]^2/[time], which determines how easily M diffuses. If an additional *reaction* term $R(x,t)$ is added to the right hand side of eq. (E.1), this is known as a *reaction-diffusion* problem. The reaction term describes the local production and decay of M.

To be useful on a lattice, eq. (E.1) must first be re-written into a discrete finite-difference form. By discretizing time $t = (t_1, t_2, \ ... \ , t_n, \ ... \ , t_{N_T-1}, t_{N_T})$, using a forward Euler approximation of the temporal first derivative, and a central difference approximation of the spatial second derivative, the discrete version of eq. (E.1) with an added reaction term is:

$$\frac{[\mathrm{M}]_s^{(n+1)} - [\mathrm{M}]_s^{(n)}}{\Delta t} = D_{\mathrm{M}} \frac{[\mathrm{M}]_{s+1}^{(n)} + [\mathrm{M}]_{s-1}^{(n)} - 2[\mathrm{M}]_s^{(n)}}{(\Delta x)^2} + R_s^{(n)}, \tag{E.2}$$

where $\Delta x = |x_{s+1} - x_s|$ is the lattice spacing and $\Delta t = |t_{n+1} - t_n|$ is the time step. With a specific form of $R_s^{(n)}$, together with initial and boundary conditions for [M], one can numerically solve the reaction-diffusion problem constituted by eq. (E.2) for all lattice sites and times. By extension, this allows one to solve the patterning ODE:s on the lattice.

Like in models of gene transcription (see appendix B), a reasonable first assumption is a decay of M which is proportional to its current level, with a decay constant δ_{M}. Furthermore, M should only be produced on the lattice sites which constitute the morphogen source regions. Even with these requirements met, there is a lot of freedom in choosing a particular form of $R_s^{(n)}$ - the two forms chosen in this thesis, equations (3.3) and (3.4), are among the simpler ones.

E.2 The bi-stable switch in a patterning context

The right kind of network architecture can act to sharpen the boundary between two differently patterned regions, and can maintain the pattern even after the loss of morphogen signalling. One such architecture is the *bi-stable switch* or *toggle switch*. For the toy network shown in figure B.1, the mutual repressions between genes A and B together with their self-activation constitute such a switch. The switching behaviour can be understood as follows. If the expression of A is initially sufficiently much higher than that of B, A's self-regulation will lead to [A] settling into a steady-state value. Due to A's repression of B, this would also suppress B towards [B] = 0, or at least towards a low steady-state value. Due to the symmetry of the AB sub-circuit, an initially high-enough value of [B] would suppress the expression of A in the same way. Figure E.1 below illustrates such a situation for the toy network in figure B.1, where the time-evolution of the bi-stable switch AB is regulated by a steady-state spatial morphogen gradient $[M]^*(x)$ on a 1D lattice with 1000 points. We see that there indeed is a very sharp boundary where the A/B dominance shifts rapidly in space, and that A dominates where [M] is high (despite the initial conditions) and vice versa as expected, since M both induces A and represses B. The boundary location depends on how strongly M affects A and B.

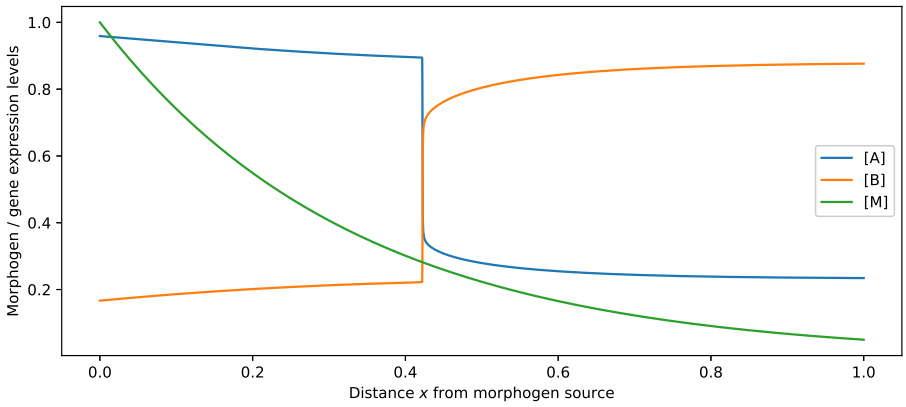


Figure E.1: Steady-state values of [A] and [B] in equation (B.5), which is solved with initial conditions $[A]_0 = 0$ and $[B]_0 = 1$ for 1000 different morphogen levels $[M]^*(x)$ to simulate the variation in A/B dominance one might observe along a spatial morphogen gradient for a bi-stable switch topology. The ODE:s were solved using all cooperativities $n = 2$, and all δ, b, $c = 1$ except for $c_3 = c_6 = 6$, the point of which is to capture a situation where there is a strong mutual repression between A and B. Strikingly, there is a critical value of $[M] \approx 0.3$ at $x \approx 0.42$ where the switch "flips", which leads to the establishment of a very sharp spatial boundary separating the two distinct stable states of the A-B switch.

The bi-stable switch is a commonly appearing sub-feature, or *motif*, in gene regulatory networks [43], [44]. Perhaps due to the mechanism outlined above, it appears to play a key role in several networks which govern early developmental patterning, for example in the network which patterns early *Drosophila* embryos due to the *Bicoid* morphogen [2]. Model 2 in this thesis contains a bi-stable motif in OTX2 and GBX2, and model 1 contains a generalized *tri*-stable motif in FB, MB, and HB.

E.3 Simulating patterning in an arbitrary 3D geometry

The goal of the 3D spatial simulations made in this thesis is to find the time-evolution of regulatory network genes in a particular 3D geometry, given a morphogen distribution [M] across this geometry. If there is a steady-state morphogen gradient $[\mathrm{M}]^*(x, y, z)$ in the geometry, this is simple: one simply computes the time evolution the network in all geometry voxels, using the local $[\mathrm{M}]^*$ as input to the network. However, if morphogens are produced over time from particular source voxels, there is coupling between voxels due to morphogen diffusion. The particular form of this coupling is geometry-specific, since different voxels may only neighbour other voxels in particular directions. It can be non-trivial to keep track of this in a complex geometry.

To address the challenges arising due to the situation described above, it is useful to re-write the 3D diffusion problem into a 1D diffusion problem. We thus want to map an (x, y, z)-coordinate to a *site number* s_i, which acts as a unique identifier for the position of the i:th voxel, or site, in (x, y, z)-space. This allows one to collect the information about the neighbours of all sites into a matrix, which is denoted $\mathbf{A}$. This can be understood as follows.

The 3D analogue of the diffusion terms in eq. (E.2) is, in a slightly re-cast form:

$$[\mathrm{M}]^{(n+1)}_{x,y,z} = [\mathrm{M}]^{(n)}_{x,y,z} + D_\mathrm{M} \cdot \Delta t \frac{[\mathrm{M}]^{(n)}_{x+1,y,z} + [\mathrm{M}]^{(n)}_{x-1,y,z} + [\mathrm{M}]^{(n)}_{x,y+1,z} + [\mathrm{M}]^{(n)}_{x,y-1,z} + [\mathrm{M}]^{(n)}_{x,y,z+1} + [\mathrm{M}]^{(n)}_{x,y,z-1} - 6[\mathrm{M}]^{(n)}_{x,y,z}}{a^2},$$

where a is the voxel spacing. If we use the mapping $(x, y, z) \to s_i$, the above can be written as a matrix equation, with the notation $[\mathbf{M}]$ for a vector containing the morphogen levels in s-space for all voxel sites:

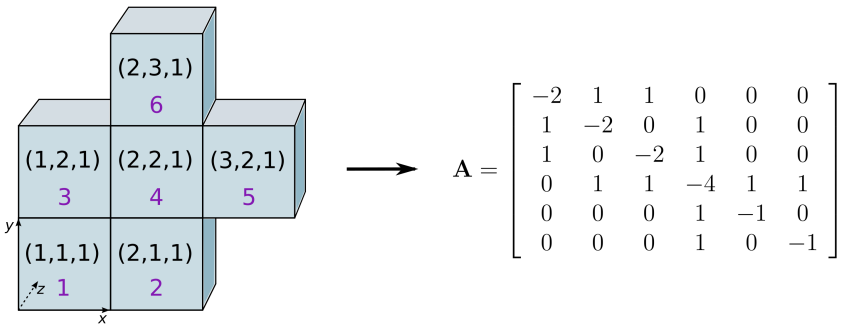
$$[\mathbf{M}]^{(n+1)} = \left(\boldsymbol{I} + \frac{D_M \cdot \Delta t}{a^2} \mathbf{A} \right) [\mathbf{M}]^{(n)}, \tag{E.3}$$

where $\boldsymbol{I}$ is the identity matrix. The matrix $\mathbf{A}$ will then consist of the following

- diagonal elements $A_{s_i s_i}$: minus the number of nearest neighbours of s_i in (x, y, z)-space,
- off-diagonal elements $A_{s_i s_j}$, $i \neq j$: one if s_j neighbours s_i in (x, y, z)-space, zero otherwise.

The above ensures that the morphogen never diffuses outside of the simulation geometry: no-flux (Neumann) boundary conditions are imposed on the geometry edges. Figure E.2 on the following page shows a very simple example 3D geometry, its (x, y, z) coordinates, the mapped s-values in purple, and the corresponding $\mathbf{A}$-matrix.

With the A-matrix found for the geometry of interest, the relevant morphogen and gene regulatory equations are solved for all voxel sites, with a particular choice of time step



$$\mathbf{A} = \begin{bmatrix} -2 & 1 & 1 & 0 & 0 & 0 \\ 1 & -2 & 0 & 1 & 0 & 0 \\ 1 & 0 & -2 & 1 & 0 & 0 \\ 0 & 1 & 1 & -4 & 1 & 1 \\ 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 \end{bmatrix}$$

Figure E.2: Example showing how to find the **A**-matrix for a particular geometry, with a specific choice of a map $(x, y, z) \to s$.

Δt and total integration time. Figure E.3 below outlines the full Python-based patterning simulation scheme in a flow chart. Morphogen production and decay was solved in together with the diffusion terms, i.e. eq. (E.3), using the explicit Euler scheme. The gene regulatory equations were solved using a 4^{th} order Runge-Kutta scheme. As eq. (E.3) is only conditionally stable [34], care had to be taken when choosing the size of Δt. Since morphogen levels were always found to have reached steady-state after less than an hour, morphogen levels were only updated if $t < 2$ h. This allowed a much larger time step after $t = 2$ h, which helped in significantly speeding up the simulations. As the CT gradient in the MiSTR geometry is steady-state already from the start (see figure 4.3), the first simulation step of updating morphogen levels before $t < 2$ h was omitted for simulations made in the geometry which is a model of the MiSTR setup (i.e. the one in figure 4.3).

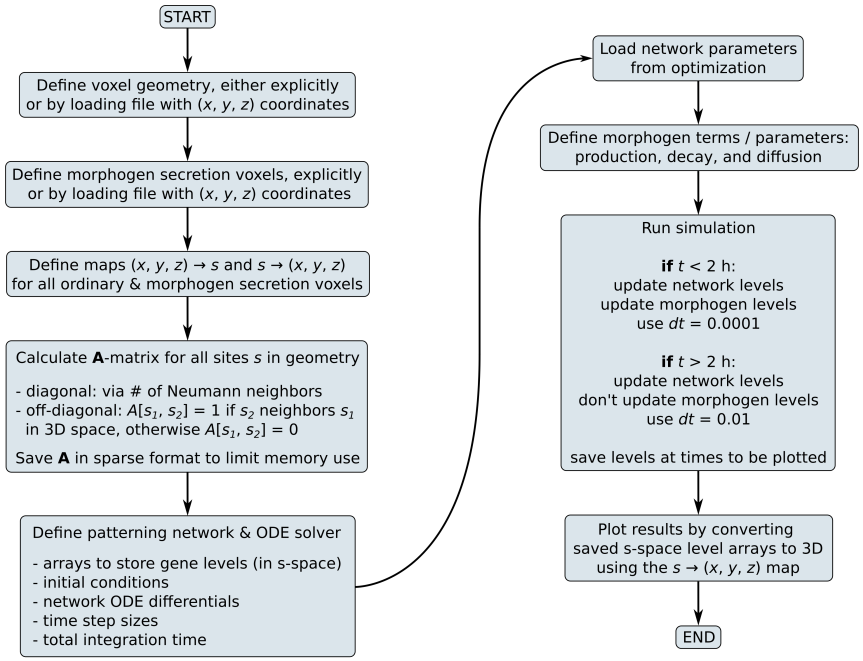


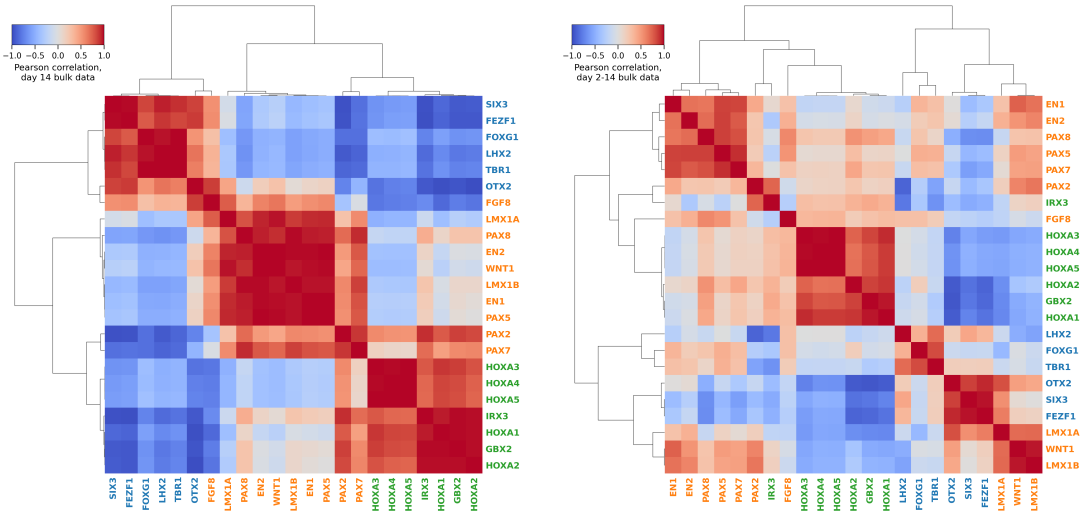
Figure E.3: Flow chart which describes the 3D simulations.

F Supplementary results

F.1 Clustering of gene expression data

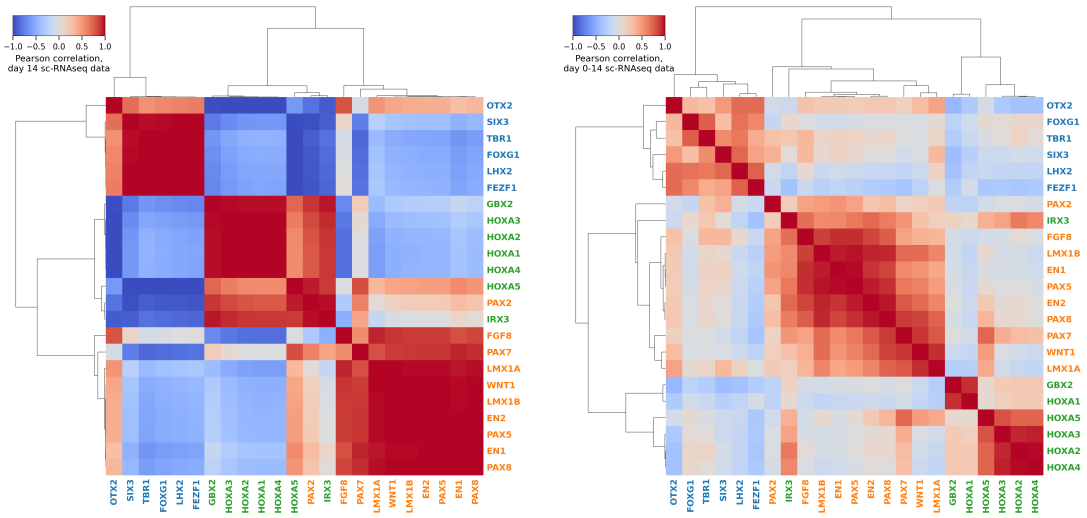
Figure F.1(a) below shows the result of bottom-up, hierarchical clustering applied on the day 14 bulk data, while figure F.1(b) shows the result of clustering on all bulk data time points. Figure F.2 on the following page shows analogous results for the sc-RNAseq data. In both figures, the heat maps show the Pearson correlation between pairs of genes, and the dendrograms show the hierarchy of similarity between the expression profiles of genes and / or clusters of genes. Note the re-ordering of gene names compared to figure 2.5.

If the grouping of genes into FB, MB, and HB categories is meaningful, most of the genes within each of these groups should be found on the same branch in the dendrograms, at a dendrogram level where there are three branches. For both bulk and sc-RNAseq data for day 14, this is generally the case, but there are a few outliers. For the bulk data, FGF8 is grouped with all the FB genes, and both PAX2 and PAX7 are in the cluster which otherwise contains only HB genes; for the sc-RNAseq data, PAX2 is ordered into the cluster containing all HB genes. However, most genes are clustered correctly according to their FB / MB / HB pre-clustering labels. Furthermore, very similar clusters emerge from two completely different data sets, and strong intra-cluster correlations imply that the three top-level clusters are quite distinct - this is particularly noticeable for the sc-RNAseq data. Taken together, these observations imply that the FB / MB / HB grouping is robust and meaningful at day 14, in agreement with Brambach et al. (2021) [11]. Discrepancies can be attributed to both noisy data and that three clusters do not perfectly capture the full patterning picture.



(a) (b)

Figure F.1: Pearson correlation coefficients and cluster dendrograms for bulk data. (a) Clustering on day 14 data only. (b) clustering on data from all time points, from day 2 to day 14.



(a) (b)

Figure F.2: Pearson correlation coefficients and cluster dendrograms for sc-RNAseq data. (a) Clustering on day 14 data only. (b) clustering on data from all time points, from day 0 to day 14.

When it comes to simultaneous clustering to all time points, correlations within clusters are generally weaker, and it is not as apparent that the data can be meaningfully organized into FB, MB, and HB genes. This is especially the case for the bulk data - the sc-RNAseq data better retains the FB / MB / HB categories. The weaker intra-cluster correlations likely stem from that additional time points bring forward finer details in gene expression dynamics, and thereby makes genes less similar to each other when clustering. This is supported by the fact that this effect is stronger for the sc-RNAseq data set, which has data for more time points than the bulk data. Despite this, the sc-RNAseq clustering retains the FB / MB / HB categories better than the bulk data clustering. This could at least in part be due to selection bias, as some of the clustered FB / MB / HB genes were used as marker genes in the sc-RNAseq data processing workflow.

It is clear that the amount of heterogeneity in expression profiles is significantly higher when the full dynamics from day 2 / day 0 to day 14 are considered. This raises the question of whether it justified to treat all genes within one of the FB / MB / HB categories on equal footing in a regulatory network model, as is done in the tri-stable network model in this thesis: a more sophisticated model is required to capture the dissimilarities in dynamics between different genes of the same FB / MB / HB category.