Modeling and simulation of transcriptional networks in murine cortical development

Author(s):
Pfister, Sabina Sara

Publication Date:
2013

Permanent Link:
https://doi.org/10.3929/ethz-a-010105015

Rights / License:
In Copyright - Non-Commercial Use Permitted
Modeling and simulation of transcriptional networks in murine cortical development

A dissertation submitted to the
ETH ZURICH

for the degree of
DOCTOR OF SCIENCES

presented by
SABINA SARA PFISTER
Master of Science ETH, ETH Zürich
born 10 September 1982
citizen of Switzerland, Roggwil (BE)

accepted on the recommendation of
Prof. Dr. Rodney James Douglas, examiner
  Dr. Colette Dehay, co-examiner
  Dr. Oliver Raineteau, co-examiner

2013
Abstract

Conventional engineering is about design of systems that are completely defined within a limited range of operating conditions. The system’s blueprint is dawned in a top-down fashion by an external architect, and there is a clear distinction between construction phase and operational phase. Unlike explicit design, biological entities self-construct through the processes of development: structural and functional order result collectively from energy-dissipative processes involving cell replication and local interactions among cells or molecular components embedded in the physical environment.

In this Thesis, I focus on a remarkable example of a self-constructing process, namely the development of the murine neocortex. The mouse neocortex is a highly organized structure composed of sheets of glial and neuronal cells characterized by great diversity in morphology, molecular expression profile and function. The elegant disposition of identifiable neural types into cortical laminae and their genetic characterization makes the cortex a convenient model to address the question of how such a diversity of cell types is generated.

The novelty that I wish to present, is a hypothesis-generating method to study development, and corticogenesis in particular. In particular, I address three main aspects: (i) inference of transcriptional network models (in form of algorithmic rules) from sparse biological data during corticogenesis; (ii) the development of a formal description of gene regulatory rules consistent with biology; and (iii) the simulation of transcriptional networks for the generation of a laminated cortical column in Cortex3D, a Java agent-based software.

At its core, development is the process by which multipotent progenitor cells undergo multiple rounds of symmetric and asymmetric cell divisions to finally produce differentiated cells with specific phenotypes, commonly referred to as terminal cell types. I describe this process at four different levels of abstraction, going from biological data to simulation of genetic control during development.

At the first level of abstraction, insight into biological developmental programs is provided by the genealogical history of each precursor cell, the Cell Lineage Tree (CLT). Cell lineages are the key to understand development, since they describe the number of cells generated, the distribution of cell fates, and the topology or patterns of cell divisions. The collection of cell lineage data is technically challenging, therefore I relay on reconstructions from experimentally observed cell distributions.

At the second level of abstraction, cell lineage data is described by a Markov Model with hidden states, where the characteristic features of a cell, and it’s daughter cells, are
Abstract
determined by an internal unobservable (hidden) cell state. Here, I propose a novel approach to model cellular replication and differentiation based on spectral decomposition, which exploits cell lineage and sparse phenotypic information. The result is a compact state model that identifies statistically significant cell states and cell state transitions.

At the third level of abstraction, I propose an abstract formalism to encode logic interactions between transcription factors. The inferred cell state diagram gives insights into the relationships between transcription factor activation and the ontogenetic space. The model makes two predictions: which transcription factor patterns should be active and when; and a list of causal links (logic rules) of how transcription factors interact, either directly or indirectly, with each other.

At the forth level of abstraction, the genetic code is implemented in Cortex3D, an agent based software, where the consistency and usability of the model is tested by building a laminated cortical column. The implementation has been integrated into a computational model for programming growing structures (G-code) developed by Frederic Zubler and Andreas Hauri.

Finally, I give several examples of the model predictions, which are supported by experimental evidence. To my knowledge, this is the first time an algorithmic description of cortical neurogenesis is generated directly from sparse biological data.
Zusammenfassung


Die Neuheit, die ich präsentiere, ist eine Hypothese-generierendes Verfahren, um die biologische Entwicklung zu studieren, und von Grosshirn Entwicklung im Besonderen. Ich berücksichtige drei wesentliche Aspekte: (i) die Inferenz von transkriptionelle Netzwerk-Modellen (in Form von algorithmischen Regeln) aus biologischen Teildaten; (ii) die Entwicklung einer formalen Beschreibung der genetischen Regeln im Einklang mit der Biologie, und (iii) die Simulation der genetischen Regulierung für die Erzeugung einer geschichteten Grosshirn-Säule in Cortex3D, ein agenten-basierten Java Software.

Biologische Entwicklung ist charakterisiert durch multipotente Vorläuferzellen, die mehrere Zyklen von symmetrischer und asymmetrischer Zellteilung durchlaufen, um sich schliesslich in definierte Zelltypen zu differenzieren. Ich beschreibe diesen Prozess auf vier verschiedenen Abstraktionsebene, von den biologischen Daten bis zur Simulation der genetischen Modelle der Entwicklung.


Zweite Abstraktionsebene: ich stelle den Zellstammablauf mit einem Markov Modell mit unbeobachtete Zelzustände dar, in dem die Merkmale einer Zelle, und deren
Zusammenfassung

Tochterzellen, durch eine internen, verborgenen, Zustand charakterisiert sind. Ich schlage einen neuen Ansatz basierend auf Spectralzerlegung vor, der Zelllinien und phänotypische Informationen nutzt, um ein kompaktes Zustandsmodell zu generieren, das statistisch signifikante Zellzustände und Zell Zustandsübergänge identifiziert.


Schlussendlich, gebe ich einige Beispiele von Modellvorhersagen, die durch experimentelle Beweise gestützt sind. Meines Wissens nach, ist diese das erste Mal, dass eine algorithmische Beschreibung der Neurogenese direkt aus biologischen Teildaten generiert wird.
Acknowledgements

This Thesis could have not be achieved without the guidance and support of my advisor, collaborators, friends, and family.

I would like to express my deepest gratitude to my advisor, Prof. Dr. Rodney J. Douglas, for excellent discussions, infinite patience, and providing me with the possibility to work on a very interesting, and challenging Research project. I’m also in dept to all members of the Institute of Neuroinformatics, of our Research team, and especially to Frederic Zubler and Andreas Hauri, for their useful inputs (and programming lessons!), and to Roman Bauer for collaboration and interesting discussions. A special thanks goes to Michael Pfeiffer, for his support and personal input in the SECO project.

In particular, I’m very grateful for having the opportunity to collaborate with Dr. Colette Dehay. She gave me the opportunity to approach Neuroscience from the experimental perspective, and I deeply enjoyed the interactions with her team members, Marion Betizeau foremost for our close collaboration.

I would also like to thank my parents, my sister, and my little brother. They were always supportive and encouraging.
# Contents

Abstract iii  
Zusammenfassung v  
Acknowledgments vii  
Curriculum Vitae ix  
List of Figures xvii  
Abbreviations xix  

## 1 Introduction to the Thesis 1  
1.1 Brief introduction to cortical organization 2  
1.2 Motivation – The logic behind cortical development 3  
1.3 Problem statement – Looking for cortical growth rules 4  
1.4 Approach to understand cortical development 6  
1.4.1 Information representation for developmental data 7  
1.4.2 A model of cortical development based on local rules 8  
1.4.3 Developmental programs are specified by the genetic code 10  
1.5 Contribution 11  
1.6 Outline of the Thesis 12  

## 2 Introduction to cortical development: experiments and models 15  
2.1 Introduction to cortical architectures and cortical development 16  
2.1.1 Principles of cortical organization 16  
2.1.2 Canonical microcircuits 18  
2.1.3 The cerebral cortex originates from different transient embryonic regions 19  
2.1.3.1 Preplate neurons 19  
2.1.3.2 Subplate neurons 20  
2.1.3.3 Excitatory neurons 20  
2.1.3.4 Glial cells 22  
2.1.3.5 Interneurons 22  
2.1.4 Genetic characterization of neural progenitors 23  
2.2 Models of development: from neurons to cortical architectures 24  
2.2.1 The protomap versus protocortex hypothesis 24  
2.2.2 Ontogenetic columns and the radial unit hypothesis 24
2.2.3 Evolution of cortical architectures ............................ 26
2.3 Experimental approaches for the study of cortical histogenesis .... 28
  2.3.1 Dissociated cultures of neural precursors .......................... 28
  2.3.2 Transplantation .................................................. 30
  2.3.3 Mitotic history of neuronal populations ............................ 31
  2.3.4 Retroviral clonal analysis ....................................... 31
  2.3.5 Lineage analysis .................................................. 32

3 Inferring cell lineages underlying cortical development \textbf{35}
  3.1 Introduction to models of cortical development .................... 36
    3.1.1 Two different approaches to model cortical development ....... 36
      3.1.1.1 Population models ........................................ 36
      3.1.1.2 Single-cell models ....................................... 40
  3.2 Objectives of this Chapter ...................................... 41
  3.3 Methods: Models of corticogenesis ............................... 42
    3.3.1 Mitotic history of neuronal populations ....................... 42
    3.3.2 Global fitting of cell cycle parameters ....................... 43
    3.3.3 Ordinary differential equations model of cell cycle progression . 44
    3.3.4 Distribution of cell division modes .......................... 45
    3.3.5 Lineage trees reconstruction by generative functions ........ 46
  3.4 Results: Reconstruction of cortical cell lineages .................. 47
    3.4.1 ODE model of cell cycle progression and cell differentiation .. 47
    3.4.2 Distribution of cell division modes .......................... 53
    3.4.3 Lineage trees reconstruction with probability generative functions . 55
  3.5 Discussion ......................................................... 59
    3.5.1 The final cortical cytoarchitecture is defined by mode of cell divisions . 59
    3.5.2 Why are lineage trees important? ............................. 60

4 Analysis of developmental pathways in murine cortical development \textbf{63}
  4.1 Introduction ....................................................... 64
  4.1.1 Development as a self-construction process ...................... 64
  4.1.2 Lineage trees are composed of cell division patterns .......... 65
  4.1.3 Models of lineage evolution .................................... 65
  4.1.4 Inference of transcriptional networks .......................... 66
  4.2 Objectives of this Chapter ...................................... 69
  4.3 Methods: Inference of Markov Branching Processes .................. 70
    4.3.1 Graphical representation of cell lineages ...................... 70
    4.3.2 Spectral decomposition ....................................... 70
    4.3.3 Multi-type Markov Branching Process .......................... 72
    4.3.4 Dimensionality reduction ..................................... 73
      4.3.4.1 Homogeneous Markov Model (HM model) ..................... 74
      4.3.4.2 Time-dependent Markov Model (TM model) .................. 74
      4.3.4.3 Non-homogeneous Markov Model (NM model) ............... 74
  4.4 Results: A state machine model of cortical development ............ 75
    4.4.1 The Cell Lineage Tree ........................................ 75
    4.4.2 The State Diagram ............................................ 75
    4.4.3 Spectral Decomposition ....................................... 77
4.4.3.1 *Label propagation* .......................... 77
4.4.3.2 *Discrete-time Markov branching process* .......... 79
4.4.3.3 *Dimensionality reduction* ........................ 80

4.4.4 Spectral Decomposition of cortical lineages .............. 81
4.4.4.1 *Cortical cell lineages reconstruction* ............. 81
4.4.4.2 *State diagrams of murine cortical neurogenesis* .... 82

4.4.5 Transcriptional networks ................................ 85
4.4.5.1 *Transcriptomic Atlas of mouse neocortical layers* .... 85
4.4.5.2 *A transcriptional model of cortical development* .... 85

4.5 Discussion ............................................. 88
4.5.1 A novel approach for cell lineage data analysis .................. 88
4.5.2 Hidden generative models ................................ 89
4.5.3 Computational models of cortical neurogenesis .......... 90

5 *A genetic language for developmental programs* 93
5.1 Introduction to gene regulatory networks ...................... 94
5.1.1 The biological principles of the genomic regulatory code .... 94
5.1.2 Computational aspects of gene regulatory networks .......... 96
5.1.2.1 *Cis-regulatory modules constitute the computational units* .... 96
5.1.2.2 *Coding properties of transcriptional networks* ............. 97
5.1.3 Structural properties of transcriptional networks ........ 99
5.1.3.1 *Modularity* ........................................ 99
5.1.3.2 *Motifs* ............................................. 100
5.1.3.3 *Path length and redundancy* .......................... 101
5.1.3.4 *Composability* ...................................... 101
5.1.4 Gene regulation and development .......................... 102
5.1.4.1 *What is the link between genes and development?* ....... 102
5.1.4.2 *Asymmetric distribution of cell fate determinants* ........ 103
5.1.4.3 *Epigenetic control of cell fate specification* ........... 103
5.1.5 The choice of the modeling formalism ..................... 104

5.2 Objectives of this Chapter ................................ 105

5.3 Methods: Mathematical models for transcriptional networks .. 106
5.3.1 Reaction kinetics of the binding of a single transcription factor ........................................ 106
5.3.2 The rate equation ..................................... 110
5.3.3 Gene Reaction Network model ............................ 112
5.3.3.1 *An algebraic representation of combinatorial transcription factor binding* .................. 112
5.3.3.2 *Gene regulation without order* .................... 114
5.3.3.3 *Gene regulation with order* ....................... 114
5.3.3.4 *Relationship with conventional logic theory* .......... 115
5.3.3.5 *Example of combinatorial transcriptional regulation* .... 119
5.3.3.6 *Asymmetric cell division* ............................. 120

5.4 Results: Genetic control of developmental processes ............ 122
5.4.1 Formal genetic language definition ........................ 122
5.4.2 Toward a genetic model of development: Gene Regulatory Motifs ........................................ 124
5.4.2.1 *A Gene Regulatory Motif: the Bistable Switch System* ........................................ 124
5.4.2.2 *A Gene Regulatory Motif: the Multistable Switch System* .... 127
A  Experimental data  169
   A.1  Cell cycle parameters of murine cortex in area 3 and 6  . . . . . . . . . . . 169
   A.2  Number of neurons in the adult rodent cerebral cortex  . . . . . . . . . . . 170

Bibliography  171
## List of Figures

1.1 Cartoon example of cell lineage tree .......................... 7  
1.2 Cartoon example of state diagram .......................... 10  
1.3 Cartoon example of a gene regulatory network ................. 11  
1.4 Thesis layout ........................................ 13  

2.1 The cytoarchitecture of the mouse cerebral cortex .............. 17  
2.2 Development of the mouse nervous system ..................... 19  
2.3 Schematic representation of cell divisions for the generation of spiny and smooth neuron subtypes in the murine cerebral cortex .......... 21  
2.4 Evolution of the cerebral cortex ............................ 27  
2.5 3 experimental approaches for the study of cerebral cortex development ........................................... 29  

3.1 Dynamics of cell division and differentiation in a cortical column ...... 37  
3.2 Proportions of types of divisions ................................ 39  
3.3 Mitotic history of neuronal populations ........................ 42  
3.4 Laminar probabilities of area 3 ................................. 48  
3.5 Laminar probabilities of area 6 ................................ 49  
3.6 Leaving fraction, cell cycle length and laminar probabilities in area 3 and 6 50  
3.7 ODE simulation ........................................ 51  
3.8 Proportions of types of divisions ................................ 54  
3.9 Probabilistic reconstruction of lineage trees .................... 55  
3.10 Lineage tree examples ...................................... 57  
3.11 Lineage tree distributions .................................. 58  

4.1 Examples of cell lineage trees and corresponding division rules model .... 67  
4.2 From Cell Lineage Trees to the State Diagram .................. 76  
4.3 Spectral decomposition: a toy model .......................... 78  
4.4 Spectral decomposition on state diagram examples ............... 79  
4.5 Quantification of spectral decomposition performance ........... 80  
4.6 Spectral decomposition: cortical area 3 and 6 .................. 81  
4.7 Comparison of spectral decomposition in cortical area 3 and 6 .... 82  
4.8 Spectral decomposition and cell distributions of cortical area 3 ........ 83  
4.9 Spectral decomposition and cell distributions of cortical area 6 ........ 84  
4.10 Reconstruction of transcriptional network: cortical area 3 and 6 ........ 86  

5.1 From DNA to proteins ....................................... 95  
5.2 Model of transcriptional control ................................ 97  
5.3 Kinetics of autocatalytic transcription factor binding ............. 108
5.4 The Hill function ........................................... 109
5.5 Transcriptional network .................................. 111
5.6 Computation of coefficient matrix $B$ .................... 116
5.7 Combinatorial transcription logic ......................... 118
5.8 Example of gene network topology ....................... 120
5.9 The symmetry of cell division ........................... 121
5.10 Parameter dependence of a 2-dimensional genetic switch .................. 126
5.11 Dynamics of a 2-dimensional genetic switch with input ....................... 127
5.12 Dynamics of a 3-dimensional genetic switch with input ....................... 128
5.13 Continuous and discrete network states .................. 129
5.14 Division angle and cell division type ................... 130
5.15 Cell division patterns .................................. 133
5.16 Example of an attractor dynamics system .............. 136

6.1 Waddington’s developmental landscape .................. 144
6.2 Kaufmann’s developmental landscape .................... 145
6.3 Attractor landscape .................................... 151
6.4 Simulation of the dynamics of a bistable genetic switch ................. 153
6.5 Noise effects on bistable switch genetic landscape ............... 154
6.6 Gene regulatory network for area 3, 4 and 6 in the murine cerebral cortex 156
6.7 Simulation of mouse cortical lamination for area 3, 4 and 6 .............. 158

7.1 Probabilistic reconstruction of lineage trees ............. 164
7.2 Spectral decomposition: cortical area 3 and 6 ............... 165
# Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Asymmetric Division</td>
</tr>
<tr>
<td>CLT</td>
<td>Cell Lineage Tree</td>
</tr>
<tr>
<td>CP</td>
<td>Cortical Plate</td>
</tr>
<tr>
<td>CRs</td>
<td>Cajal-Retzius Cells</td>
</tr>
<tr>
<td>INM</td>
<td>Interkinetic Nuclear Movement</td>
</tr>
<tr>
<td>IPC</td>
<td>Intermediate Zone</td>
</tr>
<tr>
<td>IZ</td>
<td>Intermediate Zone</td>
</tr>
<tr>
<td>GRP</td>
<td>Glial-Restricted Precursors</td>
</tr>
<tr>
<td>GRN</td>
<td>Gene Regulatory Network</td>
</tr>
<tr>
<td>MZ</td>
<td>Marginal Zone</td>
</tr>
<tr>
<td>NE</td>
<td>Neuroepithelial Cell</td>
</tr>
<tr>
<td>RGC</td>
<td>Radial Glial Cell</td>
</tr>
<tr>
<td>S.D.</td>
<td>Standard Deviation</td>
</tr>
<tr>
<td>SLT</td>
<td>State Lineage Tree</td>
</tr>
<tr>
<td>SNP</td>
<td>Short Neural Precursors</td>
</tr>
<tr>
<td>SP</td>
<td>Subplate</td>
</tr>
<tr>
<td>SVZ</td>
<td>Subventricular Zone</td>
</tr>
<tr>
<td>S/P</td>
<td>Symmetric Proliferative Division</td>
</tr>
<tr>
<td>S/T</td>
<td>Symmetric Terminal Division</td>
</tr>
<tr>
<td>ODE</td>
<td>Ordinary Differential Equations</td>
</tr>
<tr>
<td>TF</td>
<td>Transcription Factor</td>
</tr>
<tr>
<td>VZ</td>
<td>Ventricular Zone</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction to the Thesis

Biological systems display remarkable structural and organizational characteristics such as repetitive topological motifs, hierarchies and loops, and important properties such as robustness, flexibility and evolvability. The overall organization is complex despite the presence of modular patterns, and has withstand many attempts of concise description. Such complexity is not the result of the accidental combination of elements, but the consequence of growth and continuous local interactions between its molecular components.

An important step to unravel the design principles of biological entities consists in identifying the relationship that exists between structural complexity and its unfolding during the execution of developmental programs. Here I focus on a remarkable example of self-constructing systems, namely the neocortex in the mouse. The analysis of cortical development is complicated by the fact that early cortical structures are continuously overridden by more recent ones, and advanced experimental as well as data analysis techniques are required.

After a brief introduction on structural elements of the cerebral cortex (Section 1.1) and the role of development (Section 1.2), I define the research problem that will be addressed in this Thesis: the search for primitive cortical growth rules to describe neurogenesis (Section 1.3). This problem is tackled with a novel combination of experimental data collection, modeling, and simulation (Section 1.4 and Section 1.5), for which I provide an intuitive explanation with some cartoon examples.

Finally, I describe the Thesis layout (Section 1.6).
1.1 Brief introduction to cortical organization

The neocortex is the center of extraordinary cognitive abilities and, according to Pasko Rakic, the "crowning achievement of evolution" [1]. It is composed of thousands of neurons, about $4 \cdot 10^6$ in the mouse and $20 \cdot 10^9$ in the human cortex [2, 3, 4], assembled in a highly organized cytoarchitecture. This crystalline-like arrangement of cortical neurons was firstly brought to light at the beginning of the 20th century by Santiago Ramón y Cajal with the application to the nervous tissue of the silver impregnation method ("la reazione nera"), developed by Camillo Golgi [5]. It was finally possible to look at the organization of the central nervous system in terms of functional assemblies of different neural cells, linked to one another by intricate webs of synaptic connections. The discovery opened the doors for the exploration of a new universe of unbelievable complexity and beauty, and made possible the investigation of the molecular basis of the mind.

The most striking emerging property of cortical organization in mammals is the subdivision into six layers, each one easily distinguishable histologically based on the morphology and density of the constitutive neural cells. Across the cortical tangential dimension the thickness of the layers varies and defines specialized regions called areas, which deal with sensory, motor, or association operations. Irrespective of the tangential location, the cytoarchitecture is composed of about 80% excitatory neurons with ramified axonal projections, and 20% inhibitory neurons with local arborizations.

The cellular regionalization of the cortex is not clearly reflected at the level of neural connectivity, as was revealed in the 80’s through bulk injections of tracers like horseradish peroxidase, which allowed to follow individual cell projections across several tissue sections [6]. Although at the scale of few cells a sparse number of excitatory and inhibitory neurons seem to be frequently arranged into predefined local networks [7], it is unclear how the connectivity is affected by increasing scale and how it is related to different functional maps.

Despite their elusivity to the experimenters, the concepts of a recurring organizing motif across the entire cerebral cortex, and of the cortical regionalization into distinct functional areas, have always appealed neurobiologists. A cerebral cortex build by superimposition of recurring modules was proposed very early in the history of neuroscience and dates back to Cajal’s extraordinary experiments. Not surprisingly, the simple idea
that the cortex consists of variations of a fundamental circuit has received considerable attention also in the last century. Despite huge efforts and evidences in its favor [6, 7, 8, 9], the existence of such a fundamental circuit is still debated. A tentative step toward an alternative approach to tackle the relationship between structure and function represents the driving motivation of this Thesis.

1.2 Motivation – The logic behind cortical development

Conventional engineering is about design of systems that are completely defined in a limited range of operating conditions. The system’s blueprint is dawned in a top-down fashion by an external architect, and there is a clear distinction between construction phase and functional phase. Unlike direct design, biological entities self-construct through the processes of development: structural and functional order result collectively from energy-dissipative processes involving cell replication and local interactions among cells or molecular components embedded in the physical environment [10].

Even if we were given the complete detailed anatomical structure and connectivity map of the brain - what architects would refer to as the \textit{Bauplan} (blueprint) - we would be still far from understanding its basic structural and functional principles. The construction process is completely different from conventional manufacturing processes, and the essence of the cytoarchitecture can be hardly captured by the layout of its components. The quest for a fundamental cortical circuit, if such a concept really exists, has to take into consideration the unfolding of the development process.

A glimpse on the role of developmental programs is given by comparative studies of cortical architectures across different species. The cortical organization of the dorsal pallium in mammals is an evolutionary innovation. Indeed, during mammalian evolution the neocortex has undergone dramatic modifications [11]: increase in size and surface area, folding of the cortical layers in carnivores and primates, increase in neuron numbers, and generation of new specialized neural cells. In contrast to those remarkable evolutionary changes, many molecular mechanisms that regulate neurogenesis are shared among many different species.

This remark motivates the following question: to what extent are cortical architectures prespecified in the neuroepithelium, that is to which extent are cortical structures the
result of hard coded developmental sequences? And how does such code specify different cytoarchitectonic features in distinct cortical areas and distinct species?

The complexity of the final cortical circuitry has fueled the controversy between the belief of the existence of developmental prespecified cortical architectures [12] and the opponent view of the cerebral cortex as a tabula rasa” to be shaped by experience only [13, 14]. More and more evidence supports the importance of intrinsic developmental programs: cortical networks are built by the generation and positioning of distinct neural cells and the establishment of the wiring diagram [1]. The entire process relies on the striking coordination of tissue growth, maturation and patterning along the anterior-posterior and dorso-ventral axis of the developing nervous system. Undoubtedly, common developmental mechanisms used in all brain structures are involved in the production and distribution of precise numbers and types of neuronal cells: among them are the tight control over number of cell-cycles, establishment of morphogenetic gradients of signaling molecules for local position information, migration and axonal/dendritic branching cues.

1.3 Problem statement – Looking for cortical growth rules

**Hypothesis H0**: Even though the role of development in shaping the final cytoarchitecture is still elusive, an attractive hypothesis states that the emerging cortical structure results from the regulation of conserved genetic developmental programs (rules), which regulate neural cell generation and their behavior [15]. Since it is difficult to address this hypothesis directly, I prefer to tackle three related hypothesis:

**H1**: Sparse cortical development data contains hidden structures that can be exploited to infer an hidden, underlying generative mechanism, through the use of advanced data-mining algorithms.

**H2**: Cell lineage data can be cast into a compact, probabilistic Markov Model composed of states, and state transitions, where each state represent a specific cell phenotype and it’s behavior. The model can be seen as a series of local rules, i.e. actions that can be taken by cells depending on their internal states and the state of their neighbors.
H3: State transitions can be described by inferred profiles of gene expression, which provide an abstraction of the genetic code, and can thus give an insight on the underlying gene regulatory mechanisms.

In this Thesis I investigate the logic of the mechanisms behind the regulation of corticogenesis of the mouse cerebral cortex from an abstract, computational point of view. More precisely, I address the question of which sequences of rules govern the generation of a murine laminated cortical volume in different areas. The rules that we chose to describe cortical development with, consist of abstract descriptions of cellular events, for instance expression of molecular machinery, cell division, cell-cell interactions, etc. Each molecule or cell is characterized by a current state, and rules define the conditions under which the state can be changed into a new one. Depending on the modeling formalism I will use set of rules with different levels of abstraction (systemic, cellular, or molecular level), and will show how they can be translated into another.

In more detail I tackle several issues: the collection and representation of information on cortical development of a laminated mouse cortical volume (Section 1.4.1), the inference of missing information and the generation of models of cortical development composed of local rules (Section 1.4.2), and finally the investigation of the relationship to genetic mechanisms (Section 1.4.3). Each issue is described in detail, together with the solutions I propose, in the following sections.

I chose to analyze developmental processes in the particular case of neocortical development for several reasons. Firstly, the elegant establishment of identifiable neural types in a highly organized structure makes the neocortex a convenient model to address the question of how such an arrangement of neurons is orchestrated. Secondly, the cortex is not merely a structure, but has amazing computational abilities. This makes the understanding of its development even more interesting, especially from the point of view of self-constructing computational systems. Here I focus on the mouse cerebral cortex, being the most well studied model organism and whose cortex exhibits many features present in all mammals [16].
1.4 Approach to understand cortical development

In the present work I address the question of how a sequence of abstract rules can control the generation of a laminated mouse cortical column. The major challenge is to understand how progenitor cells discriminate between alternative cell fates. I use the following three different levels of abstraction to go from a global view of cortical development to a local interpretation, either abstract or inspired by genetic regulatory mechanisms:

- **Global data representation.** The global behavior of neural progenitor cells can be represented by capturing the sequence of divisions and differentiation events that lead to the generation of distinct mature neurons. Taking the viewpoint of an external global observer, progenitor’s behavior can be modeled by ordinary differential equations that take in account average rates of proliferation and differentiation.

- **Model of cortical development with local rules (Hypothesis: H1, H2).** Cortical development is an example of how emergent behavior results from the collective interactions of the individual components and their local environment. Local rules describe actions that can be taken by a cell, namely divide, differentiate, move, and die. The design of rule sequences given the global behavior of the system is non-trivial, and is the main problem I address in this Thesis.

- **Simulation of local genetic regulation (Hypothesis: H3).** Cortical developmental data is mainly in the form of cell division patterns and expression of specific cell features, usually at the genetic level. Gene regulation plays a central role in unraveling developmental programs. Implementation of genetic growth and differentiation rules in a simulation enables to verify the models and investigate the role played by physical constraints and geometrical distances.

Those three points are explained in more detail in the sections below.
Chapter 1. *Introduction to the Thesis*

1.4.1 Information representation for developmental data

Traditionally, developmental genetics has focused on the study of expression of genes that have layer and neuronal subtype specificity within the cortex. This approach has enabled investigators to qualitatively classify different progenitor subpopulations at different time points during development. Despite the numerous studies, there is still limited understanding of the logic behind the generation of different neuronal subtypes. At a closer investigation the situation is even more subtle: many observed features are highly variable among cells and may represent a continuum or some other hidden features that are not directly visible by direct experimental observation.

Insight into the mechanisms underlying corticogenesis is provided by two types of data: gene expression, morphological or behavioral features of single cells, and the relationship between mother and daughter cells. A comprehensive data structure is the cortical *cell lineage tree* (Figure 1.1), which is the genealogical tree representing the division history of every precursor cell, and is associated with features belonging to each cell. The lineage tree contains a wealth of information about the correlation between different cells, their function, their anatomical position, and the developmental process itself.

**Figure 1.1. Cartoon example of cell lineage tree.** (A) Black and white version of Pax6/Tbr2/ToPro immunostained coronal section in area 3 of E15 mouse embryo. Progenitor cells divide and differentiate in the germinal layers (GL), and from there immature neurons migrate to the cortical plate (CP). The white circle is a randomly selected progenitor cell, for which we provide on the right (B) the graphical representation of a possible cell lineage tree. Nodes represent cells with 2 features, gene red and gene blue (observed only at terminal cells), and arrows represent the cell’s genealogical relationships.
The capability to directly observe cell divisions in-vivo is quite limited in mammals – because of the opaque body and the tremendous amount of cells produced. On the other hand in-vitro slice preparations can be usually maintained in culture only for a limited period of time. Moreover, advances in this research area are severely hampered by the scarcity of cell markers and the impossibility to visualize the dynamics of more than few proteins in a single time-lapse microscopy experiment.

However, it is possible to measure cell lineage data that contains sparse information labels, corresponding to different features such as gene expression data, cell morphology or other attributes. I use here a dataset of cortical lineages from the mouse somatosensory and motor cortex. The lineages were not observed directly due to technical limitations, but reconstructed using a probability generation function computed on experimentally measured numbers of dividing and differentiating cells [17, 18]. Transcription factor expression was obtained from a openly available transcriptomic atlas of mouse neocortical layers [19]. The integration of information across the genealogical tree, even if only partially reconstructed, is a valuable representation that can be used to infer hidden variables and models.

1.4.2 A model of cortical development based on local rules

Cell lineages are a valuable dataset since they contain information not only about single cells, but also the spatial and genealogical relationship among them. Developmental programs are reflected in the structure of the cell lineage tree: the tree topology describes the relationship between all cells that existed at given time point during development and the fate of their progeny. The structure of lineage trees, and especially recurring patterns of cell division and differentiation, could be used to infer basic developmental principles.

However, the sparseness of cell lineage information, due to non trivial technical restraints, is a limiting factor to the interpretation of the cellular processes that regulate corticogenesis as a whole system. Advanced computational approaches are needed to explore, model, and interpret these particular data structures. Networks or graphs conveniently capture relationships of cell features over lineage trees. Clusters of highly related cells in terms of behavior or cell fate are an essential hallmark of the graph structure and could be used to discover a meaningful representation of the data.
I assume that cell lineages and cell features can be described in the form of a Markov Model, according to which each cell, with it’s characteristic features, can be completely described by a hidden state. The evolution of cell states is defined by the cell current state, which comprises the cell internal state and it’s immediate surroundings.

In the worst scenario, state, behavior and spatial position for every single cell and cell division would have to be explicitly specified in the genome. This is quite unrealistic given the costs that a huge genome would cause, and is in favor of a more sparse genetic code. A second argument that argues for coding sparseness, is the observation that most of the high dimensional data on development exhibits dense aggregations at a low dimensional space. This is certainly true for cell lineages, where only a very small set of all possible internal genetic expression profiles are visited by cells during development. The necessity for coding sparseness has a strong consequence: although every cell in the organism is unique, it is reasonable to consider that cells with similar genetic signatures and similar differentiative behavior are regulated by a common molecular mechanism [20, 21], and are thus represented by a unique hidden state.

Clustering over cell features and genealogical relationships is used to identify recurring cell division patterns. The transitions between identified clusters can be cast into a graph depicting the transition probabilities between hidden cell states, that is the state diagram (Figure 1.2). State diagrams are defined as computational models composed of set of states and rules, according to which biological entities such as cells behave by state transformation and are ideal for the qualitative understanding of cortical development. While cells themselves do not have explicit knowledge about their hidden state (it is represented by their characteristic phenotypic expression), state models are very useful to capture the complex computation that occurs during development.

The problem of finding an optimal mapping between local rules and cell behaviors suggests the analogy with lossy compression in information theory. The mapping quality (or performance) is a trade-off between code length (the amount of instructions needed to specify all the required functions) and the expression of the code (the ability to build a structure with given characteristics). Developmental programs make large use of local information, for instance by the generation of a labeled internal environment. This strategy drastically reduces the number of rules needed to specify a developmental sequence, since rules can be reused in different contexts and times.
Chapter 1. *Introduction to the Thesis*

1.4.3 Developmental programs are specified by the genetic code

Finally, I consider the expression of computational models at the level of gene expression. An operational description of cortical architectures, more interesting than their cell lineage schematics or the sequence of hidden cell states, is the genetic code embedded in the genome, which represents the algorithmic description of how the neocortex is built and constrains the shapes it can assume.

The *genotype* of a cell is merely its genetic code, the sequence of base pairs that code for specific instructions. Genes are responsible for the execution of the code by regulating the expression of the molecular machinery composed of interacting proteins (this regulatory networks are referred to as *gene regulatory networks* or *transcriptional networks*). The *phenotype* is the display of all cellular characteristics that results from the execution of the genetic code.

Unfortunately, the genetic description is not easily understandable in terms of the final structure and its function, since several layers of transformations separate it from the cytoarchitecture it describes. In order to address these issues, I investigate the relationship between gene regulatory networks and the detailed sequence of possible cell fate
1.5 Contribution

Decades of biological focused studies have provided a considerable advance in the understanding of the processes regulating cortical development. This studies have accumulated an impressive amount of detail biological information and have also been successful in elucidating some of the molecular mechanisms.

The novelty that I wish to present in the present Thesis, is an approach based on automatic model generation and simulations to study cortical development from a new perspective: I shift the analysis of cortical development from a global systemic level toward the underlying mechanisms through the identification of recurring developmental patterns. The approach is motivated by the interest in studying an outstanding biological example of self-constructing system and the rules that govern its behavior, namely...
genetic rules. Although the idea of self-constructing organisms or tissues has a long history in the literature, few attempts have been made to directly link the biological data to more theoretical self-construction principles. To my knowledge this is the first time an algorithmic description of cortical neurogenesis have been generated directly from biological data.

1.6 Outline of the Thesis

In this Thesis I demonstrate how partial information about neurons, progenitor cells, their observed features and the genealogical relationships, can be used to build generative models of neurogenesis based on local rules only. Moreover, I investigate how those local rules can be mapped to genetic regulation. The Thesis is organized as following (Figure 1.4):

- **Chapter 1 – Introduction.** Introduces the research problem, with a brief description of the different questions I tackle, and an explanation of the novelty of the approach.

- **Chapter 2 – Strategies for studying cortical development.** Provides detailed literature review on cortical development, and available experimental techniques to understand corticogenesis.

- **Chapter 3 – Inferring cell lineages underlying cortical development.** Reviews models of neurogenesis and presents the experimental data on dividing populations of precursor cells, and how I reconstruct cell lineages in the murine sensomotor and motor cortex.

- **Chapter 4 – Analysis of developmental motifs.** Propose a novel algorithm that can be used to uncover lower dimensional models for cell lineages generation. The interpretation of the clustering provide also an approach to infer the molecular mechanisms underlying cell differentiation directly from biological data.

- **Chapter 5 – A genetic language for the simulation of cerebral cortex development.** Explores a genetic language to describe concisely molecular regulatory events and how to apply it to the control of developmental sequences.
Figure 1.4. **Thesis layout.** Chapter 1 introduces the themes addressed in this Thesis. Chapter 2 gives an introduction in the biology of cortical development. Chapters 3, 4, and 5 lay out the inference of genetic regulatory rules from collection of cell lineage data, inference of models for developmental processes. Chapter 6 presents the results of the simulation in a agent-base environment. Chapter 7 contains a summary of the main results, and discussion for future research.

- **Chapter 6 – Simulation of cortical lamination in CX3D.** Provides the implementation of the code for the generation of a laminated murine cortical column into a physical interactive environment, namely Cortex3D, in which the models are tested.

- **Chapter 7 – Conclusions.** Presents a summary of the main results, conclusions, and discussion of open problems.

Each Chapter, except Chapter 2 and Chapter 7, is structured as following: (i.0) a brief introductory text that describes the content layout, (i.1) a detailed introduction to the Chapter’s topics, (i.2) a statement of the objectives, (i.3) a detailed description of the methods, (i.4) a report on the most significant results, and (i.5) a discussion section.
Chapter 2

Introduction to cortical development: experiments and models

Understanding cortical development in terms of molecular mechanisms requires a deep exploration of the different processes involved. In this Chapter I lay out some of the knowledge and speculations that results from more than a century of research in the field, with a focus on cell fate decisions.

I start with a detailed introduction on principles of mouse cortical architectures (Section 2.1.1, Section 2.1.2), the role of cortical development in the production of different types of neurons (Section 2.1.3), and genetic specification of cortical precursors (Section 2.1.4). These observations where achieved by cellular, molecular biology, and genetic manipulation approaches (Section 2.3).

Observation of cortical architectures has lead to qualitative models of corticogenesis (Section 2.2 and Section 2.3), among them the influential radial unit hypothesis by Pasko Rakic. However powerful, these models endow a more descriptive view of cortical development, since based at the level of cell populations rather than at the single cell behavior. Clearly, a more profound grasp on the genetic mechanisms that control corticogenesis are necessary to unveil the logic behind cortical developmental programs.

This Chapter provides the knowledge foundation on which this Thesis is build. Literature review on specific topics is included directly in the introduction of each Chapter.
2.1 Introduction to cortical architectures and cortical development

2.1.1 Principles of cortical organization

The neocortex (Latin form for “new rind”) identifies the outermost part of the cerebrum, which is mainly composed of gray matter (cell bodies and unmyelinated axonal projections), and surrounds the deeper white matter. Sensory information coming from the peripheral nervous system is relayed by the thalamus to the cortex, which projects back into subcortical regions. Although the first primordial cortical structures can be found in amphibians and reptiles, the neocortex—as the Latin origin of the name suggests—is the most recent part of the cerebral cortex and a unique outstanding feature of mammals [11]. In primates, it is responsible for sensory perception, higher cognitive functions, generation of motor commands, spatial reasoning, and consciousness.

The adult mammalian neocortex is an highly organized structure composed of sheets of glial and neuronal cell types characterized by great diversity in morphology, molecular expression profile and function [5, 22]. Glial cells comprise astrocytes and oligodendrocytes, which have the function to support, nourish and maintain the cortical structure, whereas neurons are the basic computational units of the brain.

Despite their diversity, neurons and their precursors can be classified based on genetic, morphological and behavioral studies [23]. Cortical neurons are commonly subdivided into two broad classes according to their morphology: spiny neurons and smooth neurons [24]. About 80-85% of neurons are spiny excitatory, since they exhibit a high density of dendritic spines, form asymmetrical synapses and are glutamatergic. Excitatory neurons are mostly represented by pyramidal projection neurons with two axonal domains: a vertical domain extends from the soma to deeper cortical laminae or the contralateral cortex via the corpus callosum and to subcortical/subcerebral locations (e.g. thalamus, striatum, pontine nuclei, spinal chord); a horizontal domain projects into the cortical laminae. A minority of excitatory cells are star pyramidal neurons and spiny stellate cells, the latter having local dendritic and axonal arborizations. The remaining 15-20% (mouse) or 15-30% (human) [25] are smooth non-pyramidal neurons, which are inhibitory GABA-ergic, have short axons (interneurons), and form exclusively local connections. Spiny neurons account for only 15% of the total number of synapses because of their small axonal branches.
Distinct populations of neurons are found in different cortical regions and the dominant features of the cortical cytoarchitecture are undoubtedly its horizontal and vertical organization. Parallel to the pial surface the mammalian cortex is stratified in six layers, labeled I, II/III, IV, V and VI (with VI the innermost and I the outermost layer), each identified by neurons with specific density and morphology (Figure 2.1). Pyramidal neurons are present in high densities in layers II/III, V and the upper section of layer VI, whereas the lower section of layer IV and layer VI accommodate also non-pyramidal neurons. Across the tangential dimension the cortex is partitioned into areas, characterized by specific cytoarchitectonic features and long range intra-areal connections. Each region has distinct functional properties ranging from motor and sensory to cognitive.

An example of areal differences is given by structural dissimilarities between somatosensory area 3 and motor area 6. Going from frontal area 6 to parietal area 3, there is an increase in thickening and higher packing density of pyramidal neurons in layer IV. Not surprisingly, layer IV is the target region where a substantial percentage of thalamic afferents terminate, and is more developed in areas that receive visual sensory input.

This is a first example, as we shall see later on, on how cortical development is central for the diversification of cortical architectures. During cortical development, thalamocortical axons originate from the dorsal thalamus, extend ventrally along the surface of the ventral thalamus, and innervate the neocortex providing sensory information. Ingrowth of thalamic afferents is partially responsible for the generation of a more conspicuous layer IV, providing a mechanism by which external information can influence the local specification of the neural circuitry [26, 27].
2.1.2 Canonical microcircuits

The radial cytoarchitecture is remarkably repetitive across neocortical areas and suggests the presence of a common circuitry [9]. Evidence for this hypothesis comes from the presumed canonical organization of excitatory connections, first described in cat visual cortex [6, 8, 30, 31], in which local neural networks can be understood in terms of connections between relatively few types of excitatory and inhibitory neurons in a given spatial range [7, 32, 33]. Indeed, the pyramidal neuron projection patterns are highly layer-specific: by looking at the cortex across the radial dimension, we find bundles of thalamic afferents innervating layer IV, whose pyramidal neurons project to the superficial layers II/III. In turn axons from layers II/III terminate in layer V, and occasionally layer VI. Eventually layer V projections target several subcortical targets including the spinal cord, pons and superior colliculus, and neurons in layer VI project back to the thalamus.

From the functional perspective there is evidence that the mammalian neocortex (especially concerning superficial layers) is organized into interconnected columns, whose neurons form highly specific synaptic connections [34, 35, 36, 37, 38], display similar stimulus-response properties, and are believed to function as minimal processing centers of the cerebral cortex [9, 39]. Intriguingly, the functional architecture may emerge as early as the onset of cortical development. Progenitor cells divide and migrate into the cortical plate forming columns of clonally related neurons, albeit dispersed in the tangential dimension. Recent evidence suggests that sister neurons are more likely to display similar functional characteristics than unrelated cells, such as orientation preference in the visual cortex [40, 41].

Altogether, these findings point toward a developmental basis for the cortical columnar architecture, but how do these patterns emerge and how are various projection neuron subtypes generated during corticogenesis? The elegant disposition of identifiable neural types in laminae and their genetic characterization makes the neocortex a convenient model to address the question of how such an arrangement of neurons is orchestrated during corticogenesis.

In the following sections I provide a detailed description of the current knowledge on cortical development.
2.1.3 The cerebral cortex originates from different transient embryonic regions

Cortical neurons, identified by the expression of β-III tubulin, are produced by neural precursor cells during late embryogenesis until shortly before birth time; in the mouse from embryonic day E11 to about E18.5. Remarkably, cortical neurons are not born within the cortex itself, but originate from different precursor cells in transient embryonic regions located on the neural plate, which is folded to form the neural tube.

Shortly before neurogenesis the neural tube sets the anterior-posterior axis of the body, and its anterior region (forebrain) expands into three vesicles, of which the most anterior forms the telencephalon (Figure 2.2). In mammals, the dorsal telencephalon (pallium), develops into the cerebral cortex (Figure 2.3), and the ventral telencephalon (subpallium) generates additional neural subtypes together with the basal ganglia, a large group of neurons lying under the cortex.

2.1.3.1 Preplate neurons

The first postmitotic cells produced in the embryonic telencephalon are Cajal-Retzius (CRs) cells, GABAergic neurons and pioneer neurons, which from the preplate. CR cells migrate from specific regions at the pallium borders, namely the cortical hem, the ventral pallium and the pallial septum [43, 44, 45]. CR cells in the superficial layers play an essential role in the control of the neural migration into the emerging cortex through the extracellular expression of the glycoprotein Reelin [43, 46], establishing a Reelin morphogenic gradient that sets the direction of migration for later born cells.
2.1.3.2 **Subplate neurons**

Neurons migrating radially from the dorsal pallium and tangentially from the ventral pallium split the preplate into a subplate (SP), mostly formed by pioneer cells, and a marginal zone (MZ), where high densities of CR cells are found, and the cortical plate (CP) forms below. The SP is a transient layer of postmitotic cells between the CP and the intermediate zone (IZ), and is important to direct the migration of the first thalamic axons into the pallium [47]. Subplate cells may undergo cell death during the course of development and the remaining cells that survive into adulthood forms layer VIb [48]. In contrast the CP and the MZ are destined to become the six-layered structure characteristic of the mature cortex.

2.1.3.3 **Excitatory neurons**

Excitatory neurons derive from the murine germinal zones located in the dorsolateral (pallial) telencephalon, the ventricular zone (VZ) and the subventricular zone (SVZ) [49, 50]. Before the onset of neurogenesis, the neural plate is composed of neuroepithelial cells (NE) with radial processes extending from the apical membrane to the pial surface. NEs self-renew by symmetric divisions leading to a lateral expansion of the germinal layers [51], and additionally give rise to preplate neurons. At this stage the growth of the proliferative layers is exponential. As neurogenesis progresses, NEs differentiate into radial glial cells (RGCs) [52, 53, 54, 55, 56] retaining neuroepithelial properties such as apical-basal polarity. RGCs are responsible for the generation of the majority of excitatory neurons, with different classes born in overlapping temporal waves: earliest born neurons settle in lower cortical layers and later born neurons migrate past them toward more superficial layers in an inside-out sequence, although there is some degree of overlap in the birthdates of different neuronal subtypes [57].

RGCs divide at the apical surface of the VZ and undergo stereotypical patterns of cell divisions. Symmetric divisions amplify the pool of precursor cells [51], whereas asymmetric divisions [58, 59] (which become predominant) give rise directly and indirectly through intermediate progenitor cells (IPCs) to the majority of cortical neurons [56, 60, 61].

During neurogenic divisions RGCs maintain the pial processes, and postmitotic cells produced in the germinal compartments migrate radially along radial glia scaffold [54, 59] to the cortical plate in an 'inside-out' sequence, according to which later born neurons
Figure 2.3. Schematic representation of cell divisions for the generation of spiny and smooth neuron subtypes in the murine cerebral cortex. Simplified scheme depicting the different types of precursor cells and mode of divisions that lead to the generation of spiny pyramidal, stellate and smooth neurons. Excitatory neurons are generated by three types of precursors: radial glial cells (RGC), short neural precursors (SNP), and intermediate progenitor cells (IPC). RGCs and SNPs divide in the apical ventricular zone (VZ), whereas IPCs divide near the basal surface of the VZ and in the subventricular zone (SVZ). Newborn pyramidal neurons migrate radially into the cortical plate (CP). At the end of neurogenesis glial-restricted precursor (GRP) produce glial cells that invade the cortex. Different classes of projection neurons are born in overlapping temporal waves, where the earliest born neurons settle in lower cortical layers and later born neurons migrate past them toward more superficial layers in an inside-out sequence. Extracortical signals is provided by ingrowing embryonic thalamic axons and tangentially migrating interneurons, respectively Cajal-Retzius cells (CR) generated in the ventral telencephalon and cortical hem by different populations of interneuron precursors.

acquire an higher position in the cortical plate [62]. The glial scaffold is a prerequisite for the construction of the cerebral cortex, as it guides the generation of stacks of horizontal layers along the radial dimension. Curiously, at least some of the progenitors in the VZ display a vertical nuclear movement along the radial glial fibers in correlation with cell cycle progression. This characteristic movement of the nuclei, referred to as interkinetic nuclear movement (INM), is a distinctive hallmark of the developing VZ.

In contrast to RGCs, neural precursors with short radial morphology, also termed short neural precursors (SNPs) [63, 64], are localized in the VZ and retract their basal processes during division at the apical surface. Although they contribute to all cortical layers and neural glutamatergic morphologies, they predominantly generate neurons fated to layer IV [65].

At early developmental stages most progenitors, either RGCs or SNPs, divide at the ventricular surface. As development proceeds, mitosis at more basal locations increases: apical-dividing precursors give rise through asymmetric division to a second population
of precursors cells, referred to as basal or intermediate progenitor cells (IPCs) [58, 59, 66]. IPCs may divide in the basal VZ before migrating into the SVZ, where they lack apical-basal polarity and have a multipolar shape. IPCs are present in the developing cortex at all stages and seem to contribute to neurons in all cortical layers [67], although at mid and end-neurogenesis they have limited proliferative potential and divide symmetrically only once or twice before generating neurons of granular (IV) and supragranular layers (II/III) [58, 59, 66, 68].

### 2.1.3.4 Glial cells

RGCs are present until late neurogenesis before birth, when they differentiate further into glial-restricted precursors (GRPs) and produce, by rapid symmetrical divisions, a cohort of glial cells invading the laminated cortex [69, 70, 71, 72]. A small subpopulation of RGCs retain apical contact and continues to generate neurons and oligodendrocytes in the neonate, or converts into adult SVZ astrocytes, which differentiate into adult neural stem cells in the adult (reviewed in [73]).

### 2.1.3.5 Interneurons

The murine germinal zone produces only the excitatory spiny neurons of cortex. Interneurons are generated primarily in the ventral (subpallial) telencephalon, or medial and caudal ganglionic eminences [74, 75]. Parvalbumin and somatostatin interneurons distribute at higher densities in deep cortical layers and originate in the medial ganglionic eminences [76, 77, 78, 79]. In contrast, calretinin (CR) and vasoactive intestinal peptide interneurons originate in the caudal ganglionic eminences and cortical hem and distribute preferentially in the superficial layers II/III [74, 80, 81]. After birth, the majority of cells migrate tangentially into the subventricular/intermediate zone before turning upwards toward their final position in the various cortical laminae [82], whereas a smaller number of inhibitory cells migrate tangentially via a more dorsal route in the marginal zone and distributes homogeneously across the horizontal dimension of the cortical plate and layer specific in the radial dimensions [83].
2.1.4 Genetic characterization of neural progenitors

So far some of the basic mechanisms that control neuronal specification, migration, and connectivity within the neocortex have been identified [84, 85]. Development of different cortical areas involves a rich array of signals, which are integrated by the underlying transcriptional network.

Insights from extensive studies suggest that an early regionalization of precursor cells is established in response to the organizing centers located along the edges and midline of the neural plate and in the telencephalic vesicles [86, 87, 88, 89]. Organizing centers are sources of signaling molecules, among which we find fibroblast growth factors (FGFs), bone morphogenetic proteins (BMPs), Wnt proteins, Sonic Hedgehog (SHH), anti-Wnts, or retinoic acid (RA). Signaling molecules often induce patterning by establishing molecular gradients, which are transformed into the graded expression of transcription factors [90, 91, 92, 93].

Corticogenesis is indeed controlled to a large degree by transcription factors (TFs), such as Lhx2, Emx2, Pax6, Tbr2 [94], Tlx [95], Otx1 [96], Fex1 [97], Cux1 and Cux2 [98, 99, 100], Satb2 [101], Nex [68], Ngn1 and Ngn2 [102, 103], Id and Hes [104], as well as non-coding RNA Svet1 [105]. In the dorsal telencephalon, maintenance of self-replicating cortical progenitors is achieved through the expression of Id and Hes factors, whereas commitment to deep layer pyramidal neurogenic fate is linked with the activation of Ngn and NeuroD, as well as expression of Otx1 and Fex1, which have a crucial role in the specification of axonal projections. Other transcription factors, among which Cux2, Tbr2, Satb2, and Nex, are specifically overexpressed in the SVZ and upper cortical layers, suggesting the existence of distinct fate restricted progenitors [106].

The choice between alternative pyramidal cell fates is the result of transcriptional regulation [107, 108]: at early stages Pax6 is expressed preferentially in apical precursors and directly activates Ngn1 and Ngn2, which repress ventral fates (GABAergic) and promote early, deep layer specification. In a second phase, Pax6 promotes the specification of basal progenitors by induction of Ngn2 [58], AP2γ [109], Tbr2 [94]. While Pax6 and Tbr2 are characteristics of apical and basal progenitors respectively, Ngn2 and AP2γ specify a subset of apical progenitors with prospective basal fate.
2.2 Models of development: from neurons to cortical architectures

2.2.1 The protomap versus protocortex hypothesis

Several decades of research have led to two opposing hypotheses for how the cerebral cortex develops: on the one hand, the apparent homogeneity of cortical structures across its tangential dimension has provided the scientific basis of the *tabula rasa* or *protocortex theory* [13, 14], which assumes that all areal specifications arise from an initially homogeneous neuroepithelium molded by peripheral inputs only. On the other hand, the *protomap theory*, proposed by Rakic in 1988 [12], postulates that cortical progenitors are composed of a genetically prespecified population, which is modulated by external environmental influences. In their most extreme interpretations, these two theories incarnate the debate between genetic endowment versus environmental specification.

Although the protomap and protocortex hypothesis are often viewed as mutually exclusive, there is increasing evidence that they represent extreme expressions of a unique underlying process [90, 91], and that molecular specification of precursor cells plays a substantial role.

2.2.2 Ontogenetic columns and the radial unit hypothesis

Cohorts of cortical neurons that originate from a common neural progenitor cell (clonally related neurons) migrate radially into the cortical plate, and form structures referred to as *ontogenetic cortical columns* [12, 110], or *radial units*. Excitatory neurons follow glial fibers and align precisely into clonal radial stacks, commonly referred to as *ontogenetic cortical columns* [12, 110]. The final position of spiny neurons encodes the spatial information of their origin in the horizontal coordinates, and the time of birth in the vertical coordinate [12, 54, 111, 112]. The tangential dispersion of excitatory neurons into radially restricted clusters supports the idea of a columnar arrangement of the cortical infrastructure. In contrast, interneurons adopt the same cortical layer as pyramidal neurons born at the same time in the VZ [113], albeit they disperse uniformly across the entire tangential dimension of the cortex.

It is evident that the columnar arrangement of excitatory neurons is deeply linked to the way the cortex is constructed during the developmental phase: the reliable assembly of cortical columns relies on the tightly regulated spatial and temporal generation of
neuron subtypes, the deployment to the final location within specific layers \([15, 114]\), and the establishment of stereotypical connections \([12, 111, 115]\) (refined at subsequent stages by experience-dependent synaptic plasticity). The question is, which parts of this process are really under genetic control?

The number and diversity of neurons in each radial unit is controlled by several factors, including the initial number of founder cells, changes in the rate of neuron production, that is the control over neural birthtime \([17, 116, 117, 118]\), and the duration of the cell cycle. How these factors influence the final cortical cytoarchitecture is described by the radial unit hypothesis \([1, 111, 114]\): an increase in proliferative divisions at the onset of neurogenesis would lead to an increase in the number of founder cells, and consequently to the expansion of the cortical surface area; an increase in proliferative divisions at later stages influences the total number produced in each single cortical column, and determines the neocortical thickness\(^1\).

The radial unit hypothesis is an influential model that links the concepts of cell proliferation and ontogenic cortical columns into a comprehensive model, where the size of the neocortex is determined by the number of cortical columns. However, it relies on the assumption of an initially homogeneous population of precursor cells. Recent evidence suggests a more complex picture.

According to the classical model, the birthdate of a neuron determines its fate: early multipotent progenitors become progressively fate-restricted from producing all types of pyramidal neurons to superficial layers only \([120]\). In a game-changing scientific publication, Franco et al. \([106]\) reported the unequivocal presence of fate-restricted progenitors toward granular and supragranular layers (IV, II/III) as early as E10, resolving a 50 year old debate. In contrast to the prevailing model, their study shows that birth order is a result of molecular fate specification rather than vice versa.

These results are intriguing for several reasons. Firstly, they point to the central role played by intrinsic genetic programming of precursors cells in shaping the developmental process; and secondly, they suggest the presence of well-defined precursor subpopulations, that can be adaptively used to construct different cortical architectures (for instance in different areas). The genetic specification of defined cell populations has

\(^1\)During neurogenesis the number of apoptotic cells is relatively low, and although programmed cell death may still be used to control the number of progenitors and neurons \([119]\), there is so far no evidence that such mechanism should be responsible for shaping progenitor subpopulations and of the thickness of different laminae.
a direct impact on cortical evolution. While an ad-hoc construction process may be very difficult to evolve toward different morphologies, compact and modular genetic specifications are more plastic in terms of their potential outputs.

2.2.3 Evolution of cortical architectures

Despite the great variation in the self-organization processes in the neocortex among different species, some of the building patterns have been conserved during the course of evolution. While the neocortex is a feature unique to mammals, non-mammalian vertebrates display simpler layered cortical architectures from which the neocortex has been evolutionary derived: the primordial plexiform layer and the cortical plate (Figure 2.4). A primitive cortex is also present in mammals, for instance the hippocampus is composed of three subdivisions (CA1, CA2, CA3) derived from the VZ only [121]. The SVZ is a more recent acquisition, typical of neocortical architectures, and is dedicated to the production of superficial layers (IV, II/III) [105] on top of the existing columns.

The neocortical organization has undergone an extensive expansion during evolution and is even more prominent in primates [12, 111, 122]. The primate cortex results from a longer and more elaborated period of neurogenesis. Cortical progenitors undergo at least 28 cell division rounds [123], compared to an average of 11 in the mouse [124], and produce a larger proportion of late-born neurons as well as specific cell types unique to primates. As the radial unit hypothesis predicts, the contribution of the neocortical SVZ to neuronal production correlates with the expansion of superficial layers, cortical folding and the increasing complexity of the neocortex in terms of neuronal subtypes. Additionally, in the monkey cerebrum a considerable number of interneurons originate in the germinal zones of the dorsal telencephalon [125].

The substantial increase in neuron numbers, types and diversification of migratory pathways during cortical evolution paved the way for the establishment of new advantageous connections in primates. The radial unit hypothesis is a compelling theory that can explain how small evolutionary changes in developmental programs can lead to drastic changes in cortical architectures [111]. Conceivably, the evolutionary diversity results from modifications of the switch between proliferative symmetrical divisions and differentiative asymmetrical divisions in the precursor pool, and suggests that modifications
of cell proliferation have contributed to the evolutionary enlargement of the cerebral cortex [15].

Despite the evolutionary evidence for a relationship between ontogenetic columns and cortical organization, it is still unclear how the local microcircuit is built at the level of individual neurons to form a functional columnar architecture [115, 126]. Although the radial unit hypothesis makes a general prediction on the effect of proliferation at different stages of development, the contribution of individual progenitor subpopulations is still object of debate [51]. Moreover, models based on precursor subpopulations defined by few markers have limited prediction ability, since they do not take into account the entire lineage relationship between progenitor cells, and the relationship between genetic expression profiles along sequences of cell divisions. This type of understanding requires a change in the observer’s point of view from a global description to a deeper analysis of local cellular events and their relationship with the global behavior.
2.3 Experimental approaches for the study of cortical histogenesis

It was only at the end of the 19th century that the application of the silver impregnation method allowed Santiago Ramón y Cajal to expose the black silhouettes of nerve cells and their interconnections. From there on, a variety of techniques have been introduced to study cortical development down to the molecular level in different model organisms. The mouse, being an easily breed mammal, has been one of the most used experimental model for corticogenesis, although the fact that development occurs in the maternal uterus makes it hardly accessible to the experimenter.

Since neurogenesis results mainly from the proliferation and migration of cells, ideally methods would combine the ability to genetically target selected cell subpopulations in their natural environment, visualize global patterns and track their fate and the fate of their progeny. Traditionally, techniques rely on the analysis of cells of fixed tissue and in-vitro cultures. Two different approaches stand out: analysis of precursors within controlled environments (Section 2.3.1, Section 2.3.2), and the analysis in a partial native environment (Section 2.3.3, Section 2.3.4, Section 2.3.5).

The development of cortical architectures is not strictly explicitly coded in the genome. External influences and internal stochastic effects have a large influence on the growing structure. We distinguish between different information channels that converge to specify a final structure and can be experimentally tested (Figure 2.5): (1) intrinsic signals, defined by expressed transcription factors and epigenetic modifications of proteins and DNA, which are transmitted from one cell to the other through cell division; (2) the endogenous environment defined by the local extracellular signals generated by the self-construction process and the intracellular signaling machinery; (3) and finally the exogenous environment defined as influences from the external world. The role of the genetic specification is to define the reachable states of the system (the different subtypes of postmitotic cells), and to channel both stochastic and self-construction processes into a well defined parametric space through the expression of specific subsets of molecules.

2.3.1 Dissociated cultures of neural precursors

The simplest approach to study neural precursors is probably the use of dissociated cultures. Dissociated cortical progenitors from different embryonic stages are plated at
Figure 2.5. 3 experimental approaches for the study of cerebral cortex development. (A) Dissociated cultures of precursors plated in a controlled environment, where only intrinsic fate specifications and constant cues from the medium can influence the patterns in the cell lineages. (B) Slice culture of a coronal section. Cells are observed in a preserved environment, where both intrinsic cues and the local environment plays a role in shaping the cell division patterns. (C) In-vivo analysis by use of fluorescent or radioactive tracers injected directly in the developing embryos, where cells are surrounded by their native environment and signals from the external world is provided by innervation from thalamic afferent at early/middle-corticogenesis. Black arrows, inheritance of cell fates decision from mother to daughter cell (cell lineage), referred to as genetic specification; green arrows, influences from the local cell environment, referred to as endogenous environment; red arrows, influences from the external word, referred to as exogenous environment.

clonal density on adherent substrates and grown under controlled conditions [69, 70], such that cell-to-cell contacts are minimized and only extracellular cues contained in the medium can be responsible for the control over cell proliferation and cell fate decisions. This assay is particularly well suited to test genetically encoded differentiation capacities of cortical precursors, and additionally to test the role of different culturing conditions.

Rounds of cell divisions can be followed by time-lapse microscopy in slice cultures and the genetic identity of terminal cells is determined with cell-type specific markers. The lineage tree describes the relationships between progenitor cells and final expression of markers. The sequence of cell fates defines the mode of progenitor division. For instance, if the daughter cells have different fates (intermediate progenitor cell, glial progenitor, neuron, astrocyte, etc.) the division is identified as asymmetric, otherwise if no difference in the daughter cell behavior can be detected, it is considered to be symmetric.

After several days in culture, clones of dissociated neural precursors have been shown to be able to generate all diversity of neural and glial cell classes in the approximate correct sequence [127, 128]. The recorded lineage trees contained mostly asymmetric proliferative divisions, and asymmetric neurogenic divisions, while glia were produced later through a series of symmetric proliferative and symmetric terminal divisions.

These results not only indicate that the differentiation pathways are prespecified by
intrinsic cellular mechanisms, but most importantly they highlight a key concept of biological self-construction, namely the role of the local environment for proper morphogenesis, since the lack of an endogenous signal prevents progenitors from producing the correct numbers of neurons. An example of the importance of the local environment has been provided by an astonishing study, that showed the autonomous morphogenesis of the optic cup (retinal primordium) in a three-dimensional culture of mouse embryonic stem cells [129].

2.3.2 Transplantation

Cell fate determination in response to intrinsic and endogenous cues at different time-points can be assessed in more challenging transplantation experiments, in which progenitors cells from a donor at a specific embryonic stage are transplanted in the host tissue at a different stage. Testing the transplanted cells with specific molecular markers reveals the cell’s original developmental potential and how local cues can modify differentiation trajectories.

In a key experiment, transplantation of early progenitor cells into later developed stages has shown their ability to switch to the production of later-born neurons of layers II/III. [130]. In contrast late progenitors transplanted into earlier developmental stages, outside a small window of plasticity, have not been able to produce neurons of deep layers [131]. These results are strongly in favor of a progressive restriction in the differentiation potential of mitotic cells [132], albeit we cannot exclude that later progenitors failed to change their fate just because of absence of strong exogenous cues during early stages of development.

Other experimental approaches consist of the observation of progenitor cells in their native environment, either by in-vivo cell labeling (such as the mitotic history technique), or in-vitro tissue cultures (such as tracking of cell lineages). These techniques are not directly able to distinguish between cell intrinsic, endogenous and exogenous regulation, but have the advantage of capturing native cell behavior.
2.3.3 Mitotic history of neuronal populations

Despite their advantages, tissue culture assays do not entirely preserve the native environment for corticogenesis. The birthdating technique, also called mitotic history technique, is used to selectively label cohorts of precursors cells in the developing embryo and to track them in their native environment until their final position. The method involves injection of tritiated ($^3$H)-thymidine, which is incorporated in the DNA of precursors during replication of the genome (S phase). The cortical tissue is analyzed usually at adulthood for autoradiographic signal from stained postmitotic cells. The signal is maximal for cells that exit the cell cycle shortly after the injection (first generation labeling, FG), whereas the signal is diluted about to half at each successive division of the precursors that reenter the cell cycle (subsequent generation labeling, SG) [133].

The birthdate of a neuron population can be assessed by pulse injections of tritiated ($^3$H)-thymidine at consecutive times in development and is identified by heavily labeled cells. For a single injection, there is a relationship between the neuron’s birthtime and its radial position in the cortex: heavily labeled FG neurons have a radial distribution shifted toward lower parts of the cortical plate in comparison to weakly labeled SG neurons, which display a more dispersed distribution [17]. Variations in the FG/SG ratio reflect changes in the proportion of precursors differentiative divisions with respect to proliferative divisions, and indicate a change in the mode of cell division.

The birthdating technique has revealed some important principles of cortical development [17, 133]: firstly, the time of the last mitotic cell division is a strong predictor of the cell fate and spatial position that the newborn neurons will assume in the cortical plate; secondly, newborn neurons migrate long distances either radially or tangentially to reach the final resting place; finally, both fate specification and targeted migration result in the arrangement into distinct cortical laminae.

2.3.4 Retroviral clonal analysis

An alternative method is the labeling of selected precursor cells with retroviruses modified to carry a fluorescent reporter gene [134, 135, 136] that is stably expressed across the entire cell lineage (defined as all the cells that are derived from a single infected clone). This method has the advantage over the mitotic history technique to sparsely
label the progenitor pool, which allows a statistical separation between cells derived from the same clone.

Retroviral clone tracking experiments have showed that whereas earlier progenitors are able to give rise to neurons that span several layers of the cortex, later progenitors are more restricted to the production of more superficial layers [134, 135, 137, 138, 139, 140, 141]. However, these studies were complicated by the intermingling of labeled excitatory and inhibitory cells. Stronger evidence for clonal fate restriction was provided only later with the specific labeling of excitatory precursors [72].

2.3.5 Lineage analysis

A more challenging, but more precise technique is the lineage analysis, which is commonly used in combination with tissue slice cultures. Lineage analysis refers to the tracking of precursor cells and their progeny, the cell’s phenotypes at the level of morphological features or gene expression, and the ultimate resting position of postmitotic cells after their migration (for an extensive review on the reconstruction of embryonic development see [142]).

Cell lineage tracking methods consists mainly of sparse labeling of precursor cells by either intracellular injection of a tracer (fluorescent dye or horseradish peroxidase), electroporation (in-utero or ex-utero), viral transduction, and germ line genetic manipulation. Because multiple rounds of cell division causes dye dilution, cell lineage tracing experiments often require a more reliable marker, namely retroviruses that carry a fluorescent reporter tag [54, 59, 143]. Target progenitors and their progeny will express a unique marker that can be followed by time-lapse confocal imaging of slice cultures.

Recently, an array of promising biomarkers has been made available to study specific functional states of molecular elements, for instance determination of cell cycle stages (Fucci [144]), markers for regulators of cell polarity (Cdc42 or Rac bioprobes [145, 146]), fluorescent timers [147], photoconvertible fluorophores (Kaeda[148]), and many others.

Observation of lineages is also possible in-utero, where the embryos are immerse in their natural environment, thanks to the development of multiphoton microscopy. A promising technique is clonal barcoding [149], where labeled retroviral vectors with random sequence tags (barcodes) are integrated into the host genome at specific developmental stages through cell transplantation. The clonal progeny of each cell can be tracked in
time by performing single-cell sequencing at a later timepoint. The method has the advantages that enables the tracking of developmental processes \textit{in-vivo}, and can be combined with micro-array gene expression analysis of the final cells. In the last years, an even more advanced technology based on light-sheet fluorescence microscopy [150], as emerged as a promising approach for the reconstruction of embryonic development. Light-sheet microscopy is used to rapidly image living three-dimensional organisms at sub-cellular resolution. Although at its infancy, the ability to follow progenitors cells directly in the intact cortex promises great advances in the understanding of cortical development.

The partial tracking of cell lineages has enabled the dissection of sequences of cell division and cell fate commitments. The breakthrough was the identification of RGCs as the majority of proliferative cells in the VZ [54, 55], and of a second proliferative pool of IPCs in the SVZ [59]. The two subpopulations not only undergo different types of divisions, with RGCs dividing more asymmetrically and IPCs almost exclusively symmetrically, but also display radically different expression marker profiles [94]. Similar strategies have led to the discovery of new progenitor subpopulations in the progenitor pool in mouse [63, 151] and in monkey [151, 152, 153].

A recent publication in \textit{Science} [106] reports an unequivocal finding: the transcription factor \textit{Cux2}, expressed in a small subset of progenitor cells, causes a strong bias toward the almost exclusive generation of superficial layers. The existence of fate-committed progenitors, demonstrated for the first time, challenges the classical view that the birthdate of a neuron largely determines its fate, and underlines the importance of lineage specificity in developmental process. Indeed, the hypothesis of a sequential generation of waves of different neuronal types was the result of incomplete experimental observations. In reality, progenitor cells are preprogrammed to generate specific classes of pyramidal neurons independent of birthdate or niche. To reconcile this data with the evidence for multipotent progenitors early in cortical development [127, 128], the authors suggests that progenitor fate-commitment may be environmentally regulated, with progenitors becoming progressively restricted to selected part of the lineage.

Proper corticogenesis depends on sequences of symmetric and asymmetric divisions of precursor cells [58, 59, 154, 155]. Asymmetric divisions arise from combinations of different mechanisms: (1) unequal segregation of cell fate determinants between daughter cells; (2) integration of extrinsic and intrinsic signals, probably during G1 phase [15, 156],
before cell division; and (3) cell fate switches based on local environmental factors, such as Delta–Notch lateral inhibition [64]. Despite the importance of division symmetry across evolution [157, 158], still little is known about the mechanisms and the link with fate determination.
Chapter 3

Inferring cell lineages underlying cortical development

Cortical development is the process by which dividing progenitor cells acquire different fate specifications and eventually differentiate toward a final cell type (Section 3.1). The commitment to a cell phenotype is the result of intrinsic genetic programs and the interaction among neighbouring cells. The detailed sequence of cell fate specifications is given by the genealogical history of every precursor cell, that is the history of cell divisions from the very first cell (cell lineage).

In this Chapter I use published experimental data on rates of cell division and differentiation to investigate the effect of those parameters on the final cytoarchitecture in murine cortical somatosensory area 3 and motor area 6, which display significant structural and functional differences (Section 3.4.1).

While rate equations are useful to illustrate the dynamics of corticogenesis, the regulatory mechanism lies in the control over sequences of symmetrical and asymmetrical cell divisions (Section 3.4.2). Therefore I reconstruct cell lineage trees from the murine cortex by sampling over the distribution of different pyramidal neural types originating at different embryonic stages (Section 3.4.3). This information will be exploited in successive Chapters to develop a mechanistic model of neurogenesis, that is a model not based on the view of the external observer, but a model in which every cell locally controls its own behaviour.
3.1 Introduction to models of cortical development

3.1.1 Two different approaches to model cortical development

Two strategies have been used to model cortical development. On one hand, population based models describe the dynamics of cortical development in terms of cell populations (assumed to be homogeneous), and transition probabilities between different populations such as cell cycle exit or cell cycle progression rate across the entire duration of neurogenesis. On the other hand, single-cell based models relies on data about the mitotic and differentiative behavior of single cells, and are technically limited to the analysis of only a restricted series of cell cycle divisions. However, they provide valuable information such as the mode of cell division, which can be inferred only approximately in population based models. The aim of single-cell models is to provide a mechanistic description of the observed phenomena, rather than an external, global description.

3.1.1.1 Population models

Population based models describe neurogenesis in terms of differential equations and rates. The number of neurons produced in the germinal zones is determined by the rates of neuron production according to two parameters: the rate of cell-cycle progression and cell-cycle exit [17, 116, 117, 118]. The first raw estimates of the average number of cell cycles and the number of neurons produced (neuronal output) in the mouse cortex comes from radioactive labeling of precursors cells at different developmental stages [159]. The quantification of cell cycle length and neural differentiation rates consider the progenitor pool as a homogeneous population of progenitor cells, which leads to estimation errors as high as 10% [160]. Nonetheless, these parameters are quite useful for reasoning about the essential concepts that determine the outcome of development [124].

At each time step the number of neurons (differentiative pool $Q$) increases, and the amount of increase depends on three variables: the number of proliferating cells (proliferative pool $P$), the cell cycle length ($T_c$), and the proportion of daughter cells that exit the cell cycle at mitosis (leaving fraction $k_Q \in [0,1]$). Since cell death within the proliferative population is traceable, the percentage of cells re-entering the cell-cycle (proliferating fraction $k_P \in [0,1]$) must be equal to $1 - k_Q$. Those percentages change
Chapter 3. *Inferring cell lineages underlying cortical development* 37

Figure 3.1. **Dynamics of cell division and differentiation in a cortical column.** (A) A schematic diagram that illustrates the dynamics of the pool of proliferative cells $P$ (light gray box) and the pool of differentiated cells $Q$ (dark gray box) in a cortical column section. Arrows indicate possible transitions of cells between the two pools; the white sheet indicates the pial surface. At the onset of neurogenesis (E10), the rate of cell differentiation is almost zero and the progenitor pool is expanded until middle neurogenesis (E14-15). As development proceeds, the differentiation rate $k_Q$ increases, and more neurons are produced. At the same time the precursor pool is depleted, and at the end of neurogenesis (E18) only a thin layer of precursor cells is left. (Figure adapted from Takahashi et al., 1996 [124]). (B) Quantification of $k_Q$ in the mouse cerebral cortex from E11 to E18 according to Takahashi et al., 1996 [124], fitted to a polynomial function of second degree. (C) Numerical solution of differential equations describing the dynamics of proliferative cells and the cumulative number of cells (either progenitors or neurons) as function of the number of elapsed cell cycles according to Equation 3.1 and data from Takahashi et al., 1996 [124] (gray, $P$; black, $Q$).

dynamically with developmental time. The relationship for each timestep between proliferative and differentiative pools can be summarized by the following two-compartments model [124] (Figure 3.1A):

$$
\begin{align*}
P_{t+1} &= 2(1 - k_Q)P_t - k_QP_t \\
Q_{t+1} &= k_QP_t
\end{align*}
$$

(3.1)

The term $(1 - k_Q)$ is multiplied by a factor 2 since cells that re-enter the cell cycle will divide into two daughter cells. Depending on the values taken by $k_Q$, the system has different regimes: for $k_Q < 0.5$ the proliferative population grows exponentially and $Q$
increases without boundaries, whereas for $k_Q > 0.5$ the proliferative population shrinks until disappearing, and $Q$ will stop growing. $P$ is constant at $k_Q = 0.5$, since precisely one cell re-enters the proliferative cycle and one differentiates into a neuron. Different cortical architectures are produced depending on the change of $k_Q$ as a function of time (Figure 3.1B-C), since this parameter strongly influences the final number of neurons in each lamina.

From a mechanistic point of view differences in $k_Q$ are caused by the control over the asymmetry of cell fate decision at the moment of cell divisions, either by intrinsic signals or by external modulation. Asymmetrical divisions arise from unequal distribution of molecular components during mitosis and result in two daughter cells with different fates [161]. Mitotic division can be classified into three possible types depending on the mitotic behavior of the daughter cells: (1) symmetric proliferative division ($S/P$); (2) symmetric terminal division ($S/T$); (3) asymmetric division ($A$). Whereas symmetric proliferative divisions generate two progenitors that re-enter the cell-cycle and continue to divide in the germinal layers, neurogenic differentiative divisions result in at least one postmitotic daughter cell that migrates into the cortical plate to become an immature neuron, either by symmetric terminal or asymmetric division.

Under the assumption of independence between sister cells’ fates, after division each daughter cell has a probability to either re-enter the cell cycle or differentiate:

\[
Pr(cell_1) = k_P + k_Q = 1
\]
\[
Pr(cell_2) = k_P + k_Q = 1
\]  

(3.2)

The joint probability is thus given by the expansion of the binomial:

\[
Pr(cell_1, cell_2) = (k_P + k_Q)^2 = k_P^2 + 2k_Pk_Q + k_Q^2 = 1
\]  

(3.3)

$k_P^2$ represent the probability of symmetric proliferative divisions, whereas $k_Q^2$ represent the probability of symmetric terminal divisions, and the term $2k_Pk_Q$ represents two asymmetric divisions. Changes in the proportions of these three types of mitotic divisions are reflected in the values assumed by $k_P$ and $k_Q$ respectively, which have dramatic effects on the shape of the resulting cortical architecture [162]. However, the prediction
Chapter 3. Inferring cell lineages underlying cortical development

Figure 3.2. Proportions of types of divisions. (A-B-C) Possible distribution of cell divisions for \( k_Q = 0.5 \) and \( k_P = 0.5 \) (model 1, no asymmetric divisions; model 2, binomial distribution; model 3, maximal percent of asymmetric divisions). (D-E-F) Diagrams for the 3 models describing different changes in the proportions of symmetric proliferative (green, S/P), symmetric differentiative (violet, S/T) and asymmetric (blue, A) cell divisions as a function of changes in \( k_Q \). A similar analysis was proposed by Cai et al. [162].

... of division type proportions may be biased, since the assumption of daughter cell independence is certainly violated. For instance, for same values of \( k_Q \) and \( k_P \) different distributions of mode of division are conceivable: a \( k_Q = 0.5 \) may correspond to 100% asymmetric divisions, or to 50% symmetric proliferative and 50% symmetric terminal divisions (Figure 3.2).

The two-compartment model described in Equation 3.1 is an oversimplification: the pool of differentiated cells \( Q \) is composed of different subtypes of neurons, and conversely the proliferative pool \( P \) is an heterogeneous population of progenitor cells. The control of the cortical stratification in different areas, or different species, is due to the regulation of inhomogeneities in the proliferative population, an hypothesis that was advanced about 20 years ago [117, 163], and recently confirmed [106].

Therefore, a more accurate model would need to split the pool \( P \) and \( Q \) into many different pools corresponding to the identified cell types. However, even the decision to classify progenitors in subpopulations based on behavior or bio-markers, and the construction of models based on this proposed classifications, are based on arbitrary classifications of progenitors, and ignore the intrinsic variability among those subpopulations. For this, and other reasons, research on corticogenesis has shifted to the study of the behavior of single cells, and the development of single-cell based models.
3.1.1.2 Single-cell models

Current experimental methods for the study of cortical development are driven by the need to understand the molecular basis of cell proliferation and cell fate decisions at a single cell resolution. The emerging strategy is the tracking of cell lineages by mapping cohorts of progenitor cells in their native environment at high spatial and temporal resolution, ideally down to the molecular level. Cell lineage data describes the genealogical history of single cortical neurons and offers a powerful insight into patterns of cell division and differentiation, information that cannot be obtained with the traditional approach of considering progenitor cells composed of homogeneous subpopulations.

Single-cell based models are directly based on cell division mode probabilities, rather than proliferation and differentiation probabilities, that is on parameters that directly reflect the mechanistic control of cell fate rather than purely descriptive parameters. Cell division probabilities are defined as probabilities for pairs of progeny (the fate of two daughter cells is highly correlated, since they share the same mother cell and a proximal spatial position). This definition reflects directly the different types of mitosis, with $k_{PP}$ the probability of a S/P division, $k_{PQ}$ of an A division, and $k_{QQ}$ of a S/T division:

$$
Pr(\text{cell}_1, \text{cell}_2) = \frac{k_{PP}}{S/P} + \frac{2k_{PQ}}{A} + \frac{k_{QQ}}{S/T} = 1
$$

Approaches at the single cell level are quite limited in the amount of collected data and usually either concentrate on a very small developmental temporal window and on very few cell features. Acquisition of cell lineage data together with markers is not trivial, requires considerable experimental efforts, and results in a sparse dataset. Cell lineage analysis is greatly simplified in dissociated progenitor cultures, where sequences of mitosis are tracked by video microscopy and staining techniques. Such approaches have unveiled previously unrecognized asymmetries in progenitors division patterns and indicate that intrinsic molecular differences control cell fate decisions [106, 127, 128].

Given the sparse data, models of corticogenesis at the single-cell level are usually focused on very specific molecular mechanisms, such as the link between division plane orientation and cell fate distribution [164], or investigation of cell division sequences [59] to cite some, and there are only few attempts of building comprehensive quantitative models at a systemic level (see Chapter 4).
3.2 Objectives of this Chapter

Here I focus on two areas in the murine cortex - area 3 and 6 - that are characterized by a different proportions in their laminar structure. The comparison between the two areas, and the analysis of different cell subpopulations, allows the investigation of the mechanisms that modulate neurogenesis.

Goals for this Chapter are:

- **Build a population-based model to predict number and type of neurons produced from a single precursor cell.** Differential equations (ODE) are used to model the dynamics of neurogenesis from experimental data in the mouse and validate the data by comparing predicted number of cells with independent measurements. The ODE gives accurate prediction of the number and types of neurons produced starting from a single precursor at the onset of neurogenesis.

- **Investigation of the dynamics of neurogenesis in area 3 and 6.** Population-based model allows to investigate how different parameters describing cell proliferation, asymmetry of cell division, and differentiative stages are linked. I focus here on the effect of cell cycle parameters (probability of differentiation and cell cycle length) on the final cytoarchitecture.

- **Reconstruction of cell lineage trees with a probabilistic single-cell based model.** Although population rates are useful to describe the dynamics of cortical development, only analysis at the single cell level can reveal the intrinsic regulatory mechanisms. Towards this end, my first concern is the collection of cell lineage data of the murine neocortical development, which I directly address in the present Chapter. Due to the technical difficulties in collecting in-vivo lineage data, I reconstruct probabilistic lineage trees from data on the mitotic history of precursor cells. Analysis of cell division patterns will be treated in detail in Chapter 4.
Figure 3.3. Mitotic history of neuronal populations. (A) Architectonic map of mouse neocortex (dorsal view, adapted from Polleux et al., 1997 [17]). Full lines indicate the boundaries between architectonic fields (frontal, parietal, occipital, temporal, and retropial) whereas dashed lines indicate boundaries between cortical areas. The arrow indicates the location of examined sections. (B) Distribution of first generation (FG) neurons in the adult frontoparietal cortex after tritiated (\( ^3 \)H)-thymidine injections on embryonic day E15.5. Arrowheads indicate the positions of areas 6, 4, and 3 (adapted from Polleux et al., 1997 [17]). The labeled neurons where generated from the original labeled cohort of precursors cells after they have undergone several rounds of division and differentiation. (C) Schematic view of cell cycle of eukaryotic cells, divided into four successive phases: M-phase (mitosis), in which the nucleus and the cytoplasm divide; S-phase (DNA synthesis), in which the DNA in the nucleus is replicated, and two gap phases, G1 and G2. The G1-phase allows intra- and extracellular cues to induce either commitment to a further round of cell division or differentiation (G0).

3.3 Methods: Models of corticogenesis

3.3.1 Mitotic history of neuronal populations

To investigate regional differences in the cell-cycle kinetics of neural precursors, I use published data on in-utero pulse \( ^3 \)H-thymidine injections made throughout corticogenesis of areas 3 and 6 in the mouse [17, 18]. Single pulse injections of tritiated \( ^3 \)H-thymidine in the embryo label cohorts of precursors cells in S-phase, and make possible to track them in their native environment until their final position. The cortical tissue is analyzed usually at adulthood for autoradiographic signal from stained postmitotic cells. Cells that exit the mitotic cycle at the first division round are maximally labeled (first generation labeling, FG), whereas cells that undergo successive rounds of divisions have a diluted autoradiographic signal, proportional to the number of divisions (subsequent generation labeling, SG) [133] (Figure 3.3).

A cortical column is generated by the neurons that undergo their final mitosis in the germinal layers. Computation of the percentages of FG neurons with respect to the total population T [165] gives the cell cycle-exit probability of the labeled population, also referred to as the leaving fraction \( kQ \) (LF) [17, 18]. Because the maximum LF index
that can be measured in the cortex is 40% of the true value due to the labeling profile, it is first normalized to 100% [18]. The problem of (potential) differential cell death is overcome by dividing the cumulative FG measure by T, because larger amounts of cell death in one of the areas will influence similarly FG and T. The percentage of FG neurons $p_X$ in a specific layer $X$, or the laminar probability, is given by dividing the number of FG neurons in layer $X$ over total FG.

The labeling index $SG/T$ reflects the proportion of precursors in S-phase and equals $T_s/T_c$, where $T_s$ is the duration of S-phase and $T_c$ the total cell cycle length. Assuming that $T_s$ is constant and equals to 6 hrs throughout corticogenesis [116, 166], variation in the percentages of labeled precursors directly reflects changes in the rate of cell-cycle progression. This means that labeling indexes allow to calculate the theoretical cell cycle length $T_c$. The measured value corresponds to the average cell cycle length for any given timepoint, and differences in coexisting subpopulations of cells, e.g. ventricular and subventricular cells, are averaged out.

Variations in the FG/SG ratio reflect changes in the proportion of precursors differentiative divisions with respect to proliferative divisions and indicate a change in the mode of division. Whereas symmetric proliferative divisions generate two progenitors that re-enter the cell-cycle, differentiative divisions result in at least one daughter cell exiting the cell cycle and differentiate, either by symmetric neurogenic division or asymmetric division.

### 3.3.2 Global fitting of cell cycle parameters

The leaving fraction $k_Q$, the average cell cycle duration $T_c$, and laminar probabilities $p_X$ (Table A.1) in area 3 and 6 of the mouse cortex were experimentally measured starting from E11 to E18.5 by Polleaux and Moraillon [17]. The laminar probabilities are fitted with linear combinations of exponential functions in the form $1/[1+\exp^{at+b}]$. The global estimation of the parameters $a$ and $b$ is based on simultaneous fitting of two sigmoidal functions and thus guarantees that the laminar probabilities sum up to 1 at every time-point. The average cell cycle duration $T_c$, and laminar probabilities $p_X$, are fitted using the spline method for interpolation. Due to experimental noise, averaging across multiple individuals, and contamination from the interneuron population, I consider 10% experimental error for each data point.
Ordinary differential equations model of cell cycle progression

The dynamics of cellular division and differentiation is controlled by the length of the cell cycle and the proportion of differentiating cells. In order to prove the validity of the experimental data, I consider a simple model based on differential equations to simulate neurogenesis and analyze the behavior of different subpopulations of cells. In the simple case of an homogeneous population of progenitor cells characterized by exponential growth at rate \( k \), the number of precursors \( N \) is given by:

\[
\frac{dN}{dt} = kN(t) \quad \Rightarrow \quad N(t) = N_0 \exp^{kt} \tag{3.5}
\]

with \( N_0 \) the initial number of progenitors at time \( t = 0 \). After a constant time step of \( T_c \) the population doubles in size, and thus:

\[
\frac{N(T_c)}{N_0} = 2 = \exp^{kT_c} \quad \Rightarrow \quad k = \frac{\ln(2)}{T_c} \quad (3.6)
\]

Cortical development is slightly more complicated, since cells have to decide whether to re-enter the cell cycle, or terminally differentiate into neurons. I use a simple differential equation three-compartment model, already proposed by Polleux et al., 1997 [18]. In the germinal layers, I distinguish two compartments to simulate the progression of precursors through the cell cycle: the first compartment \( P_1(t) \) represents the pool of cycling precursors in phases G1, S, G2 or M progressing through the cell cycle at rate \( k = \frac{\ln(2)}{T_c} \); the second compartment \( P_0(t) \) represents precursors at the G1/S-phase checkpoint, when cells either differentiate and become postmitotic with probability \( k_Q \) or re-enter the cell cycle with probability \( k_P = 1 - k_Q \). The third compartment \( Q(t) \) is composed of postmitotic cells in phase G0 (Figure 3.3C). The ordinary differential equations (ODE) used to describe the system are:

\[
\begin{align*}
\frac{dP_1}{dt} &= (1 - k_Q)P_0 - \frac{\ln(2)}{T_c}P_1 \\
\frac{dP_0}{dt} &= 2\frac{\ln(2)}{T_c}P_1 - P_0 \\
\frac{dQ}{dt} &= k_QP_1
\end{align*} \tag{3.7}
\]

Notice that the derivation for \( k \) is valid only in the case \( T_c \) is a constant, and I will deal with a variable \( T_c(t) \) by applying a non-linear time-warping function to compensate for the lengthening of the cell cycle. The solution of the differential equations describe
the dynamics of a population of precursor cells during corticogenesis. \( P \) represents the proliferating cells. \( Q \) are differentiating cells and \( Q_X \) are differentiating cells specific for a particular layer \( X \) with probability \( k_{Q_X} \).

\[
P = P_0 + P_1 \\
Q_X = k_{Q_X} Q
\] (3.8)

3.3.4 Distribution of cell division modes

Knowledge about cell cycle proliferation and differentiation rates allows to make predictions on the underlying distributions of cell division modes, that is the proportion of symmetric and asymmetric distributions (where symmetry is defined in respect to the cell proliferative behavior). Since single values of \( k_Q \) constrain, but do not uniquely define the distributions, I conceive three different scenarios. The three models are used as null hypotheses to compare with measured proportions of mode of cell divisions.

**Model 1:** Fate of daughter cells is dependent and symmetrical divisions dominate the distribution. Asymmetrical divisions do not exist.

\[
Pr(cell_1, cell_2) = \frac{k_P^2 + k_Qk_P}{S/P} + \frac{0}{A} + \frac{k_Q^2 + k_Qk_P}{S/T} = 1
\] (3.9)

**Model 2:** Fate of daughter cells is independent, or distribution of symmetric versus asymmetric divisions follows a binomial distribution. In other words it is assumed that a distribution of all type of divisions is present at each cell cycle.

\[
Pr(cell_1, cell_2) = \frac{k_P^2}{S/P} + \frac{2k_Pk_Q}{A} + \frac{k_Q^2}{S/T} = 1
\] (3.10)

**Model 3:** Fate of daughter cells is dependent and asymmetrical divisions dominate the distribution. Symmetric proliferative and symmetric terminal cell divisions do not coexist.

\[
Pr(cell_1, cell_2) = \frac{k_P^2 - k_Qk_P}{S/P} + \frac{2k_Pk_Q + 2k_Qk_P}{A} + \frac{k_Q^2 - k_Qk_P}{S/T} = 1
\] (3.11)
3.3.5 Lineage trees reconstruction by generative functions

I have shown that population rate models can be quite useful to investigate the dynamics of neurogenesis. However, they can tell us little about the composition of the precursor pool at the single cell level given that they are based on average measurements. A much powerful analysis would be to collect data by observing the behavior and gene expression profile of each individual cell during corticogenesis. While this approach is partially feasible in smaller animals like *C. elegans*, recording of the whole lineage tree in mammals is cumbersome and technically very challenging. In Section 3.5.2 I discuss advanced technologies that have been developed in the recently to address such experimental constraints. Here I adopt an alternative approach by using the experimental data to constrain probabilistic cell lineage reconstructions.

In order to reconstruct possible cortical lineage trees I use a probability-generating function approach. Cell proliferation is a discrete branching process whose timestep $\Delta t$ is equal to the cell cycle length. In each timestep every cell can either divide or terminally differentiate. These possibilities can be represented with the probability-generating function (pgf):

$$f(s) = \sum_i p_i s^i = k_Q s + 2(1 - k_Q)s^2$$  \hspace{1cm} (3.12)

$p_i$ is the probability that a cell will give $i$ offspring in the next generation and $s$ is a dummy variable. In other words, a cell divides with probability $p_2 = (1 - k_Q)$ to produce two daughter cells or differentiates with probability $p_1 = k_Q$. The pgf enumerates all the possible outcomes after one timestep, and has the property $\sum_i p_i = 1$.

Single trees with multiple types of differentiated cells are generated by randomly sampling from the laminar probability distributions $k_QX$. To ensure that progenitor cells do not switch to the production of earlier developmental stages, I enforce the current layer $X$ to be equal or greater to the past current layer.

Lineage trees are visualized with the use of a custom application I developed in Python, LinAnalyzer, which takes advantage of Python bindings for the C++ graphical library igraph to display graphical information.
3.4 Results: Reconstruction of cortical cell lineages

3.4.1 ODE model of cell cycle progression and cell differentiation

The dynamics of corticogenesis is determined by the proliferative behavior of cells and the length of the cell cycle. Area 3 (primary somatosensory area, parietal) and 6 (primary motor area, frontal) have significant cytoarchitectonic differences, and are thus particularly suited to investigate the influence of those parameters on the final number of neurons in each layer: area 3 is characterized by 20-40% more neurons per column under unit surface than area 6, and a much more thicker layer IV as well. Area 3 and 6 are separated through sharp borders by area 4, which displays intermediate features.

Experimental data on rates of cell division $k_P$, differentiation $k_Q$, laminar probabilities $p_X$ and cell cycle length $T_c$ in area 3 and 6 of the mouse cortex were experimentally measured by Polleaux and Moraillon [17] by radio labeling cohorts of proliferating cells during development. Since the distribution of labeled neurons was evaluated at adulthood, this method allows to correlate developmental changes with cortical cytoarchitectonic features.

The experimental data was fitted on the measured time points as described in Section 3.3.2 (Figure 3.4, Figure 3.5 and Figure 3.6), with R values $>0.99$. Those measurements are however inevitably noisy and averaged on different individuals. Therefore, I test the accuracy of the observations by incorporating the values in a parametric model of ordinary differential equations (ODE) to model the generation of different neural types and compare the resulting cortical lamination with data on the developing and mature murine cortex (Appendix A). Here I use a three-compartment ODE model for the analysis of cell division and differentiation in area 3 and 6 as described in Section 3.3.3. For the moment I consider the population of proliferative cells as homogeneous and I will address the evaluation of different subpopulations in Chapter 4.

The numerical solution of the ODE system using the experimentally measured cell cycle length $T_c$, leaving fraction $k_Q$ and laminar probabilities $p_X$ in function of developmental time generates profiles of cumulative numbers of precursors (proliferative compartment $P$) and neurons (differentiative compartment $Q$) (Figure 3.7A, D). The time evolution of cell populations shows that, at early developmental stages, the majority of divisions are proliferative and lead to an increase in the precursor pool. As corticogenesis proceeds,
Figure 3.4. **Laminar probabilities of area 3.** (A) Laminar probabilities in area 3 for layer VIb, layer VIa, layer V, layer IV, layer II/III. (B) Pairwise fitting of sigmoidal curves to the data. R values for all fitted parameters are higher than 0.99. (C) Data and fitted laminar probabilities.
Figure 3.5. Laminar probabilities of area 6. (A) Laminar probabilities in area 6 for layer VIb, layer VIa, layer V, layer IV, layer II/III. (B) Pairwise fitting of sigmoidal curves to the data. R values for all fitted parameters are higher than 0.99. (C) Data and fitted laminar probabilities.
there is a steady increase in the proportion of differentiative divisions which causes the exhaustion of the precursor pool and an increase in the neuronal output.

According to the results, at the end of cortical neurogenesis area 6 has generated about 10% more neurons than area 3, which is consistent with the fact that area 6 lies at the border between the two hemispheres and is subject to a much greater surface expansion than area 3. The tangential expansion effect is considerable, since the mature area 3 has a higher neural density than the final area 6. For instance, in rats there are on average 48000 ± 4000/mm³ neurons in the parietal cortex, and 34000 ± 4000/mm³ neurons in the frontal cortex. In average a single precursor at E11 generates about 20-30 neurons in area 3 and 30-40 neurons in area 6. Since a cortical column in the murine area 6 contains about 110'000 neurons under a mm² of surface, 2'500 founder cells/mm² are required at E11.

Precursors at E10.5 were reported of giving rise to 600 neurons based on embryonic stem (ES) cell injections into the blastocysts [167], and three rounds of division later, precursors at E11.5 were estimated to generate about 140 neurons based on cell cycle parameter estimation [124]. Both results are likely overestimated: the first because of the highly proliferative behavior of ES cells, the second possibly because of errors in the estimation of cell cycle parameters.
Figure 3.7. ODE simulation. Simulation of proliferative and differentiative populations of cells in areas 3 and 6. We start the simulation at E11 with 50 precursor cells. Shaded areas represent the standard deviation considering 10% experimental error in the measurements of leaving fraction and average cell cycle length. Error bars represent the standard deviation over multiple measurements. (A,D) Cumulative number of cells in the proliferative pool (black) versus differentiative pool (gray) from E11 to E18.5. Datapoints correspond to measured number of cells on Ki67/ToPro immunostained cortical slices at E14, E15, E16, and E17 in a 100 μm wide column. (B,E) Instantaneous cell production rate in the proliferative (black) and differentiative (gray) pools. (C,F) Cumulative number of cells in layers VIa (blue), VIb (sea-green), V (green), IV (orange), and II/III (red). (G) Plot of all the neurons encountered in 500-μm wide strips of areas 3, 4, and 6 on a 10-μm-thick paraffin section counterstained with cresyl violet (adapted from Polleux et al. [17]). (H, I) Simulation of final laminar production (gray) compared to experimental data from the adult mice cortex (white).

Rates of cell production were evaluated by computing the derivative of the cumulative functions (Figure 3.7B, E). According to the simulation, neuron production in area 6 exceeds that of area 3 and is responsible for the increased number of cells in area 6. However, there is a transient increase in the rate of neuron production in area 3 compared to area 6 at E14.5, which is when maximal difference in the leaving fraction $k_Q$ is found between areas 3 and 6. The increase at E14.5 is compensated by a decrease at E16, which guarantees that the precursor pool is not depleted prematurely. The effects of rate changes on the final cytoarchitecture are best illustrated by the laminar cumulative distributions.
Cortical layers are generated when precursor cells undergo their final mitosis in the ventricular or subventricular zone. In order to estimate the proportion of neurons allocated to each layer I multiply the probabilities of laminar production $p_X$ characterizing corticogenesis of areas 6 and 3 with the calculated neuronal output (Figure 3.7C, F). The final numbers of cells in different layers for each area are estimated at E18.5 (Figure 3.7H, I and Table 3.1). Neural production starts in area 3 slightly before area 6 with layer VI neurons, which are subdivided into later VIb, a thin layer of highly packed cells close to the white matter, and the major fraction layer VIa. Up to E14.5 all precursors in their last cell cycle are committed to infragranular layers, and the increase in neural output in area 3 at E14.5 depletes the precursor pool destined to layer V, causing a relative reduction in the number of differentiative cells generated. The depletion is compensated by an increase of proliferative rates between E15 and E17, which causes an expansion of layer IV. In contrast area 6 is characterized by a sustained decrease in the proliferative rate between E15 and E17, which results in a very thin layer IV in agreement with experimental observations [168, 169]. At E18.5 remaining proliferating radial glial cells differentiate into glial cell precursors, and switch to a series of symmetrical divisions that will generate all glial cells (not modeled here).

The solution of the ODE system is compared against measurements of the number of the total proliferative and differentiative cells at different developmental stages counted in a 100 µm wide column of cortex at E14, E15, E16, and E17 (Figure 3.7A, D). The measurements represent a underestimation of the total number of cells produced by an initial founder cell population, especially of differentiative cells, given that I do not take into account the effect of the expansion of cortical columns in the tangential dimension.

Since the number of neurons produced strongly depends on the shape assumed by the

<table>
<thead>
<tr>
<th>Layer</th>
<th>Area 6</th>
<th>Area 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Experimental</td>
<td>ODE Model</td>
</tr>
<tr>
<td>I</td>
<td>0.9 ± 0.87</td>
<td>N.d.</td>
</tr>
<tr>
<td>III-II</td>
<td>27.1 ± 6.36</td>
<td>29.2 ± 0.5</td>
</tr>
<tr>
<td>IV</td>
<td>12.0 ± 2.12</td>
<td>12.2 ± 2.5</td>
</tr>
<tr>
<td>V</td>
<td>27.0 ± 6.07</td>
<td>27.4 ± 4.1</td>
</tr>
<tr>
<td>VIa</td>
<td>25.8 ± 2.19</td>
<td>27.2 ± 1.0</td>
</tr>
<tr>
<td>VIb</td>
<td>7.1 ± 1.64</td>
<td>4.1 ± 0.8</td>
</tr>
</tbody>
</table>

Table 3.1. Final laminar production percentage in area 6 and 3. Quantification of simulated final neuronal production in each layer compared with experimental evaluation. Values are given in % with standard deviation. Experimental values were averaged and normalized to 100%. N.d., not determined.
Chapter 3. Inferring cell lineages underlying cortical development

$k_Q$ curve, I adjusted the fitting on the experimental data in the 10% error interval such that the disparity between simulation and observed number of cells is minimal, which was done by iterating over different value combinations for $k_Q$ and computing the simulation error. This guarantees that the estimation of the number of proliferative and differentiative cells at any given time point is as closely as possible to the true underlying distribution.

3.4.2 Distribution of cell division modes

The production of daughter cell types different from each other is referred to as asymmetric cell division. This phenomenon results from the intrinsic ability of cells to perform asymmetric distribution of cell fates, or asymmetric environmental influences. Although generated by a core mechanism, asymmetric divisions have been classified in different ways depending on the observed features. A classical definition concerns the proliferative/differentiative fate of cells, that is whether the daughter cells continue to divide or not: symmetric proliferative divisions ($S/P$) produce two progenitor cells, symmetric terminal divisions ($S/T$) produce two neurons, and asymmetric divisions produce one progenitor cell and one neuron ($A$).

The values assumed by $k_Q$ during development constrains the distribution of division types, either symmetric or asymmetric that cells can undergo. However, how changes in the proliferative rate impact the topology of single lineages is ambiguous. This ambiguity arises from the fact that to a single value of $k_Q$ may correspond different distributions of modes of cell division, for instance two asymmetric divisions correspond to the same differentiation rate as the combination of a symmetric proliferative and a symmetric terminal division.

Depending on how the dependence of the two daughter cells influence cell fates, three different models (Figure 3.8) can be envisaged, and any model in between those extremes (see Section 3.3.4): the fate of daughter cells is dependent and symmetrical distributions dominate the distribution (Model 1); cell fates are independent, and/or proportion of symmetrical versus asymmetrical divisions follow the binomial distribution (Model 2); the fate of daughter cells is dependent and asymmetrical divisions dominate the distribution (Model 3). The three models are used as null hypotheses to compare with measured proportions of mode of cell divisions.
**Figure 3.8. Proportions of types of divisions.** Distribution of cell division modes in area 3 and area 6 in function of time according to three possible models: (A, D) no asymmetric divisions; (B, E), binomial distribution; (C, F), maximal percent of asymmetric divisions.

Model 1 is certainly unrealistic, since asymmetric division of radial glial cells have been often observed [58, 59]. Analysis of retrolabeled lineages have suggested that the distribution of neurons originated from a common precursor are best described by Model 2 [162]. Model 2 is also closer to measurements performed in the rat at the middle of neurogenesis, where the following proportion of divisions were observed: 10% S/P, 40% A, 45% S/T, 5% unclassified [59]. Notice that the majority of asymmetric divisions occur in the VZ, while symmetrically terminal divisions almost exclusively in the SVZ.

An interesting observation is the correlation between cell cycle length, more accurately the length of G1 phase $T_{G1}$, and the mode of division: during corticogenesis the increase in the frequencies of differentiative divisions is accompanied by a progressive slowing of $T_{G1}$ [170, 171]. G1 is a critical phase in which precursors are able to respond to extracellular cues and undergo fate commitment [118, 172], either proliferative or differentiative, and its length is believed to be correlated with the mode of cell division [15, 156]. Although the length of $T_{G1}$, and thus the cell cycle length, is correlated with the different distributions of division types, changes in the cell cycle parameter alone have very little effect on the final cytoarchitecture (as we showed in [156]), other than determining the amount of time needed for the completion of neurogenesis. In other words, the length of the cell cycle works in combination with, or is an effect of, the cell division mode.
Figure 3.9. Probabilistic reconstruction of lineage trees. Lineage trees are reconstructed by sampling from the probability distribution, measured experimentally. (A-C) Probability distribution for area 3 and 6. Points, experimental data; lines, fitted data. (C-D) Example of sampled lineage trees (time layout). Black: precursor cell; blue: layer VIb; green: layer V; yellow: layer V; orange: layer IV; red: layer II-III; dashed lines, proliferation of glial precursor cells (not modeled).

3.4.3 Lineage trees reconstruction with probability generative functions

The leaving fraction [17] (estimated and validated with the ODE model as described in the previous section) describes the probability for a proliferative cell to differentiate into a layer specific neuron in function of developmental time. The distributions for area 3 and 6 constrain considerably the topologies of possible lineages trees, and are thus useful to investigate the mechanisms behind corticogenesis.

Cell lineage trees are reconstructed by sampling from the probability distribution with a probability generating function, which describes the probability of cells to either divide or differentiate in a sequence of cell divisions (Figure 3.9). The formula is used to simulate
the generation of cell lineages from single precursor cells. After \( n \) simulations I obtain a forest of \( n \) different trees whose moments (mean and variance) are defined by the probability generating function (Figure 3.10).

Distribution of cells across 100 lineages for each area were analyzed for mode of divisions, and presence of fate-restricted progenitors (Figure 3.11). As expected, cumulative distributions in differentiative compartments, and mode of division represent a discretized version of the distributions measured in Figure 3.7 and Figure 3.8, according to Model 2. Notice that not all progenitor cells produce pyramidal neurons: a restricted pool of precursors is maintained and will either give rise to glial cells by switching to rounds of symmetric divisions, or become quiescent as adult neural precursors (not simulated here for lack of quantitative data).

Progenitor fate restriction was measured by computing the percentage of superficial layer neurons (IV-II/III) versus deeper layer neurons (VI-V) produced by each precursor cell. 0% indicates a precursor that produces exclusively deep layers, 50% a precursor that produces an equal amount of superficial and deeper neurons, 100% a precursor that produces only superficial layers. As described by in-vivo retrolabeling experiments [134, 135, 137, 138, 139, 140, 141], and later confirmed in in-vitro dissociated cultures [127], we can observe that earlier progenitors contribute to the production of neurons for all cortical layers.

Notably, the lineage topologies cause the generation of fate-restricted progenitors toward superficial layers (80-100%) of about 1.8% (E12.5), 11.1% (E13.5), 77.9% (E14.5), and 98.3% (E15.5). Recent experiments have reported the presence of about 4.6% (E10.5), 17.9% (E13.5), and 34.7% (E15.5) superficial layers restricted progenitors (more precisely at \( \sim 87\% \)), independently of niche or birthdate [106]. At early developmental stages the percentage of fate restricted progenitors is closed to the experimental observations (slightly underestimated since the lineage reconstruction starts only at E11). The evident discrepancy of percentages during late developmental stages is due to the fact that we do not consider the proliferation of glial precursors cells and adult precursor cells.

In overall, the structure of reconstructed lineage trees (constrained by the target probability distribution) seems to be in agreement with most experimental quantification on cell numbers, percentages of cell division modes, and percentages of fate-restricted progenitors. Those results indicate that a hybrid model containing both multipotent early
Figure 3.10. Lineage tree examples. Examples of 4 probabilistic lineage trees reconstructed from the probability generating function of (A) area 3 and (B) 6. Black: precursor cell; blue: layer VIb; green: layer VIb; yellow: layer V; orange: layer IV; red: layer II-III. Lineages are disposed in circle, with the initial precursor cell positioned near the center of the image. Some lineages overlap because of the layout algorithm.
Figure 3.11. **Lineage tree distributions.** Cumulative distribution of proliferative/differentiative cells in (A) area 3 and (D) 6. Mode of division in (B) area 3 and (E) 6. Fate-restricted progenitors toward superficial layers, that is the ratio upper layers neurons over lower layer neurons produced by each precursor cell, in (C) area 3 and (F) 6. Distributions were measured over 100 reconstructed lineages in each area.

Progenitors, and fate-restricted progenitors (whose percentage increases over time), is compatible with the measured distribution of different cell types both in area 3 and 6. Albeit realistic, not all reconstructed lineages may actually arise in reality, and a further selection would require the use of additional constraints or the collection of higher resolution data. Moreover, an exhaustive identification of specific fate-restricted precursors and their contribution to the lineage three, can be achieved only by the precise tracing of the lineage tree until the production of different neuronal types.
3.5 Discussion

3.5.1 The final cortical cytoarchitecture is defined by mode of cell divisions

Area 3 and 6 of the adult mouse cortex are characterized by a radically different laminar organization, which reflects their distinct functional roles. Experimental findings show how the temporal changes in the cell-cycle parameters generate different numbers of neurons in successive cortical layers [17]. I use a mathematical model of cell cycle regulation based on ordinary differential equations (ODE) [18] to explore how cell-cycle parameters (rate of cell-cycle progression and cell-cycle exit probability) influence neuron number in the murine sensimotor area 6 and 3.

Mathematical modeling has been used in the past to explore how cell-cycle parameters relate to the neuron production and the final cytoarchitecture [18, 118, 173]. Their influence on the neuronal output is complex: an increase in the cell cycle exit probability leads to only a transient increase in neuron production because of a rapid exhaustion of the precursor pool. In contrast, increase in the rate of proliferation results in a dramatically higher neurogenic output through the amplification of the precursors. Given the indirect effect that cell cycle parameters have on the neuronal output, differential equations are a valuable help to understand their impact on the final cytoarchitecture.

The ODE simulation uses the experimentally observed $T_c$ and $k_Q$ functions to generate profiles of neuron production between the onset of neurogenesis (E11) until the switch to the production of glial cells (E18.5). The generated cumulative profiles of proliferative and differentiative cells determine the final cytoarchitecture, which is compared to independent experimental data in the adult murine cortex [168, 169].

The main ODE model prediction is a consistent expansion of the granular layer (IV) at the expense of layer V in area 3 in respect to area 6. Area 6 is characterized by a unique proliferative peak at E13.5, followed by a sustained differentiative peak at E16. In contrast, area 3 has two differentiative peaks at E14.5 and E17, with an intermediate proliferative peak at E15.5. These differences have two main effects: firstly, the higher differentiation rates in area 3 at early stages of neurogenesis reduce the pool of cells for infragranular layers, layer V in particular; secondly, the proliferative peak beginning at E16 contributes to the expansion of layer IV.
As indicated by the standard deviation, cumulative numbers are quite sensitive to errors in the estimation of the cell cycle parameters, especially the leaving fraction index ($k_Q$). In contrast, deviations in cell cycle length do not have a significant impact on the shape of the cortical architecture. Since targeted experimental modification of the cell cycle length have been shown to have dramatic effects on the laminar production [156], the control over cell-cycle length has undoubtedly an indirect effect on the cytoarchitecture by modulating the mode of cell division.

These observations lead to the hypothesis that the core regulatory mechanisms responsible to shape the final cortical architecture is the asymmetric distribution of cell fates among daughter cells. Under this assumption, changes in the cytoarchitecture – in accordance with the radial glia unit hypothesis [111] – are the direct result of changes in the regulatory network responsible for cell fate determination.

3.5.2 Why are lineage trees important?

Research on in-vivo cortical development has mainly focused on discrete stages of cortical development, or on dissociated cultures. Although illuminating, these experiments are not able to capture a cohesive picture of neurogenesis. Advances in the understanding of the logic behind neurogenesis, and biological developmental processes in general, has been hampered by the limited access to information on individual cell features and their relative position in the cell lineage tree.

Direct observations of cells in *C. elegans* have led to the reconstruction of the complete lineage tree of the nematode [174], also referred to as fate map [175]. Unfortunately, data on cell lineages in the mouse cortex is very sparse: direct observation of cell division is not applicable to mammals – because of the opaque body and the tremendous amount of cells produced – and in-vitro brain slice preparations can be usually maintained in culture only for a limited period of time. A variety of experimental approaches rely on cell labeling and tracking of the progeny at the microscope, but they are also limited on the time scale and are highly invasive. A novel promising approach is based on reconstructing the cell mitotic history by exploiting stochastic DNA mutations that occur every time a cell divides [176] or injection of a pool of recombinant retroviruses [149].
Given the current technical limitations, and also because of the more theoretical orientation of this Thesis, I reconstruct cell lineage trees through sampling of the probability distribution described by experiments on mitotic history [17]. The dataset of lineage trees is probabilistic, in the sense that although constrained by the experimental data, it may display features that do not occur in real corticogenesis. The accuracy of fit to the data points is validated by the ODE model by comparing the final number of neurons in the adult murine cortex.

The topology of lineage trees in both area 3 and 6 makes an interesting prediction. According to the emerging model, lineages are characterized by mixed populations of progenitor cells: multipotent progenitors, and fate-restricted progenitors toward deeper or superficial layers. Moreover, superficial layer fate-restricted progenitor cells increase in number over time. This finding is rather interesting, since it seems to reconcile two opposite experimental evidences, that is the concomitant presence of multi-potent [127] and fate restricted progenitors in the neural precursor pool [106].

I use the collection of lineages trees to illustrate in the next Chapters how to recover developmental pathways (statistically representative patterns of cell divisions and differentiations, that we express with a list of division rules) from partial lineage and genetic expression information, and infer the existence of particular proliferative subpopulations. The power of the method proposed in Chapter 4 is it’s ability to generate data-driven models, as opposed to imposing an ad-hoc model, and could be easily improved by higher resolution datasets.

In conclusion, the cell lineage information is crucial since it directly captures the sequence and modes of cell divisions, which – as we have seen from population based models - lies at the heart of developmental processes. I envisage that such approaches are going to be decisive to decipher the genetic regulation of developmental processes.
Chapter 4

Analysis of developmental pathways in murine cortical development

The generation of cortical pyramidal neurons is described by cell lineage trees in the form of binary trees, that is the genealogical history of every precursor cell. Cell lineages offer a unique insight into the process of neurogenesis, and repetitive patterns of division are reminiscent of the tight control over cell fate decisions in invertebrate development. However, cell lineage data is not easily interpretable.

Here, I present a novel approach for mapping differentiation trajectories to the underlying transcriptional network. The detailed sequence of cell fate specifications is described by the cell lineage. I exploit lineage and sparse phenotypic information to classify progenitor cells into different subpopulations by means of spectral decomposition (Section 4.4.3). This approach enables to identify the most likely differentiation paths that leads to the production of different neuronal cell types (Section 4.4.4 and Section 4.4.5). The resulting model constitutes a compact probabilistic state model of the developmental process. To my knowledge this is the first time an algorithmic description of cortical neurogenesis is generated directly from sparse biological data.
4.1 Introduction

4.1.1 Development as a self-construction process

Conventional engineering is about design of systems that are completely defined in a limited range of operating conditions. The system’s blueprint is dawned in a top-down fashion by an external architect, and there is a clear distinction between construction phase and functional phase. Unlike direct design, biological entities self-construct through the processes of development: structural and functional order result collectively from energy-dissipative processes involving cell replication and local interactions among cells or molecular components embedded in the physical environment [10].

Although contextual information from the environment is certainly important, the specification of how biological systems are constructed is stored in the genome [177], which is composed up to tens of thousands protein-coding genes, depending on the species. The genetic information is identical in each cell, yet progenitor cells can assume specialized functions by expressing different patterns of gene products (transcription factors and functional proteins). Gene expression profiles follow specific trajectories in different cells during development and are responsible for their specialization and their diverse localization in the organism.

How gene expression unfolds in spatial and temporal patterns is captured by the cell lineage tree, which describes the genealogical history of each individual cell. Lineage data offers a unique insight into the developmental process by describing the number of cells generated, the distribution of cell fates, and the topology or patterns of cell division. For instance, neurogenesis in nematodes and ascidians is achieved by an invariant sequence of cell divisions [174], that is the division patterns are the same in every organism. Not surprisingly, the invertebrate neural circuitry is mostly invariant across individuals of the same species, with distinct neuronal types and their stereotyped connections [178]. This is a clear example where the developmental program has a tight control over the rounds of mitosis and cell fate segregation depends on asymmetrically inherited intrinsic determinants or cell-cell communication. Molecular mechanisms that operate in the invertebrate development are found also in the vertebrate brain, but in mammals there is high individual variability in the number and types of neuron produced [179].
4.1.2 Lineage trees are composed of cell division patterns

Cell lineages are the key to understand development, since they describe the generation of different cells and of the appropriate environment into which development takes place. The ordered sequence of cell divisions is controlled by transcriptional networks and epigenetic regulation, which operate as a cellular memory. In the worst coding scenario we would have to explicitly specify in the genome the state, behavior and spacial position for every single cell and cell division. A brief calculation shows that for instance the specification of the 80 billion neurons that are found in the human brain would require \(46\) billion possible molecular combinations, one for each cell division event. However, such an expensive combinatorial expression of genetic markers is unrealistic. Only a fraction of the information that a cell harbors is encoded into molecular signals. Even if we would consider that all cells of an organism are different from each other, the sparseness in their molecular profile leads inevitably to some cells being much similar than others (cell types). It is reasonable to assume that those similarities are exploited during the construction process: cells with related expression profiles and spatial positions are probably generated by a common molecular mechanism. Under this assumption, the lineage description defines the reachable states or cell types in which a cell can be found, and the possibility of transition between states. Formally, it can be described as a series of unique rules, each corresponding to a cell division, for instance \(X \rightarrow (Y, Z)\), which means: 'cell in state \(X\) divides into cells in states \(Y\) and \(Z\)'. If it is possible to find division rules that are repetitively reused in different part of the lineage, we would be able to describe the growing process with a compact model composed of a minimal set of rules (or a compact genetic specification).

4.1.3 Models of lineage evolution

The size and shape of lineage trees reflects dynamic changes in the mode of division, and reveal repeated patterns of mitosis. A first step in the modeling of single cell behavior during neurogenesis has been the use of two-type Galton–Watson branching processes \[180\] to study decision making and variability along individual cell lineages recorded \emph{in-vitro} \[181\]. The model is based on cell division probabilities, defined as probabilities for pairs of progeny (the fate of two daughter cells is highly correlated since they share the same mother cell and a proximal spacial position). This definition reflects directly the
different types of mitosis, with $k_{PP}$ the probability of a symmetric proliferative (S/P) division, $k_{PQ}$ of a asymmetric (A) division, and $k_{QQ}$ of a symmetric terminal (S/T) division:

$$Pr(\text{division}) = k_{PP} + 2k_{PQ} + k_{QQ} = 1$$

$$S/P \quad 2A \quad S/T$$

(4.1)

The Galton–Watson branching processes is a pattern-generator model used as a stochastic null model, which assumes that the data can be described by the branching probabilities independently on the previous branch history, that is the progenitor pool is considered to be an homogeneous population of cells without memory of previous states (null hypothesis). Families of lineage trees are mathematically defined by probability sequences, for instance the probability of lineage trees $L_n$ with $n$ terminal cells are described by the following equations [181] (Figure 4.1A):

$$L_2 = k_{QQ}$$

$$L_3 = k_{QQ}k_{PP}$$

$$L_4 = k_{PP}k_{QQ} + k_{QQ}k_{PQ}^2$$

$$\ldots$$

(4.2)

Interestingly, branching processes decompose lineage trees into sequences of branching patterns, like algorithmic sequences of instructions. Nevertheless, they are not very useful to investigate local mechanisms specific to each single cell, since they consider global probabilities applied to an homogeneous precursor pool. Indeed, branching probabilities do depend on cell’s previous history and their environment, and their investigation requires branching models with conditional dependencies on the mother cells, for instance by introducing a generation dependency (Figure 4.1B) or more complicated models with state memory, as I describe in this Chapter.

4.1.4 Inference of transcriptional networks

Advances in cell lineage research are severely hampered by the scarcity of cell markers and the impossibility to visualize the dynamics of more than few proteins in a single time-lapse microscopy experiment. It is however possible to observe partial cell lineages, and sparse labeling information in form of observed features such as gene expression data,
Figure 4.1. Examples of cell lineage trees and corresponding division rules model. Examples of cell lineage trees, each producing 4 terminal neurons (N). The distribution of different divisions (colored circles) depend on the branching process considered. (A) Two-type Galton–Watson branching processes and corresponding (B) state model. (C) Generation-dependent branching process and corresponding (D) state model.

cell morphology or other features. The working hypothesis is that the characteristics features of a cell, and of the daughter cells, are determined by an internal unobservable cell state. Under this assumption, each lineage tree can be described by probabilistic models composed of states, and sequences of state transitions.

Formally, a cell lineage can be represented as a directed graph, where each node represents a cell instance, and each edge a genealogical relationship. In this description, each node is associated with a vector, whose elements correspond to measured features. Statistical regularities in large, unordered dataset, can be inferred by clustering, an unsupervised machine learning method. Usually, clustering is used to find groups of data points that have similar distribution of features. To include lineage information, I introduce conditional dependencies between data points.

A rather popular clustering approach is represented by Hidden Markov Trees, which model Markov Tree processes over a set of observed variables, their conditional dependency, and infer unobserved (hidden) states, as we have previously shown \[182\]. Under very sparse sampling conditions this approach is not straightforward since it attempts to solve a problem (finding an appropriate mapping function between observed features and hidden states), which is more complex than the desired result (predicting hidden
states of unlabeled data). According to Vapnik’s principle, also referred as transductive learning, it is possible to exploit relationships between datapoints to infer missing labels, since highly connected objects are likely to share similar labels.

For this purpose, I consider a class of clustering algorithms referred to as spectral clustering [183, 184], which uses information from the local data structure to infer the model’s hidden states. The algorithm as been used in several fields such as image segmentation [185, 186], nonlinear dimensionality reduction [187, 188, 189, 190], manifold approximation [187, 190], random walk estimation [188] and multiple-view learning [191]. The central idea is to define an appropriate distance measure that captures not only the similarity between different molecular cell configurations (structural states) but also consider the transitions that allow one state to morph into the next (dynamic states).

In the present work, I exploit sparse cell lineage information to generate state models of development, specifically in the context of the generation of a laminated cortical column in the mouse. The elegant disposition of identifiable neural types into lamina and their genetic characterization makes the cortex a convenient model to address the question of how such a diversity of cell types is generated.

Firstly, I reconstruct cell lineages during mouse corticogenesis by sampling from cell number distributions obtained experimentally by radioactive labeling of cohorts of precursor cells at several time points during development [18], as described in Chapter 3. To each terminal state, I associate feature vector information in the form of transcription factor expression profile. Secondly, I develop a novel approach to analyze cell lineage data and sparse feature information. The algorithm is validated on a database of artificial data, and used to analyze the reconstructed murine cell lineages. The obtained state diagram describes the relationship between states and expression profiles of transcription factors, and identifies statistically representative pathways of cell differentiation. Finally, I reconstruct plausible logic models of transcriptional regulation compatible with the data. Computational models [192] consists of algorithmic descriptions, according to which biological entities are characterized by states and state transitions.
4.2 Objectives of this Chapter

How does the compressed genetic information unfold itself in order to control the execution of developmental processes? And how are the different cell types encoded and related to each other? In order to address these issues, I have chosen to investigate the relationship between the detailed sequence of possible cell fate specifications, which is provided by the genealogical history of individual precursors (cell lineage), and plausible generative models.

Cell lineages are the key to understand development, since they describe the generation of different cells and of the appropriate environment into which development take place. Formally, a cell lineage can be represented as a directed graph, where each node represents a cell instance, and each edge a genealogical relationship. In this description, each node is associated with a vector, whose elements correspond to measured features.

Statistical regularities in large, unordered dataset, can be automatically extracted by clustering, an unsupervised machine learning method. I propose a method inspired on spectral clustering to identify statically significant recurrence of division patterns, and express cell lineage data with a compressed representation of states associated with feature distributions, and state transitions.

In the present work, I exploit sparse cell lineage information to generate transcriptional network models of development, specifically in the context of the generation of a laminated cortical column in the mouse. The elegant disposition of identifiable neural types into lamina and their genetic characterization makes the cortex a convenient model to address the question of how such a diversity of cell types is generated.

Objectives for this Chapter are:

- **Algorithmic analysis of cell lineages.** I propose a novel approach based on spectral decomposition to analyze patterns of cell division on a sparse cell lineage dataset. The method is validated on artificially generated lineages, for which the generative model is know.

- **Transcriptional model of corticogenesis.** I show the application of spectral decomposition to reconstruct plausible models for the generation of area 3 and 6 of the murine neocortex.
4.3 Methods: Inference of Markov Branching Processes

4.3.1 Graphical representation of cell lineages

Let $G = \{V, \mathcal{E}\}$ be a directed graph with a finite set of vertexes $V = \{v_1, v_2, \ldots, v_n\}$ and a finite set of directed edges $\mathcal{E} = \{e_{ij}\} \subseteq V \times V$. For weighted directed graphs, there is an asymmetric, non-negative adjacency matrix $W$ that associates each edge with a weight as following: $w_{ij} = 1$ if there is a direct link that connects node $i$ to node $j$ or $w_{ij} = 0$ otherwise. We define the in-degree matrix $D_{in}$ as the diagonal matrix of the sum of weights on incoming edges and the out-degree matrix $D_{out}$ as the diagonal matrix of the sum of weights on outgoing edges:

$$D_{in}(i, i) = \sum_i w_{ij}, \quad D_{out}(i, i) = \sum_j w_{ij}$$ \hspace{1cm} (4.3)

Given a directed weighted graph, there is a natural random walk on the graph defined by the transition probability matrix $P$, where $p_{ij} = w_{ij}/d_{out}(i)$ for all edges, and 0 otherwise. The random walk on a connected, acyclic graph converges to a unique stationary distribution $\pi$.

It is convenient to represent cell lineages and feature information in the form of a direct graph. Each vertex $V$ of the graph corresponds to a cell state, and each edge $E$ encodes genealogical relationships between states. The transition probability matrix $P$ represents the strength of the links (proportion of cells, multiplied by 2 to account for cell division). Moreover, cell states are labeled by a feature vector, describing the occurrence of observed features on the given node. Unobserved features are referred to as unknown (?) and the corresponding node is considered unlabeled.

4.3.2 Spectral decomposition

Given a directed graph $G = \{V, \mathcal{E}\}$ and vectors of observed features $(f_1, f_2, \cdots, f_L)$ for a set of labeled nodes, I consider the task of inferring missing feature values of unlabeled nodes on acyclic directed graphs. Nodes with the same (or similar) feature vectors are assigned to a unique label, and are represented by a single node, which inherits all the incoming edges.
For undirected graphs, a widely used method in semi-supervised learning is spectral clustering, a technique that make use of the spectrum (eigenvectors and eigenvalues) of a similarity matrix to cluster data into highly connected groups. While spectral clustering has been studied extensively, there are only few attempts to extend the theory to directed graphs, in particular strongly directed graphs [193].

Here I propose an approach based on the spectral decomposition of the Laplacian $L$ of the normalized directed random-walk matrix:

$$L = I - O^{-1}P I = U \Lambda U^T$$

where $P$ is the directed transition probability matrix, $O$ the out-degree matrix, $I$ the in-degree matrix, and $I$ the identity matrix. $\Lambda = \text{diag} [\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n]$ is the diagonal matrix, and $U = [U_1 U_2 \ldots U_n]$ is the orthonormal matrix with eigenvectors of $L$ in each column. By taking only the last $k$ eigenvectors, we obtain $U_k$, where $k$ corresponds to the number of terminal nodes contained in the graph. $U_k : \mathcal{V} \rightarrow \mathbb{R}^n$ provide an embedding for each vertex of the graph in a $k$-dimensional space, where each row corresponds to a vertex. The main idea of spectral decomposition is to map each graph node into a space where distances between nodes are proportional to the directed random-walk distances on the graph.

The mapping of the graph to a $k$-dimensional space is particularly useful, since in this space conventional machine learning algorithms, such as $k$-means or hierarchical clustering, can be used. I compute the feature vectors for each node by simple linear algebra, since I assume that each node projection can be represented by a linear combination of feature vectors:

$$F = U_k F$$

where $F$ is a $k \times l$ matrix containing the distribution of $l$ features on observed nodes, $U_k$ is a $n \times k$ matrix of eigenvalues of $L$, and $F$ is a $n \times l$ matrix with observed and estimated feature distributions. For the case where observed features are also available for a subset of non-terminal nodes, more complex models to map $U_k$ to $F$ are required, such as the logistic model.
4.3.3 Multi-type Markov Branching Process

The graphical representation of cell lineages can be interpreted as a Markov branching process with multiple states. A branching processes is a discrete-time random process that models a population in which each particle in generation $t$ produces some number of individuals in generation $t+1$, each of which can assume one over $m$ different states. Let denote a finite set of states $Q = \{q_1, q_2, \ldots, q_m\}$, and $Z_n = (z_1, z_2, \ldots, z_m)$ the vector of variables describing the population size at the $n$th generation in each state.

The time-invariant transition probability $p_{ij}$ describes the probability for each particle to transit from state $i$ to state $j$ (Markov property):

$$p_{ij} = \mathbb{P}(Z_{n,j} = z_j | Z_{n-1,i} = z_i) \quad (4.6)$$

The system evolution is completely characterized by the set of states, the marginal distribution of its initial state $Z_0$, and the transition probabilities between states. We write the joint probability distribution of $Z_n$:

$$\mathbb{P}(Z_n) = \mathbb{P}(Z_0) \prod_{t=1}^{n} \mathbb{P}(Z_t | Z_{t-1}) \quad (4.7)$$

By setting the elements of the transition probability matrix $P$ equal to the probability of moving from state $i$ to a state $j$, the equation may be rewritten in matrix representation:

$$\mathbb{P}(Z_n) = \mathbb{P}(Z_0) \prod_{t=1}^{n} \mathbb{P}(Z_t | Z_{t-1}) = Z_0 P^n \quad (4.8)$$

The conditional expectation of the Markov process is given by:

$$\mathbb{E}(Z_{n,j} = z_j | Z_{n-1,i} = z_i) = \sum_i p_{ij} Z_i \quad (4.9)$$

and the conditional variance:

$$\mathbb{V}(Z_{n,j} = z_j | Z_{n-1,i} = z_i) = \sum_i p_{ij} Z_i^2 - \left( \sum_i p_{ij} Z_i \right)^2 \quad (4.10)$$

The variance estimation is an upper boundary, since they do not take into consideration dependencies between sister cells.
4.3.4 Dimensionality reduction

The Euclidean distance $\delta$ between each pair of nodes $(r, s)$ is referred to as the distance matrix, where distances are computed as following:

$$\delta_{rs}^2 = (\mathbf{f}_r - \mathbf{f}_s)(\mathbf{f}_r - \mathbf{f}_s)^T$$  \hspace{1cm} (4.11)

with $\mathbf{f}_i$ the row vector of $\mathcal{F}$ corresponding to the feature vector of node $i$.

Binary clustering on the distance measure is computed by the single linkage algorithm, and nodes whose distance is less than a given threshold, are clustered together by iteratively combining the corresponding transition probabilities. How probabilities should be combined is given by the computation of the log likelihood of transition probabilities. From Equation 4.8, we write the likelihood of the transition matrix:

$$L(\mathbf{P}) = P(Z_0) \prod_{i=1}^{m} \prod_{j=1}^{m} p_{ij}^{n_{ij}}$$  \hspace{1cm} (4.12)

where $m$ is the number of states, and $n_{ij}$ the transition counts from $i$ to $j$. The corresponding log-likelihood is defined as:

$$logL(\mathbf{P}) = logP(Z_0) + \sum_{i,j} n_{ij} p_{ij}$$  \hspace{1cm} (4.13)

By taking the derivative and setting it to zero at $\hat{p}_{ij}$, we obtain the maximum likelihood estimator (MLE):

$$\hat{p}_{ij} = \frac{n_{ij}}{\sum_{i,j} n_{ij}}$$  \hspace{1cm} (4.14)

Not surprisingly, the probability of a transition to node $j$ from mode $i$ is given by the transition count $n_{ij}$ over the total count number of transitions from that node $i$. This result justifies that state transition probabilities for merged states are obtained by averaging the transition probabilities from the original individual states, since:

$$\hat{p}_{1+2} = \frac{n_1 \sum_{i,j} n_{1j}}{\sum_{i,j} n_{1j} n_1} + \frac{n_2 \sum_{i,j} n_{2j}}{\sum_{i,j} n_{1j} n_2}$$

$$= \frac{n_1 + n_2}{\sum_{i,j} n_{1j} n_1 + \sum_{i,j} n_{2j} n_2}$$  \hspace{1cm} (4.15)

$$= \frac{n_{1+2}}{\sum_{i,j} n_{1j} n_{1+2}}$$
Below I consider three different models for computing the averaged transition probability matrix $P$. Dependent on the model, different trade-offs between model accuracy and model complexity are obtained.

4.3.4.1 *Homogeneous Markov Model (HM model)*

In this approach, the system is modeled with a time-independent transition probability matrix $P$ (classical Markov branching process). When two states are merged into a single one, the corresponding transition probabilities are averaged as previously described. In other words $P$, or it’s reduced form, completely describes the Markov branching process. While the probability distributions over cell states of a reduced HM model is guaranteed to converge asymptotically to the original distribution, the process dynamics may not be modeled as accurately, especially in presence of sharp changes in the state distribution.

4.3.4.2 *Time-dependent Markov Model (TM model)*

In this approach, the time-dependent transition probability matrix $P(t)$ is a three-dimensional matrix, with the third dimension encoding time. That is, only transition probabilities corresponding both to the same clustered state and the same time point are averaged. The time-dependent model is much more precise in capturing the dynamics of Markov branching processes also after a consistent reduction in the number of states. However, this is achieved at the expense of an increase in the number of parameters, since we need to store transition probabilities between each states for each time step.

4.3.4.3 *Non-homogeneous Markov Model (NM model)*

As a compromise, the non-homogeneous Markov model combines properties from the homogeneous and time-dependent models. This approach is sparse in the number of parameters, since I consider a two-dimensional time-independent probability transition matrix $P$ and a two dimensional matrix $N$ storing the number of maximal division rounds allowed in each state. The combination of the two matrices is represented is called the age-dependent probability transition matrix $P(a)$. The NM model is more apt at capturing sudden changes in the process dynamics with few parameters.
4.4 Results: A state machine model of cortical development

4.4.1 The Cell Lineage Tree

The genealogical relationship between cells that existed at any given time point during development is described by the Cell Lineage Tree (CLT). A CLT is an acyclic connected graph composed of nodes and edges, and takes the form of a rooted binary tree: the nodes represent cell instances, whereas edges defines the tree topology by connecting daughter and mother cells (Figure 4.2A). Nodes internal to the tree are dividing multipotent or pluripotent proliferative cells and nodes at the leaves of the tree are terminally differentiated cells. By definition the CLT is rooted, which means that there is only one path that links the root with each leaves.

Each cell in the lineage is characterized by its genotype, which is equal in all cells, and by its phenotype, that is the observable characteristics that represents the cell’s current state. The internal state is drawn from the set of features $\Omega = \{f_1, f_2, \ldots f_k, \ldots, f_n\}$, where $f_i$ with $i \in [1, l]$ are observable features, for instance gene expression levels, and $f^*_j$ with $j \in [l+1, n]$ are hidden. The cell state vector is thus defined by $Q = \langle f_{1:1}, f^*_{l+1:n} \rangle$. This internal state contributes to the cell phenotypic expression and for convenience I will use these two terms synonymously.

At the moment of mitosis, two new cell instances are generated, and internal state transitions occur: if the precursor cell divides symmetrically it will produce two daughter instances that have identical internal states, but if it divides asymmetrically, it will produce instances with different internal states. Thus, cell division of an arbitrary cell $c$ can be seen as encoding a branching state transition rule $Q^{(c)} \rightarrow \{Q^{(2c)}, Q^{(2c+1)}\}$ (division rule). While we may potentially measure the feature vector $Q$ on all cell instances, we assume here to be able to observe it over the terminal cells only.

4.4.2 The State Diagram

An alternative, more functional view of the lineage tree is one that describes the underlying cell states and their transitions rather than the relationship between cell instances. I call this alternative view the State Diagram (SD). Here the vertexes of the state diagram represent all possible observable internal cell states, and edges between vertexes’s represent the ability to transit from one state to the other at cell division (Figure 4.2B).
Figure 4.2. From Cell Lineage Trees to the State Diagram. (A) Toy example of sublineage trees, where each node corresponds to a cell and lines to sequences of cell divisions. Two progenitor cells (dark gray) divide into proliferative cells (light gray) and give rise to 22 terminally differentiated cells. Colors represent the observed feature vector \( \langle f_A, f_B, f_C \rangle \). (B) The state transition diagram shows how cell states are connected. Directed arrows indicate the transition probability \( p_{ij} \) from state \( i \) to \( j \). (C) State diagram with reduced number of intermediate states.

The SD starts from a set of initial precursor cells (roots). When progenitor cells divide, they generate two daughter cells, which will eventually divide and evolve subtrees or terminally differentiate. I account for the number of cells transiting from one state to the other in terms of the weights of the edges in the SD, which will be interpreted as state transition probabilities. The CLT describes both terminal cell identities and their individual ontogenies, whereas the SD explains the experimentally observed numbers and dynamics of production of neuronal types in terms of state transitions.

If we consider every cell and cell division as unique, each cell phenotype defined by experimentally observable features is assigned to its own state. However identical and very similar states can be clustered into classes (cell types). Hence, a CLT may produce many instances that share the same state configuration \( Q \), particularly at its leaves. In the example, I categorize terminal states in three classes \( (A, B, C) \) based on the set of markers \( \Omega = \{ f_A, f_B, f_C \} \), which are observed in the terminal states only.

Notably, the decision to group terminal cell instances into unique states is rather arbitrary, since based uniquely on the set of observed markers \( f_i \). Depending on the scientific question being asked, state diagram models at different level of resolution can be used.

In the following sections I shall explain the spectral decomposition algorithm on a series of small toy example cases, before addressing the analysis of reconstructed murine cortical lineages.
4.4.3 Spectral Decomposition

Any set of lineage trees (CLT) can be fully encoded by its corresponding state diagram (SD). For instance, the sublineages in Figure 4.2A can be completely described by a state diagram in which each cell instance is assigned to a unique state. From a biological perspective this means that a developing system would require unique genetic mechanisms to control each single cell division, which is highly unlikely. Indeed, it is plausible that cells in different part of the lineage tree undergo similar state transitions as the result of a common molecular mechanisms. For instance, the SD for the generation of cells belonging to three cell types described previously (Figure 4.2B) has a statistically equivalent, more compact representation (Figure 4.2C), that describes the same developmental process with fewer states.

How can we identify similar states in complex lineages? Given a SD composed of a set of labeled (terminal differentiated) states \( Q = \langle f_{1:n} \rangle \) and unlabeled (progenitor) states \( Q = \langle f^*_1:n \rangle \), we want to infer missing labels and cluster vertexes by exploiting the global structure of the graph. In other words, we want to classify cell states into different subgroups based on their expression profile (node feature vector) and on their genealogical relationships (directed edges).

4.4.3.1 Label propagation

In the presence of labeled and unlabeled instances on undirected graphs, the problem reduces to spectral clustering on graphs. Spectral graph theory is used to characterize the structural properties of undirected graphs using information conveyed by the eigenvalues and eigenvectors of the Laplacian or the Laplacian pseudo-inverse [194, 195]. The model exploits the graph connectivity to compute a dissimilarity measure among states and propagate labels across the unlabeled vertexes.

Spectral clustering is generally not defined for directed graphs, and I thus propose here an alternative approach for acyclic directed graphs based on the spectral decomposition of the Laplacian of the normalized directed weight matrix:

\[
\mathbf{L} = \mathbf{I} - \mathbf{O}^{-1} \mathbf{P} = \mathbf{U} \Lambda \mathbf{U}^T
\]  

(4.16)
Chapter 4. Analysis of developmental pathways in mouse cortical development

State Diagram  

Distance Matrix  

Distance Matrix  

Distance Matrix  

Figure 4.3. Spectral decomposition: a toy model. (A) State diagram of example sublineages. Nodes represent cell states, arrows state transition probabilities. States are labeled according to 3 observed features: \( A = (1,0,0) \) (blue), \( B = (0,1,0) \) (green), \( C = (0,0,1) \) (orange), and \( \# = (?,?,?) \) (gray) for states with unobserved features. Initial states are depicted in dark gray. (B-D) State clustergram displaying the computed distance between every state pair with dimensions \( D = 23 \), \( D = 13 \), \( D = 11 \), and \( D = 7 \). The dendrogram on the left indicates hierarchical binary linkage of states. (E) Spectral label propagation on models, where each node is colored according to the estimated feature distribution (percentage of data represented in parenthesis). (F) Model error as percentage of the correct final cell states distribution for spectral decomposition (black) versus random model (gray). Standard deviation on 100 trials. HM, Homogeneous Markov model.

where \( P \) is the directed transition probability matrix, \( O \) the out-degree matrix, \( I \) the in-degree matrix, and \( I \) the identity matrix. \( U_k \) is the orthonormal matrix with last \( k \) eigenvectors arranged columnwise. Interestingly, rows of \( U_k \) provide an embedding for each vertex of the graph in a \( k \)-dimensional space, and distances between vertexes are proportional to the random-walk distance on the directed graph from unlabeled to labeled nodes (Figure 4.3A-D and Figure 4.4 for additional examples).

Each node in the graph is associated with a unique feature vector, such that a label corresponds to a linear combination of features. The feature matrix \( F \) is computed from the eigenvalues and \( F \), the matrix containing the feature distribution of observed terminal nodes.

\[
F = U_k F
\]
Chapter 4. Analysis of developmental pathways in mouse cortical development

Figure 4.4. Spectral decomposition on state diagram examples. (A) Original state diagrams of example sublineages. Nodes represent cell states, arrows state transition probabilities. (B) Spectral label propagation, where each node is colored according to the estimated feature distribution. (C) Dimensionality reduction. (D) State clustergram displaying the computed distance between every state pair. The dendrogram on the left indicates hierarchical binary linkage of states.

4.4.3.2 Discrete-time Markov branching process

Formally, we can describe a SD model as a branching process. Let the row vector $Z_t = (z_1, z_2, \ldots, z_n)$ denote the expected number of cells that we can find in each of the cell states $Q = (q_1, q_2, \ldots, q_n)$ at generation $t$. $Z_0$ specify the number of initial cells in each state. If we assume that progenitors divide at a uniform rate, the distribution of cells in each generation follows a discrete-time Markov branching process.

Every precursor give rise to two daughter cells and the state of each cell is determined by a time-invariant transition matrix $P$, such that the next possible states depend only on the current states and not on the past ones. The entry $p_{ij} \leq 1$ is the probability that a daughter cell in state $i$ transition to state $j$. The number of cells in each unique state $q_x$ can be computed by the following matrix operation:

$$Z_t = Z_0 P^t$$  (4.18)
Chapter 4. *Analysis of developmental pathways in mouse cortical development* 80

**Figure 4.5. Quantification of spectral decomposition performance.** Performance, that is the ability to recover the correct Markov branching process, was assessed on 100 lineages generated with 10 random 5-state models. Spectral decomposition assigns to each cell a unique class, which are compared to the original model class. (A) Confusion matrix of spectral decomposition on deterministic model (0 ± 0% classification error). (B) Confusion matrix of spectral decomposition on probabilistic model (20.3 ± 17.8% classification error). (C) Histogram of errors for deterministic ($p = 1$) and probabilistic ($p < 1$) models.

Given the limited ability of Markov models to describe usefully time-dependent processes, I use the following approaches: homogeneous (HM, probability $P$), the non-homogeneous (NM, age-dependent probability $P(a)$) and the time-dependent (TM time-dependent probability $P(t)$) Markov branching process. To compare two branching processes of a SD model with different dimensions, I compute the error percentage as the number of misclassified cells (cells in the wrong states) over the total number of cells produced at the end of the developmental process (Figure 4.3F).

### 4.4.3.3 Dimensionality reduction

The dissimilarity matrix (pairwise distance) encodes the degree of similarity between states, which is computed directly on the feature matrix $\mathcal{F}$. When the distance between two states is less than a certain threshold, they are considered equivalent and clustered into a single state. I define the model dimensionality $D$ as the number of states required to specify a particular lineage or a set of lineages. The goal is to find an optimal threshold such that the model of dimension $d << D$ can describe a maximal percentage of the original dataset while preserving it’s statistical properties (Figure 4.3E).

I validated the spectral decomposition by measuring the performance on a set of artificially generated lineages given an initial, known random model (Figure 4.5). In the deterministic case, the method recovers the hidden generative model in 100% of cases, whereas in the probabilistic case, success rate is at about 80%.
Figure 4.6. Spectral decomposition: cortical area 3 and 6. (A) State diagram of cortical lineages in area 3 and 6 combined. Nodes represent cell states, arrows state transition probabilities. Cell states are labeled according to the following classification: LayerVIb (blue), LayerVIa (sea green), LayerV (green), LayerIV (orange), LayerII/III (red), Glia (pink), and Unknown (gray). Initial states are depicted in dark gray. (B-D) State clustergram displaying the computed distance between every cell state with model dimension $D = 519$, $D = 158$, $D = 31$, and $D = 10$. The dendrogram on the left indicates hierarchical binary linkage of states. (E) Spectral label propagation, where each node is colored according to the estimated feature distribution (percentage of data represented in parenthesis). (F) Model error as percentage of the correct final cell states distribution for spectral decomposition (black) versus random model (gray). Standard deviation on 100 trials. HM, Homogeneous Markov model; NM, Non-Homogeneous Markov Model.

4.4.4 Spectral Decomposition of cortical lineages

4.4.4.1 Cortical cell lineages reconstruction

The recording of the whole lineage tree in mammals is cumbersome and technically very challenging. Only recently, appealing techniques have been developed to address this question. Here I adopt an alternative approach. The dynamics of cellular division and differentiation during development in murine cortical area 3 and 6 was quantified by Polleux et al. using the mitotic history technique, which selectively monitors the proliferative behavior of a defined cohort of precursor cells that are generated at a particular time point [18]. Cortical lineage trees are reconstructed by sampling from the experimentally determined cell distributions (Chapter 3).
Chapter 4. Analysis of developmental pathways in mouse cortical development

Figure 4.7. Comparison of spectral decomposition in cortical area 3 and 6. (A) State diagram of cortical lineages in area 3 and 6 combined. Nodes represent cell states, arrows state transition probabilities. Cell states are labeled according to the following classification: Layer VIb (blue), Layer VIA (sea green), Layer V (green), Layer IV (orange), Layer II/III (red), Glia (pink), and Unknown (gray). Initial states are depicted in dark gray. (B-D) State clustergram displaying the computed distance between every cell state with model dimension \( D = 519 \), \( D = 158 \), \( D = 31 \), and \( D = 10 \). The dendrogram on the left indicates hierarchical binary linkage of states. (E) Spectral label propagation on models, where each node is colored according to the estimated feature distribution (percentage of data represented in parenthesis). (F) Model error as percentage of the correct final cell states distribution for spectral decomposition (black) versus random model (gray). Standard deviation on 100 trials. NM, Non-Homogeneous Markov Model.

4.4.4.2 State diagrams of murine cortical neurogenesis

I analyzed 60 reconstructed lineages from area 3 and 6 of the mouse cortex, for a total of 3214 cell instances (1549 in area 3 and 1714 in area 6). Cell instances are labeled according to the following classification: Layer VIb, Layer VIA, Layer V, Layer IV, Layer II/III, Glia, and Unknown. This accounts for a state diagram composed of 6 terminal states, 765 unknown states in area 3, and 848 unknown states in area 6.

Spectral label propagation and dimensionality reduction was performed for both area 3 and 6 on the combined dataset, which allow to exploit the similarities between the two areas (Figure 4.6). For visualization purposes, I associate terminal fate labels with a unique color. How the color distributes across the network by label propagation is a visual indication of how the different progenitors are close to a given final differentiated state, and whether they are fate restricted.

While the model is computed over the lineage dataset for combined areas, it is easy to extract individual models for each area separately (Figure 4.7). Remarkably, the reduced models for area 3 and 6 are strikingly similar. This suggests that a unique model with minimal parametric differences is able to explain the generation of both areas. The
parametric difference between the model for area 3 and 6 is significant in a restricted
time-window in the first half of neurogenesis (E13-E14), which interestingly coincides
with the time at which thalamic afferents innervate the cortical plate.

The accuracy of the SD models for area 3 and 6 were assessed for the homogeneous (HM),
the non-homogeneous (NM) and the time-dependent (TM) Markov branching process
(Figure 4.8 and 4.9). In HM models, transition probabilities are time independent, and
at low model dimensions instantaneous cell output distributions are characterized by
long tails, which means that converge to the original distribution occurs only after a
great number of time steps. I introduce time dependence by considering age-dependent
probability distributions in NM models, that is each state is characterized by unique
outgoing transition probabilities and by a maximal number of possible self-replicative
divisions. Both areas are best described by the NM model, which is able to follow closely
the system dynamics and is an optimal trade off between model complexity (31 or 11
dimensions) and model accuracy (11% or 18% model error).

Figure 4.8. Spectral decomposition and cell distributions of cortical area 3. (A) State
diagrams with dimensions $D = 257$ and $D = 30$ of cell lineages in cortical area 3. Nodes represent
cell states, arrows state transition probabilities. (B) Mean cumulative number of differentiated cells
produced at every time step. (C) Mean instantaneous number of differentiated cells produced at
every time step. Dashed lines, original distribution; colored lines, model distribution; shaded area,
standard deviation. HM, Homogeneous Markov model; NM, Non-homogeneous Markov model; TM,
Time-dependent Markov model.
Figure 4.9. Spectral decomposition and cell distributions of cortical area 6. (A) State diagrams with dimensions $D = 292$ and $D = 30$ of cell lineages in cortical area 6. Nodes represent cell states; arrows state transition probabilities. (B) Mean cumulative number of differentiated cells produced at every time step. (C) Mean instantaneous number of differentiated cells produced at every time step. Dashed lines, original distribution; colored lines, model distribution; shaded area, standard deviation. HM, Homogeneous Markov model; NM, Non-homogeneous Markov model; TM, Time-dependent Markov model.

The 10 dimension SD model explains 82% of the data, and it is the most visually intuitive to reason about the logic behind the development of area 3 and 6. The black node represents an originally homogeneous population of precursor cells, which divide into subpopulations of precursor cells with different neurogenic potential: a small proportion of cells fated exclusively to infragranular layers (VIa, VIb, and V), cells fated very early on exclusively toward granular/supragranular layers (IV, II/III), and a big pool of heterogeneous proliferative cells that are less fate restricted. The 31 dimension SD model is more precise since it explains 89% of the data, but less intuitive to understand. A striking difference from the 10 dimension SD model is the presence of a starting heterogeneous population of cells composed of 2 different precursor types. Those precursors differentiate themselves in their fate restriction: one is more fated toward granular and supragranular layers, the other less fate specific.

SD models are a useful to capture statistically relevant differentiation pathways from the cell lineage dataset. Whether those state transitions are the result of intrinsic genetic
specifications, or of environmental signals, its however not possible to discern clearly. The results suggest that purely intrinsically specified mechanisms (HM model) are not good descriptors of the data, and there is the need of additional modulations of the transition probabilities (NM model), possibly achieved by inter-cellular communication, such as quorum-sensing mechanisms.

4.4.5 Transcriptional networks

4.4.5.1 Transcriptomic Atlas of mouse neocortical layers

I have so far showed label propagation on state diagrams with fictitious color codes for visualization purposes. The goal is however to (i) understand how the expression of different genes is distributed across the network, and (ii) extract from the activation and deactivation sequence mechanistic rules of how transcription factors (TFs) interact with each other.

Gene expression levels in cortical neurons were obtained from a transcriptome Atlas of cortical layers in the adult mouse somatosensory cortex (area 3) [19]. I used here calibrated layer enrichment probabilities (obtained by use of a naive Bayes classifier) provided by the authors. I identify 12 classes of transcription factors (TFs) with characteristic laminar patterning by k-means clustering. The reduced dataset is composed of 1751 TFs from an original dataset of 11411 gene probes, each one associated with its unique Ensembl identifier.

TFs feature vectors with expression probabilities are estimated by label propagation on the state diagram for area 3 (Figure 4.10). The expression patterns across precursor states gives a visual indication of TFs and the cell fate restriction. Very few TF Clusters show an almost homogeneous distribution across layers (5, 8, 9, and 11), whereas most of them displays a very characteristic distribution in different precursor subpopulations (1, 2, 3, 4, 6, 7, 10, and 12).

4.4.5.2 A transcriptional model of cortical development

The reduced state diagram provides both a description of states, state transitions, and transcription factor activation profiles in pools of identified cell subpopulations. We can simplify the model with a cartoon description of the major differentiation pathways
that were recovered (Figure 4.10A-B). Precursors can be subdivided into two big classes. Firstly, fate unrestricted cells, present very early in development, and certainly corresponding to the experimentally described radial glial cells (RGCs) in the ventricular
zone (VZ) [54, 55]. Secondly, a class of granular and supragranular layers restricted precursors, which seem to arise quite early during development, and increase in number as development progresses, whose presence has also been confirmed experimentally [106]. Those cells correspond to intermediate precursors, usually found in the subventricular zone (SVZ). The transition between RGCS and IPCs seems not to be as clear as currently believed, and we depict it as a plastic (and partially reversible!) transition between RGCs, basal precursors (BPCs) dividing at the basal border of the ventricular zone, and intermediate precursor cells (IPC). Plastic transitions correlate with graded expression of several transcription factors, for instance Cluster 1 and 12.

The inference of the gene steady-state distribution along lineage trees gives an insight into the relationship between gene activation and the lineage space. The model makes a prediction on which genes should be active, when, and possibly a list of causal links (logic rules) of how transcription factors interact (either directly or indirectly) with each other. In the next Chapters I address the question of how to encode states of transcriptional regulation into genetic network models.
4.5 Discussion

4.5.1 A novel approach for cell lineage data analysis

The key to understand the complexity of the adult cerebral cortex architecture, both in mouse and in the primate, resides in the developmental process. The patterns of intra- and interareal connectivity are a direct consequence of the number, specificity, timing and position of the neurons generated and the connectivity itself is used to shape the functional structure of the cortex. An interesting question is how the regulatory system, a network composed of more than thousand of genes, determines cell fate diversity and leads to a defined cytoarchitecture.

One way to address this question would be a systematic analysis of in vivo gene expression profiles of dividing cells during the whole developmental process. The capability to directly observe cell divisions in-vivo is quite limited in mammals – because of the opaque body and the tremendous amount of cells produced. On the other hand in-vitro slice preparations can be usually maintained in culture only for a limited period of time. Moreover, advances in this research area are severely hampered by the scarcity of cell markers and the impossibility to visualize the dynamics of more than few proteins in a single time-lapse microscopy experiment.

It is however possible to measure partial cell lineage data that contains sparse label information, corresponding to different observed features such as gene expression data, cell morphology or other attributes. A recent, promising technique to collect such data is clonal barcoding [149], where labeled retroviral vectors with random sequence tags (barcodes) are integrated into the host genome at specific developmental stages through cell transplantation. The clonal progeny of each cell can be tracked in time by single-cell sequencing and can be combined with micro-array gene expression analysis of the final cells.

Here, I propose a novel strategy based on spectral decomposition to analyze cell lineages. The proposed method applies to directed acyclic graphs and has the advantage that requires only semi-labeled data (cell lineage in which only terminal cell are labeled according to observed features) and could be applied to sparse databases of cell division patterns. The integration of information across the genealogical tree, even if only partially reconstructed, is a valuable representation that can be used to infer hidden
variables and models. Obviously, the quality of the classification strongly depends on the labels of the differentiated cells. A more extensive collection of features results in a finer classification. This implies that the model can be refined as more different cell fates are discovered.

A similar approach used to quantify lineage topologies as already been proposed [196]. In contrast to previous work, I don’t restrict the model to a deterministic one, but rather use probabilistic final state machines to describe developmental programs. The advantage is the ability to recover statistically significant recurring division patterns even in the presence of noise or stochastic processes: indeed cell states are clustered based on a similarity measure and not on exact matches of graph connectivities.

The drawback of spectral analysis is the requirement of fully connected lineages, since the current implementation does not support disconnected graph components. A second disadvantage is the current requirement of a clear separation between set of unobserved nodes, and set of observed nodes. An extension toward a more flexible implementation would be the combination of Hidden Markov Tree Models, as we have developed in [182], with the spectral approach.

It should be noted that in the proposed approach the model is not a set of differential equations that can be numerically solved, but rather a computational model, a state machine relating different cellular configurations to each other. Computational models have the advantage of being qualitative and thus particularly useful to test hypothesis without the need of huge parametric searches, as is the case of genetic algorithms.

4.5.2 Hidden generative models

Observed random variables are usually the results of an hidden generative model. In order to test the spectral decomposition performance, I have analyzed artificially generated dataset of cell lineages (for which the generative model is known) and showed that the underlying model can be successfully recovered. The algorithm selects a list of states and state transitions that can be recursively used to specify the sequences of developmental events.

I show that hidden models are recovered with 100% success rate in case of deterministic models. The reason is the fact that the recovery of a deterministic generative model requires only geometrical operations, which are implicit in the spectral decomposition.
In contrast, success rate for probabilistic models lies at about 80%. The reason of this drop is not only classifications, but also due to the existence of multiple equally likely solutions, or the existence of a more compressed solution than the hidden model itself.

In overall, the performance of spectral decomposition over random model selection is evident, and justifies the use of this algorithm for analyzing cell lineage division patterns.

4.5.3 Computational models of cortical neurogenesis

I apply the described reconstruction of a state machine diagram to cortical lineages from area 3 and 6 of murine developing cortex. Notably, the spectral decomposition was performed on the combined cell lineage dataset of both areas, such as to identify similarities and differences between generative models for one area versus the other. I recover models for both areas at dimensionality 31 and 10, where an optimal trade-off between model complexity and model accuracy is found. Surprisingly, the recovered models for area 3 and 6 display many similarities, and few significant differences, suggesting that a unique model with few control parameters is able to generate both cortical structures.

The obtained models are remarkable for their predicting abilities. While obtained an combination of cell lineages and transcription factor expression only, the models predict the presence of an heterogeneous population of precursors pool regarding their fate potential: multipotent progenitor cells coexist with a specific population of cells restricted to either infragranular or granular/supragranular layers, as recently proved experimentally [106]. Moreover, the models relies on plastic progenitors, which have a limited (but statistically significant) ability to revert to less differentiated states. Cell fate decisions correlate with predicted transcription factor expression levels, and plastic decisions are probably achieved by exploiting the noise in the expression machinery. An even better model could be obtained if we would be able to apply it on fully experimentally recorded cell lineage trees.

Both models were tested for their ability to approximate the original cell distributions: at dimensions 31 approximation error were 30% for the homogeneous Markov Branching model (HM), and 10% when considering an age dependent Markov Branching model (NM). In the HM case, the Markov process correctly converges to the original distributions only after a considerable number of time steps, and thus is not able to capture the dynamics of the process. The ability of the NM model to approximate both steady
state and dynamic distributions lies in the fact that state age dependence eliminates the log-normal tail typical of probabilistic homogeneous processes.

In biological terms, we interpret this finding as the demonstration that developmental processes cannot be modeled by homogeneous statistical processes, but need additional parameters to control the transition probabilities. Whether this mechanism is achieved by an internal aging mechanism on single states, or by quorum sensing mechanisms that synchronize the behaviors of cells, we cannot say.

The inferred sequences of states, and state transitions, with the associated feature distributions, allows to make some predictions about the underlying transcriptional network. Although I am not able to infer direct interactions between transcription factors, it is possible to identify patterns of commonly regulated transcription factors. I use this information to reconstruct a plausible transcriptional network models in the following Chapters.
Chapter 5

A genetic language for developmental programs

The biological information of how organisms and tissues are constructed is stored in the genome. The genetic information is identical in each cell, yet cells can assume specialized functions by expressing different patterns of gene products. The gene expression profiles follow specific trajectories during development and are responsible for the specialization of different neurons and their diverse localization in the cortex.

In this Chapter I attempt to discern the computational primitives of the genetic code. For this purpose, Section 5.3 is dedicated to the investigation of an appropriate model: I consider basic concepts from reaction kinetics and cis-regulatory structures to understand the rules of gene regulation. I propose a symbolic framework for transcriptional networks able at the same time to describe the continuous dynamics of translational processes and the logic of interactions between gene products, an aspect that is not explicitly present in previously proposed models in the literature.

The framework is used in Section 5.4 to investigate the role of genetic networks in controlling developmental processes. From an analysis of the gene’s computational space I extract some basic motifs that can be used as modules to reverse engineer regulatory networks with a predefined behavior.
5.1 Introduction to gene regulatory networks

5.1.1 The biological principles of the genomic regulatory code

An organism’s architecture and function is the result of the integration over space and time of sequential developmental processes. The production of different cell types requires the coordination of five primary processes: cell growth, cell replication, cell division (mitosis), cell differentiation, and morphogenesis. The orchestration of these processes is achieved through many layers of regulation [177] (Figure 5.1), with increasing diversity of components the farther away from the process core.

At the outermost layer proteins are responsible for the cell morphology and its functional behaviors. Proteins, which represents the nuts and bolts, sensors and microprocessors of the cellular machinery, are designed to fulfill specialized tasks, such as structural support, enzymatic activity, gene expression regulation, etc. Deeper into the regulatory machinery we find DNA-binding regulatory proteins (transcription factors), which dynamically define the progression through different cell activity states by regulating the gene expression profile of each cell [197].

At the core is the genome, which encodes information that directs cellular functions. The genome is the sum of coding and non-coding sequences. Physically, the genome consists of a sequence of modular DNA elements, or genes, that interact with transcription factors. During gene expression, the information encoded by a gene is transcribed into mRNA and translated into a functional protein. Genes are indirectly responsible for the control of most cellular processes, like cell proliferation and growth, genome replication at cell division, response to environmental signals, and cell differentiation.

A gene is composed of different parts: a coding region, which codes for the primary amino acid sequence of a protein, and a regulatory region, which determines under which conditions the gene is expressed. Gene expression is regulated by a core promoter region upstream of the coding region, where the transcriptional complex forms and initiates the transcription of the coding sequence. Transcription initiation is additionally regulated by transcription factors that bind to cis-regulatory elements, that is elements located on the same chromosome. Binding of different transcription factors can initiate or inhibit the mRNA transcription from the DNA template. In prokaryotes the regulatory region is usually contiguous and upstream of the coding region; in eukaryotes regulatory
elements can be found at a considerable distance, either upstream or downstream of the coding region. Distal regulatory regions are likely to interact with the formation of the transcription complex by looping of the DNA strand, such that regions loaded with transcription factors bend and are brought into contact with the basal transcription machinery, composed of promoter-proximal tethering elements that recruit distal enhancers, and insulator DNAs, which prevents cross-talking between neighboring genes.

A gene regulatory network (GRN), also referred to as transcriptional regulatory network, is a collection of genes that interact with each other through their protein expression. The network can be represented with a directed graph, where a node is a gene and a directed edge between a gene A and B corresponds to the regulation of gene activity B by the product of gene A. The interactions regulate the rates at which genes in the network are transcribed and thus define the dynamics of the system [198].

In prokaryotes clusters of genes with related structural or functional properties are grouped into operons [199], units of genomic DNA controlled by a single promoter. Operons are usually translated into polycistronic mRNA (a single mRNA molecule that codes for more than one protein) and thus guarantee that related proteins are translated into proteins at proximal spacial positions. Although operons also occur occasionally in eukaryotes [200], combinatorial expression of functional genetic units is greatly increased in the latter. Eukaryotic genes are composed of exons, which mainly constitute the protein coding sequence, and introns, which are cut out from the mRNA sequence in the process of alternative splicing. The different possible recombinations of exons provide a mechanism, by which single genes can code for multiple protein isoforms [201].
5.1.2 Computational aspects of gene regulatory networks

5.1.2.1 Cis-regulatory modules constitute the computational units

Genetic regulation can be viewed as a computational process with inputs, input processing, and outputs. Transcription factors that bind to cis-regulatory elements represent individual inputs to the central computational core, the transcription machinery. The regulatory elements constitute a reading device about the current regulatory state of the cell. This information is processed by regulating (directly or indirectly) the transcription rate of different coding regions, which determine whether individual genes are silent or active. The outputs are the different possible isoforms of proteins encoded in the DNA sequence, with different isoforms being expressed under different conditions.

The computational units are cis-regulatory modules (Figure 5.2A). The input of each cis-regulatory module is represented by the occupancy status of the core promoter and regulatory elements, on which combinatorial integration is performed. The output, which is causally dependent on the inputs, is usually a protein or a set of protein isoforms (Figure 5.2B). Proteins that do not bind to the DNA are the leaves of the transcriptional network, while transcription factors constitute the links in the regulatory network that connect the output and inputs of cis-regulatory modules.

Since the distribution of cis-regulatory modules across the genome is mostly sequence-independent, the term genetic network refers to the topology of the links between genes encoding transcription factors, their corresponding proteic products, and the binding sites on the genome. For simplicity, I will refer sometimes to genes regulating other genes, where it is implicitly understood that this regulation happens through the production of transcription factors and their binding behavior.

Highly linked networks of genetic elements consist of a relative small set of genes, in the range of ten to hundred. In contrast to conventional logic circuits, where information processing is performed on an high number of simple and fast interconnected units, deep architectures built of many layers of biochemical cascades have strong limitations because of slow computational speed and potential error propagation. Genetic networks concentrate their computational power in the regulation of individual genes and on a limited number of signal processing layers, and they have proved capable of slow but complex input integration, noise filtering and parallel asynchronous computation.
5.1.2.2 Coding properties of transcriptional networks

In metazoans, about 5-10% of the total coding sequence is dedicated to transcription factors, which are required to regulate the gene expression rate of all genes [197, 202]. The number of encoded transcription factors in different species scales with the number of genes, whereas the complexity of organisms is not correlated with gene number but with greater elaboration in cis-regulatory modules and a more sophisticated transcriptional machinery [202].

The structure and function of transcriptional networks are dictated by physical constraints. The binding of transcription factors to the corresponding regulatory elements follows the laws of thermodynamics (stability behavior) and kinetics (dynamical behavior), according to which transcription factors diffuse in the intracellular medium by facilitated diffusion [203], and assemble at the site of transcriptional initiation.

Stochastic binding has several consequences. Firstly, the low binding specificity of many transcription factors leads to the interaction with multiple proteins, and a substantial number of combinatorial interaction possibilities [204]. Secondly, the total number of transcription factors is limited by the ability to recognize distinct binding sites. Binding sequences are usually palindromic 5-6 base pair repeats, which result into $4^6/2 = 2048$. 

**Figure 5.2. Model of transcriptional control.** (A) A typical eukaryotic gene consists of a coding region (Gene), a core promoter region (P), and cis-regulatory sequences (E). (B) Schematic diagram of transcriptional control, where genes are translated into functional proteins P or transcription factors $T_i$ (black arrows), which influence each others expression through binding to regulatory regions (dashed arrows). (C) Coding scheme for binding sequences for different transcription factors. Regulatory elements (dots) correspond to binding sequences, the binding abilities of transcription factors spans multiple regulatory elements (shaded disks). Black dots outside the transcription factor range represent potential regulatory sites to which none of the three transcription factors binds.
Table 5.1. Genes and transcription factors activation in function of transcriptional network depth. Number of potentially expressed genes $G(n)$ and transcription factors $T(n)$ in function of transcriptional network depth $d$ and number of binding sites $n$ per cis-module. I consider the case of the most simple network topology, a directed tree. Bold numbers represent networks that activate more than 30'000 genes and 2'000 transcription factors.

<table>
<thead>
<tr>
<th>Depth $d$</th>
<th>$G(n=2)$</th>
<th>$T(n=2)$</th>
<th>$G(n=3)$</th>
<th>$T(n=3)$</th>
<th>$G(n=4)$</th>
<th>$T(n=4)$</th>
<th>$G(n=5)$</th>
<th>$T(n=5)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>3</td>
<td>9</td>
<td>4</td>
<td>16</td>
<td>5</td>
<td>25</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>7</td>
<td>27</td>
<td>13</td>
<td>64</td>
<td>21</td>
<td>125</td>
<td>31</td>
</tr>
<tr>
<td>4</td>
<td>16</td>
<td>15</td>
<td>81</td>
<td>40</td>
<td>256</td>
<td>85</td>
<td>625</td>
<td>156</td>
</tr>
<tr>
<td>5</td>
<td>32</td>
<td>31</td>
<td>243</td>
<td>121</td>
<td>1024</td>
<td>341</td>
<td>3125</td>
<td>781</td>
</tr>
<tr>
<td>6</td>
<td>64</td>
<td>63</td>
<td>729</td>
<td>364</td>
<td>4096</td>
<td>1365</td>
<td>15625</td>
<td>3906</td>
</tr>
<tr>
<td>7</td>
<td>128</td>
<td>127</td>
<td>2187</td>
<td>1093</td>
<td>16384</td>
<td>5461</td>
<td>78125</td>
<td>19531</td>
</tr>
<tr>
<td>8</td>
<td>256</td>
<td>255</td>
<td>6561</td>
<td>3280</td>
<td>65536</td>
<td>21845</td>
<td>390625</td>
<td>97656</td>
</tr>
<tr>
<td>9</td>
<td>512</td>
<td>511</td>
<td>19683</td>
<td>9841</td>
<td>262144</td>
<td>87381</td>
<td>1953125</td>
<td>488281</td>
</tr>
<tr>
<td>10</td>
<td>1024</td>
<td>1023</td>
<td>59049</td>
<td>29524</td>
<td>1048576</td>
<td>349525</td>
<td>9765625</td>
<td>2441406</td>
</tr>
</tbody>
</table>

In human, automatic sequence annotation predicts about 1’000-2’000 transcription factors [206], which have the control over about >30’000 protein-coding genes. We can represent the regulatory network as a graph of nodes (genes) connected by directed links. Nodes with outgoing links are transcription factor genes, while nodes with incoming links only represent genes coding for other proteins. In the most simple case, in which each transcription factor can activate on average $n$ genes without overlaps and in a feed-forward network (no loops), the graph becomes a directed tree, where the leaves are the 30’000 genes, and internal nodes are transcription factor genes. A simulation of the number of activated terminal genes and transcription factors in function of the tree depth is given in Table 5.1: for instance, a transcriptional network with average connectivity $n = 4$ can activate 30’000 genes with a minimal network depth between 7-8 layers, and require the activation of more than 10’000 transcription factors! Overlap of binding sites, combinatorial integration of inputs and feed-back loops in the network, are responsible for the drastic reduction in the number of transcription factors needed.
5.1.3 Structural properties of transcriptional networks

5.1.3.1 Modularity

Albeit combinatorial transcription assures more specificity in the protein interactions and greatly reduces the number of transcription factors needed, at the same time it reduces the number of control points that can steer the network activity. This limitation has interesting implications on the topology of transcriptional networks: firstly, individual transcription factors can activate many different genes, usually coding for proteins with similar expression patterns and functions [177]. Secondly, genes are reused in different contexts, for instance many of the genes encoded in the genome are used at some point during development to execute unrelated functions [207].

Undoubtedly molecular programs are likely being carried out by modules [208, 209], that is network parts whose function is at least partially isolated from other modules either by chemical specificity or chemical/spatial isolation\(^1\). Modularity at the level of molecular interactions has many examples: for instance, biochemical pathways are usually isolated by high specificity in the proteins interaction, whereas protein synthesis relies on the subcellular localization of ribosomes inside the cell. At the level of genetic regulation, an outstanding example is the cell cycle [210, 211].

Modularity at the genetic level (activation of related genes) has to be propagated to the functional level (activation of the related molecular machinery, that is ribozymes and proteins). Physically the construction of modular molecular components is limited by the free diffusion of molecules in the cellular medium, and formation of protein assemblies follows the laws of kinetics and thermodynamics. This limitation is cleverly bypassed in eukaryotes by the process of alternative splicing [201], by which different combinations of exons code for different proteins isoforms. In this perspective exons represent the fundamental genetic information unit under the control of a common cis-regulatory module, and map the combinatorial and modular properties of gene regulatory networks to the proteome.

The empirical study of transcriptional networks and mathematical tools from graph theory [212] have been used to investigate the effect of particular network topologies

\(^1\)I refer here to modules as a subset of genes whose function is separable from those of other genes. This term is not to be confused with cis-regulatory modules, which refer to the ensemble of regulatory elements that regulate the expression of a single gene.
on the computational capabilities of the network. A Erdős-Rényi random graph with \( n \) nodes is a graph with randomly connections according to a probability \( p \), and the expected connectivity degree given by \( \left( \begin{array}{c} n \\ 2 \end{array} \right) p \). In contrast to random graphs, transcriptional networks are statistically inhomogeneous. A widely used simplified model for genetic networks have been scale-free graphs, which are characterized by a power law distribution of connectivity degrees and random in all other aspects. Scale-free graphs embody the potential of transcription factors to regulate a multitude of downstream target genes. However, there is increasing evidence that this view is flawed by observation biases due to the inherent variability of biological data, and artifacts in subnetworks sampling.

5.1.3.2 Motifs

While the degree and significance of modularity of components in transcriptional networks is still controversial, there is growing evidence for a higher abundance of interaction motifs in cellular networks. Motifs are small connected subgraphs whose nodes are connected to each other by a specific wiring diagram, and whose occurrence frequencies are higher compared to a randomized version of the same network. For instance, the directed triangle motif known as feed-forward loop is overrepresented in many transcriptional networks.

Motifs are very small subgraphs usually composed of 3 or 4 nodes. As the number of combinatorial subgraphs increases exponentially with the increase of numbers of genes in the subgraph, the study of larger motifs becomes intractable. The only alternative is to study how small motifs are integrated into larger networks, and assign to those topological structures a functional role. Since many molecular components of motifs interact with nodes that lies outside the motif, they often overlap with each other and belong to multiple clusters. For instance, in the theoretical case of scale-free networks, topological motifs combine into a hierarchical network of components: individual nodes are firstly grouped into motifs, and motifs aggregate into larger modules, whose interconnection forms the final regulatory network.
5.1.3.3 Path length and redundancy

An additional property of transcriptional networks consists in the existence of short path lengths between related genes, which is a direct consequence of the limited number of available transcription factors and thus of the limited computational depth (number of computational layers that separate input from output). Short path lengths ensure fast reaction to network perturbations, as well as limited noise amplification. Robustness to random variations in the protein concentrations is also achieved through path redundancy (the availability of multiple paths between a pair of transcription factors [221]) in the wiring of components.

5.1.3.4 Composability

The logic composition of transcriptional networks is largely an unsolved puzzle and, at a first approach, its sheer complexity may seem incompatible with rational design: molecular networks are characterized by the cooperative re-arrangement and modification of preassembled protein complexes, extensive pathway cross-talk, expression noise, and finally unspecific binding. Despite the complexity of gene regulation, the basic rules governing molecular interactions are well understood, and make a good foundation for the artificial design of enzymes and genetic circuits [222].

In engineering terms, the logic of genetic circuits is represented by the stitching and wiring of linear pieces of DNA. Transcriptional networks can be seen as a set of interconnected small modular networks: each module or motif consists of few input nodes, some internal nodes for computation, and few output nodes [223]. Isolated modules can be connected by sparse links between outputs and inputs, which guarantees functional autonomy and ability of selective signal transduction between modules. This topology is less optimal than in highly interconnected networks such as neural networks [224], since modules limit the number of possible connections in the network. However, they display robustness to noise, component failure tolerance and are advantageous from an evolutionary point of view. New components can be designed by combining already existing modules [225] as pieces of LEGO® can be rearranged in always new constructions.
5.1.4 Gene regulation and development

5.1.4.1 What is the link between genes and development?

All multicellular organisms develop from a single embryonic cell that goes through the processes of growth, internal replication, cell division, and terminal differentiation. During cell growth, cells increase in size. At the time of cell division the entire genome is replicated to produce two identical copies, where each of the daughter cells inherits one complete copy of the genome, and a portion of the molecular machinery of the mother cell. The signals that instruct cells when to switch from growth to replication, from replication to division, and when to terminally differentiate are controlled by the specific expression of genes.

The complete description of the regulatory circuit is embedded in the DNA sequence and in the rules that govern molecular interactions. How does the compressed genetic information unfold itself in order to control all the basic behaviors of living cells? And more specifically, how does transcriptional control regulate developmental processes and guide the organism construction? As with any code, it can be understood by looking at the sequence of processes that are used to translate the encrypted information in an executable program.

Development results from the execution of genetic programs, that is the precise temporal and spatial control over the expression of transcription factors and functional proteins (gene expression profile). Transcription factors are the key mechanism that regulate the transition from one expression profile (or state) to the other, since their concentration determines successive regulatory states. The regulatory control over many genes is achieved by the specification of alternative choices of states, and the chosen regulatory state depends on the cell environment.

Genetic regulation of development is complicated by the requirement of multicellularity: the state of gene activation in a cell depends not only on it’s local environment, but also on the decisions taken by neighboring cells. While most of the regulation involves intracellular mechanisms, communication between cells is essential to coordinate the unfolding of development in different parts of the organism. Therefore we categorize signaling mechanisms into three different classes: (1) by internal expression of transcription factors; (2) by direct physical contact between neighboring cells, where messages
are transmitted either through membrane receptors or gap-junctions; and (3) by diffusion of morphogens, a class of signaling molecules used to provide cells with positional information [226, 227].

The fate of a proliferative cell and the fate of its progeny are strongly correlated with the internal gene expression profile. As development progresses, the number of states that a cell can reach becomes progressively restricted toward terminally differentiated phenotypes. We can identify two key processes, that are essential for the distribution of different phenotypes into the generated cells, as described in the following subsections.

5.1.4.2 *Asymmetric distribution of cell fate determinants*

Regulation of self-renewal and the establishment of cell diversity can be controlled by the symmetry of cell divisions. Asymmetric divisions are the result of an asymmetrical partitioning of molecules, such as cell determinants, to the daughter cells, and often lead to differing behaviors, whereas symmetrical cell divisions give rise to equivalent daughter cells in terms of the constitutive components.

Asymmetric divisions arise from the asymmetric localization of molecules inside the cells by phosphorylation, and attachment to the membrane or differential diffusion. The relative orientation of the mitotic spindle determines the division plane and how subcellular components are partitioned among the daughter cells, as well as the final daughter cell size. The balance between symmetric and asymmetric divisions has profound effects on the number and types of cells produced, since symmetrical division amplifies the precursor pool, whereas asymmetrical divisions usually lead directly or indirectly to differentiation.

5.1.4.3 *Epigenetic control of cell fate specification*

Complex epigenetic regulation – modifications of the genetic information other than by changes of the nucleotide sequence – orchestrates the spatial and temporal expression of genes at the level of chromatin (DNA packaging, methylation, histone acetylation and deacteylation) and post-transcriptional regulation. Epigenetics provides a useful mechanism to confer memory states on top of the coding sequence. For instance, environmental signals can provoke the silencing or de-silencing of genes, and those states will persist even after the perturbation has vanished.
In pluripotent embryonic stem cells, key developmental genes are kept silent or minimally expressed by reversible trimethylation at two different amino acid positions on histone H3 (H3K27 and H3K4) \[228\]. The repression is easily reversible since trimethylation can be removed in response to specific developmental cues \[229\] (poised state). As cells progress toward more differentiated states, transcription of lineage-specific genes is increased (active state) and alternative fates are permanently silenced through direct DNA methylation at CpG dinucleotides \[230, 231\] (stably repressed state).

### 5.1.5 The choice of the modeling formalism

Different approaches have been used to investigate the dynamical properties of genetic networks \[232, 233\]. The first approach consists in a detailed description of molecular interactions by means of non-linear differential equations \[234, 235\] and kinetic models of transcriptional dynamics in specific molecular networks. While this method is particularly useful for specific networks, it usually does not provide insight into the key dynamical aspects because of the high number of parameters. Indeed the determination of quantitatively precise interactions between gene products and genes is a notoriously difficult task. A second method is the use of simple logical models such as discrete Boolean networks \[20, 236, 237\], piecewise-linearized models \[237, 238, 239\], probabilistic Boolean networks \[240\], Petri nets \[241\], and hybrid models \[241\]. What those alternative approaches have in common is that they focus more on the design principles and less on specific biological interaction parameters.

While different modeling formalisms describe gene regulatory networks at different levels of abstraction, gene regulation is fundamentally a dynamic process characterized by time-changing gene expression profiles. As such, the behavior of gene regulatory networks can be described in terms of states, and state transitions. A state corresponds to a unique or a set of gene expression profiles, and to each state we can associate state transitions that describe the system evolution in time. Notably, genetic networks are characterized by multiple stable states or attractors, toward which the systems tends over time depending on its initial conditions. It has been proposed that attractors correspond to cell types \[236\] or alternate cell fates \[21, 242\]. The analogy between cell fates and attractor states provides a useful concept for understanding the relationship between the dynamics of the system and its function.
5.2 Objectives of this Chapter

Abstract models of gene regulatory networks (GRNs) are essential for the analysis, experimental manipulation and, most fundamentally, for the comprehension of often overwhelming complex molecular mechanisms. The optimal level of abstraction is dictated by the details and features that the model should be able to capture, and the questions to address.

Given the great variety of existing models, why do I need a new one? Here I propose a symbolic framework for transcriptional networks able at the same time to describe the translational processes and the logic of interactions between gene products, an aspect that is not explicitly present in other models in the literature. I will then use this formalism to investigate and encode regulatory networks to control developmental processes.

Objectives for this Chapter are:

- **Mathematical description.** Give a formal mathematical description of gene regulation and transcription factor interactions as described by the laws of thermodynamics and kinetics.

- **Formal genetic language.** Provide a higher level language describing in a compact and unambiguous way complex regulatory transcriptional network models and compare the usefulness of the model with previous work on the design of genetic networks.

- **Model for developmental process.** Use the proposed language to investigate the logic of genetic control of developmental processes.
5.3 Methods: Mathematical models for transcriptional networks

5.3.1 Reaction kinetics of the binding of a single transcription factor

I derive a generalized model of gene regulation directly from reaction kinetics \[243\]. The chemical equation for a gene controlled by a single regulator is described by the following three-step model \[244\]: a gene $G$ binds to a single transcriptional inducer $T$, which promotes the synthesis of mRNA $R$ and translation into proteins $P$. Thus the flow of information from the DNA to proteins is a non-reversible process.

$$ T + G \overset{k_1}{\underset{k_{-1}}{\rightleftharpoons}} T \cdot G \overset{k_2}{\rightarrow} R \overset{k_3}{\rightarrow} P \quad (5.1) $$

where $G$ is a unbound gene, and conversely $T$ is an unbound transcription factor. $T \cdot G$ is the gene-transcription factor complex, which results into the transcription of mRNA $R$ and eventually translation into a protein $P$. $k_1, k_{-1}, k_2, \text{ and } k_3$ are the constant reaction rates that dictate the direction of the reactions.

The dynamics of the system is described by a simple set of ordinary differential equations:

$$ \begin{align*}
\frac{d[T \cdot G]}{dt} &= k_1[T][G] - k_{-1}[T \cdot G] - k_{T \cdot G\text{deg}}[T \cdot G] \\
\frac{d[R]}{dt} &= k_2[T \cdot G] - k_{R\text{deg}}[R] \\
\frac{d[P]}{dt} &= k_3[R] - k_{P\text{deg}}[P]
\end{align*} \quad (5.2) $$

which takes into consideration the degradation rates $k_{T \cdot G\text{deg}}$ of $T \cdot G$, $k_{R\text{deg}}$ of $R$, and $k_{P\text{deg}}$ of $P$. The total gene concentration is given by $[G_{\text{tot}}] = [G] + [T \cdot G]$, whereas the total transcription factor concentration is given by $[T_{\text{tot}}] = [T] + [T \cdot G]$.

I take the following assumption: the concentration of the intermediate complex $[T \cdot G]$ does not change over time and is at thermodynamical equilibrium. Indeed it is commonly assumed that the dissociation rates are much higher than any other reaction, which gives a fast relaxation time of the activated complex concentration. Under this assumption we can consider $d[T \cdot G]/dt = 0$ even if the system is far away from steady state. The equations reduce to the Michaelis-Menten model of enzyme kinetics.

$$ [T \cdot G] = [G_{\text{tot}}] \frac{T}{K_M + [T]} \approx [G_{\text{tot}}] \frac{T_{\text{tot}}}{[T_{\text{tot}}] + K_M} \quad (5.3) $$
with:

\[ K_M = \frac{k_{-1} + k_T G_{deg}}{k_1} \]  (5.4)

\([T \cdot G]\) is governed by a non linear function on \([T]\), whereas RNA and protein synthesis are linearly dependent on the concentration of the intermediate complex. Note that the assumption \([T] \approx [T_{tot}]\) is valid only at high concentration of \(T\).

Regulation of protein expression is a complex process, which can be modulated at many different levels. On a coarse-grained level of description we can assume that the dynamics of the mRNA molecules is much faster than that of proteins [245, 246] and thus \(R\) concentration reaches the steady state quickly in comparison to protein levels. Since extensive studies on circuits modeling both mRNA and protein levels agree with this simplification, I consider transcription and translation as a single process:

\[ [T] + [G] \xrightleftharpoons[k_{-1}]{k_1} [T \cdot G] \xrightarrow[k_2]{k_2} [P] \]  (5.5)

A more realistic model of binding of a transcriptional inducer \(T\) to the gene promoter region takes in consideration transcription factors composed of \(m\) multiple subunits (Figure 5.3A). According to the conservation law the equation becomes:

\[ [mT] + [G] \xrightleftharpoons[k_{-1}]{k_1} [mT \cdot G] \xrightarrow[k_2]{k_2} [P] \]  (5.6)

and:

\[ [T \cdot G] = [G_{tot}] \frac{T_m^m}{K_m + T_{free}^m} \]  (5.7)

or the fractional saturation (concentration of bound gene):

\[ \frac{[T \cdot G]}{[G_{tot}]} = \frac{T_m^m}{K_m + T_{free}^m} \]  (5.8)

In the case of a gene regulated by a repressor \(T\) composed of \(m\) multiple subunits (Figure 5.3D) I derive the concentration of an unbound gene, which is proportional to the rate of protein expression:

\[ [mT \cdot G] \xrightleftharpoons[k_{-1}]{k_1} [mT] + [G] \xrightarrow[k_2]{k_2} [P] \]  (5.9)

and:

\[ \frac{[G]}{[G_{tot}]} = \frac{1}{1 + \frac{T_m^m}{K_M}} \]  (5.10)
Since the total gene concentration does not change over time, I consider from now on the partition function $[T \cdot G]/[G_{\text{tot}}]$ or $[G]/[G_{\text{tot}}]$, which represent the binding, respectively unbinding probability. The partition function depends only on the free transcription factor concentration $x = [T]$, the dissociation constant $\theta = K_M$ and the cooperativity constant $m$. The equation for transcriptional activation (5.8) takes the form of the well known Hill function $Z$ in function of the concentration $x$.

$$Z(x, m) = \frac{x^m}{\theta^m + x^m} \quad (5.11)$$

At high levels of $x$ the Hill function reaches saturation, since the probability of binding cannot exceed 1. The Hill constant $\theta$ at 50% response corresponds to the substrate concentration, since for $Z = 1/2$ we have $\theta^m = x^m$. The Hill coefficient $m$ is a positive integer which describes the cooperativity binding between different transcription factors and determines the steepness of the function: when $m > 1$, $Z$ is a sigmoidal function and when $m \to \infty$, $Z$ approximates the Heaviside step function. Remarkably for $m = 1$ the equation reduces to the Michaelis-Menten model for enzymatic reactions, which has been
Figure 5.4. The Hill function. The Hill function $Z([T])$ describes the transcriptional activity of a gene in function of binding of a transcriptional inducer $T$. (A) Dependence on cooperativity coefficient $m$ (from thick to thin line: $m = 1$, $m = 2$, $m = 4$, $m = 10$ and with $\theta = 1$). (B) Dependence on dissociation coefficient $\theta$ (from thick to thin line: $\theta = 3$, $\theta = 1$, $\theta = 0.5$, $\theta = 0.1$ and with $m = 4$). (C) Dependence on basal transcriptional activity, represented by the independent binding of an additional transcription factor with steady concentration $c$ (from thick to thin line: $c = 1$, $c = 0.5$, $c = 0.33$, $c = 0$ and with $m = 4$, $\theta = 1$). (D) Dependence on binding site saturation, represented by the binding of a second transcription factor to the same binding sequence with constant concentration $c$ (from thick to thin line: $c = 1$, $c = 0.5$, $c = 0.33$, $c = 0$ and with $m = 4$, $\theta = 1$).

applied to the simulation of transcriptional regulation [247]. For a repressor (5.10), the Hill input function is a decreasing S-shaped curve $\overline{Z}$ and can be interpreted as inhibition of protein synthesis for a constitutive active gene (Figure 5.3E):

$$\overline{Z}(x, m) = 1 - Z(x, m) = \frac{\theta^m}{\theta^m + x^m}$$  \hspace{1cm} (5.12)

Note that the binding probability of a repressor can be written as the binding of an enhancer with negative cooperativity $m < 0$:

$$Z(x, -m) = \frac{x^{-m}}{\theta^{-m} + x^{-m}} = \frac{x^{-m} \cdot \theta^m x^m}{\theta^{-m} + x^{-m} \cdot \theta^m x^m}$$

$$= \frac{\theta^m}{\theta^m + x^m} = \overline{Z}(x, m)$$
The function $Z$ (and its complement $\overline{Z}$) describes the binding kinetics of a single transcription factor to a corresponding DNA strand. The influence of $m$ and $\theta$ values on the shape of the Hill function are depicted in Figure 5.4. Note that a change in the incoming input weights $w_{ij}$ is equivalent to an inverse proportional change in the affinity constant $\theta_j$ for the transcription factor $j$ to its binding sequence.

$$Z_{ij} = \frac{(w_{ij}x_j)^m}{\theta^m + (w_{ij}x_j)^m} = \frac{(\frac{\theta}{w_{ij}})^m}{x_j^m + x_j^m}$$

(5.13)

Since regulatory sequences often have multiple binding sites, I use combinations of Hill functions to model multiple inputs and generalize this formalism to an entire transcriptional network.

5.3.2 The rate equation

Genetic regulation consists of networks of genes and their proteic products, or transcription factors (TFs), which can influence each other's expression over time by binding onto specific gene regulatory regions (Figure 5.5). The network dynamics depends on the connections and the update rules for each gene. The regulation of the concentration $x_i$ of the transcription factor $T$ encoded by gene $x_i$ follows the rate equation:

$$\dot{x}_i = k_1 F_i(x_i) - k_2 G_i(x_i) x_i$$

(5.14)

I use bold letters to indicate vectors (or matrices) and normal letters to represent scalar values. $x_i$ stands for the concentration of a gene product and $X$ for the respective gene, whereas the elements of the vector $x_i = \{x_1, x_2, \ldots, x_n\}$ are the concentrations of internal transcription factors influencing gene $X_i$ or eventually an external input.

The first term of the equation describes the rate of gene synthesis with maximal rate $k_1$. The function $F_i$ expresses how the synthesis rate of the protein encoded by a gene depends on the total concentrations $x_i$ of other transcription factors, either intra- or extracellular. Different models can be implemented according to the form taken by $F_i$. The choice of the mathematical formalism has been usually dictated by the nature...
Figure 5.5. Transcriptional network. (A) A typical eukaryotic gene consists of a coding region (Gene), a core promoter region (P), which is the minimal sequence of bases required to properly initiate transcription, and a cis-regulatory sequence (E), a proximal sequence upstream of the gene that contains primary regulatory elements. Cis-regulatory modules receive and process informational inputs in form of transcription factors. Function $Z$ describes the binding probability of transcription factor $T$ or any other transcription factors to the $E$ sequence. Function $F$ computes the protein $T$ synthesis rate given a combination of enhancing or repressing transcription factors bounded to $E$, whereas $G$ describes the dynamic of regulated degradation. (B) Example of a small regulatory network of genes with interactions between gene regulatory regions and transcription factors.

of the experimental data and by the time scale in consideration. In the most simple assumption $F_i$ is directly proportional to some linear combination of transcription factor concentrations. A more realistic, and interesting model, is characterized by a non-linearity in the function $F$.

The second term represents the degradation with constant rate $k_2$. The function $G_i$ describes the dynamics of regulated degradation [243], although it is generally assumed that $G_i = 1$ and the term $k_2 x_i$ is called auto-regulated degradation. The basal production of a constitutive active gene is not explicitly included in the formula since it can be represented by a transcription factor, in this case the transcription initiation complex, that spontaneously initiates transcription with a certain probability.

The system of differential equations defines the dynamics of the network. Depending on the protein concentrations $x_i$, the regulatory functions $F_i$ and $G_i$, the next state of the network for every gene is given by $x_i(t + dt) = x_i(t) + \dot{x}_i(t)dt$. Interestingly, $x_i(t)$ is not necessarily bound to be interpreted as a concentration and could also represent the average copy number of proteins per cell under the assumption of a constant cell volume.
5.3.3 Gene Reaction Network model

5.3.3.1 An algebraic representation of combinatorial transcription factor binding

Transcription factors regulate gene expression by binding to specific sites in the genomic cis-regulatory regions [214, 248]. Many genes are controlled by a number of different transcription factors and different arrangements of binding sites can compute logic operations on multiple inputs.

The model relies on two assumptions: firstly the dissociation constant $\theta$ of transcription factors defined in (5.11) can be continuously tuned by choosing the appropriate binding sequence; secondly interactions between transcription factors are possible if the binding sites are close enough. Both assumptions are realistic since dissociation constants range typically across a relevant range of protein concentrations and transcription factors are able to interact via steric, hydrophobic and electrostatic interactions. Therefore I express the dynamics of transcriptional networks with the following formalism:

$$ F_i^n(Z) = \sum_{S \subseteq \{1,2,\ldots,n\}} \beta_S \prod_{j \in S} Z_j $$

(5.15)

$F_i(Z)$ is a multilinear polynomial function, or sigma-pi function, since a sum and a product operation appear on the right hand side. The function describes the binding of $n$ different transcription factors to a promoter region and their interactions. Formally the function is a linear combination of monomials in $Z = (Z_1, Z_2, \ldots, Z_n)$, where each term $Z_j$ represents the effect of transcription factor $j$ on gene $i$ according to the Hill function. The coefficients $\beta_S \in \{0,1\}$ determine which one of the monomials is present.

While the equation may appear complicated, the intuition behind it is quite simple: a multilinear formula can be seen as a binary tree with the edges directed towards the root. Every node can compute either addition (plus gate +) or multiplication (product gate $\times$). For a gene $i$ the leaves represent the input variables $Z_1, Z_2, \ldots, Z_n$ and the output is the polynomial $F_i$ evaluated by the root. The computation performed by different combinations of transcription factors are easily represented by multidimensional linear functions. For instance, non-competitive binding of two transcriptional enhancers $A$ and $B$ on the same DNA sequence is described by $Z_A Z_B$. Only one molecule at time can be bound to it’s specific site, and thus the power of every input variable in each monomial
is at most one (the non-linearity caused by transcription factor oligomerization is given by the constant \( m \) of the Hill function).

Not all combinations of monomials defined by (5.15) have a straightforward interpretation as molecular interactions between transcription factors and binding sites. In other words, we have to find a suitable general formulation for \( F_i \) such that all plausible combinations can be represented in a single formula. I consider a set of \( n \) inputs, representing the activity of transcriptional enhancers \( Z \) and transcriptional repressors \( \overline{Z} = 1 - Z \) in the domain \([0, 1]\), as defined in equations (5.8) and (5.10). The set of transcription factor combinations influencing a gene \( x_i \) can be expressed in matrix form using the Kronecker product\(^2\) (\( \otimes \)), also known as direct product, tensor product, or outer product.

\[
F^i_n(Z) = B \bigotimes_{j=1}^{n} \left( \begin{array}{c} Z_j \\ \overline{Z}_j \end{array} \right) = B \left( \begin{array}{c} Z_1 \\ \overline{Z}_1 \end{array} \right) \otimes \left( \begin{array}{c} Z_2 \\ \overline{Z}_2 \end{array} \right) \otimes \cdots \otimes \left( \begin{array}{c} Z_n \\ \overline{Z}_n \end{array} \right) \right) \tag{5.16}
\]

The Kronecker product defines all the possible combinations of transcription factors that influence the expression of a gene. The Kronecker product is multiplied with \( B \), a column vector composed of the coefficients \( \beta_S \in \{0, 1\} \). Depending on the values assumed by each \( \beta_S \), I can describe different models of transcription factor binding.

The number of combinations is given by all possible combinations of \( \beta_S \) coefficients, that is \( 2^n \) combinations for the binding of \( n \) transcription factors. Given that a gene has \( \leq n \) transcription factors binding sites, the combinatorial possibilities results from the possible combinations of binding of either 1 transcription factor, 2 transcription factors, or up to \( n \) transcription factors: \( 2^1 + 2^2 + \cdots + 2^n \). Since the number of combinations increase exponentially with the number of transcription factors, searching for network topologies that have interesting computational characteristics in this space becomes soon unfeasible. Interestingly, many combinations represent the same logic function, or are equivalent if we assume, for instance, that the sequence of transcription binding sites on the DNA strand does not have an influence. The course of dimensionality can be partially avoided by considering the physics of transcription factors and binding sites interaction, and thus by setting some of the coefficients \( \beta_S \) to be equal to each other.

---

\(^2\)Kronecker product (\( \otimes \)): given two matrices \( A = [a_{ij}] \) of order \( m \times n \), and \( B = [b_{ij}] \) of order \( p \times q \), the Kronecker product is defined as a matrix \( A \otimes B = [a_{ij}B] \) of order \( (mp) \times (nq) \).
5.3.3.2 Gene regulation without order

Firstly I consider the case of $n$ TF binding sites regulating the production of a gene, for which the sequential order of binding sites on the DNA strand is irrelevant. This is necessarily true for two binding sites, since their order does not influence the nature of their mutual interaction. We can assume that this is equally true for $n$ binding sites since the DNA strand can bend and TFs interactions in space are not necessarily constrained by the order on their sequence (Figure 5.6A-B). In this case the Kronecker product reduces to:

$$\mathcal{F}_i^n(Z) = \mathbf{B} \bigotimes_{j=1}^{n} \begin{pmatrix} Z_j \\ \overline{Z}_j \end{pmatrix} = \beta_1 (Z_1 Z_2 \ldots Z_n) + \beta_2 (Z_1 \overline{Z}_2 \ldots Z_n + \cdots + Z_1 Z_2 \ldots \overline{Z}_n) + \cdots + \beta_{d-1} (\overline{Z}_1 \overline{Z}_2 \ldots Z_d + \cdots + Z_1 \overline{Z}_2 \ldots \overline{Z}_d) + \beta_d (Z_1 Z_2 \ldots \overline{Z}_d)$$

(5.17)

The indexes of the coefficients $\beta_S$ are computed based on the binomial coefficients (Pascal triangle) to compensate for the geometrical symmetries. The advantage of this representation is clear, since we have reduced the number of dimensions from $2^n$ to $d$. The number $d$ of variables $\beta_S$ scales linearly with increasing number of inputs $n$, since $n = d + 1$ instead of exponentially.

5.3.3.3 Gene regulation with order

Gene regulation without order constrains considerably the number of possible TF logic combinations. I consider here the case in which the order of the TF binding sites on the DNA is decisive to compute the regulatory output $\mathcal{F}_i$: more precisely, the order of the transcription binding sites on the DNA strand is important, although the orientation of the sequence in respect to the coding region (from left to right or right to left) is not (Figure 5.6C-D).
For \( n = 2 \) the model is simply the same as above, since the interaction of two transcription factors is not dependent on the left/right order of the binding sites:

\[
\mathcal{F}_1^2(Z) = 2 \otimes_{j=1}^2 \left( \begin{array}{c} Z_j \\ Z_j \end{array} \right) \\
= \beta_1 (Z_1 Z_2) + \\
+ \beta_2 (Z_1 Z_2 + Z_1 Z_2) + \\
+ \beta_3 (Z_1 Z_2)
\]

(5.18)

For \( n = 3 \) the model is build by recursive use of the Kronecker product. Let’s say on a cis-regulatory module, transcription factor \( Z_1 \) interact with it’s neighboring transcription factor \( Z_2 \), and the result of their interaction is detected by a third transcription factor \( Z_3 \). All possible unique logic functions given \( Z = \{Z_1, Z_2, Z_3\} \) are described by:

\[
\mathcal{F}_1^3(Z) = 3 \otimes_{j=1}^3 \left( \begin{array}{c} Z_j \\ Z_j \end{array} \right) \\
= \mathcal{F}_1^2 (\mathcal{F}_1^2(Z_1, Z_2), Z_3) \\
= \beta_4 (\mathcal{F}_1^2(Z_1, Z_2) Z_3) + \\
+ \beta_5 \left( \mathcal{F}_1^2(Z_1, Z_2) Z_3 + \overline{\mathcal{F}_1^2(Z_1, Z_2)} Z_3 \right) + \\
+ \beta_6 \left( \overline{\mathcal{F}_1^2(Z_1, Z_2)} Z_3 \right)
\]

(5.19)

\( \mathcal{F}_1^3 \) represents originally \( 2^n = 256 \) different combinations of logic functions, that we can reduce to \( 2^d = 64 \), where \( d = 6 \). Since we also have to take into account the possible permutations of inputs in the logic tree (either \( ((Z_1, Z_2)Z_3) \) or \( ((Z_1, Z_3)Z_2) \) or \( ((Z_2, Z_3)Z_1) \)), the total number of combinations is \( 3 \cdot 2^d = 197 \), which is still far less than the complete parametric space.

5.3.3.4 Relationship with conventional logic theory

The tensor algebra provides a model to formally express transcriptional networks as compositions of computational primitives, since to a Kronacker product with defined coefficients \( \mathbf{B} \) corresponds a unique logic function. Depending on the coefficient values, the equation can describe the independent binding of transcription factors as well as
Figure 5.6. Computation of coefficient matrix $B$. The coefficients corresponding to different combinations of 2 inputs $Z_1$ and $Z_1$ or 3 inputs $Z_1$, $Z_2$ and $Z_3$ (and their complement $\overline{Z}_1$, $\overline{Z}_2$ and $\overline{Z}_3$). In the case where the binding site order is not important (A-B) the coefficients for each monomial can be computed using the Pascal triangle algorithm. The corresponding matrix is shown at the bottom. If the binding site sequence has an effect on the possible interactions between TFs (C-D), the coefficient matrix is obtained by recursively applying the matrix for two inputs. The corresponding matrix is shown at the bottom.

steric interactions such as in competitive binding. Given the binding strength, the steric interactions, the cooperativity and the numerical factors $\beta_S$, the binding probability can be calculated straightforwardly. The reverse is also true, to each logic function corresponds a kinetic representation of the molecular process.

Logic functions can be described as combinations of the elementary gates AND, OR, NOR, NAND, XOR, EQV, TRUE, and FALSE. According to Boolean algebra, different 2-input logic gates can be constructed by choosing appropriate values for the numerical parameters $\beta_i$, as shown in Table 5.2A. Depending on the value of the constants, different forms of computation can be obtained. The same procedure apply to multiple inputs gates, for instance for 3-input gates as shown in Table 5.2B.

AND, OR, NOR and NAND logic gates are four logic operations referred as canalizing functions. Canalizing gates are Boolean functions in which at least one of the input variables is able to determine the output regardless of the other variables. These functions are abundantly utilized in vertebrate gene regulatory networks and provides stability to
dynamical systems [249]. When competitive binding or in general interaction between two binding transcription factors occur, the respective logic gates obtained are: XOR and EQV, which are non-canalizing functions. In this case knowledge of both inputs is always required to compute the output. Additionally we have the logic gates FALSE and TRUE, which are independent of any input (Figure 5.7).

Complex regulatory functions with more inputs can be easily implemented by the general equation (5.16). The described gene-like structure is modular, in the sense that transcription factors can be combined into independent computational subunits. The formalism I use to describe regulation in transcriptional networks is compatible with the grammar of structured programming languages. Notice that although the equation extends to higher dimensions, it is more convenient to use the formulation for two inputs recursively.

I will refer from now on to such transcriptional network as gene reaction network (GRN) since it relays on a set of chemical reactions. I would like to emphasize that the complexity of molecular interactions such as transcription, mRNA splicing, translation, alternative splicing, enzymatic catalysis, phosphorylation and degradation can be modeled with the proposed mathematical framework and the parameters can be matched to experimental measurements since the model is directly derived from the laws of reaction kinetics.
Chapter 5. A genetic language for developmental programs

Figure 5.7. Combinatorial transcription logic. Cis-regulatory constructs can implement conventional canalizing logic gates (A) AND, (B) NAND, (C) OR, (D) NOR and non-canalizing (E) XOR, (F) EQV, (G) FALSE, (H) TRUE. The z-axis represents the output partition function $P$ given $[X]$ and $[Y]$. The computation depends on the steepness of the sigmoidal function $H$, ranging from (top to bottom row) continuous, approximately Boolean and discrete Boolean.
5.3.3.5 Example of combinatorial transcriptional regulation

I shall illustrate the utility of the proposed transcriptional network model and its relation with conventional computation theory with elementary examples. Consider a gene $i$ regulated by two different transcription factors $A$ and $B$, either enhancers or repressors. The resulting multilinear polynomial is defined by:

$$F^2(Z) = \begin{bmatrix} \beta_1 & \beta_2 & \beta_2 & \beta_3 \end{bmatrix}^T \left[ \begin{pmatrix} Z_A \\ Z_A \end{pmatrix} \otimes \begin{pmatrix} Z_B \\ Z_B \end{pmatrix} \right]$$

$$= \beta_1 Z_A Z_B + \beta_2 Z_A Z_B + \beta_2 Z_A Z_B + \beta_3 Z_A Z_B$$

$$= \beta_1 Z_A Z_B + \beta_2 (1 - Z_B) + \beta_2 (1 - Z_A) Z_B + \beta_3 (1 - Z_A)(1 - Z_B)$$

The first monomial $Z_A Z_B$ represents the effect of the binding of two enhancing transcription factors, the last the effect of two repressors. $Z_A Z_B$ and $Z_A Z_B$ are equivalent formulations for the binding of one repressor and one enhancer and are thus associated to the same parameter $\beta_2$. We can think of $F_i$ as the binding probability of a particular configuration of transcription factors in function of their cellular concentrations.

As an example, in the gene network represented in Figure 5.8, regulation of gene $A$ is described by the following formulation (the formulation for gene $B$ is equivalent):

$$F^2_A(Z) = \begin{bmatrix} 1 & 0 & 0 & 1 \end{bmatrix}^T \left[ \begin{pmatrix} Z_A \\ Z_A \end{pmatrix} \otimes \begin{pmatrix} Z_B \\ Z_B \end{pmatrix} \right]$$

$$= Z_A Z_B + Z_A Z_B$$

$$= Z_A Z_B + (1 - Z_A)(1 - Z_B)$$

$$= 1 - Z_A - Z_B + 2Z_A Z_B$$

$$= Z_A + (1 - Z_B) - 2Z_A(1 - Z_B)$$

$$= OR(Z_A, Z_B)$$

The advantage of the present formulation in respect to standard kinetic differential equations, is that it is now much easier to visualize, understand and manipulate networks of interacting elements. By choosing an appropriate restricted set of free parameters, I reduce considerably the dimensionality of genetic networks: the network topology, rather than exact parameters, determines the system dynamics.
Chapter 5. A genetic language for developmental programs

5.3.3.6 Asymmetric cell division

The model described so far deals with the logic of interactions between transcription factors and genes, but how can cells containing the same genetic information behave differently? This can be explained by the process of cell division\(^3\). In the trivial case of symmetric divisions, substances are distributed equally in the daughter cells, and concentrations are the same as in the mother cell. Symmetric division imply that, in absence of asymmetric input from the external world, the two daughter cells display the same behavior of their mother, and therefore are regarded as being of the same cell type.

A more interesting situation arises with asymmetric distribution of substances, since now daughter cells receive different molecular components and therefore may perform different types of computation. For proper morphogenesis, cell divisions and cell fate decisions must be tightly regulated. One elegant way to accomplish this is to couple the asymmetry of cell division with the distribution of cell fate into the daughter cells (Figure 5.9). The division of a progenitor produces two daughters cells with often divergent cell fates: usually one daughter retains progenitor characteristics while the other commits to a defined cell lineage through differentiation. In this way asymmetric cell divisions assure the generation of a variety of cell types.

I assume that during cell division, different TFs can be distributed asymmetrically to the two daughter cells $2c$ and $2c + 1$ depending on their corresponding asymmetry constant $\alpha \in [-1, 1]$, so defining new (and possibly different) internal states. For TFs with $\alpha = 0$ both daughter cells receive the same amount, whereas for $\alpha = \pm 1$ the TFs are distributed

---

\(^3\)Additional mechanisms for cellular diversification exists: either by influence of external inputs, or by an asymmetry breaking processes such as interaction between Notch and Delta membrane ligands.
A genetic language for developmental programs

Figure 5.9. The symmetry of cell division. The distribution of a transcription factor $X$ depends on its internal distribution in the cell and the orientation of the division plane (dashed line). (A) Scheme of asymmetric division mechanism ($\alpha = -1$), where the division plan is rotated 90 degrees to the left in respect to the vertical axes. (B) Scheme of symmetric division mechanism ($\alpha = 0$), where the orientation of the division plan is parallel to the vertical axes. (C) Scheme of asymmetric division mechanism ($\alpha = 1$), where the division plan is rotated 90 degrees to the right in respect to the vertical axes. The vertical axes is arbitrarily defined as the axes of symmetry given by the internal distribution of molecular components.

asymmetrically with only one daughter cell receiving the entire amount.

$$
egin{align*}
    x_{2c} &= x_c + \alpha X x_c = (1 + \alpha X) x_c \\
    x_{2c+1} &= x_c - \alpha X x_c = (1 - \alpha X) x_c
\end{align*}
$$

Upon asymmetric distribution, the inputs to some of the genes $X_i$ have now changed, and thus the regulatory element functions $F^+_i(Z)$ may deliver different outputs.
Chapter 5. A genetic language for developmental programs

5.4 Results: Genetic control of developmental processes

The ordered pattern of cell divisions and differentiations into defined cell types is implicitly specified in the primary genetic sequence (genetics), modifications in the secondary DNA structure (epigenetics), and influenced by cell-cell interactions. Despite the numerous studies on cellular processes regulating development, there is still poor understanding of how molecular mechanisms that control cell fate specification are linked to the final phenotype.

Here, I present the results regarding the investigation of computation models suited to represent development processes. In Section 5.4.1 I briefly summarize the genetic language formalism I developed; in Section 5.4.2 I use the algebraic framework to model basic mechanisms regulating cell division and differentiation (cell division asymmetry and irreversible differentiation); finally in Section 5.4.4 I address the question on how, given a desired developmental sequence, it’s possible to reverse engineer plausible regulatory networks.

5.4.1 Formal genetic language definition

The genetic language is defined by a set of variables $x \in \mathbb{R}_{\leq 0}$ that represent substance concentrations, and a set of allowed operations on the substance concentration values. These operations completely define the genetic language.

Read. Information about transcription factor concentrations is obtained through the Hill function $Z$, which computes the binding probability of a transcription factor to a regulatory region given the affinity constant $\theta$, cooperativity $m$ and binding bias $b$.

$$Z(x + b, \theta, m) = \frac{(x + b)^m}{\theta^m + (x + b)^m} \quad (5.23)$$

Write. Information can be written to the environment by production of a given substance according to the rate equation, which influences the substance current concentration. $F$ takes the form of one of the possible logic operations, or combinations thereof.

$$\dot{x} = k_1 F[Z(x)] - k_2 x \quad (5.24)$$

Distribute. Information is encapsulated by the cell membrane, which prevents external agents to directly interact/modify the cellular molecular components, and provides a
protected environment where the cell can perform it’s computation. During development, a cell $c$ divides and is able to distribute asymmetrically its internal components to daughter cells $2c$ and $2c + 1$ with gene asymmetry constant $\alpha_X \in [-1, 1]$.

$$
\begin{align*}
    x_{2c} &= x_c + \alpha_X x_c = (1 + \alpha_X) x_c \\
    x_{2c+1} &= x_c - \alpha_X x_c = (1 - \alpha_X) x_c
\end{align*}
$$

(5.25)

**Logic operations.** Logic operations perform combinatorial computation on values, where $y$’s can be either the output of $Z$ or the output of another logic operation.

$$
\begin{align*}
    \text{AND}(y_1, y_2) &= y_1 \cdot y_2 \\
    \text{OR}(y_1, y_2) &= y_1 + y_2 - \text{AND}(y_1, y_2) \\
    \text{NOT}(y) &= 1 - y
\end{align*}
$$

(5.26) (5.27) (5.28)

**Derived logic operations.** The elementary operations can be composed into derived operations, for instance:

$$
\begin{align*}
    \text{XOR}(y_1, y_2) &= \text{AND}(\text{NOT}(\text{AND}(y_1, y_2)), \text{OR}(y_1, y_2)) \\
    \text{NAND}(y_1, y_2) &= \text{NOT}(\text{AND}(y_1, y_2)) \\
    \text{NOR}(y_1, y_2) &= \text{NOT}(\text{OR}(y_1, y_2)) \\
    \text{NXOR}(y_1, y_2) &= \text{NOT}(\text{XOR}(y_1, y_2)) \\
    \text{TRUE}(y) &= \text{AND}(y, y) \\
    \text{FALSE}(y) &= \text{NOT}(\text{TRUE}(y))
\end{align*}
$$

(5.29) (5.30) (5.31) (5.32) (5.33) (5.34)

Another useful derived operation is the threshold function $Z_o$, that will perform threshold at any given desired value $tr \in [0, 1]$:

$$
Z_o(y, tr, \theta, m \to \infty) = Z(y + \theta - tr, \theta, m \to \infty)
$$

(5.35)

Notice that for cooperativity $m \to \infty$, values of $x$ are bounded to the set $\{0, 1\}$, logic operations behave as Boolean logic gates, and our genetic language reduces to conventional Boolean algebra.

The proposed formalism, to which I refer to as *genetic reaction network* model, greatly simplifies the task for searching and constructing regulatory network for developing systems, since it is sufficient to explore the network topology and removes the need of parameter tuning. Although abstract, the formalism can be directly cast into the corresponding – but less understandable – kinetic differential equations.
5.4.2 Toward a genetic model of development: Gene Regulatory Motifs

Unluckily, the link between genetic networks and development is complex and unclear. The genome is a product of evolution, and is not designed under the laws of parsimony or simplicity. Redundancy, evolvability, robustness and strong feedback interactions are the prominent features of most transcriptional networks. Gene regulatory network models have been used to investigate patterns of gene activation. Identification of network motifs for simple mechanisms such as bistability and oscillations have proved successful approaches, providing a conceptual framework for understanding the network design principles [214].

Among a variety of interesting network topologies to control developmental processes, a simple interaction pattern stands out, namely auto-excitation and mutual inhibition in presence of cooperative binding. Inhibition provides a mechanism by which genes compete for expression, whereas auto-excitation is necessary to stabilize winning states.

An example of bistable phenotype is how glutamatergic fates are chosen over GABA-ergic fates. Glutamate and GABA (gamma-aminobutyrate) are the predominant neurotransmitters that specify excitatory versus inhibitory neurons in vertebrates. The expression of those neurotransmitter is mutually exclusive, and is achieved through the mutual repression of \( Tlx3 \) and \( Tlx1 \) [250].

5.4.2.1 A Gene Regulatory Motif: the Bistable Switch System

Cell decisions regarding acquisition of an appropriate cell fate rely on the ability to commit to different stable states. The most simple system that can perform such a task is a bistable switch [242, 251], in which two autocatalytic transcription factors \( X_1 \) and \( X_2 \) negatively regulate the expression of the other. Such a biological observed system has been described by the following system of differential equations [242, 245]:

\[
\begin{align*}
\dot{x}_1 &= a_1 \frac{x_1^m}{\theta^m + x_1^m} + b_2 \frac{\theta^m}{\theta^m + x_2^n} - k_2 x_1 \\
\dot{x}_2 &= a_2 \frac{x_2^m}{\theta^m + x_2^m} + b_1 \frac{\theta^m}{\theta^m + x_1^n} - k_2 x_2
\end{align*}
\] (5.36)

where \( x_1 \) and \( x_2 \) refer respectively to the concentrations of \( X_1 \) and \( X_2 \), and \( a_1, b_1, a_2, b_2 \) the respective weights.
Given a constant input, the bistable switch has two stable states. In the first state, the gene for one repressor is turned on while the synthesis of the second repressor is turned off. The reverse is true for the second state. A third meta-stable state forms at the border between the two attractor basins. I approximate the above system with my compact formalism:

\[
\begin{align*}
\dot{x}_1 &= k_1 \text{OR}[Z(x_1), Z(x_2)] - k_2 x_1 \\
\dot{x}_2 &= k_1 \text{OR}[Z(x_2), Z(x_1)] - k_2 x_2
\end{align*}
\] (5.37)

This formulation gives a straightforward view on the mechanisms at play. Gene \(X_1\) contains an enhancer sequence with two distal binding sites: one for the inducer \(X_1\) itself, the other for a repressor \(X_2\). In the absence of binding the gene is expressed with a basal production proportional to the input strength, which is progressively decreased by the binding of the repressor \(X_2\). Binding of \(X_1\) is independent from \(X_2\) (given by the function OR). Notice that the parameters \(a\) and \(b\) are now represented by the dissociation constants of the TFs.

The Michaelis-Menten model, in which the cooperative binding between proteins is negligible \((m = 1)\), is characterized by a single stable attractor state. However, using the Hill function \(Z\) with cooperativity \(m > 1\) we find a particular parameter space where multiple attractor states are present (Figure 5.10). This binary switch has three stable states, one of which can become a saddle point for specific \(\theta\) values. Moreover the system reaches saturation at a concentration \([X] \approx 1\), thus providing a useful limitation of maximal equilibrium concentration for any given gene \(X_i\).

We can rewrite the equation such that the system can be driven toward a specific state by an input \(I\).

\[
\begin{align*}
\dot{x}_1 &= k_1 \text{AND}[\text{OR}[Z(x_1), Z(x_2)], Z(I)] - k_2 x_1 \\
\dot{x}_2 &= k_1 \text{AND}[\text{OR}[Z(x_2), Z(x_1)], Z(I)] - k_2 x_2
\end{align*}
\] (5.38)

The input can be either an intrinsic mechanism or an external stimulation. The first mechanism is encoded in the genetic code and represents genetically defined developmental pathways. The second mechanism is important to confer flexibility and the possibility of coordinating fate choices between neighboring cells.

An interesting variation that takes into consideration the input \(I\) is given by:
Figure 5.10. Parameter dependence of a 2-dimensional genetic switch. Vector field $V$ in function of $X_1$ and $X_2$ concentrations for a simple bistable switch with mutual inhibition and self-excitation depending on the cooperativity $n$ and TF dissociation constant $\theta$. The system has either 1 or 3 attractors. Red traces are simulated trajectories from different initial starting points.

\[
\begin{align*}
\dot{x}_1 &= k_1 \text{OR}[Z(x_1), \text{AND}[\text{OR}(Z(x_1), I), Z(x_2)]] - k_2 x_1 \\
\dot{x}_2 &= k_1 \text{OR}[Z(x_2), \text{AND}[\text{OR}(Z(x_2), I), Z(x_1)]] - k_2 x_2
\end{align*}
\]

(5.39)

In presence of a strong input the switch $I$ has 3 attractor states, and in its absence has 4 attractor states (Figure 5.11), which means that the system displays hysteretic behavior upon input withdrawal. During cell division cells are able to jump from one place in the attractor space to another depending on the values assumed by the respective asymmetry constants (or by a asymmetric external input), thus enabling cells to acquire determined expression patterns that are highly stable even in presence of noise.
Chapter 5. *A genetic language for developmental programs*

5.4.2.2 *A Gene Regulatory Motif: the Multistable Switch System*

The bistable switch can be extended to higher dimensions by adding new genes and new self-excitatory and inhibitory connections. For instance in a 3 gene system, the equations for a simple multistable switch are defined as following:

\[
\begin{align*}
\dot{x}_1 &= k_1 \text{AND}[\text{OR}[Z(x_1), \text{AND}[Z(x_2), Z(x_3)]], Z(I)] - k_2 x_1 \\
\dot{x}_2 &= k_1 \text{AND}[\text{OR}[Z(x_2), \text{AND}[Z(x_1), Z(x_3)]], Z(I)] - k_2 x_2 \\
\dot{x}_3 &= k_1 \text{AND}[\text{OR}[Z(x_3), \text{AND}[Z(x_2), Z(x_3)]], Z(I)] - k_2 x_3
\end{align*}
\]

(5.40)

And for the more interesting case with hysteresis:

\[
\begin{align*}
\dot{x}_1 &= k_1 \text{OR}[Z(x_1), \text{AND}[\text{OR}(Z(x_1), I), \text{AND}[Z(x_2), Z(x_3)]], Z(I)] - k_2 x_1 \\
\dot{x}_2 &= k_1 \text{OR}[Z(x_2), \text{AND}[\text{OR}(Z(x_2), I), \text{AND}[Z(x_1), Z(x_3)]], Z(I)] - k_2 x_2 \\
\dot{x}_3 &= k_1 \text{OR}[Z(x_3), \text{AND}[\text{OR}(Z(x_3), I), \text{AND}[Z(x_1), Z(x_2)]], Z(I)] - k_2 x_3
\end{align*}
\]

(5.41)

in which the system without input is stable in its *off* state (neither of the two genes is expressed). The behavior of the system is similar to the simple bistable switch and is illustrated in Figure 5.12. Again during cell asymmetric division cells are able to jump from one place in the attractor space to another depending on the values assumed by the respective asymmetry constants $\alpha_1$, $\alpha_2$ and $\alpha_3$. 

---

**Figure 5.11. Dynamics of a 2-dimensional genetic switch with input.** (A) Top: scheme of the network with mutual inhibition between the two transcription factors $X_1$ and $X_2$ with auto-stimulation and external input $I$. Bottom: logic circuit with continuous Boolean gates. (B) Vector field $V$ in function of $X_1$ and $X_2$ concentrations for the switch without input $I = 0$. The system has 4 attractor states, which means that the attractor states exhibit hysteresis. (C) Vector field $V$ in function of $X_1$ and $X_2$ concentrations for the switch with input $I = 1$. Attractors at either high $X_1$ or $X_2$ represent downstream differentiation pathways. Red traces are simulated trajectories from different initial starting points.
5.4.3 State-dependent computation with Gene Regulatory Motifs

5.4.3.1 Encoding state transitions with bistable switches

Genetic networks are dynamical systems. At any given time point, the dynamic system is associated with a state as a point in the state space, that is the vector $\mathbf{x}$ of real numbers that describe the current concentration of each gene product. The evolution of the system is determined by the corresponding kinetic differential equations, and can be represented by:

$$ \dot{\mathbf{x}} = V(\mathbf{x}) $$

where the vector field $V(\mathbf{x})$ is a smooth function that provides the velocity vector of the dynamical system at every point of the state space (Figure 5.13). The evolution of the concentrations of gene products is continuous through time, and their trajectories are characterized by the presence of regions of unstable and stable gene expression. Stable regions act as basins of attraction (a wide range of initial conditions lead to the same expression patterns).

In the discrete case, we can associate a set of discrete values to each system variable, with as many values as stable states. Each point in the state space is mapped to a unique set of discrete variables corresponding to the system evolution after an infinite
amount of time. While a continuous network state gives information about the current position of the system, a discrete network state describes the expected behavior of the system. For instance, a 2-gene network \((A\text{ and } B)\) with \([A] = 0.8\) and \([B] = 0.2\) will evolve to the stable state \([A] \simeq 1\) and \([B] \simeq 0\) even in presence of noise. The discrete state is given by:

\[
\{A \simeq 1, B \simeq 0\}, \text{ or simply } \{A\}
\]

Each gene \(X\) is characterized by an asymmetry constant parameter \(\alpha_X\), corresponding to the asymmetric division constant of its protein. At mitosis the type of division is the result of the partitioning of all the substances contained in the cell. Most of them will be distributed symmetrically (\(\alpha = 0\)), but some have the ability to dock on the membrane, preferentially on one side of the cell in respect to the position of the mitotic spindle (\(\alpha > 0\) or \(\alpha < 0\)). Biologically, asymmetry constants could be seen as the probabilities of a substance to get tagged and sequestered in a specific location in the cell.

Because of the law of mass conservation in Equation (5.22), jump trajectories taken by cells at the time of cell division have opposite directions in the gene expression space. A representation of the behavior of cells regulated by a single switch with asymmetry constants \(\alpha_A\) and \(\alpha_B\) can be seen in Figure 5.14. Jumps lead to different distribution of transcription factors, and thus to different gene expression profiles and often to different cell states. A systematic description of division patterns is provided in Table 5.3.
Figure 5.14. Division angle and cell division type. (A) Schematic representation of asymmetric distribution of fate determinants $A$ and $B$ by internal asymmetric distribution, that is control of division angle $\omega$ in respect to the symmetry axis. (B) Distributions of different division types in function of the division angle $\omega$. Different division patterns arise: (C) $\{AB\} \rightarrow \{AB\}, \{AB\}$; (D) $\{AB\} \rightarrow \{A\}, \{AB\}$; (E) $\{AB\} \rightarrow \{A\}, \{B\}$; (F) $\{AB\} \rightarrow \{B\}, \{AB\}$. Red straight traces are simulated jumps at different angles and red curvilinear trajectories show the time evolution after the jump. Blue lines indicate the $\omega$ angle in respect to the internal distribution of proteins.
Table 5.3. Deterministic state transition table. Mapping of different discrete state transitions (division rule) in function of division angle or asymmetric constants for a bistable switch composed of gene $A$ and $B$.

Although $\alpha_A$ and $\alpha_B$ can assume a range of different values independently, the effect on the cell division outcome is better described by the parameter $\omega$, which describes the angle of the division in respect to some arbitrary reference. Given the division angle, we can compute the $\alpha$ constant for each substance in the bistable switch. The values taken by $\omega$ completely describe symmetric and asymmetric partitioning of substances into the daughter cells.

By setting:

$$
\alpha_A = N \left( \frac{\sin(\omega)}{\cos(\omega) + \sin(\omega)} \right)
$$

$$
\alpha_B = N \left( 1 - \frac{\sin(\omega)}{\cos(\omega) + \sin(\omega)} \right)
$$

with $N$ a normalization constant defined by:

$$
N = \frac{1}{\text{Max} \left( \frac{\sin(\omega)}{\cos(\omega) + \sin(\omega)}, 1 - \frac{\sin(\omega)}{\cos(\omega) + \sin(\omega)} \right)}
$$
5.4.3.2 Epigenetic control over cell state transitions

The asymmetrical repartitioning of gene activity permits the formation of diverse cell fates and results into patterns of cell fate decisions (Figure 5.15A-G). However, it does not explain how cells are prevented to jump back to earlier, more permissive, differentiation states. Despite some window of plasticity, once cells have activated specific differentiation markers, there is a very small probability that they may subsequently activate patterns of earlier developmental stages, or alternative cell fates.

Irreversibility can be achieved by inhibiting asymmetric distribution once a threshold concentration is reached (Figure 5.15H-P). I implement this mechanisms by introducing a dependency in the asymmetry constant on the gene protein concentrations, such that the expression of a transcription factor prevents the reactivation of genes of previous states by inhibiting directly their asymmetric distribution. In biological terms this can be viewed as epigenetic programming of differentiation memory on top of the DNA coding sequence.

5.4.3.3 Probabilistic cell divisions encode probabilistic state transitions

So far the description of genetic circuits has been deterministic. In a deterministic network, each gene transcription rate depends on other genes by means of a unique combination of continuous Boolean functions and asymmetric distribution coefficients are constant. Conceptually this assumes an environment without uncertainty. A more interesting case is when the genetic circuit does not represent exact transitions but is rather probabilistic. The main idea is to accommodate a distribution of asymmetric division angles weighted by their probabilities $p_{\alpha_i}$:

$$x_{2c} = \sum_{i} p_{\alpha_i} (x_c + \alpha_i x_c)$$

$$x_{2c+1} = \sum_{i} p_{\alpha_i} (x_c - \alpha_i x_c)$$

(5.47)

As we will see in the next section, those probability distributions can be encoded by a single parameter (angle of cell division $\omega$). Under this perspective, the equation has a very direct interpretation: repartition of molecules in the cellular environment is subject to noise and stochastic decisions, which causes uncertainty in the orientation of the mitotic spindle and thus the ability of a cell to undergo asymmetrical divisions. Stochastic distribution of substance allows to encode probabilistic state transitions.
Figure 5.15. Cell division patterns. Cell lineage distribution of two mutually exclusive transcription factors A and B in function of their own asymmetry distribution constants $\alpha_A$ and $\alpha_B$. Circles represent cells expressing A (red), B (blue), or both AB (violet). (A-F) Lineages corresponding to asymmetric and symmetric repartition of transcription factors A and B. (H) Coupled bistable switches. (G-N) Lineages corresponding to asymmetric and symmetric repartition of transcription factors A and B, where transition to previous states is inhabited through epigenetic memory.
5.4.4 Reverse engineering of gene regulatory networks

5.4.4.1 The inverse problem

I am interested in the synthesis of regulatory systems by explicit construction of networks, also referred as the *inverse problem*. The difference between inference and reverse engineering consists in the fact that, in the first, we produce a generative model based on the data that we observe, whereas in the second, we construct a system that generates the desired behavior. The inverse problem is ill-posed since many distinct network topologies may display the same desired behavior.

The inference of the gene steady-state distribution along lineage trees gives us an insight into the relationship between gene activation and the lineage space (Chapter 4). The model makes a prediction on which genes should be active and when, and possibly a list of causal links (logic rules). The goal is therefore to create genetic networks that display behavior described by a specific state machine description based on simple mechanisms that control cell division, call fate distribution, and epigenetic regulation.

For this purpose I have designed an algebraic representation of gene regulation that makes intuitive the exploration and reverse engineering of genetic circuits for the control of cell divisions and differentiation events. From a variety of basic developmental motifs mechanisms, I have selected multistable switches since they are compact, robust, can display memory behavior, and can be regulated by epigenetic-like mechanisms. Multistable switches are used to control cell fate decision at the moment of cell division, and are used to link the activation of different functional genes, for instance genes responsible for terminal differentiation, or arborizations growth, to different developmental stages along the lineage. It is likely that real genetic networks are composed of multiple variations on this theme.

5.4.4.2 Sequences of multistable switches

Single switches motifs control the patterning of the lineage and can be used in combination with local environment feedback motives to completely control the evolution of cell lineage trees. For instance, the state diagram of the small lineage in Figure 5.16 can be cast into a network consisting of genetic bistable switches that controls the transition of cell states at every decision point (cell division). The translation of state lineage graphs
into a gene regulatory network follows some simple rules. I illustrate the idea in the case of bistable switches, although the same principle generalizes to multiple attractor networks.

- **Bistable switches are the regulators of key asymmetric division points in the lineage tree.** Every cell division is controlled by a bistable switch, which depending on some parameters will determine how the daughter cells will divide, and transitions are robust in respect to noise because of the attractor dynamics. Since we have shown in the previous Chapter that lineage trees can be decomposed into repetitive modular sequences of cell division and differentiation, only a small subset of switches are needed to control a greater number of divisions. Additionally, some of the cell divisions may be rather under the control of external cues, which further reduce the number of control switches needed.

- **Asymmetric divisions are used to encode state transitions.** Every bistable switch is characterized by the asymmetry constant parameters $\alpha_1$ and $\alpha_2$, corresponding respectively to the asymmetric division constants of TFs $X_1$ and $X_2$. Depending on the values assumed by the specific asymmetry constants, the division leads to defined state transitions.

- **Cell fate transitions occur only at the moment of cell division.** During the initial phase of the cell cycle the switch mechanisms becomes active, usually because of an incoming input. Both concentrations of TFs increase and are maintained equal by the attractor. Although active, the switch will have to wait until the moment of cell division to actuate its differentiation decision. This guarantees that the computation does not depend on cell cycle length, although it may be correlated. In this model, I do not consider state transitions during the cell life.

- **Probabilistic distribution of division angles.** Biological genetic networks are characterized by intrinsic noise, and stochastic decision at the molecular level. This results in uncertainties in the gene expression levels, and in the values assumed by the mitotic division angle. The presence of uncertainties is not necessarily detrimental, since stochastic decisions are very useful to encode probabilistic state transitions. I represent those transitions with a probability distribution over the division angle range.
Figure 5.16. Example of an attractor dynamics system. (A) Example lineage tree for the production of neurons (N) and glial cells progenitors (G). (B) State diagram, where nodes represent attractor states, and edges indicate transition between states, assumed to occur at the moment of cell division. Squares represent the joint probability of a cell division producing a given pair of daughter cells. Boxes represent inactive and active genes for each state. (C) Gene regulatory network composed of 3 interconnected bistable switches, for the selection of alternative fates in progenitor cells, and read-out genes for final phenotypes (N and G).

An example of an implementation of a small genetic network is illustrated in Figure 5.16. The original lineage tree consists of 5 progenitor cells, and 6 terminally differentiated cells (4 neurons and 2 glial cells). Because of the repetition of equivalent pattern of divisions, the system is decomposed of 3 proliferative states (P0, P1, and P2) and 2 differentiative states (N for neurons and G for glial cell) according to the method described in Chapter 4. In this case, each state fully describes the cell fate of the daughter cells, depicted as the joint probability of symmetric or asymmetric division.

The reverse engineering of a deterministic gene regulatory network consists in the design of a dynamic system with the appropriate state transitions. A series of bistable switches is used for modeling the transition from one cell type to the next. The network is composed of two parts: a core transcriptional network, and a read-out network. 6 genes coding for transcription factors (P0A, P0B, P1A, P1B, P2A, and P2B) are used to implement 3 decision points between alternative cell fates. State transitions occur only at the moment of cell division, when the asymmetric distribution of substances will determine the trajectory of the gene expression profiles in the two daughter cells. 2 read-out genes (N and G) are used to read the network configuration, and express appropriate functional proteins responsible for differentiation into neuron or glial cell. Epigenetic inhibition of alternative states assures that the differentiation process proceed forwards toward differentiated cell states (irreversibility).
5.5 Discussion

5.5.1 A formal model of development

Many elegant theoretical models have been proposed for molecular regulatory networks, and some have addressed the question of how cell fate decision are regulated in the context of biological development [21, 242, 245, 251, 252]. However, a comprehensive model of genetic controlled self-organization has to deal with many issues that more theoretical approaches do not address. Cells with different fates arise from the concomitant influence of asymmetric repartitioning of cell fate determinants, regulation of the cell cycle machinery, integration of continuous inputs from internal programs or external cues, changes in concentration and quantities due to the process of cell division, fluctuations in the transcription factor concentrations, locality of molecular interactions, combinatorial integration of inputs, and many others.

I propose here a formal model that, being derived directly from the laws of molecular interactions and reaction kinetics, is close enough to biology to be plausible and abstract enough to be understandable. The model works on two levels: on the first we have an highly abstract language that makes the explicit programming of developmental programs straightforward. On a second level the high level language can be projected down to the kinetic differential equations of the system, thus enabling to connect the abstract representation of the network with its molecular counterpart.

5.5.2 Genetic modules based on gene cooperation and competition

The regulation of gene expression plays a key role in development and function of an organism or a tissue, and it is widely believed that cellular differentiation is the result of a combination of epigenetic memory (to control the progression in the developmental pathways) and asymmetric distribution of cell fate determinants (to control the diversification of cell types). An interesting candidate regulatory mechanism is represented by bistable or multistable switches [245].

I developed a computational model that describes how cells can progress into the developmental pathway and eventually terminally differentiate. This is achieved by mathematical formulation of the interactions of transcription factors with cis-regulatory modules the control gene expression. The model is based on the idea that development can be
described by more or less conserved cell division patterns, and that those sequences are regulated by multistable gene regulatory modules.

Gene modules are small gene network which perform simple computation with well defined input and output, and can be easily combined in more complex networks. Their formulation has the following characteristics:

- **Non-linearity in the activation function.** Gene activation is described by the well known Hill function, which represent the binding probability of transcription factors. High non-linearity, which results from the cooperativity term, is necessary for stable attractor states to emerge.

- **Leak.** Each genetic element undergoes non-regulated degradation, modeled as an exponential decay. More complex form of stability could be achieved with autoregulated degradation.

- **Cooperative binding.** Binding of transcription factors to regulatory element is characterized by cooperative binding, which is necessary for the emergence of multiple attractor states.

- **Mutual inhibition with autocatalysis.** Gene elements repress each other and autocatalyze their own expression (bistable or multistable switch). Several examples of such transcription factors are known, especially in the widely studied hematopoietic differentiation pathway (GATA-1 and PU-1 [21]).

- **Susceptibility to asymmetrical division or external regulatory input.** While inputs induce the expression of antagonist genes, the decision toward one of the states necessitate a symmetry breaking event (cell division, synchronization to external influence). This characteristic is particularly useful, since it guarantees that cells perform computational processing on their state only in response to particular events, and provides a simple way to synchronize cells.

5.5.3 Counting divisions without a counter

One of the mysteries of development is how cells are able to coordinate their behavior without a global supervision, such that a defined and completely functional cytoarchitecture is built. Obviously, such a coordination is the results of both genetically encoded
programs and diffusion of positional information in the form of morphogen gradients, although their implementation is less obvious.

I have addressed the issue of the production of a reliable cortical architecture in two ways: firstly, by extracting statistically relevant construction rules from sparse experimental data, and secondly, by proposing a model according to which gene programs can tightly regulate the number of cell divisions, and the differentiation pathways selected by cells. This particular modeling choice has the advantage that it does not require a clock mechanism to explicitly count the number of cell divisions. In contrast, the control over cell numbers is an implicit result of the asymmetric distribution of gene substances, and how gene influence each other expression. Moreover, the composability of the genetic framework eliminates the need of extensive parametric searches, in favor of a topological-directed design.

5.5.4 Reverse engineering of gene regulatory networks

Observations from biological experiments give partial information on the sequences of cell states, and how gene activity correlates with different states. However those observations do not directly address the question on how are genes wired together and on the dynamics of the gene regulatory network.

Based on a simple model of interacting substances and asymmetric distribution, and a mathematical formalisms that links laws of molecular kinetics with a symbolic interpretation of developmental processes, I show that is possible, starting from a known state diagram, to reverse engineer genetic networks to control cell fate specification. An interesting case is when the state diagram does not represent exact transitions but is rather probabilistic, where probabilistic transitions are encoded with a probabilistic distribution over cell division angles.

Although I have described models in which the cell fate distributions are intrinsically encoded, the model can be easily adapted to incorporate external influences. Asymmetry constants can be modulated by external factors and cell-cell interactions. Modulation of transition probabilities is important to impose constraints on the probabilistic process, in other words impose stability (produce a reliable distribution of cells) at the expense of flexibility (the variability in the distribution of cells).
In Chapter 6 I present a more general description of genetic attractor landscapes, and the implementation of a gene regulatory network to control cortical development.
Chapter 6

Simulation of cortical self-construction

At the core of biological self-construction processes lies the DNA, which codes both for functional proteins and the rules on how they are expressed. A collection of DNA elements and their indirect interactions through the expression of transcription factors is referred to as gene regulatory network. Gene regulation has been described with the help of theoretical frameworks in many different modeling and computational approaches. While a popular approach is the use of differential equations, it is useful to have a discrete representation of the system, where genes can assume discretized states in discretized timesteps. Such systems are computationally more tractable than those based on a continuous parametric space, as I have showed in the previous Chapter.

I provide here a review of the relevant literature in the context of biological self-construction processes (Section 6.1), of which cortical development is an example. After a description of the computational model developed in this Thesis, I show the reconstruction of the genetic network responsible for the generation of the mouse cortex in area 3, 4 and 6 (Section 6.3) As a proof of concept, the gene regulatory model has been implemented in Cortex3D (Section 6.4), a java agent-based 3D environment used to simulate growth of cells and tissues.
6.1 Introduction to self-organization in developmental processes

6.1.1 The origin of order

An adult organism, or a tissue, is composed of multiple parts, each originating during the process of embryonic development. In contrast to conventional manufacturing processes, in biology there is no conventional *Bauplan*, in which prefabricated components are combined into a unique structure by a global constructor. Development is the process by which undifferentiated proliferative cells undergo sequential fate restriction, either dictated by their internal genetic program or by external modulation, and organize themselves into functional structures. This construction strategy implies that the process is recursive: previous structures are built on top of previous structures, interact with them and can even modify them. It is this process of replication and continuous feedback that makes development unlike any human manufacturing process.

Combinations of different experimental techniques have substantially improved the understanding of the mechanisms involved in corticogenesis, especially regarding the expression of key transcriptional regulators and their correlations with cell behavior and morphology, although the way in which they are intertwined is still an open question. Theoretical approaches have been used in the recent past to get an insight into how combinations of cellular processes can result in a functional cytoarchitecture [253]: formal models, either in the form of mathematical equations or computational algorithms, can represent cellular or molecular components and predict the result of their interactions. The fundamental question that theoretical models address is to understand, even in some limited context, how the final structural organization is orchestrated by developmental programs. Here, I briefly review the main theoretical contributions to the understanding of self-constructing process, of which cortical development constitute a beautiful example. The most amazing property of biological organisms is their complex structural and functional order¹, which as never ceased to fascinate scientists and led them to ask one of the most fundamental questions in biology: where does molecular order come from?

¹Here I refer to order as a recognizable set of structural patterns, which are reflected at the functional level. Order is generally not a property of chemical reactions, where ordered states may proceed towards a disordered state depending on thermodynamic parameters.
6.1.2 Self-organization has many definitions

6.1.2.1 Self-organization according to Erwin Schrödinger

In the inspiring book *What is Life?*, Schrödinger discusses the origin of order in biology. His conclusion is that there is a fundamental difference between the laws governing physics, and the laws governing biology. In non-living physical systems order arises at the macroscopic level from molecular disorder. Because of thermal energy, atoms and molecules behave probabilistically, but the variance is negligible according to the law of large numbers and at the microscope scale the system can be described by Fick’s deterministic equations. In contrast, he argues that organization in biology cannot proceed from statistical physics, since the estimated number of molecules in each cell is too low. Biology has to rely on the guidance of well defined molecular programs, which defines the molecules that are built, their temporal and spatial position, and finally their interactions.

Schrödinger envisaged biology as a purely deterministic process, in which order arises from order. Genes\(^2\) would be the physical mean by which information gets propagated into cells and regulate their behavior in a mechanisms very similar to the precise regulation of a clock. This metaphor reminds of the molecular biology notion of stereospecificity. In chemistry, stereospecificity refers to the propriety of a reaction mechanisms that specifies how molecules can interact with each other based on their specific 3D conformation. According to this deterministic view, cells functions are explained by the specificity of interactions between molecules, and their ability to exclude chance.

6.1.2.2 Self-organization according to Watson, Crick, Monod, and Fisher

Between 1950 and 1970 a series of important discoveries allowed a more clear formulation of a theory of self-construction: firstly, the discovery of the DNA as the molecule for hereditary information by James D. Watson, Francis Crick and Rosalind Franklin [254]; secondly the discovery of the central dogma of biology [255], according to which information flows from genetic information to mRNA and proteins, and never in the reverse by Jacques Monod; thirdly, the discovery of stereospecificity in molecular interactions, which arose from Fisher’s lock and key model of enzymatic reactions.

\(^2\)At that time DNA was not yet identified as the fundamental material for heredity, and Schrödinger refers to a sort of aperiodic crystal.
The three-dimensional conformation of proteins depends on their linear sequence of amino acids, which in turns depend on the nucleotide sequence of the corresponding gene. According to the central dogma of biology, structure and function arise from the transformation of genetic information to material actuators that represent the cellular machinery. However, for the construction process to be reliable, interactions between molecules have to be highly specific, and in low numbers to avoid combinatorial binding and cross-talk.

6.1.2.3 Self-organization according to Waddington

In the course of his career, the biologist Conrad Hal Waddington made the first attempt to connect the concepts of molecular interactions at the level of genes to their expression in molecular structure and function. Or, in modern terms, the connection between genomics (genes) and phenomics (the gene’s phenotypic expression). Waddington main contribution has been the introduction of the concept of epigenetic landscape: a ball rolling on a surface composed of branching ridges and valleys as the metaphor of a cell undergoing changes in its internal gene expression profile and potentially cell fate decisions [256] (Figure 6.1).

From a system dynamics perspective, cell trajectories are the result of how genes influence each other expression. The epigenetic landscape represent the potential energy surface, and trajectories on this space can be seen has differentiation processes that drive cells from multipotent states toward final differentiated states, or attractors. Epigenetic
mechanisms can influence the conformation of the energy landscape, and are thus responsible for actively channeling cells into specific developmental pathways. Whereas a compelling metaphor for developmental processes, Waddington’s energy landscape has hardly a rigorous physical interpretation.

6.1.2.4 Self-organization according to Stuart Kaufmann

A systematic analysis of genetic space and its effect on the global network behavior has been presented by Stuart Kaufmann with the use of Random Boolean Networks (RBNs) [257, 258]. RBNs represent a simple model for gene regulatory networks. Genes are represented by a set of $N$ nodes, each with two possible states (on or off), indicating whether the gene is active or not. Nodes are connected with direct edges, where each node receives input from $K$ randomly chosen genes. The update function of the gene activity in the next timestep is chosen from a set of logic functions with a given probability, whereas the network topology and Boolean functions do not change in time.

The behavior of RBNs is described by depicting the trajectories of each state toward the attraction basins (Figure 6.2). RBNs display three different types of behavior, either ordered networks with short attractors or limit cycles, disordered networks with highly sensitivity to initial conditions, and critical networks, at the edge between the two previous behaviors. Despite their simplicity, critical RBNs have a developmental interpretation: different basins of attraction represent different cell differentiation pathways, which lead to distinct final cell types. Unfortunately, Kaufmann’s models do not yet explain how decisions between alternative cell fates are taken.
6.1.2.5  **Self-organization according to Von Neumann**

While most of the research and philosophical thinking has been concerned with the link between genes and and their phenotypic expression in distinct cell types, a crucial aspect is the quantization of genetic information by virtue of the cell membrane and the need of molecular and cellular replication.

Von Neumann’s theory of self-replicating machines (automata) [259], which predates the discovery of the DNA structure, has many analogies with cellular replication in living organisms. An automata is composed of a *universal constructor* $A$, which is responsible to replicate the hardware $D$ (composed by parts $A$, $B$, and $C$) according to the *explicit instructions* encoded in a tape $I$. A copier $B$ is responsible to copy the information tape, and both $A$ and $B$ are controlled by a controller $C$.

From a biological perspective, the code stored in $I$ corresponds to the DNA, $A$ is the transcriptional/translational machinery, $B$ is the DNA replication machinery, $C$ is the regulation by proteins and mRNA, and $D$ represents the entire molecular environment of the cell. However, similarities ends here. Von Neumann used an explicit one-to-one mapping between the description of the system (or blueprint) and the target system, whereas the encoding of an organism in the DNA is implicit. The indirect mapping of the genotype to the phenotype provides a powerful mechanism for incremental construction, where previous part of the system can guide the construction process by feedback loops, and is thus more resistant to noise and failure.

6.1.3  **Cortical development as a self-constructing process**

I address the problem of self-construction in the context of cortical development, since self-constructing rules play a decisive roles into shaping different cortical architectures, both in structure and function. In contrast to the more theoretical work I have reviewed, I approach the problem from a practical perspective: given a set of experimental sparse data, I wish to infer growth rules and describe them with an appropriate configuration language for developmental processes.

Cortical development requires the coordination of five primary processes: cell growth, cell replication, cell division (mitosis), cell differentiation, and morphogenesis, whereas the progression through the different phases of the cell life is controlled by the cell cycle.
During corticogenesis, cascades of self-organizing processes arise from local interactions between molecular components. The interplay between different mechanisms makes difficult to grasp the effect of local rules on the global system behavior, thus the need of computational models [192]. The characteristic of computational models is that they are composed of rules, according to which biological entities such as cells are characterized by states and state transitions depending on internal and neighboring components, and are thus ideal for the qualitative understanding of complex processes.

The sequence-dependent control characteristic of developmental programs is very different from conventional human constructing procedures. Conventional construction relies on a global plan (the “blueprint”) which describes each single component and their relative positions to each other. In contrast, self-constructing structures have the following characteristics:

- **Local.** Development is regulated on a local scale only. There is no single component that perform the role of a global controlling center.

- **Distributed.** Genetic programs are distributed asymmetrically among cells, which guarantees that different cells perform different, specific tasks.

- **Event-based.** No general clock controls the behavior of cells, but rather the responsiveness to asynchronous events and the mutual synchronization between cells.

- **Programmable.** Genes and their regulatory sequences are the computational units of the genetic code, which can be reused, recombined and modified by evolution to create new shapes.

- **Robust.** The product of development is never exactly the same, because is the result of concurring events. Despite variation, construction of biological systems is extremely reliable in term of the specification of the final target structure.
6.2 Objectives of this Chapter

Building on top of the previous Chapters, I will now present the reconstruction of a gene regulatory network for the cortical development in the mouse and its simulation in an artificial physical environment, where locality constraints cannot be bypassed.

Objectives for this Chapter are:

- **Reverse engineering of genetic regulatory networks.** In the context of this Thesis, I give a theoretical description of how genetic programs control developmental processes. As an excellent example, I show the reconstruction of an abstract genetic network for the control of cortical lamination in area 3 and 6 of the mouse.

- **Implementation in Cortex3D.** As a proof of concept I implemented the model in an agent-based software, Cortex3D, developed by Frederic Zubler and Andreas Hauri, where we have simulated the generation of murine cortical columns in area 3 and 6 starting from an homogeneous sheet of precursor pools.
6.3 Methods: A gene regulatory network for cortical areas 3 and 6

I artificially construct regulatory networks by interconnecting small modules that can control the progression of cells through developmental sequences. The model is formulated in an high level language that projects directly to the kinetic differential equations, such as scripting languages can be complied in running code.

Here I provide the G-code implementation as series of interconnected bistable switches that code for the gene regulatory network controlling the lamination in area 3, 4 and 6 of the murine cortex. The network replicates the state transitions obtained in previous Chapters, but was manually simplified in order to reduce the production overlap in time of neurons fated to different cortical laminae. As I will explain later on, this modification was necessary to improve the stability of the cortical growth simulation.

The gene network is divided into two sections: a list of interacting genes coding for transcription factors, and a list of read-out genes, which become active when specific gene expression profiles occur, and are responsible for instantiating the cell differentiation and/or migration machinery. The function bistableSwitch refers to the bistable switch described in the previous Chapter, the function monostableSwitch is used to encode symmetric divisions, and the function bistableSwitchWithControl is a bistable switch with a limit on the maximal number of cell divisions.

```plaintext
// run transcription factors
startGene("g0");
bistableSwitchWithControl("gM","g0",div(7));
bistableSwitchWithControl("gS","gM_C",div(7));
monostableSwitch("g1","gS_C",div(0));
bistableSwitchWithControl("g231","g1",div(1));
monostableSwitch("g2","g231_A","g231_B");
bistableSwitchWithControl("g23","g1",div(1));
bistableSwitch("g45","g23_A");
bistableSwitch("g67","g23_C");
bistableSwitch("g1213","g67_A");
bistableSwitch("g89","g45_A");
bistableSwitch("g2425","g1213_A");
bistableSwitch("g1617","g89_A");
bistableSwitch("g1819","g1617_B");
bistableSwitch("g3839","g2425_A","g1617_A");
bistableSwitch("g4849","g3839_A","g3839_B");
bistableSwitch("g3132","g89_B");
bistableSwitch("g3334","g3132_A");
```
monostableSwitch("g7","g45_B","g67_B",div(2));
monostableSwitch("g13","g1213_B",div(2));
monostableSwitch("g17","g1617_B",div(2));
monostableSwitch("g24","g3839_A","g3839_B",div(2));

// run read-out genes
if(instanciateStopMachine("gM_A", val(0.99)).eval())
    instantiate(new Differentiate(Color.DARK_GRAY, "L1")); // layer1 promoter
if(instanciateStopMachine("g3_A", val(0.99)).eval())
    instantiate(new Differentiate(Color.LIGHT_GRAY, "sub")); // subplate promoter
if(instanciateMachine("g23_A", "g23_C", val(0.99)).eval())
{
    if(LNot(readMorphIsPropertyThisValue("rgc","t")).eval())
    {
        instantiate(new Differentiate(Color.BLACK, "rgc")); // RGC promoter
        writeMorphProperty("rgc", "t");
    }
}
if(instanciateStopMachine("g7", val(0.99)).eval())
    instantiate(new Differentiate(Color.BLUE, "L6")); // layer 6 promoter
if(instanciateStopMachine("g13", "g3334_B", "g1819_B", val(0.99)).eval())
    instantiate(new Differentiate(Color.CYAN, "L5")); // layer 5 promoter
if(instanciateStopMachine("g2425_B", "g89_B", val(0.99)).eval())
    instantiate(new Differentiate(Color.PINK, "glia")); // glia promoter
if(instanciateMachine("g3839_A", "g3839_B", val(0.99)).eval())
{
    if(LNot(readMorphIsPropertyThisValue("svz","t")).eval())
    {
        instantiate(new Differentiate(Color.YELLOW, "svz")); // svz promoter
        writeMorphProperty("svz", "t");
    }
}
if(instanciateStopMachine("g4849_B", "g17", val(0.99)).eval())
    instantiate(new Differentiate(Color.ORANGE, "L4")); // layer 4
if(instanciateStopMachine("g24", val(0.99)).eval())
    instantiate(new Differentiate(Color.RED, "L2-3")); // layer 2/3

// run cell cycle
cellCycle();
Chapter 6. Simulation of cortical self-construction

6.4 Results: Simulation of cortical lamination

6.4.1 The attractor landscape

I represent the concept of genetic programs with dynamical system theory. A dynamical system is composed of a set of functions that describe the evolution of the system’s variables in time in dependence of each other. Each variable correspond to the expression of a gene (or of an epigenetic marker), and at every timepoint a cell is associated with a specific location in the genetic space (gene expression profile). Every point in this high dimensional space completely define the internal state of the cell, is associated with a vector that represents the direction and velocity of the system evolution, that is the evolution of the internal gene expression profile.

Few isolated points in the genetic space are attractors toward which the system, i.e. cells, move spontaneously. During development all cells obey to the same physical laws and bear the same genetic information, but because of the ability of cells to asymmetrically divide their own molecular content in the daughter cells, and because of the existence of inducing signals, cells can jump between different regions, follow different pathways and end up into different final attractors (terminal differentiation). In this way an homogeneous pool of cells can lead to the generation of diverse cell types, defined as attractor regions in the genetic landscape (Figure 6.3A). The genetic attractor landscape
is defined as a potential function $P$, which defines the amount of energy required to displace a cell of a unitary point in the gene concentration space.

While very appealing conceptually, continuous genetic landscapes are notoriously difficult to design other than for very simple genetic circuits. The reason for their limited practicality reside in the huge parametric space in which they are defined, since slight modifications in gene constants or interactions could cause dramatic shifts in the landscape. I circumvent this problem by constructing genetic networks in a discretized space composed of defined states, and by combining genetic modules (bistable and multistable switches), which are characterized by a well defined behavior.

Dynamic systems that have discrete behavior can be described in term of states (basins of attraction), and state transitions (jumps between basins of attraction). Under this assumption, the dynamics is analyzed in the mathematical setting of state transition systems, where both space and time are discretized. There are several reasons that justify the use of discrete models to describe dynamic systems. Among them, the necessity to approximate the system behavior with a more computationally tractable model, and the possibility of quantitative measurements of the system properties.

Similar to Kaufmann’s models, I describe the dynamical system with a set of discrete variables, where in this case variables can assume more than two values, and variable combinations describe the system in terms of its attractor basins (Figure 6.3B). The specific wiring of genes creates an attractor landscape, where cells can either move smoothly on its surface, or perform jumps at the time of cell division. Depending on whether jumps are symmetric or asymmetric, cells can bypass high energy barriers and reach new states previously inaccessible.

6.4.2 The attractor landscape of bistable switches

I illustrate the idea of the genetic landscape on the simple case of a bistable switch (Figure 6.4). In the previous Chapter, I have proposed a formal model that can link kinetic equations with continuous Boolean abstract representation of network topologies. The advantage of a continuous Boolean network model is the ability to link continuous gene expression to discrete behavior. Such systems is computationally more tractable than those based on differential equations, but closed enough to molecular biology to be used in detail cases.
Figure 6.4. Simulation of the dynamics of a bistable genetic switch. (A) Structure of the network with mutual inhibition between the two transcription factors $X_1$ and $X_2$ and with auto-stimulation. The switch can be modulated by external or internal inputs and can provide some outputs for downstream signaling. (B) Attractor landscape representing the state probability distribution in function of $X_1$ and $X_2$ concentrations for the switch with input $I = 0$. The system as a unique attractor state $C$ at $[X_1] = 0$ and $[X_2] = 0$. (C) Attractor landscape representing the state probability distribution in function of $X_1$ and $X_2$ concentrations for the switch with input $I = 1$. The saddle point $C$ corresponds to the original cell type, whereas attractor $A$ and $B$ represent downstream differentiation pathways. (D) 2-dimension Sammon projection of the genetic landscape, where nodes represent cells, and vectors their trajectory. (E) Schematics of 2-dimension Sammon projection with highlighted attractor states (dashed black lines) and transitions between attractor states (black arrows).

Since the visualization of the gene network in a high dimensional space is not easy to grasp, we can imagine of projecting all dimensions into two dimension only, by discarding irrelevant information and preserving topological shape. Here I use the Sammon projection algorithm, which is used to map a high-dimensional space to a space of lower dimensionality by preserving the structure of inter-point distances. The two dimensional space is the potential surface, whose folds determines the behavior of cells.

Bistable switches are a particularly interesting mechanism also from a biological point of view. Firstly, they are highly resistant to noise perturbations. The existence of attractor states guarantees that cells will tend to aggregate in a discrete number of stable states in the genetic state space, and that small variations in gene concentration do not
Figure 6.5. **Noise effects on bistable switch genetic landscape.** Simulation on the effect on noise on a bistable genetic switch. Cell trajectories are displayed with a 2-dimension Sammon projection, where nodes represent cells, and vectors their trajectory. (A) 0% noise, (B) 10% noise, (C) 30% noise, (D) 50% noise.

influence the behavior of the genetic network. For instance, my implementation displays robustness up to 10% of noise, which is rather remarkable and guarantees reliable computation even in perturbed environments (Figure 6.5). A second useful feature of bistable switches is that jumps across the landscape can happen only at the time of cell division, and thus the computation of the genetic network does not need to be perfectly synchronized with the cell cycle progression (variations in the duration of cell cycle time do not influence the choice of cell fates).

In this particular model it is very simple to control the number of cells and their distribution across different states even under noisy conditions. Cells can either travel across the genetic landscape because of the interaction between genes, or perform jumps at
the moment of cell division, which allow them to assume states that would be otherwise unreachable. Therefore, simply by the process of cell division, more states became accessible as development unfolds. The partial irreversibility of jumps in the genetic landscape (DNA primary structure) is assured by epigenetic modifications (DNA secondary structure). Here it is implemented as a dependency of the asymmetry constants on the gene product concentrations.

6.4.3 Building of a laminated volume of cortex in CX3D

Biological development can be viewed as the sequential progression of cells through different gene expression profiles in time and space. The attractor landscape explains how cells can switch between cell fates during the course of development, but does not explain how cells can progress along differentiation pathways, and necessitate the introduction of epigenetic events that prevent the reactivation of previous or alternative cell fates. The progression of cells into the differentiation cascades can be modeled with multiple bistable switches, where every switch represents a branch in the differentiation decision tree. For instance the concentration of a gene can function as input of a second bistable switch and so on. This procedure permits the design of genetic programs for the growth of a wide range of cell populations, beginning from a single progenitor cell.

The bistable switch system is used here for modeling the transition form one cell type to the next in the context of cortical lamination in area 3 and 6 of the murine neocortex. The gene regulatory network was designed by identification of statistically recurring patterns of cell divisions, and assigning to them a control switch for cell fate decisions. The resulting series of bistable switches used to model the generation of 5 different neuronal types is depicted in Figure 6.6. State transitions are occurring only at the moment of cell division, when the asymmetric distribution of substances will determine the trajectory of the gene expression profiles in the two daughter cells.

In the present work I use a CX3D high-level machine code to implement cell division, differentiation and migration and obtain a laminated volume of murine cortex (Figure 6.7). The programming language has been designed on a set of carefully chosen primitive functions (read, write, die, morph, etc.), which guarantees the locality of the computation and restrict the possible actions. A complete description of the language and simulations details is in Frederic Zubler and Andreas Hauri Thesis.
We start the simulation with an array of progenitor cells of the VZ, which undergo symmetrical and asymmetrical divisions to form the desired number of neuronal precursors. Expression of intracellular substances is used to represent the internal state of the cell and changes in concentrations at cell division allows the transition to other states. Precursors cells that are instructed to exit the cell cycle will then migrate along the radial glial cell processes and form the cortical plate in a inside-out manner. Migration is directed by local integration of guidance cues and stopping signals, which are established by the cells themselves according to the internal gene profile of expression.
6.4.3.1 Preplate formation

Future neurons are formed in the ventricular zone and subventricular zone, and migrate through the preplate, composed of the subplate and of the future the marginal zone. In multicellular organisms, the initial cell has a polarity. The two poles contain different concentrations of maternal proteins and mRNAs which are specifically partitioned during the first divisions. The simulation requires similar initial conditions. I suppose in our first cell, the existence of an North-South internal axis, which corresponds experimentally to radial glial fibers connecting RCG precursors to the pia.

The first steps in the simulation process is be to form an initial neural epithelium by a series of symmetrical divisions (with the division axis staying in one plane). This step doesn’t require any external guidance, provided that the cells can use their internal polarity to organize the division axis. At this point cells progress into a sequence of asymmetric divisions that lead to the production of the marginal zone and subplate cells, which from the early preplate.

6.4.3.2 Cortical plate formation

After the preplate formation, the cells in the ventricular zone start to divide and initiate the cortical plate formation, i.e. the establishment of a layered structure of neural precursors. This developmental phase can be subdivided into two parts: (1) the generation of the correct number of neuron precursors of each cell type and (2) their migration at the right position to form a layered structure. The generation of cells with different behaviors is determined by the gene transcriptional network and by epigenetic regulatory mechanisms. The distinction between area 3 and 6 is induced by diffusible morphogens, that pattern the cortical surface with a predefined left/right gradient. Ideally those substance would be produced by incoming thalamic afferents.

Cell division is controlled by a cell cycle machine that induces cell cycle progression in precursor cells. As long as it is not stopped, cells divide at with an average cell cycle time, and distributes intracellular substances according to their asymmetry constants. When certain conditions on the gene regulatory network genes are met, few read-out genes are then activated, the cells stop dividing and express the cellular machinery necessary for terminal differentiation and migration upward in the cortical plate.
Figure 6.7. Simulation of mouse cortical lamination for area 3, 4 and 6. (A) Neural plate (E9). (B) Formation of the marginal zone, subplate and radial glial cells (E10). (C) Proliferation of progenitors in the ventricular zone (E11). (D) Generation of cortical plate (deep layers) and subventricular zone (E13). (E) Generation of cortical plate (superficial layers) (E16). Gray, radial glial cells; Brown, subventricular zone; Red, Layer VI; Green, Layer V; Blue, Layer VI; Cyan, Layer II/III; Yellow, Marginal Zone; Pink, apoptotic cells. Figures from Andreas Hauri Thesis.
6.5 Discussion

6.5.1 Reverse engineering of cortical gene regulatory networks

Cortical development is a remarkable example of self-constructing system, where cell proliferation, differentiation and migration play an essential role in establishing a premature cortex, where connections between neurons are eventually established. The regulation of gene expression plays a key role in development and function of the cortical tissue. Every cell is characterized by a specific gene expression profile, which defines the cell behavior and its responsiveness to external cues. Cell decisions regarding acquisition of an appropriate cell fate relay on the ability to activate particular gene expression configurations.

The epigenetic landscape first proposed by Waddington describes branching pathways of cell fate determination and differentiation. It is represented as a surface, attached to array of pegs (genes) by a network of ropes (gene interactions), on which balls (cells) roll according to the surface conformation. This metaphoric image conveys the idea that the differentiation potential surface is controlled by how gene interact and their epigenetic modifications.

Biological development can be viewed as the sequential progression of cells through different gene expression profiles in time and space. The potential landscape explains how cells can switch between cell fates during the course of development. The underlying regulatory network defines molecular attractor states on the landscape, toward which cells are drawn, whereas fluctuations and asymmetries in the gene expression profiles causes cells to transit between different attractor regions.

Despite interesting, those metaphors require a more explicit mathematical and computational formulation. Here I propose a suitable gene regulatory network model, referred to as Gene Reaction Networks based on continuous Boolean Networks. The Gene Reaction Networks model is derived directly from the laws of reaction kinetics. The proposed models has three advantages: (i) it captures the combinational power of gene transcriptional regulation (continuous mathematical formulation), but (ii) has a direct representation at the logic level (discrete computational formulation), and (iii) can be explicitly used to encode state models, as the one derived in Chapter 4.
6.5.2 Simulation of the mouse cortical lamination

The genetic specification of the intracellular regulatory network defines the reachability of particular gene expression profiles. Cells are allowed to reach only a limited number of states, which are specified in the regulatory genome. Therefore, a large number of possible input conditions can lead to only a limited number of terminal states.

Under this assumption, I build a simple model composed of interacting genes, and asymmetric distribution of gene substances at the moment of cell division to emulate differentiation processes. Specifically, I make extensive use of bistable genetic switches to encode cell fate branching points, since binary cell fate decisions based on two competing transcription factors have been observed repetitively, for instance in hematopoietic lineages [242, 251].

As an example, I reconstructed the gene regulatory network for the generation of murine cortical columns in area 3, 4, and 6, as described by the state model obtained in Chapter 4. The generation of a laminated cortical structure was achieved by implementing the gene regulatory network into Cortex3D, a java agent-based 3D environment used to simulate growth of cells and tissues developed by Frederic Zubler and Andreas Hauri [260]. A full description of the cortical simulation is provided in Andreas Hauri Thesis.

A single gene regulatory network is responsible for the generation of all areas, and the area specificity is determined by an environmental signal that regulates the balance between different precursor cells. In the current simulation, the specification of the area is provided by an initial artificial gradient aligned with the medial/lateral axes, whose concentration is detected by a gene in the gene regulatory network. Ideally, the environmental cue would be provided by incoming thalamic afferents, which have been showed to regulate cellular proliferation in the germinal layers during development [26, 27].

Because the quality of the cortical lamination (correct positioning of cells and clear separation between different layers) is affected by the temporal generation of different neuronal types in the germinal layers, we considered a simplified genetic network for which the production overlap in time of cells fated to different layers is greatly reduced. A more realistic implementation would require the introduction of additional synchronization mechanisms to coordinate the migration of newborn neurons into the cortical plate.
Chapter 7

Conclusions

On December 29, 1959, the physicist Richard Feynman gave an inspiring and visionary lecture on the future of computation entitled *There is Plenty of Room at the Bottom*. He described the possibility of building computing machines at the sub-microscopic level, or nanoscale. The machine would work thanks to direct manipulation of individual atoms, which implied nanomolecules able to replicate and self-assemble. In contrast to conventional building of computers (top-down approach), nano-computers would be expected to assemble from the results of their molecular interactions (bottom-up approach). At the time, this talk was revolutionary. Half a century later, it still is.

Major advances has been done since in the field of molecular, and recently, synthetic biology, which represent a direct biological counterpart of Feynman’s dream. In this Thesis I studied cellular and genetic self-construction processes in the context of cortical development. Self-construction is defined in terms of a set of local rules that express themselves at the genetic level, and whose influences propagate at different scales to organize a plate of progenitor cells into a cortical architecture.

The main outcomes of this Thesis are discussed below, with reference to the wider research context and future work directions.
7.1 Summary

7.1.1 Symmetric and asymmetric divisions define the final cortical cytoarchitecture

The dynamics of corticogenesis is determined by the proliferative behavior of cells and the length of the cell cycle. In this Thesis I have focused specifically on two areas in the murine neocortex, area 3 and 6, which are characterized by different proportions in the number of cells produced for each lamina. The comparison between the two areas allows the investigation of the mechanisms that modulate neurogenesis.

Experimental findings show how the temporal changes in the cell-cycle parameters generate different numbers of neurons in successive cortical layers [17]: an increase in the cell cycle exit probability leads to a transient increase in neuron production because of a rapid exhaustion of the precursor pool. In contrast, increase in the rate of proliferation results in a dramatically higher neurogenic output through the amplification of the precursor pool.

I have used differential equations (ODE) to model the dynamics of neurogenesis from experimental data on cell cycle parameters at different timepoint during development. The ODE model gives an accurate prediction of the number and types of neurons produced starting from a single precursor at the onset of neurogenesis. The main ODE model prediction is a consistent expansion of the granular layer (IV) at the expense of layer V in area 3 with respect to area 6. Area 6 is characterized by a unique proliferative peak at E13.5, followed by a sustained differentiative peak at E16. In contrast, area 3 has two differentiative peaks at E14.5 and E17, with an intermediate proliferative peak at E15.5. These differences have two main effects: firstly, the higher differentiation rates in area 3 at early stages of neurogenesis reduce the pool of cells for infragranular layers, layer V in particular; secondly, the proliferative peak beginning at E16 contributes to the expansion of layer IV.

More interestingly, while the predicted cumulative number of cells is quite sensitive to errors in the estimation of the cell cycle exit probability, deviations in cell cycle length do not have a significant impact on the shape of the cortical architecture. In contrast, experimental perturbation of cell cycle length has in a contrast dramatic effects on the numbers and proportions of cells produced [156]. This results lead us to the hypothesis that the core regulatory mechanisms responsible to shape the final cortical architecture
is the asymmetric distribution of cell fates among daughter cells, and change in cell cycle length have to cause changes in the mode of divisions in order to perturb neurogenesis.

### 7.1.2 Reconstruction of the Cell Lineage Tree

Current experimental methods for the study of cortical development are driven by the need to understand the molecular basis of modes of cell division and cell fate decisions at a single cell resolution. The information of the mapping between divisions and cell fates determinants are explained by the *cell lineage tree*, which describes the genealogical history of each individual cell.

I have chosen to investigate models of corticogenesis in the discretized space of lineage trees for two reasons: firstly, the cell lineage tree represents a very concise description of developmental processes; and secondly, given the increasing interest and experimental effort to collect information on branching processes (cell lineage trees, dendritic and axonal arborizations, etc.), I propose a general algorithmic approach that can be used to model any type of experimentally observable branching growth process.

Cell lineage data is notoriously difficult to track, especially in mammals – because of the opaque body and the tremendous amount of cells produced. It is however possible to measure partial cell lineage data that contains sparse label information, corresponding to different observed features such as gene expression data, cell morphology or other attributes. A recent, promising technique to collect such data is clonal barcoding [149], where labeled retroviral vectors with random sequence tags (barcodes) are integrated into the host genome at specific developmental stages through cell transplantation. The clonal progeny of each cell can be tracked in time by single-cell sequencing and can be combined with micro-array gene expression analysis of the final cells.

Given the current technical limitations, and also because of the more theoretical orientation of this Thesis, I reconstruct cell lineage trees through sampling of the probability distribution described by experiments on mitotic history [17] (Figure 7.1). The dataset of lineage trees is probabilistic, in the sense that although constrained by the experimental data, it may display features that do not occur in real corticogenesis. Despite uncertainties, lineages of area 3 and 6 are statistically significantly different.
Chapter 7. Conclusions

Figure 7.1. Probabilistic reconstruction of lineage trees. Lineage trees are reconstructed by sampling from the probability distribution, measured experimentally. (A,C) Probability distribution for area 3 and 6. Points, experimental data; lines, fitted data. (B,D) Example of sampled lineage trees (time layout). Black: precursor cell; blue: layer VIb; green: layer VIb; yellow: layer V; orange: layer IV; red: layer II-III; dashed lines, proliferation of glial precursor cells (not modeled).

7.1.3 State models of corticogenesis

Cell lineages are a valuable dataset since they comprise information not only about single cells, but also the spatial and genealogical relationship among them. Developmental programs are reflected in the structure of the cell lineage tree: the tree topology describes the relationship between all cells that existed at given time point during development and the fate of their progeny.

Formally, a cell lineage can be represented as a directed graph, where each node represents a cell instance, and each edge a genealogical relationship. In this description, each node is associated with a vector, whose elements correspond to measured features. Statistical regularities in large, unordered dataset, can be automatically extracted by clustering, an unsupervised machine learning method. Usually, clustering is used to find groups of data points that have similar distribution of features. To include lineage information, we have to introduce conditional dependencies between data points.

I propose a method inspired on spectral clustering to identify statistically significant recurrence of division patterns, and express cell lineage data with a compressed representation of states associated with feature distributions, and state transitions. The proposed method applies to directed acyclic graphs and has the advantage that requires only semi-labeled data (cell lineage in which only terminal cell are labeled according to observed features). In order to test the spectral decomposition performance, I have analyzed artificially generated dataset of cell lineages (for which the generative model is known) and showed that the underlying model can be successfully recovered (100% success rate for deterministic models, and 80% for probabilistic models).
I have analyzed 60 reconstructed lineages from area 3 and 6 of the mouse cortex, for a total of 3214 cell instances (1549 in area 3 and 1714 in area 6). Spectral decomposition and dimensionality reduction was performed for both area 3 and 6 on the combined dataset, which allow to exploit the similarities between the two areas. While the model is computed over the combined cell lineage dataset, it is easy to extract individual models for each area separately. Remarkably, the reduced models for area 3 and 6 are strikingly similar. This suggests that a unique model with minimal parametric differences is able to explain the generation of both areas.

The 10 dimension state model explains 82% of the data, and it is the most visually intuitive to reason about the logic behind the development of area 3 (Figure 7.2). The black node represent an originally homogeneous population of precursor cells, which divide into subpopulations of precursor cells with different neurogenic potential: a small proportion of cells fated exclusively to infragranular layers (VIa, VIb, and V), cells fated very early on exclusively toward granular/supragranular layers (IV, II/III), and a big pool of heterogeneous proliferative cells that are less fate restricted. Transitions between those pools appear to be plastic, and partially reversible.

According to the emerging model, lineages are characterized by mixed populations of progenitor cells: multipotent progenitors, and fate-restricted progenitors toward deeper or superficial layers. Moreover, superficial layer fate-restricted progenitor cells increase in number over time. This finding is rather interesting, since it seems to reconcile two opposite experimental evidences, that is the concomitant presence of multi-potent [127] and fate restricted progenitors in the neural precursor pool [106].
7.1.4 Transcriptional network models of corticogenesis

Biological development can be viewed as the sequential progression of cells through different gene expression profiles in time and space. One of the mysteries of development is how cells are able to coordinate their behavior without a global supervision, such that a defined and completely functional cytoarchitecture is built. Obviously, such a coordination is the results of both genetically encoded programs and diffusion of positional information in the form of morphogen gradients, although their implementation is less obvious.

I have addressed the issue of the production of a reliable cortical architecture in two ways: firstly, by extracting statistically relevant construction rules from sparse experimental data, and secondly, by proposing a model according to which gene programs can tightly regulate the number of cell divisions, and the differentiation pathways selected by cells. This particular modeling choice has the advantage that it does not require a clock mechanism to explicitly count the number of cell divisions. In contrast, the control over cell numbers is an implicit result of the asymmetric distribution of gene substances, and how gene influence each other expression.

The specific wiring of gene regulatory network defines the possible cell states, and their relationships in term of state transitions. I visualize the behavior of this dynamical system in an attractor landscape, where for each point in the genetic expression space (concentration of genes) corresponds a potential value, proportional to the likelihood of occupancy of that point. The landscape explains how cells can switch between cell fates during the course of development. The underlying regulatory network defines also the molecular attractor states, toward which cells are drawn, whereas fluctuations and asymmetries in the gene expression profiles causes cells to transit between different attractor regions.

Despite interesting, those metaphors require an explicit mathematical and computational formulation. Here I propose a suitable gene regulatory network model, referred to as Gene Reaction Networks based on continuous Boolean Networks. The Gene Reaction Networks model is derived directly from the laws of reaction kinetics. By using a simple model of interacting substances and asymmetric distribution, I show that is possible, starting from a given state diagram, to reverse engineer genetic networks controlling cell fate specification. Specifically, I make extensive use of bistable genetic switches to
encode cell fate branching points, since binary cell fate decisions based on two competing transcription factors have been observed repetitively, for instance in hematopoietic lineages [242, 251]. The composability of the genetic framework eliminates the need of extensive parametric searches, in favor of a topological-directed design.

The inferred sequences of state, and state transitions, with the associated feature distributions, allows to make some predictions about the underlying transcriptional network. Although I am not able to infer direct interactions between transcription factors, it is possible to identify patterns of commonly regulated transcription factors. I use this information to reconstruct a plausible transcriptional network models. As an example, I reconstructed the gene regulatory network for the generation of murine cortical columns in area 3, 4, and 6 and provide an implementation in Cortx3D, which is a programming language particularly suited to simulate developmental programs. A single gene regulatory network is responsible for the generation of all areas, and the area specificity is determined by an environmental signal that regulates the balance between different precursor cells.

The idea of a cell as a dynamical system governed by gene regulatory networks has undergone a revival in light of recent discoveries regarding reprogrammability of mitotic cells [261]. Even terminally differentiated cells have been (partially) reprogrammed to express behavior of more plastic cell states or alternative lineages. These findings challenge the dogma of irreversible cellular differentiation, and suggests that models of transcriptional regulation at the systemic level would be highly valuable as tools to investigate the control of fates, and fate plasticity in stem and multipotent cells.
7.2 Outlook

Current experimental methods for the study of cortical development are driven by the need to understand the molecular basis of cell proliferation and cell fate decisions at a single cell resolution. The emerging strategy is the tracking of cell lineages by mapping cohorts of progenitor cells in their native environment at high spatial and temporal resolution, ideally down to the molecular level. Cell lineage data describes the genealogical history of single cortical neurons and offers a powerful insight into patterns of cell division and differentiation, information that cannot be obtained with the traditional approach of considering progenitor cells composed of homogeneous subpopulations.

The power of the method proposed in this Thesis, referred to as spectral decomposition, relies in its ability to generate data-driven models of transcriptional regulation from sparse cell lineage information, as opposed to imposing an ad-hoc model. More generally, the proposed method is interesting for it’s ability to generate hypothesis for the growth of tree-shaped structures, of which cell lineages are just an example.

Spectral decomposition identifies statistically relevant branching patterns by taking into consideration the branching structure and the final phenotypes of the three leaves. I have shown that identification of recursive branching patterns can be used to infer a compact Hidden Markov Model, which can be used has a generative model. Obviously, the quality of the model strongly depends on the accuracy of the collected lineage information, and on the labels of the differentiated cells. This implies that the model can be refined by collection of more precise datasets.

While it is not possible to infer direct causal link between transcription factors, I can infer regulatory rules between family of genes differentially regulated in different cell states. This information is not only interesting to investigate the link between gene regulation and developmental programs, but this information could be exploited to experimentally manipulate cell fate determination.

In conclusion, the cell lineage information is crucial since it directly captures the sequence and modes of cell divisions, which – as we have seen from population based models - lies at the heart of developmental processes. I envisage that such approaches are going to be decisive to decipher the genetic regulation of developmental processes.
Appendix A

Experimental data

A.1 Cell cycle parameters of murine cortex in area 3 and 6

Cell cycle parameters such as average cell cycle exit probability (leaving fraction $k_Q$), cell cycle rate (cell cycle length $T_c$), and laminar probability (laminar fraction $p_X$), were experimentally measured during murine corticogenesis from E11 till E18.5 by Polleux et al. [17] in area 3 (parietal) and 6 (frontal) of the mouse cortex (Table A.1).

<table>
<thead>
<tr>
<th>Leaving fraction $k_Q$</th>
<th>11.0E</th>
<th>12.0E</th>
<th>12.5E</th>
<th>13.5E</th>
<th>14.0E</th>
<th>14.5E</th>
<th>15.5E</th>
<th>16.0E</th>
<th>16.5E</th>
<th>17.5E</th>
<th>18.5E</th>
</tr>
</thead>
<tbody>
<tr>
<td>Area 3</td>
<td>0.00</td>
<td>0.10</td>
<td>0.30</td>
<td>0.54</td>
<td>0.49</td>
<td>0.30</td>
<td>0.93</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Area 6</td>
<td>0.00</td>
<td>0.10</td>
<td>0.30</td>
<td>0.54</td>
<td>0.68</td>
<td>0.93</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cell cycle length $T_c$</th>
<th>11.0E</th>
<th>12.0E</th>
<th>12.5E</th>
<th>13.5E</th>
<th>14.0E</th>
<th>14.5E</th>
<th>15.5E</th>
<th>16.0E</th>
<th>16.5E</th>
<th>17.5E</th>
<th>18.5E</th>
</tr>
</thead>
<tbody>
<tr>
<td>Area 3</td>
<td>8.00</td>
<td>15.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Area 6</td>
<td>10.00</td>
<td>15.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Laminar fraction $p_X$</th>
<th>11.0E</th>
<th>12.0E</th>
<th>12.5E</th>
<th>13.5E</th>
<th>14.0E</th>
<th>14.5E</th>
<th>15.5E</th>
<th>16.0E</th>
<th>16.5E</th>
<th>17.5E</th>
<th>18.5E</th>
</tr>
</thead>
<tbody>
<tr>
<td>Area 3 Vlb</td>
<td>1.00</td>
<td>0.90</td>
<td>0.75</td>
<td>0.45</td>
<td>0.20</td>
<td>0.05</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Vlb</td>
<td>0.20</td>
<td>0.30</td>
<td>0.65</td>
<td>0.71</td>
<td>0.46</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>VI</td>
<td>0.21</td>
<td>0.20</td>
<td>0.27</td>
<td>0.49</td>
<td>0.23</td>
<td>0.07</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>IV</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.05</td>
<td>0.74</td>
<td>0.78</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>II/III</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Area 6 Vlb</td>
<td>1.00</td>
<td>0.90</td>
<td>0.75</td>
<td>0.45</td>
<td>0.20</td>
<td>0.05</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Vlb</td>
<td>0.25</td>
<td>0.55</td>
<td>0.60</td>
<td>0.74</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>VI</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>IV</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>II/III</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Table A.1. Cell cycle parameters. Leaving fraction $k_Q$, cell cycle length $T_c$, and laminar probabilities $p_X$ in function of embryonic day from E11 to E18.5 in murine area 3 and 6.
A.2 Number of neurons in the adult rodent cerebral cortex

Laminar thickness (Table A.2), densities of neurons in different laminae of the adult cerebral cortex (Table A.3) and number of neurons in a column under a cortical surface of 1 \( \text{mm}^2 \) (Table A.4) has been estimated stereologically for different species: mouse [262], rat [168, 169], cat, monkey, and human. Measured areal cell densities \( D_A \) are converted into the corresponding volume densities \( D_V \) using the well-known formula [263]:

\[
D_V = \frac{D_A}{t + d}
\]  

(A.1)

where \( d \) denotes the average cell body diameter and \( t \) the section thickness.

Table A.2. Cortical thickness. Mean thickness in \( \mu\text{m} \pm \text{S.D.} \) of individual laminae in occipital (primary visual area), parietal (somatosensory barrelfield area) and frontal (primary motor area) of the mouse and rat adult cortex. N.d.: not determined.

<table>
<thead>
<tr>
<th>Layer</th>
<th>Mouse</th>
<th>Rat</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Frontal</td>
<td>Occipital</td>
</tr>
<tr>
<td>I</td>
<td>126.00 ±24.58</td>
<td>118.33 ±19.60</td>
</tr>
<tr>
<td>II</td>
<td>64.33 ±8.08</td>
<td>54.33 ±14.50</td>
</tr>
<tr>
<td>III</td>
<td>207.50 ±10.61</td>
<td>192.00 ±1.00</td>
</tr>
<tr>
<td>IV</td>
<td>133.67 ±58.05</td>
<td>105.33 ±24.79</td>
</tr>
<tr>
<td>V</td>
<td>321.00 ±58.59</td>
<td>190.00 ±36.29</td>
</tr>
<tr>
<td>VIa</td>
<td>285.00 ±20.22</td>
<td>205.33 ±15.70</td>
</tr>
<tr>
<td>VIb</td>
<td>N.d.</td>
<td>73.99 N.d.</td>
</tr>
</tbody>
</table>

Table A.3. Numerical density of neocortical neurons. Number of 10^3 neurons per \( \text{mm}^3 \pm \text{S.D.} \) in occipital (primary visual area), parietal (somatosensory barrelfield area) and frontal (primary motor area) of the mouse and rat adult cortex. N.d.: not determined.

<table>
<thead>
<tr>
<th>Layer</th>
<th>Mouse</th>
<th>Rat</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Frontal</td>
<td>Occipital</td>
</tr>
<tr>
<td>I</td>
<td>9.64 ±0.53</td>
<td>11.44 ±6.50</td>
</tr>
<tr>
<td>II</td>
<td>131.37 ±20.48</td>
<td>166.99 ±20.17</td>
</tr>
<tr>
<td>III</td>
<td>94.04 ±22.55</td>
<td>133.12 ±15.11</td>
</tr>
<tr>
<td>IV</td>
<td>141.11 ±44.55</td>
<td>210.46 ±54.64</td>
</tr>
<tr>
<td>V</td>
<td>79.43 ±8.51</td>
<td>88.28 ±16.05</td>
</tr>
<tr>
<td>VIa</td>
<td>139.78 ±22.18</td>
<td>143.29 ±31.86</td>
</tr>
<tr>
<td>VIb</td>
<td>78.16 N.d.</td>
<td>75.15 N.d.</td>
</tr>
</tbody>
</table>

Table A.4. Number of neurons per cortical column. Number of 10^3 neurons beneath 1 \( \text{mm}^2 \) ± S.D. of pial surface in occipital (primary visual area), parietal (somatosensory barrelfield area) and frontal (primary motor area) of the mouse and rat adult cortex. N.d.: not determined.

<table>
<thead>
<tr>
<th>Layer</th>
<th>Mouse</th>
<th>Rat</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Frontal</td>
<td>Occipital</td>
</tr>
<tr>
<td>I</td>
<td>1.22 ±0.07</td>
<td>1.35 ±0.77</td>
</tr>
<tr>
<td>II</td>
<td>8.45 ±1.32</td>
<td>9.07 ±1.10</td>
</tr>
<tr>
<td>III</td>
<td>19.51 ±4.68</td>
<td>25.56 ±2.90</td>
</tr>
<tr>
<td>IV</td>
<td>18.86 ±5.95</td>
<td>22.17 ±5.76</td>
</tr>
<tr>
<td>V</td>
<td>25.50 ±2.73</td>
<td>16.77 ±3.05</td>
</tr>
<tr>
<td>VIa</td>
<td>29.78 ±6.32</td>
<td>23.87 ±6.54</td>
</tr>
<tr>
<td>VIb</td>
<td>10.06 N.d.</td>
<td>5.56 N.d.</td>
</tr>
</tbody>
</table>
Bibliography


