amU
Aix Marseille Université

NNT : 2024AIXM0001

THÈSE DE DOCTORAT

Soutenue à AMU — Aix-Marseille Université
le 7 mai 2025 par

Nadine BEN BOINA

From logical modelling to transcriptomic data analysis

A systems biology perspective

Discipline
Mathématiques

École doctorale
ED 184 Mathématiques et Informatique

Laboratoire/Partenaires de recherche
Marseille Medical Genetics (MMG)
Institut de Mathématiques de Marseille (I2M)
The Turing Centre for Living Systems (Centuri)
Institut Marseille Maladies Rares (MarMaRa)

Composition du jury

Laurence CALZONE Ingénieure de recherche, Institut Curie	Rapportrice
Madalena CHAVES Directrice de recherche, INRIA	Rapportrice
Paul RUET Chargé de recherche, CNRS	Examinateur
Elias VENTRE Chargé de recherche, INRIA	Examinateur
Denis THIEFFRY Professeur, IBENS	Examinateur
Jacques VAN HELDEN Professeur, AMU	Président du jury
Élisabeth REMY Directrice de recherche, CNRS	Directrice de thèse
Anaïs BAUDOT Directrice de recherche, CNRS	Co-directrice de thèse

INSTITUT DE MATHÉMATIQUES DE MARSEILLE
Marseille Medical Genetics
CENTURI TURING CENTRE FOR LIVING SYSTEMS

Affidavit

I, undersigned, Nadine Ben Boina, hereby declare that the work presented in this manuscript is my own work, carried out under the scientific supervision of Élisabeth Remy and Anaïs Baudot, in accordance with the principles of honesty, integrity and responsibility inherent to the research mission. The research work and the writing of this manuscript have been carried out in compliance with both the french national charter for Research Integrity and AMU charter on the fight against plagiarism.

This work has not been submitted previously either in this country or in another country in the same or in a similar version to any other examination body.

Marseille, 04/03/2025



This work is licensed under Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International Public License

List of Publications and Conference Participation

List of publications produced as part of the PhD project :

1. Nadine Ben Boina, Brigitte Mossé, Anaïs Baudot, Elisabeth Remy, Refining Boolean models with the partial most permissive scheme, Bioinformatics, Volume 41, Issue 4, April 2025, btaf123, https ://doi.org/10.1093/bioinformatics/btaf123

Participation in conferences and summer schools during the PhD period :

1. Journées Ouvertes en Biologie, Informatique et Mathématiques (JOBIM), Rennes university, 5th-8th July 2022 (Poster)
2. CNRS Thematic School "Modélisation Formelle de Réseaux de Régulation Biologique", Porquerolles, 4th-9th June 2023 (Poster)
3. Centuri Scientific Meeting, 16th-20th October 2023 (Oral presentation)
4. Annual GT Bioss Meeting, Campus Pierre-et-Marie-Curie (Jussieu), Paris, 27th-28th May 2024 (Oral presentation)

Résumé

Les systèmes biologiques sont par nature complexes. Ils sont composés de nombreux éléments dont les interactions donnent lieu à des propriétés émergentes. L'étude des systèmes biologiques repose souvent sur des approches qui prennent en compte l'ensemble du système et les interactions entre ses différents éléments. Notamment, la modélisation mathématique constitue un outil puissant pour simuler et analyser le comportement des systèmes biologiques. Parmi les modèles mathématiques existants, les modèles logiques (booléens et multivalués) se distinguent par leur faible dépendance aux données quantitatives. Ces modèles mettent l'accent sur les relations causales qualitatives entre les composants.

Ma thèse porte sur l'étude des systèmes biologiques à l'aide de modèles logiques. Après une introduction à la biologie des systèmes et à la modélisation logique, je présente ma contribution principale : une méthode pour affiner les modèles booléens, appelée « Multivalued Refinements of Boolean Models » (MRBM). Les modèles booléens utilisent des variables binaires pour représenter l'activité des composants biologiques. Cette simplification peut aboutir à une représentation incomplète de la dynamique d'un système. Les modèles multivalués peuvent remédier à cette limitation en permettant aux composants de prendre plus de deux niveaux d'activité. Néanmoins, déterminer quels composants multivaluer reste un défi. MRBM propose une solution en exploitant une nouvelle méthode de mise à jour de la dynamique, appelée « most permissive partielle ». Cette mise à jour est basée sur le schéma classique « most permissive ».

Cette thèse inclut aussi une étude collaborative sur des données de séquençage d'ARN issues de souris atteintes du syndrome de Hutchinson-Gilford Progeria. Cette étude illustre une approche différente de la modélisation logique pour l'étude des systèmes biologiques.

Mots clés : Biologie des systèmes, Modélisation mathématique, Formalisme logique, Raffinement multivalué, Séquençage d'ARN.

Abstract

Biological systems are inherently complex. They are composed of numerous elements whose interactions give rise to emergent properties. The study of these systems often relies on holistic approaches that account for both the entire system and the interactions among its various components. In particular, mathematical modelling is a powerful tool for simulating and analysing the behaviour of biological systems. Among the existing mathematical models, logical models (both Boolean and multivalued) distinguish themselves by their minimal reliance on quantitative data.

My thesis focuses on studying biological systems using logical models. After introducing systems biology and logical modelling, I present my main contribution: a method for refining Boolean models, termed "Multivalued Refinements of Boolean Models" (MRBM). Boolean models use binary variables to represent the activity of biological components, but this simplification can lead to an incomplete representation of a system's dynamics. Multivalued models can overcome this limitation by allowing components to adopt more than two activity levels. However, determining which components should be modelled as multivalued remains a challenge. The MRBM method addresses this challenge by leveraging a new updating strategy, called "partial most permissive" updating, which is based on the classical "most permissive" scheme.

This thesis also includes a collaborative study on RNA sequencing data from mice affected by Hutchinson-Gilford Progeria syndrome. This study illustrates a different approach to studying biological systems — other than logical modelling.

Keywords: Systems biology, Mathematical modelling, Logical formalism, Multivalued refinement, RNA sequencing.

Acknowledgements

First, I would like to thank Laurence Calzone and Madalena Chaves for agreeing to be the rapporteurs of my thesis and for taking the time to evaluate my work. I am also grateful to Elias Ventre, Paul Ruet, Denis Thieffry, and Jacques Van Helden for agreeing to be examiners of my thesis.

I'd also like to express my gratitude to my thesis directors, Anaïs Baudot and Élisabeth Rémy, as well as my supervisor, Brigitte Mossé. Thank you for choosing to work with me on this thesis, for your time and guidance, and for sharing your scientific expertise. Your knowledge in bioinformatics, computational biology, discrete mathematics and logical modelling has been invaluable. More than that, your patience and encouragement — especially during moments of self-doubt and when I struggled with writing — helped me see this project through. I truly appreciate your support over these four years. I'm also grateful for the opportunities you provided, from seminars to conferences, which allowed me to connect with the broader biological modelling community.

I would like to thank the members of my thesis committee, whose input has been invaluable. A special thanks to Philippe Roudot for your tips on better organisation and focus. Your book recommendation was appreciated — I bought it, never read it, but still!

I would also like to thank Patrice Roll and his team at MMG for giving me the opportunity to collaborate on the HGPS project. This experience greatly enriched my research and broadened my perspective in the field.

A huge thank you to all the past and present members of MMG and I2M. I'm especially grateful to Saran, José, and Clément for the many discussions that shaped my thinking throughout this journey. We shared an office for a while, until I abandoned you all to join Safae, who turned out to be a great office mate and an even better cook.

Finally, to my friends and family, thank you. Your unwavering support, encouragement, and understanding have meant the world to me. Whether through

conversations, motivation, or simply being there, you've been a constant source of strength. You know who you are, I am not listing names. A special thanks to my sisters for feeding me and doing my dishes — I really needed that!

Contents

Affidavit 2

List of Publications and Conference Participation 3

Résumé 4

Abstract 5

Acknowledgements 6

Contents 8

List of Figures 11

List of Tables 12

1 Introduction: Mathematical modelling in systems biology 13

1.1 Biological systems: definition and properties 15

1.1.1 What is a biological system? 15

1.1.2 Biological systems are complex systems 15

1.1.3 The example of the cell 20

1.2 Systems Biology: a complex-system approach to biology 23

1.2.1 What is systems biology? 23

1.2.2 Top-down systems biology 24

1.2.3 Bottom-up systems biology 25

2 Logical Modelling : Principles and Applications 31

2.1 What is a logical model? 32

2.1.1 The regulatory graph 32

2.1.2 The logical functions 32

2.2 Updating schemes . . . 35
2.2.1 Synchronous updating scheme . . . 36
2.2.2 Asynchronous updating scheme . . . 36
2.2.3 Alternative updating schemes . . . 38
2.2.4 Most permissive updating scheme for Boolean models . . . 39
2.3 Properties and features of the dynamics: what do they all mean? . 42
2.3.1 Reachability . . . 43
2.3.2 Attractor . . . 43
2.3.3 Basin of Attraction . . . 44
2.3.4 Model perturbation . . . 44
2.4 Conclusion . . . 45

3 Paper: Refining Boolean models with the partial most permissive scheme 47

4 Analysis of RNA-sequencing data from vascular smooth muscle cells in wild-type and HGPS model mice 75

4.1 Introduction . . . 76
4.1.1 What is Hutchinson-Gilford Progeria Syndrome? . . . 76
4.1.2 RNA-sequencing . . . 78
4.1.3 Aim of the study . . . 83
4.2 Material and methods . . . 84
4.2.1 Preliminary work . . . 84
4.2.2 Exploratory data analysis . . . 85
4.2.3 Differential expression analysis . . . 85
4.2.4 mRNAs functional analysis . . . 86
4.2.5 miRNA-mRNA Interaction Analysis . . . 87
4.2.6 Multi-omics factor analysis . . . 87
4.3 Results . . . 87
4.3.1 mRNA expression analysis . . . 87
4.3.2 miRNA expression analysis . . . 96
4.3.3 miRNA-mRNA joint analysis . . . 100
4.4 Discussion and conclusion . . . 105

5 Conclusion 108

5.1 MRBM: A novel revision method . . . 108

5.2 RNA sequencing for the study of HGPS 110
5.3 Omics data analysis and logical modelling in systems biology 111

Bibliography **113**

List of Figures

1.1 Hierarchical organisation of biological systems 17
1.2 Schematic representation of the interplay between molecular networks 21
1.3 Bottom-up mathematical modelling 26
1.4 Network representations of biological processes 28

2.1 Toy model 33
2.2 Gene activity as a function of its expression level 34
2.3 State transition graphs of the Toy model 37
2.4 Transitions in a node activity level allowed by the most permissive scheme 40
2.5 Boolean model, asynchronous and most permissive STGs 41

4.1 mRNA data exploration 89
4.2 Hierarchical clustering of the top 2,000 mRNAs with the highest count variability across samples 90
4.3 UpSet plot of the DEmRs identified across all comparisons 92
4.4 Top 10 enriched biological processes and pathways in DEmRs HGPS vs WT. 94
4.8 miRNA data exploration 99
4.9 Hierarchical clustering of the top 200 miRNAs with the highest count variability across samples 100
4.11 miRNA-mRNA interaction network for HGPS vs WT 102
4.12 miRNA-mRNA interaction network for HGPSP8 vs WTP8 103
4.13 MOFA variance decomposition result 104
4.14 Distribution of mRNA and miRNA weights along Factor 1 104
4.15 MOFA and differential expression correlation for mRNAs 105

List of Tables

2.1 **Configurations and associated target levels of the Toy model**. The table lists each possible configuration $x = (x_A, x_B, x_C)$ of the Toy model (Figure 2.1), and the corresponding values of the logical functions $f_A(x)$, $f_B(x)$, and $f_C(x)$. 35

4.1 mRNA samples metadata 88
4.2 Number of differentially expressed mRNAs identified in each comparison 91
4.3 Metadata for miRNA samples used in the analysis, detailing genotype and passage for each dataset. 98
4.4 Number of differentially expressed miRNAs identified in each comparison. 99

1 Introduction: Mathematical modelling in systems biology

Summary

1.1 Biological systems: definition and properties 15
1.1.1 What is a biological system? 15
1.1.2 Biological systems are complex systems 15
1.1.2.1 Diverse and numerous interconnected components . 16
1.1.2.2 Hierarchical organisation 16
1.1.2.3 Complex dynamical behaviours 18
1.1.2.4 Robustness, adaptability, and sensitivity 19
1.1.3 The example of the cell 20
1.1.3.1 Cellular processes drive organism functions 20
1.1.3.2 Cells are complex systems 22
1.2 Systems Biology: a complex-system approach to biology 23
1.2.1 What is systems biology? 23
1.2.2 Top-down systems biology 24
1.2.3 Bottom-up systems biology 25

The goal of biological research is to understand the natural world, including the mechanisms that enable living organisms to function. Living organisms are autonomous, organised entities capable of carrying out metabolic reactions, responding to stimuli, reproducing, growing, and developing (Campbell et al. 2018). A living organism can refer to a single cell, such as a unicellular bacterium, or a more complex multicellular organism, like the human body. Any living organism can be considered as a system: a network of interacting components working together to sustain the system overall function and structure (Bertalanffy 1988). In the context of living organisms, such a system is referred to as a biological system. For instance, the human body is a biological system composed of trillions of interacting

cells, which form tissues, which in turn form organs. These organs work together to maintain the body's functions. Not only can entire organisms be considered biological systems, but so can a group of organs, individual organs, tissues, cells, and even networks of molecules.

To fully understand the behaviour, properties, and functions of any biological system, we must examine both its individual components and their interactions. Indeed, components interactions drive the behaviour of the entire system. For example, to understand how skeletal muscle contracts, we must first examine the behaviour, properties, and functions of its key components: muscle fibres and motor neurons. Equally important is understanding how these components interact. When motor neurons are activated, they transmit electrical signals that trigger the release of calcium ions within muscle fibres. This calcium release facilitates the interaction between the contractile proteins myosin and actin, leading to muscle contraction (Frontera et al. 2015). It is only by studying both the individual components and their interactions that we can fully decipher how a biological system operates.

This integrative perspective (combining the examination of individual components with an analysis of their interactions) is the foundation of systems biology. Systems biology is a field of scientific research dedicated to understanding the behaviour of biological systems by studying how biological components interact with one another. Systems Biology integrates biological observations at the component, interaction, and system levels, and relies on computational tools and mathematical models to elucidate the mechanisms controlling biological systems' functions (Kohl et al. 2010). Mathematical models are central in systems biology. They aim to provide a simplified representation of a biological system (components and their interactions), using sets of mathematical equations or formulas. Models can be used to analyse and explain the mechanisms controlling biological systems' functions, such as cellular processes, and to predict possible outcomes of a system in different conditions (Ingalls 2013).

This first chapter aims to motivate, and explain, the use of mathematical models to study biological systems. It is organised in two sections. The first section defines biological systems and explains why they are considered as complex systems. I provide a detailed example of the cell as a complex system, as cell biology and modelling of cellular processes will be the focus of the rest of the thesis. The second section introduces the field of systems biology. I define and explain how

mathematical models are used, and describe the two main approaches to study biological systems : top-down and bottom-up approaches. I detail how to build a mathematical model of cellular processes, using a bottom-up approach.

1.1 Biological systems: definition and properties

1.1.1 What is a biological system?

A biological system is a set of interacting biological components (e.g., molecules, cells, tissues, or organs) that collectively perform specific functions. A biological system is delimited by a boundary that separates it from its environment (i.e., everything outside the system). Biological components self-organise, meaning they spontaneously organise into systems (Alcocer-Cuarón et al. 2014; "Chapter 1 – What Is Self-Organization?" 2001). As stated previously, biological systems exist at all levels of biological organisation, from molecules to ecosystems. All biological systems perform functions essential to maintaining the system's integrity. These functions vary depending on the system. For example, an organism carries out multiple functions supporting its survival, such as locomotion, respiration, communication, growth, and nutrient absorption. Individual organs have specialised functions: skeletal muscles enable locomotion, the heart pumps blood to circulate oxygen and nutrients, and the brain processes sensory information and cognitive tasks.

To understand how a biological system performs specific functions, it is necessary to study the individual components of the system as well as how they interact with each others. In addition, it is important to take into account the fact that biological systems are complex systems (Mazzocchi 2008).

1.1.2 Biological systems are complex systems

While there isn't a unique definition of complex systems, the consensus is the following. A complex system consists of many interconnected components, whose interactions and mutual influences lead to emergent, system-wide, properties. These emergent properties cannot be easily predicted by studying components in isolation (Estrada 2023). Most biological systems fit this definition.

In biological systems, emerging properties are the attributes that arise from the cooperative interactions among biological components. These properties vary

across systems. For example, emerging properties might include biological functions (e.g., cell division), structural characteristics (e.g., cell shape and size), or phenotypic traits (e.g., eye colour and shape). In addition to emerging properties, complex biological systems exhibit other features, including diverse and numerous interconnected components, hierarchical organisation, complex dynamic behaviours, robustness, adaptability, and sensitivity (Ladyman et al. 2013).

1.1.2.1 Diverse and numerous interconnected components

Biological components are constantly interacting with one another. Interaction refers to any form of physical and functional relationship or influence between two or more components, where the behaviour of one component affects and/or is affected by another. There is a high diversity of interactions at every level of biological organisation. For instance, at the molecular level, proteins interact with DNA by binding to DNA sequences, regulating essential processes such as gene expression. At the cellular level, cells interact with each other; for example, neural cells transmit signals via electrical impulses. At the organism level, humans interact with each other through social interactions. These interactions lead to emergent properties that are not intrinsic to the interacting components. For example, at the molecular level, when a signalling molecule binds to a receptor on the cell membrane, the complex they form cantrigger a cascade of intracellular events, an outcome that neither the signalling molecule nor the receptor can achieve independently (Schlessinger 2000).

In addition to this diversity in biological component interactions, the complexity of a biological system can be described by the number of diverse components. For instance, the human organism contains trillions of cells, organised in 400 distinct cell types (Khan et al. 2024). Each cell types possess unique molecular components, shapes, sizes, and functions. There is a division of labour between cells, that can form specialised tissues and organs. The diversity and high number of cell types contribute to the complexity of the human organism.

1.1.2.2 Hierarchical organisation

Another aspect of biological systems complexity is their hierarchical organisation. They are structured in nested levels, ranging from groups of molecules to entire ecosystems. Each level of organisation (whether molecular, cellular, tissue, organ,

organism, or population) builds upon the previous one. Each level is characterised by its own set of components and interactions (Alcocer-Cuarón et al. 2014). Cellular structures and functions are governed by molecular components. In multicellular organisms, cells organise into increasingly large structures: they interact to form tissues, which in turn form organs. Organs work together within organ systems, each performing specific functions to maintain the overall functionality and structure of an organism.

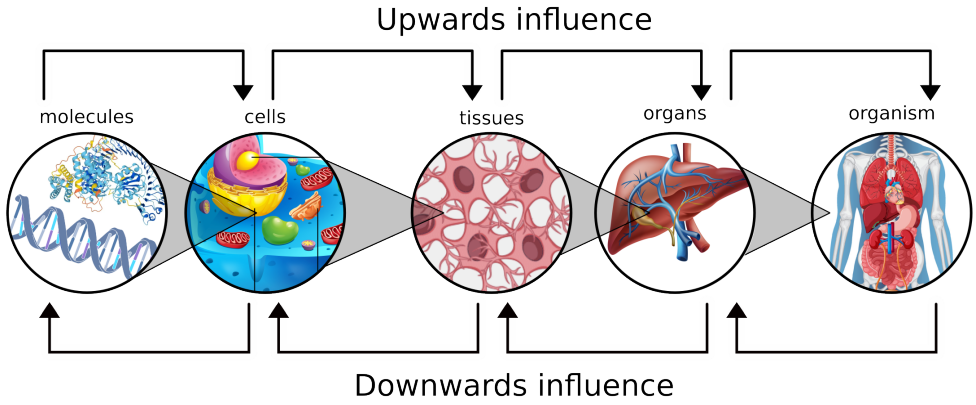


Figure 1.1 – **Hierarchical organisation of biological systems**. The diagram shows that biological systems are built upon several levels of organisation. Each level corresponds to one or several systems that are characterised by their own components and interactions. On top of that, each level of organisation can interact with and influence other levels: from lower to higher levels of organisation (upwards influence), and conversely (downwards influence).

Given this organisation, all biological systems are composed of different levels. Each level encompasses one or several (sub)-systems. The different levels interact with each other (Figure 1.1). Those interactions span a range of spatiotemporal scales, reflecting variations in component sizes and the rate at which biological processes occur (Hunter et al. 2003). Indeed, interactions range from the microscopic molecular level to the macroscopic organism level. Consequently, interactions at the molecular level can trigger changes at the organism level, and vice versa (Noble et al. 2019). For example, in dysferlinopathies — a group of muscular dystrophies caused by mutations in the *DYSF* gene — the absence of the dysferlin protein impairs muscle membrane repair, leading to progressive muscle degeneration at the tissue level. This degeneration affects the muscular system, resulting in severe physical limitations at the organism level. Physical activity can exacerbate the

condition: increased stress on muscles amplifies molecular defects, accelerating disease progression (Anwar et al. 2024).

1.1.2.3 Complex dynamical behaviours

Biological systems are constantly evolving, with the states of their components and their interactions continuously changing. Over time, biological components undergo changes in quantity, structure, and functional status. At the cellular level, for instance, macromolecules such as DNA, RNA, proteins, and lipids are continually being synthesised and degraded. This dynamic nature of biological systems extends to higher levels of organisation. Tissues and organs exhibit dynamic changes in both shape and functional state, and the same holds true at the organism and population levels. The timescales of these changes vary across levels of organisation: molecular and cellular processes occur rapidly, while changes at the tissue or organ level may take days, weeks, or even years (Barbulescu et al. 2015).

An important feature of biological systems' dynamic (and complex systems in general) is their nonlinearity. Non-linearity refers to a characteristic of a system in which the response to a combination of external inputs is not equal to the sum of the responses to each individual input (Janson 2012). A perfect example of nonlinearity in biological systems is the relationship between gene expression (the input) and phenotypic traits (the output). This relationship is often far from straightforward: small changes in gene expression can lead to disproportionately large or unpredictable phenotypic effects, while large changes in gene expression may have little to no visible impact (Félix et al. 2015). Non-linear biological behaviour exist in variety of biological systems, from cellular to ecological (Kapitaniak et al. 2018).

The dynamical behaviour of biological systems is influenced by both internal and external stimuli. Considering an organism, internal stimuli may include fluctuations in nutrient levels or hormonal changes, while external stimuli can involve variations in temperature or exposure to light. Responses to these stimuli are governed by regulatory mechanisms. A regulatory mechanism is any biological process that ensures systems respond appropriately to internal and external changes, maintaining proper function. These mechanisms typically involve feedback loops (Wolkenhauer; Mesarović 2005) and feedforward mechanisms (Konieczny et al. 2023; Rahman et al. 2018). Feedback loops are processes in which the response to a stimulus either inhibits (negative feedback) or amplifies (positive feedback) the original

stimulus. Negative feedback helps maintain balance, such as insulin regulating blood sugar. In contrast, positive feedback can enhance processes, such as in the case of memorisation, where repeated retrieval of information strengthens neural connections, making it easier to recall that information in the future. Feedforward mechanisms, on the other hand, anticipate changes in the environment and initiate responses before the actual changes occur.

Regulatory mechanisms involve all levels of biological organisation, and all have a molecular basis. At the molecular level, regulation occurs through control of gene expression, signalling pathways, and metabolic reactions. These molecular-level regulations ripple through the system, influencing higher organisational levels such as tissues and organs. Conversely, changes at higher levels, such as alterations in tissue function or organ activity, can have impacts at the molecular level.

By integrating nonlinear dynamic behaviour, tightly controlled by regulatory mechanisms, biological systems exhibit complex system properties at every level of organisation, from molecules to entire organisms.

1.1.2.4 Robustness, adaptability, and sensitivity

The dynamical behaviour of biological systems is modulated by internal and external stimuli, with regulatory mechanisms ensuring functionality under varying conditions. These mechanisms contribute to three characteristics of biological systems: robustness, adaptability, and sensitivity (Tzafestas 2018; Whitacre 2010). Robustness refers to a system's ability to maintain function despite fluctuations or environmental perturbations. A common example is maintaining stable body temperature despite external changes. Adaptability, on the other hand, allows systems to adjust their behaviour in response to environmental shifts, such as brain plasticity, where neurons adapt based on learning and memory. While robustness maintains stability, adaptability enables systems to change state or function for better survival in new conditions. Biological redundancy — the presence of duplicate structures or pathways, or components with multiple and similar functions — contributes to these traits. For instance, humans have pairs of organs such as eyes, kidneys, and lungs. Losing one organ does not eliminate its function entirely. At the molecular level, gene redundancy is one of many factors (e.g., non-viable mutations, reduced life expectancy, etc.) that help explain why monogenic diseases are rare. Some genes have overlapping functions, allowing critical processes to continue even if one gene is mutated or lost.

Despite robustness and adaptability, biological systems can also be highly sensitive. Biological sensitivity is the degree to which a system reacts to internal or external conditions. A highly sensitive system may respond significantly to minor changes, while a less sensitive system requires larger inputs for a response. High sensitivity in biological systems plays a crucial role in enabling organisms to detect and respond rapidly to small changes in their internal or external environments. However, high sensitivity can also lead to disproportionate responses, which can sometimes be deleterious, such as when minor mutations cause diseases.

In conclusion, biological systems are inherently complex due to several factors. They are composed of numerous, diverse components that interact to sustain the functions of the system. These components are organised into hierarchical (sub-)systems, ranging from molecular structures to entire ecosystems. Each level of biological organisation exhibits dynamic behaviours that are modulated in response to various stimuli through regulatory mechanisms. Moreover, biological systems demonstrate robustness, adaptability, and sensitivity to stimuli, which collectively enable the maintenance of system integrity amidst fluctuations in internal and external environments. However, biological systems are not infallible; in certain cases, changes can profoundly disrupt their functionality.

Critical questions to address when studying biological processes include: What is the cellular basis of the process? What biological components are involved, and how do they interact? Is the process robust, or is it susceptible to disruption? Due to the inherent complexity of biological systems, particularly in multicellular organisms, it may not always feasible to account for all these elements simultaneously.

While all levels of biological organisation contribute to the functioning of biological systems, biological functions are fundamentally grounded in cellular processes. Cells are the basic units of life, and cellular processes govern organism behaviour, which in turn modulates cellular processes. The following section will provide a more detailed description of the cell as a complex system. The remainder of the chapter will focus on cells and cellular processes.

1.1.3 The example of the cell

1.1.3.1 Cellular processes drive organism functions

Cells are the fundamental building blocks of all living organisms. Cellular processes drive cellular functions such as energy production, cell division, and cell

contraction. Those functions govern not only the survival and reproduction of cells but also coordinate higher-level functions in multicellular organisms, such as organ development, immune responses, or tissue repair (Bich et al. 2019).

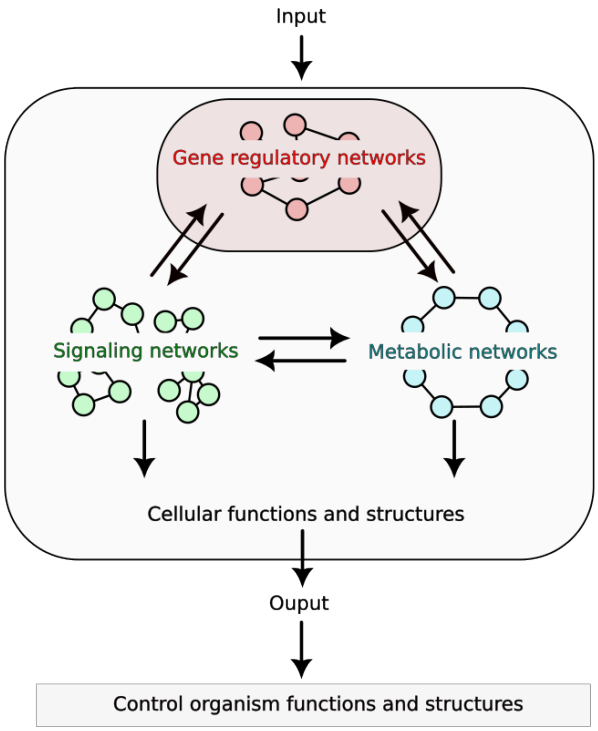


Figure 1.2 – **Schematic representation of the interplay between molecular networks — genes regulatory networks, signalling networks and metabolic networks — involved in cellular and organism functions**. The networks represented are composed of various macromolecules, including gene, RNAs, proteins, and metabolites. External inputs modulate the interactions within and between intracellular networks, consequently influencing cellular functions.

Cellular processes are orchestrated by networks of molecular components such as DNA, RNA, proteins, metabolites, and lipids (Figure 1.2). For instance, signal transduction plays a critical role in cell-to-cell communication and is supported by signalling networks. These networks typically involve protein receptors located on the cell surface, which, when activated, trigger a cascade of biochemical events within the cell. This cascade ultimately initiates a cellular response, leading to changes in gene expression, protein activity, or metabolic processes. In addition to signalling networks, metabolic networks ensure that cells receive the necessary nutrients to function effectively. These networks facilitate the breakdown of sub-

strates to generate energy and synthesise essential biomolecules. Gene regulatory networks are pivotal in controlling gene expression, thereby influencing protein production and cellular behaviour. Together, these interconnected networks enable cells to respond to internal and external stimuli, and support overall cellular and organism functions (Konieczny et al. 2023).

Understanding how cells function provides insights into how organisms develop, maintain homeostasis, and respond to environmental stimuli. Disruptions at the cell level can have far-reaching consequences; a single gene mutation can impair protein function, leading to cellular malfunction and ultimately systemic disease. Diseases such as cancer, where normal regulatory mechanisms governing cell growth and division are impaired, illustrate the direct connection between cellular function and overall organismal health. This highlights the importance of studying biological systems at the cellular level, as cellular dysfunction is often the root cause of larger physiological disorders (Somvanshi et al. 2014).

1.1.3.2 Cells are complex systems

Cells are complex systems that exhibit emergent properties. For example, the formation of cell membrane emerges from the coordinated action of various molecules. In addition, like any complex biological systems, cells also possess all the characteristics described in section 1.1.2 (Wolkenhauer; Muir 2011):

1. Diverse and numerous interconnected components: Cells possess numerous, diverse components interacting together. These include thousands of distinct macromolecules displaying a large diversity of functions: proteins, nucleic acids, lipids, and metabolites.
2. Hierarchical organisation: The hierarchical organisation in cells can be considered in two ways: the flow of genetic information and the structural organisation. The flow of genetic information refers to the process in which genes are transcribed into RNA, which are in turn translated into protein. Additionally, cells have a hierarchical structural organisation. They are composed of specialised organelles, each performing distinct tasks that contribute to overall cell function. For example, the nucleus houses the genetic material, while the mitochondria are responsible for generating energy through cellular respiration.
3. Complex dynamical behaviours: Cellular processes are highly dynamic and

responsive to external changes. Cells continuously adjust their behaviour in response to fluctuations in nutrient availability, change in temperature, exposure to stressors, and signal from surrounding cells. Such external changes can trigger signalling cascades that alter gene expression, protein synthesis, and metabolic activity.

4. Robustness, Adaptability, and Sensitivity: Cells exhibit robustness and adaptability, allowing them to maintain function despite environmental fluctuations. Additionally, their sensitivity to external signals enables quick responses to changes, which is crucial for survival.

The complexity of cellular systems makes it difficult to fully understand their behaviour by studying individual components in isolation. Instead, integrative approaches are needed to analyse how cells function as complex systems. This is where systems biology and mathematical modelling become essential.

1.2 Systems Biology: a complex-system approach to biology

1.2.1 What is systems biology?

Systems biology is the science that studies biological systems as complex systems. It aims to understand biological systems by analysing how their components interact over time, and under different condition. Rather than focusing on the properties and functions of isolated components, it emphasises their interactions across different scales, from molecules to whole organisms (Bhalla et al. 1999; Bruggeman et al. 2007; Kesić 2016). Systems biology is interdisciplinary, it draws on principles and methodologies from biology, mathematics, bioinformatics, and computer science. Biological research relies on wet lab experimentation to recover and interpret biological data. The need of biological research drives the development of technologies and computational tools, that can broaden the scope of biological exploration. For example, progress in genome sequencing (Lander et al. 2001) have facilitated detailed mapping of genetic material, while high-throughput omics technologies (Ideker et al. 2001) (such as transcriptomics and proteomics) have revolutionised the capacity to collect and analyse extensive datasets related to molecular components. The consequent substantial increase in biological data has

driven the creation of robust computational methods and mathematical modelling methods for data analysis and interpretation.

Scientific literature often classifies systems biology approaches in two categories: top-down and bottom-up approaches (Kohl et al. 2010; Tavassoly et al. 2018). This classification provides a framework for understanding the different methodologies in systems biology, but are not rigid separations. In practice, researchers often integrate aspects of both types of approaches, depending on the research questions, the data available, and the experimental possibilities (e.g., material and cost).

1.2.2 Top-down systems biology

Top-down systems biology approaches begin with the collection of system-wide, large-scale experimental datasets, and works towards understanding the patterns and molecular mechanisms hidden in the data. Large-scale datasets are typically omics datasets: high-throughput measurements related to various classes of biomolecules. Genomics profiles of DNA sequences intend to identify genetic mutations, transcriptomics focuses on measuring gene expression with sequencing of RNA transcripts, and proteomics measure protein abundance or protein-protein interactions. By using multiple omics datasets, researchers can obtain detailed omics profiles of a cell.

A classical (and complete) top-down analysis approach can be divided in four main steps (Shahzad et al. 2012):

1-Data acquisition: The approach begins by collecting omics data on the system, sometimes at various time points or under specific experimental conditions (e.g., after a mutation). In Chapter 4, I present RNA-sequencing datasets reflecting RNA and miRNA expression levels.

2-Data analysis: Once data is acquired, statistical tests and bioinformatics methods are used to identify meaningful patterns, correlations, and relationships within the datasets. A common analysis of transcriptomics data is differential gene expression analysis, which identifies genes whose expression levels differ significantly under specific experimental conditions. I conducted such an analysis on the datasets presented in the chapter 4. In addition to differential expression, other statistical techniques are used to interpret omics data. Principal component analysis helps reveal the main sources of variation in the dataset and cluster similar samples

together. Functional enrichment analysis is used to identify biological processes perturbed among differentially expressed genes.

3-Data integration: Different omics datasets can be integrated to obtain a more comprehensive view of the biological system. Data integration consist in combining datasets from genomics, transcriptomics, proteomics, and other omics approaches to reveal correlation or interactions between different molecular layers, such as genes, RNA transcripts, and proteins. The integration process relies on bioinformatics methods such as Multi-Omics Factor Analysis (MOFA) (Argelaguet et al. 2018). MOFA models the data with latent variables that capture shared and omics-specific patterns of variation. It uncovers factors driving biological variability that could be associated for instance with experimental condition or genetic background, and distinguishing between shared variations across different omics data types and those unique to each omics (see Chapter 4). Data integration often leads to the formulation of hypotheses regarding potential regulatory mechanisms, such as signalling pathways or gene regulatory networks.

4-Experimental validation The final step involves conducting wet-lab experiments to test the formulated hypotheses. These experiments help to confirm or refine the predicted correlations and mechanisms identified during the previous steps. For example, gene knockdown experiments may be performed to test the involvement of specific genes in the observed biological process.

Top-down approaches offer the advantage of providing a potentially comprehensive, genome-wide view of a biological system. However, they require large datasets, often comprising thousands of data points across various conditions. Additionally, top-down approaches are able to infer correlations, which represent hypothetical mechanisms that may not be accurate, and thus require further experimental validation.

1.2.3 Bottom-up systems biology

Bottom-up systems biology approaches use detailed knowledge on molecular components and their interactions, and assembles them to understand and model the behaviour of biological systems. Unlike top-down approaches, which begins with system-wide omics data and deduces molecular mechanisms, bottom-up approaches start with the characterisation of individual molecular components and

their dynamical interactions. In bottom-up systems biology approaches, mathematical models are used to describe biological systems by capturing the dynamical interactions between molecular components (Álvarez-Buylla Roces et al. 2018). Mathematical models are formal representations of real-world systems expressed through mathematical expressions (equations, algorithms, or logical operators) (Michelsen 2006; Torres et al. 2015). These models serve as simplified versions of complex biological systems, allowing researchers to simulate, analyse, and predict system behaviour under different conditions.

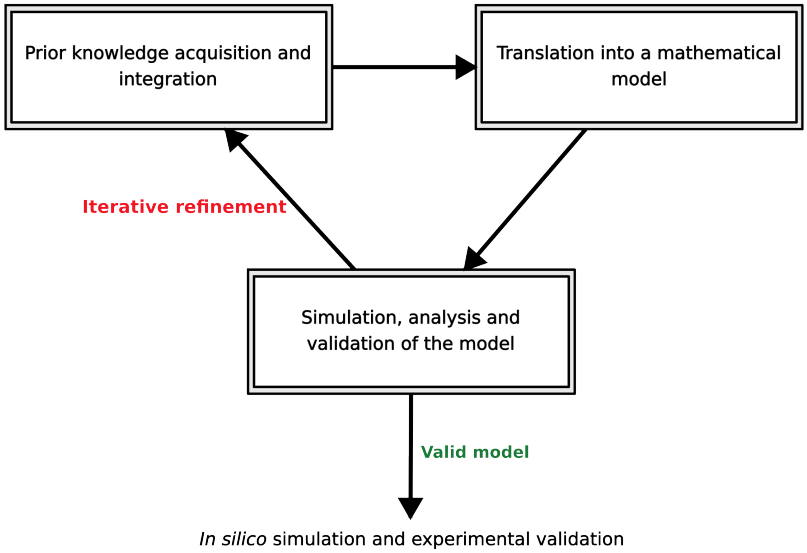


Figure 1.3 – **Bottom-up mathematical modelling**. A flowchart outlining the four steps of the bottom-up systems biology approach, detail in Section 1.2.3.

In practice, a bottom-up approach that incorporates mathematical modelling typically follows these steps (Shahzad et al. 2012)(Figure 1.3): (1) acquiring and integrating prior knowledge of components and their interactions, (2) translation or abstraction of this knowledge into a mathematical model, (3) analysis and validation of the model, and (4) experimental validation of hypothesis obtained from the model.

1-Prior knowledge acquisition and integration: The first step of the bottom-up approach involves gathering existing knowledge about the molecular components and their interactions involved in the biological process of interest. Prior knowledge

typically refers to insights or interpretations derived from the analysis of biological data. These insights or interpretation include details about molecular components, such as their abundance (e.g., protein concentrations or gene expression levels), their function and their regulators. Such knowledge can be sourced from scientific literature, curated biological databases, or collaborations.

Curated biological databases provide organised knowledge on biological components, interactions, and systems (International Society for Biocuration 2018). For instance, Expression Atlas (Papatheodorou et al. 2018) contains gene and protein expression data obtained in different conditions. The databases DisGeNET (Piñero et al. 2017) and OMIM (Hamosh et al. 2002) link genetic mutations with diseases and phenotypes. Pathways can be retrieved in REACTOME (Fabregat et al. 2018) or KEGG (Kanehisa et al. 2000). The SIGNOR database provides detailed information on biological entities and their interaction: causal effect (up/down-regulation), mechanisms (e.g. binding, phosphorylation, transcription, etc.), and cellular location (Perfetto et al. 2016).

Similar knowledge can also be gleaned from scientific literature, which is however not as organised as curated databases, and the process is more time-consuming. Collaborations with experts in the biological system of interest can also provide insights based on experimental findings. By synthesising knowledge from these diverse sources, one can determine several aspects of molecular components such as: their functions, the pathways in which they are involved (i.e., possible interactions and regulation), activity or expression conditions (i.e., regulatory influences), cellular and tissue localisation, and any known mutations associated with diseases.

Integrating diverse knowledge is used to understand molecular interactions and processes relevant to the biological system under investigation. For readability and visualisation, systems knowledge integration is often represented as a network (Le Novère 2015). Depending on the type of knowledge available, different network representations can be employed (see Figure 1.4):

— Interaction networks, which simply depict the interactions between components without detailing their regulatory influence.
— Regulatory networks, or activity flow, where the influence of one entity on another is represented (e.g., gene A activates gene B). These networks are often used to represent gene regulation or protein signalling.
— Process descriptions, which offer a more detailed depiction, showing both variables and the biological reactions that link them.

All of the aforementioned networks have been standardised using the Systems Biology Graphical Notation (SBGN) (Novère et al. 2009), a set of standardised symbols designed to represent the entities within a biochemical network and their interactions.

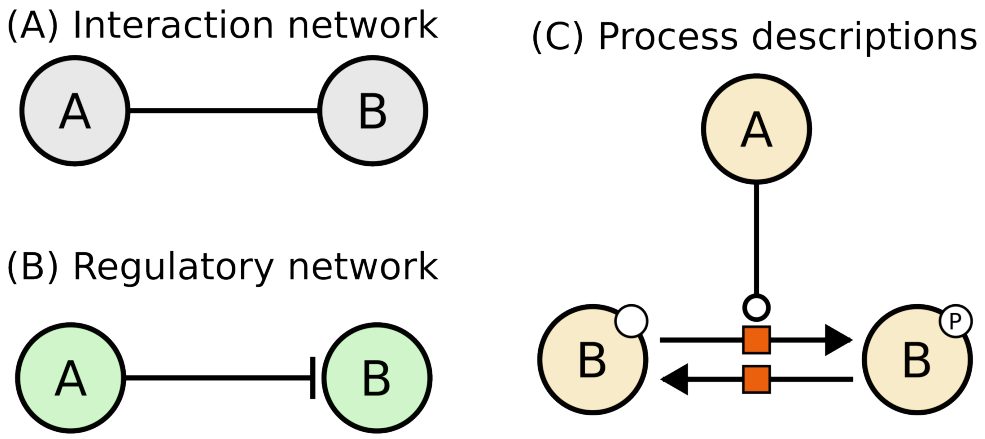


Figure 1.4 – **Network representations of biological processes**. Biological processes can be represented through different types of networks. (A) Interaction networks depict physical or functional interactions between components, as shown by the connection between the protein A and B. (B) Regulatory networks, or activity flow, illustrate influences between components, where protein A inhibits the activity of protein B. (C) Process descriptions capture specific molecular mechanisms, such as protein A triggering the phosphorylation of protein B. The figure is adapted from (Le Novère 2015) .

This initial step provides a comprehensive view of the system, usually represented as a network.

2-Translation into a mathematical model: The next step of a bottom up approach is to translate the network built from prior knowledge integration into a mathematical model. This translation involves representing the components activity and their dynamical interactions into a mathematical framework that can be used to describe, simulate, and predict the system's behaviour. The quantities and activity of a component can be described by quantitative values, such as concentration changes over time, or qualitative values that provide a state interpretation through terms like "ON" or "OFF". In this section, I present

two frequently used mathematical formalisms to model biological systems, one qualitative and one quantitative (Machado et al. 2011).

Quantitative mathematical formalism captures component changes with continuous values. In models of cellular processes, these changes can correspond to protein concentration or mRNA expression levels. Ordinary differential equations (ODEs) are the most prevalent formalism within the quantitative mathematical formalism category (Machado et al. 2011; Torres et al. 2015). ODE-based models assign an equation to each component of the system. Each equation gives the quantity of a component (protein or mRNA concentration), which are modulated by processes such as production and degradation. These processes are represented by rates that depend on other components of the system (Zheng et al. 2010). Using ODEs to model system dynamics requires a high level of detail regarding biological species and how concentrations and rates impact the system. However, the information is not always available and may be challenging to estimate or experimentally determine, especially for large systems.

Qualitative formalism capture the qualitative states and relationships of components of a system. For example, Boolean models represent biological components as being either in an "ON" or "OFF" states. Logical rules are associated with each component, defining how a component is activated (switches from OFF to ON) based on interactions with other components (Karlebach et al. 2008). For instance, a gene is "ON" if its expression level triggers a noticeable change in the system, and "OFF" if it has no effect.

By simplifying the state of a component as being either "ON" or "OFF", it is possible to model biological systems for which the precise mechanism of action are not fully understood. This might be useful when the knowledge available from the literature is not detailed enough for quantitative modelling. Generalisations of Boolean models, known as multivalued models, allow the incorporation of more than two state levels for each component (e.g., "OFF", "LOW-ON", "HIGH-ON"), offering a more detailed representation of biological processes (Didier et al. 2011). Boolean and multivalued network are refereed to as logical models.

Quantitative (e.g., ODE-based) and qualitative (e.g., logical model) mathematical formalisms offer different modelling possibilities for biological systems. Smaller models, well-suited for quantitative approaches like ODE-based models, require extensive data on the concentration, location, regulation, degradation, and other parameters of all components. Conversely, larger systems, with data describing

causal relationships and limited quantitative data available are more suited for qualitative approaches like logical modelling. Ultimately, the data available and the model's purpose dictates the choice of modelling formalism.

3-Analysis and validation of the model: The next step is the simulation and analysis of the dynamics of the model to ultimately validate its accuracy and reliability. Simulations correspond to the model's behaviour over time, providing insights into how the biological system operates under various conditions. The validation process is the comparison of the model's outputs with experimental data and prior knowledge (Torres et al. 2015). This comparison allows researchers to evaluate the model's accuracy in representing the biological system. Based on the results of this validation, the model may require refinement. This could involve adjusting parameters, revising mathematical equations, or even modifying underlying assumptions, such as incorporating additional components into the model.

4-Experimental validation: Similarly to the top-down approach, the final step can involve conducting wet-lab experiments to test the predictions made by the mathematical models. Experimental validation is essential to confirm the insights gained from the previous steps

Bottom-up systems biology approaches emphasise understanding the individual molecular components and their interactions to reconstruct biological systems. By following these steps (prior knowledge acquisition, model construction, mathematical translation, and optionally experimental validation) researchers can create robust models that accurately reflect biological behaviour and contribute to advancements in various scientific fields.

My thesis work focusing on the mathematical modelling of biological systems using the logical formalism. In Chapter 3, I present a novel method for refining logical models to better align with biological observations. Before delving into the core of my work, the next chapter provides a detailed introduction to the logical formalism to model cellular processes.

2 Logical Modelling : Principles and Applications

Summary

2.1 What is a logical model? 32
2.1.1 The regulatory graph 32
2.1.2 The logical functions 32
2.2 Updating schemes 35
2.2.1 Synchronous updating scheme 36
2.2.2 Asynchronous updating scheme 36
2.2.3 Alternative updating schemes 38
2.2.4 Most permissive updating scheme for Boolean models 39
2.3 Properties and features of the dynamics: what do they all mean? . 42
2.3.1 Reachability 43
2.3.2 Attractor 43
2.3.3 Basin of Attraction 44
2.3.4 Model perturbation 44
2.4 Conclusion 45

As already discussed in the previous chapter, biological systems are inherently complex. Their behaviour is driven by the intricate regulatory interactions between various components. Mathematical modelling offers a valuable approach for describing and analysing the dynamics of these systems. The choice of formalism used to model a biological system depends on the available biological data, prior knowledge and the biological question. The logical formalism stands out due to its advantage of requiring minimal quantitative data, relying instead on qualitative causal relationships between components — information that is often readily available in the literature. This second chapter introduces the logical formalism in detail. It is organised into three sections. In the first section, I define the core

elements of logical models. The second section focuses on updating schemes, which govern the simulation of a logical model's dynamics. I define and compare several updating schemes. The third section covers the key properties and features of a logical model's dynamics. The characterisation of these properties and features is essential for the analysis and interpretation of a logical model's dynamics.

2.1 What is a logical model?

We define a logical model by a regulatory graph — depicting the components and their regulations — associated with a set of logical functions describing the dynamics.

2.1.1 The regulatory graph

In the context of biological modelling, a logical model is generally built on the basis of a **regulatory graph**. The regulatory graph represent the interactions between selected components of a biological system that are involved in a biological process of interest (A. Naldi et al. 2009).

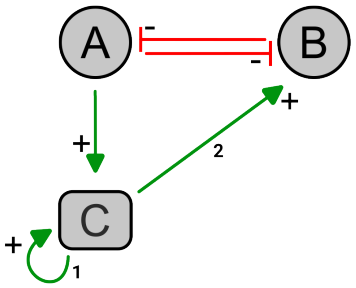
Definition 1 (Regulatory graph) *A regulatory graph $G = (V, E, s)$ is a signed directed graph in which the nodes $V = \{v_1, ..., v_n\}$ represent the biological components, and the edges $E \subseteq V \times V$ the regulations. To each edge is attached a sign $s \in \{+, -\}$ denoting the nature of the regulation: a positive sign ("+") for an activation, a negative sign ("-") for an inhibition.*

Figure 2.1-A depicts the regulatory graph of a Toy model with three nodes (A, B, and C) and five regulations (three activations and two inhibitions): A inhibits B and activates C, B inhibits A, and C activates both itself and B.

2.1.2 The logical functions

To each node $i \in \{1, \ldots, n\}$ of the regulatory graph is associated a discrete variable representing its **activity level**, $x_i \in X_i = \{0, \ldots, m_i\}$, where $m_i \in \mathbb{N}^*$ is the maximal activity level of node i. When $m_i = 1$, the node i is a Boolean node, otherwise it is a multivalued node. When $m_i = 1$ for all the nodes i, the model is a **Boolean model**. Otherwise it is a **multivalued model**.

(A) Regulatory graph



(B) Logical functions

$$f_A(x) = 1 \text{ if } (x_B = 0)$$
$$f_B(x) = 1 \text{ if } (x_A = 0) \wedge (x_C = 2)$$
$$f_C(x) = \begin{cases} 2 \text{ if } (x_A = 1) \wedge (x_C \geq 1) \\ 1 \text{ if } (x_A = 1) \wedge (x_C = 0) \end{cases}$$

Figure 2.1 – **Toy model**. **(A) Regulatory graph**: The logical model contains three nodes — A, B, and C — and five edges. Rectangles represent multivalued nodes, and circles Boolean nodes. Node C is multivalued (3 activity levels), and the nodes A and B are Boolean. Positive edges (in green) correspond to activations and the negative edges (in red) to inhibitions. The outgoing edges associated with the multivalued node C are labelled with the threshold of regulation.**(B) The logical functions**: For each node $i \in \{A, B, C\}$, the logical function f_i gives the condition of activation of the node i depending on the configurations $x = (x_A, x_B, x_C)$.

The state of the model, referred to as the **configuration**, is described by a vector $x = (x_1, \ldots, x_n)$ that provides the activity levels of all nodes. The entire configuration space is given by $X^n = \prod_{i=1}^{n} X_i$.

The use of discrete variables to model the regulation of biological components is directly tied to the non-linear nature of biological systems' dynamics (Section 1.1.2). Indeed, the relationship between the expression level of a gene A (i.e., mRNA concentration) and the activity of its target B (i.e., other gene or protein) can be modelled using a Hill function (Mestl et al. 1995): small increases in the expression level of A produce minimal changes in the activity of its target B, except around a specific level (threshold) where the activity of B changes (increases or decreases) sharply. After the sharp change, the response gradually levels off, reaching a plateau (Figure 2.2).

The Hill function can be approximated by a step function (Figure 2.2) that naturally discretizes the state space : below the threshold, the discrete level of the gene is 0 (gene considered inactive, as it does not impact its target), and above the threshold, the discrete value is 1 (the gene is active as it affects its target). Nodes having k levels of activity allow to distinguish $k - 1$ different regulation thresholds (Thomas 1991). Thus, in Boolean models ($k = 2$ for all nodes), we assume that all

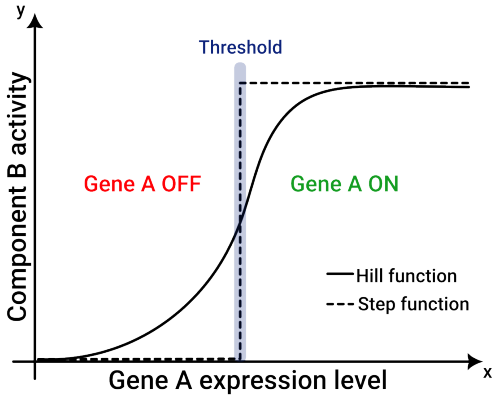


Figure 2.2 – **Gene activity as a function of its expression level**. The graph illustrates the relationship between the expression level of a gene A (X-axis) and the activity of its target B (Y-axis). The gene is considered "OFF" when the expression level is below a certain threshold, and "ON" when it exceeds this threshold.

regulations occur at the same threshold. In multivalued models (some $k > 2$), we can differentiate the thresholds at which a multivalued node regulates its targets.

Each node of the regulatory graph is associated to a **logical function** that determines its target activity level according to the state of the model. Defined using the logical operators "and" ($\wedge$), "or" ($\vee$) and "not" ($\neg$), logical functions specify for each node both which are its regulators, and what activity levels the regulators must have for the node to reach a specific activity level (Kauffman 1993). For example, in the Toy model (Figure 2.1), node B is associated with the logical function $(x_A = 0) \wedge (x_C = 2)$. This function specifies that node B is called to increase its activity level to 1 ($f_B(x) = 1$) when its activator (node C) is present at level 2 and its inhibitor (node A) is absent. Thus, given a configuration $x = (x_1, \ldots, x_n)$, each logical function f_i provides the target activity level (i.e., level to reach) of node i (see Table 2.1 which provide the target levels for the three nodes A, B and C of the Toy model in each of the 12 configurations).

In summary, we define a logical model by a regulatory graph associated to logical functions which determine the dynamics of the system. Note that the regulatory graph can be reconstructed from the logical functions alone, while the reverse is not true (Chaouiya; Remy 2013).

$x = (x_A, x_B, x_C)$	$f_A(x)$	$f_B(x)$	$f_C(x)$
000	1	0	0
010	0	0	0
001	1	0	0
011	0	0	0
002	1	1	1
012	0	1	1
100	1	0	1
110	0	0	1
101	1	0	2
111	0	0	2
102	1	0	2
112	0	0	2

Table 2.1 – **Configurations and associated target levels of the Toy model**. The table lists each possible configuration $x = (x_A, x_B, x_C)$ of the Toy model (Figure 2.1), and the corresponding values of the logical functions $f_A(x)$, $f_B(x)$, and $f_C(x)$.

2.2 Updating schemes

The **dynamics** of a logical model refers to the evolution of the model configurations. This dynamics is discrete, the evolution is represented by a succession of configurations (Saadatpour et al. 2013). The dynamics of a logical model can be described using a **state transition graph** (STG), with nodes representing the model configurations, and directed edges representing transitions between configurations. Thus, the STG describes all possible trajectories, i.e., sequences of consecutive configurations.

Definition 2 (State transition graph) *A STG is a directed graph $S = (X^n, T)$ where X^n is the set of nodes corresponding to all the possible configurations, and $T \subset X^n \times X^n$ is the set of edges representing transitions between configurations.*

To generate the dynamics of a biological system, i.e. describe its trajectories, a logical model is associated with an **updating scheme**. While the logical functions provide the target activity level of each node for a given configuration, the updating scheme dictates how these activity levels are updated to produce the next configuration(s) (Paulevé; Sené 2022). Consequently, the STG of a model is

highly dependent on the updating scheme used to simulate its dynamics (Figure 2.3).

In this section, I first present two classical updating schemes used to model biological systems' dynamics: the **synchronous** and **asynchronous** updating schemes. Then, I provide a brief overview of alternative updating schemes, which allow for the specification of the precise order in which nodes are updated. Finally, I introduce a recently proposed updating scheme, the **most permissive** updating scheme (Paulevé; Kolčák, et al. 2020), which I used in the development of the method presented in Chapter 3.

2.2.1 Synchronous updating scheme

The synchronous updating scheme allows all the nodes of a model to be updated simultaneously at each transition (Kauffman 1993).

Definition 3 (Synchronous updating scheme) *Given a logical model f of dimension n, the synchronous updating scheme allows a transition from a configuration $x \in X^n$ to a configuration $y \in X^n$ if and only if $x \neq y$ and $y = f(x)$.*

The synchronous updating scheme is deterministic, meaning there is at most one successive configuration (the target configuration) for every configuration of a model (Figure 2.3-A).

The number of configurations grows exponentially with the number of nodes in the regulatory graph. Consequently, the dynamics of a model exhibits a combinatorial explosion that hinders its visualisation and analysis. The synchronous scheme is easy to implement and more computationally efficient than other updating schemes. However, in the context of biological regulatory networks, the synchronous assumption is not realistic, as biological processes typically involve regulatory interactions that occur at different speeds (Shamir et al. 2016). Thus, trajectories generated with a synchronous update scheme do not faithfully simulate the behaviour of biological systems.

2.2.2 Asynchronous updating scheme

The asynchronous updating scheme allows only one node of the model to be updated at each transition (Thomas 1991).

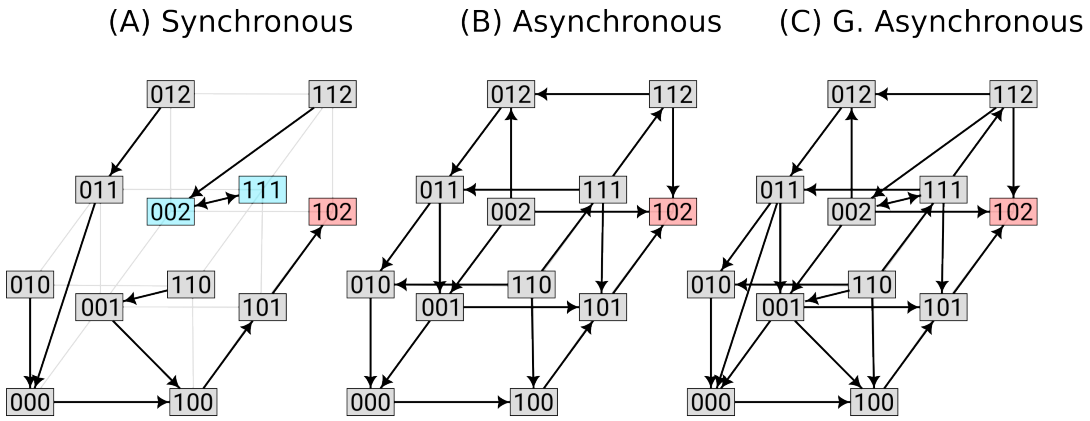


Figure 2.3 – **State transition graphs of the Toy model (Figure 2.1), updated with the synchronous (A), asynchronous (B) or generalised asynchronous (C) updating scheme**. Each node of the STGs represents a configuration ($x = (x_A, x_B, x_C)$) of the Toy model, and arrows correspond to transitions between configurations. The red configurations correspond to fixed points ($f(x) = x$), the blue configurations correspond to cyclic attractor (see Section 2.3.2).

Definition 4 (Asynchronous updating scheme) *Given a logical model f of dimension n, the asynchronous updating scheme allows a transition from a configuration $x \in X^n$ to a configuration $y \in X^n$ if and only if $\exists i_0 \in \{1, \ldots, n\}$ such that $x_{i_0} \neq f_{i_0}(x)$, and*

$$y_{i_0} = x_{i_0} + sign(f_{i_0}(x) - x_{i_0}),$$
$$y_j = x_j, \quad for \ j \neq i_0.$$

The asynchronous scheme is non-deterministic, each configuration can lead to multiple possible successive configurations (as shown in Figure 2.3-B).

This scheme is often considered more realistic than the synchronous scheme for modelling biological systems, as it can distinguish the different speeds of regulatory interactions that occur in biological systems. The asynchronous dynamics of a model includes all possible orders in which only one node is updated at a time.

The **generalised asynchronous scheme** allows any combination of nodes to be updated in a single transition. This scheme captures all possible update orders, including synchronous and asynchronous updates (Figure 2.3-C). The generalised asynchronous scheme can be very computationally demanding due to the large number of transitions it contains.

In practice, the asynchronous scheme is often the chosen updating mode, because of its balance between biological realism and computational efficiency (Paulevé;

Sené 2022).

2.2.3 Alternative updating schemes

The synchronous and asynchronous updating schemes are the classical approaches for simulating the dynamics of logical models, each with its strengths and limitations. As previously discussed, neither scheme explicitly specifies the order of regulatory events. The synchronous scheme forces all events to occur simultaneously, regardless of biological timing, which can lead to oversimplified dynamics. On the other hand, the asynchronous scheme assumes only one event can happen at a given moment, which may under-represent co-occurring processes. There are alternative schemes that allow the incorporation of biological knowledge about the timing of regulatory events, as discussed in (Paulevé; Sené 2022).

For example, **block-sequential updating schemes** partition the nodes of a model into separate blocks. Nodes within the same block are updated simultaneously, while the blocks themselves are updated in a predefined order (Goles et al. 2010). Consider the following Boolean model f with three nodes:

$$f_A(x) = 1 \text{ iff } x_B = 0,$$
$$f_B(x) = 1 \text{ iff } x_A = 0,$$
$$f_C(x) = 1 \text{ iff } x_A = 1.$$

and the following partitioning: block $1 = \{A\}$, block $2 = \{B, C\}$. A transition from one configuration to the next appears as a single step, but it is composed of k intermediate steps (k being the number of blocks) : first the node A is updated, then the nodes B and C are updated simultaneously. Thus, with $x = (1, 1, 1)$ the next configuration is $(0, 1, 0)$.

Another approach is **block-parallel updating schemes**, nodes are also partitioned into blocks. However, in this case, nodes within each block are updated sequentially according to a specified order, while the blocks themselves are updated simultaneously (Demongeot et al. 2022). The size of each block determines the update period: a block of size 2 is updated every two steps, a block of size 3 every three step, and so on. Using the same Boolean model and partitioning, first A and B are updated, then A and C. The updating stops once the update sequence start to repeat. Thus, with $x = (1, 1, 1)$, the next configuration is $(1, 0, 0)$.

General block-parallel updating schemes build upon block-parallel schemes by introducing a delay mechanism. This delay allows some nodes to pause their

updates, enabling synchronization between blocks. For instance, one block might update immediately, while another block waits before updating. Consider the following partitioning with a delay denoted Ω: block 1 = $\{A, \Omega\}$, block 2 = $\{B, C\}$. With the addition of a delay, the node A is updated, then there is a delay, resulting in the following sequence of updates: first, nodes A and B are updated simultaneously, then there is a delay and the node C is updated. From $x = (1, 1, 1)$, the next configuration is $(0, 0, 1)$.

Other updating schemes incorporating information about update timing are discussed in the literature (Paulevé; Sené 2022).

2.2.4 Most permissive updating scheme for Boolean models

More recently, **the most permissive updating scheme** (m.p. scheme) was proposed in (Paulevé; Kolčák, et al. 2020) to update Boolean models. In Boolean models, each node is associated to a binary activity level. The most permissive scheme has the particularity of adding intermediate activity levels, i and d. Hence, each node of a Boolean model updated with the m.p. scheme can take four activity levels: 0, 1, i, and d. The state space of the model is given by $X_{m.p.} = \{0, 1, i, d\}^n$, in which we distinguish Boolean configurations — configurations that only consist of binary values — from non-Boolean configurations.

The activity levels i and d correspond to increasing and decreasing activity level between the binary levels. A node must go through these intermediate levels to reach its target level 0 or 1. For example, a node with an activity level of 0 (respectively, 1) will take on the value i (respectively, d) if it is called to increase (respectively, decrease), as shown in Figure 2.4.

An important aspect of the m.p. scheme is that, when a node is in an intermediate level, other nodes of the model can arbitrarily evaluate it as being either at 0 or 1. Thus, to each state x of $X_{m.p.}$ is associated a set of Boolean configurations given by :

$$\gamma(x) = \{x' \in \{0,1\}; \forall j \in \{1, \ldots, n\}, (x_j = 0 \Rightarrow x'_j = 0) \text{ and } (x_j = 1 \Rightarrow x'_j = 1)\}.$$

For example, a configuration $x = i0$ is associated with the set of Boolean configurations: $\gamma(i0) = \{(10), (00)\}$.

The m.p. scheme is defined as follows:

Definition 5 (Most permissive updating scheme) *Given a Boolean model f*

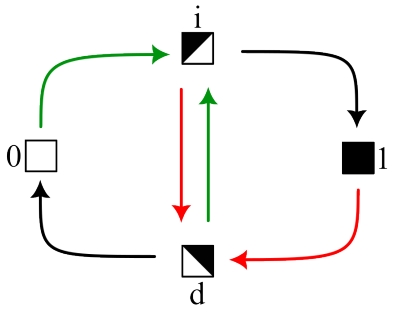


Figure 2.4 – **Change in a node's activity level allowed by the most permissive scheme.** A node updated with the m.p. scheme is associated to four activity levels: 0, 1, i and d. The diagram illustrates the conditions under which the activity level of a node is updated. Green arrows indicate that a node with an activity level of 0 or d is updated to i when it is called to increase. Red arrows indicate that a node with an activity level of 1 or i is updated to d when it is called to decrease. Black arrows show that, from d or i, a node is updated to 0 or 1 respectively, without condition.

of dimension n, the m.p. scheme allows a transition from a configuration $x \in X_{m.p.}$ to a configuration $y \in X_{m.p.}$ if there exists $j_0 \in \{1, ..., n\}$ such that:

$$y_{j_0} = \begin{cases} i, & \text{if } x_{j_0} \in \{0, d\} \text{ and } \exists x' \in \gamma(x) : f_{j_0}(x') = 1, \\ d, & \text{if } x_{j_0} \in \{1, i\} \text{ and } \exists x' \in \gamma(x) : f_{j_0}(x') = 0, \\ 1, & \text{if } x_{j_0} = i, \\ 0, & \text{if } x_{j_0} = d, \end{cases}$$

$$y_j = x_j, \quad \text{for } j \neq j_0.$$

One major advantage of the m.p. scheme is the following: the intermediate levels allow representing the lack of information regarding the thresholds that must be reached for an interaction (or regulation) between two biological components to occur. Consider the Boolean model in Figure 2.5. The node B activates itself and inhibits the node A. Suppose that node B is at an intermediate level i. Both A and B will evaluate B as being at level 0 or level 1. The resulting dynamics will take into account the following scenarios: both A and B evaluate B at 0, both evaluate it at 1, or one node evaluates it at 1 while the other evaluates it at 0. Thus, the resulting m.p. STG encompasses the possibility that B may have different activation thresholds depending on the nodes in the model.

Because of the intermediate levels (i and d), transitions that were not possible for a Boolean model updated using the synchronous or asynchronous schemes are possible with the m.p. scheme. For example, the m.p. STG (Figure 2.5-B/C) of the Boolean

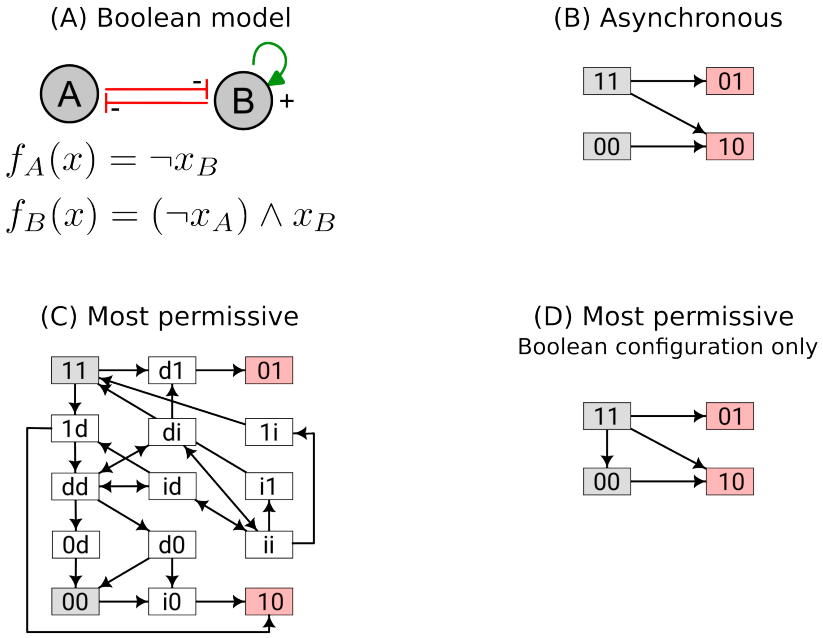


Figure 2.5 – **Boolean model, asynchronous and most permissive STGs. (A) Regulatory graph and associated logical functions**: The model is composed of two nodes; node A is inhibited by the node B, node B activates itself and is inhibited by node A. **(B) Asynchronous STG**: Each node represents a configuration of the model $x = (x_A, x_B) \in \{0,1\}^n$ and edges correspond to transitions between configurations. **(C) Most permissive STG**: Each node represents a configuration $x = (x_A, x_B) \in \{0, i, d, 1\}^n$. **(D) Most permissive STG restricted to Boolean configurations**: Each node corresponds to a Boolean configuration, the edges signal the existence of a trajectory between two configurations. Those trajectories, represented in the panel (C), contain non-Boolean configurations. This visualisation facilitates comparison between STGs. Red nodes indicate fixed points (see Section 2.3.2)

model in Figure 2.5 allows a transition between the configurations 11 and 00, which is not observable in the asynchronous STG (Figure 2.5-D). Note that, to maintain clarity in the representation of the m.p. STG, the representation is restricted to Boolean configurations. The edges between two Boolean configurations in the m.p. STG indicate that there exists a trajectory between them. The corresponding trajectory includes configurations that contain intermediate activity levels (Figure 2.5-B/C).

In conclusion, the m.p. updating scheme provides a richer dynamics compared to classical updating schemes. Indeed, the m.p. STG contains transitions that are observable only within the m.p. dynamics. These additional transitions can capture

relevant biological behaviours of the system being modelled. Additionally, it has been demonstrated in (Paulevé; Kolčák, et al. 2020) that, given a Boolean model, the m.p. scheme is guaranteed to reproduce all dynamical behaviours achievable by any finer modelling framework, such multivalued refinement or ODE-based model. This means that if a trajectory is impossible with the m.p. scheme, there is no more precise qualitative or quantitative refinement of the model that allow this trajectory. However, this increased number of transitions can present challenges for biological interpretation.

Chapter 3 of my thesis introduces a novel method called MRBM (Multivalued Refinements of Boolean Models), which leverages the m.p. scheme. Essentially, given a Boolean model, the MRBM method help to build a multivalued refinement for which the asynchronous dynamics capture trajectories of only observable in the m.p. dynamics of the Boolean model.

2.3 Properties and features of the dynamics: what do they all mean?

First, let's recall that the STG represents the transitions between configurations of a model according to the logical functions and the chosen updating scheme. The analysis of the dynamic of a model is done by analysing the properties of the STG. This analysis is necessary, first, for validating the model by ensuring that its dynamics align with prior biological knowledge; then to use the model for application such as predicting the outcomes of biological processes, the impact of perturbation, and generating new hypotheses.

The analysis of logical model STG is typically focused on two main properties: reachability, which consist on analysis whether a model is able to reach a configuration x from a configuration y, and attractors, which correspond to a model's long-term behaviours.

In this section, I define the properties of the STG and related features (reachability, attractor, and basins of attraction) and explain their significance in the context of biological systems. Additionally, I provide a definition of model perturbations and explain how they impact the aforementioned properties.

2.3.1 Reachability

Given a logical model f and its STG obtained using an updating scheme α, a configuration y is said to be reachable from a configuration x if and only if there exists a trajectory from x to y. The existence of a reachability is denoted by $x \overset{f}{\underset{\alpha}{\rightsquigarrow}} y$.

When analysing model dynamics, modellers often verify reachability properties to assess whether the model aligns with observed biological behaviours. Specifically, they determine whether one configuration can reach another, with both configurations representing distinct states of the biological system that are relevant to the study.

In most cases, establishing general reachability between configurations is sufficient, without needing to specify the exact trajectory. However, certain applications require verifying that the reachability between two configurations satisfies a specific order of events.

2.3.2 Attractor

Attractors are single configurations or sets of configurations that describe the long-term behaviour of the model. Once the the system reaches an attractor, its configuration can no longer get out of the attractor.

Definition 6 (Attractor) *Given a logical model f, and an updating scheme α, the (non-empty) set of configurations $\mathcal{A}$ is an attractor of the model f if:*

— for all $x, y \in \mathcal{A}$, there is a reachability from x to y denoted by $x \overset{f}{\underset{\alpha}{\rightsquigarrow}} y$,

— for all $x \in \mathcal{A}$, and $z \notin \mathcal{A}$, z is not reachable from x, i.e., $x \overset{f}{\underset{\alpha}{\not\rightsquigarrow}} z$.

*Attractors of size one are called **fixed points**. Attractors containing multiple configurations are called **cyclic attractor**. The attractors are the terminal strongly connected nodes of an STG.*

While fixed-point are independent of the updating scheme, cyclic attractors vary depending on the scheme used (Figure 2.3). Additionally, in non-deterministic dynamics, a single configuration can lead to multiple attractors, whereas in deterministic dynamics, a configuration can only reach one attractor.

Attractors are central to the analysis of biological models as they are closely associated with biological phenotypes. In biological systems, attractors represent stable states or recurring patterns of gene or protein expression that the system eventually settles into. These patterns can correspond to specific cellular states,

such as distinct cell types, phases of the cell cycle, or cellular functions (e.g., apoptosis, differentiation, or proliferation).

2.3.3 Basin of Attraction

The basin of attraction of an attractor $\mathcal{A}$, is the set of configurations that can reach the attractor $\mathcal{A}$.

Definition 7 (Basin of attraction) *Given a logical model f, an updating scheme α, an attractor $\mathcal{A}$ and any configuration $a \in \mathcal{A}$, the basin of attraction of $\mathcal{A}$ is defined as:*

$$\mathcal{B}^f_\alpha(\mathcal{A}) = \{x \in X^n \mid \exists x \overset{f}{\underset{\alpha}{\rightsquigarrow}} a\}.$$

In deterministic dynamics (e.g., synchronous dynamics), each configuration always reaches a unique attractor. Thus, each of the attractors has its own basin of attraction. However, in non-deterministic dynamics (e.g., asynchronous and m.p. dynamics), configurations of the model can reach multiple attractors. Consequently, non-deterministic dynamics show two types of basin: strong and weak (Klarner; Streck, et al. 2017). A weak basin includes all configurations that can reach a particular attractor. A strong basin only includes configurations that lead exclusively to a particular attractor, without the possibility of reaching others.

From a biological perspective, weak basins represent a form of commitment to attractors that is less definitive. A configuration that is part of the weak basin of attraction of an attractor $\mathcal{A}$ still has the possibility of reaching other attractors. Transitions from weak to strong basins indicate a commitment to an attractor. This transition can suggest decision processes: the change in activity of a node — or group of nodes — makes the system only able to reach one attractor (Klarner; Heinitz, et al. 2020).

2.3.4 Model perturbation

Model perturbations are modifications to a logical model that simulate alterations in the system being modelled. Such perturbations can modify the dynamics of the model, potentially eliminating existing attractors or creating new ones relative to the unperturbed model. These perturbations are implemented by overriding the logical functions that govern the activity levels of selected nodes (Abou-Jaoudé et al. 2016). For a given node i, the activity is fixed to a value

$p_i \in \{0, \ldots, m_i\}$ for any configuration of the model, so that $f_i(x) = p_i$. Based on the value of p_i, two types of node perturbations can be defined:

— **Knock-down mutation**: Constraining the activity level of a node i to 0 ($p_i = 0$).

— **Ectopic expression**: Constraining the activity level of a node i to its highest value ($p_i = m_i$), or potentially to a range of values greater than zero in the case of a multivalued node ($p_i \in \{1, \ldots, k\}$, where $k \leq m_i$).

More nuanced perturbations can be achieved by allowing modifications to the logical function, such as removing or adding the influence of one node on another, or changing the effect of a specific regulation (H.-C. Trinh et al. 2019). Various perturbations can be defined to reflect experimental observations or to generate predictions regarding the dynamical roles of specific regulatory factors or interactions. For example, model perturbation can represent a gene mutation that leads to loss of expression, or the effects of a drug that inhibits a protein's activity.

2.4 Conclusion

In this chapter, we defined logical models, how to simulate their dynamics, and how to analyse their behaviour. We explored various updating schemes used to model the dynamics of biological systems, highlighting the different ways in which the dynamics evolve based on the updating mechanism employed. We discussed the key properties and features of logical model dynamics, such as reachability, attractors, and basins of attraction. These properties are crucial for validating a model and using it as a predictive tool.

Numerous tools are available for the construction of logical models and the analysis of their dynamics. Notable examples include GINsim (A. Naldi et al. 2009), BoolNet (Müssel et al. 2010), pyBoolNet (Klarner; Streck, et al. 2017), CellCollective (Helikar et al. 2012)], bioLQM (Chaouiya; Bérenguier, et al. 2013), MaBoSS (Stoll et al. 2017), and mpbn (V.-G. Trinh et al. 2024). These tools are also integrated into the CoLoMoTo Interactive Notebook (Aurélien Naldi et al. 2018), which provides a unified environment to edit, execute, share, and reproduce analyses of logical models of biological networks. The notebook facilitates the use of the aforementioned tools in combination. The wide range of available tools allows for the construction of regulatory graphs, defining and editing logical functions, identifying attractors and their reachability, and simulating dynamics

under different updating schemes, including the simulation of model perturbations. Additionally, repositories of logical models ensure their accessibility to the scientific community.

While the analysis of logical model dynamics provides valuable insights into system behaviour, accurately parameterising these models to accurately reflect the dynamics of a system remains a challenge. In the next chapter, I present the novel method I developed to guide the parameterisation of logical models.

3 Paper: Refining Boolean models with the partial most permissive scheme

In this paper, I present a method that aims to help revising Boolean models with multivalued refinements. Boolean models have proven valuable for studying various cellular processes (Chapter 1). However, the modelling process poses significant challenges. First, it requires the acquisition and integration of biological data and knowledge (obtained through wet-lab experimentation, biological databases, and literature) to construct a comprehensive overview of the biological system of interest. This data and knowledge need then to be translated into a mathematical model (using Boolean formalism) with the goal of accurately representing the biological system and its dynamics. Overall, modelling requires iterative revision steps to improve model accuracy, ultimately leading to a validated model that can be used for diverse application (e.g., drug target identification, drug synergy prediction).

The iterative revision of a model typically consists in identifying and addressing regulatory (i.e., interaction between components) and dynamic inconsistencies. These inconsistencies can arise from inaccurate or incomplete biological knowledge or from modelling oversimplifications or mistake (Saadatpour et al. 2013). The revision process can be long and tedious, especially if done manually: the modeller must first identify inconsistencies, then determine which modifications to make. Modification of the model can involve adding or removing components, modifying logical rules, or introducing multivalued variables to model the activity of certain components. Current automated tool for Boolean model revision rely on adjusting logical rules, simplifying the model by removing components, or adding new components (Aleixo et al. 2023; Gouveia et al. 2020).

In the present paper, I propose a novel method, MRBM (Multivalued Refinements of Boolean Models), that guides the revision of Boolean models by identifying components that require multivalued activity levels. MRBM identifies the minimal

sets of components that need to be multivalued to satisfy specific reachability properties within the model's asynchronous dynamics. The method works by comparing the dynamics of the Boolean model when updated with different schemes: the partial most permissive (a new updating scheme presented in the paper), the most permissive, and the asynchronous scheme. Thus, in contrast to current model revision methods, the MRBM method preserves the Boolean model's component and interaction while enriching its dynamics. The MRBM method is supported by rigorous mathematical proofs and has been validated through applications to two published models of stem cell differentiation.

Refining Boolean models with the partial most permissive scheme

Nadine Ben Boina[1,2], Brigitte Mossé[1], Anaïs Baudot[2,3,4], and Elisabeth Remy[1]

[1] *Aix Marseille Univ, CNRS, I2M (UMR 7373), Turing Center for Living systems, Marseille, France*
[2] *Aix Marseille Univ, INSERM, MMG, Marseille, France*
[3] *CNRS, Marseille, France*
[4] *Barcelona Supercomputing Centre, Barcelona, Spain*

Abstract

Motivation In systems biology, modelling strategies aim to decode how molecular components interact to generate dynamical behaviour. Boolean modelling is more and more used, but the description of the dynamics generated by discrete variables with only two values may be too limited to capture certain dynamical properties. Multivalued logical models can overcome this limitation by allowing more than two levels for each component. However, multivaluing a Boolean model is challenging.

Results We present MRBM, a method for efficiently identifying the components of a Boolean model to be multivalued in order to capture specific fixed-point reachabilities in the asynchronous dynamics. To this goal, we defined a new updating scheme locating reachability properties in the most permissive dynamics. MRBM is supported by mathematical demonstrations and illustrated on a toy model and on two models of stem cell differentiation.

Availability and Implementation Data and code available at:
https://github.com/NdnBnBn/MRBM.git.

Contact elisabeth.remy@univ-amu.fr

1

1 Introduction

Systems biology aims to understand how the interactions between biological components such as molecules, cells, tissues, or organs, give rise to specific phenotypes and behaviours (Bruggeman and Westerhoff, 2007). Mathematical models, integrating data from experiments and literature to represent biological systems and facilitate their study, play a pivotal role in systems biology. They have been instrumental in studying a broad spectrum of biological systems, ranging from signal transduction (Heinrich et al., 2002) to complex developmental processes (Tomlin and Axelrod, 2007). Importantly, mathematical models reduce the cost and time associated with *in vivo* or *in vitro* experiments.

Various mathematical frameworks are used to model biological interaction networks (Chen et al., 2010; Ay and Arnosti, 2011; Torres and Santos, 2015). We focus here on logical modeling, a formalism that qualitatively captures the main dynamical properties that explain the overall behaviour of the system, without requiring precise parameters. In this formalism, the activity levels of the components of a biological system are represented as discrete variables. Logical models encompass both Boolean models (BMs) and multivalued models. In BMs, the activity level of the components are represented as binary variables (inactive/active) (Thomas, 1973). However, BMs may not be sufficient to capture the complex dynamics of biological systems. Multivalued models can address this limitation by allowing some components of the model to have more than two levels of activity, while retaining the conceptual simplicity of logical modeling (Thomas, 1978). When a multivalued model is derived from an existing BM, without adding new components nor modifying the effects of the regulations (activation or inhibition), we refer to this multivalued model as a multivalued refinement of the BM (Paulevé et al., 2020).

The dynamics of logical models is driven by logical functions and an updating scheme. The logical functions specify the conditions of activation of each component according to the current activity levels of the components of the system , while the updating scheme dictates how the system is updated. The choice of updating scheme is crucial as it impacts the dynamics of the system. Two updating schemes are commonly used, the synchronous and the asynchronous updating schemes[1] (Garg et al., 2008). Under the synchronous scheme, the activity levels of all the components are updated simultaneously, resulting in a deterministic dynamics. Under the asynchronous scheme, a single component's activity level is updated at a time. The resulting dynamics is not deterministic, as some states may have several successors. The asynchronous scheme is often preferred in modeling biological systems. Indeed, the asynchronous scheme enables the emergence of complex behaviours that are not revealed with the synchronous scheme (Paulevé and Sené, 2022). The most permissive updating scheme (m.p. scheme), recently introduced in (Paulevé et al., 2020), is a non-deterministic scheme that only applies to BMs. The m.p. scheme adds to each component two intermediate activity levels which mark the transition between the two Boolean levels. These intermediate levels represent the fact that the component is in the process of increasing its level from 0 to 1 (level 'i') or decreasing it from 1 to 0 (level 'd'). When a regulator of a component is at an intermediate level, its logical function takes into account both the situation where the regulator is present and the situation where it is absent.

Attractors, which represent the long-term behaviours of a model, are key features of the dynamics.

[1] sometimes denoted fully asynchronous

Once identified, they can be associated with distinct phenotypic readouts (Schwab et al., 2020). The existence of a trajectory from a given state to a specific attractor proves the reachability of the attractor from this state. The number of attractors present in the dynamics, and their reachability, are impacted by the choice of the updating scheme. In particular, the dynamics of a BM under the m.p. scheme captures all the trajectories obtained under the synchronous and asynchronous schemes, as well as all the trajectories obtained from any multivalued refinement of the BM. As a result, more reachability properties are satisfied in the m.p. dynamics than in the other dynamics (see (Paulevé et al., 2020)).

We introduce here MRBM (Multivalued Refinement of Boolean Model), a method which aims at identifying components to be multivalued in a refinement of a BM in order to provide the desired reachabilities within the asynchronous dynamics. MRBM utilizes the partial m.p. scheme, a new updating scheme that we are proposing here (Section 2.4) and that we adapted from the m.p. scheme. In the partial m.p. , only a subset of the model's components is updated using the m.p. scheme, the remaining components being updated with the asynchronous scheme. The resulting dynamics helps to pinpoint the components of the BM that need to be multivalued. In this manuscript, we give a detailed description of the method MRBM. We illustrated its potential with a toy model (Section 3.1). We then use the MRBM method to propose multivalued refinements of two BMs of stem cell differentiation. The first model (Section 3.2) describes the aging effects on hematopoietic stem cell differentiation. The second model (Section 3.3) explores the differentiation of stem cells in the root of *Arabidopsis thaliana*. In both cases, we find a multivalued refinements which display reachabilities that were not present in the BMs.

2 Method

2.1 Boolean Modeling

Set $\mathbb{B} = \{0, 1\}$ and n an integer > 0. A *Boolean Model* (BM) of dimension n is a map $f = (f_1, \ldots, f_n) : \mathbb{B}^n \longrightarrow \mathbb{B}^n$, where each Boolean state $x = (x_1, \ldots, x_n) \in \mathbb{B}^n$ specifies the activity level of n components of the model g_1, ..., g_n. The values of the logical regulatory function $f_i(x) \in \mathbb{B}$ provide the target levels of the i-th component. This model is associated with a *regulatory graph*, i.e. a directed signed graph where the n nodes $g_1, ..., g_n$ of the graph are the components of the model, and the edges between nodes represent the regulations between components. Each edge has a sign, positive for activation, negative for inhibition (Figure 1A). Note that dual edges are allowed. Dual edges represent interactions that can be activations or inhibitions depending on the context.

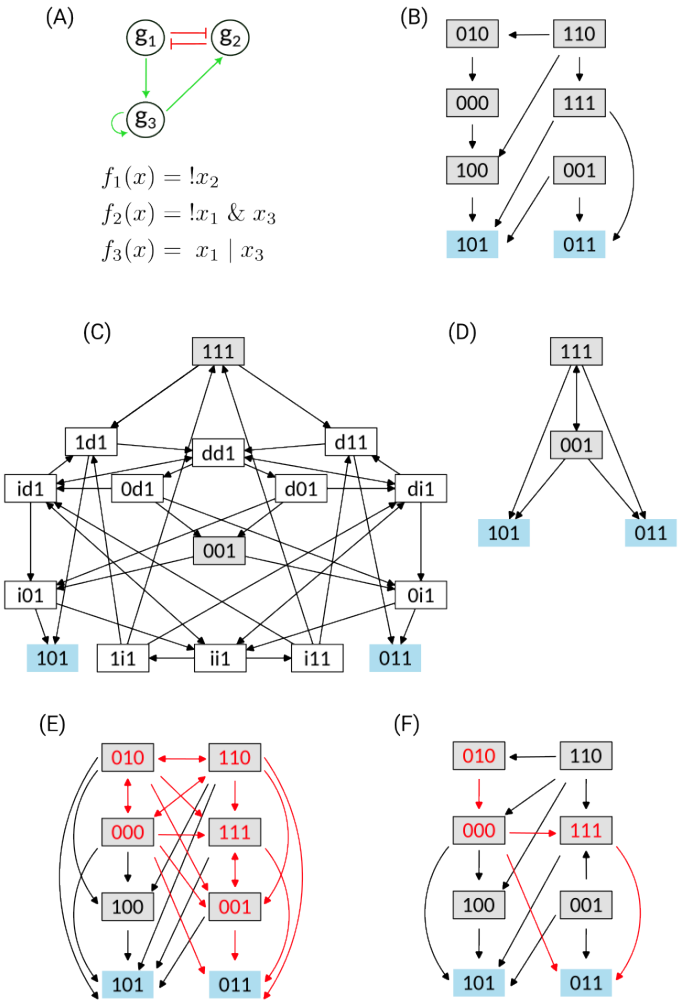


Figure 1: **Toy Boolean Model (BM)**. (**A**) Regulatory graph associated with a BM f composed of three components g_1, g_2, g_3, five interactions (red edges for inhibition, green edges for activation), and logical regulatory functions (operators "&","|","!" stand for AND, OR, NOT respectively). (**B**) The full State Transition Graph (STG) of the toy model in the asynchronous scheme. Each node represents a state $x = (x_1, x_2, x_3)$ of the model and each edge represents an asynchronous transition between two consecutive states. Fixed points are colored in blue. (**C**) STG of the toy model in the m.p. dynamics, representing all the m.p. trajectories from the initial state 111. (**D**) STG in the m.p. dynamics of the toy model, representing all the m.p. trajectories from the initial state 111, restricted to the Boolean states. (**E**) The full STG of the toy model in the m.p. dynamics, displaying only the Boolean states. (**F**) The ${}^{g_1}m.p.$ STG (full STG in the partial m.p. dynamics applied on gene g_1) of the toy model, displaying only the Boolean states. The trajectories from the state 010 to the state 011 are highlighted in red (see Section 3.1.2).

The dynamics of the model is represented by the *State Transition Graph* (STG). The STG is a directed graph $(\mathbb{B}^n, T)$ in which $\mathbb{B}^n$ is the set of nodes that represent the states of the model, and $T \subset \mathbb{B}^n \times \mathbb{B}^n$ is the set of edges that represent the transitions between states. The full STG contains all states of the system. We can also represent a sub-graph of the STG, which contains all the states that can be reached from a given initial state. Given a BM, an updating scheme determinates the stepwise evolution of the system. Within the *asynchronous scheme*, the activity level of only one component can be updated

at a time according to the function f (Figure 1B). More precisely, there is a transition from the state $x = (x_1, \ldots, x_n)$ to the state $y = (y_1, \ldots, y_n)$ if there exists $i_0 \in \{1, \ldots, n\}$ such that $x_{i_0} \neq f_{i_0}(x)$ and

$$y_{i_0} = x_{i_0} + \text{sign}(f_{i_0}(x) - x_{i_0}),$$
$$y_j = x_j, \text{ for } j \neq i_0.$$

Note that the asynchronous scheme generates a non-deterministic dynamics, as each state can have several successors.

A transition from state x to state y using the asynchronous scheme is denoted by $x \xrightarrow[\text{asyn}]{f} y$. A *trajectory* from state x to state y is a sequence of transitions $x \xrightarrow[\text{asyn}]{f} x' \xrightarrow[\text{asyn}]{f} \ldots \xrightarrow[\text{asyn}]{f} y$. A state y is *reachable* from x if there exists a trajectory from x to y. We denote this reachability by $x \overset{f}{\underset{asyn}{\rightsquigarrow}} y$.

The *attractors* of the STG are the terminal strongly connected components (SCCs) of the graph, gathering the asymptotic states of the model. We distinguish two types of attractors, fixed points and cyclic attractors. Fixed points are terminal SCCs of size 1, whereas cyclic attractors are terminal SCCs of size ≥ 2. Given an attractor ω, the basin of attraction $\mathcal{B}^f_{\text{asyn}}(\omega)$ is the set of states from which a state of ω is reachable.

The above concepts on trajectories and attractors are directly adaptable to multivalued models and other updating schemes.

2.2 Multivalued Refinements of Boolean Models

A natural extension of BMs is to have multivalued variables attached to some components, enabling these components to have more than two levels of activity.
Given $m_j \in \mathbb{N}^*$ for $j \in \{1, \ldots, n\}$, with at least one $m_j > 1$, let $X = \prod_{j=1}^n \{0, 1, \ldots, m_j\}$. A *multivalued model* on X is a map $h = (h_1, \ldots, h_n) : X \to X$ satisfying the condition $h_j(x) - x_j \in \{-1, 0, +1\}$, for each $x = (x_1, \ldots, x_n) \in X$ and $j \in \{1, \ldots, n\}$.

Let us consider $f = (f_1, \ldots, f_n)$ a BM of dimension n. Following (Paulevé et al., 2020), and given $X = \prod_{j=1}^n \{0, 1, \ldots, m_j\}$ as above, we introduce for $x \in X$ the set
$\alpha(x) = \{x' \in \mathbb{B}^n \ ; \ \forall j \in \{1, \ldots, n\}, (x_j = 0 \Rightarrow x'_j = 0) \text{ and } (x_j = m_j \Rightarrow x'_j = 1)\}$. For instance, if $X = \{0, 1, 2\}^3$, $\alpha(021) = \{010; 011\}$. A multivalued model h on X is said to be a *refinement of* f if, for any $x = (x_1, \ldots, x_n) \in X$ and $j \in \{1, \ldots, n\}$,

$$h_j(x) < x_j \Longrightarrow \exists x' \in \alpha(x), f_j(x') < x'_j,$$
$$h_j(x) > x_j \Longrightarrow \exists x' \in \alpha(x), f_j(x') > x'_j.$$

This means that, in a multivalued refinement, the level of activity of a component may decrease (resp. increase) at a given state only if its value goes to 0 (resp. 1) under f at some Boolean state of $\alpha(x)$.

Starting from f, we use the following approach to parameterize a multivalued refinement, i.e. to define the function h. For each multivalued component g_j, we consider the logical formulas of its targets

g_k (this means that g_k is regulated by g_j). For each conjunctive clause (i.e. conjunctive logical assertion) of these formulas, we choose a threshold s in $\{0, \ldots, m_j - 1\}$. This threshold specifies that the regulation of g_k occurs when $x_j \in \{s+1, \ldots, m_j\}$. The threshold s is chosen according to the target and to the clause (see details in Supplementary Data Section A.

In the remaining sections of this manuscript, all the refinements we consider are built as described above. In the Results section, we describe these refinements h through the functions $\mathcal{H}_j$ of $x \in X$ with values in $\{-1, +1\}$. These functions $\mathcal{H}_j$ indicate whether gene j is called to decrease or to increase its level x_j under h (when possible, i.e. remaining between 0 and m_j).

Note that in most cases, given a BM f, if there exists an asynchronous path from $x \in \mathbb{B}^n$ to $y \in \mathbb{B}^n$, a corresponding path also exists in the asynchronous dynamics of any refinement of f. Moreover, for any fixed point in the asynchronous dynamics of f, a corresponding fixed point exists in the asynchronous dynamics of any refinement of f belonging to $\prod_{j=1}^{n} \{0, m_j\}$, and conversely. There is one exception. When any multivalued component is self-inhibited, these results are wrong: reachability properties are modified, and even the number and nature of attractors may change (Supplementary Data Section B.

Finally, it should be noted that there is no need to multivalue an output (component without target) in a refinement. Indeed, the multivaluation of an output would have no consequences in terms of reachability properties.

2.3 Most Permissive Scheme

The *most permissive scheme* (m.p. scheme) is a non-deterministic updating scheme for BM proposed by (Paulevé et al., 2020). With this updating scheme, each component can take four activity levels: 0, 1, i, and d. We set $X_{m.p.} = \{0, 1, i, d\}^n$. The increasing and decreasing levels i and d refer to intermediate levels. A component must go through these intermediate levels to reach its target level, either 1 or 0. A component at an intermediate level can be considered by its targets as being at either level 0 or level 1. We associate a set of Boolean states to each state x of $X_{m.p.}$:

$$\gamma(x) = \{x' \in \mathbb{B}^n; \forall j \in \{1, \ldots, n\}, (x_j = 0 \Rightarrow x'_j = 0) \text{ and } (x_j = 1 \Rightarrow x'_j = 1)\}.$$

For instance, $\gamma(di1) = \{001;\ 011;\ 101;\ 111\}$. Then, given a BM f, there is a transition from a state $x \in X_{m.p.}$ to a state $y \in X_{m.p.}$ if there exists $j_0 \in \{1, ..., n\}$ such that one of the following occurs

- $x_{j_0} \in \{0, d\}$ and $\exists x' \in \gamma(x)$ such that $f_{j_0}(x') = 1$ and $y_{j_0} = i$,
- $x_{j_0} \in \{1, i\}$ and $\exists x' \in \gamma(x)$ such that $f_{j_0}(x') = 0$ and $y_{j_0} = d$,
- $x_{j_0} = i$ and $y_{j_0} = 1$,
- $x_{j_0} = d$ and $y_{j_0} = 0$,

and $y_j = x_j$, for $j \neq j_0$.

All states with at least one component activity level equal to i or d will be referred to as *m.p. states*, and other states as *Boolean states*. For the sake of clarity, when representing the m.p. STG we do not

show the m.p. states. For instance, the trajectory : $111 \to 1d1 \to dd1 \to 0d1 \to 001$ (Figure 1C) is represented by the transition $111 \to 001$ (Figure 1D). The full m.p. STG of the toy model is represented (Figure 1E).

An interesting property emerges from the m.p. scheme: given a BM, the m.p. scheme captures all the trajectories present in any of its multivalued refinements. This completeness property of the m.p. is described in (Paulevé et al., 2020). Consequently, if a reachability is not achievable using the m.p. scheme, then no multivalued refinement will present this reachability.

2.4 Partial Most Permissive Scheme

We introduce here the *partial m.p. scheme*, derived from the m.p. scheme. Each of these new updating schemes is associated with the choice of a subset J of $\{1, \ldots, n\}$. We then assign the four possible levels 0, 1, i, and d to components g_j such that $j \in J$. We refer to these chosen components as "m.p. components", and we keep the Boolean levels for all other components. So we consider the set of states $X_{J_{m.p.}} = \prod_{j=1}^{n} \mathcal{A}_j$, where $\mathcal{A}_j = \{0, 1, i, d\}$ if $j \in J$, and $\mathcal{A}_j = \{0, 1\}$ otherwise.

Given a BM f and J as above, we then define the *partial ^{J}m.p. scheme* on $X_{J_{m.p.}}$ in the following way: there is a transition from a state $x \in X_{J_{m.p.}}$ to a state $y \in X_{J_{m.p.}}$ if there exists $j_0 \in \{1, ..., n\}$ such that one of the following occurs

$$
\begin{array}{l}
- \ j_0 \in J \ \text{ and } \ x_{j_0} \in \{0, d\} \ \text{ and } \ \exists x' \in \gamma(x) \text{ such that } f_{j_0}(x') = 1 \quad \text{and } \ y_{j_0} = i, \\
- \ j_0 \in J \ \text{ and } \ x_{j_0} \in \{1, i\} \ \text{ and } \ \exists x' \in \gamma(x) \text{ such that } f_{j_0}(x') = 0 \quad \text{and } \ y_{j_0} = d, \\
- \ j_0 \in J \ \text{ and } \ x_{j_0} = i \quad \text{and } \ y_{j_0} = 1, \\
- \ j_0 \in J \ \text{ and } \ x_{j_0} = d \quad \text{and } \ y_{j_0} = 0, \\
- \ j_0 \notin J \ \text{ and } \ x_{j_0} = 0 \ \text{ and } \ \exists x' \in \gamma(x) \text{ such that } f_{j_0}(x') = 1 \quad \text{and } \ y_{j_0} = 1, \\
- \ j_0 \notin J \ \text{ and } \ x_{j_0} = 1 \ \text{ and } \ \exists x' \in \gamma(x) \text{ such that } f_{j_0}(x') = 0 \quad \text{and } \ y_{j_0} = 0,
\end{array}
$$

and $y_j = x_j$, for $j \neq j_0$.

Note that if the size of the set J is equal to n (the dimension of the model), then the partial ^{J}m.p. scheme is the m.p. scheme. If J is empty, the partial ^{J}m.p. scheme is the asynchronous scheme. The ${}^{g_1}m.p.$ STG of the toy model is represented (Figure 1F).

The completeness property of the m.p. scheme can be generalized to the partial m.p. scheme: given a BM f and a set J, the ^{J}m.p. dynamics encompasses the dynamics of any multivalued refinement of f with $\{g_j \ ; \ j \in J\}$ as the set of multivalued components (see proof in Supplementary Data Section C.

We can emphasize the following properties :

- If J_1 and J_2 are subsets of $\{1, \ldots, n\}$ such that J_1 is contained in J_2, then $X_{J_1 m.p.}$ is contained in $X_{J_2 m.p.}$; moreover, if there exists a trajectory between two states in the ${}^{J_1}m.p.$ dynamics of f, the same holds in the ${}^{J_2}m.p.$ dynamics. In particular, if there exists a trajectory between two Boolean states in the asynchronous dynamics of f, the same property holds in any partial m.p. dynamics.

- The asynchronous dynamics and the partial m.p. dynamics of f share the same fixed points.

For the sake of simplicity, in the Results section, we will identify the set J as the set of m.p. components.

2.5 Assessment of fixed points reachability property

Assessing the reachability of attractors involves determining whether specific states can reach an attractor, or assessing the entirety of states that can reach an attractor (i.e. characterizing the basin of attraction of an attractor). In this section, we will describe the methods we used to identify the existence of reachabilities in the dynamics of a logical model and to compute the basins of attraction.

The MRBM approach is based on comparing the reachability properties of fixed points across the dynamics obtained with different updating schemes for the same BM. The fixed points of a BM do not change with the updating scheme. However, the basins of attraction of these fixed points differ according to the updating scheme, reflecting changes in reachability. So, to identify these changes in reachability, we compare the basins of attraction. The state spaces are not the same depending on the choice of update: the m.p. states are not present in the Boolean asynchronous dynamics. In order to make a relevant comparison between asynchronous, partial m.p. and m.p. dynamics, we have only considered Boolean states.

In the same way, there is a one-to-one correspondence between the fixed points of a BM and those of any of its multivalued refinements, provided there is no self-inhibited multivalued component (see Section 2.2). To compare the basins of attractions in a BM with the basins in a multivalued refinement, we only took into account the states of the subspace $\prod_{j=1}^{n} \{0, m_j\}$. This subspace is in correspondence with the set $\mathbb{B}^n = \prod_{j=1}^{n} \{0, 1\}$.

2.5.1 Reachability Identification

The tools available to identify reachabilities do not support all the possible updating schemes. In this context, for BMs updated with the asynchronous or partial m.p. schemes, we verified reachability with model checking (Baier and Katoen, 2008). To do so, we used the NuSMV model checker software extension (Cimatti et al., 2002) provided in the PyBoolNet Python package (Klarner et al., 2020) and we expressed the attractor reachability property to be verified using computational tree logic (CTL) formulas, as described in (Baier and Katoen, 2008; Klarner et al., 2020). For BMs updated with the m.p. scheme, we used the Python package mpbn (Paulevé et al., 2020). This package is based on Answer-Set Reprogramming (Gebser et al., 2013) and on the solver Clingo (Gebser et al., 2014).

2.5.2 Basins of Attraction Identification

Given f a logical model (Boolean or any multivalued refinement) and its asynchronous dynamics, we use model checking to compute the basin of attraction of a fixed point ω. This is done by identifying, for any state x, whether there exists a path from x to ω.

In the partial m.p. scheme, basins of attraction are computed in the same way. There is however a preliminary step to generate the partial m.p. dynamics of a BM using the bioLQM toolkit (Naldi, 2018).

In the m.p. scheme, the basin of attraction are computed using the Python package mpbn (Paulevé et al., 2020).

We use the size of the basins of attraction of fixed points as a measure of attractor reachability. For a logical model f, we define the size of the basin of attraction of ω, denoted by $\#\mathcal{B}^f(\omega)$, as the number of Boolean states (or equivalent states for a multivalued refinement) present in the basin. We report the sizes of the basins as percentages relative to the sizes of the Boolean states space ($\frac{\#\mathcal{B}^f(\omega)}{2^n} \times 100\%$).

2.6 Availability and Implementation

The MRBM method and the BMs used in this manuscript are available on GitHub at: `https://github.com/NdnBnBn/MRBM` This repository contains the scripts necessary to run the MRBM method within the MRBM directory. Three supplementary directories contain the toy model and the two stem cell differentiation models presented in the Results section. Each of these directories contains the original BM files, the output generated after running MRBM, and the multivalued refinements described in the results section. The method relies on the Colomoto Docker environment (Naldi et al., 2018).

3 Results

3.1 The MRBM Method: defining Multivalued Refinements of Boolean Models

3.1.1 Description of the MRBM method

Let f be a BM such that its asynchronous dynamics displays several fixed points and no cyclic attractor.

Suppose that a reachability property $\mathcal{P}$ is satisfied in the m.p. dynamics of the BM f, but not in the asynchronous dynamics. The MRBM method aims to identify multivalued refinements of f such that the asynchronous dynamics exhibits the reachability property $\mathcal{P}$.

We know that $\mathcal{P}$ is satisfied when the most permissive update is applied to the whole system. The question behind the MRBM method is: can $\mathcal{P}$ be satisfied by applying the most permissive update to only a subset of the components of the system (the other components being asynchronously updated)? If yes, this subset will contain the components that need to be multivalued for the asynchronous dynamics to satisfy $\mathcal{P}$. Thus, the MRBM method is based on the use of the partial ^{J}m.p. dynamics of the BM f (defined in section 2.4), where J is the set of components to be identified.

Since self-inhibited and output components are excluded from the sets J (see Section 2.2), we will only consider the remaining components, called the admissible components. We initiate the method by examining whether there are sets J of admissible components of size 1 that satisfy the reachability property $\mathcal{P}$ in the ^{J}m.p. dynamics of f. If there are no such set J of size 1, we increment the set size by one and reexamine. This iterative process continues until we identify set(s) of size $l \leq n$ that fulfill the reachability property $\mathcal{P}$.

The next step of the MRBM method is to build a multivalued refinement with the components $g_j \in J$

multivalued and the other components remaining Boolean. Towards this goal, the logical functions of each multivalued node and its targets need to be parameterised. We favor refinements where the maximum activity of the multivalued components is as low as possible (principle of parsimony). Hence, we decided to start this parameterisation with 3 levels of activity (i.e. $m_j = 2$ if $g_j \in J$) and to increase the maximal level of activity only if necessary. Finding the multivalued functions of the refinement consists in assigning a threshold (between 1 and m_j) to each of the outgoing regulations of the multivalued components. Defining multivalued logical functions such that the asynchronous dynamics of the model satisfies the reachability property $\mathcal{P}$ is not an easy task and several strategies exist. One strategy, that we call exhaustive strategy, assesses all possible logical functions, i.e. assesses all possible threshold assignations to the clauses of these logical functions (see Supplementary Data Section B). The main issue with this strategy is the combinatorial explosion of possibilities. Another strategy, that we call the *ad hoc* strategy, is to identify the sequence of regulations occurring in a path of interest in the ^{J}m.p. dynamics, and deduce the thresholds of regulation from it. Importantly, in both strategies, the proposed multivalued refinements can be supported or validated by the use of available biological data. This is especially relevant when using exhaustive strategy, as this process can actually generate multiple solutions. These solutions can be filtered using biological data to retain the most biologically relevant refinements. Once the logical functions have been defined, model checking approaches are used to verify the reachability property $\mathcal{P}$ (see Section 2.5).

3.1.2 Illustration of the MRBM method on a toy Boolean Model

To illustrate the MRBM method, we considered the toy BM f of Figure 1A, which describes regulations between three components g_1, g_2, g_3. Both the asynchronous and the m.p. dynamics of the BM contain two fixed points and no cyclic attractor.

First, we considered a specific reachability property $\mathcal{P}_1$, the existence of a trajectory from the state 010 to the fixed point 011. We can observe that there are trajectories from 010 to 011 in the m.p. STG (in red in Figure 1F), but not in the asynchronous STG (Figure 1B). Our aim was to identify a multivalued refinement of the toy BM for which the reachability property $\mathcal{P}_1$ is satisfied in the asynchronous dynamics. To do this, we used the MRBM method. Following the method described in Section 3.1.1, we first searched for sets of admissible components J of size one for which the reachability $010 \overset{f}{\underset{J_{m.p.}}{\rightsquigarrow}} 011$ is satisfied in the partial ^{J}m.p. dynamics of the BM. Among all the sets of size 1 $\{g_1\}$, $\{g_2\}$ and $\{g_3\}$, only the set $J = \{g_1\}$ fulfils the required property (Figure 1G).

Therefore, component g_1 has to be multivalued. To define the multivalued functions of g_1 and its targets, we applied the *ad-hoc* strategy. We investigated the shortest trajectory from state 010 to the fixed point 011 in the ^{g_1}m.p. dynamics of the toy model : $010 \to 0\mathit{0}0 \to \mathit{i}00 \to \mathit{i}0\mathit{1} \to \mathit{i}1\mathit{1} \to \mathit{d}11 \to 0\mathit{1}1$ (in each state, the number in red indicates the component that changed during the state transition). In this trajectory, we observed an increase in the activity level of g_3 when g_1 is at an intermediate level ($\mathit{i}00 \to \mathit{i}0\mathit{1}$). Hence, in a multivalued refinement with $m_1 = 2$, we proposed a threshold 1 for the activation from g_1 to g_3. Additionally, the level of activity of g_2 increases when its activator g_3 is present and its inhibitor g_1 is at an intermediate level ($\mathit{i}0\mathit{1} \to \mathit{i}1\mathit{1}$). We deduced that g_1 needs to be fully active in order

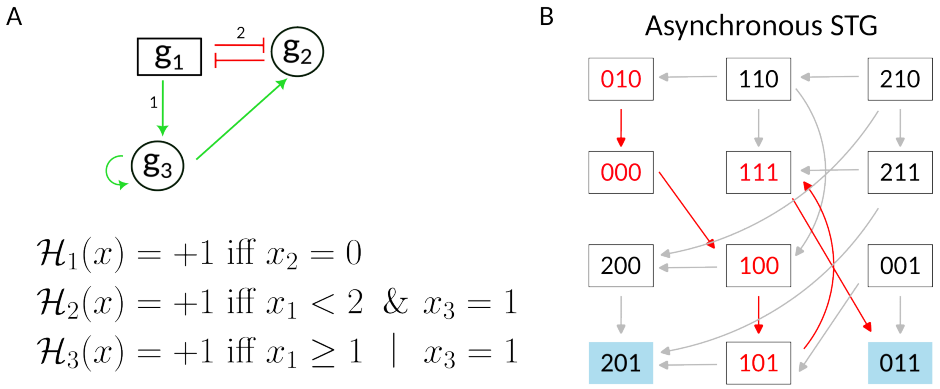


Figure 2: **Multivalued Refinement of the toy model**. (**A**) Regulatory graph of the multivalued refinement of the BM of Figure 1A, and associated logical functions. The rectangle represents the multivalued component g_1 ($m_1 = 2$); outgoing edges from g_1 have labels that specify the threshold levels from which the regulations occur. (**B**) Asynchronous STG of the multivalued refinement h. Each node represents a state (x_1, x_2, x_3). Trajectories representing the reachability 010 $\overset{h}{\underset{asyn.}{\rightsquigarrow}}$ 011 are highlighted in red.

to inhibit g_2; so we set a threshold at level 2 for the inhibition from g_1 to g_2. The resulting proposed multivalued refinement is described in Figure 2A. We verified that its asynchronous dynamics satisfies the reachability property $\mathcal{P}_1$, that is the reachability from the state 010 to the fixed point 011 (see Figure 2B).

	Asyn	m.p	^{g1}m.p.	^{g2}m.p.	^{g3}m.p.
$\#\mathcal{B}^f(101)$	87.5	87.5	87.5	87.5	87.5
$\#\mathcal{B}^f(011)$	50	75	75	50	50

Table 1: **Sizes of the Basins of Attraction of the toy Model Under Different Updating Schemes**. Sizes of the basins of attraction of the fixed points 101 and 011 (as a percentage of the size of the Boolean state space) obtained when using either the asynchronous, m.p. , or partial ^{J}m.p. schemes ($J = \{g_1\}$, $\{g_2\}$, or $\{g_3\}$). Red values signal any increase in the size of a basin of attraction when compared to the size of the basin of attraction of the asynchronous dynamics.

Secondly, we considered a a set of reachability properties $\mathcal{P}_2$. We observed that the size of the basin of attraction of the fixed point 101 does not change with the updating scheme, whereas the size of the basin of attraction of the fixed point 011 changes ($\#\mathcal{B}^f_{m.p.}(011) > \#\mathcal{B}^f_{asyn}(011)$; see Table 1). As all the asynchronous trajectories are captured in the m.p. STG (Section 2.3), this change in size can be interpreted as a loss of reachability in the asynchronous dynamics. Our aim was to build a refinement restoring all these lost reachabilities (property $\mathcal{P}_2$).

Using MRBM, we found that we get $\#\mathcal{B}^f_{J_{m.p.}}(011) = \#\mathcal{B}^f_{m.p.}(011)$ with the partial m.p. scheme with $J = \{g_1\}$. Hence, the asynchronous dynamics of the multivalued refinement defined above allows us to recover a basin of attraction of the same size as for the BM with the m.p. dynamics.

3.2 Application of MRBM to refine a BM of Early Hematopoietic Stem Cell differentiation

The BM depicted in Figure 3A is a model of Hematopoietic Stem Cell (HSC) differentiation detailed in (Hérault et al., 2023). The model contains 15 components. Its dynamics displays five attractors, all of which are fixed points, each representing a distinct HSC fate (i.e. a distinct final stage of differentiation): lymphoid (pLymph), neutrophils and mastocytes (pNeuMast), erythrocytes (pEr), megakaryocytes (pMk), and an inactive state (named zeros). We excluded the fixed point zeros from our study as it was not discussed in (Hérault et al., 2023). The authors of (Hérault et al., 2023) have defined (from data analysis) five transient key stages in HSC differentiation: initial HSCs (iHSC), self-renewal HSCs (srHSC), quiescent HSCs (qHSC), pre-differentiating HSCs (preDiff), and pME (denoted "Transient states" in Figure 3B). We tested the fixed points reachabilities from these transient states in the m.p. and asynchronous dynamics (Figure 3C). The two dynamics display seven differences in terms of reachability (Figure 3C). With the asynchronous scheme, pLymph is not reachable from any of the transient stages of HSC differentiation, pNeuMast is not reachable from qHSC, and pEr is not reachable from qHSCs, pre-diff, and pME. We aimed to identify a multivalued refinement that would provide all those seven reachability properties ($\mathcal{P}$) under the asynchronous scheme.

Using the MRBM method, we identified three admissible sets J of size one for which the partial m.p. dynamics captures part of the desired reachabilities (Figure 3D):

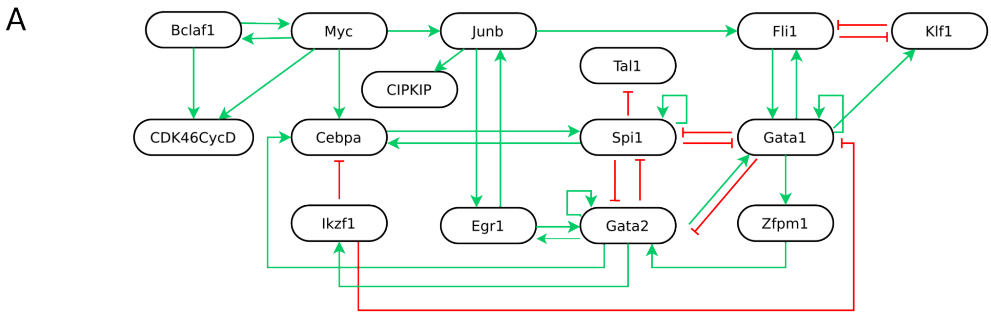
- With $J = \{Fli1\}$, pEr is reachable from pME and preDiff.
- With $J = \{Gata2\}$ or $J = \{Spi1\}$, pLymph is reachable from srHSC.

Increasing the size of the sets J by 1, we identified additional sets capturing additional reachabilities of interest:

- With $J = \{Fli1, Gata1\}$, pNeuMast and pEr are reachable from qHSC.
- With $J = \{Egr1, Gata2\}$, $J = \{Egr1, Spi1\}$, $J = \{Gata1, Gata2\}$, $J = \{Gata2, Junb\}$ or $J = \{Junb, Spi1\}$, pLymph is reachable from the state iHSC.

However, pLymph is still not reachable from qHSC. We found that the set $J = \{Egr1, Fli1, Gata2, Gata1\}$ is required for the property $\mathcal{P}$ to be satisfied in the partial ^{J}m.p. dynamics.

We then built a multivalued refinement of the BM that satisfies properties ($\mathcal{P}$) under the asynchronous scheme. Finding the multivalued functions of the refinement consists in assigning a threshold 1 or 2 to each of the outgoing regulations of the 4 multivalued components. First we set to 2 the maximum activity level of the four components of J ($Egr1$, $Fli1$, $Gata2$, and $Gata1$). By analyzing in the partial ^{J}m.p. dynamics the paths corresponding to the seven reachabilities we want to recover (*ad-hoc* strategy, similarly to the



B

	pLymph	pNeuMast	pMk	pEr	iHSC	srHSC	qHSC	preDiff	pME
	Fixed Points				Transient States				
Egr1	0	0	0	0	0	1	1	0	0
Junb	0	0	0	0	0	0	1	0	1
Bclaf1	0	0	0	0	1	1	0	1	1
Myc	0	0	0	0	0	0	1	1	1
Fli1	0	0	1	0	1	1	1	0	0
Gata2	1	0	0	0	1	0	1	0	0
Spi1	1	1	0	0	0	0	0	1	0
Cebpa	0	1	0	0	0	0	0	0	0
Gata1	0	0	1	1	0	0	0	0	0
Klf1	0	0	0	1	0	0	0	0	0
Tal1	0	0	1	1	1	*	1	*	*
Ikzf1	1	0	0	0	0	*	0	*	*
Zfpm1	0	0	1	1	*	*	*	*	*
CDK46CycD	0	0	0	0	0	1	1	0	0
CIPKIP	0	0	0	0	0	0	1	0	0

C

m.p. / asyn	pLymph	pNeuMast	pMk	pEr
iHSC	✓ / ✗	✓	✓	✓
srHCS	✓ / ✗	✓	✓	✓
qHSC	✓ / ✗	✓ / ✗	✓	✓ / ✗
preDiff	✗	✓	✓	✓ / ✗
pME	✗	✗	✓	✓ / ✗

D

	pLymph	pNeuMast	pMk	pEr
iHSC	J = {Egr1, Gata2} J = {Egr1, Spi1} J = {Gata1, Gata2} J = {Gata2, Junb} J = {Junb, Spi1}	✓	✓	✓
srHCS	J = {Gata2} J = {Spi1}	✓	✓	✓
qHSC	J = {Egr1, Fli1, Gata2, Gata1}	J = {Fli1, Gata1}	✓	J = {Fli1, Gata1}
preDiff	✗	✓	✓	J = {Fli1}
pME	✗	✗	✓	J = {Fli1}

Figure 3: **Boolean Model of Early Hematopoïetic Stem Cell Differentiation**. (**A**) Regulatory graph of the BM (edges in red represent inhibitions and edges in green activations). The logical rules are detailed in (Hérault et al., 2023). (**B**) Configuration (i.e. activity level of each component) of the 5 fixed points and the 5 transient key states (in column). Each cell of the table represents the activity level of a component (0=inactive, in white, 1 = active in dark green, * = 0 or 1, in light green). (**C**) Reachabilities of the fixed points (in column) from transient states (in row) in m.p. and asynchronous ("asyn) dynamics. ✓ stands for "reachable", and ✗ for "not reachable". The cells are divided if there is a difference between the m.p. and the asynchronous dynamics. (**D**) Reachabilities of the fixed points from transient states (in row) in the partial ^{J}m.p. dynamics. The accessibility of transient states at fixed points in the partial dynamics of ^{J}m.p. is satisfied for the J sets specified in the table cell.

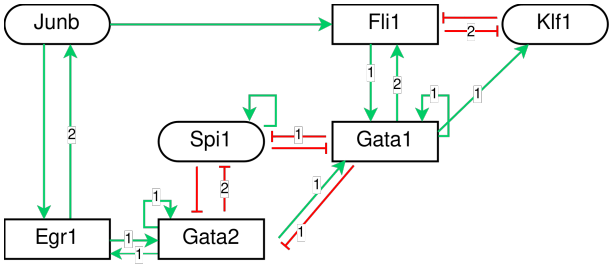


Figure 4: **Subgraph of a Multivalued Refinement of the BM of Early HSC Aging**. Multivalued components are represented as rectangles, while other components are depicted as ovals. The label on arrows indicates the threshold level from which the regulation by a multivalued component occurs.

path analysis in the case of the toy model), we decided to set the threshold of 4 regulations at level 2: the inhibition of $Klf1$ by $Fli1$, the inhibition of $Spi1$ by $Gata2$, the activation of $Fli1$ by $Gata1$, and the activation of $Junb$ by $Egr1$ (Figure 4). The logical functions of the multivalued refinement are available in Supplementary Data Section D.

As described in (Hérault et al., 2023), available biological data support part of this parameterisation. Indeed, starting from a pre-differentiated stage (preDiff) cells can differentiate into either neutrophils/mastocytes (pNeuMast), erythrocytes (pEr) or megakaryocytes (pMk). The circuit composed of $Gata1$, $Fli1$, and $Klf1$ controls the cell fate decision between pEr and pMk. Moreover, an analysis of the dynamics of the BM shows that the switch between the pEr fixed point and pMk fixed point from the preDiff state depends on the existence of two distinct thresholds for $Fli1$'s influence on its targets $Klf1$ and $Gata1$ (see (Hérault et al., 2023) and references therein).

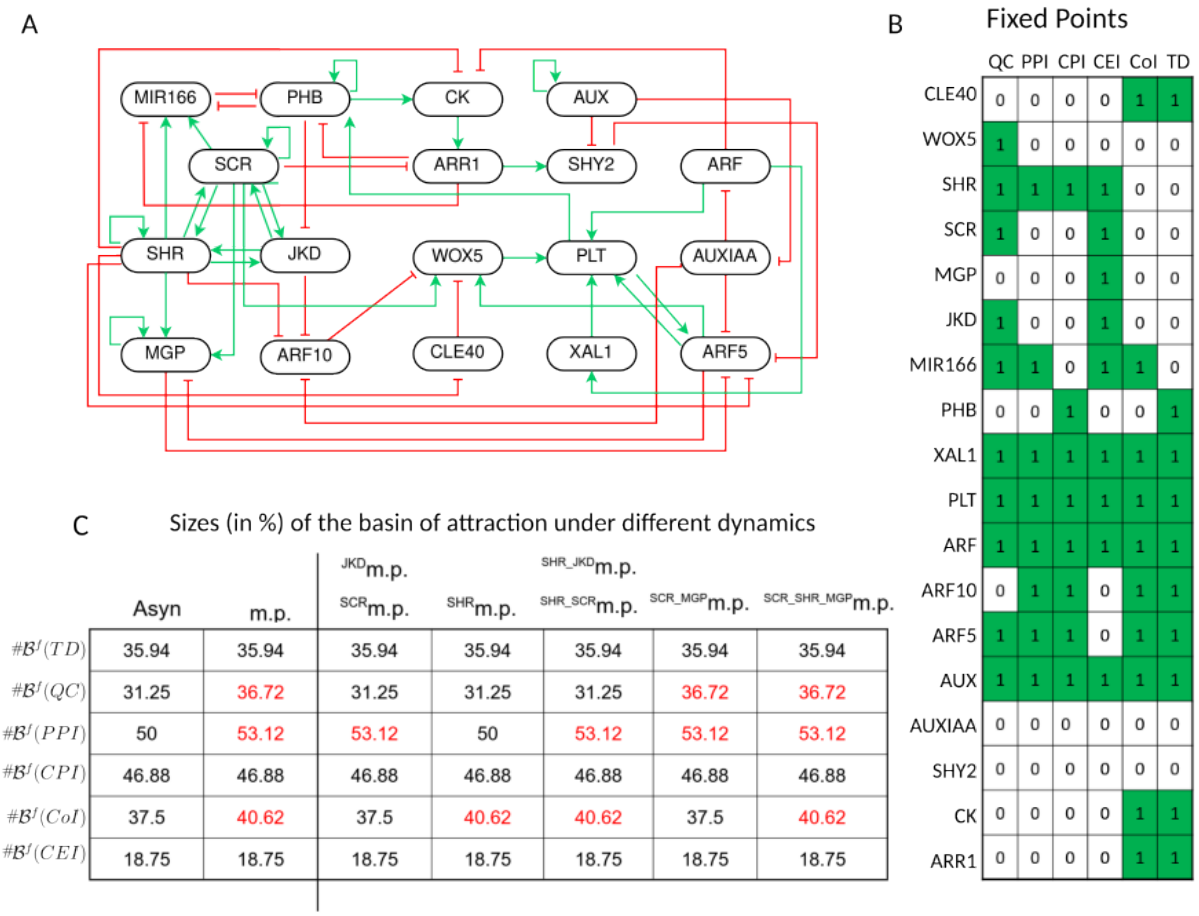
3.3 Application of MRBM to refine a Boolean Model of Asymmetric Stem Cell Division in *Arabidopsis Thaliana* Root

As a second case study, we considered the BM of asymmetric stem cell division in the root of *Arabidopsis thaliana* presented in (García-Gómez et al., 2020) (Figure 5A). The model has six attractors, all of which are fixed points, each of them corresponding to different cell types observed within the root of the plant, including the quiescent center (QC), peripheral and central pro-vascular initials (PPI and CPI), cortex/endodermis initials (CEI), columella initials (CoI), and transition domain (TD) (Figure 5B).

By comparing the asynchronous dynamics to the m.p. dynamics of the model, we noted that the sizes of the basins of attraction of three fixed points (QC, CoI, PPI) are larger within the m.p. dynamics than within the asynchronous dynamics (Figure 5C). As already mentioned in the toy model description (Section 3.1.2), one can interpret this size difference as a loss of reachability in the asynchronous dynamics. To identify a multivalued refinement of the BM whose asynchronous dynamics provides basins of attraction as large as in the m.p. dynamics, we applied the MRBM method. Considering sets J of admissible m.p. components of size one, we obtained:

- if $J = \{JKD\}$ or $J = \{SCR\}$, then $\#\mathcal{B}^f_{J_{m.p.}}(PPI) = \#\mathcal{B}^f_{m.p.}(PPI)$,
- if $J = \{SHR\}$, then $\#\mathcal{B}^f_{J_{m.p.}}(CoI) = \#\mathcal{B}^f_{m.p.}(CoI)$.

14



	Asyn	m.p.	^{JKD}m.p. ^{SCR}m.p.	^{SHR}m.p.	^{SHR_JKD}m.p. ^{SHR_SCR}m.p.	^{SCR_MGP}m.p.	^{SCR_SHR_MGP}m.p.
$\#\mathcal{B}^f(TD)$	35.94	35.94	35.94	35.94	35.94	35.94	35.94
$\#\mathcal{B}^f(QC)$	31.25	36.72	31.25	31.25	31.25	36.72	36.72
$\#\mathcal{B}^f(PPI)$	50	53.12	53.12	50	53.12	53.12	53.12
$\#\mathcal{B}^f(CPI)$	46.88	46.88	46.88	46.88	46.88	46.88	46.88
$\#\mathcal{B}^f(CoI)$	37.5	40.62	37.5	40.62	40.62	37.5	40.62
$\#\mathcal{B}^f(CEI)$	18.75	18.75	18.75	18.75	18.75	18.75	18.75

Figure 5: **Boolean model of the Asymmetric Stem Cell Division in *Arabidopsis Thaliana* Root**. (**A**) Regulatory graph of the BM; the logical functions are available in (García-Gómez et al., 2020). (**B**) Summary of the 6 fixed points. Each row represents a component, each column corresponds to a fixed point, and each cells to the activity level (1 for active, in dark green, 0 for inactive, in white). (**C**) Sizes (in percentage) of the basins of attraction of the BM under different updating schemes. The rows represent basins of attraction and the columns correspond to the updating schemes: asynchronous ("Asyn"), m.p. and partial m.p. . For each partial ^{J}m.p. dynamics, the contributing set J is indicated in the column headers. We grouped in the same column the partial ^{J}m.p. dynamics leading to same sizes of basin of attraction. We only report the smallest set(s) of m.p. components resulting in a specific size of basin of attraction. Values in red correspond to increases in the sizes of the basin of attraction as compared to the sizes obtained in the asynchronous dynamics.

A set J of size two is required for the basin of attraction of the fixed point QC: if $J = \{MGP, SCR\}$, then $\#\mathcal{B}^f_{J_{m.p.}}(QC) = \#\mathcal{B}^f_{m.p.}(QC)$. Thus, $J = \{SHR, MGP, SCR\}$ is sufficient to ensure identical basin of attraction sizes between the partial m.p. and the m.p. dynamics (Figure 5C).

Thus, the use of MRBM efficiently restricts the search space for determining a multivalued model. In this example, the use of MRBM enabled us to identify precisely the three components on which it was relevant to search for multi-level parameterisation, without having to test all 816 possibilities (number of 3-element parts of an 18-element set).

The next step is the specification of multivalued refinements. We started looking for multivalued

refinements with the maximum value of each multivalued component (that are the components in the previously identified sets J) fixed at 2. We used the exhaustive strategy, testing systematically all possible multivalued functions, i.e. all possible assignments of the thresholds (1 and 2) to each of the outgoing edges of the multivalued components. We thus provided multivalued refinements which led to expected sizes of asynchronous basin of attraction for all cases except two ($J = \{SCR, MGP\}$ and $J = \{SCR, SHR, MGP\}$). Indeed, for $J = \{SCR, MGP\}$, when the maximum value of the multivalued components is restricted to 2, only one fixed point among QC and PPI has the same size of basins of attraction in both m.p. and asynchronous dynamics (top panel of Figure 6 represents all the possible situations). We thus increased the maximum value of activity of SCR to 3. Doing so, we were able to find a refinement which recovers the same sizes for the two basins of attraction (PPI and QC) in both dynamics (Figure 6, bottom panels). This 3-levels parameterisation of SCR also makes it possible to find a refinement with $J = SCR, SHR, MGP$ with which we found the same size for the three basins of attraction (PPI, QC and CoI) in both dynamics.

The resulting multivalued refinements are available in the Supplementary Data, Section D.

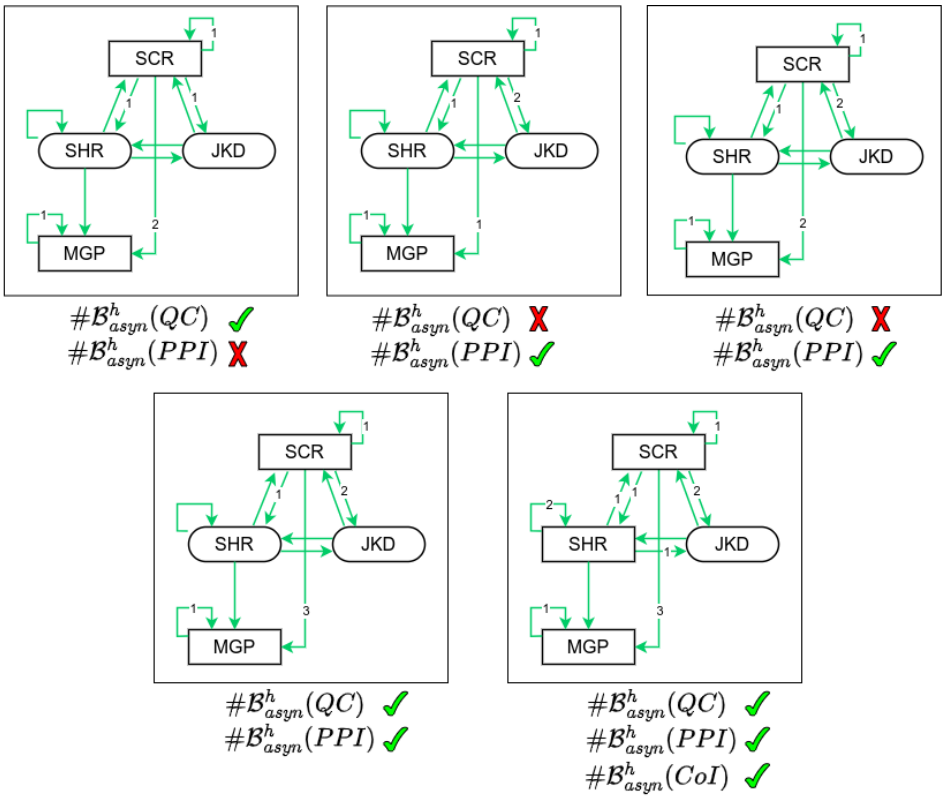


Figure 6: **Subgraph of Multivalued Refinements of the BM of Asymmetric Stem Cell Division in *Arabidopsis thaliana* Root**. Multivalued components are depicted as rectangles, while Boolean components are shown as ovals. The label on arrows indicates the threshold level from which the regulation by a multivalued component occurs. On the top panels $m_{SCR} = 2$, and on the bottom panels $m_{SCR} = 3$. We noted below each panel if we managed to recover the expected size(s) of basin of attraction depending on the set of multivalued components (see Figure 5C). The name of the basin of attraction is followed by a checkmark for a success, or a cross for a failure.

4 Discussion

Although Boolean models have proved to be a valuable and effective tool for modeling biological interaction networks, they have certain limitations due to their high level of abstraction (Wang et al., 2012). In particular, they capture the salient properties of the dynamics but might overlook other finer-grained properties (Chamseddine and Rejniak, 2020). Multivalued logical models can be useful to overcome such limitation, as they enable the consideration of more situations (Flobak et al., 2015; Selvaggio et al., 2021). However, the refinement of a Boolean model into a multivalued model is not straightforward. This involves finding the components that need to be multivalued, as well as parameterising these components and their targets. The possibilities are very numerous and we face a combinatorial explosion. When multivaluing a Boolean model, the aim is to capture the expected properties sparingly, by minimising the number of multivalued components and their maximum level.

In this paper, we present MRBM, an original method to refine Boolean models. The aim of these refinements of Boolean models is to encompass specific dynamical properties $\mathcal{P}$ that appear in the most permissive dynamics but not in the asynchronous Boolean dynamics. We have defined a new updating scheme, the partial most permissive dynamics. This updating scheme applies the most permissive update only to a subset of network components, while asynchronous updating is applied to the other components. This allows to identify the components of the network that control the dynamical properties $\mathcal{P}$, and that would need to be multivalued. Hence, we created a multivalued refinement by multivaluing these components in the Boolean model. Doing so, we induce the dynamical properties $\mathcal{P}$ in the asynchronous dynamics of a multivalued refinement model. Importantly, the method is supported by mathematical arguments and proofs (see Supplementary Data). The effectiveness of MRBM is demonstrated on a toy model and two published Boolean models of cell differentiation, highlighting its ability to handle complex biological systems.

The MRBM method systematically identifies the components to be multivalued but the parameterisation itself remains challenging. We propose two main strategies: 1) An exhaustive strategy, in which all possible parameterisations are tested and then checked to verify if they satisfy the desired property. This strategy can be time consuming, especially as the maximum activity levels allowed for each components increase. 2) An *ad hoc* strategy, which analyses precisely relevant paths within the selected partial dynamics and thereby provides clues for deciphering the regulation thresholds. In both strategies, available a priori biological knowledges are used to guide the definition of the logical functions.

We applied the MRBM method to two published Boolean models of differentiation: a Boolean model of mammalian early hematopoietic stem cell differentiation and a Boolean model of asymmetric stem cell division in *Arabidopsis thaliana*. In both cases, we have identified reachability properties of biological interest that appear in the most permissive dynamics but not in the asynchronous dynamics. The MRBM method allowed us to successfully identify multivalued refinements of the Boolean models exhibiting these reachability properties under the asynchronous updating scheme. In the case of the early hematopoietic stem cell differentiation Boolean model, we proposed a multivalued refinement with four multivalued components (with maximal levels of 2). Existing biological data provide support for part of the parameterisation. For the Boolean model representing asymmetric stem cell division in *Arabidopsis*

thaliana, we had to multivalue three components, one of them with a maximum activity level set at 3, while the others were set at 2. The maximal level of components is a feature that can be used to identify multivalued nodes requiring more detailed regulatory changes to recover the expected sizes of basins of attraction.

There are some limitations for the use of MRBM. The method is restricted to dynamics with only stable states (and no cyclic attractors) and focuses on the reachability properties of these stable states. A generalisation to the reachability properties of cyclic attractors is not straightforward. Indeed, cyclic attractors depend on the updating scheme. Therefore there may not be a one-to-one correspondence between the cyclic attractors of the most permissive dynamics and those of the asynchronous dynamics. This makes it difficult to carry out the global analysis of the size of the basins of attraction, but a study of specific reachabilities between an initial state and a state belonging to a cyclic attractor remains possible. In this work, we have forbidden the multivaluation of a self inhibited component. Indeed, the multivaluation of an auto-inhibited component leads to a change in dynamics. It would be necessary, when studying a particular case, to observe this new dynamics in detail before opting for the multivaluation of such a component.
The method faces computational issues for large models. We used the python package mpbn to check the dynamical properties. The calculation of the size of the basins of attractions was possible for relatively small models, but when studying larger models (of more than 15 nodes), MRBM method was limited to the study of specific reachability properties.

Finally, we can point out an intrinsic combinatorial limitation of the method. The number of components to be multivalued (i.e. the size of the set J) can be as large as n, the number of components of the model. Of course, the larger the set J, the more difficult it is to find a parameterisation. It should be noted that difficulties in finding an appropriate refinement may indicate that the initial Boolean model is incomplete, reflecting gaps in our understanding of the system.

MRBM overall provides a systematic and innovative approach to refine Boolean models with multivaluation. It exploits the richness of most permissive dynamics, extracts its significant properties, and integrates them in the asynchronous Boolean model to refine it. Importantly, the MRBM method does not rely on external data sources, making it useful when experimental data are limited.

Acknowledgement

The project leading to this publication has received funding from France 2030, the French Government program managed by the French National Research Agency (ANR-16-CONV-0001), from Excellence Initiative of Aix-Marseille University - A*MIDEX (Turing Centre for Living Systems). This project was supported also supported by the ANR through the CAMUDI project (ANR-21-CE45-0001).

References

Ay, A. and Arnosti, D. N. (2011). Mathematical modeling of gene expression: a guide for the perplexed biologist. *Critical Reviews in Biochemistry and Molecular Biology*, 46(2):137–151.

Baier, C. and Katoen, J.-P. (2008). *Principles of model checking*. The MIT Press, Cambridge, Mass. OCLC: ocn171152628.

Bruggeman, F. J. and Westerhoff, H. V. (2007). The nature of systems biology. *Trends in Microbiology*, 15(1):45–50.

Chamseddine, I. M. and Rejniak, K. A. (2020). Hybrid modeling frameworks of tumor development and treatment. *Wiley Interdisciplinary Reviews. Systems Biology and Medicine*, 12(1):e1461.

Chen, W. W., Niepel, M., and Sorger, P. K. (2010). Classic and contemporary approaches to modeling biochemical reactions. *Genes & Development*, 24(17):1861–1875.

Cimatti, A., Clarke, E., Giunchiglia, E., Giunchiglia, F., Pistore, M., Roveri, M., Sebastiani, R., and Tacchella, A. (2002). NuSMV 2: An OpenSource Tool for Symbolic Model Checking. In Brinksma, E. and Larsen, K. G., editors, *Computer Aided Verification*, Lecture Notes in Computer Science, pages 359–364, Berlin, Heidelberg. Springer.

Flobak, , Baudot, A., Remy, E., Thommesen, L., Thieffry, D., Kuiper, M., and Lægreid, A. (2015). Discovery of Drug Synergies in Gastric Cancer Cells Predicted by Logical Modeling. *PLoS computational biology*, 11(8):e1004426.

García-Gómez, M. L., Ornelas-Ayala, D., Garay-Arroyo, A., García-Ponce, B., Sánchez, M. d. l. P., and Álvarez Buylla, E. R. (2020). A system-level mechanistic explanation for asymmetric stem cell fates: Arabidopsis thaliana root niche as a study system. *Scientific Reports*, 10(1):3525.

Garg, A., Di Cara, A., Xenarios, I., Mendoza, L., and De Micheli, G. (2008). Synchronous versus asynchronous modeling of gene regulatory networks. *Bioinformatics (Oxford, England)*, 24(17):1917–1925.

Gebser, M., Kaminski, R., Kaufmann, B., and Schaub, T. (2013). *Answer Set Solving in Practice*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Springer International Publishing, Cham.

Gebser, M., Kaminski, R., Kaufmann, B., and Schaub, T. (2014). Clingo = ASP + Control: Preliminary Report.

Heinrich, R., Neel, B. G., and Rapoport, T. A. (2002). Mathematical models of protein kinase signal transduction. *Molecular Cell*, 9(5):957–970.

Hérault, L., Poplineau, M., Duprez, E., and Remy, (2023). A novel Boolean network inference strategy to model early hematopoiesis aging. *Computational and Structural Biotechnology Journal*, 21:21–33.

19

Klarner, H., Heinitz, F., Nee, S., and Siebert, H. (2020). Basins of Attraction, Commitment Sets, and Phenotypes of Boolean Networks. *IEEE/ACM transactions on computational biology and bioinformatics*, 17(4):1115–1124.

Naldi, A. (2018). BioLQM: A Java Toolkit for the Manipulation and Conversion of Logical Qualitative Models of Biological Networks. *Frontiers in Physiology*, 9:1605.

Naldi, A., Hernandez, C., Levy, N., Stoll, G., Monteiro, P. T., Chaouiya, C., Helikar, T., Zinovyev, A., Calzone, L., Cohen-Boulakia, S., Thieffry, D., and Paulevé, L. (2018). The CoLoMoTo Interactive Notebook: Accessible and Reproducible Computational Analyses for Qualitative Biological Networks. *Frontiers in Physiology*, 9:680.

Paulevé, L., Kolčák, J., Chatain, T., and Haar, S. (2020). Reconciling qualitative, abstract, and scalable modeling of biological networks. *Nature Communications*, 11(1):4256.

Paulevé, L. and Sené, S. (2022). Boolean Networks and Their Dynamics: The Impact of Updates. In De Maria, E., editor, *Systems Biology Modelling and Analysis*, pages 173–250. Wiley, 1 edition.

Schwab, J. D., Kühlwein, S. D., Ikonomi, N., Kühl, M., and Kestler, H. A. (2020). Concepts in Boolean network modeling: What do they all mean? *Computational and Structural Biotechnology Journal*, 18:571–582.

Selvaggio, G., Chaouiya, C., and Janody, F. (2021). In Silico Logical Modelling to Uncover Cooperative Interactions in Cancer. *International Journal of Molecular Sciences*, 22(9):4897.

Thomas, R. (1973). Boolean formalization of genetic control circuits. *Journal of Theoretical Biology*, 42(3):563–585.

Thomas, R. (1978). Logical analysis of systems comprising feedback loops. *Journal of Theoretical Biology*, 73(4):631–656.

Tomlin, C. J. and Axelrod, J. D. (2007). Biology by numbers: mathematical modelling in developmental biology. *Nature Reviews. Genetics*, 8(5):331–340.

Torres, N. V. and Santos, G. (2015). The (Mathematical) Modeling Process in Biosciences. *Frontiers in Genetics*, 6:354.

Wang, R.-S., Saadatpour, A., and Albert, R. (2012). Boolean modeling in systems biology: an overview of methodology and applications. *Physical Biology*, 9(5):055001.

20

Supplementary data

Nadine Ben Boina[1,2], Brigitte Mossé[1], Anaïs Baudot[2,3,4], and Elisabeth Remy[1]

[1] *I2M, CNRS, Aix Marseille University, Marseille, France.*
[2] *Aix Marseille Univ, INSERM, MMG, Marseille, France*
[3] *CNRS, Marseille, France*
[4] *Barcelona Supercomputing Centre, Barcelona, Spain*

04 - 2023

S1 Construction of Multivalued Refinements of a BM

Given a BM f on $\mathbb{B}^n$, we recall that given $j_0 \in \{1,\dots,n\}$, the map f_{j_0} can be set in a disjunctive normal form (DNF). In other words, $f_{j_0}(x)$ is written as a disjunction of clauses that are themselves conjunctions of radicals $w\,x_j$ as below:

$$f_{j_0}(x) = \bigvee_{h\in\{1,\dots,s_{j_0}\}} \bigwedge_{k\in\{1,\dots,r_{j_0,h}\}} w_{j_0,h,k}\, x_{j_{0,h,k}} \quad , \tag{1}$$

where operators $\wedge$, $\vee$ stand for AND, OR respectively, and the $w_{j_0,h,k}$ are equal to ε (empty string), or $\neg$ (NON).
Moreover, this DNF is supposed to be a shortest one - this ensures that components involved in the expression of f_{j_0} are effective regulators of g_{j_0}, and avoids redundancy.

A way to build a refinement h of f on $X = \prod_{j=1}^{n} \{0,1,\dots,m_j\}$ is then the following:

- Starting from (1), and the current variable becoming $x \in X$, for each $j = j_{j_0,h,k}$ in (1) for which $m_j > 1$, we choose a partition of $\{0,\dots,m_j\}$ of the form $\{0,\dots,s\}\cup\{s+1,\dots,m_j\}$ and we replace $w_{j_0,h,k}\, x_{j_{j_0,h,k}}$ with $x_{j_{j_0,h,k}} \geq s+1$ if $w_{j_0,h,k} = \varepsilon$, and with $x_{j_{j_0,h,k}} < s+1$ if $w_{j_0,h,k} = \neg$. Clauses with indices j such that $m_j = 1$ remain unchanged.

- In this way, we obtain a function of the variable $x \in X$ with values in $\mathbb{B}$; let $\mathcal{H}_{j_0}$ be defined by $\mathcal{H}_{j_0}(x) = +1$ when this value is 1, and $\mathcal{H}_{j_0}(x) = -1$ when it is 0.

- Finally, we define the refinement $h = (h_1 \dots, h_n)$ setting for $j \in \{1,\dots,n\}$

$$h_j(x) = max(0, min(m_j, x_j + \mathcal{H}_j(x))),$$

that ensures that the map h is actually a map from X to itself.

S2 Comparison of Reachability Properties in a BM and Multi-valued Refinements

Proposition 1. *Let f be a BM on $\mathbb{B}^n$, the map h a refinement of f on $X = \prod_{j=1}^{n} \{0, 1, \ldots, m_j\}$, and $J = \{j \in \{1, \ldots, n\} \; ; \; m_j > 1\}$. We suppose that no component g_j of J is self-inhibited.*

Set a and a' two Boolean states of $\{0,1\}^n$, and b and b' the elements of X obtained replacing the coordinate a_j of a (resp. a'_j of a') by m_j when $a_j = 1$ (resp. $a'_j = 1$), for each $j \in J$. If there exists a trajectory from a to a' in the asynchronous STG of f, then there exists a trajectory from b to b' in the asynchronous STG of h.

Proof. It is sufficient to prove this property in the case where there is a transition $a \to a'$ in the asynchronous STG of f, and $a_1 \neq a'_1$. Suppose for instance that $a_1 = 0$ and $a'_1 = 1$ (the case $a_1 = 1$ and $a'_1 = 0$ can be treated in a similar way). Then the states b and b' differ only by their first coordinates, $b_1 = 0$ and $b'_1 = m_1$.

In the previous notation, a conjunctive clause of the form $\bigwedge_{k \in \{1,\ldots,r\}} w_k \, x_{j_k}$ in the logical formula of f_1 is made true by a. This implies that the corresponding clause of $\tilde{h}$ built as above is made true by b.

- In the case $m_1 = 1$, this gives a transition from b to b'.

- In the case $m_1 > 1$, there is a transition from b to the state $b^{(1)}$ obtained from b by changing b_1 into 1. Considering the hypothesis that x_1 does not occur in the clause $\bigwedge_{k \in \{1,\ldots,r\}} w_k \, x_{j_k}$, this can be iterated until getting all the transitions $b \to b^{(1)} \to \cdots \to b^{(m_1)} = b'$, as expected. □

Remark 1. *If a component g_j of J is self-inhibited, this proposition is irrevocably defeated. This can be illustrated for instance setting $n = 2$, and for $x \in \mathbb{B}^2$,*
$f_1(x) = \neg x_1 \; \vee \; \neg x_2 \, , \quad f_2(x) = \neg x_1 .$

- An example of refinement h of f on $X = \{0, 1, 2\} \times \{0, 1\}$ is given by
$\mathcal{H}_1(x) = +1$ *iff* $x_1 < 2 \; \vee \; \neg x_2 \, , \quad \mathcal{H}_2(x) = +1$ *iff* $x_1 < 1 \, .$
There is a transition from 11 *to* 01 *in the asynchronous STG of f, but no trajectory from* 21 *to* 01 *in the asynchronous STG of h.*

- Another example of refinement h' of f on $X = \{0, 1, 2\} \times \{0, 1\}$ is given by
$\mathcal{H}'_1(x) = +1$ *iff* $x_1 < 1 \; \vee \; \neg x_2 \, , \quad \mathcal{H}'_2(x) = +1$ *iff* $x_1 < 2 \, .$
The state 10 *is a fixed point of f, and the lonely attractor of its asynchronous dynamics. There is a trajectory from* 01 *to* 10 *in the asynchronous STG of f, but no trajectory from* 01 *to* 20 *in the asynchronous STG of h'. Moreover, there is a new attractor in this STG of h', that is a cycle of length* 2 *between* 01 *and* 11.

x	$f(x)$
0 0	*1 1*
0 1	*1 1*
1 0	*1 0*
1 1	*0 0*

x	$h(x)$
0 0	*1 1*
1 0	*2 0*
2 0	*2 0*
0 1	*1 1*
1 1	*2 0*
2 1	*1 0*

x	$h'(x)$
0 0	*1 1*
1 0	*2 1*
2 0	*2 0*
0 1	*1 1*
1 1	*0 1*
2 1	*1 0*

We are especially interested in the fixed points of the dynamics, for which the way we build refinements gives the following.

Proposition 2. *Let f be a BM on $\{0,1\}^n$, the map h a refinement of f on $X = \prod_{j=1}^{n} \{0, 1, \ldots, m_j\}$, and $J = \{j \in \{1, \ldots, n\} \; ; \; m_j > 1\}$. We suppose that no component g_j of J is self-inhibited.*

- *Set ω be a fixed point of f, and ω' the element of X obtained replacing the coordinates ω_j of ω equal to 1 by m_j, for each $j \in J$. Then ω' is a fixed point of h, and all the fixed points of h are obtained in this way.*
- *Let suppose that the attractors of the asynchronous dynamics of f are l fixed points $\omega^{(1)}$, ..., $\omega^{(l)}$. Then the l fixed points $\omega'^{(1)}$, ..., $\omega'^{(l)}$ of h obtained as above are the lonely attractors of the asynchronous dynamics of h.*

Proof.

- The fact that ω' is a fixed point of h comes from the inequalities required to be a refinement and from $\alpha(\omega') = \{\omega\}$.

  Conversely, suppose that a state $x \in X$ is a fixed point of h. Then, by the hypothesis, all the coordinates x_j with $j \in J$ have to be equal to 0 or m_j, and $\alpha(x)$ is reduced to one Boolean state y. Finally, the state x being a fixed point of h, the state y is necessarily a fixed point of f, as expected.

- Set $x \in X$. By the hypothesis, there is a trajectory from x to a state z whose coordinates z_j with $j \in J$ are all equal to 0 or m_j. Then, $\alpha(z)$ is reduced to one Boolean state y. There is a trajectory from y to some fixed point $\omega^{(k)}$ of f. By Proposition 1, there is thus a trajectory from z to $\omega'^{(k)}$. In conclusion, all the states x of X lead to a fixed point of h, that achieves the proof.

□

S3 Generalization of the Completeness Property of the m.p. Scheme to the Partial m.p. Schemes

For the sake of convenience, in this supplementary, trajectories from a state x to a state y will be denoted in a contracted manner $x \xrightarrow[- - -]{f}{}^{*} y$.

S3

Completeness property of the m.p. scheme

We begin recalling Paulevé & al theorem ([9]), and give a detailed constructive proof.

Let us consider an integer $n \geq 1$, a BM f of dimension n, and a refinement h of f defined on $X = \prod_{j=1}^{n} \{0, 1, \ldots, m_j\}$, where $m_j \in \mathbb{N}^*$ for each $j \in \{1, \ldots, n\}$, and at least one of the m_j is > 1.

For $j \in \{1, \ldots, n\}$, we denote by F_j the map defined by $F_j(x) = f_j(x) - x_j$ for $x \in \mathbb{B}^n$, and by H_j the map defined by $H_j(x) = h_j(x) - x_j$ for $x \in X$ (these maps are with values in $\{-1, 0, +1\}$).

Let x be an element of X. Let us call *m.p. state compatible with* x any element $\hat{x}$ of $X_{m.p.} = \{0, 1, i, d\}^n$ such that for each $j \in \{1, \ldots, n\}$,

- if $x_j = 0$, then $\hat{x}_j = 0$,
- if $x_j = m_j$, then $\hat{x}_j = 1$,
- if $x_j \notin \{0, m_j\}$, then $\hat{x}_j = i$ or d.

Remark 2. *In both definitions of the m.p. scheme related to f and of the multivalued refinements of f, the associated components are supposed to assume that any intermediate level of some component g (the levels i and d, the levels l such that $0 < l < m_j$ for some j) can be considered as levels where g is, or is not, active.*

For $x \in X$, the set
$\alpha(x) = \{x' \in \mathbb{B}^n \,;\, \forall j \in \{1, \ldots, n\},\, (x_j = 0 \Rightarrow x'_j = 0)\ and\ (x_j = m_j \Rightarrow x'_j = 1)\}$,
and for $x \in X_{m.p.}$ the set
$\gamma(x) = \{x' \in \mathbb{B}^n \,;\, \forall j \in \{1, \ldots, n\},\, (x_j = 0 \Rightarrow x'_j = 0)\ and\ (x_j = 1 \Rightarrow x'_j = 1)\}$
are introduced to this end.

Hence, the proof of the following theorem is essentially based on the fact that if $x \in X$, and $\hat{x}$ is an element of $X_{m.p.}$ compatible with x, then $\alpha(x) = \gamma(\hat{x})$.

Theorem 1. *In the previous notation, let $x \xrightarrow[asyn]{h}{}^{*} y$ be a trajectory in the asynchronous STG of h. For any element $\hat{x}$ of $X_{m.p.}$ compatible with x, there exists in the STG of the m.p. dynamics of f a trajectory $\hat{x} \xrightarrow[m.p.]{f}{}^{*} \hat{y}$ such that $\hat{y}$ is compatible with y.*

Proof. It is clear that it is sufficient to prove the result in the case where the asynchronous trajectory is reduced to one transition, that is $x \xrightarrow[asyn]{h} y$.

Let $\hat{x}$ be an element of $X_{m.p.}$ compatible with x, and $j \in \{1, \ldots, n\}$ the integer such that $x_j \neq y_j$.

For convenience, we suppose that $j = 1$, and we detail the construction of $\hat{x} \xrightarrow[m.p.]{f}{}^{*} \hat{y}$ in the case $H_1(x) > 0$, the case $H_1(x) < 0$ being similar.

- If $m_1 = 1$, $x_1 = \hat{x}_1 = 0$ and $y_1 = 1$, then the existence of $x' \in \beta(x) = \gamma(\hat{x})$ such that $F_1(x') > 0$ gives
$$\hat{x} = (0, \hat{x}_2, \ldots, \hat{x}_n) \;\rightarrow\; (i, \hat{x}_2, \ldots, \hat{x}_n) \;\rightarrow\; (1, \hat{x}_2, \ldots, \hat{x}_n)\,.$$
- If $m_1 > 1$, $x_1 = \hat{x}_1 = 0$ and $y_1 = 1$, then the existence of $x' \in \beta(x) = \gamma(\hat{x})$ such that $F_1(x') > 0$ gives
$$\hat{x} = (0, \hat{x}_2, \ldots, \hat{x}_n) \;\rightarrow\; (i, \hat{x}_2, \ldots, \hat{x}_n)\,.$$

S4

- If $m_1 > 1$, $x_1 = l$, where $0 < l < m_1 - 1$ and $y_1 = l + 1$, then $\hat{x}_1 = i$ or d and $\hat{x}$ is compatible with y :

$$\text{stay on } \hat{x}.$$

- If $m_1 > 1$, $x_1 = m_1 - 1$, $y_1 = m_1$ and $\hat{x}_1 = i$, then the existence of $x' \in \beta(x) = \gamma(\hat{x})$ such that $F_1(x') > 0$ gives

$$\hat{x} = (i, \hat{x}_2, \ldots, \hat{x}_n) \ \rightarrow \ (1, \hat{x}_2, \ldots, \hat{x}_n) \, .$$

- If $m_1 > 1$, $x_1 = m_1 - 1$, $y_1 = m_1$ and $\hat{x}_1 = d$, then the existence of $x' \in \beta(x) = \gamma(\hat{x})$ such that $F_1(x') > 0$ gives

$$\hat{x} = (d, \hat{x}_2, \ldots, \hat{x}_n) \ \rightarrow \ (i, \hat{x}_2, \ldots, \hat{x}_n) \ \rightarrow \ (1, \hat{x}_2, \ldots, \hat{x}_n) \, .$$

□

Completeness property of the partial m.p. schemes

We consider now an integer $n \geq 1$, a BM f of dimension n, a non-empty subset J of $\{1, \ldots, n\}$, and a refinement h of f defined on $X = \prod_{j=1}^{n} \{0, 1, \ldots, m_j\}$, where $m_j \in \mathbb{N}^*$ for each $j \in \{1, \ldots, n\}$, and $m_j > 1$ if and only if $j \in J$.

As above, for $j \in \{1, \ldots, n\}$, we denote by F_j the map defined by $F_j(x) = f_j(x) - x_j$ for $x \in \mathbb{B}^n$, and by H_j the map defined by $H_j(x) = h_j(x) - x_j$ for $x \in X$.

Let x be an element of X. We call *^{J}m.p. state compatible with* x any element $\hat{x}$ of $X_{{}^J m.p.}$ such that for each $j \in \{1, \ldots, n\}$,

- if $x_j = 0$, then $\hat{x}_j = 0$,
- if $x_j = m_j$, then $\hat{x}_j = 1$,
- if $x_j \notin \{0, m_j\}$, then $\hat{x}_j = i$ or d.

Theorem 2. *In the previous notation, let $x \xrightarrow[asyn]{h} {}^* y$ be a trajectory in the asynchronous STG of h. For any element $\hat{x}$ of $X_{{}^J m.p.}$ compatible with x, there exists in the STG of the ^{J}m.p. dynamics of f a trajectory $\hat{x} \xrightarrow[{}^J m.p.]{f} {}^* \hat{y}$ such that $\hat{y}$ is compatible with y.*

Proof. The proof is an easy adaptation of the one of Theorem 1: we restrict ourselves to the case where the considered trajectory is reduced to one transition $x \xrightarrow[asyn]{h} y$, with $x_1 \neq y_1$ and $H_1(x) > 0$. Let $\hat{x}$ be an element of $X_{{}^J m.p.}$ compatible with x. A suitable trajectory $\hat{x} \xrightarrow[{}^J m.p.]{f} {}^* \hat{y}$ is obtained on the following way:

- If $1 \in J$ the construction of $\hat{x} \xrightarrow[{}^J m.p.]{f} {}^* \hat{y}$ is the same than in the proof of Theorem 1.
- If $1 \notin J$, then $x_1 = 0$, $y_1 = 1$ and the existence of $x' \in \beta(x) = \gamma(\hat{x})$ such that $F_1(x') > 0$ gives

$$\hat{x} = (0, \hat{x}_2, \ldots, \hat{x}_n) \ \rightarrow \ (1, \hat{x}_2, \ldots, \hat{x}_n) \, .$$

S5

□

S4 Multivalued refinement of the BM of Early Hematopoietic Stem Cell Aging

You can find the model in ".bnet" format within the GitHub repository's "Examples/Hérault, Léonard et al (2022)" directory. The entirety of the logical rules for the multivalued refinement presented in the results section is as follows:

$Egr1 = Gata2 \,\&\, Junb$
$Junb = Egr1:2 \,|\, Myc$
$Blacf1 = Myc$
$Myc = Cebpa \,\&\, Blacf1$
$Fli1:2 = Junb \,|\, Gata1:2 \,\&\, !Klf1)$
$Gata2:2 = (Gata2 \,\&\, !Gata1 \,\&\, !Zfpm1) \,|\, (Egr1 \,\&\, !Gata1 \,\&\, !Zfpm1 \,\&\, !Spi1)$
$Spi1:2 = (Spi1 \,\&\, !Gata1) \,|\, (Cebpa \,\&\, !Gata1 \,\&\, !Gata2:2)$
$Cebpa = (Gata2 \,\&\, !Ikzf1) \,|\, (Spi1 \,\&\, !Ikzf1)$
$Gata1:2 = Fli1 \,|\, (Gata2 \,\&\, !Spi1) \,|\, (Gata1 \,\&\, !Ikzf1 \,\&\, !Spi1)$
$Klf1 = Gata1 \,\&\, !Fli1:2$
$Tal1 = Gata1 \,\&\, !Spi1$
$Ikzf1 = Gata2$
$Zfpm1 = Gata1$
$CDK46CycD = Bclaf1 \,|\, Myc$
$CIPKIP = Junb$

To simplify notation, we denote the activity of a node by its label (e.g., $Gata2$, not x_{Gata2}). Additionally, for multivalued components, when the threshold of regulation is other than 1, it is denoted by ":level". For instance, $Gata2:2$ signifies that $x_{Gata2} = 2$ whereas $Gata2$ means $x_{Gata2} = 1$.

S5 Multivalued refinement of the BM of Asymmetric Stem Cell Division in *Arabidopsis Thaliana* Root

You can find the multivalued refinement discussed in the corresponding result section within the GitHub repository's "Examples/García-Gómez, Mónica L et al (2020)" directory.

4 Analysis of RNA-sequencing data from vascular smooth muscle cells in wild-type and HGPS model mice

Summary

4.1 Introduction . . . 76
4.1.1 What is Hutchinson-Gilford Progeria Syndrome? . . . 76
4.1.2 RNA-sequencing . . . 78
4.1.3 Aim of the study . . . 83
4.2 Material and methods . . . 84
4.2.1 Preliminary work . . . 84
4.2.1.1 VSMC preparation and RNA extraction . . . 84
4.2.1.2 RNA-sequencing . . . 84
4.2.2 Exploratory data analysis . . . 85
4.2.3 Differential expression analysis . . . 85
4.2.4 mRNAs functional analysis . . . 86
4.2.5 miRNA-mRNA Interaction Analysis . . . 87
4.2.6 Multi-omics factor analysis . . . 87
4.3 Results . . . 87
4.3.1 mRNA expression analysis . . . 87
4.3.1.1 Exploratory data analysis . . . 88
4.3.1.2 Differential mRNA expression analysis with DESeq2 90
4.3.1.3 Functional analysis of the differentially expressed mRNAs . . . 92
4.3.2 miRNA expression analysis . . . 96
4.3.2.1 Data quality assessment . . . 97

4.3.2.2 Differential miRNA expression analysis 98
4.3.3 miRNA-mRNA joint analysis 100
4.3.3.1 Interaction network 100
4.3.3.2 Multi-omics factor analysis 101
4.4 Discussion and conclusion 105

In this chapter, I present the analysis that I performed on a dataset obtained from an RNA sequencing (RNA-seq) experiment related to Hutchinson-Gilford Progeria Syndrome (HGPS). This dataset is part of a work led by Patrice Roll and his team at the MMG unit. I begin the chapter with an overview of the biological background, including the clinical and molecular features of HGPS, followed by an introduction to the principles of RNA-seq and an explanation of the aim of the study. I then provide a detailed description of the materials and methods used in the analysis, explicitly distinguishing the tasks that I performed from those performed by Patrice Roll's team. Finally, I present the results and conclude with a discussion of the findings and limitations.

4.1 Introduction

4.1.1 What is Hutchinson-Gilford Progeria Syndrome?

HGPS is a rare genetic disorder affecting approximately 1 per 4 to 8 million live births. It is characterised by clinical symptoms resembling accelerated physiological ageing (Harhouri et al. 2018). HGPS is caused by a point mutation in the *LMNA* gene (Eriksson et al. 2003), which encodes Lamin A and C, two proteins essential for maintaining nuclear structure and integrity. This mutation results in the production of Progerin, an aberrant form of Lamin A. Progerin differs from normal Lamin A by a deletion of 50 amino acids near its C-terminal region. This deletion prevents Progerin from undergoing the final cleavage by the metalloprotease ZMPST24 in the normal post-translational processing of Lamin A. The protein remains permanently anchored to the inner nuclear membrane due to the persistence of a farnesyl group. The accumulation of Progerin at the nuclear membrane is toxic and causes structural defects, including inward bulging of the nuclear envelope and increased nuclear stiffness. These alterations disrupt nuclear architecture and leads to an array of cellular defects (Cisneros et al. 2023; Goldman et al. 2004; Lamis et al. 2022).

At the cellular level, patients with HGPS exhibit reduced cell proliferation, premature senescence, impaired DNA repair, increased reactive oxygen species production, mitochondrial dysfunction, loss of peripheral heterochromatin, and telomere shortening, among other (Cisneros et al. 2023). These cellular defects collectively drive the accelerated ageing phenotype observed in HGPS, including growth impairment, thin skin, loss of subcutaneous fat, alopecia, osteoporosis, hearing loss, cardiovascular complications, and other age-related symptoms (Harhouri et al. 2018).

To date, there is no cure for HGPS. Proposed treatment strategies focus either on genetic approaches to eliminate the disease-causing mutation, on ways to reduce Progerin production, or on mitigating its toxic effects (Harhouri et al. 2018). Notably, the molecule lonafarnib has been approved by the FDA for the treatment of HGPS and related syndromes caused by the accumulation of farnesylated prelamin A. Clinical studies have shown that HGPS patients treated with lonafarnib experience a survival benefit of 2.5 years compared to untreated individuals (Suzuki et al. 2023).The average life expectancy of individuals with HGPS is approximately 14.6 years, with myocardial infarction being the leading cause of death (Harhouri et al. 2018).

The main cause of myocardial infarction in patients with HGPS is atherosclerosis, a disease characterised by arterial hardening and narrowing. Most patients develop severe forms of this atherosclerosis. Atherosclerosis begins with the accumulation of low-density lipoproteins in the arterial walls, triggering inflammation and the recruitment of immune cells. Over time, plaques of debris and cholesterol are formed. Vascular smooth muscle cells (VSMCs) are specialised muscle cells that primarily compose the tunica media, the middle layer of the blood vessel wall. During atherosclerosis formation, VMSCs migrate to stabilise the plaques by forming a fibrous cap. Plaque rupture can expose its contents, leading to blood clots and potentially fatal cardiovascular events. As they age, patients with HGPS experience significant VSMC depletion in their arteries, contributing to advanced atherosclerosis and subsequent cardiovascular complications (Coll-Bonfill et al. 2023; Hamczyk et al. 2019). The mechanisms by which Progerin induces widespread VSMC loss and vessel alterations remain poorly understood (Coll-Bonfill et al. 2023). Given the critical role of VSMC depletion in HGPS pathology, Patrice Roll and his team investigated the mechanisms driving VSMC disruption in HGPS, and how they contribute to the disease progression. One of their strategies is to use a

global approach by RNA-seq. Before discussing the aim of the study, I recall the principle of RNA-seq.

4.1.2 RNA-sequencing

RNA-seq is a technique used to identify the sequence and abundance of RNA molecules in cells and tissues. RNA-seq has proven to be very effective in transcriptomic studies to uncover which gene, encoding a protein or any type of regulatory RNAs such as micro RNA (miRNA), are involved in a biological process of interest. A classical RNA-seq experiment can be divided into the following six steps (Conesa et al. 2016).

Step 1: Experimental Design The first step in an RNA-seq experiment is to establish the experimental design. This includes defining the biological question, selecting appropriate controls, and determining experimental conditions (e.g., mutant vs normal cells). Two very important aspects to consider are determining the number of biological replicates to ensure statistical power and addressing potential batch effects (Conesa et al. 2016).

RNA-seq experiments yield more reliable results with an increased number of biological replicates, with three being considered the consensus minimum. Having more replicates enhances statistical power, reducing the likelihood of false positives and increasing the ability to detect true differences between experimental conditions. The choice of replicate number should also consider the variability within biological samples, as high variability may necessitate additional replicates.

Another critical aspect is managing batch effects, which can arise when experiments are conducted in multiple batches — for example, at different times or by different experimenters. These batch effects can introduce systematic biases that obscure true biological signals. To mitigate such biases, experimental batches must be clearly identified and recorded as part of the metadata.

Thorough planning and consideration of these points during the experimental design phase allow to maximise the quality and interpretability of RNA-seq data.

Step 2: RNA Preparation Once the experimental design is defined, the next step is the extraction of RNA from the biological samples. To begin, cells from the organism or tissue of interest, under the specific condition being studied, are lysed. This process breaks the cells, releasing their RNA content. The RNA is

then isolated using specialised kits or methods to ensure purity and quality for downstream use.

RNA integrity is crucial to obtain reliable results in RNA-seq, as degraded RNA can lead to incomplete sequencing data. The RNA Integrity Number (RIN) is commonly used to measure RNA integrity, with a higher RIN value (typically above 7) indicating intact RNA. Low RIN values (typically below 5) suggest degraded RNA, which may compromise the accuracy of sequencing results. Depending on the research focus, the extracted RNA may undergo enrichment/filtering to isolate specific types of RNAs (Schroeder et al. 2006). For studies focusing on mRNA (the protein-coding RNAs), mRNA is typically enriched by selecting polyadenylated (poly-A) sequences. To capture a broader range of RNA types, including regulatory RNAs (like miRNA), ribosomal RNA (rRNA) depletion is used. Indeed, since rRNA makes up the majority of RNA in a cell, removing it helps focus on other RNA species (S. Zhao; Zhang, et al. 2018).

Once the RNA content is isolated and enriched, it is reverse-transcribed into complementary DNA (cDNA). This process converts the RNA content into a more stable DNA form, suitable for sequencing. The resulting cDNA is then fragmented into smaller pieces adapted to the sequencing platform. After fragmentation, adapter sequences are added to the ends of DNA fragments. These adapters serve two purposes: they provide a binding site for the sequencing primers and ensure the correct orientation for sequencing (Kukurba et al. 2015; Satam et al. 2023).

Step 3: Sequencing and data quality assessment Next-generation sequencing (NGS) platforms enable the parallel sequencing of millions of DNA fragments. High-depth sequencing ensures accurate detection of gene expression. Sequencing depth refers to the number of times a fragment is sequenced, with deeper sequencing providing greater confidence in the results (Satam et al. 2023). The sequencer outputs raw data in the form of sequence reads, each representing a fragment of the original RNA that has been reverse-transcribed and fragmented into cDNA. The raw sequencing data consists of the nucleotide sequences (the reads) and associated quality scores, known as Phred scores. These scores indicate the reliability of each base call, with higher scores corresponding to greater accuracy in sequencing (Cock et al. 2010). To ensure data reliability, this raw data undergoes quality assessment using tools like FastQC (Andrews et al. 2012). Fore instance, FastQC provide various quality metrics, including:

— Adapter contamination: Ensuring no unwanted adapter sequences remain in the data after fragmentation and adapter ligation.
— Sequence quality: Examining the base calling quality across the reads to identify low-quality regions or positions.
— GC content: A check on the overall GC content, which can be informative in identifying biases in the sequencing process.
— Overrepresented k-mers: Detecting sequences that appear disproportionately often, which might indicate contamination or over-amplification of certain fragments.
— Duplicated reads: Identifying duplicate reads that may arise from PCR amplification, which can skew downstream analyses.

Quality metrics such as duplication rates, k-mer frequencies, and GC content should be consistent across samples within the same experiment. Significant deviations in these metrics between samples might indicate technical issues or biases that need to be addressed before proceeding with downstream analyses. After quality assessment, the raw data is cleaned using software tools like Trimmomatic (Bolger et al. 2014). This step involves removing low-quality bases at the ends of reads and trimming adapter sequences that may have been retained during sequencing. Additionally, low-quality reads, which could introduce errors into the analysis, are discarded.

Step 4: Alignment of reads Once the sequencing reads have been generated, the next step is to align these reads to a reference genome or transcriptome. Popular alignment algorithms include Bowtie, a fast and memory-efficient aligner suitable for short reads (Langmead et al. 2009), and STAR, which performs well with longer reads and is known for its high sensitivity and accuracy in mapping spliced RNA sequences (Dobin et al. 2013).

The reads are mapped to specific genomic regions. Unaligned reads may suggest sequencing errors, or may indicate the presence of genomic regions that are not well-characterised by the reference genome or transcriptome. Aligned reads are saved in Sequence Alignment/Map (SAM) or Binary Alignment/Map (BAM) formats. These formats contain not only the sequence data, but also information about the alignment quality and the position of the reads in the reference genome or transcriptome. To ensure high-quality mappings, several tools, RSeQC (Wang et al. 2012), Qualimap (Okonechnikov et al. 2016), can be used to assess the

alignment results. These tools provide important mapping quality metrics, including the percentage of mapped reads to the reference genome. For typical RNA-seq experiments, between 70% and 90% of the reads are expected to map to the genome, depending on the aligner used. A lower mapping percentage may indicate issues such as contamination or poor RNA quality. Another key metric is the uniformity of read coverage on exons. An even distribution of reads across exonic regions is desirable. If reads accumulate primarily at the 3' end of transcripts, this could suggest poor RNA quality from the starting material, potentially indicating degradation or other quality control issues (Conesa et al. 2016; Gallego Romero et al. 2014).

Step 5: Quantification After aligning the sequencing reads to a reference genome or transcriptome, the next step is to quantify transcript expression. Quantification consist in measuring the abundance of each transcript, which is the number of reads that map to a transcript. First, reads are mapped to annotated genomic regions, including features like genes, exons, or transcripts. Then, reads are quantified to generate a count matrix, which summarises the number of reads assigned to each feature. Common tools used for this summarization step include featureCounts (Liao et al. 2014) and **HTSeq-count** (Putri et al. 2022). Both tools require an annotated reference genome to match the reads to features like genes or exons.

Raw read counts alone are not sufficient for comparing transcript expression across samples because they are influenced by factors such as sequence length and total number of reads. Normalization methods are applied to adjust for these factors, ensuring that expression values (i.e., counts) are comparable across different samples. Common normalization methods include TPM (transcripts per million), FPKM (fragments per kilobase of transcript per million mapped reads), and RPKM (reads per kilobase of exon model per million reads), which all normalizes for both sequencing depth and gene length. Expression values normalized with these methods are suitable for comparison within samples. For comparison across samples, other methods are necessary, such as TMM (trimmed mean of M-values) and size factor normalization (S. Zhao; Ye, et al. 2020; Y. Zhao et al. 2021).

Cross-sample comparisons are used for differential expression analysis (see the next section). The DESeq2 tool is commonly used for this analysis and applies size factor normalization. This method corrects for differences in sequencing depth between samples but does not account for gene length. Since each sample in RNA-

seq may have a different total number of reads (library size), direct comparison of expression levels across samples can be challenging. DESeq2 calculates a size factor for each sample, which reflects its relative library size (total number of reads). The size factor is determined by comparing the read counts of genes expected to have stable expression across all samples. After calculating the size factors, DESeq2 scales the read counts for each sample by dividing them by their corresponding size factor. This adjustment compensates for differences in total read counts between samples, making the expression levels more comparable. By applying this scaling factor, the expression levels of genes are corrected so that variations in expression across samples are driven by true biological differences, rather than technical variations such as differences in sequencing depth (Love et al. 2014).

Step 6: Differential expression analysis Once the transcript quantification and normalization steps are completed, the next step is differential expression analysis. This analysis consists in examining the count matrix to identify genes or transcripts that exhibit significant changes in expression between different experimental conditions (Love et al. 2014). Several statistical tools are available to perform this analysis, DESeq2 is one of the most widely used. DESeq2 uses a negative binomial model to account for the higher-than-expected variability in count data, providing robust statistical methods for detecting differential expression. DESeq2 also includes built-in methods for data normalization and transformation, making it a comprehensive tool for RNA-seq analysis (Love et al. 2014).

DESeq2 performs hypothesis testing to determine which genes or transcripts exhibit significant expression changes. It compares the observed counts for each gene between experimental conditions and assesses whether the differences are larger than what would be expected due to random variation. The result is a list of genes, each associated with a p-value and a log fold change. The log fold change indicates the magnitude and direction of gene expression differences between conditions, with positive values representing upregulation and negative values representing downregulation. The p-value is used to assess the statistical significance of these changes, with a threshold (typically 0.05) used to determine significance (Conesa et al. 2016).

The resulting list of differentially expressed genes (DEGs) can be further analysed to investigate their biological functions and involvement in specific pathways. Functional analysis tools, such as over-representation analysis (ORA) or Gene Set

Enrichment Analysis (GSEA), identify biological terms associated with a set of genes or transcripts (Chen et al. 2023). These tools assess whether certain biological terms, such as Gene Ontology (GO) terms or pathway identifiers, are enriched in the given gene set. The terms are typically derived from reference databases such as Gene Ontology (Gene Ontology Consortium 2004), KEGG (Kanehisa et al. 2000), or Reactome (Fabregat et al. 2018). Alternatively, network analysis can be performed to explore potential interactions between genes, or to investigate regulatory mechanisms that may be influencing the observed changes in gene expression (Conesa et al. 2016).

4.1.3 Aim of the study

Recall that HGPS is a genetic disorder with symptoms resembling accelerated physiological ageing. HGPS is caused by a point mutation in the *LMNA* gene. This mutation results in the production of Progerin, an aberrant and toxic form of Lamin A that disrupts the nuclear membrane. Currently, there is no cure for HGPS, and myocardial infarction is the leading cause of death (Eriksson et al. 2003; Harhouri et al. 2018). The main cause of myocardial infarction in patients with HGPS is atherosclerosis (arterial hardening and narrowing). Studies have shown that HGPS patients exhibit VSMC depletion in the arteries, which contributes to advanced atherosclerosis and subsequent cardiovascular complications (Coll-Bonfill et al. 2023; Hamczyk et al. 2019). Despite these findings, the mechanisms by which Progerin induces widespread VSMC loss and vascular alterations remain poorly understood.

Therefore, the aim of this study is to investigate the molecular mechanisms underlying VSMC loss and its contribution to HGPS progression, using RNA-seq. Specifically, we are interested in the following questions: first, what is the impact of Progerin accumulation on gene expression in VSMCs? Second, given that HGPS is characterised by accelerated ageing, how does ageing affect HGPS VSMCs compared to wild-type VSMCs?

To tackle these questions, Patrice Roll and his team performed RNA-seq on primary VSMC cultures from four groups of mice: HGPS and wild-type (WT), at two different passages of cell culture. Passages are a succession of subculture, passage number serves as a proxy for cellular age. I analysed the resulting RNA-seq to compare gene expression levels across these groups, aiming to identify biological

functions impaired in HGPS VSMCs, at different passages. The study focused on two types of RNA: messenger RNA (mRNA, code for protein) and miRNA (regulatory RNA). Given that previous studies have highlighted the role of specific miRNAs in HGPS pathology, their inclusion in the analysis was deemed relevant (Frankel et al. 2022).

4.2 Material and methods

4.2.1 Preliminary work

This preliminary work section describes the procedures conducted by Patrice Roll's team.

4.2.1.1 VSMC preparation and RNA extraction

VSMCs were isolated from the aortas of one-month-old HGPS mouse models ($Lmna^{G609G/G609G}$) and wild-type (WT) mice (from the same litter). Aortas from two mice were pooled to establish each VSMC lineage. For each group of mice (HGPS and WT VSMCs), three lineages were established, and primary cultures of these six lineages were maintained. Biological samples were collected at two passage numbers: passage 8 (P8), representing early passages, and passage 12 (P12), representing late passages, to study the effects of cellular ageing. Total RNA, including miRNA, was extracted using the miRNeasy kit (QIAGEN).

4.2.1.2 RNA-sequencing

RNA samples were prepared according to the library kit manufacturer's protocol, indexed, pooled, and sequenced on an Illumina sequencer. Initial quality assessment of raw sequencing reads was performed using *FastQC*. Reads were aligned to the *GRCm38 Mouse Genome Reference* using *STAR*. The quality of alignments was assessed using both *Qualimap* and *SAMtools*. Mapping rates consistently exceeded 70%, indicating high-quality alignments. The mRNA and miRNA counts were quantified using *StringTie* with annotations from the *GENCODE vM15* database.

The following sections detail the bioinformatics analyses that I conducted using R. The code is available in "RNA-seq analysis" file of the GitHub repository: https://github.com/NdnBnBn/Enrich-Vis. I used the mRNA and miRNA count matrix as input.

4.2.2 Exploratory data analysis

To assess the pairwise similarity between samples, I first computed their Euclidean distances. I applied the variance-stabilising transformation from the *DESeq2* package to stabilise the variance across expression levels. To visualize the transformed data with batch variation removed, I used the *removeBatchEffect* function from the *limma* package. This step is used solely for visualisation, as the variance-stabilising transformation does not account for batch-related variation. I then calculated the Euclidean distances between the mean count values of each sample. To visualise these relationships, I generated a hierarchical clustering heatmap using the *ComplexHeatmap* R package (v1.0.12). The heatmap highlighted sample clusters based on genotype and passage number, and helped identify potential outliers. To further investigate clustering patterns, I performed principal component analysis (PCA) using the *plotPCA* function from the *DESeq2* package.

I also performed hierarchical clustering on the miRNA and mRNA with the highest variability across the samples. This step aimed to identify patterns or groups within the data by clustering mRNA or miRNAs with similar expression profiles. To determine the most variable mRNA or miRNAs, I calculated variance for each mRNA/miRNA across all samples. The top-ranking mRNA or miRNAs based on their variance (2000 mRNA and 200 miRNA) were selected for clustering.

4.2.3 Differential expression analysis

I performed the differential expression analysis on the mRNA and miRNA count matrix using *DESeq2*. I excluded mRNA and miRNA with low expression, i.e., having fewer than 10 counts across all samples. After this filtering step, approximately 30,000 mRNA and 740 miRNAs remained for downstream analysis. The differential expression analysis done by Deseq2 start by normalised raw counts using the method implemented in DESeq2. This method adjusts for differences in sequencing depth and library size across samples. DESeq2 uses a generalised linear model to test for differences in expression between conditions (HGPS vs. WT). It incorporates experimental design information (genotype, passage, batch effects) via a design matrix. Then, DESeq2 uses the Wald test to determine whether observed differences in expression level are statistically significant. The test provides a p-value for each mRNA/miRNA, representing the probability that the observed difference is due to random variation. DESeq2 adjusts the p-value for

multiple hypothesis testing using the Benjamini-Hochberg method. The differences in expression level between condition is given by the log2 fold change (log2FC). The mRNAs and miRNAs with a log2FC >1 (overexpression) or log2FC < -1 (underexpression) and an adjusted p-value <0.01 were considered significantly differentially expressed.

4.2.4 mRNAs functional analysis

I performed over-representation enrichment analysis on significantly differentially expressed mRNAs (DEmR) using the *clusterProfiler* R package. Using the enrichGO and enrichKEGG functions, I identified which biological terms from the GO database and which pathways terms from the KEGG database were overrepresented in the sets of DEmRs. Both function uses a hypergeometric test to assess the probability of observing a certain number of DEmRs associated with a specific term. The test outputs a p-value, that is adjusted for multiple testing (using the Benjamini-Hochberg method). The output is a table including the enriched terms or pathways, the associated adjusted p-value and the associated mRNA (from the DEmRs list). Terms and pathways with an adjusted p-value above 0.05 were considered significantly enriched.

Shiny app for functional analysis To facilitate the exploration and interpretation of the enrichment results, I developed a Shiny application that generates various visualisations of the differentially expressed genes (DEGs) and their enriched terms, sourced from the Gene Ontology (GO) and KEGG pathway databases. The application offers several types of visualisations, including bar plots, dot plots, and network diagrams, to display the most significantly enriched biological terms and pathways. These visualisations help to highlight key biological processes or pathways that may be dysregulated in the disease context.

In addition to these visual tools, the app provides interactive features that allow users to filter and sort the enriched terms, making it easier to focus on the most relevant biological categories. The GitHub repository containing the source code for the Shiny app is available at: https://github.com/NdnBnBn/Enrich-Vis.

4.2.5 miRNA-mRNA Interaction Analysis

MicroRNAs are regulatory sequences that modulate gene expression by binding to complementary regions on mRNAs. With the set of DEmR and DEmiR as input, I used the *miRTarBase* database (Cui et al. 2025) to identify the potential mRNA targets of the DEmiR. Then, I used *Cytoscape*, to visualise the resulting miRNA-mRNA interaction network.

4.2.6 Multi-omics factor analysis

I performed a multi-omic analysis using the multi-omics factor analysis (MOFA) method (Argelaguet et al. 2018). MOFA is a statistical approach designed to integrate and analyse multiple types of omics data, to uncover latent (hidden) factors that can explain the variation observed across different datasets. I used MOFA to integrate both mRNA and miRNA data obtained from HGPS and WT VSMCs.

MOFA works by modelling the data in a way that captures the major sources of variation across multiple omics layers. By uncovering shared latent factors, MOFA can reveal patterns that may not be evident in separate analyses of mRNA or miRNA data separately.

MOFA takes as input the expression matrix of mRNA and miRNAs, transformed using the variance stabilising transformation function. The initial datasets included 30,000 mRNAs and 740 miRNAs. To reduce dimensionality, I retained the 7,000 mRNAs with the highest variance across samples for the analysis. MOFA was configured to use Gaussian likelihood. The number of factors was set to five. This means that MOFA tries to identify five factors that explain the variation across both the mRNA and miRNA data.

4.3 Results

4.3.1 mRNA expression analysis

In the first section of the results, I present the analysis of mRNA expression data obtained from HGPS VSMCs and WT VSMCs using RNA-seq. I analysed a total of 12 RNA-seq datasets:

— 3 datasets for HGPS VSMCs at passage 8 (HGPSP8)

— 3 datasets for HGPS VSMCs at passage 12 (HGPSP12)

— 3 datasets for WT VSMCs at passage 8 (WTP8)

— 3 datasets for WT VSMCs at passage 12 (WTP12)

The Table 4.1 summarises the metadata for each sample. Genotype, passage, and batch are indicated for each sample, along with their respective identifiers.

Sample	Genotype	Passage	Batch
X2021A40	WT1	P8	ONE
X2021A41	WT2	P8	ONE
X2021A92	WT3	P8	TWO
X2021A45	WT1	P12	ONE
X2021A46	WT2	P12	ONE
X2021A93	WT3	P12	TWO
X2021A42	HGPS1	P8	ONE
X2021A43	HGPS2	P8	ONE
X2021A44	HGPS3	P8	ONE
X2021A47	HGPS1	P12	ONE
X2021A48	HGPS2	P12	ONE
X2021A49	HGPS3	P12	ONE

Table 4.1 – mRNA samples metadata

4.3.1.1 Exploratory data analysis

To evaluate the distance and variation between samples, I analysed the mRNA count matrix (after filtering lowly expressed mRNA). After normalising the mRNA counts using the variance-stabilising transformation, I performed hierarchical clustering and PCA (Figure 4.1).

Hierarchical clustering and PCA revealed clear patterns in mRNA expression profiles. Samples clustered primarily by genotype (HGPS versus WT) and secondarily by passage (P8 versus P12), reflecting differences in gene expression driven by both factors. The WT samples showed close clustering patterns, with minimal variation and clear separation by passage (Figure 4.1). In contrast, HGPS samples exhibited greater variability, with some samples failing to cluster strictly by passage. For instance, samples X20211A44 (HGPSP8) and X20211A49 (HGPSP12) clustered more closely with each other than with other HGPS samples, suggesting increased heterogeneity within the HGPS group. These observations indicate that genotype strongly influences mRNA expression patterns, with WT samples showing

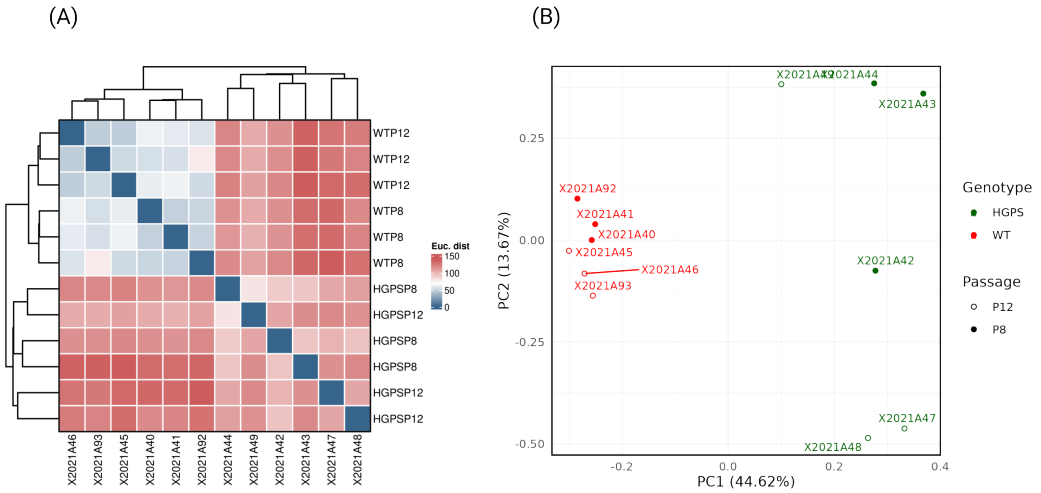


Figure 4.1 – **(A)** Hierarchical clustering of mRNA sequencing data from all experimental samples. Distances between samples were computed using the Euclidean distance of their mean mRNA count values. Rows represent the genotype (WT or HGPS) and passage (P12 or P8) of each sample, while columns correspond to sample identifiers. **(B)** PCA of mRNA sequencing data from all experimental samples after VST-transformation of DESeq2-normalised expression data. The X-axis represents the first principal component, and the Y-axis represents the second. Samples are coloured by genotype and passage, and labels indicate sample identifiers. Both analyses (clustering and PCA) accounted for batch effects, as samples X2021A92 (WTP8) and X2021A93 (WTP12) were part of a separate experimental batch (Table 4.1).

consistent passage-based clustering, while HGPS samples exhibit additional sources of variability.

To further investigate this variability, I performed hierarchical clustering of the 2,000 most variable mRNAs across all samples. Variability was assessed based on standard deviation of counts after variance-stabilising transformation. This analysis confirmed the earlier patterns: WT samples clustered distinctly by passage, while most HGPS samples clustered by genotype but exhibited less consistent separation by passage. Two HGPS samples (X20211A44 and X20211A49) remained distinct from the rest, further emphasising the higher variability in the HGPS group (Figure 4.2).

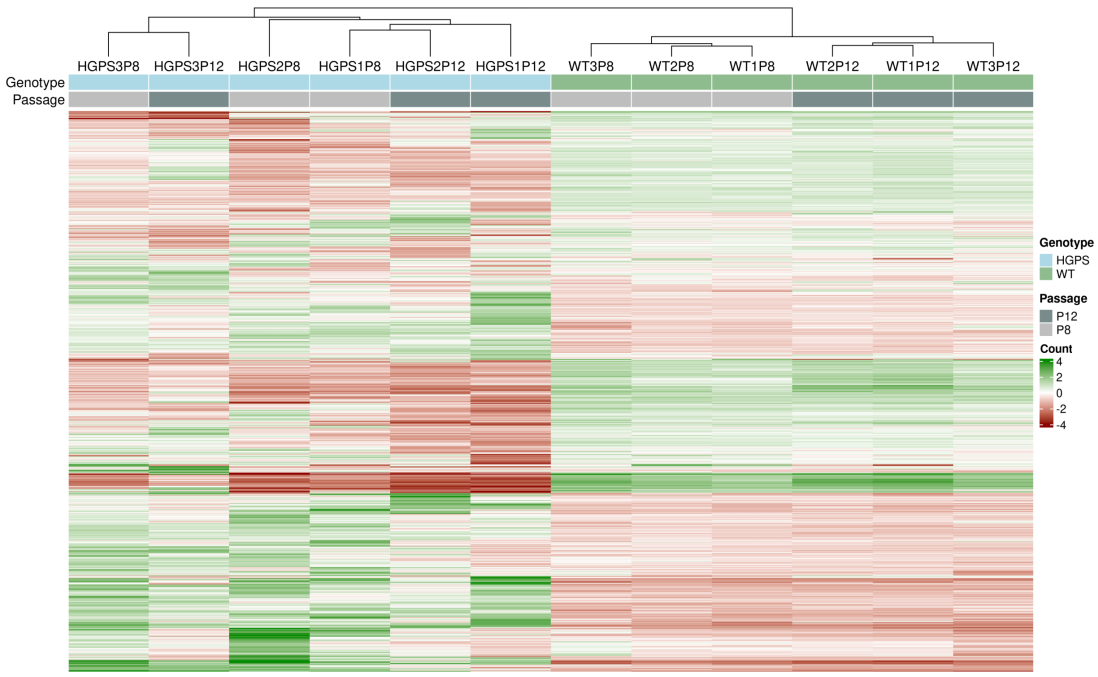


Figure 4.2 – **Hierarchical clustering of the top 2,000 mRNAs with the highest count variability across samples**. The heatmap displays clustering of the top 2,000 mRNAs with the highest count variability. Distances were computed using Euclidean distance. Each row represents an mRNA, and each column represents a sample.

4.3.1.2 Differential mRNA expression analysis with DESeq2

To identify mRNAs with altered expression levels across experimental conditions, I performed differential expression analysis using DESeq2. The analysis accounted for observed batch effects by incorporating them into the experimental design matrix. I conducted five distinct comparisons:

— Comparison I: HGPS versus WT (genotype effect, independent of passage)
— Comparison II: HGPSP8 versus WTP8 (genotype effect at passage 8)
— Comparison III: HGPSP12 versus WTP12 (genotype effect at passage 12)
— Comparison IV: HGPSP12 versus HGPSP8 (effect of passage within HGPS)
— Comparison V: WTP12 versus WTP8 (effect of passage within WT)

For all comparisons, I considered mRNAs with a $\log_2$ fold change (log2FC) greater than 1 or less than -1, and an adjusted p-value <0.01, as significantly differentially expressed.

Table 4.2 summarises the number of differentially expressed mRNAs (DEmRs) identified in each comparison. In Comparison I (HGPS versus WT), 1,459 DEmRs were detected, with a predominance of upregulated mRNAs (983 upregulated and 476 downregulated). For Comparisons II and III, which examined genotype effects at specific passages, I identified 1,089 DEmRs for HGPSP8 versus WTP8 and 777 DEmRs for HGPSP12 versus WTP12. Notably, the number of DEmRs decreased at later passages, suggesting reduced genotype-specific differences as cells aged. In both passage-specific comparisons, upregulated mRNAs outnumbered downregulated ones.

Comparison	Total DEmRs	Downregulated	Upregulated
HGPS vs WT	1,459	476	983
HGPSP8 vs WTP8	1,089	339	750
HGPSP12 vs WTP12	777	284	493
HGPSP12 vs HGPSP8	117	99	18
WTP12 vs WTP8	13	2	11

Table 4.2 – **Number of differentially expressed mRNAs identified in each comparison**.

In Comparisons IV and V, I observed far fewer DEmRs. A total of 117 mRNAs were differentially expressed between HGPSP12 and HGPSP8, with 99 being downregulated and 18 upregulated. Only 13 DEmRs were identified between WTP12 and WTP8, suggesting that cellular ageing has a more pronounced effect on mRNA expression in HGPS cells compared to WT cells.

Using an UpSet plot, I determined intersections of DEmRs across comparisons were visualised (Figure 4.3). The plot reveals a substantial number of shared DEmRs (487) between Comparisons I, II, and III. These shared DEmRs indicate consistent genotype-driven expression changes across passages. Additionally, unique DEmRs were identified in each comparison, highlighting condition-specific expression changes.

For each intersection set, I evaluated the direction of expression changes (upregulated, downregulated, or both). Most shared DEmRs showed consistent regulation across comparisons. However, in intersection sets containing DEmRs from both genotype and passage comparisons, some mRNAs exhibited opposite regulation (upregulated in one comparison and downregulated in another).

To gain further insights into the biological functions and pathways associated

with the identified DEmRs, I performed functional enrichment analyses, detailed in the next section.

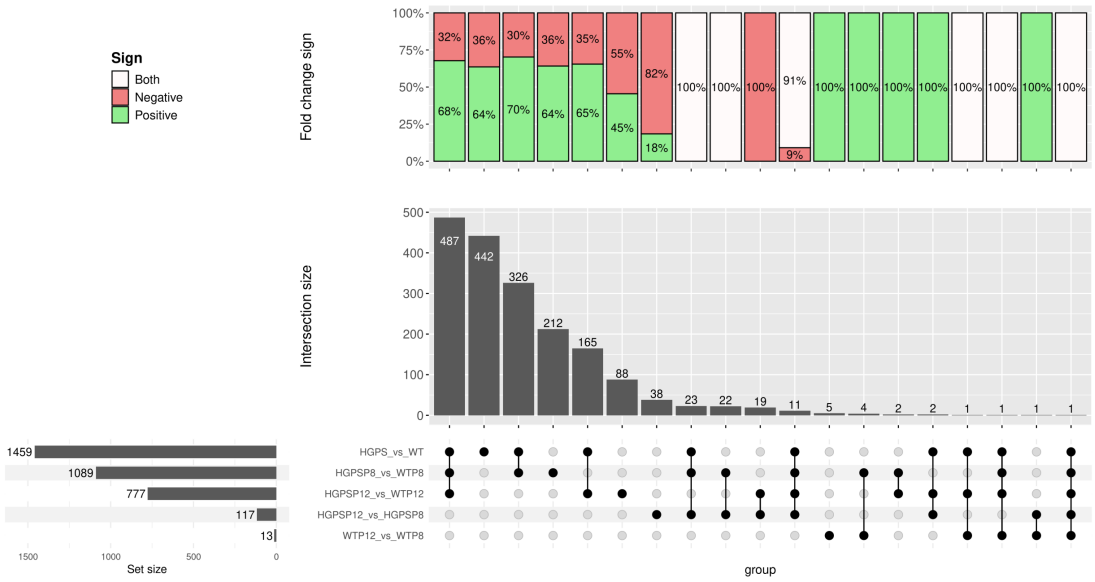


Figure 4.3 – **UpSet plot of the DEmRs identified across all comparisons**. The bottom panel of the plot represents intersections of DEmRs across different comparisons (HGPS vs WT, HGPSP8 vs WTP8, HGPSP12 vs WTP12, HGPSP12 vs HGPSP8, WTP12 vs WTP8). Rows correspond to datasets, and columns represent shared genes across comparisons. Set sizes (total DEmRs) and intersection sizes (shared genes) are illustrated as bar charts.The top panel indicates the sign of the log2FC for each mRNA within intersection sets: negative for downregulated, positive for upregulated, and "both" for mRNAs showing opposite regulation in different comparisons.

4.3.1.3 Functional analysis of the differentially expressed mRNAs

To better understand the biological relevance of the DEmRs, I performed over-representation analyses using terms from the GO biological processes database and KEGG pathways database. I performed the analyses on the DEmR sets identified in each comparison. I divided the sets of DEmRs into three categories: the complete set of DEmRs, upregulated DEmRs, and downregulated DEmRs. Terms and pathways with an adjusted p-value < 0.05 were considered significantly enriched. For each analysis, I selected the top 10 most significant results based on their p-values.

Overall genotype comparison Figure 4.4 shows the results of the over-representation analysis for DEmRs from Comparison I (HGPS vs WT). The functional analysis revealed that GO biological processes were predominantly associated with viral infection. KEGG pathway analysis showed enrichment in pathways related to viral infection, cardiomyopathies, various signalling pathways, and pathways involving the cytoskeleton and extracellular matrix.

Separate analyses of upregulated and downregulated DEmRs provided additional insights. Downregulated DEmRs were strongly enriched in GO terms and KEGG pathways related to viral processes, suggesting a potential suppression of viral response pathways in HGPS cells. Conversely, upregulated DEmRs were enriched in muscle development, tissue formation, and cardiomyopathies, reflecting the genotype-specific effects on muscle and cardiac-related functions.

Passage-specific comparisons For the passage-specific comparison at passage 8 (Comparison II: HGPSP8 vs WTP8), the results were consistent with those observed in the overall genotype comparison (Figure 4.5). Downregulated DEmRs were enriched in viral processes, while upregulated DEmRs were associated with muscle development and cardiomyopathies.

At passage 12 (Comparison III: HGPSP12 vs WTP12), a similar enrichment pattern was observed, but KEGG pathways related to viral processes appeared to be enriched in the upregulated DEmRs (Figure 4.6). This shift suggests that as passage progresses, the disruption of viral-related pathways intensifies in HGPS cells, potentially reflecting ageing-associated regulatory changes.

Passage-dependent comparisons within genotypes Finally, I performed over-representation analysis on the DEmRs from Comparisons IV (HGPSP12 vs HGPSP12) and V (WTP12 vs WTP8). Significant results were only obtained from Comparison IV. The DEmRs were associated with cytoskeletal organisation and extracellular matrix regulation, indicating a structural impact of cellular ageing (Figure 4.7). I obtained similar results for the downregulated mRNAs. No significant enrichment was observed for upregulated mRNAs, likely due to the small number of upregulated genes identified in this comparison.

These over-representation analyses reveal distinct functional roles for upregulated and downregulated DEmRs in HGPS versus WT VSMCs and highlight passage-dependent variations in mRNA expression. The results suggest that cellular ageing

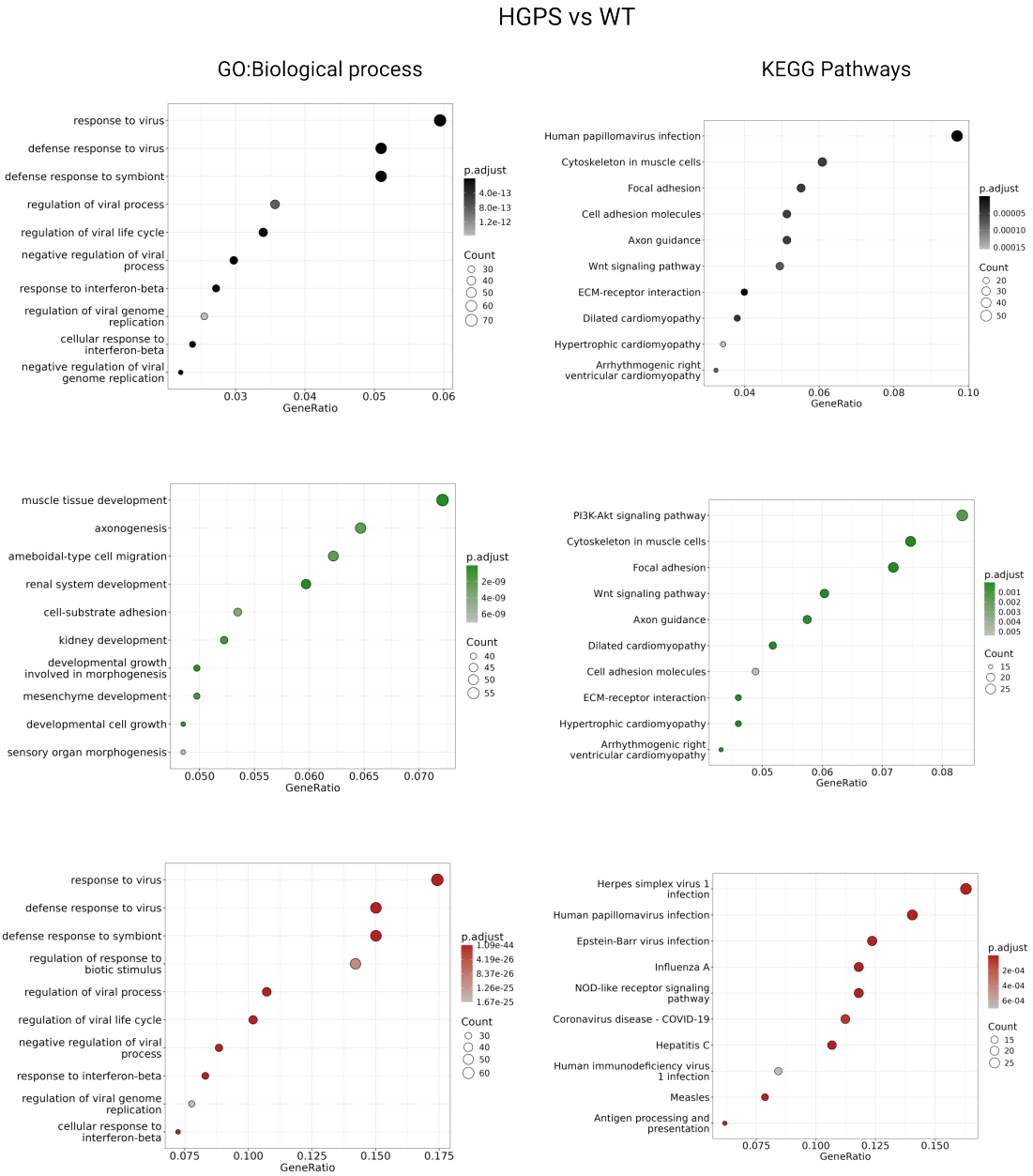


Figure 4.4 – **Top 10 enriched biological processes and pathways in DEmRs HGPS vs WT.** Each plot represents the 10 most significant enriched terms in either the GO biological processes (left) or KEGG pathways (right). The gene ratio indicates the proportion of genes in the term relative to the total number of DEmRs tested. Plots with black dots show the enrichment results for all DEmRs, green dots for upregulated DEmRs, and red dots for downregulated DEmRs.

exerts a significant impact on mRNA regulation in HGPS cells, particularly in processes tied to viral infection, muscle development, and cytoskeletal integrity.

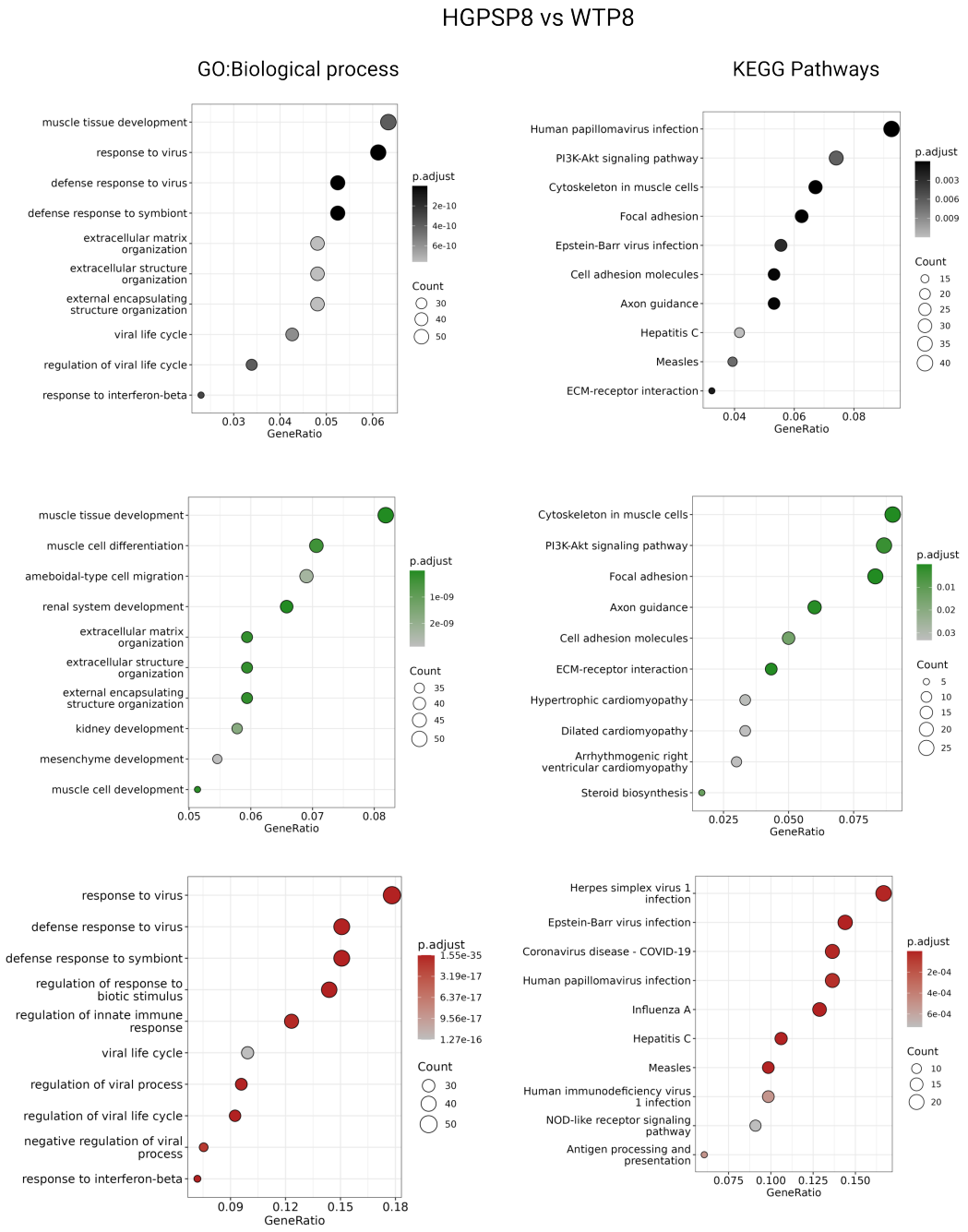


Figure 4.5 – **Top 10 enriched biological processes and pathways in DEmRs HGPS vs WT at passage 8.**

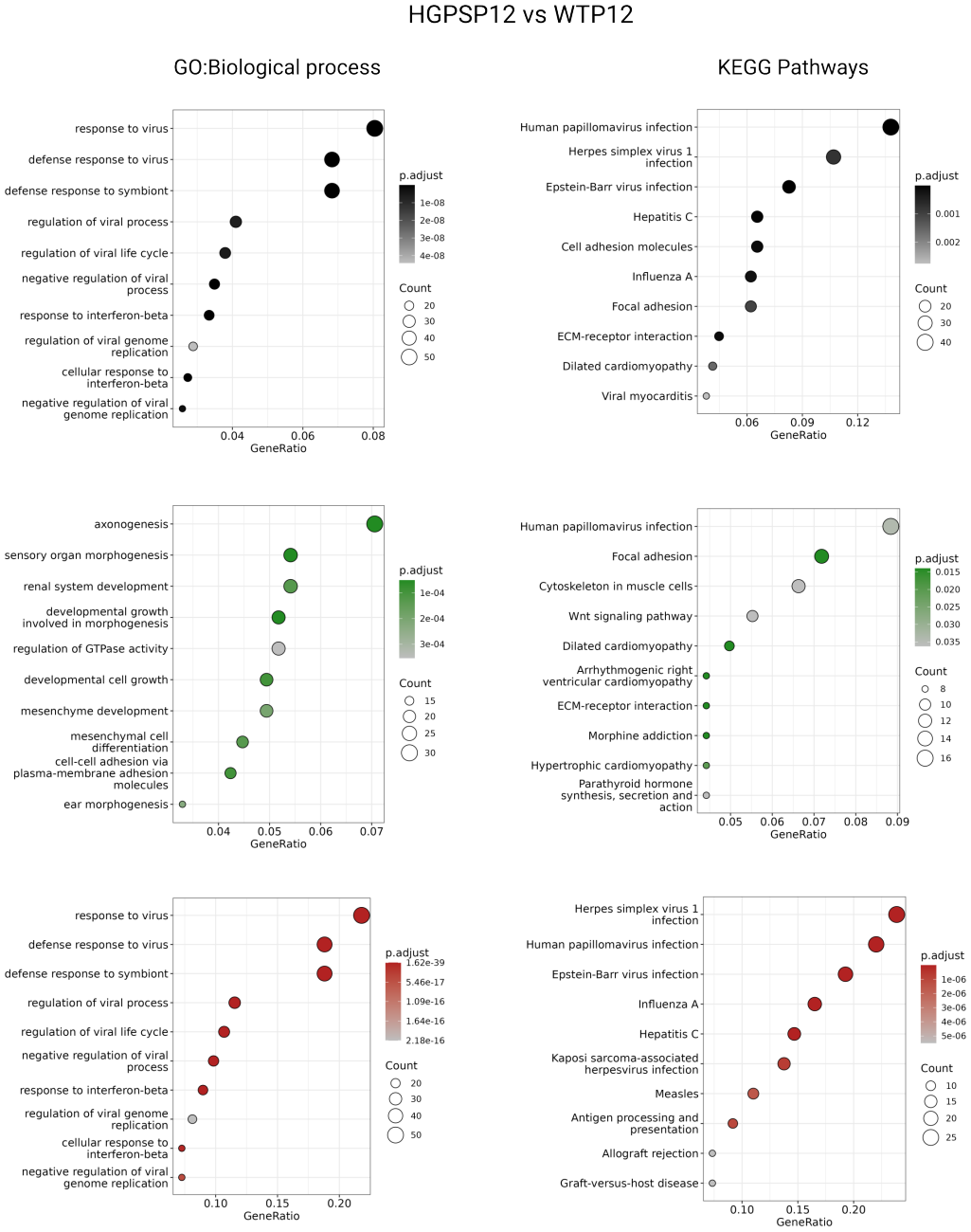


Figure 4.6 – **Top 10 enriched biological processes and pathways in DEGs HGPS vs WT at passage 12.**

4.3.2 miRNA expression analysis

In the second section of the results, I present the analysis of miRNA expression data. A total of 10 RNA-seq datasets were analysed:

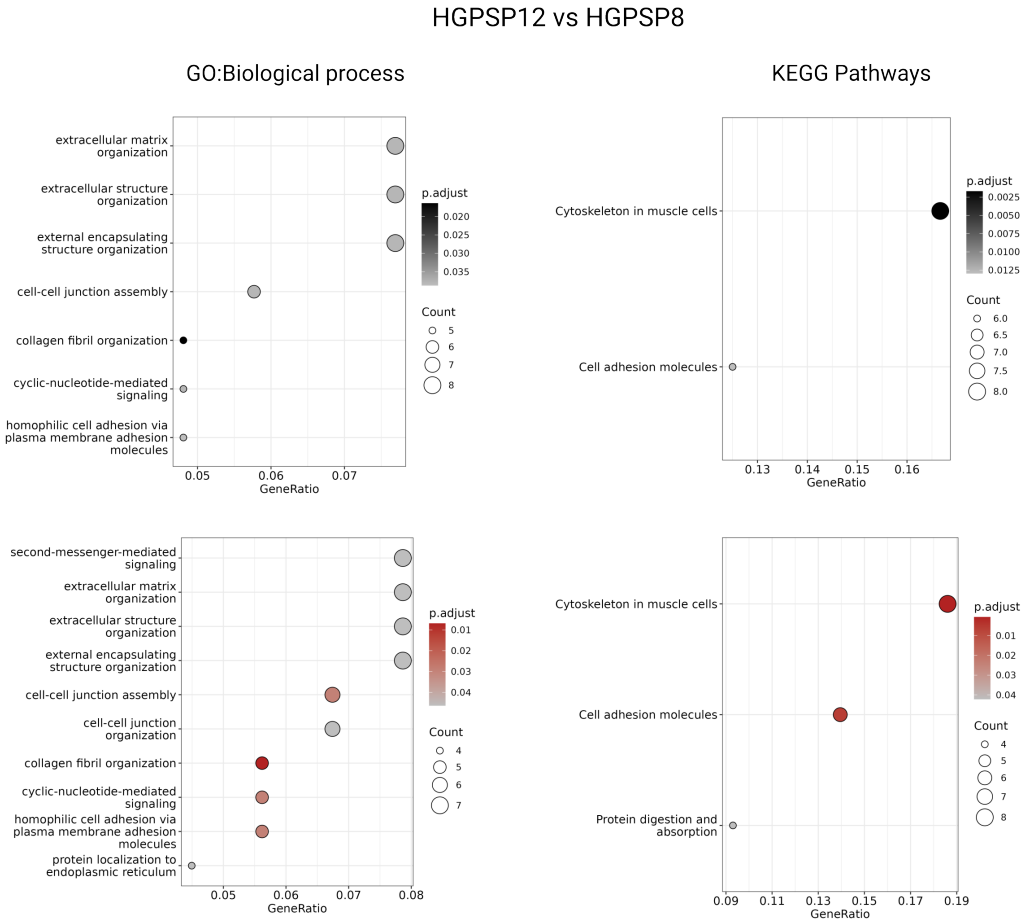


Figure 4.7 – **Top 10 enriched biological processes and pathways in DEGs for passage 12 vs passage 8 in HGPS.** No significant enrichment was observed for upregulated mRNAs.

— 3 datasets for HGPS VSMCs at passage 8 (HGPSP8)
— 3 datasets for HGPS VSMCs at passage 12 (HGPSP12)
— 2 datasets for WT VSMCs at passage 8 (WTP8)
— 2 datasets for WT VSMCs at passage 12 (WTP12)

Two WT samples were discarded due to contamination. Table 4.3 summarises the metadata for each sample, indicating the genotype, passage, and sample identifiers.

4.3.2.1 Data quality assessment

To assess the quality of the miRNA expression data, I applied a similar approach as for the mRNA analysis. The count matrix was transformed using the variance stabilising transformation to perform hierarchical clustering and PCA.

Sample	Genotype	Passage
X2021A55	HGPS1	P8
X2021A56	HGPS2	P8
X2021A57	HGPS3	P8
X2021A60	HGPS1	P12
X2021A61	HGPS2	P12
X2021A62	HGPS3	P12
X2021A53	WT1	P8
X2021A54	WT2	P8
X2021A58	WT1	P12
X2021A59	WT2	P12

Table 4.3 – Metadata for miRNA samples used in the analysis, detailing genotype and passage for each dataset.

Hierarchical clustering revealed that WT samples clustered together, with a clear separation by passage. However, the clustering pattern for the HGPS samples was unexpected. One HGPS sample, HGPS3P12, clustered more closely with WT samples than with the other HGPS samples. Among the remaining five HGPS samples, four clustered by passage, while one sample (HGPS2P8) clustered separately (Figure 4.8-A). The PCA plot provided a clearer distinction between WT and HGPS samples. However, the sample HGPS3P12 appeared to deviate from its expected clusters (Figure 4.8-B). This deviation is possibly due to the biological variability in the HGPS samples.

To further investigate this variability, I clustered the top 200 miRNAs with the highest count variance across samples. This analysis yielded similar results, highlighting the biological heterogeneity in the HGPS samples (Figure 4.9).

4.3.2.2 Differential miRNA expression analysis

I performed differential miRNA expression analysis using the DESeq2 R package, focusing on the same comparisons as those used for mRNA expression analysis. I considered miRNAs with log fold changes greater than 1 or less than -1 and an adjusted p-value of <0.01 as differentially expressed.

The results are summarised in Table 4.4. I identified the highest number of differentially expressed miRNAs (DEmiRs) in the HGPS versus WT comparison, with 148 DEmiRs, of which 121 were upregulated and 27 were downregulated. The passage-specific comparisons revealed fewer DEmiRs. In the HGPSP12 versus

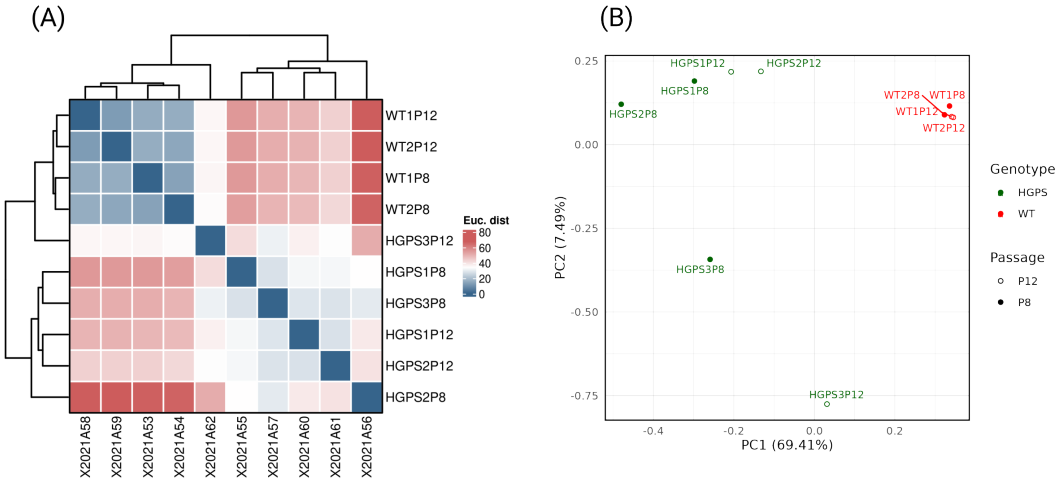


Figure 4.8 – **miRNA data exploration (A)** Hierarchical clustering of mRNA sequencing data from all experimental samples. The heatmap shows clustering of WT samples by passage and higher biological variability among HGPS samples. **(B)** PCA plot.

Test	Total	Downregulated	Upregulated
HGPS vs WT	148	27	121
HGPSP8 vs WTP8	105	14	91
HGPSP12 vs WTP12	69	11	58
HGPSP12 vs HGPSP8	5	4	1
WTP12 vs WTP8	0	0	0

Table 4.4 – Number of differentially expressed miRNAs identified in each comparison.

HGPSP8 comparison, I identified only five DEmiRs, with four downregulated and one upregulated. I found no significant DEmiRs in the WTP12 versus WTP8 comparison.

In Figure 4.10, I present the UpSet plot, which illustrates the intersections of DEmiRs across the different comparisons. The plot shows that most of the DEmiRs identified in the HGPS versus WT comparison were also found in the passage-specific comparisons.

These results show the miRNA dysregulation in HGPS compared to WT. The lack of significant DEmiRs in WTP12 versus WTP8 suggests a stable miRNA profile in WT samples over time.

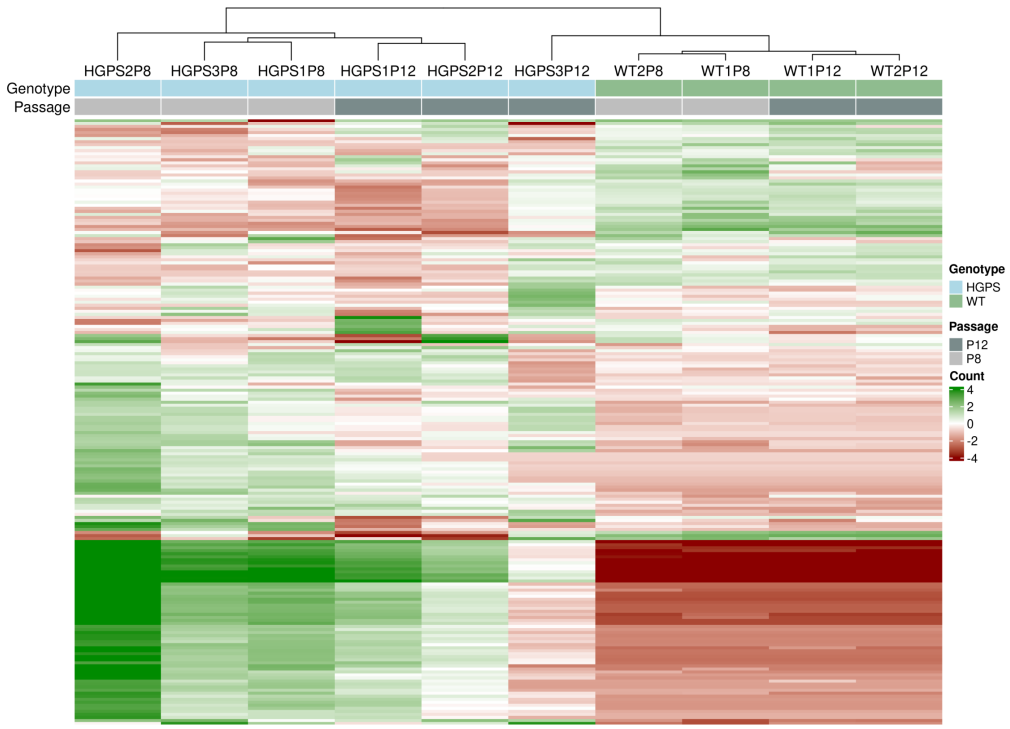


Figure 4.9 – **Hierarchical clustering of the top 200 miRNAs with the highest count variability across samples**.

4.3.3 miRNA-mRNA joint analysis

4.3.3.1 Interaction network

miRNAs are regulatory RNAs that bind to mRNA sequence in order to modulate gene expression (in general inhibits gene expression). Using the lists of DEmiRs and DEmRs obtained from the differential expression analysis, I identified DEmRs that are targeted by the DEmiRs identified in all comparisons. To do so, I used information retrieved in miRTarBase, a database cataloguing miRNA target genes.

Among all the conditions analysed, I identified two miRNA-mRNA interaction networks. These networks were built from the DEmiRs and DEmRs in the comparisons HGPS vs WT and HGPP8 vs WTP8. I did not identify any interaction between DEmiRs and DEmRs obtain from other comparisons. The identified networks were visualised and constructed using Cytoscape, enabling a clear representation of the interactions. The resulting networks are shown in Figures 4.11-A and 4.12-A.

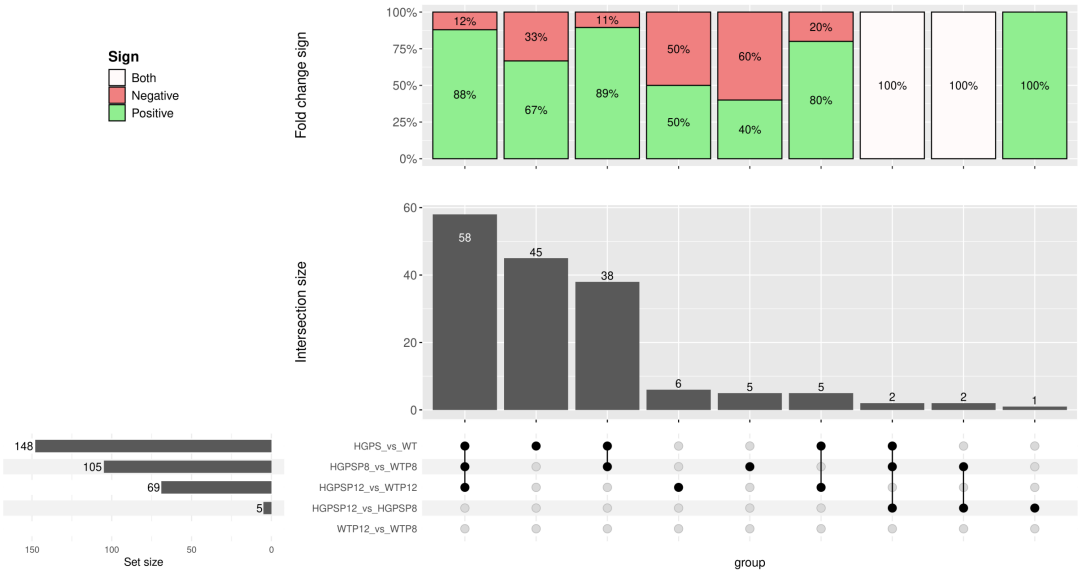


Figure 4.10 – **UpSet plot of the differentially expressed miRNA obtained in all comparison.** The comparison WTP12 versus WTP8 did not yield any DEmiR

For both networks, I used the set of mRNA of the network and performed an over-representation analysis of the GO biological process term. The mRNA in network Figure 4.11-A are linked to cardiac cell development and differentiation and osteoblast differentiation (Figure 4.11-B).

The over-representation analysis of the mRNA of the second network gave similar result (Figure 4.12-B).

4.3.3.2 Multi-omics factor analysis

In addition to performing differential expression analysis, I conducted a Multi-Omics Factor Analysis (MOFA) to jointly analyse miRNA and mRNA expression data. The primary objectives of this analysis were to: (1) identify groups of miRNAs and mRNAs that vary together, and (2) uncover factors that explain the variability across samples.

I used the miRNA and mRNA count matrices as the inputs. The mRNA count matrix contains data from 12 samples, while the miRNA matrix contains data from 10 samples. To match the number of samples between datasets, I removed two WT samples from the mRNA dataset. I removed the samples that were part of the

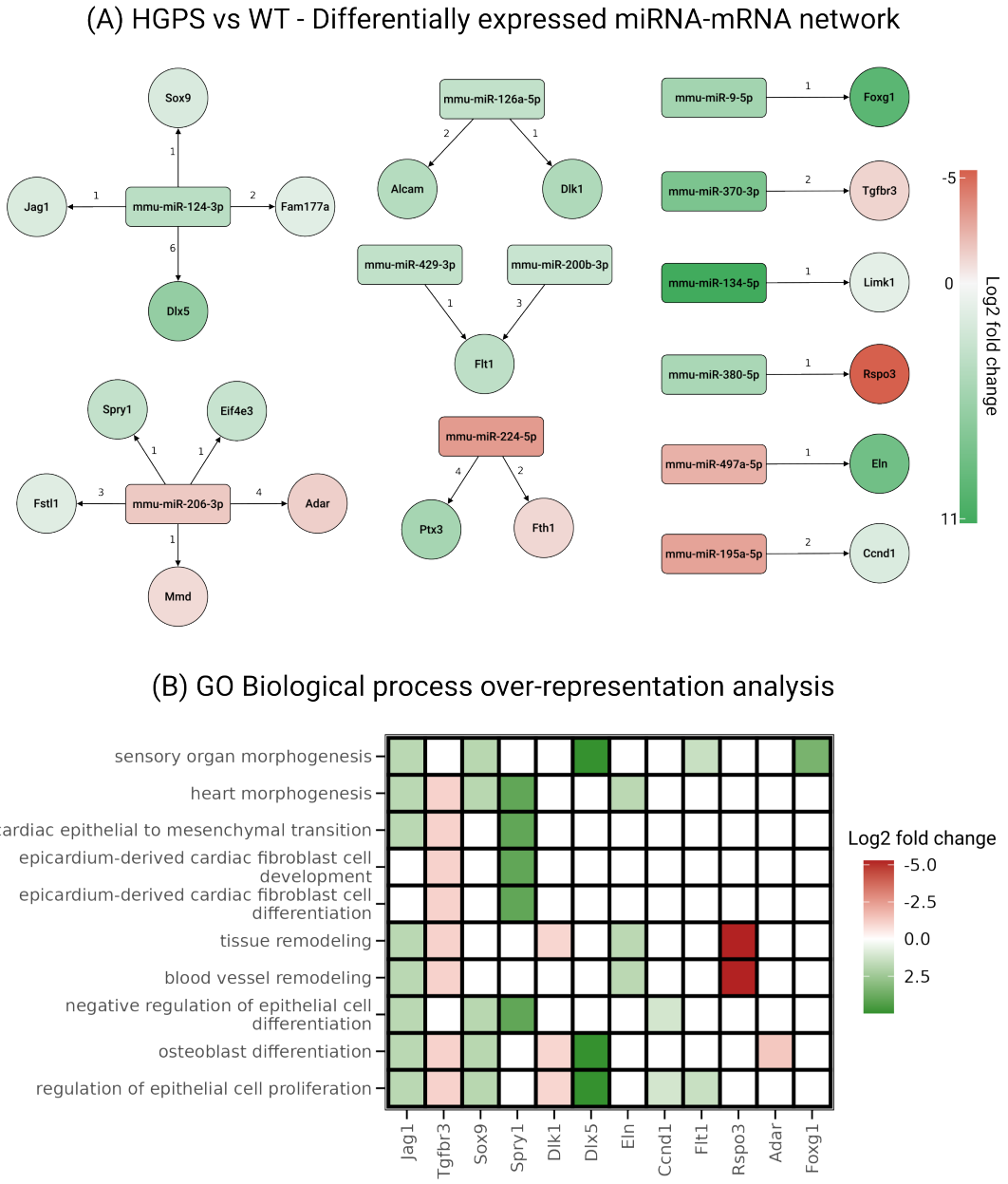


Figure 4.11 – **miRNA-mRNA interaction network for HGPS vs WT. (A)** The network represents interactions between the DEmiR and DEmR identified in the HGPS vs WT comparison. Round nodes are mRNAs and square ones are miRNAs.**(B)** Heatmap of the over-representation analysis done on the mRNA from the network in the panel A. In both panels, the colour gradient correspond to the log2FC obtained in the differential expression analysis.

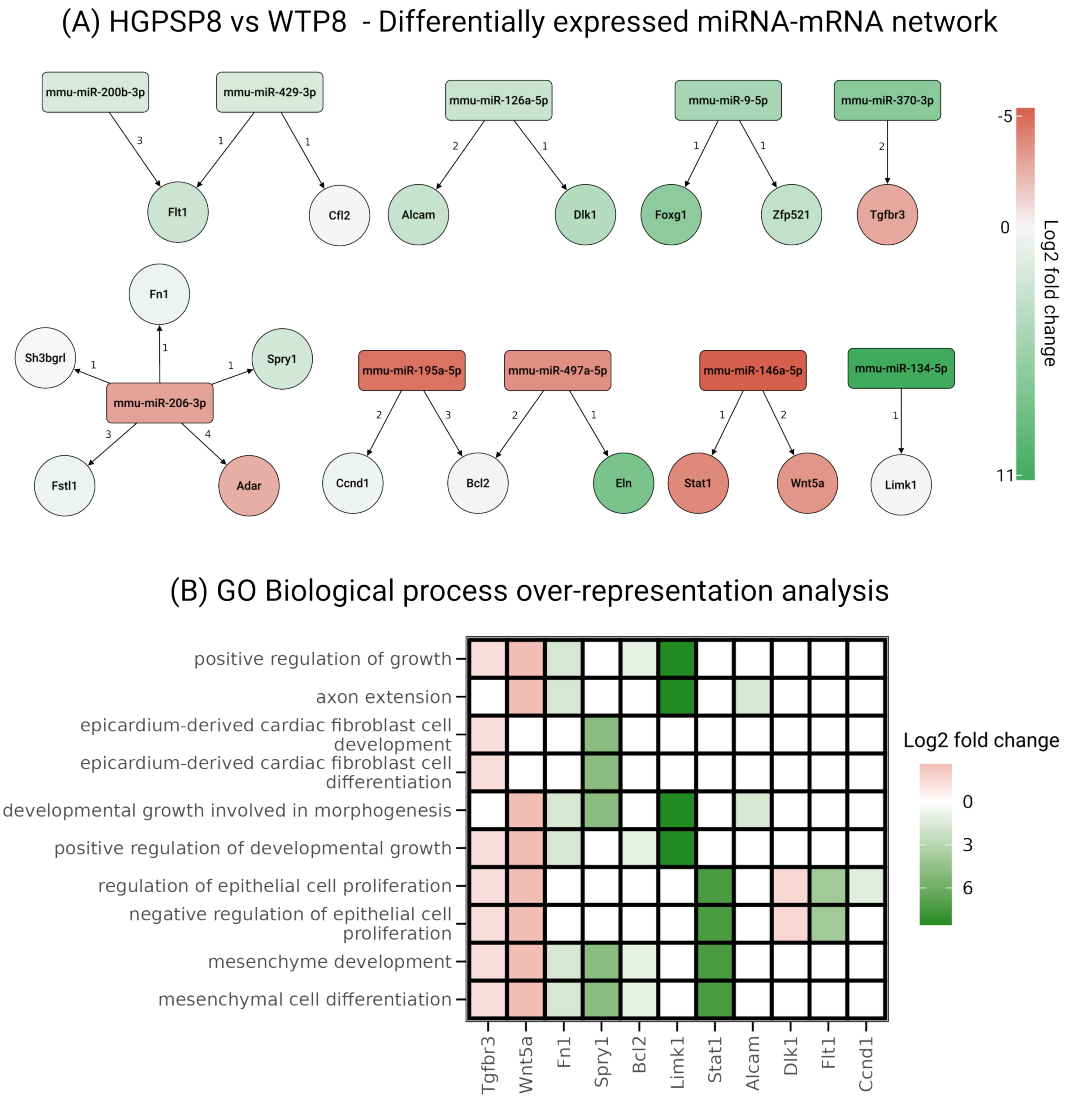


Figure 4.12 – **miRNA-mRNA interaction network for HGPSP8 vs WTP8. (A)** The network represents interactions between the DEmiR and DEmR identified in the HGPSP8 vs WTP8 comparison. **(B)** Heatmap of the over-representation analysis done on the mRNA from the network in the panel A.

experimental second batch (see metadata in Table 4.1).

The initial datasets consisted of approximately 30,000 mRNAs and 740 miRNAs. The large disparity in the number of features between these datasets exceeded the recommendations in the MOFA user guidelines. Thus, I reduced the mRNA dataset to the 7,000 mRNAs with the highest variability across samples.

The factor 1 explained the largest proportion of variability in both datasets

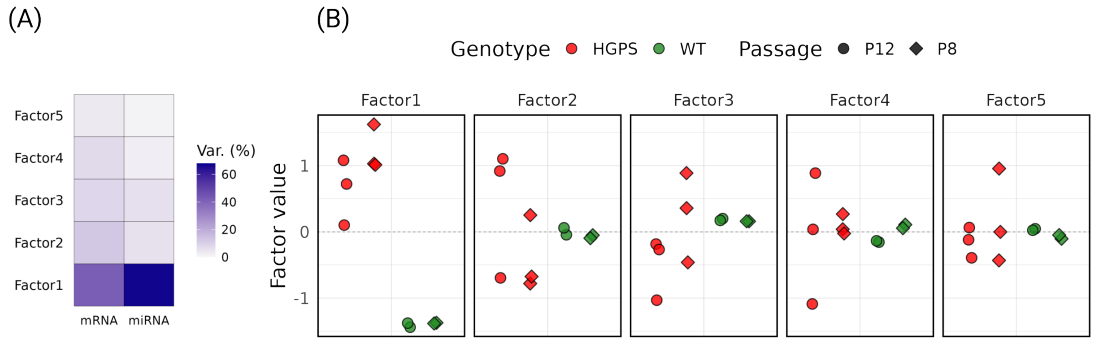


Figure 4.13 – **MOFA variance decomposition result (A)** Variance explained by MOFA factors across omics datasets. Factor 1 explains the largest proportion of variability and separates samples by genotype. **(B)** Factors value per sample.

(Figure 4.13-A) and separated samples by genotype, demonstrating its biological significance. The factor 2 seem to separate 4 of the 6 HGPS samples by passage, but not the WT sample. The remaining three factors did not reveal any notable patterns (Figure 4.13-B). Thus, I focused the remainder of the analysis on the factor 1.

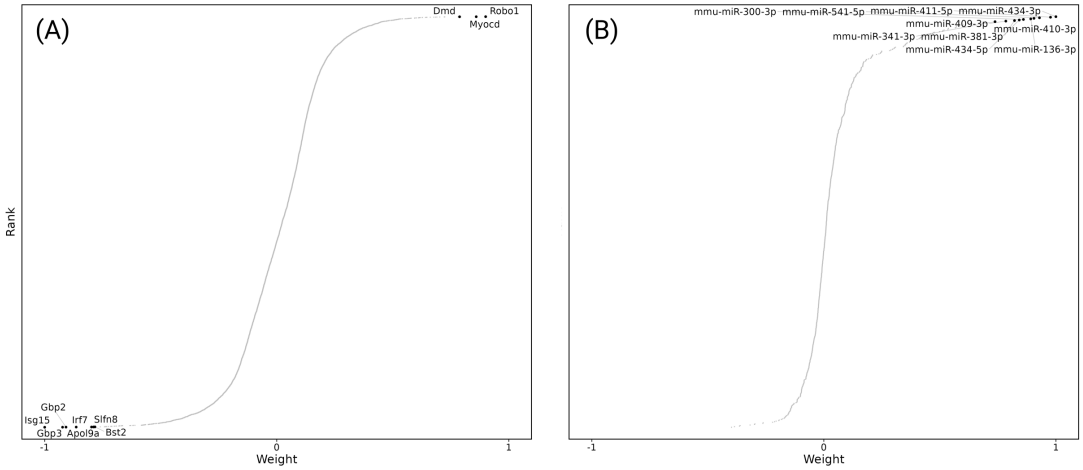


Figure 4.14 – **Distribution of mRNA (A) and miRNA (B) weights along Factor 1**.

The Figure 4.14 shows the distribution of mRNA and miRNA weights along Factor 1, highlighting the top 10 features (i.e., mRNA and miRNA names) with the highest absolute weights. These features represent the primary drivers of variability within each omics dataset. We can see that in the mRNA data, feature with both negative and positive weights seem to contribute equally samples variability (Figure

4.14-A). In contrast, the feature withe positive weights in the miRNA dataset are the main driver (Figure 4.14-B).

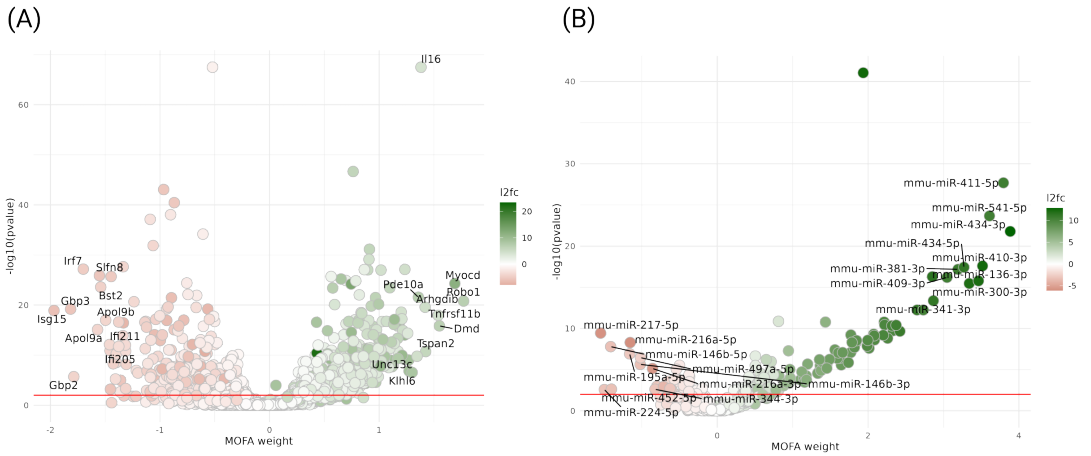


Figure 4.15 – **Integration of MOFA weights and differential expression analysis.** MOFA weights for mRNAs **(A)** and miRNAs **(B)** along Factor 1 are plotted against their differential expression p-values (HGPS vs WT). Dots are coloured by log2FC, and miRNA and mRNA with the highest weights are labelled.

As detailed above, I used differential expression analysis to identify which of mRNAs and miRAs are differentially expressed between samples, and the amplitude of expression change. MOFA provides complementary insights by revealing the main drivers of variability across datasets. I plotted the MOFA weights against the p-values obtained from the differential expression analysis of HGPS versus WT (Figure 4.15). Each data point is coloured by its log2FC, and the top 10 genes (mRNAs in Figure 4.15-A, and miRNAs in Figure 4.15-B) with the highest and lowest weights are labelled. The combination of MOFA weights and differential expression analysis provides complementary insights, facilitating the selection of mRNAs and miRNAs that contribute to the observed differences between samples.

4.4 Discussion and conclusion

The results of this study reveal differences in the mRNA and miRNA expression profiles between HGPS and WT VSMCs at different culture passages.

In the HGPS versus WT comparisons (HGPS vs WT, HGPSP8 vs WTP8, HGPSP12 vs WTP12), downregulated mRNAs were associated with functions

related to viral response, while upregulated mRNAs were linked to muscle cell development and differentiation, as well as pathways involving the extracellular matrix and cardiomyopathy.

Interestingly, functions known to be disrupted in HGPS (including ageing, senescence, apoptosis, DNA repair, and ossification) (Hennekam 2006) did not appear prominently affected. Upon discussion with Patrice Roll, we concluded that the observed gene expression changes might be influenced by the infection history of the mice from which the VSMCs were derived. Viral infections are known to significantly impact gene expression (Desfarges et al. 2012; Dexheimer et al. 2023; Dong et al. 2024). Additionally, it was shown recently that viral infection in HGPS mice result in a decline in the immune response compared to young and old WT mice (Haoyu et al. 2024). Therefore, the changes in gene expression in HGPS VSMCs compared to WT VMSCs could result from the combined effect of viral infection could and Progerin accumulation. Indeed, VSMCs are inherently plastic and highly responsive to environmental changes. Under normal conditions, VSMCs function primarily as contractile cells, maintaining vascular tone and regulating blood flow. However, upon physical or chemical injury, VSMCs can transition from a contractile to a synthetic phenotype. Synthetic VSMCs lose contractility but gain proliferative and migratory abilities, contributing to vascular repair. This phenotypic shift also enables further differentiation into osteogenic, macrophage-like, mesenchymal-like, and fibroblast-like phenotypes (Cao et al. 2022). Both viral infections and Progerin can trigger these phenotypic changes (Coll-Bonfill et al. 2023; Gopal et al. 2020). The combination of Progerin accumulation and viral infection hinders on the process.

The comparison of HGPSP12 versus HGPSP8 provides particularly valuable information. This comparison, which controls for genotype and infection status, highlights genes and miRNAs associated with accelerated ageing in HGPS. The enrichment terms identified in this analysis pertain to cytoskeletal organisation and extracellular matrix regulation, both of which are significantly impacted in HGPS. Similar enrichment results were found when considering only downregulated mRNAs, aligning more closely with functions known to be disrupted in HGPS. These findings emphasise the structural and functional vulnerabilities in HGPS VSMCs during cellular ageing. The fact that few (only 5) differentially expressed mRNAs were found in WTP12 versus WTP8 further underscores the relevance of the observed differences in the HGPS genotype.

I used the lists of DEmiRs and DEmRs to build networks of miRNA-mRNA. Only DEmiRs and DEmRs from the comparisons HGPS vs WT and HGPSP8 vs WTP8 showed possible miRNA-mRNA interactions. The resulting networks include mRNAs related to biological processes such as cardiac tissue remodelling and development, as well as osteoblast and mesenchymal differentiation. These networks show the potential impact of certain miRNAs on mRNA expression patterns.

I completed my analysis with a joint analysis of both omics dataset using MOFA. For each omics, I was able to identify the drivers (mRNAs and miRNAs) of the variability between genotype (HGPS or WT).

This study demonstrates how integrative bioinformatics approaches can help to understand the molecular perturbation underlying HGPS. By leveraging mRNA and miRNA sequencing data, combined with network and multi-omics factor analysis (MOFA). I was able to identify key expression patterns, enriched pathways, and regulatory interactions that distinguish HGPS VSMCs from WT VSMCs, in the context of viral infection.

5 Conclusion

As discussed in the introduction (Chapter 1), biological systems are inherently complex, and various approaches have been developed to facilitate their study. These approaches are generally divided into two categories: top-down and bottom-up (Kohl et al. 2010; Shahzad et al. 2012; Tavassoly et al. 2018). Top-down approaches rely on large-scale datasets — such as RNA-seq data — to infer system-level properties and identify components associated with these properties. In contrast, bottom-up approaches rely on detailed mechanistic knowledge on individual components to build predictive models. My thesis presents two projects that contribute to either a bottom-up (i.e., Boolean model revision) or a top-down system approach (i.e., RNA-seq data analysis).

The main work of my thesis was the development of a method called MRBM for the revision of Boolean models. In parallel, I collaborated with Patrice Roll and his team on the analysis of an RNA-seq dataset. In the following sections, I discuss both works, outline key findings and limitations, and explore future directions. I conclude with reflections on the complementary roles of omics data analysis and logical modelling in systems biology.

5.1 MRBM: A novel revision method

Logical modelling is an iterative process in which models are continuously refined as new data and findings emerge. Model revision — typically performed after the model fails to meet the validation criterion — involves modifying the model so it aligns more closely with experimental data and established biological knowledge (i.e., attractors reflecting the correct phenotype and reachabilities and trajectories coherent with biological event as detailed in Chapter 1.2.3 and Chapter 2).

While conceptually straightforward, model revision can be as challenging as model construction. Discrepancies between the model and biological observations

may stem from missing regulations, incorrect regulatory signs, or the need for more nuanced activity levels for some components (i.e., multivalued components). Each potential correction must be carefully evaluated, as addressing one discrepancy might disrupt another validated behaviour.

When done manually — which is often the case — model revision is a time-consuming and iterative process. Only a few tools have been developed to automatise model revision, including: ModRev (Gouveia et al. 2020) and ARBoLoM (Aleixo et al. 2023). Both tools revise Boolean models primarily by modifying regulatory interactions — such as adding or removing regulations, changing activation to inhibition (or vice versa) — or alter logical functions (e.g., replacing "OR" with "AND").

MRBM takes a different approach (see Chapter 3). It guides the revision of Boolean models by determining which nodes must be multivalued to recover missing attractor reachabilities. MRBM relies on the analysis of the partial most permissive dynamics of the Boolean model. Once we have identified the components to be multivalued, we propose two strategies for parameterising the model:

— Exhaustive strategy: Systematically tests all possible parameterisations to determine which satisfy the desired property.

— Targeted strategy: Analyses relevant paths within the selected partial most permissive dynamics to infer the appropriate parameterisation.

The resulting model is a multivalued refinement, in which neither the logical rules nor the sign of existing regulatory interactions are altered. The core dynamical properties (existing reachabilities and attractors) of the original Boolean model are preserved.

However, as discussed in Chapter 3, MRBM's main limitation is its combinatorial complexity — especially when many components require multivalued representation. Moreover, MRBM propose only one revision strategy, that is, make some component multivalued.

Thus, if no suitable refinement is found using MRBM, this may indicate either a fundamental gap in the current understanding of the system or the need to consider alternative revision strategies that adjust interactions or logical functions. In this regard, MRBM can be used in combination with tools like ModRev and ARBoLoM.

5.2 RNA sequencing for the study of HGPS

The initial aim of this study was to investigate the mechanisms underlying VSMC alterations and loss in HGPS using RNA-seq.

First, the individual analyses of mRNA and miRNA datasets showed different expression profiles between HGPS and WT VSMCs. Hierarchical clustering and PCA consistently differentiated the two groups, while differential expression analysis identified specific mRNAs and miRNAs that were significantly up- or down-regulated in HGPS. Functional enrichment analysis of these deregulated mRNAs further highlighted alterations in pathways related to immune responses, cardiomyopathies, and cytoskeletal organisation.

Next, I constructed networks linking differentially expressed miRNAs to their potential mRNA targets. These networks provided candidate regulatory modules that may contribute to HGPS progression. Finally, I used MOFA to integrate mRNA and miRNA data. This analysis confirmed the differences observed in individual analyses: HGPS and WT VSMCs have distinct miRNA and mRNA expression profiles. MOFA assigns weights for each feature (mRNA/miRNAs), reflecting their contribution to the latent factors that capture major sources of variation in the data. While differential expression analysis identifies statistically significant changes in gene expression, MOFA highlights features that contribute most to the variance between samples. Combining both approaches — selecting features with high MOFA weights and significant differential expression — allows narrowing down the mRNAs and miRNAs driving the HGPS phenotype (Chapter 4.3.3, Figure 4.15).

In summary, this study demonstrates how integrative bioinformatics approaches can improve our understanding of the molecular perturbations underlying HGPS. By leveraging multiple types of omics data and complementary analytical methods, we identified differences in gene expression between HGPS and WT samples that may contribute to the HGPS phenotype. However, we later discovered that the mice (both HGPS and WT) were carrying a viral infection. As a result, the observed alterations in immune response pathways must be analysed in the context of this infection. This additional analysis could provide new insights into the antiviral response in this HGPS mouse model.

Future work could focus on experimentally validating the predicted miRNA-mRNA interactions and investigating whether the observed immune dysfunction in

HGPS VSMCs contributes to disease progression.

5.3 Omics data analysis and logical modelling in systems biology

As discussed in Chapter 1.2, systems biology seeks to understand biological systems by analysing how their components interact over time and under different conditions. It integrates methodologies from biology, mathematics, bioinformatics, and computer science.

This thesis demonstrates how mathematical modelling and omics data analysis can be used to address the complexity of biological systems.

On one hand, logical models are typically constructed using a bottom-up approach. They are built from detailed knowledge of individual components —such as genes or proteins — and their interactions (see Chapter 1.2.3). By abstracting complex regulatory interactions into discrete logical rules, these models enable the simulation and analysis of dynamic system behaviour. Logical model serve as predictive tools to assess how biological systems respond to perturbations (e.g., gene knockouts or overexpression). Ensuring that logical model are valid is essential to use then, hence the development of MRBM.

On the other hand, omics analyses — specifically RNA-seq in this work—represent a top-down approach. Instead of focusing on individual components, this strategy captures the global behaviour of the system (e.g., change RNA expression profiles). This allows for the identification of molecular drivers and regulatory pathways underlying observed phenotypic changes.

Although I used logical modelling and omics data analysis for two distinct project, they are highly complementary.

Omics analyses provide the empirical data required to construct and validate logical models. For instance, expression patterns and regulatory pathways identified through RNA-seq can guide the refinement of logical models. In fact, revision methods such as ModRev and ARBoLoM can incorporate time-series RNA-seq data as input. Additionally, there exist inference methods that use omics data to build regulatory networks (Hill et al. 2016; Margolin et al. 2006; Qiu et al. 2020), and even logical model (Chevalier et al. 2019; Ostrowski et al. 2016; Terfve et al. 2012). Conversely, logical models offer mechanistic hypotheses that guide experimentation and even the interpretation of omics data. By suggesting which

molecular interactions are most likely responsible for observed changes, these models can direct future experimental investigations.

Overall, the iterative interplay between omics data and modelling — where omics data analysis inform model, and model predictions, in turn, refine the interpretation of omics results — deepens our understanding of complex biological systems.

In addition to the work presented in this thesis, I have been developing a logical model of dysferlinopathies, which I was unable to complete in time to include in this manuscript. Dysferlinopathies are a group of rare myopathies caused by mutations in the dysferlin gene. Patients with dysferlinopathies experience progressive skeletal muscle weakness and degeneration. Currently, there is no cure for dysferlinopathies, only palliative treatments.

My aim is to build a model that integrates and explains the molecular alterations observed in these dysferlinopathies. Ultimately, I want to provide a model that can be used as a predictive tool to better understand the disease and support the development or testing of treatments.

The current version of the model, along with the code to analyse its dynamics and preliminary results, can be accessed on the following GitHub repository: https://github.com/NdnBnBn/Dyf-model.This model is constructed using a combination of literature reviews, data from pertinent databases, and omics analysis results from existing studies, highlighting the interaction between modelling and omics data.

Bibliography

ABOU-JAOUDÉ, Wassim; TRAYNARD, Pauline; MONTEIRO, Pedro T., et al., 2016. Logical Modeling and Dynamical Analysis of Cellular Networks. *Front. Genet.* [en ligne]. Vol. 7 [visité le 2024-11-13]. ISSN 1664-8021. Disp. à l'adr. DOI: 10.3389/fgene.2016.00094 (cit. on p. 44).

ALCOCER-CUARÓN, Carlos; RIVERA, Ana L.; CASTAÑO, Victor M., 2014. Hierarchical structure of biological systems: a bioengineering approach. *Bioengineered.* Vol. 5, no. 2, p. 73–79. ISSN 2165-5987. Disp. à l'adr. DOI: 10.4161/bioe.26570 (cit. on pp. 15, 17).

ALEIXO, Frederico; KNORR, Matthias; LEITE, João, 2023. Revising Boolean Logical Models of Biological Regulatory Networks. In: *Proceedings of the Twentieth International Conference on Principles of Knowledge Representation and Reasoning* [en ligne]. Rhodes, Greece: International Joint Conferences on Artificial Intelligence Organization, p. 12–22 [visité le 2024-10-23]. ISBN 978-1-956792-02-7. Disp. à l'adr. DOI: 10.24963/kr.2023/2 (cit. on pp. 47, 109).

ÁLVAREZ-BUYLLA ROCES, María Elena; MARTÍNEZ-GARCÍA, Juan Carlos; DÁVILA-VELDERRAIN, José, et al., 2018. Modeling Procedures. In: *Modeling Methods for Medical Systems Biology* [en ligne]. Cham: Springer International Publishing. Vol. 1069, p. 35–134 [visité le 2024-05-30]. ISBN 978-3-319-89353-2 978-3-319-89354-9. Disp. à l'adr. DOI: 10.1007/978-3-319-89354-9_2. Series Title: Advances in Experimental Medicine and Biology (cit. on p. 26).

ANDREWS, Simon; KRUEGER, Felix; SEGONDS-PICHON, Anne, et al., 2012. *FastQC* [Babraham Institute]. Babraham, UK (cit. on p. 79).

ANWAR, Saeed; YOKOTA, Toshifumi, 2024. The Dysferlinopathies Conundrum: Clinical Spectra, Disease Mechanism and Genetic Approaches for Treatments. *Biomolecules* [en ligne]. Vol. 14, no. 3, p. 256 [visité le 2025-01-14]. ISSN 2218-273X. Disp. à l'adr. DOI: 10.3390/biom14030256. Publisher: MDPI AG (cit. on p. 18).

ARGELAGUET, Ricard; VELTEN, Britta; ARNOL, Damien, et al., 2018. Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology* [en ligne]. Vol. 14, no. 6, e8124 [visité le 2025-01-09]. ISSN 1744-4292, ISSN 1744-4292. Disp. à l'adr. DOI: 10.15252/msb.20178124 (cit. on pp. 25, 87).

BARBULESCU, Ruxandra; IOAN, Daniel, 2015. Multi-scale and reduced modelling of biological systems. In: *2015 9th International Symposium on Advanced Topics in Electrical Engineering (ATEE)* [en ligne]. Bucharest, Romania: IEEE, p. 597–602 [visité le 2024-07-22]. ISBN 978-1-4799-7514-3. Disp. à l'adr. DOI: 10.1109/ATEE.2015.7133876 (cit. on p. 18).

BERTALANFFY, Ludwig von, 1988. *General system theory: foundations, development, applications*. Rev. ed., 10. print. New York: Braziller. ISBN 978-0-8076-0452-6 978-0-8076-0453-3 (cit. on p. 13).

BHALLA, Upinder S.; IYENGAR, Ravi, 1999. Emergent Properties of Networks of Biological Signaling Pathways. *Science* [en ligne]. Vol. 283, no. 5400, p. 381–387 [visité le 2024-06-07]. ISSN 0036-8075, ISSN 1095-9203. Disp. à l'adr. DOI: 10.1126/science.283.5400.381 (cit. on p. 23).

BICH, Leonardo; PRADEU, Thomas; MOREAU, Jean-François, 2019. Understanding Multicellularity: The Functional Organization of the Intercellular Space. *Front. Physiol.* [en ligne]. Vol. 10, p. 1170 [visité le 2025-03-04]. ISSN 1664-042X. Disp. à l'adr. DOI: 10.3389/fphys.2019.01170 (cit. on p. 21).

BOLGER, Anthony M.; LOHSE, Marc; USADEL, Bjoern, 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* [en ligne]. Vol. 30, no. 15, p. 2114–2120 [visité le 2025-01-09]. ISSN 1367-4811, ISSN 1367-4803. Disp. à l'adr. DOI: 10.1093/bioinformatics/btu170 (cit. on p. 80).

BRUGGEMAN, Frank J.; WESTERHOFF, Hans V., 2007. The nature of systems biology. *Trends Microbiol.* Vol. 15, no. 1, p. 45–50. ISSN 0966-842X. Disp. à l'adr. DOI: 10.1016/j.tim.2006.11.003 (cit. on p. 23).

CAMPBELL, Neil A.; URRY, Lisa A.; CAIN, Michael L., et al., 2018. *Biology: a global approach*. Eleventh edition, global edition. New York, NY: Pearson. ISBN 978-1-292-17043-5 (cit. on p. 13).

CAO, Genmao; XUAN, Xuezhen; HU, Jie, et al., 2022. How vascular smooth muscle cell phenotype switching contributes to vascular disease. *Cell Commun Signal* [en ligne]. Vol. 20, no. 1, p. 180 [visité le 2024-12-06]. ISSN 1478-811X. Disp. à l'adr. DOI: 10.1186/s12964-022-00993-2 (cit. on p. 106).

CHAOUIYA, Claudine; BÉRENGUIER, Duncan; KEATING, Sarah M, et al., 2013. SBML qualitative models: a model representation format and infrastructure to foster interactions between qualitative modelling formalisms and tools. *BMC Syst Biol* [en ligne]. Vol. 7, no. 1, p. 135 [visité le 2024-07-21]. ISSN 1752-0509. Disp. à l'adr. DOI: 10.1186/1752-0509-7-135 (cit. on p. 45).

CHAOUIYA, Claudine; REMY, Elisabeth, 2013. Logical Modelling of Regulatory Networks, Methods and Applications. *Bull Math Biol* [en ligne]. Vol. 75, no. 6, p. 891–895 [visité le 2024-11-18]. ISSN 0092-8240, ISSN 1522-9602. Disp. à l'adr. DOI: 10.1007/s11538-013-9863-0 (cit. on p. 34).

Chapter 1 – What Is Self-Organization?, 2001. In: *Self-Organization in Biological Systems* [en ligne]. Princeton University Press, p. 7–14 [visité le 2024-07-01]. ISBN 978-0-691-21292-0. Disp. à l'adr. DOI: 10.1515/9780691212920-003 (cit. on p. 15).

CHEN, Jiung-Wen; SHRESTHA, Lisa; GREEN, George, et al., 2023. The hitchhikers' guide to RNA sequencing and functional analysis. *Briefings in Bioinformatics* [en ligne]. Vol. 24, no. 1, bbac529 [visité le 2025-01-09]. ISSN 1467-5463, ISSN 1477-4054. Disp. à l'adr. DOI: 10.1093/bib/bbac529 (cit. on p. 83).

CHEVALIER, Stephanie; FROIDEVAUX, Christine; PAULEVE, Loic, et al., 2019. Synthesis of Boolean Networks from Biological Dynamical Constraints using Answer-Set Programming. In: *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)* [en ligne]. Portland, OR, USA: IEEE, p. 34–41 [visité le 2025-02-26]. ISBN 978-1-72813-798-8. Disp. à l'adr. DOI: 10.1109/ICTAI.2019.00014 (cit. on p. 111).

CISNEROS, Bulmaro; GARCÍA-AGUIRRE, Ian; DE ITA, Marlon, et al., 2023. Hutchinson-Gilford Progeria Syndrome: Cellular Mechanisms and Therapeutic Perspectives. *Arch Med Res*. Vol. 54, no. 5, p. 102837. ISSN 1873-5487. Disp. à l'adr. DOI: 10.1016/j.arcmed.2023.06.002 (cit. on pp. 76, 77).

COCK, Peter J. A.; FIELDS, Christopher J.; GOTO, Naohisa, et al., 2010. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research* [en ligne]. Vol. 38, no. 6, p. 1767–1771 [visité le 2025-01-09]. ISSN 0305-1048, ISSN 1362-4962. Disp. à l'adr. DOI: 10.1093/nar/gkp1137 (cit. on p. 79).

COLL-BONFILL, Nuria; MAHAJAN, Urvashi; SHASHKOVA, Elena V., et al., 2023. Progerin induces a phenotypic switch in vascular smooth muscle cells and triggers replication stress and an aging-associated secretory signature. *GeroScience* [en ligne]. Vol. 45, no. 2, p. 965–982 [visité le 2024-12-06]. ISSN 2509-2715, ISSN 2509-2723. Disp. à l'adr. DOI: 10.1007/s11357-022-00694-1 (cit. on pp. 77, 83, 106).

CONESA, Ana; MADRIGAL, Pedro; TARAZONA, Sonia, et al., 2016. A survey of best practices for RNA-seq data analysis. *Genome Biol* [en ligne]. Vol. 17, no. 1, p. 13 [visité le 2025-01-09]. ISSN 1474-760X. Disp. à l'adr. DOI: 10.1186/s13059-016-0881-8 (cit. on pp. 78, 81–83).

CUI, Shidong; YU, Sicong; HUANG, Hsi-Yuan, et al., 2025. miRTarBase 2025: updates to the collection of experimentally validated microRNA–target interactions. *Nucleic Acids Research* [en ligne]. Vol. 53, no. D1, p. D147–D156 [visité le 2025-01-09]. ISSN 0305-1048, ISSN 1362-4962. Disp. à l'adr. DOI: 10.1093/nar/gkae1072 (cit. on p. 87).

DEMONGEOT, Jacques; SENÉ, Sylvain, 2022. About block-parallel Boolean networks: a position paper [en ligne] [visité le 2024-06-17]. Disp. à l'adr. DOI: 10.48550/ARXIV.2204.10712. Publisher: arXiv Version Number: 1 (cit. on p. 38).

DESFARGES, Sébastien; CIUFFI, Angela, 2012. Viral Integration and Consequences on Host Gene Expression. In: WITZANY, Günther (ed.). *Viruses: Essential Agents of Life* [en ligne]. Dordrecht: Springer Netherlands, p. 147–175 [visité le 2025-01-10]. ISBN 978-94-007-4898-9 978-94-007-4899-6. Disp. à l'adr. DOI: 10.1007/978-94-007-4899-6_7 (cit. on p. 106).

DEXHEIMER, Phillip J; PUJATO, Mario; ROSKIN, Krishna M, et al., 2023. VExD: a curated resource for human gene expression alterations following viral infection. *G3: Genes, Genomes, Genetics* [en ligne]. Vol. 13, no. 10, jkad176 [visité le 2025-01-10]. ISSN 2160-1836. Disp. à l'adr. DOI: 10.1093/g3journal/jkad176 (cit. on p. 106).

DIDIER, Gilles; REMY, Elisabeth; CHAOUIYA, Claudine, 2011. Mapping multivalued onto Boolean dynamics. *Journal of Theoretical Biology* [en ligne]. Vol. 270, no. 1, p. 177–184 [visité le 2024-07-21]. ISSN 00225193. Disp. à l'adr. DOI: 10.1016/j.jtbi.2010.09.017 (cit. on p. 29).

DOBIN, Alexander; DAVIS, Carrie A.; SCHLESINGER, Felix, et al., 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* [en ligne]. Vol. 29, no. 1, p. 15–21 [visité le 2025-01-09]. ISSN 1367-4811, ISSN 1367-4803. Disp. à l'adr. DOI: 10.1093/bioinformatics/bts635 (cit. on p. 80).

DONG, Weixia; WANG, Huiqin; LI, Menghui, et al., 2024. Virus-induced host genomic remodeling dysregulates gene expression, triggering tumorigenesis. *Front. Cell. Infect. Microbiol.* [en ligne]. Vol. 14, p. 1359766 [visité le 2025-01-10]. ISSN 2235-2988. Disp. à l'adr. DOI: 10.3389/fcimb.2024.1359766 (cit. on p. 106).

ERIKSSON, Maria; BROWN, W. Ted; GORDON, Leslie B., et al., 2003. Recurrent de novo point mutations in lamin A cause Hutchinson–Gilford progeria syndrome. *Nature* [en ligne]. Vol. 423, no. 6937, p. 293–298 [visité le 2025-03-05]. ISSN 0028-0836, ISSN 1476-4687. Disp. à l'adr. DOI: 10.1038/nature01629 (cit. on pp. 76, 83).

ESTRADA, Ernesto, 2023. What is a Complex System, After All? *Found Sci* [en ligne] [visité le 2024-08-29]. ISSN 1233-1821, ISSN 1572-8471. Disp. à l'adr. DOI: 10.1007/s10699-023-09917-w (cit. on p. 15).

FABREGAT, Antonio; JUPE, Steven; MATTHEWS, Lisa, et al., 2018. The Reactome Pathway Knowledgebase. *Nucleic Acids Res.* Vol. 46, no. D1, p. D649–D655. ISSN 1362-4962. Disp. à l'adr. DOI: 10.1093/nar/gkx1132 (cit. on pp. 27, 83).

FÉLIX, Marie-Anne; BARKOULAS, Michalis, 2015. Pervasive robustness in biological systems. *Nat Rev Genet* [en ligne]. Vol. 16, no. 8, p. 483–496 [visité le 2024-09-19]. ISSN 1471-0056, ISSN 1471-0064. Disp. à l'adr. DOI: 10.1038/nrg3949 (cit. on p. 18).

FRANKEL, Diane; DELECOURT, Valérie; NOVOA-DEL-TORO, Elva-María, et al., 2022. miR-376a-3p and miR-376b-3p overexpression in Hutchinson-Gilford progeria fibroblasts inhibits cell proliferation and induces premature senescence. *iScience* [en ligne]. Vol. 25, no. 2, p. 103757 [visité le 2024-11-18]. ISSN 25890042. Disp. à l'adr. DOI: 10.1016/j.isci.2022.103757 (cit. on p. 84).

FRONTERA, Walter R.; OCHALA, Julien, 2015. Skeletal Muscle: A Brief Review of Structure and Function. *Calcif Tissue Int* [en ligne]. Vol. 96, no. 3, p. 183–195 [visité le 2025-03-04]. ISSN 0171-967X, ISSN 1432-0827. Disp. à l'adr. DOI: 10.1007/s00223-014-9915-y (cit. on p. 14).

GALLEGO ROMERO, Irene; PAI, Athma A; TUNG, Jenny, et al., 2014. RNA-seq: impact of RNA degradation on transcript quantification. *BMC Biol* [en ligne]. Vol. 12, no. 1, p. 42 [visité le 2025-01-09]. ISSN 1741-7007. Disp. à l'adr. DOI: 10.1186/1741-7007-12-42 (cit. on p. 81).

GENE ONTOLOGY CONSORTIUM, 2004. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research* [en ligne]. Vol. 32, no. 90001, p. 258D–261 [visité le 2025-01-09]. ISSN 1362-4962. Disp. à l'adr. DOI: 10.1093/nar/gkh036 (cit. on p. 83).

GOLDMAN, Robert D.; SHUMAKER, Dale K.; ERDOS, Michael R., et al., 2004. Accumulation of mutant lamin A causes progressive changes in nuclear architecture in Hutchinson–Gilford progeria syndrome. *Proc. Natl. Acad. Sci. U.S.A.* [en ligne]. Vol. 101, no. 24, p. 8963–8968 [visité le 2024-12-06]. ISSN 0027-8424, ISSN 1091-6490. Disp. à l'adr. DOI: 10.1073/pnas.0402943101 (cit. on p. 76).

GOLES, Eric; NOUAL, Mathilde, 2010. Block-sequential update schedules and Boolean automata circuits. *Discrete Mathematics & Theoretical Computer Science* [en ligne]. Vol. DMTCS Proceedings vol. AL,... No. Proceedings, p. 2762 [visité le 2024-11-13]. ISSN 1365-8050. Disp. à l'adr. DOI: 10.46298/dmtcs.2762 (cit. on p. 38).

GOPAL, Radha; MARINELLI, Michael A.; ALCORN, John F., 2020. Immune Mechanisms in Cardiovascular Diseases Associated With Viral Infection. *Front. Immunol.* [en ligne]. Vol. 11, p. 570681 [visité le 2025-01-10]. ISSN 1664-3224. Disp. à l'adr. DOI: 10.3389/fimmu.2020.570681 (cit. on p. 106).

GOUVEIA, Filipe; LYNCE, Inês; MONTEIRO, Pedro T., 2020. ModRev - Model Revision Tool for Boolean Logical Models of Biological Regulatory Networks. In: ABATE, Alessandro; PETROV, Tatjana; WOLF, Verena (eds.). *Computational Methods in Systems Biology* [en ligne]. Cham: Springer International Publishing.

Vol. 12314, p. 339–348 [visité le 2024-10-23]. ISBN 978-3-030-60326-7 978-3-030-60327-4. Disp. à l'adr. DOI: 10.1007/978-3-030-60327-4_18. Series Title: Lecture Notes in Computer Science (cit. on pp. 47, 109).

HAMCZYK, Magda R.; ANDRÉS, Vicente, 2019. Vascular smooth muscle cell loss underpins the accelerated atherosclerosis in Hutchinson-Gilford progeria syndrome. *Nucleus* [en ligne]. Vol. 10, no. 1, p. 48–54 [visité le 2024-11-18]. ISSN 1949-1034, ISSN 1949-1042. Disp. à l'adr. DOI: 10.1080/19491034.2019.1589359 (cit. on pp. 77, 83).

HAMOSH, Ada; SCOTT, Alan F.; AMBERGER, Joanna, et al., 2002. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* Vol. 30, no. 1, p. 52–55. ISSN 1362-4962. Disp. à l'adr. DOI: 10.1093/nar/30.1.52 (cit. on p. 27).

HAOYU, Wu; MEIQIN, Liu; JIAOYANG, Sun, et al., 2024. Premature aging effects on COVID-19 pathogenesis: new insights from mouse models. *Sci Rep* [en ligne]. Vol. 14, no. 1, p. 19703 [visité le 2025-01-20]. ISSN 2045-2322. Disp. à l'adr. DOI: 10.1038/s41598-024-70612-2 (cit. on p. 106).

HARHOURI, Karim; FRANKEL, Diane; BARTOLI, Catherine, et al., 2018. An overview of treatment strategies for Hutchinson-Gilford Progeria syndrome. *Nucleus* [en ligne]. Vol. 9, no. 1, p. 265–276 [visité le 2024-11-18]. ISSN 1949-1034, ISSN 1949-1042. Disp. à l'adr. DOI: 10.1080/19491034.2018.1460045 (cit. on pp. 76, 77, 83).

HELIKAR, Tomáš; KOWAL, Bryan; MCCLENATHAN, Sean, et al., 2012. The Cell Collective: Toward an open and collaborative approach to systems biology. *BMC Syst Biol* [en ligne]. Vol. 6, no. 1, p. 96 [visité le 2024-07-21]. ISSN 1752-0509. Disp. à l'adr. DOI: 10.1186/1752-0509-6-96 (cit. on p. 45).

HENNEKAM, Raoul C.M., 2006. Hutchinson–Gilford progeria syndrome: Review of the phenotype. *American J of Med Genetics Pt A* [en ligne]. Vol. 140A, no. 23, p. 2603–2624 [visité le 2024-11-18]. ISSN 1552-4825, ISSN 1552-4833. Disp. à l'adr. DOI: 10.1002/ajmg.a.31346 (cit. on p. 106).

HILL, Steven M.; HEISER, Laura M.; COKELAER, Thomas, et al., 2016. Inferring causal molecular networks: empirical assessment through a community-based effort. *Nat Methods*. Vol. 13, no. 4, p. 310–318. ISSN 1548-7105. Disp. à l'adr. DOI: 10.1038/nmeth.3773 (cit. on p. 111).

HUNTER, Peter J.; BORG, Thomas K., 2003. Integration from proteins to organs: the Physiome Project. *Nat Rev Mol Cell Biol* [en ligne]. Vol. 4, no. 3, p. 237–243 [visité le 2024-09-19]. ISSN 1471-0072, ISSN 1471-0080. Disp. à l'adr. DOI: 10.1038/nrm1054 (cit. on p. 17).

IDEKER, T.; THORSSON, V.; RANISH, J. A., et al., 2001. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*. Vol. 292, no. 5518, p. 929–934. ISSN 0036-8075. Disp. à l'adr. DOI: 10.1126/science.292.5518.929 (cit. on p. 23).

INGALLS, Brian P., 2013. *Mathematical modeling in systems biology: an introduction*. Cambridge, Massachusetts: MIT Press. ISBN 978-0-262-01888-3 978-0-262-54582-2 (cit. on p. 14).

INTERNATIONAL SOCIETY FOR BIOCURATION, 2018. Biocuration: Distilling data into knowledge. *PLoS Biol*. Vol. 16, no. 4, e2002846. ISSN 1545-7885. Disp. à l'adr. DOI: 10.1371/journal.pbio.2002846 (cit. on p. 27).

JANSON, Natalia B., 2012. Non-linear dynamics of biological systems. *Contemporary Physics* [en ligne]. Vol. 53, no. 2, p. 137–168 [visité le 2024-09-23]. ISSN 0010-7514, ISSN 1366-5812. Disp. à l'adr. DOI: 10.1080/00107514.2011.644441 (cit. on p. 18).

KANEHISA, M.; GOTO, S., 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. Vol. 28, no. 1, p. 27–30. ISSN 0305-1048. Disp. à l'adr. DOI: 10.1093/nar/28.1.27 (cit. on pp. 27, 83).

KAPITANIAK, Tomasz; JAFARI, Sajad, 2018. Nonlinear effects in life sciences. *Eur. Phys. J. Spec. Top.* [en ligne]. Vol. 227, no. 7-9, p. 693–696 [visité le 2024-09-23]. ISSN 1951-6355, ISSN 1951-6401. Disp. à l'adr. DOI: 10.1140/epjst/e2018-800104-6 (cit. on p. 18).

KARLEBACH, Guy; SHAMIR, Ron, 2008. Modelling and analysis of gene regulatory networks. *Nat Rev Mol Cell Biol*. Vol. 9, no. 10, p. 770–780. ISSN 1471-0080. Disp. à l'adr. DOI: 10.1038/nrm2503 (cit. on p. 29).

KAUFFMAN, Stuart A, 1993. *The Origins of Order: Self-Organization and Selection in Evolution* [en ligne]. Oxford University PressNew York, NY [visité le 2024-11-13]. ISBN 978-0-19-507951-7 978-0-19-770175-1. Disp. à l'adr. DOI: 10.1093/oso/9780195079517.001.0001 (cit. on pp. 34, 36).

KESIĆ, Srdjan, 2016. Systems biology, emergence and antireductionism. *Saudi Journal of Biological Sciences* [en ligne]. Vol. 23, no. 5, p. 584–591 [visité le 2024-05-30]. ISSN 1319562X. Disp. à l'adr. DOI: 10.1016/j.sjbs.2015.06.015 (cit. on p. 23).

KHAN, Yusuf S.; FARHANA, Aisha, 2024. Histology, Cell. In: *StatPearls* [en ligne]. Treasure Island (FL): StatPearls Publishing [visité le 2024-07-20]. Disp. à l'adr.: http://www.ncbi.nlm.nih.gov/books/NBK554382/ (cit. on p. 16).

KLARNER, Hannes; HEINITZ, Frederike; NEE, Sarah, et al., 2020. Basins of Attraction, Commitment Sets, and Phenotypes of Boolean Networks. *IEEE/ACM Trans Comput Biol Bioinform*. Vol. 17, no. 4, p. 1115–1124. ISSN 1557-9964. Disp. à l'adr. DOI: 10.1109/TCBB.2018.2879097 (cit. on p. 44).

KLARNER, Hannes; STRECK, Adam; SIEBERT, Heike, 2017. PyBoolNet: a python package for the generation, analysis and visualization of boolean networks. *Bioinformatics* [en ligne]. Vol. 33, no. 5, p. 770–772 [visité le 2024-07-21]. ISSN 1367-4803, ISSN 1367-4811. Disp. à l'adr. DOI: 10.1093/bioinformatics/btw682 (cit. on pp. 44, 45).

KOHL, P; CRAMPIN, E J; QUINN, T A, et al., 2010. Systems Biology: An Approach. *Clin Pharmacol Ther* [en ligne]. Vol. 88, no. 1, p. 25–33 [visité le 2024-05-21]. ISSN 0009-9236, ISSN 1532-6535. Disp. à l'adr. DOI: 10.1038/clpt.2010.92 (cit. on pp. 14, 24, 108).

KONIECZNY, Leszek; ROTERMAN-KONIECZNA, Irena; SPÓLNIK, Paweł, 2023. Regulation in Biological Systems. In: *Systems Biology* [en ligne]. Cham: Springer International Publishing, p. 159–203 [visité le 2024-09-23]. ISBN 978-3-031-31556-5 978-3-031-31557-2. Disp. à l'adr. DOI: 10.1007/978-3-031-31557-2_4 (cit. on pp. 18, 22).

KUKURBA, Kimberly R.; MONTGOMERY, Stephen B., 2015. RNA Sequencing and Analysis. *Cold Spring Harb Protoc* [en ligne]. Vol. 2015, no. 11, pdb.top084970 [visité le 2025-01-09]. ISSN 1940-3402, ISSN 1559-6095. Disp. à l'adr. DOI: 10.1101/pdb.top084970 (cit. on p. 79).

LADYMAN, James; LAMBERT, James; WIESNER, Karoline, 2013. What is a complex system? *Euro Jnl Phil Sci* [en ligne]. Vol. 3, no. 1, p. 33–67 [visité le 2024-08-29]. ISSN 1879-4912, ISSN 1879-4920. Disp. à l'adr. DOI: 10.1007/s13194-012-0056-8 (cit. on p. 16).

LAMIS, Aselah; SIDDIQUI, Shiza W; ASHOK, Tejaswini, et al., 2022. Hutchinson-Gilford Progeria Syndrome: A Literature Review. *Cureus* [en ligne] [visité le 2025-01-05]. ISSN 2168-8184. Disp. à l'adr. DOI: 10.7759/cureus.28629 (cit. on p. 76).

LANDER, E. S.; LINTON, L. M.; BIRREN, B., et al., 2001. Initial sequencing and analysis of the human genome. *Nature*. Vol. 409, no. 6822, p. 860–921. ISSN 0028-0836. Disp. à l'adr. DOI: 10.1038/35057062 (cit. on p. 23).

LANGMEAD, Ben; TRAPNELL, Cole; POP, Mihai, et al., 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* [en ligne]. Vol. 10, no. 3, R25 [visité le 2025-01-09]. ISSN 1474-760X. Disp. à l'adr. DOI: 10.1186/gb-2009-10-3-r25 (cit. on p. 80).

LE NOVÈRE, Nicolas, 2015. Quantitative and logic modelling of molecular and gene networks. *Nat Rev Genet*. Vol. 16, no. 3, p. 146–158. ISSN 1471-0064. Disp. à l'adr. DOI: 10.1038/nrg3885 (cit. on pp. 27, 28).

LIAO, Yang; SMYTH, Gordon K.; SHI, Wei, 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* [en ligne]. Vol. 30, no. 7, p. 923–930 [visité le 2025-01-09]. ISSN 1367-4811, ISSN 1367-4803. Disp. à l'adr. DOI: 10.1093/bioinformatics/btt656 (cit. on p. 81).

LOVE, Michael I; HUBER, Wolfgang; ANDERS, Simon, 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* [en ligne]. Vol. 15, no. 12, p. 550 [visité le 2025-01-09]. ISSN 1474-760X. Disp. à l'adr. DOI: 10.1186/s13059-014-0550-8 (cit. on p. 82).

MACHADO, Daniel; COSTA, Rafael S; ROCHA, Miguel, et al., 2011. Modeling formalisms in Systems Biology. *AMB Express* [en ligne]. Vol. 1, no. 1, p. 45 [visité le 2024-05-21]. ISSN 2191-0855. Disp. à l'adr. DOI: 10.1186/2191-0855-1-45 (cit. on p. 29).

MARGOLIN, Adam A.; NEMENMAN, Ilya; BASSO, Katia, et al., 2006. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*. Vol. 7 Suppl 1, no. Suppl 1, S7. ISSN 1471-2105. Disp. à l'adr. DOI: 10.1186/1471-2105-7-S1-S7 (cit. on p. 111).

MAZZOCCHI, Fulvio, 2008. Complexity in biology. Exceeding the limits of reductionism and determinism using complexity theory. *EMBO Rep*. Vol. 9, no. 1, p. 10–14. ISSN 1469-221X. Disp. à l'adr. DOI: 10.1038/sj.embor.7401147 (cit. on p. 15).

MESTL, Thomas; PLAHTE, Erik; OMHOLT, Stig W., 1995. A mathematical framework for describing and analysing gene regulatory networks. *Journal of Theoretical Biology* [en ligne]. Vol. 176, no. 2, p. 291–300 [visité le 2024-10-29]. ISSN 00225193. Disp. à l'adr. DOI: 10.1006/jtbi.1995.0199 (cit. on p. 33).

MICHELSEN, Claus, 2006. Functions: a modelling tool in mathematics and science. *Zentralblatt für Didaktik der Mathematik* [en ligne]. Vol. 38, no. 3, p. 269–280 [visité le 2024-06-07]. ISSN 1863-9690, ISSN 1863-9704. Disp. à l'adr. DOI: 10.1007/BF02652810 (cit. on p. 26).

MÜSSEL, Christoph; HOPFENSITZ, Martin; KESTLER, Hans A., 2010. BoolNet—an R package for generation, reconstruction and analysis of Boolean networks. *Bioinformatics* [en ligne]. Vol. 26, no. 10, p. 1378–1380 [visité le 2024-07-21]. ISSN 1367-4811, ISSN 1367-4803. Disp. à l'adr. DOI: 10.1093/bioinformatics/btq124 (cit. on p. 45).

NALDI, A.; BERENGUIER, D.; FAURÉ, A., et al., 2009. Logical modelling of regulatory networks with GINsim 2.3. *Biosystems* [en ligne]. Vol. 97, no. 2, p. 134–139 [visité le 2024-07-21]. ISSN 03032647. Disp. à l'adr. DOI: 10.1016/j.biosystems.2009.04.008 (cit. on pp. 32, 45).

NALDI, Aurélien; HERNANDEZ, Céline; LEVY, Nicolas, et al., 2018. The CoLoMoTo Interactive Notebook: Accessible and Reproducible Computational Analyses for Qualitative Biological Networks. *Front Physiol*. Vol. 9, p. 680. ISSN 1664-042X. Disp. à l'adr. DOI: 10.3389/fphys.2018.00680 (cit. on p. 45).

NOBLE, Raymond; TASAKI, Kazuyo; NOBLE, Penelope J., et al., 2019. Biological Relativity Requires Circular Causality but Not Symmetry of Causation: So, Where, What and When Are the Boundaries? *Front. Physiol.* [en ligne]. Vol. 10, p. 827 [visité le 2024-09-19]. ISSN 1664-042X. Disp. à l'adr. DOI: 10.3389/fphys.2019.00827 (cit. on p. 17).

NOVÈRE, Nicolas Le; HUCKA, Michael; MI, Huaiyu, et al., 2009. The Systems Biology Graphical Notation. *Nature Biotechnology* [en ligne]. Vol. 27, no. 8, p. 735–741 [visité le 2025-08-06]. ISSN 1087-0156, ISSN 1546-1696. Disp. à l'adr. DOI: 10.1038/nbt.1558 (cit. on p. 28).

OKONECHNIKOV, Konstantin; CONESA, Ana; GARCÍA-ALCALDE, Fernando, 2016. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* [en ligne]. Vol. 32, no. 2, p. 292–294 [visité le 2025-01-09]. ISSN 1367-4811, ISSN 1367-4803. Disp. à l'adr. DOI: 10.1093/bioinformatics/btv566 (cit. on p. 80).

OSTROWSKI, M.; PAULEVÉ, L.; SCHAUB, T., et al., 2016. Boolean network identification from perturbation time series data combining dynamics abstraction and logic programming. *Biosystems* [en ligne]. Vol. 149, p. 139–153 [visité le 2025-02-26]. ISSN 03032647. Disp. à l'adr. DOI: 10.1016/j.biosystems.2016.07.009 (cit. on p. 111).

PAPATHEODOROU, Irene; FONSECA, Nuno A.; KEAYS, Maria, et al., 2018. Expression Atlas: gene and protein expression across multiple studies and organisms. *Nucleic Acids Res.* Vol. 46, no. D1, p. D246–D251. ISSN 1362-4962. Disp. à l'adr. DOI: 10.1093/nar/gkx1158 (cit. on p. 27).

PAULEVÉ, Loïc; KOLČÁK, Juri; CHATAIN, Thomas, et al., 2020. Reconciling qualitative, abstract, and scalable modeling of biological networks. *Nat Commun* [en ligne]. Vol. 11, no. 1, p. 4256 [visité le 2025-03-04]. ISSN 2041-1723. Disp. à l'adr. DOI: 10.1038/s41467-020-18112-5 (cit. on pp. 36, 39, 42).

PAULEVÉ, Loïc; SENÉ, Sylvain, 2022. Boolean Networks and Their Dynamics: The Impact of Updates. In: DE MARIA, Elisabetta (ed.). *Systems Biology Modelling and Analysis* [en ligne]. 1st ed. Wiley, p. 173–250 [visité le 2024-04-08]. ISBN 978-1-119-71653-2 978-1-119-71660-0. Disp. à l'adr. DOI: 10.1002/9781119716600.ch6 (cit. on pp. 35, 37–39).

PERFETTO, Livia; BRIGANTI, Leonardo; CALDERONE, Alberto, et al., 2016. SIGNOR: a database of causal relationships between biological entities. *Nucleic Acids Res.* Vol. 44, no. D1, p. D548–554. ISSN 1362-4962. Disp. à l'adr. DOI: 10.1093/nar/gkv1048 (cit. on p. 27).

PIÑERO, Janet; BRAVO, Àlex; QUERALT-ROSINACH, Núria, et al., 2017. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.* Vol. 45, no. D1, p. D833–D839. ISSN 1362-4962. Disp. à l'adr. DOI: 10.1093/nar/gkw943 (cit. on p. 27).

PUTRI, Givanna H; ANDERS, Simon; PYL, Paul Theodor, et al., 2022. Analysing high-throughput sequencing data in Python with HTSeq 2.0. *Bioinformatics* [en ligne]. Vol. 38, no. 10, p. 2943–2945 [visité le 2025-01-09]. ISSN 1367-4803, ISSN 1367-4811. Disp. à l'adr. DOI: 10.1093/bioinformatics/btac166 (cit. on p. 81).

QIU, Xiaojie; RAHIMZAMANI, Arman; WANG, Li, et al., 2020. Inferring Causal Gene Regulatory Networks from Coupled Single-Cell Expression Dynamics Using Scribe. *Cell Syst.* Vol. 10, no. 3, 265–274.e11. ISSN 2405-4720. Disp. à l'adr. DOI: 10.1016/j.cels.2020.02.003 (cit. on p. 111).

RAHMAN, Anisur; TIWARI, Abhinav; NARULA, Jatin, et al., 2018. Importance of Feedback and Feedforward Loops to Adaptive Immune Response Modeling. *CPT Pharmacom & Syst Pharma* [en ligne]. Vol. 7, no. 10, p. 621–628 [visité le 2024-09-23]. ISSN 2163-8306, ISSN 2163-8306. Disp. à l'adr. DOI: 10.1002/psp4.12352 (cit. on p. 18).

SAADATPOUR, Assieh; ALBERT, Réka, 2013. Boolean modeling of biological regulatory networks: A methodology tutorial. *Methods* [en ligne]. Vol. 62, no. 1, p. 3–12 [visité le 2024-06-20]. ISSN 10462023. Disp. à l'adr. DOI: 10.1016/j.ymeth.2012.10.012 (cit. on pp. 35, 47).

SATAM, Heena; JOSHI, Kandarp; MANGROLIA, Upasana, et al., 2023. Next-Generation Sequencing Technology: Current Trends and Advancements. *Biology* [en ligne]. Vol. 12, no. 7, p. 997 [visité le 2025-01-09]. ISSN 2079-7737. Disp. à l'adr. DOI: 10.3390/biology12070997 (cit. on p. 79).

SCHLESSINGER, Joseph, 2000. Cell Signaling by Receptor Tyrosine Kinases. *Cell* [en ligne]. Vol. 103, no. 2, p. 211–225 [visité le 2024-09-19]. ISSN 00928674. Disp. à l'adr. DOI: 10.1016/S0092-8674(00)00114-8 (cit. on p. 16).

SCHROEDER, Andreas; MUELLER, Odilo; STOCKER, Susanne, et al., 2006. The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC Molecular Biol* [en ligne]. Vol. 7, no. 1, p. 3 [visité le 2025-01-09]. ISSN 1471-2199. Disp. à l'adr. DOI: 10.1186/1471-2199-7-3 (cit. on p. 79).

SHAHZAD, Khuram; LOOR, Juan J., 2012. Application of Top-Down and Bottom-up Systems Approaches in Ruminant Physiology and Metabolism. *Curr Genomics.* Vol. 13, no. 5, p. 379–394. ISSN 1389-2029. Disp. à l'adr. DOI: 10.2174/138920212801619269 (cit. on pp. 24, 26, 108).

SHAMIR, Maya; BAR-ON, Yinon; PHILLIPS, Rob, et al., 2016. SnapShot: Timescales in Cell Biology. *Cell* [en ligne]. Vol. 164, no. 6, 1302–1302.e1 [visité le 2024-11-13]. ISSN 00928674. Disp. à l'adr. DOI: 10.1016/j.cell.2016.02.058 (cit. on p. 36).

SOMVANSHI, Pramod Rajaram; VENKATESH, K. V., 2014. A conceptual review on systems biology in health and diseases: from biological networks to modern therapeutics. *Syst Synth Biol* [en ligne]. Vol. 8, no. 1, p. 99–116 [visité le 2025-03-04]. ISSN 1872-5325, ISSN 1872-5333. Disp. à l'adr. DOI: 10.1007/s11693-013-9125-3 (cit. on p. 22).

STOLL, Gautier; CARON, Barthélémy; VIARA, Eric, et al., 2017. MaBoSS 2.0: an environment for stochastic Boolean modeling. *Bioinformatics* [en ligne]. Vol. 33, no. 14, p. 2226–2228 [visité le 2024-11-19]. ISSN 1367-4803, ISSN 1367-4811. Disp. à l'adr. DOI: 10.1093/bioinformatics/btx123 (cit. on p. 45).

SUZUKI, Mari; JENG, Linda J.B.; CHEFO, Solomon, et al., 2023. FDA approval summary for lonafarnib (Zokinvy) for the treatment of Hutchinson-Gilford progeria syndrome and processing-deficient progeroid laminopathies. *Genetics in Medicine* [en ligne]. Vol. 25, no. 2, p. 100335 [visité le 2025-01-16]. ISSN 10983600. Disp. à l'adr. DOI: 10.1016/j.gim.2022.11.003 (cit. on p. 77).

TAVASSOLY, Iman; GOLDFARB, Joseph; IYENGAR, Ravi, 2018. Systems biology primer: the basic methods and approaches. *Essays in Biochemistry* [en ligne]. Vol. 62, no. 4, p. 487–500 [visité le 2024-05-21]. ISSN 0071-1365, ISSN 1744-1358. Disp. à l'adr. DOI: 10.1042/EBC20180003 (cit. on pp. 24, 108).

TERFVE, Camille; COKELAER, Thomas; HENRIQUES, David, et al., 2012. CellNOptR: a flexible toolkit to train protein signaling networks to data using multiple logic formalisms. *BMC Syst Biol* [en ligne]. Vol. 6, no. 1, p. 133 [visité le 2025-02-26]. ISSN 1752-0509. Disp. à l'adr. DOI: 10.1186/1752-0509-6-133 (cit. on p. 111).

THOMAS, René, 1991. Regulatory networks seen as asynchronous automata: A logical description. *Journal of Theoretical Biology* [en ligne]. Vol. 153, no. 1, p. 1–23 [visité le 2024-04-08]. ISSN 00225193. Disp. à l'adr. DOI: 10.1016/S0022-5193(05)80350-9 (cit. on pp. 33, 36).

TORRES, Nestor V.; SANTOS, Guido, 2015. The (Mathematical) Modeling Process in Biosciences. *Front Genet*. Vol. 6, p. 354. ISSN 1664-8021. Disp. à l'adr. DOI: 10.3389/fgene.2015.00354 (cit. on pp. 26, 29, 30).

TRINH, Hung-Cuong; KWON, Yung-Keun, 2019. RMut: R package for a Boolean sensitivity analysis against various types of mutations. *PLoS ONE* [en ligne]. Vol. 14, no. 3, e0213736 [visité le 2024-11-13]. ISSN 1932-6203. Disp. à l'adr. DOI: 10.1371/journal.pone.0213736 (cit. on p. 45).

TRINH, Van-Giang; BENHAMOU, Belaid; PAULEVÉ, Loïc, 2024. mpbn: a simple tool for efficient edition and analysis of elementary properties of Boolean networks [en ligne] [visité le 2024-07-21]. Disp. à l'adr. DOI: 10.48550/ARXIV.2403.06255 (cit. on p. 45).

TZAFESTAS, Spyros G., 2018. Adaptation, Complexity, and Complex Adaptive Systems. In: *Energy, Information, Feedback, Adaptation, and Self-organization* [en ligne]. Cham: Springer International Publishing. Vol. 90, p. 409–460 [visité le 2024-06-27]. ISBN 978-3-319-66998-4 978-3-319-66999-1. Disp. à l'adr. DOI: 10.1007/978-3-319-66999-1_8. Series Title: Intelligent Systems, Control and Automation: Science and Engineering (cit. on p. 19).

WANG, Liguo; WANG, Shengqin; LI, Wei, 2012. RSeQC: quality control of RNA-seq experiments. *Bioinformatics* [en ligne]. Vol. 28, no. 16, p. 2184–2185 [visité le 2025-01-09]. ISSN 1367-4811, ISSN 1367-4803. Disp. à l'adr. DOI: 10.1093/bioinformatics/bts356 (cit. on p. 80).

WHITACRE, James M., 2010. Degeneracy: a link between evolvability, robustness and complexity in biological systems. *Theor Biol Med Model*. Vol. 7, p. 6. ISSN 1742-4682. Disp. à l'adr. DOI: 10.1186/1742-4682-7-6 (cit. on p. 19).

WOLKENHAUER, Olaf; MESAROVIĆ, Mihajlo, 2005. Feedback dynamics and cell function: Why systems biology is called Systems Biology. *Mol. Biosyst.* [en ligne]. Vol. 1, no. 1, p. 14 [visité le 2024-09-23]. ISSN 1742-206X, ISSN 1742-2051. Disp. à l'adr. DOI: 10.1039/b502088n (cit. on p. 18).

WOLKENHAUER, Olaf; MUIR, Allan, 2011. The Complexity of Cell-Biological Systems. In: *Philosophy of Complex Systems* [en ligne]. Elsevier, p. 355–385 [visité le 2024-06-27]. ISBN 978-0-444-52076-0. Disp. à l'adr. DOI: 10.1016/B978-0-444-52076-0.50013-4 (cit. on p. 22).

ZHAO, Shanrong; YE, Zhan; STANTON, Robert, 2020. Misuse of RPKM or TPM normalization when comparing across samples and sequencing protocols. *RNA* [en ligne]. Vol. 26, no. 8, p. 903–909 [visité le 2025-01-09]. ISSN 1355-8382, ISSN 1469-9001. Disp. à l'adr. DOI: 10.1261/rna.074922.120 (cit. on p. 81).

ZHAO, Shanrong; ZHANG, Ying; GAMINI, Ramya, et al., 2018. Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: polyA+ selection versus rRNA depletion. *Sci Rep* [en ligne]. Vol. 8, no. 1, p. 4781 [visité le 2025-01-09]. ISSN 2045-2322. Disp. à l'adr. DOI: 10.1038/s41598-018-23226-4 (cit. on p. 79).

ZHAO, Yingdong; LI, Ming-Chung; KONATÉ, Mariam M., et al., 2021. TPM, FPKM, or Normalized Counts? A Comparative Study of Quantification Measures for the Analysis of RNA-seq Data from the NCI Patient-Derived Models Repository. *J Transl Med* [en ligne]. Vol. 19, no. 1, p. 269 [visité le 2025-01-09]. ISSN 1479-5876. Disp. à l'adr. DOI: 10.1186/s12967-021-02936-w (cit. on p. 81).

ZHENG, Yuting; SRIRAM, Ganesh, 2010. Mathematical Modeling: Bridging the Gap between Concept and Realization in Synthetic Biology. *Journal of Biomedicine and Biotechnology* [en ligne]. Vol. 2010, p. 1–16 [visité le 2024-05-30]. ISSN 1110-7243, ISSN 1110-7251. Disp. à l'adr. DOI: 10.1155/2010/541609 (cit. on p. 29).