Préparée à l'Institut Curie

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

From the mechanistic modeling of signaling pathways in cancer to the interpretation of models and their contributions: clinical applications and statistical evaluation

Soutenance prévue par
Jonas BEAL
Le 23/09/2020

École doctorale n°515
Complexité du Vivant

Spécialité
Génomique

Composition du jury :

Test NOM	Titre, Établissement	*Rapporteur*
Test NOM	Titre, Établissement	*Rapporteur*
Test NOM	Titre, Établissement	*Examinateur*
Test NOM	Titre, Établissement	*Examinateur*
Emmanuel BARILLOT	Institut Curie, INSERM, Mines ParisTech, PSL	*Directeur de thèse*
Aurélien LATOUCHE	Institut Curie, Cnam	*Directeur de thèse*
Laurence CALZONE	Institut Curie, INSERM, Mines ParisTech, PSL	*Co-encadrante de thèse*

Abstract

Beyond its genetic mechanisms, cancer can be understood as a network disease that often results from the interaction between different perturbations in a cellular regulatory network. The dynamics of these networks and associated signaling pathways are complex and require integrated approaches. One approach is to design mechanistic models that translate the biological knowledge of networks in mathematical terms to simulate the molecular features of cancers in a computer-readable form. However, these models only reflect the general mechanisms at work in cancers.

This thesis proposes to define personalized mechanistic models of cancer. A generic model is first defined in a logical (or Boolean) formalism, before using omics data (mutations, RNA, proteins) from patients or cell lines in order to make the model specific to each one profile. These personalized models can then be compared with the clinical data of patients in order to validate them. The response to treatment is investigated in particular in this thesis. The explicit representation of the molecular mechanisms by these models allows to simulate the effect of different treatments according to their targets and to verify if the sensitivity of a patient to a drug is well predicted by the corresponding personalized model. An example concerning the response to BRAF inhibitors in melanomas and colorectal cancers is thus presented.

The comparison of mechanistic models of cancer, those presented in this thesis and others, with clinical data also encourages a rigorous evaluation of their possible benefits in the context of medical use. The quantification and interpretation of the value of certain prognostic models is briefly presented before focusing on the particular case of models able to recommend the best treatment for each patient according to his molecular profile. A theoretical framework is defined to extend causal inference methods to the evaluation of such precision medicine algorithms. An illustration is provided using simulated data and patient derived xenografts.

All the methods and applications put forward a possible path from the design of mechanistic models of cancer to their evaluation using statistical

models emulating clinical trials.

Key-words: Modeling, Cancer, Mechanistic model, Biostatistics, Causal inference, Precision medicine

Résumé

Au delà de ses mécanismes génétiques, le cancer peut-être compris comme une maladie de réseaux qui résulte souvent de l'interaction entre différentes perturbations dans un réseau de régulation cellulaire. La dynamique de ces réseaux et des voies de signalisation associées est complexe et requiert des approches intégrées. Une d'entre elles est la conception de modèles dits mécanistiques qui traduisent mathématiquement la connaissance biologique des réseaux afin de pouvoir simuler le fonctionnement moléculaire des cancers informatiquement. Ces modèles ne traduisent cependant que les mécanismes généraux à l'oeuvre dans certains cancers en particulier.

Cette thèse propose en premier lieu de définir des modèles mécanistiques personnalisés de cancer. Un modèle générique est d'abord défini dans un formalisme logique (ou Booléen), avant d'utiliser les données omiques (mutations, ARN, protéines) de patients ou de lignées cellulaires afin de rendre le modèle spécifique à chacun. Ces modèles personnalisés peuvent ensuite être confrontés aux données cliniques de patients pour vérifier leur validité. Le cas de la réponse clinique aux traitements est exploré en particulier dans cette thèse. La représentation explicite des mécanismes moléculaires par ces modèles permet en effet de simuler l'effet de différents traitements suivant leur mode d'action et de vérifier si la sensibilité d'un patient à un traitement est bien prédite par le modèle personnalisé correspondant. Un exemple concernant la réponse aux inhibiteurs de BRAF dans les mélanomes et cancers colorectaux est ainsi proposé.

La confrontation des modèles mécanistiques de cancer, ceux présentés dans cette thèse et d'autres, aux données cliniques incite par ailleurs à évaluer rigoureusement leurs éventuels bénéfices dans la cadre d'une utilisation médicale. La quantification et l'interprétation de la valeur de certains modèles à visée pronostique est brièvement présentée avant de se focaliser sur le cas particulier des modèles capables de sélectionner le meilleur traitement pour chaque patient en fonction des ses caractéristiques moléculaires. Un cadre théorique est proposé pour étendre les méthodes d'inférence causale à l'évaluation de tels algorithmes de médecine de précision. Une illustra-

tion est fournie à l'aide de données simulées et de xénogreffes dérivées de patients.

L'ensemble des méthodes et applications décrites tracent donc un chemin, de la conception de modèles mécanistiques de cancer à leur évaluation grâce à des modèles statistiques émulant des essais cliniques.

Mots-clés: Modélisation, Cancer, Modèle mécanistique, Biostatistiques, Inférence causale, Médecine de précision

Acknowledgements

Many persons to thanks. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum

Preface

The present thesis is structured in three parts, each subdivided into three chapters. Since the whole thesis is about cancer modeling, the first part aims at defining the type of model to be referred to, and in particular models that will be called mechanistic, as well as the object of the modeling, i.e. the molecular networks involved in cancer. So the first part answers the question: **what is a cancer model and what is its purpose?**

The second part will be devoted to the methods developed during this thesis to transform qualitative models of molecular networks, known as logic models, into personalized models that can be interpreted clinically. In short, **how can a mathematical representation of biological knowledge be transformed into a tool that contributes to the understanding of the clinical manifestations of cancer?**

Finally, the third and last part will look at how the clinical relevance of all the above-mentioned models can be rigorously evaluated, both in their ability to predict the evolution of the disease and in their ability to recommend the most appropriate treatments for each patient. **How to quantify and interpret the value of the clinical information delivered by these models?**

Moreover, this thesis also exists in an online version that allows to take advantage of the interactivity of some graphs and applications: https://jonasbeal.github.io/thesis/.

Table of contents

Page

List of Tables **xiv**

List of Figures **xv**

I Cells and their models **1**

1 Scientific modeling: abstract the complexity **3**
1.1 What is a model? . . . 3
1.1.1 In your own words . . . 3
1.1.2 Physical world and world of ideas . . . 6
1.1.3 Preview about cancer models . . . 8
1.2 Statistics or mechanistic . . . 9
1.2.1 The inside of the box . . . 9
1.2.2 A tale of prey and predators . . . 13
1.3 Simplicity is the ultimate sophistication . . . 16

2 Cancer as deregulation of complex machinery **19**
2.1 What is cancer? . . . 19
2.2 Cancer from a distance: epidemiology and main figures . . . 22
2.3 Basic molecular biology and cancer . . . 23
2.3.1 Central dogma and core principles . . . 23
2.3.2 A rogue machinery . . . 26
2.4 The new era of genomics . . . 27
2.4.1 From sequencing to multi-omics data . . . 27
2.4.2 State-of-the art of cancer data . . . 28
2.5 Data and beyond: from genetic to network disease . . . 30

3 Mechanistic modeling of cancer: from complex disease to systems biology **35**
3.1 Introducing the diversity of mechanistic models of cancer . . . 35

3.2 Cell circuitry and the need for cancer systems biology 38
3.3 Mechanistic models of molecular signaling 40
3.3.1 Networks and data 41
3.3.2 Different formalisms for different applications 42
3.3.3 Some examples of complex features 45
3.4 From mechanistic models to clinical impact? 46
3.4.1 A new class of biomarkers 46
3.4.2 Prognostic models 47
3.4.3 Predictive models 49

II Personalized logical models of cancer 51

4 Logical modeling principles and data integration 53
4.1 Logical modeling paradigms for qualitative description . . . 54
4.1.1 Regulatory graph and logical rules 55
4.1.2 State transition graph and updates 56
4.1.3 Tools for logical modeling 58
4.2 The MaBoSS framework for logical modeling 59
4.2.1 Gillespie algorithm 59
4.2.2 A stochastic exploration of model behaviours 61
4.2.3 From theoretical models to data models? 61
4.3 Data integration and semi-quantitative logical modeling . . . 62
4.3.1 Build the regulatory graph 62
4.3.2 Define the logical rules 65
4.3.3 Validate the model 66

5 Personalization of logical models: method and prognostic validation 69
5.1 From one generic model to data-specific models with PROFILE method 70
5.1.1 Gathering knowledge and data 70
5.1.2 Adapting patient profiles to a logical model 72
5.1.3 Personalizing logical models with patient 77
5.2 An integration tool for high-dimensional data? 81
5.2.1 Biological relevance in cell lines 81
5.2.2 Validation with patient data 83
5.2.3 Perspectives 85

6 Personalized logical models to study an interpret drug response 87

6.1 One step further with drugs . . . 88
6.1.1 Methods . . . 88
6.1.2 Brute force . . . 88
6.2 A case study on BRAF in melanoma and colorectal cancers 88
6.3 Limitations and perspectives illustrated by a prostate cancer study . . . 88

Appendix **88**

A About datasets **89**
A.1 Cell lines . . . 89
A.1.1 Omics profiles . . . 89
A.1.2 Drug screenings . . . 89
A.1.3 CRISPR-Cas9 screening . . . 91
A.2 Patient-derived xenografts . . . 92
A.3 Patients . . . 92
A.3.1 METABRIC . . . 92
A.3.2 TCGA: Breast cancer . . . 93
A.3.3 TCGA: Prostate cancer . . . 93

B About logical models **95**
B.1 Generic logical model of cancer pathways . . . 95
B.2 Logical model of BRAF pathways in melanoma and colorectal cancer . . . 96
B.3 Logical model of prostate cancer . . . 96

C About causality **97**
C.1 Theoretical framework . . . 97

References **99**

Bibliography **101**

List of Tables

Table **Page**

1.1 **Some pros and cons for mechanistic and statistical modeling**. Adapted from Baker et al. [2018]. 12

3.1 **Features of quantitative and qualitative modeling applied to biological molecular networks** (adapted from Le Novere [2015]) 45

List of Figures

Figure **Page**

1.1 A scientist and his model 4
1.2 Network visualization of *model* thesaurus entries 5
1.3 Scientists talk about their models: words cloud. 6
1.4 Orrery, planets and models 7
1.5 Tree visualization of *model* semantic context in cancer-related literature 9
1.6 Different modeling strategies. 10
1.7 Some analyses around Lotka-Volterra model of a prey-predator system 15

2.1 Cancer is an old disease 20
2.2 World map and national rankings of cancer as a cause of premature death 22
2.3 Incidence, mortality and survival per cancer types 24
2.4 Central dogma of molecular biology 25
2.5 Hallmarks of cancer 27
2.6 Genetic alterations frequencies for cancer types from TCGA data 29
2.7 Simplistic representation of cellular circuit and pathways 31
2.8 Genetic alterations frequencies from TCGA data mapped on a schematic signaling network 33

3.1 Dissecting a biological phenomenon using a non-computational model 36
3.2 The different scales of cancer modeling 37
3.3 PubMed trends in cancer studies. 40
3.4 Modeling a biological network: an iterative and cyclical process 42
3.5 Schematic example of logical and ODE modeling around MAPK signaling 43
3.6 Schematic example of logical and ODE modeling around MAPK signaling 48

3.7 Network model of oncogenic signal transduction in ER+ breast cancer, including some drugs and their targets 50

4.1 A simple example of a logical model 56
4.2 A simple example of a logical model 57
4.3 A simple example of a logical model 58
4.4 Main principles of MaBoSS simulation framework and Gillespie algorithm 60
4.5 Main principles of MaBoSS simulation framework and Gillespie algorithm 63

5.1 Graphical abstract of PROFILE method to personalize logical models with omics data 71
5.2 Bimodal distribution of ERG gene in TCGA prostate cancer cohort 74
5.3 Bimodality criteria and their combinations 76
5.4 Normalization of continuous data for logical modeling 78
5.5 Methods for personalization of logical models 79
5.6 Validation of personalized *Proliferation* scores in cell lines . . 82
5.7 Prognostic value of *Proliferation* scores for breast cancer patients in METABRIC cohort 83
5.8 Hazard ratios for *Proliferation* and *Apoptosis* in a survival Cox model in METABRIC cohort 84
5.9 Prognostic value of *Proliferation* scores for breast cancer patients in METABRIC cohort 85

A.1 Distribution of cancer types and data types in GDSC-associated dataset 90
A.2 Drug screening metrics in cell lines 91
A.3 Available omics and survival in METABRIC Breast Cancer dataset 92
A.4 Available omics and survival in METABRIC Breast Cancer dataset 93

B.1 Graphical abstract of PROFILE method to personalize logical models with omics data 96

Part I

Cells and their models

Chapter 1

Scientific modeling: abstract the complexity

"Ce qui est simple est toujours faux. Ce qui ne l'est pas est inutilisable."
Paul Valéry (Mauvaises pensées et autres, 1942)

The notion of modeling is embedded in science, to the point that it has sometimes been used to define the very nature of scientific research. What is called a model can, however, correspond to very different realities which need to be defined before addressing the object of this thesis which will consist, if one wants to be mischievous, in analyzing models with other models. This semantic elucidation is all the more necessary as this thesis is interdisciplinary, suspended between systems biology and biostatistics. In order to convince the reader of the need for such a preamble, he is invited to ask a statistician and a biologist how they would define what a model is.

1.1 What is a model?

1.1.1 In your own words

A model is first of all an ambiguous object and a polysemous word. It therefore seems necessary to start with a semantic study. Among the many meanings and synonymous proposed by the dictionary (Figure 1.2), while

Figure 1.1: **A scientist and his model.** Joseph Wright of Derby, *A Philosopher Giving a Lecture at the Orrery (in which a lamp is put in place of the sun)*, c. 1763-65, oil on canvas, Derby Museums and Art Gallery

some definitions are more related to art, several find echoes in scientific practice. It is sometimes a question of the physical representation of an object, often on a reduced scale as in Figure 1.1, and sometimes of a theoretical description intended to facilitate the understanding of the way in which a system works [Collins, 2020]. It is even sometimes an ideal to be reached and therefore an ambitious prospect for an introduction.

The narrower perspective of the scientist does not reduce the completeness of the dictionary's description to an unambiguous object [Bailer-Jones, 2002]. In an attempt to approach these multi-faceted objects that are the models, Daniela Bailer-Jones interviewed different scientists and asked them the same question: what is a model? Across the different profiles and fields of study, the answers vary but some patterns begin to emerge (Figure 1.3). A model must capture the essence of the phenomenon being studied. Because it eludes, voluntarily or not, many details or complexity, it is by nature a simplification of the phenomenon. These limitations may restrict its validity to certain cases or suspend it to the fulfilment of some hypotheses. They are not necessarily predictive, but they must be able to generate new hypotheses, be tested and possibly questioned. Finally, and funda-

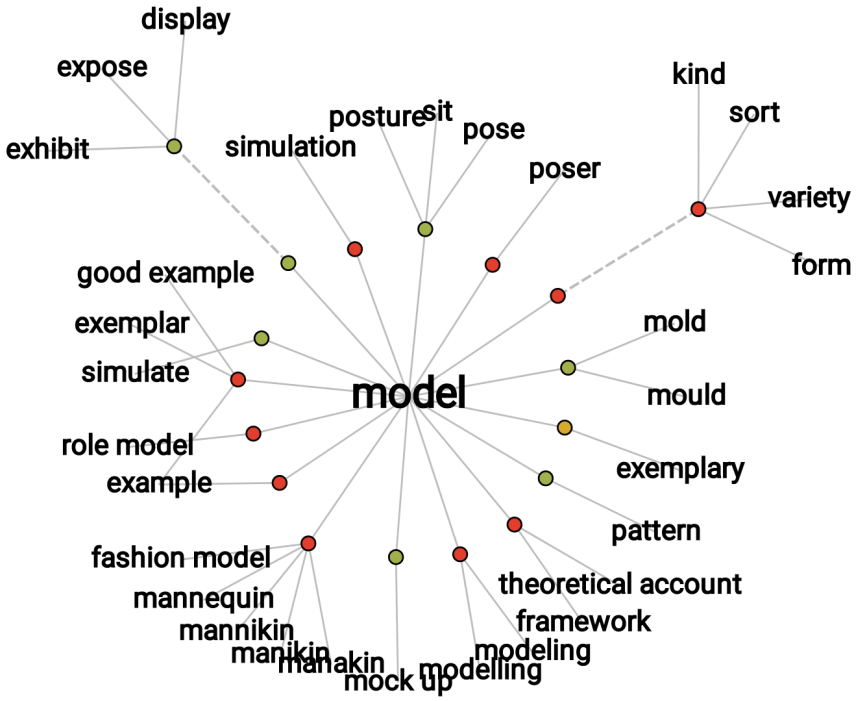

Figure 1.2: **Network visualization of *model* thesaurus entries.** Generated with the 'Visual Thesaurus' ressource

mentally, they **must provide insights about the object of study and contribute to its understanding**.

These definitions circumscribe the *model* object, its use and its objectives, but they do not in any way describe its nature. And for good reason, because even if we agree on the described contours, the biodiversity of the models remains overwhelming for taxonomists:

> *Probing models, phenomenological models, computational models, developmental models, explanatory models, impoverished models, testing models, idealized models, theoretical models, scale models, heuristic models, caricature models, exploratory models, didactic models, fantasy models, minimal models, toy models, imaginary models, mathematical models, mechanistic models, substitute models, iconic models, formal models, analogue models, and instrumental models are but some of the notions that are used to categorize models.*
> [Frigg and Hartmann, 2020]

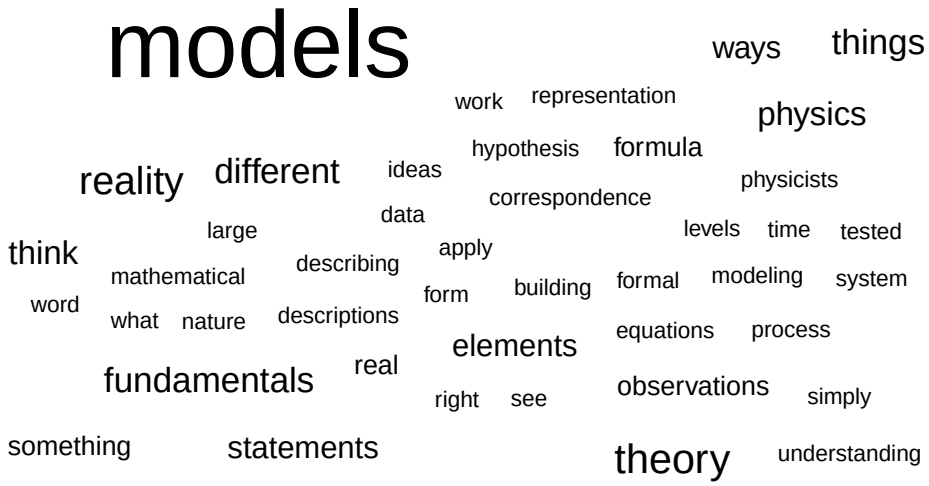

Figure 1.3: **Scientists talk about their models: words cloud.** Cloud of words summarizing the lexical fields used by scientists to talk about their models in dedicated interviews reported by Bailer-Jones [2002].

1.1.2 Physical world and world of ideas

Without claiming to be exhaustive, we can make a **first simple dichotomy between physical/material and formal/intellectual models** [Rosenblueth and Wiener, 1945]. The former consist in replacing the object of study by another object, just as physical but nevertheless simpler or better known. These may be models involving a change of scale such as the simple miniature replica placed in a wind tunnel, or the metal double helix model used by Watson and Crick to visualize DNA. In all these cases the model allows to visualize the object of study (Figure 1.4 A and B), to manipulate it and play with it to better understand or explain a phenomenon, just like the scientist with his orrery (Figure 1.1). In the case of biology, there are mainly model organisms such as drosophila, zebrafish or mice, for example. We then benefit from the relative simplicity of their genomes, a shorter time scale or ethical differences, usually to elucidate mechanisms of interest in humans. Correspondence between the target system and its model can sometimes be more conceptual, such as that ones relying on mechanical–electrical analogies: a mechanical system (e.g. a spring-mass system) can sometimes be represented by an electric network (e.g. a RLC circuit with a resistor, a capacitor and an inductor).

The model is then no longer simply a mimetic replica but is based on an intellectual equivalence: we are gradually moving into the realm of formal models [Rosenblueth and Wiener, 1945]. These are of a more symbolic

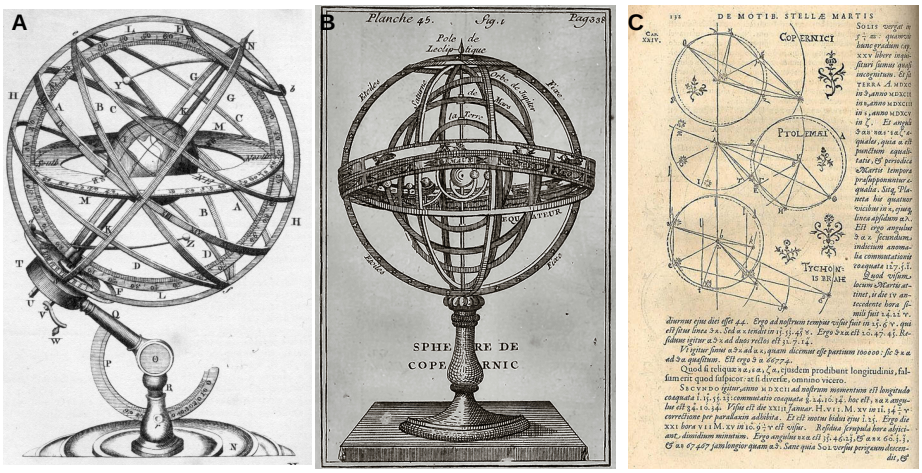

Figure 1.4: **Orrery, planets and models**. Physical models of planetary motion, either geocentric (Armillary sphere from *Plate LXXVII* in *Encyclopedia Britannica*, 1771) or heliocentric in panel B (Bion, 1751, catalogue Bnf) and some geometric representations by Johannes Kepler in panel C (in *Astronomia Nova*, 1609)

nature and they **represent the original system with a set of logical or mathematical terms**, describing the main driving forces or similar structural properties as geometrical models of planetary motions summarized by Kepler in Figure 1.4C. Historically these models have often been expressed by sets of mathematical equations or relationships. Increasingly, these have been implemented by computer. Despite their sometimes less analytical and more numerical nature, many so-called computational models could also belong to this category of formal models. There are then many formalisms, discrete or continuous, deterministic or stochastic, based on differential equations or Boolean algebra [Fowler et al., 1997]. Despite their more abstract nature, they offer similar scientific services: it is possible to play with their parameters, specifications or boundary conditions in order to better understand the phenomenon. One can also imagine these formal models from a different perspective, which starts from the data in a bottom-up approach instead of starting from the phenomenon in a top-down analysis. These models will then often be called statistical models or models of data [Frigg and Hartmann, 2020]. This distinction will be further clarified in section 1.2.

To summarize and continue a little longer with the astronomical metaphor, the study of a particularly complex system (the solar system) can be broken down into a variety of different models. Physical and

mechanical models such as armillary spheres (1.4A and B) make it possible to touch the object of study. In addition, we can observe the evolution of models which, when confronted with data, have progressed from a geocentric to a heliocentric representation to get closer to the current state of knowledge. Sometimes, models with more formal representations are used to give substance to ideas and hypotheses (1.4C). One of the most conceptual forms is then the mathematical language and one can thus consider that the previously mentioned astronomical models find their culmination in Kepler's equations about orbits, areas and periods that describe the elliptical motion of the planets. We refer to them today as Kepler's laws. The model has become a law and therefore a paragon of mathematical modeling [Wan, 2018].

1.1.3 Preview about cancer models

As we get closer to the subject of our study, and in order to illustrate these definitions more concretely, we can take an interest in the meaning of the word *model* in the context of cancer research. For this, we restrict our corpus to scientific articles found when searching for "cancer model" in the PubMed article database. Among these, we look at the occurrences of the word *model* and the sentences in which it is included. This cancer-related context of model is represented as a tree in Figure 1.5. Some of the distinctions already mentioned can be found here. The *mouse* and *xenograft* models, which will be discussed later in this thesis, represent some of the most common physical models in cancer studies. These are animal models in which the occurrence and mechanisms of cancer, usually induced by the biologist, are studied. On the other hand, *prediction*, *prognostic* or *risk score* models refer to formal models and borrow from statistical language.

Another way to classify cancer models may be to group them into the following categories: *in vivo*, *in vitro* and *in silico*. The first two clearly belong to the physical models but one uses whole living organisms (e.g. a human tumour implanted in an immunodeficient mouse) and the other separates the living from its organism in order to place it in a controlled environment (e.g. tumour cells in growth medium in a Petri dish). **In the thesis, data from both *in vivo* and *in vitro* models will be used. However, unless otherwise stated, a model will always refer to a representation *in silico*.** This third category, however, contains a very wide variety of models [Deisboeck et al., 2009], to which we will come back in chapter 3. A final ambiguity about the nature of the formal models used in this thesis needs to be clarified beforehand.

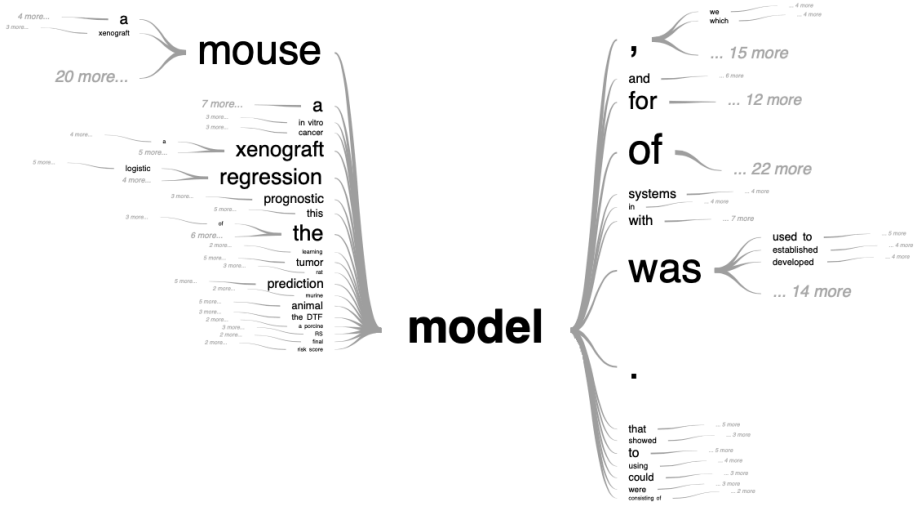

Figure 1.5: **Tree visualization of *model* semantic context in cancer-related literature** Generated with the 'PubTrees' tool by Ed Sperr, and based on most relevant PubMed entries for "cancer model" search.

1.2 Statistics or mechanistic

A rather frequent metaphor is to compare formal models to black boxes that take in input X predictors, or independent variables, and output response variable(s) Y, also named dependent variables. The models then split into two categories (Figure 1.6) depending on the answer to the question: are you modeling the inside of the box or not?

1.2.1 The inside of the box

The purpose of this section is to present in a schematic, and therefore somewhat caricatural, manner the two competing formal modeling approaches that will be used in this thesis and that we will call mechanistic modeling and statistical modeling. Assuming the unambiguous nature of the predictors and outputs we can imagine that the natural process consists in defining the result Y from the inputs X according to a function of a completely unknown form (Figure 1.6A).

The first modeling approach, that we will call **mechanistic**, consists in **building the box by imitating what we think is the process of data generation** (Figure 1.6B). This integration of a priori knowledge can take different forms. In this thesis it will often come back to presupposing certain relations between entities according to what is known about their

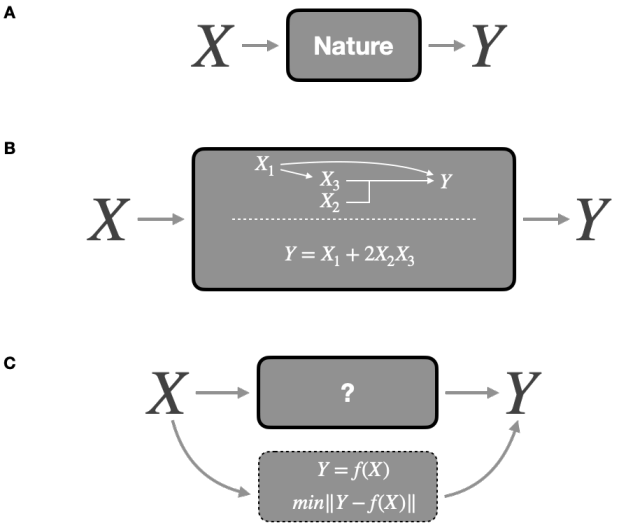

Figure 1.6: **Different modeling strategies.** (A) Data generation from predictors X to response Y in the natural phenomenon. (B) Mechanistic modeling defining mechanisms of data generation inside the box. (C) Statistical modeling finding the function f that gives the best predictions. Aadapted from Breiman [2001b].

behaviour. X_1 which acts on X_3 may correspond to the action of one biological entity on another, supposedly unidirectional; just as the joint action of X_2 and X_3 may reflect a known synergy in the expression of genes or the action of proteins. Mathematically this is expressed here with a perfectly deterministic model defined a priori. All in all, in a purely mechanistic approach, the nature of the relations between entities should be linked to biological processes and the parameters in the model all have biological definitions in such a way that it could even be considered to measure them directly. For example, the coefficient 2 multiplying X_2X_3 can correspond to a stoichiometric coefficient or a reaction constant which have a theoretical justification or are accessible by experimentation. In some fields of literature these models are sometimes called mathematical models because they propose a mathematical translation of a phenomenon, which does not start from the data in a bottom-up approach but rather from a top-down theoretical framework. In this thesis we will adhere to the *mechanistic model* name, which is more transparent and less ambiguous compared to other approaches also based on mathematics, without necessarily the other characteristics described above.

The second approach, often called **statistical modeling**, or sometimes

machine learning depending on the precise context and objective, does not necessarily seek to reproduce the natural process of data generation but to **find the function allowing the best prediction of Y from X** (Figure 1.6C). Pushed to the limit, they are "idealized version of the data we gain from immediate observation" [Frigg and Hartmann, 2020], thus providing a phenomenological description. The methods and algorithms used are then intended to be sufficiently flexible and to make the fewest possible assumptions about the relationships between variables or the distribution of data. Without listing them exhaustively, the approaches such as support vector machines [Cortes and Vapnik, 1995] or random forests [Breiman, 2001a], which will sometimes be mentioned in this thesis, fall into this category which contains many others [Hastie et al., 2009].

Several discrepancies result from this difference in nature between mechanistic and statistical models, some of which are summarized in the Table 1.1. In a somewhat schematic way, we can say that the mechanistic model first asks the question of *how* and then looks at the result for the output. The **notion of causality is intrinsic to the definition of the model**. Conversely, the statistical model first tries to approach the Y and then possibly analyses what can be deduced from it, regarding the importance of the variables or their relationships in a *post hoc* approach [Ishwaran, 2007, Manica et al., 2019]. The causality is then not a by-product of the algorithm and must be evaluated with dedicated frameworks [Hernán and Robins, 2020]. The greater flexibility of statistical methods makes it possible to better accept the heterogeneity of the variables, but this is generally done at the cost of a larger number of parameters and therefore requires more data. Moreover, statistical models can be considered as inductive, since they are able to use already generated data to identify patterns in it. Conversely, mechanistic models are more deductive in the sense that they can theoretically allow to extrapolate beyond the original data or knowledge used to build the model [Baker et al., 2018]. Finally, the most relevant way of assessing the value or adequacy of these models may be quite different. A statistical model is measured by its ability to predict output in a validation dataset different from the one used to train its parameters. The mechanistic model will also be evaluated on its capacity to approach the data but also to order it, to give a meaning. If its pure predictive performance is generally inferior, **how can the value of understanding be assessed?** This question will be one of the threads of the dissertation.

Mechanistic and statistical models are not perfectly exclusive and rather form the two ends of a spectrum. The definitions and classification of some examples is therefore still partly personal and arbitrary. For instance, the example in 1.6B can be transformed into a model with a more ambiguous

Table 1.1: **Some pros and cons for mechanistic and statistical modeling**. Adapted from Baker et al. [2018].

Mechanistic modeling	Statistical modeling
Definition	
Seeks to establish a mechanistic relationship between inputs and outputs	Seeks to establish statistical relationships between inputs and outputs
Pros and cons	
Presupposes and investigates causal links between the variables	Looks for patterns and establishes correlations between variables
Capable of handling small datasets	Requires large datasets
Once validated, can be used as a predictive tool in new situations possibly difficult to access through experimentation	Can only make predictions that relate to patterns within the data supplied
Difficult to accurately incorporate information from multiple space and time scales due to constrained specifications	Can tackle problems with multiple space and time scales thanks to flexible specifications
Evaluated on closeness to data and ability to make sense of it	Evaluated based on predictive performance

status:

$$logit(P[Y = 1]) = \beta_1 X_1 + \beta_{23} X_2 X_3$$

This model is deliberately ambiguous. As a logistic model, it is therefore naturally defined as a statistical model. But the definition of the interaction between X_2 and X_3 denotes a mechanistic presupposition. The very choice of a logistic and therefore parametric model could also result from a knowledge of the phenomenon, even if in practice it is often a default choice for a binary output. Finally, the nature of the parameters β_1 and β_{23} is likely to change the interpretation of the model. If they are deduced from the data and therefore optimized to fit Y as well as possible, one will think of a statistical model whose specification is nevertheless based on knowledge of the phenomenon. On the other hand, one could imagine that these parameters are taken from the biochemistry literature or other data. The model will then be more mechanistic. The boundary between these

models is further blurred by the different possibilities of combining these approaches and making them complementary [Baker et al., 2018, Salvucci et al., 2019a].

1.2.2 A tale of prey and predators

The following is a final general illustration of the concepts and procedures introduced with respect to statistical and mechanistic models through a famous and characteristic example: the Lotka-Volterra model of interactions between prey and predators. This model was, like many students, my first encounter with what could be called mathematical biology. The Italian mathematician Vito Volterra states this system for the first time studying the unexpected characteristics of fish populations in the Adriatic Sea after the First World War. Interestingly, Alfred Lotka, an American physicist deduced the exact same system independantly, starting from very generic process of redistribution of matter among the several components derived from law of mass action [Knuuttila and Loettgers, 2017]. A detailed description of their works and historical formulation can be found in original articles [Lotka, 1925, Volterra, 1926] or dedicated reviews [Knuuttila and Loettgers, 2017].

The general objective is to understand the evolution of the populations of a species of prey and its predator, reasonably isolated from outside intervention. Here we will use Canada lynx (*Lynx canadensis*) and snowshow hare (*Lepus americanus*) populations for which an illustrative data set exists [Hewitt, 1917]. In fact, commercial records listing the quantities of furs sold by trappers to the Canadian Hudson Bay Company may represent a proxy for the populations of these two species as represented in Figure 1.7A. Denoting the population of lynx $L(t)$ and the population of hare $H(t)$ it can be hypothesized that prey, in the absence of predators, would increase in population, while predators on their own would decline in the absence of preys. A prey/predator interaction term can then be added, which will positively impact predators and negatively impact prey. The system can then be formalized with the following differential questions with all coefficients $a_1, a_2, b_1, b_2 > 0$:

$$\frac{dH}{dt} = a_1 H - a_2 HT$$

$$\frac{dL}{dt} = -b_1 L + b_2 HL$$

$a_1 H$ represents the growth rate of the hare population (prey), i.e. the population grows in proportion to the population itself according to usual birth modeling. The main losses of hares are due to predation by lynx, as represented with a negative coefficient in the $-a_2 HT$ term. It is therefore assumed that a fixed percentage of prey-predator encounters will result in the death of the prey. Conversely, it is assumed that the growth of the lynx population depends primarily on the availability of food for all lynxes, summarized in the $b_2 HL$ term. In the absence of hares, the lynx population decreases, as denoted by the coefficient $-b_1 L$. Important features of mechanistic models are illustrated here: the equations are based on a priori knowledge or assumptions about the structure of the problem and the parameters of the model can be interpreted. a_1, for example, could correspond to the frequency of litters among hares and the number of offspring per litter.

This being said, the structure of the model having been defined a priori, it remains to determine its parameters. Two options would theoretically be possible: to propose values based on the interpretation of the parameters and ecological knowledge, or to fit the model to the data in order to find the best parameters. For the sake of simplicity, and because this example has only a pedagogical value in this presentation, we propose to determine them approximately using the following Taylor-based approximation:

$$\frac{1}{y(t)}\frac{dy}{dt} \simeq \frac{1}{y(t)}\frac{y(t+1)-y(t-1)}{2}$$

By applying this approximation to the two equations of the differential system and plotting the corresponding linear regressions (Figures 1.7B and C), we can obtain an evaluation of the parameters such as $a_1 = 0.82$, $a_2 = 0.0298$, $b_1 = 0.509$, $b_2 = 0.0129$. By matching the initial conditions to the data, the differential system can then be fully determined and solved numerically (Figures 1.7D). Comparison of data and modeling provides a good illustration of the virtues and weaknesses of a mechanistic model. Firstly, based on explicit and interpretable hypotheses, the model was able to recover the cyclical behaviour and dependencies between the two species: the increase in the lynx population always seems to be preceded by the increase in the hare population. However, the amplitude of the oscillations and their periods are not exactly those observed in the data. This may be related to approximations in the evaluation of parameters, random variation in the data or, of course, simplifications or errors in the structure of the model itself.

Besides, if one tries to carry out a statistical modeling of these data, it is very likely that it is possible to approach the curve of populations evolution

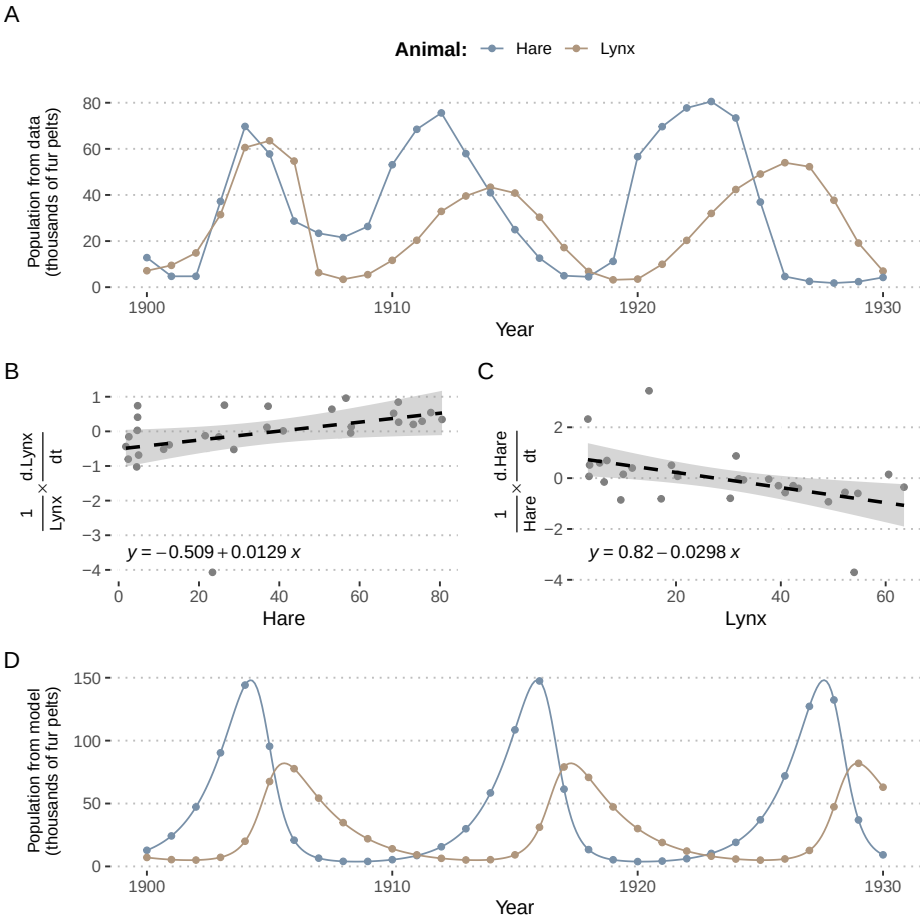

Figure 1.7: **Some analyses around Lotka-Volterra model of a prey-predator system**. (A) Evolution of lynx and hares populations based on Hudson Bay Company data about fur pelts. (B) and (C) Linear regression for estimation of parameters. (D) Evolution of lynx and hare populations as predicted by the model based on inferred parameters and initial conditions.

much closer, especially for the hares. But should it be expressed simply as a function of time or should a joint modeling be proposed? The nature of the causal link between prey and predators will be extremely difficult to establish without strong hypotheses such as those of the mechanistic model. On the other hand, if populations in later years had to be predicted as accurately as possible, it is likely that a sufficiently well-trained statistical model would perform better. Finally, and this is a fundamental difference, the **mechanistic model enables to test cases or hypotheses that go beyond the scope of the data**. Quite simply, by playing with the variables or parameters of the model, we can predict the exponential decrease of predators in the absence of prey and the exponential growth of prey in the absence of prey. More generally, it is also possible to study analytically or numerically the bifurcation points of the system in order to determine the families of behaviours according to the relative values of the parameters [Flake, 1998]. It is not possible to infer these new or hypothetical behaviours directly from the data o of the statistical model. This is theoretically possible on the basis of the mechanistic model, provided that it is sufficiently relevant and that its operating hypotheses cover the cases under investigation. Now that the value of mechanistic models has been illustrated in a fairly theoretical example, all that remains is to explore in the next chapters how they can be built and used in the context of cancer.

1.3 Simplicity is the ultimate sophistication

Before concluding this modeling introduction, it is important to highlight one of the most important points already introduced in a concise manner by the poet Paul Valéry at the beginning of this chapter. **Whatever its nature, a model is always a simplified representation of reality and by extension is always wrong to a certain extent**. This is a generally well-accepted fact, but it is crucial to understand the implications for the modeller. This simplification is not a collateral effect but an intrinsic feature of any model:

> *No substantial part of the universe is so simple that it can be grasped and controlled without abstraction. Abstraction consists in replacing the part of the universe under consideration by a model of similar but simpler structure. Models, formal and intellectual on the one hand, or material on the other, are thus a central necessity of scientific procedure.*
> [Rosenblueth and Wiener, 1945]

Therefore, a model exists only because we are not able to deal directly with the phenomenon and simplification is a necessity to make it more tractable [Potochnik, 2017]. This simplification appeared many times in the studies of frictionless planes or theoretically isolated systems, in a totally deliberate strategy. However, this idealization can be viewed in several ways [weisberg2007three]. One of them, called Aristotelian or minimal idealization, is to eliminate all the properties of an object that we think are not relevant to the problem in question. This amounts to lying by omission or making assumptions of insignificance by focusing on key causal factors only [Frigg and Hartmann, 2020]. We therefore refer to the *a priori* idea that we have of the phenomenon. The other idealization, called Galilean, is to deliberately distort the theory to make it tractable as explicited by Galileo himself:

> *We are trying to investigate what would happen to moveables very diverse in weight, in a medium quite devoid of resistance, so that the whole difference of speed existing between these moveables would have to be referred to inequality of weight alone. Since we lack such a space, let us (instead) observe what happens in the thinnest and least resistant media, comparing this with what happens in others less thin and more resistant.*

This fairly pragmatic approach should make it possible to evolve iteratively, reducing distortions as and when possible. This could involve the addition of other species or human intervention into the Lotka-Volterra system described above. A three-species Lotka-Volterra model can however become chaotic [Flake, 1998], and therefore extremely difficult to use and interpret, thus underlining the importance of simplifying the model.

We will have the opportunity to come back to the idealizations made in the course of the cancer models but it is already possible to give some orientations. The biologist who seeks to study cancer using cell lines or animal models is clearly part of Galileo's lineage. The mathematical or *in silico* modeler has a more balanced profile. The design of qualitative mechanistic models based on prior knowledge, which is the core of the second part of the thesis, is more akin to minimal idealization, which seeks to highlight the salient features of a system. But the Galilean pragmatism consisting in creating computationnaly-tractable models is also quite widespread, particularly in highly dimensional statistical approaches.

Because of the complexity of the phenomena, simplification is therefore a necessity. The objective then should not necessarily be to make the model more complex, but to **match its level of simplification with its**

assumptions and objectives. Faced with the temptation of the author of the model, or his reviewer, to always extend and complicate the model, it could be replied with Lewis Carrol words[1]:

> *"That's another thing we've learned from your Nation," said Mein Herr, "map-making. But we've carried it much further than you. What do you consider the largest map that would be really useful?"*
> *"About six inches to the mile."*
> *"Only six inches!" exclaimed Mein Herr. "We very soon got to six yards to the mile. Then we tried a hundred yards to the mile. And then came the grandest idea of all! We actually made a map of the country, on the scale of a mile to the mile!"*
> *"Have you used it much?" I enquired.*
> *"It has never been spread out, yet," said Mein Herr: "the farmers objected: they said it would cover the whole country, and shut out the sunlight! So we now use the country itself, as its own map, and I assure you it does nearly as well."*
> Lewis Carroll, *Sylvie and Bruno* (1893)

[1]More concisely stated by Rosenblueth and Wiener [1945]: "best material model for a cat is another cat, or preferably the same cat."

CHAPTER 2

Cancer as deregulation of complex machinery

"Does not the entireness of the complex hint at the perfection of the simple?"

Edgar Allan Poe (Eureka)

Armed with all these models, whether statistical or mechanistic, we are going to look at cancer, a particularly complex system that fully justifies their use. Since the first chapter recalled how important prior knowledge of the phenomenon under study is for designing models, whatever their nature, this chapter will briefly summarize some of the most important characteristics of this disease before returning to the models themselves in the next chapter. Without aiming for exhaustiveness, and after an epidemiological and statistical description, we will focus on the most useful information for the modeller, i.e. the underlying biological mechanisms and available data.

2.1 What is cancer?

Cancer can be described as a group of diseases characterized by **uncontrolled cell divisions and growth which can spread to surrounding**

Figure 2.1: **Cancer is an old disease.** Rembrandt, *Bathsheba at Her Bath*, c. 1654, oil on canvas, Louvre Museum, Paris

tissues. Descriptions of this disease, especially when associated with solid tumours, have been found as far back as ancient Egyptian documents, at least 1600 BC and we know from the first century A.D. with Aulus Celsus that it is better to remove the tumors and this as soon as possible [Hajdu, 2011a]. Progress will accelerate during the Renaissance with the renewed interest in medicine, and anatomy in particular, which will advance the knowledge of tumour pathology and surgery [Hajdu, 2011b]. The progress of anatomical knowledge has also left brilliant testimonies in the field of painting, which make the renown of the Renaissance today. The precision of these artists' traits has also allowed some retrospective medical analyses, some of them going so far as to identify the signs of a tumour in some of the subjects of these paintings [Bianucci et al., 2018]. Such is the

bluish stain on the left breast of the Bathsheba painted by Rembrandt (Figure 2.1) which has been subject to controversial interpretations, sometimes described as an example of "skin discolouration, distortion of symmetry with axillary fullness and peau d'orange" [Braithwaite and Shugg, 1983] and sometimes spared by photonic and computationnal analyses [Heijblom et al., 2014]. The mechanisms of the disease only began to be elucidated with the appearance of the microscope in the 19th century, which revealed its cellular origin [Hajdu, 2012a]. The classification and description of cancers is then gradually refined and the first non-surgical treatments appear with the discovery of ionising radiation by the Curies [Hajdu, 2012b]. The 20th century is then the century of understanding the causes of cancer [Hajdu and Darvishian, 2013, Hajdu and Vadmal, 2013]. Some environmental exposures are characterized as asbestos or tobacco. Finally, the biological mechanisms become clearer with the identification of tumour-causing viruses and especially with the discovery of DNA [Watson and Crick, 1953]. The foundations of our current understanding of cancer date back to this period, which marks the beginning of the molecular biology of cancer. It is this branch of biology that contains the bulk of the knowledge that will be used to build our mechanistic models, and it will be later detailed in Section 2.3.

One of the ways to read this brief history of cancer is to see that theoretical and clinical progress has not followed the same timeframes.The medical and clinical management of cancers initially progressed slowly but surely, and this in the absence of an understanding of the mechanisms of cancer. Conversely, the theoretical progress of the last century has not always led to parallel medical progress, except on certain specific points. The interaction between the two is therefore not always obvious. The **transformation of fundamental knowledge into medical and clinical impact is therefore of particular importance**. This is what is called *translational medicine*, the aim of which is to go from laboratory bench to bedside [Cohrs et al., 2015]. It is in this perspective that we will analyze the mechanistic models studied in this thesis. Their objective is to integrate biological knowledge, or at least a synthesis this knowledge, in order to transform it into a relevant clinical information.

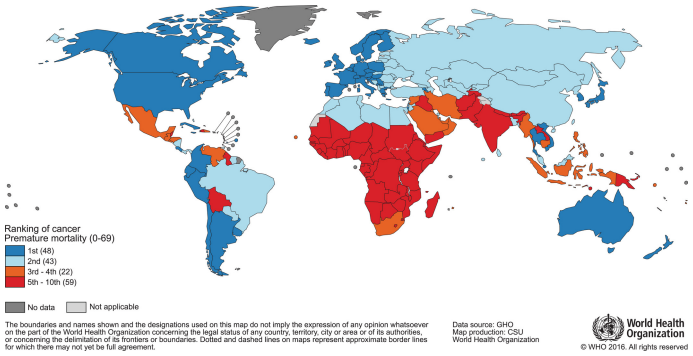

Figure 2.2: **World map and national rankings of cancer as a cause of premature death.** Classification of cancer as a cause of death before the age of 70, based on data for the year 2015. Original Figure, data and methods from Bray et al. [2018].

2.2 Cancer from a distance: epidemiology and main figures

Before going down to the molecular level, it is important to detail some figures and trends in the epidemiology of cancer today. Following the description in the previous section, cancer is first and foremost defined as a disease. Considered to be a unique disease, it caused 18.1 million new cancer cases and 9.6 million cancer deaths in 2018 according to the Global Cancer Observatory affiliated to World Health Organization [Bray et al., 2018]. However, these aggregated data conceal disparities of various kinds. The first one is geographical. Indeed, mortality figures make cancer one of the leading causes of premature death in most countries of the world but its importance relative to other causes of death is even greater in the more developed countries (Figure 2.2). All in all, cancer is the first or second cause of premature death in almost 100 countries worldwide [Bray et al., 2018]. These differences call for careful consideration of the impact of population age structures and health-related covariates.

A second disparity lies in the different types of cancer. If we classify tumours solely according to their location, i.e. the organ affected first, we already obtain very wide differences. First of all, the incidence varies considerably (Figure 2.3A)). Cancers do not occur randomly anywhere in the body and certain environments or cell types appear to be more favourable [Tomasetti and Vogelstein, 2015]. Mortality is also highly variable but is not directly inferred from incidence. Not all types of cancer have the same prognosis (Figure 2.3A and B) and survival rates [Liu et al., 2018]. Although

breast cancer is much more common than lung cancer, it causes fewer deaths because its prognosis is, on average, much better. The mechanisms at work in the emergence of cancer are therefore not necessarily the same as those that will govern its evolution or its response to treatment. And still on the response to treatment, Figure 2.3B highlights another disparity: not only are the survival prognosis associated with each cancer very different, but the evolution (and generally the improvement) of these prognoses has been very uneven over the last few decades. This means that theoretical and therapeutic advances have not been applied to all types of cancer with the same success. It is one more indication of the **diversity of cancer mechanisms in different tissues and biological contexts**, which make it impossible to find a panacea, and which, on the contrary, encourage us to carefully consider the particularities of each tumour, both to understand them and to treat them. Under a generic name and in spite of common characteristics, the cancers thus appear as extremely heterogeneous. And to understand the sources of this heterogeneity, it is necessary to consider the disease on a smaller scale.

2.3 Basic molecular biology and cancer

If it is not possible and desirable to summarize here the state of knowledge about the biology of cancer, we are going to give a very partial vision focused on the main elements used in this thesis, thus aiming to make it a self-sufficient document. The details necessary for a finer and more general understanding can be found in dedicated textbooks such as Alberts et al. [2007] and Weinberg [2013].

2.3.1 Central dogma and core principles

Some of the principles that govern biology can be described at the level of one of its simplest element, the cell. Let us consider for the moment a perfectly healthy cell. It must ensure a certain number of functions necessary for its survival and, if necessary, for its division/reproduction. These functions are encoded in its genetic information in the form of DNA, which is stable and shared by the different cells since it is defined at the level of the individual. Most biological functions, however, are not performed by DNA itself which remains in the nucleus of the cell. The DNA is thus transcribed into RNA, another nucleic acid which, in addition to performing some biological functions, becomes the support of the genetic information in the cell. The RNA is then itself translated into new molecules composed of long

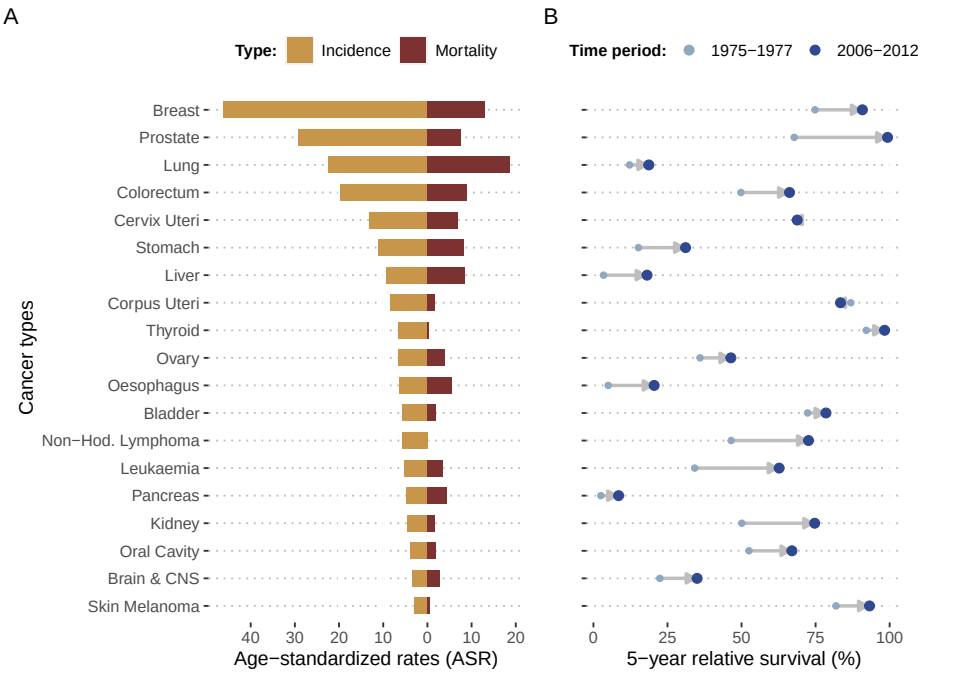

Figure 2.3: **Incidence, mortality and survival per cancer types**. (A) World incidence and mortality for the 19 most frequent cancer types in 2018, expressed with age-standardized rates (adjusted age structure based on world population); data retrieved from Global Cancer Observatory. (B) Evolution of 5-years relative survival for the same cancer types based on US data from SEER registries in 1975-1977 and 2006-2012; data retrieved from Jemal et al. [2017].

chains of amino acid residues and called proteins. They are the ones that execute most of the numerous cellular functions: DNA replication, physical structuring of the cell, molecule transport within the cell etc. A rather simplistic but fruitful way to understand this functioning is to consider it as a **progressive transfer of biological information from DNA to proteins**, which has sometimes been summarized as the central dogma of the molecular biology (2.4), first stated Francis Crick [Crick, 1970].

However, many changes would be necessary to clarify this scheme and the uni-directional nature was questioned early on. Above all, a large number of regulations interact with and disrupt this master plan. The genes are not always all transcribed, or at least not at constant intensities, interrupting or varying the chain upstream. This modulation in the transcription of genes can be induced by proteins, called transcription factors. After a gene transcription, its expression can still be regulated at various stages. RNAs

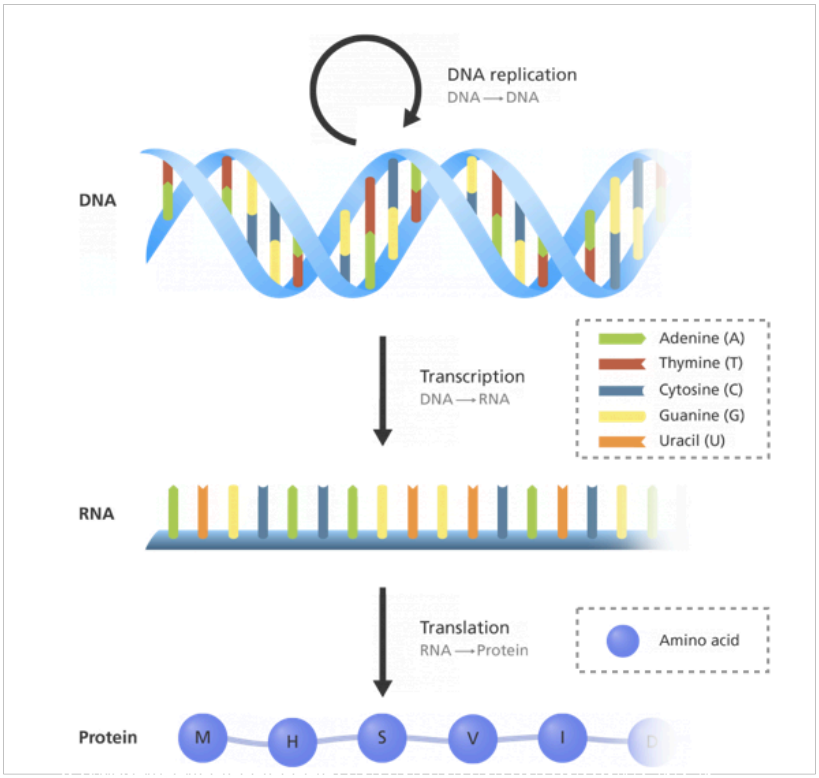

Figure 2.4: **Central dogma of molecular biology.** Schematic representation of the information flow within the cell, from DNA to proteins through RNA, more precisely described in this video (Image credit *Genome Research Limited*).

can also be degraded more or less rapidly. RNAs can be reshaped in their structure by a process called splicing, which varies the genetic information they carry. Finally, proteins are subject to all kinds of modifications referred to as post-translational, which can change the chemical nature of certain groups or modify the three-dimensional structure of the whole protein. For instance, some proteins perform their function only if a specific amino acid residue is phosphorylated. In addition, these modifications can be transmitted between proteins, further complicating the flow of information. **All these possibilities of regulation play an absolutely essential role in the life of the cell by allowing it to adapt to different contexts and situations**. From the same genetic material, a cell of the eye and a cell of the heart can thus perform different functions. Similarly, the same cell subjected to different stimuli at different times can provide different responses because these molecular stimuli trigger a regulation of

its programme. But all these regulatory mechanisms can be corrupted.

2.3.2 A rogue machinery

With the above knowledge we can now return to the definition of cancer as an uncontrolled division of cells that can lead to the growth of a tumour that eventually spreads to the surrounding tissues. Therefore, this corresponds to normal processes, like cell division and reproduction, that are no longer regulated as they should be and are out of control. Experiments on different model organisms have gradually identified genetic mutations as a major source of these deregulations [Nowell, 1976, Reddy et al. [1982]] until cancer was clearly considered as a **genetic disease** making Renato Dulbecco, Nobel Laureate in Medicine for his work on oncoviruses, say:

> *If we wish to learn more about cancer, we must now concentrate on the cellular genome.*
> [Dulbecco, 1986].

However, cancer is not a Mendelian disease for which it would be sufficient to identify the one and only gene responsible for deregulation. Indeed, the cell has many protective mechanisms. For example, if a genetic mutation appears in the DNA, it has a very high chance of being repaired by dedicated mechanisms. And if it is not repaired, other mechanisms will take over to trigger the programmed death of the cell, called apoptosis, before it can proliferate wildly. So a cancer cell is probably a cell that has learned to resist this cell death. Similarly, in order to generate excessive growth, a cell will need to be able to replicate itself many times. However, there are pieces of sequences on chromosomes called telomeres that help to limit the number of times each cell can replicate. A cancer cell will therefore have to manage to bypass this protection. Thus we can schematically define the properties that must be acquired by the cancereous cells in order to truly deviate the machinery. In an influential article, these properties were summarized in six hallmarks (Figure 2.5) which are: resisting cell death, enabling replicatve immortality, sustaning proliferative signaling, evading growth suppressors, activating invasion and inducing angiogenesis [Hanahan and Weinberg, 2000]. Two new ones were subsequently added in the light of advances in knowledge [Hanahan and Weinberg, 2011]: deregulating cancer energetics and avoiding immne destruction. The acquisition of these capacities generally requires many genetic mutations and is therefore favoured by an underlying genome instability.

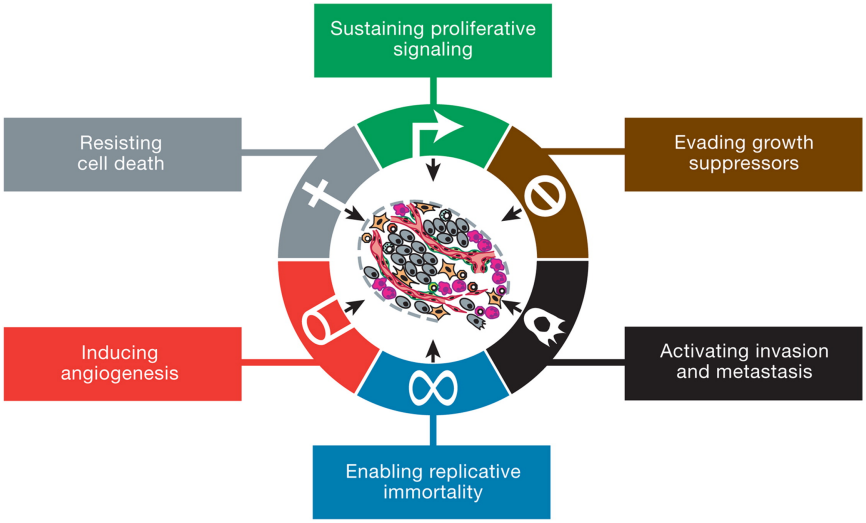

Figure 2.5: **Hallmarks of cancer.** The different biological capabilities acquired by cancer cells, as described in Hanahan and Weinberg [2000]. Reprinted from Hanahan and Weinberg [2011].

Each of these characteristics, or hallmarks, constitutes a research program in its own right. And for each one there are genetic alterations. These are tissue-specific or not, specific to a hallmark or common to several of them [Hanahan and Weinberg, 2000]. In any case, **cancer can only result from the combination of different alterations that invalidate several protective mechanisms** at the same time. This is often part of a multi-step process of hallmark acquisition that has been experimentally documented in some specific cases [Hahn et al., 1999] or more recently inferred from genome-wide data for human patients [Tomasetti et al., 2015]. In summary, it appears that in order to study the functioning of cancer cells it is necessary to look at several mechanisms and to be able to consider them not separately but together, in as many different patients as possible. This ambitious programme has been made possible by a technological revolution.

2.4 The new era of genomics

2.4.1 From sequencing to multi-omics data

In 2001, the first sequencing of the human genome symbolized the beginning of a new era, that of what will become **high-throughput genomics** [Lander et al., 2001, Venter et al., 2001]. From the end of the 20th century,

biological data started to accumulate at an ever-increasing rate [Reuter et al., 2015], feeding and accelerating cancer research in particular [Stratton et al., 2009, Meyerson et al., 2010]. The ability to sequence the human genome as a whole, for an ever-increasing number of individuals, has enabled **less biased and more systematic studies of the causes of cancer** [Lander, 2011]. The number of genes associated with cancer increased drastically and some very important genes such as BRAF of PIK3CA have been identified [Davies et al., 2002, Samuels et al., 2004]. Progress also extended to the gene expression data. Gene-expression arrays have made an important contribution by providing access to transcriptomic data (RNA), i.e. what has been transcribed from DNA and is therefore one step further in terms of biological information. This information has made it possible to further explore the differences betwween normal and tumour cells [Perou et al., 1999], or even to refine the classification of cancers, which until now has been done mainly according to the tumour site. Breast cancers are thus divided into subtypes with different combinations of molecular markers that facilitate the understanding of clinical behavior [Perou et al., 2000]. One step further, we also note the appearance of prognostic gene signatures such as gene expression patterns correlated with the survival of patients [Van't Veer et al., 2002]. This revolution was then extended to other types of data such as proteins (proteomics), reversible modifications of DNA or DNA-associated proteins (epigenomics), metabolites (metabolomics) and others, each representing a perspective that can complement the others to better understand biological mechanisms, particularly in the case of diseases [Hasin et al., 2017]. We have thus entered the era of multi-omics data [Vucic et al., 2012].

2.4.2 State-of-the art of cancer data

With respect to cancer in particular, this wealth of data is particularly represented by a family of **studies conducted by The Cancer Genome Atlas (TCGA) consortium**, started in 2008 [Network et al., 2008]. Cohorts of several hundred patients are thus sequenced over the years for different types of cancer [Network et al., 2012], resulting today in a total of 11,000 tumors from 33 of the most prevalent forms of cancer [Ding et al., 2018]. Figure 2.6 provides a partial but striking overview of the depth of data available under this program. We can see the frequencies of alterations of certain groups of genes for a list of cancer types, making it possible to visualize the disparities already anticipated in section 2.2 based on patient survival. There are indeed important differences between the organs but also between the different subtypes associated with the same organ. And

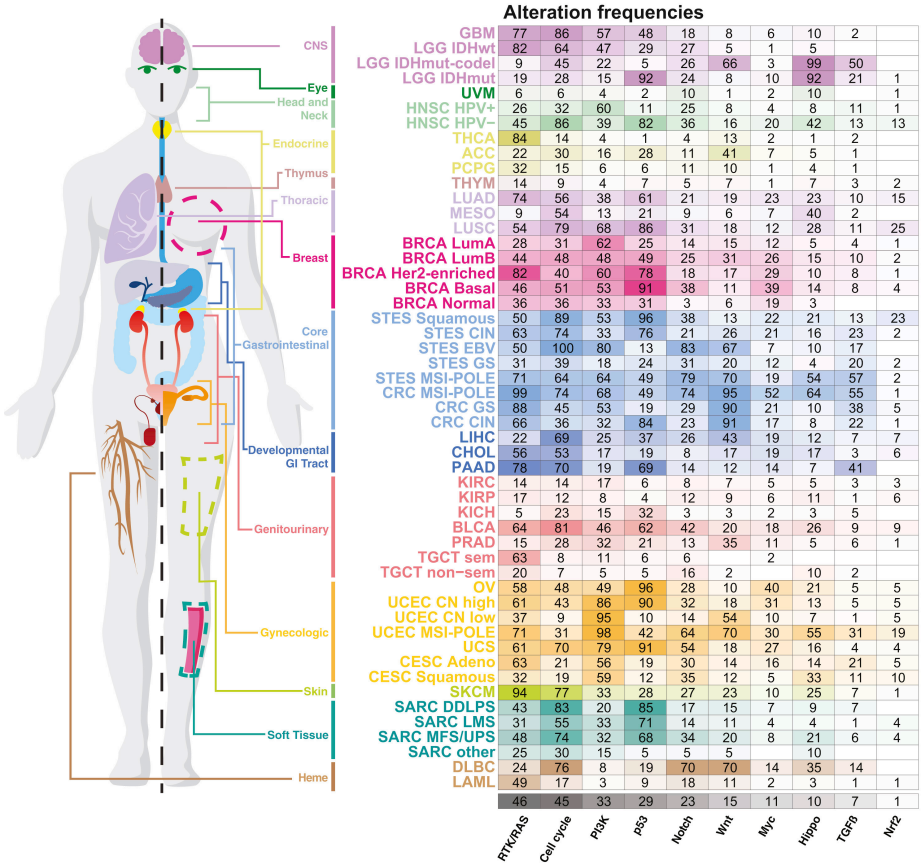

Figure 2.6: **Genetic alterations frequencies for cancer types from TCGA data.** Frequencies of alteration per pahway and tumour types as summaried in Pan-cancer analyses from TCGA data. Reprinted from Sanchez-Vega et al. [2018].

this representation only corresponds to one layer of data, that of genetic alterations. It could be used for transcriptomic, epigenomic or proteomic data, thus giving rise to an incredibly complex photography.

However, the diversity of data available for cancer research extends far beyond this, both in terms of technology and type of data. This may be data from model organisms such as mice or tumours of human origin made more suitable for experimentation. In the latter category, it is crucial to mention the **huge amount of data available on cell lines**, extracted from human tumours and transformed to be studied in culture. It is then possible to go beyond descriptive data and vary the experimental conditions in order to study the responses of these cells to perturbations and to enrich

our knowledge. This provides an opportunity to know the response to more than 100 drugs of about 700 cell lines [Yang et al., 2012]. The richness of these data, coupled with the omic profiling of each cell line, enables to study the determinants of response to treatment with unprecedented scope [Iorio et al., 2016]. More recently, but following a similar logic, other types of inhibition screenings have been proposed based on a more specific technique called CRISPR-Cas9 [Behan et al., 2019]. The simplicity of the cell lines in relation to the original tumours makes all these studies possible but sometimes hinders the clinical application of the knowledge acquired. For this reason, other types of biological models have been developed, including patient-derived xenografts (PDX) which is an implant of human tumours in mice to maintain the existence of a certain tumour microenvironment [Hidalgo et al., 2014], while maintaining drug screening possibilities [Gao et al., 2015]. These two types of data, cell lines and PDX, have been used in this thesis, in addition to TCGA patient data, thus justifying the limitation of this presentation, which could otherwise be extended to other types of biological models. Similarly, other technologies are becoming increasingly important in the generation of cancer data, such as single-cell sequencing [Navin, 2015], but will not be used in this work.

2.5 Data and beyond: from genetic to network disease

All that remains to be done now is to make sense of all these data, to organize it, because **cancer understanding does not flow directly from the abundance of data**, and the ability to produce it may have been outpaced by the ability to analyze it [Stadler et al., 2014]. A striking example is that of the prognostic signatures mentioned above. The many signatures or lists of genes proposed, even for the same cancer type, share relatively few genes, are difficult to interpret and their efficiency is sometimes poorly reproducible [Domany, 2014]. Even more surprisingly, most signatures composed of randomly selected genes were also found to be associated with patient survival [Venet et al., 2011]. One of the main avenues for improving the interpretability of the data is the **integration of the prior knowledge** we have of the phenomena, especially in the case of cancer [Domany, 2014].

This *a priori* knowledge is in fact already present in Figure 2.6 since genetic alterations have been grouped in several categories called pathways. A pathway is group of biological entities, and associated chemical reactions,

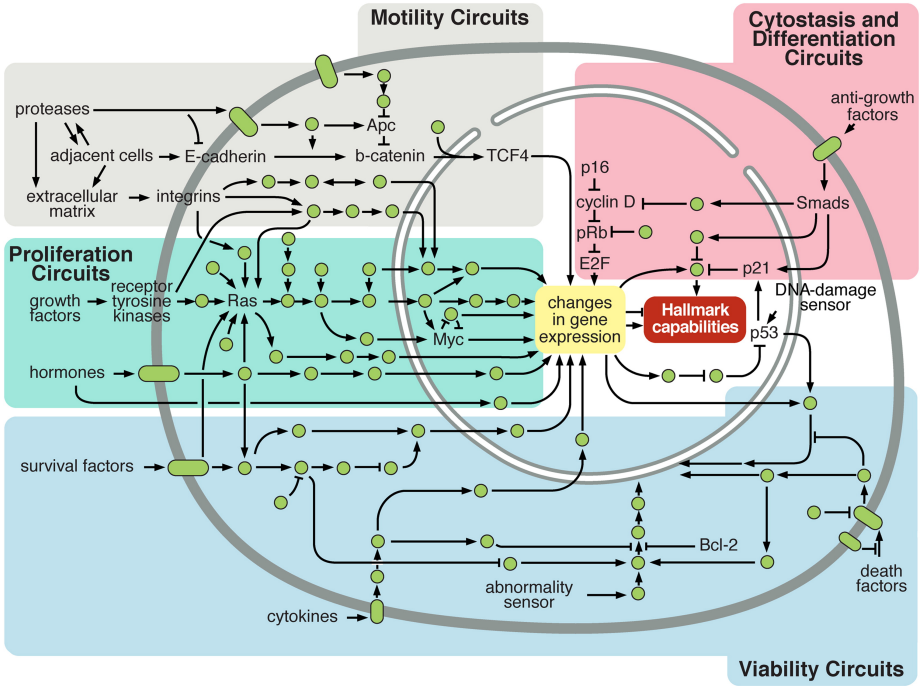

Figure 2.7: **Simplistic representation of cellular circuitry.** Normal cellular circuit sand sub-circuits (identified by colours) can be reprogrammed to regulate hallmark capabilities within cancer cells. Reprinted from Hanahan and Weinberg [2011].

working together to control a specific cell function like apoptosis or cell division. The interest of these groupings may be understood based on the description of hallmarks. Indeed, if the "aim" of a cancer cell is to inactivate each of the protective functions, then it is more relevant to think not by gene but by function. Inactivating only one of the genes associated with the function may be sufficient and it is no longer necessary to inactivate the others. Numerous alterations in a large number of genes in various patients result often in the same key impaired pathways, like alterations of cell cycle or angiogenesis for instance [Jones et al., 2008]. It is therefore possible to improve the stability and interpretability of analyses by moving **from the gene scale to the pathway scale** [Drier et al., 2013]. More generally, the integration of biological knowledge often leads to improved performance in various cancer-related prediction tasks, either through the selection of variables or by taking into account the structure of the variables [Bilal et al., 2013, Ferranti et al., 2017]. Increasingly, the biological variables are not interpreted separately but in relation to each other [Barabasi and

Oltvai, 2004]. This is reflected in the emergence of more and more resources to summarize and represent signaling pathways and associated networks such as SIGNOR [Perfetto et al., 2016], OmniPath [Türei et al., 2016] or the Atlas of Cancer Signaling Network [Kuperstein et al., 2015]. Like other diseases, cancer then goes **from a genetic disease to a network disease** [Del Sol et al., 2010] and one can study how all kinds of genetic alterations affect the wiring of these networks [Pawson and Warner, 2007], and modify the cellular functions leading to the previously described cancer hallmarks as depicted schmatically in Figure 2.7. In short, the richness of the data did not make it less necessary to use prior knowledge in order to make the analyses more interpretable and more robust.

The final step, to obtain one of the most complete and integrated visions of cancer biology, is then to integrate omics knowledge with knowledge about the structure of pathways to try to understand in detail how their combinations can lead to so many cancers that are both similar and different. An example of such a representation is given by mapping the TCGA data about genetic alterations, presented in Figure 2.6, on a representation of the different pathways showing not only their internal organization but also their cross-talk [Sanchez-Vega et al., 2018]. This representation is proposed in Figure 2.8 and is one the most recent and comprehensive view of the kind of tools and data available to the modeller who wants to dissect more deeply the mechanisms involved in cancer.

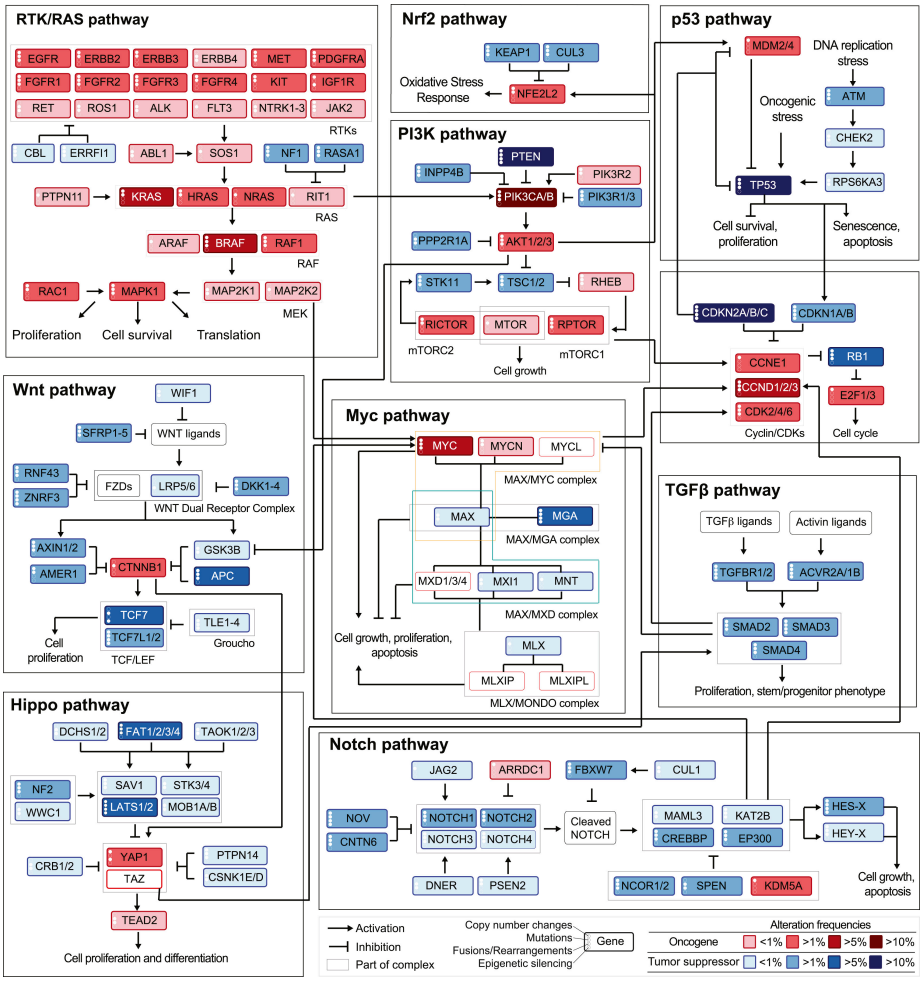

Figure 2.8: **Genetic alterations frequencies from TCGA data mapped on a schematic signaling network.** Frequencies of alteration per pathway and tumour types as summarized in Pan-cancer analyses from TCGA data. Reprinted from Sanchez-Vega et al. [2018].

CHAPTER 3

Mechanistic modeling of cancer: from complex disease to systems biology

"How remarkable is life? The answer is: very. Those of us who deal in networks of chemical reactions know of nothing like it... How could a chemical sludge become a rose, even with billions of years to try."

George Whitesides

The previous chapter identified the need to organize cancer knowledge and data. The integration of biological knowledge, particularly in the form of networks, is a first step in this direction. The deepening of knowledge, however, requires the ability to manipulate objects even more, to experiment, to dissect their behaviour in an infinite number of situations, such as the astronomer with his orrery or physicians with their old anatomical models (Figure 3.1). Is it then possible to create mechanistic models of cancer in the same way?

3.1 Introducing the diversity of mechanistic models of cancer

Modeling cancer is not a new idea. And the diversity of biological phenomena involved in cancer has given rise to an equally important diversity of

Figure 3.1: **Dissecting a biological phenomenon using a non-computational model.** Rembrandt, *The Anatomy Lesson of Dr Nicolaes Tulp*, 1634, oil on canvas, Mauritshuis museum, The Hague

models and formalisms, which we seek here to give a brief overview in order to better identify the specific models that we will focus on later. One way to order this diversity is to consider the scales of these models (Figure 3.2). Indeed, **cancer can be read at different levels, from the molecular level of DNA and proteins, to the cellular level, to the level of tissues and organisms** [Anderson and Quaranta, 2008]. Models have been proposed at all these scales, using different formalisms [Bellomo et al., 2008] and answering different questions.

Consistent with the evolution of knowledge and data, the early models were at the **macroscopic level**. While methods and terminologies may have changed, there are nevertheless traces of these models as early as the 1950s. We then speak rather of mathematical modeling with a meaning that is nevertheless intermediate between what we have defined as mechanistic models and statistical models [Byrne, 2010]. First, the initiation of tumorigenesis was theorized with biologically-supported mathematical expressions in order to make sense of cancer incidence statistics [Armitage and Doll, 1954, Knudson [1971]]. These models, however, remained relatively

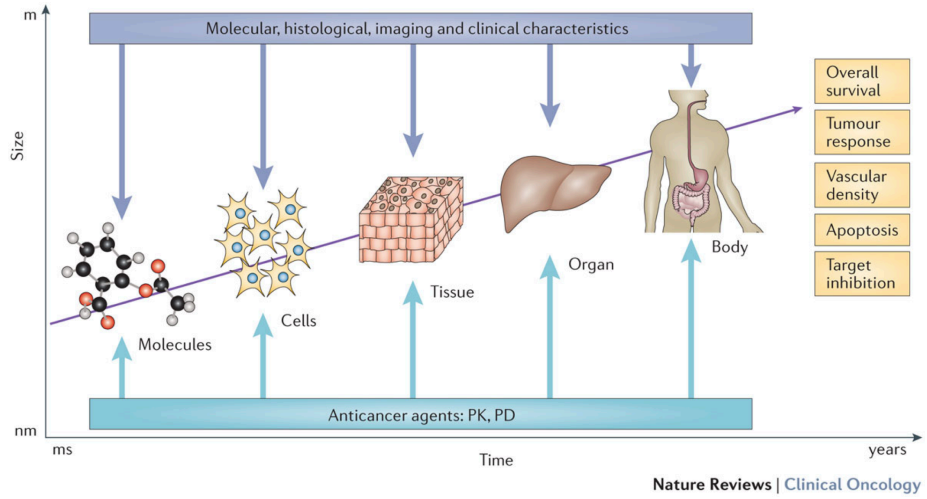

Figure 3.2: **The different scales of cancer modeling.** Cancer can be approached at different scales, from molecules to organs, using different data (dark blue), but often with the direct or indirect objective of contributing to the study of clinically interpretable phenomena (yellow boxes), in particular by studying the influence of anticancer agents (pale blue). Reprinted from Barbolosi et al. [2016].

descriptive in that they did not shed any particular light on the biological mechanisms involved and focused on gross characteristics of tumours. The integration of more advanced knowledge as well as the progressive refinement of mathematical formalisms has nevertheless allowed these models to proliferate while gaining in interpretability, with for instance mechanistic models of metastatic relapse [Nicolò et al., 2020]. Always on a macroscopic scale, the study of tumor growth has also been the playground of many mathematicians [Araujo and McElwain [2004]; byrne2010dissecting], even predicting invasion or response to surgical treatments using spatial modeling [Swanson et al., 2003]. This line of research is still quite active today and provides a mathematical basis for comparison with tumour experimental growth [Benzekry et al., 2014].

Taking it down a step further, it is also possible to model cancer at the **cellular level**, for example by looking at the clonal evolution of cancer [Altrock et al., 2015]. The aim is then to understand the impact of the processes of mutation, selection, expansion and cohabitation of different populatons of cells, at specifc rates. The accumulation of a mutation in a population of cells can thus be studied [Bozic et al., 2010]. Modeling at the cellular level is well suited to the study of interactions between cells,

between cancer cells and their environment or with the immune system. Similar to other kinds of studies of population dynamics, formalisms based on differential equations are quite common [Bellomo et al., 2008]; but there are many other methods such partial differential equations or agent-based modeling [Letort et al., 2019].

Finally, at an even smaller scale, it is possible to model the **molecular networks** at work in cells [Le Novere, 2015]. The aim is then to simulate mathematically how the different genes and molecules regulate each other, transmit information and, in the case of cancer, end up being deregulated [Calzone et al., 2010]. These models will be the subject of the thesis and will therefore be defined more precisely and used to detail the concepts and tools of systems biology in the following sections. It can already be noted that while these models can integrate the most fundamental biological mechanisms of living organisms, one of the most burning questions is whether it is possible to link them to the larger scales that are clinically more interesting (tissues, organs etc.). Can these models tell us something about the molecular nature of cancer? About patient survival? Their response to treatment? These questions apply to all of the above models, whatever their scales (Figure 3.2), but are more difficult to answer for models defined at molecular scale that are further from the clinical data of interest. The aim of this thesis is to provide potential answers to these questions. One of the ways of approaching these issues has been to propose multi-scale models, which are nevertheless very complex [Anderson and Quaranta, 2008, Powathil et al., 2015]. We will focus here on the use of models defined almost exclusively at the molecular scale, which is assumed to be prominent, to study what can be inferred on the larger scales.

3.2 Cell circuitry and the need for cancer systems biology

Most biological systems, and certainly cells, fall into the category of **complex systems**. These are systems made up of many interacting elements. While these systems can be found in many different scientific fields, the cell as a complex system is characterized by the diversity and multifunctionality of its constituent elements (genes, proteins, small molecules, enzymes), which nevertheless contribute to organized and a priori non-chaotic behaviour [Kitano, 2002]. Thus, the role of a protein such as the p53 tumour suppressor can only be understood by taking into account the interplay between its relationships with transcription factors and biochemical modi-

fications of the molecule itself [Kitano, 2002]. In a cell, as in any complex system, the multiplication of components and interactions can make the response or behaviour of the system unexpected or unpredictable. Non-linear responses, such as abrupt changes in the state of a system, called critical transitions, can be observed in response to a moderate change in the signal [Trefois et al., 2015]. Generally speaking, it is possible to observe **emergent behaviours**, i.e. behaviours of the system as a whole that were not trivially deducible from the individual behaviours of its components. This has been documented, through experiments and simulations, in the study of cell signalling pathways and the resulting biological decisions [Bhalla and Iyengar, 1999, Helikar et al., 2008]. These considerations have thus given rise to **system-level or holistic approaches that aim to integrate data and knowledge into more comprehensive representations, often called systems biology**.

What is true for the cell in general is just as true for cancer in particular. Understanding the intertwining of signaling pathways is necessary to study their contributions to different cancer hallmarks, as shown in Figure 2.7. The concepts described above can thus be transposed to **cancer systems biology** [Hornberg et al., 2006, Kreeger and Lauffenburger, 2010, Barillot et al., 2012]. Indeed, it is often a question of understanding or predicting the impact of perturbations on cellular networks. Understanding how a single genetic mutation disrupts and reprograms networks, or even predicting the responses triggered by a drug on a presumably promising molecular target, makes little sense without integrated approaches. In addition, cancers are characterized by the accumulation of numerous mutations and alterations over time that must be considered concomitantly. These points of view of biologists and modellers reinforce the observation already made in the previous chapter of cancer as a network disease, as a system disease (Figure 2.8).

Finally, to conclude this general presentation, it is important to understand that while small molecular network modeling is not recent, the rise and multiplication of wide range systems biology approaches is very much related to the production of biological data [De Jong, 2002]. The last few decades have seen the emergence of high-throughput data that have made it possible to identify and link hundreds of genes or proteins involved in cancer. Exploring the interaction and back and forth between these models and the data they use or predict is therefore of utmost importance. In addition, the now ** massive amount of data has also imposed mathematical or computational approaches as a central element in the management of this profusion** and more and more modeling approaches are focused on data integration or inference [Fröhlich et al., 2018, Bouhaddou et al., 2018].

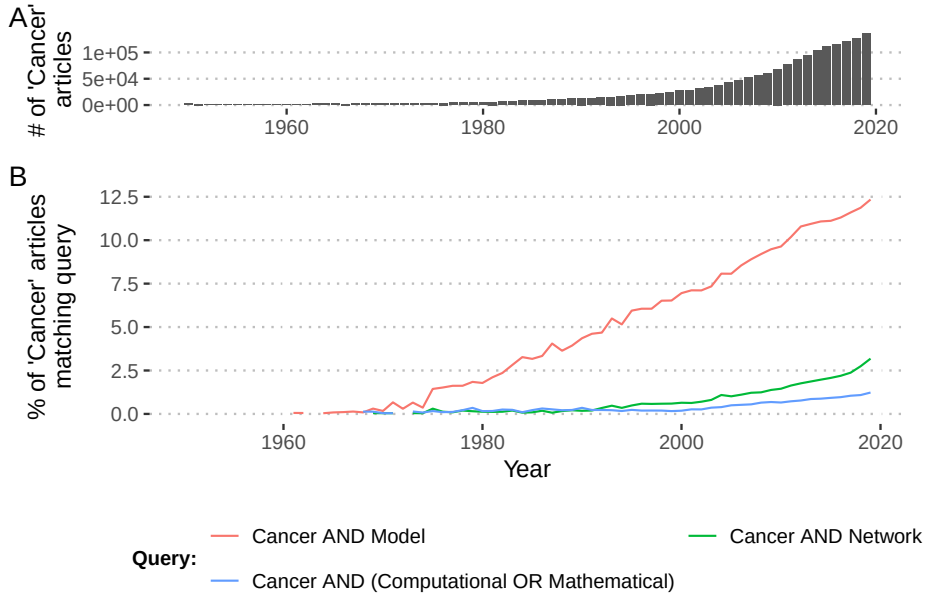

Figure 3.3: **PubMed trends in cancer studies.** (A) PubMed articles with the word *Cancer* in either title or abstract from 1950 to 2019. (B) Proportion of the *Cancer* articles with additional keywords expressed as PubMed logical queries.

More generally, Figure 3.3 shows that while the number of scientific articles devoted to cancer has increased drastically since the 1950s (panel A), the proportion of these same articles mentioning *models*, *networks* or *computational* approaches has also increased (panel B), illustrating a change in paradigms.

3.3 Mechanistic models of molecular signaling

Once the context has been defined, both biologically and methodologically, it is possible to begin the exploration of the models that will constitute the core of this thesis: the **mechanistic models of molecular networks** and signaling pathways. Before describing and illustrating some of the existing mathematical formalisms, it is possible to describe the common fundamental elements of this family of approaches.

3.3.1 Networks and data

The first step is to identify the relevant biological entities from a question or system of interest (e.g. tumor suppressor genes, signaling cascades of proteins) and then to model their interactions, the regulatory relationships that link them. At this stage the model can generally be represented by a network but this word can cover different realities [Le Novere, 2015]. The simplest network just represents undirected interactions between entities, which therefore only establishes relationships and not causal mechanisms. But modeling requires more precise definitions, in particular concerning the direction of the interaction (is it A that acts on B or the opposite) and its nature (type of chemical reaction, activation/inhibition etc.). This is usually summarized as **activity flows (or influence diagrams) with activation and inhibition arrows** as in Figure 2.7 or Figure 3.5A. These arrows emphasize the transormation of static networks into dynamic objects that can be manipulated and interpreted mechanistically. This work can be taken further by writing bipartite graphs, known as process descriptions, which explicitly show the different states of each variable (first type of nodes), depending on their phosphorylation sate for instance, and the reactions that link them (second type of nodes) as in Figure 3.5B. A more precise description of these different representations and their meanings can be found in Le Novere [2015]. **Once the network structure of the model has been defined, it is possible write the corresponding mathematical formalism** and potentially to refine certain parameters. Finally, the model is often confronted with new data to check its consistency with the biological behaviour studied or possibly make new predictions.

However, all these steps are not linear and sequential, but rather iterative and cyclical. This **modeling cycle**, with back and forth to the data, is not specific to molecular network models, but it is possible to specify it in this case (Figure 3.4). The names of the key players involved in the question of interest are thus first extracted from adapted data or from the literature. A first mathematical translation of the relationships between the entities is then proposed before verifying the compatibility of this model with the observations, whether qualitative or quantitative. If the compatibility is not good, we come back to the definition or the parameterization of the model. If compatibility is correct, the model can be used to make new predictions or study phenomena that go beyond the initial data set. Ideally, these predictions will be tested afterwards. This cyclic approach with two successive checks is analogous to the use of validation and test data in the evaluation of most learning algorithms. This analogy can sometimes be masked by the qualitative nature of the predictions or by the lack of explicit fitting of the

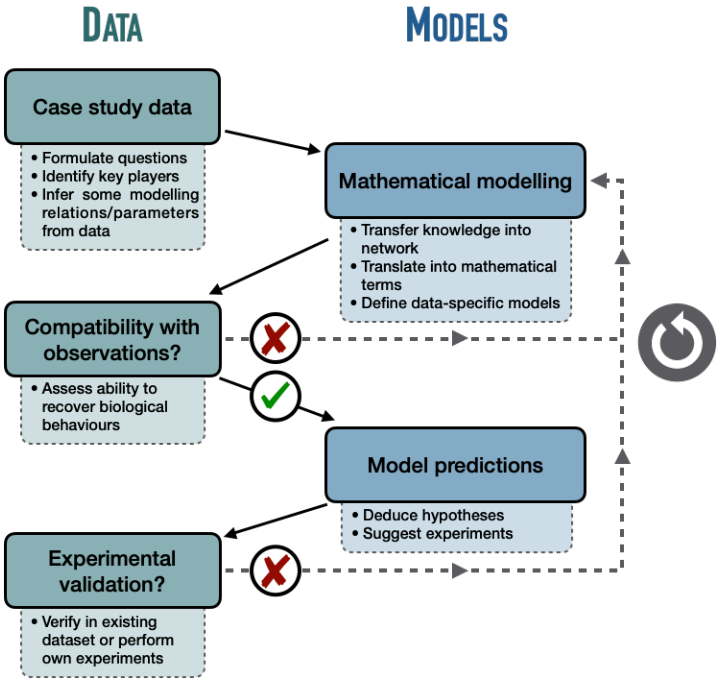

Figure 3.4: **Modeling a biological network: an iterative and cyclical process.** Reprinted from [Béal et al., 2020]. A different and simpler version of this cycle is described in [Le Novere, 2015].

parameters.

3.3.2 Different formalisms for different applications

Beyond these similarities in the construction and representation of models, the precise mathematical formalism that underlies them varies according to the type of question and the data [De Jong, 2002]. For the sake of simplicity, and without exhaustiveness, we propose to divide into quantitative and qualitative formalisms which will be essentially illustrated respectively by **ordinary differential equation (ODE)** models and **logical (or Boolean) models** for which a graphical and schematic comparison is proposed in Figure 3.5.

One of the most frequent approaches is the use of **chemical kinetics** equations to construct ODE systems which are a fairly natural translation of the process descritption networks described in the previous section [Polynikis et al., 2009]. Each biological interaction is treated as a reaction governed by the law of mass action and, under certain hypotheses, as a differential equation (Figure 3.5B); the set of reactions in the system then

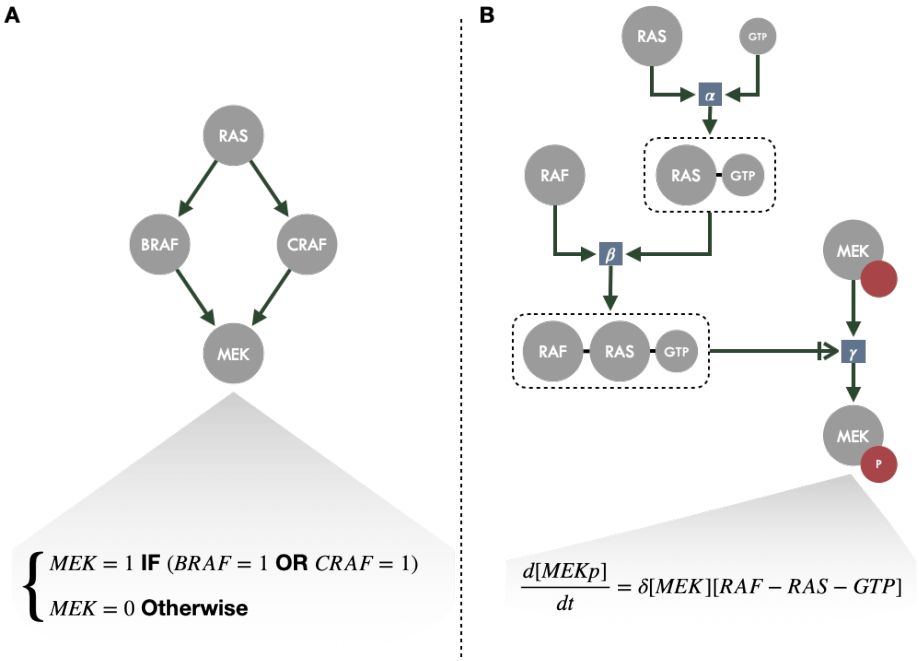

Figure 3.5: **Schematic example of logical and ODE modeling around MAPK signaling.** (A) Activity flow diagram of a small part of MAPK signaling, each node representing a gene or protein, with an example of logical rule for MEK node for the corresponding logical model. (B) Process description of the same diagram with BRAF and CRAF merged in RAF for the sake of simplicity; each square representing a reaction and the correspondong rate; an example differential equation is provided for the phosphorylated (active) form of MEK.

generates a set of differential equations with coupled variables, in an analogous way to the Lotka Volterra system presented in section 1.2.2. Thus the variables generally represent quantities of molecular species, for example concentrations of RNA or proteins, and the stoichiometric coefficients and reaction rates are used to define the system parameters. Approximations are sometimes made to simplify the equations, for example by assuming that they can be written as Michaelis-Menten's enzymatic reactions, which have a simple and well known behaviour. However, the theoretical accuracy of quantitative models has a cost since **each differential equation requires parameters**, such as reaction constants or initial conditions, to which the system is very sensitive [Le Novere, 2015]. The biochemical interpretation of the parameters sometimes allow to find their value in the literature, if the reactions are well characterized, even if possible variations in a given

biological or physical context are often unknown. Since knowledge of the values of these parameters is often limited or even non-existent, it may require a very large volume of data (including time series) to fit the many missing parameters which can be difficult if the number of parameters is large [Villaverde and Banga, 2014]. However, recent work has demonstrated the feasibility and scalability of this type of inference with sufficiently rich data [Fröhlich et al., 2018].

At the same time, more qualitative approaches to modeling biological networks have been proposed with discrete variables linked together by rules expressed as logical statements [Abou-Jaoudé et al., 2016]. These models are both more abstract since variables do not have a direct biological interpretation (e.g. concentration of a species) but are more versatile since they can unify different biological realities under the same formalism (e.g. activation of a gene or phosphorlation of a protein). The discrete nature of the variables can then be seen as an asymptotic case of the sigmoidal (e.g. Hill function) relationships often found in biology [Le Novere, 2015]. The step function thus obtained can keep a natural interpretation in the context of biological phenomena: genes activated or not, protein present or absent etc. Similarly, interactions between species are not quantified but are based on a qualitative statements (e.g. A will be active if B and C are active), drastically reducing the number of parameters (Figure 3.5A). If the theoretical interest of this formalism to study biological mechanisms was proposed quite early [Kauffman, 1969, Thomas, 1973], many concrete applications have also been developed over the years, particularly in cancer research [Saez-Rodriguez et al., 2011a, Remy et al., 2015]. This **logical formalism will constitute the core of the work presented in Part II**, where it will therefore be discussed in greater detail.

These two formalisms, which are among the most frequent for modelling biological networks, share many similarities, in particular the propensity to be built according to bottom-up strategies based on knowledge of the elementary parts of the model, i.e. biological entities and reactions. However, they differ in their implementation and objectives, **one aiming at the most accurate representation possible, the other seeking to capture the essence of the system's dynamics in a parsimonious way** (Table 3.1). The opposition is not irrevocable, as illustrated by the numerous hybrid formalisms that lie within the spectrum delimited by these two extremes such as fuzzy logic or discrete-time differential equations [Le Novere, 2015, Calzone et al., 2018]. To conclude, a comparison between the two approaches applied to the same problem is proposed by Calzone et al. [2018], studying the epithelio-mesenchymal transition (EMT, a biological process involved in cancer), to illustrate in concrete terms their complemen-

Table 3.1: **Features of quantitative and qualitative modeling applied to biological molecular networks** (adapted from Le Novere [2015])

	Quantitative modeling	Qualitative modeling
Example formalism	Ordinary differential equation (ODE) models	Logical models
Type of variables	Direct translation of biological quantities, usually continuous	Abstract representation of activity levels, usually discrete
Objective	Quantitatively accurate and temporal simulation of an experimental phenomenon	Coarse-grained simulation of qualitative phenotypes
Advantages	Direct confrontation with experimental data; precise; linear representation of time	Faster design; easy translation of literature-based assertions; simulation of perturbations
Drawbacks	Difficulty determining or fitting parameters	More difficult to link to data; lower precision

tarity.

3.3.3 Some examples of complex features

With the help of these models, both qualitative and quantitative, many complex behaviours have been identified. Benefiting from the knowledge accumulated in the study of dynamic systems, a whole zoo of patterns with complex and non-intuitive behaviours such as non-linearities have been highlighted [Tyson et al., 2003]. The MAPK pathway, coarsely described in Figure 3.5, and often simplified as a rather unidirectional cascade, shows switch or bistability behaviors generated by the complexity of its multiple phosphorylation sites [Markevich et al., 2004]. These models have also been put at the service of understanding cancer and the erroneous decision-making by cells resulting from impaired signaling pathways. Thus, Tyson et al. [2011] summarize superbly well the complexity that can be hidden in the dynamics of smallest molecular networks as soon as they contain more than two entites and crossed regulations or feedback loops. Logical models have also made it possible to better dissect some complex phenomena at play in the cell such as emergent behaviours [Helikar et al., 2008] or mechanisms behind mutation patterns in cancer [Remy et al., 2015].

3.4 From mechanistic models to clinical impact?

Mechanistic models have therefore undeniably led to a better understanding of the complex molecular machinery of signalling pathways. But beyond the interest that this understanding represents, do these models also have a clinical utility? In other words, **are they of clinical or only scientific value?**

3.4.1 A new class of biomarkers

Throughout this thesis, the clinical value of mechanical models will often be analyzed by analogy to that of biomarkers. Throughout this thesis, the clinical value of mechanical models will often be analyzed by analogy to that of biomarkers. Biomarkers are usually defined as measurable indicators of patient status or disease progression, such as prostate-specific antigen (PSA) for prostate cancer screening or BRCA1 mutation for breast cancer risk [Henry and Hayes, 2012]. Biomarkers also encompass multivariate signatures that identify more complex patterns with clinical significance. Taking the logic even further, it was therefore proposed that mechanistic models, which also reveal complex molecular behaviours, could be considered as biomarkers, capturing perhaps even dynamic information [Fey et al., 2015].

Like oncology biomarkers, the models will be divided into two categories according to their clinical objectives: **prognostic models and predictive models** [Oldenhuis et al., 2008]. Prognostic biomarkers and models are those that provide information on the evolution of cancer independently of treatment. They are therefore generally confronted with survival or relapse data. The protein Ki-67 for example, encoded by the MKI67 gene, is known to be indicative of the level of proliferation and high levels of expression are thus associated with a poorer prognosis in many cancers [Sawyers, 2008]. Predictive biomarkers and models, on the other hand, give an indication of the effect of a therapeutic strategy. The simplest example, but not the only one, concerns biomarkers that are themselves the target of treatment: treatments based on monoclonal antibodies directed against HER2 receptors in breast cancer are only effective if the HER2 receptor has been detected in the patient [Sawyers, 2008]. Without attempting to be exhaustive, some logical and ODE models, with either prognostic or predictive claims, will be described.

3.4.2 Prognostic models

One of the first mechanical models of cell signalling to have been explicitly presented as a prognostic biomarker is the one proposed by Fey et al. [2015] and describing c-Jun N-terminal kinase (JNK) pathway in neuroblastoma cells. A summary of the study is provided in Figure 3.6). The model is an ODE translation of the process description network of Figure 3.6)A, further determined and calibrated with molecular biology experimental data obtained using neuroblastoma cell lines. We thus observe the non-linear switch-like dynamics of JNK activation as a function of cellular stress (Figure 3.6)B). The precise characteristics of this sgmoidal response can, however, vary from one individual to another as captured by the network output descriptors A, K_{50} and H. Fey et al. proposed to perform neuroblastoma patient–specific simulations of the model, using patient gene expressions for ZAK, MKK4, MKK7, JNK and AKT genes to specify the initial conditions of the ODE system. Since JNK activation induces cell death through apoptosis, the patient-specific A, K_{50} and H derived from patient-specifc models are then analyzed as prognostic biomarkers (Figure 3.6)C). Readers are invited to refer to the original article for details on model calibration or binarization of network descriptors [Fey et al., 2015]. The authors also showed that in the absence of positive feedback from JNK^{**} to ${}^{P}MKK7$, an important component of non-linearity, the prognostic value is drastically decreased. All in all, this pipeline from ODE model to survival curves, thus provides a **paradigmatic example of the clinical interpretation of mechanistic models of molecular networks** that will be reused in later chapters for illustration purposes. Other ODE models following a similar rationale have been proposed by the same group for colorectal cancer [Hector et al., 2012, Salvucci et al., 2017] or glioblastoma [Murphy et al. [2013]; salvucci2019system]. Machine learning approaches have also been proposed to ease the clinical implementation of this kind of prognostic models by dealing with the potential lack of patient data needed to personalize them [Salvucci et al., 2019a].

On the logical modeling side, there are also studies including prognostic value validation. Thus, Khan et al. [2017] proposed two logical models of epitelio-mesenchymal transition (EMT) in bladder and breast cancers. These models are inferred from prior mechanisms knowledge and data analysis with particular attention to potential feedback loops. Using these models, it is possible to study the behaviour of them for all combinations of model inputs (growth factors and receptor proteins) and derive subsequent signatures for good or bad prognosis. These signatures are later validated with cohorts of patients. In this case, the mechanistic model does not seek

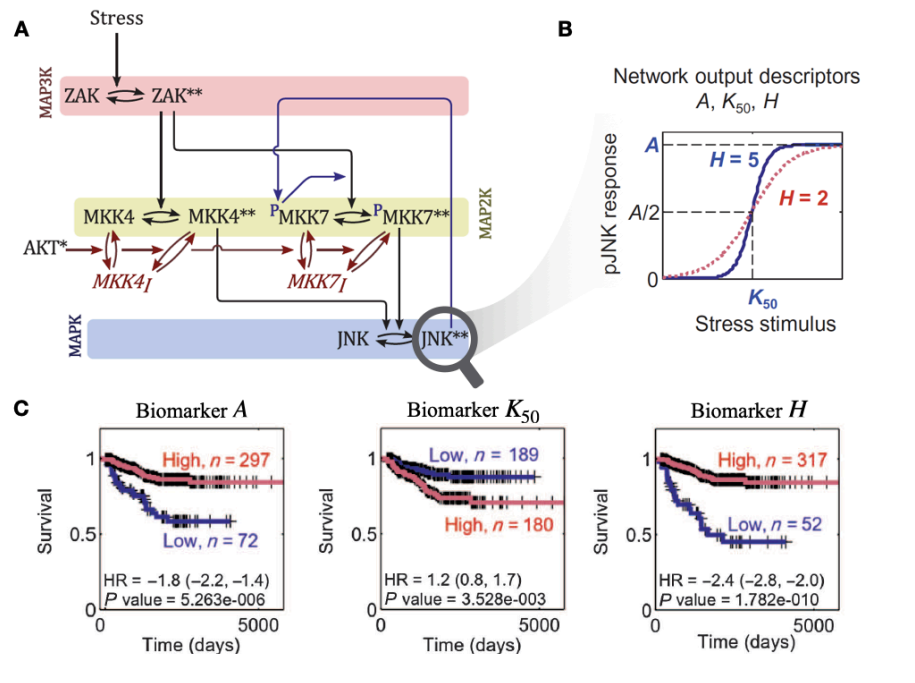

Figure 3.6: **Mechanistic modeling of JNK pathway and survival of neuroblastima patients, as described by Fey et al. [2015].** (A) Schematic representation, as a process description, for the ODE model of JNK pathway. (B) Response curve (phosphorylated JNK) as a function of the input stimulus (Stress) and characterization of the corresponding sigmoidal function with maximal amplitude A, Hill exponent H and activation threshold K_{50}. (C) Survival curves for neuroblastoma patients based on binarized A, K_{50} and H; binarization thresholds having been defined based on optimization screening on calibration cohort.

to capture a dynamic behavior but to facilitate and **make understandable the exploration of combinations of input signals that grow exponentially with the number of inputs considered**. Other formalisms, called pathway activity analysis and following the same activity flows principles (Figure 3.5A), have been analysed in the light of their prognostic value. Their greater flexibility enables the direct use of networks of several hundred or thousands of genes, such as those present in the KEGG database [Kanehisa et al., 2012]. The benefit of mechanistic modeling is then to organize high-dimensional data and to facilitate the *a posteriori* analysis of the results.

3.4.3 Predictive models

But the explicit representation of biological entities in mechanistic models makes them particularly **suitable for the study of well-defined perturbations such as drug effects**. Indeed, by assuming that the mechanism of action of a drug is at least partially known, it is possible to integrate this mechanism into the model if it contains the target of the drug (Figure 3.7). One can therefore simulate the effect of one drug or even compare several. These strategies have already been implemented in a qualitative way with logical models used to explain resistance to certain treatments of breast cancer [Zañudo et al., 2017] or even highlight the synergy of certain combinations of treatments in gastric cancer [Flobak et al., 2015]. The value of these models, however, is more scientific than clinical in that they focus on a single cell line or a restricted group of cell lines. The possibility to personalize the predictions or recommendations for different molecular profiles of cell lines or patients is therefore not obvious. Still within the context of logical formalism, Knijnenburg et al. [2016] proposed a broader approach: if their model needs to be trained, it can nevertheless provide an analytical framework for several hundred cell lines, while remaining within the scope of the training data to ensure the validity of predictions.

Conceptually comparable strategies can be found on the side of differential equations where large mechanical models of cell signalling are also trained to predict the response to different treatments [Bouhaddou et al., 2018, Fröhlich et al., 2018]. A calibrated model can then predict the response to a combination of treatments not tested in the training data, thereby proving the ability of mechanistic models to extend their predictive value beyond the data [Fröhlich et al., 2018]. As with prognostic models, mechanical approaches other than logical formalisms and ODEs have been proposed and validated [Jastrzebski et al., 2018]. What can be learned from these predictive models is that they require **significant training data to be able to go beyond qualitative predictions and dissect treatment response mechanisms of many cell lines simultaneously**. For obvious practical and ethical reasons, the validation of these models is for the moment limited to preclinical data since they require data for many uncertain therapeutic interventions.

This first bridge between mechanistic models of cell signalling and clinical applications concludes this introductory part. The next part will be devoted to the definition of new methods to establish this connection based on logical formalism, before the third part proposes a more statistical evaluation of the prognostic and predictive values of the models presented in the previous parts.

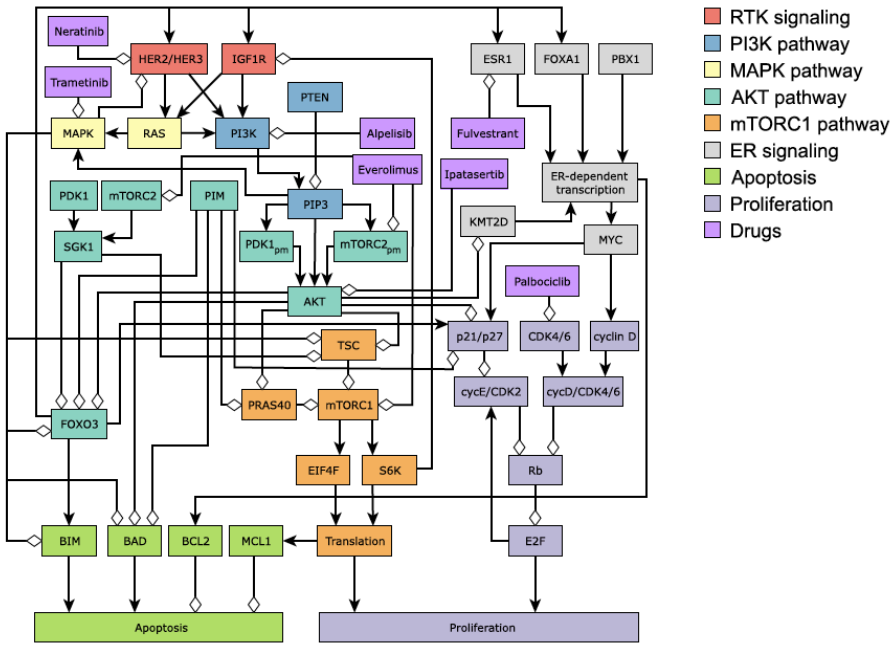

Figure 3.7: **Network model of oncogenic signal transduction in ER+ breast cancer, including some drugs and their targets.** Reprinted from Zañudo et al. [2017].

Part II

Personalized logical models of cancer

CHAPTER 4

Logical modeling principles and data integration

"Je suis l'halluciné de la forêt des Nombres.
Ils me fixent, avec leurs yeux de leurs problèmes ;
Ils sont, pour éternellement rester : les mêmes.
Primordiaux et définis,
Ils tiennent le monde entre leurs infinis ;
Ils expliquent le fond et l'essence des choses,
Puisqu'à travers les temps planent leurs causes."

Émile Verhaeren (Les nombres)

Another way of ordering the diversity of mechanistic models presented above is to consider their relationship to biological data. Those that make little use of these data are essentially theoretical scope models that describe the general functioning of signaling pathways and associated systems [Calzone et al., 2010]. Other models propose more quantitative models but require much more data, either from databases or experimental data generated for this purpose in order to fit the parameters. In the latter case, the necessary data is usually perturbation data: How does my system react to this or that inhibition or activation? For a single cell line this already corresponds to a large amount of data [Razzaq et al., 2018]. And if we want to extend these approaches to many cell lines the amount of data

becomes massive [Fröhlich et al., 2018]. For patient-specific models, access to this perturbation data is even more difficult.

Between theoretical models that are not very demanding in terms of data but not very applicable clinically and models with a clinical focus but very demanding in terms of data, an intermediate alternative is missing. **Can patient-specific mechanistic models be developed that would provide qualitative clinical interpretation with a small amount of data, accessible even in patients?** In this part, a middle way will be described to answer positively to this question. This methodology will be based on a historically qualitative mathematical formalism already presented in the previous chapter under the name of logical modeling. Logical modeling in general will be detailed in this chapter before describing an original customized approach in the next two chapters.

Publications

This chapter presents the theoretical bases of logical modeling and the tools used thereafter. It does not present any original work but refers to the synthesis and analyses of logical modeling as described in Béal et al. [2019] and Béal et al. [2020].

4.1 Logical modeling paradigms for qualitative description

Mathematical models serve as tools to answer a biological question in a formal way, to detect blind spots and thus better understand a system, to organize, into a consensual and compact manner, information dispersed in different articles. In the light of this definition, logical formalism (also called Boolean) may seem one of the closest to natural language in that it **can translate quite directly the statements present in the literature** such as "protein A activates protein B" or "the expression of gene C requires the joint presence of factors D and E". Indeed, shortly after the first descriptions of control circuits by Jacob and Monod [1961], the interest of logical models to describe biological systems was put forward by Kauffman [1969] and Thomas [1973]. Since then, studies have multiplied [Thomas and d'Ari, 1990], varying the fields of biological applications and also the mathematical and computational implementations [Naldi et al., 2018b]. The two subsections below summarize the characteristics common to most of the log-

ical formalisms, before detailing the implementation chosen in this thesis in section 4.2.

4.1.1 Regulatory graph and logical rules

A logical model is based on a network called **regulatory graph** (Figure 4.1), where each node represents a component (e.g. genes, proteins, complexes, phenotypes or processes), and is associated with discrete levels of activity (0, 1, or more when justified). The use of a discrete formalism in molecular network modeling relies on the highly non-linear nature of regulation, and thus on the existence of a regulatory threshold. Assuming that each variable represents a level of expression: it will take the value 0 if the level of expression of the entity is below the regulation threshold, i.e. insufficient to carry out the regulation; and the value 1 if it is above the threshold and regulation is possible. In other words, the control threshold discretizes the state space, here the expression levels. It is therefore possible to distinguish several thresholds for the same variable, corresponding to distinct controls that do not take place at the same expression levels. The variable is then multivalued. This extension greatly enriches the formalism, because it allows to distinguish situations that are qualitatively different and that would be confused with boolean variables. In the continuation of this thesis, we will consider by default that the activity levels are binary, 0 corresponding to an inactive entity and 1 to an active entity.

The edges of this regulatory graph correspond to influences, either positive or negative, which illustrate the possible interactions between two entities. Positive edges can represent the formation of active complexes, mediation of synthesis, catalysis, etc. and they will be later depicted as green arrows (←). Negative edges on the othe rhand can represent inhibition of synthesis, degradation, inhibiting (de)phosphorylation, etc. and they will be depicted as red turnstile (⊢).

Then, each node of the regulatory graph has a corresponding Boolean variable associated to it. The variables can take two values: 0 for absent or inactive (OFF), and 1 for present or active (ON). These variables change their value according to a logical rule assigned to them. The state of a variable will thus depend on its logical rule, which is based on logical statements, i.e., on a function of the node regulators linked with logical connectors AND (&), OR (|) and NOT (!). These operators can account for what is known about the biology behind these edges. If two input nodes are needed for the activation of the target node, they will be linked by an AND gate; to list different means of activation of a node, an OR gate will be used. For negative influences, a NOT gate will be utilized. Thus for each node, a logical

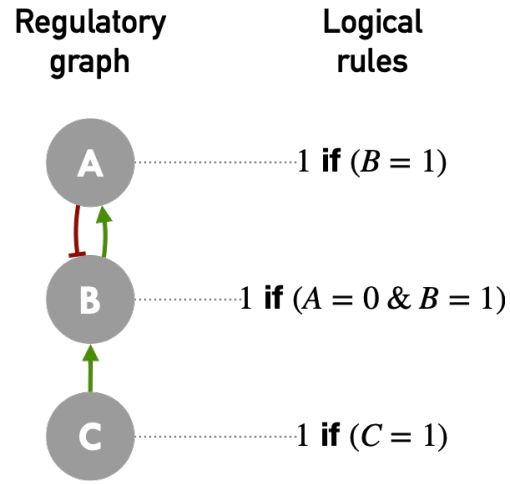

Figure 4.1: **A simple example of a logical model.** Regulatory graph on the left with positive (green) and negative regulations (red); a set of possible corresponding logical rules on the right.

rule is associated. The rules corresponding to the toy model in Figure 4.1 could be interpreted literally like this: A is acivated to 1 if B is active; B is updated to 1 in the absence of A and the presence/activity of C; C is an input of the model and therefore not regulated. It can be noted that the logical rules cannot be deduced only from the regulatory graph, which can be less precise and ambiguous. One could thus imagine that B is activated if C is, OR if A is not, thus changing the behavior of the model.

4.1.2 State transition graph and updates

In a Boolean framework, the variables associated to each node can take two values, either 0 or 1. We define a model state as a vector of all node states. All the possible transitions from any model state to another are dependent on the set of logical rules that define the model. These transitions can be viewed into a graph called a **state transition graph** (STG), where nodes are model states and edges are the transitions from one model state to another. STG nodes will be later depicted with rounded squares instead of circles in order to emphasize the difference with regulatory graphs. That way, trajectories from an initial condition to all the final states can be determined. In a model with n nodes, the state transition graph can contain up to 2^n model state nodes; thus, if n is too big, the construction and the visualization of the graph becomes difficult.

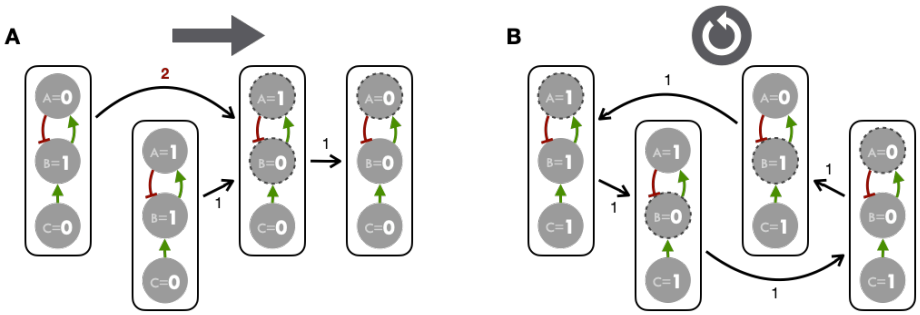

Figure 4.2: **State transition graph and synchronous updates.** Stable state (A) and limit cycle (B) attractors obtained for the example logical model with synchronous updates (all possible updates simultaneously). Figures above/below STG edges correspond to the number of nodes updated in each transition.

Based the simple logical model of Figure 4.1 it is nevertheless possible to represent the STG comprehensively. The idea for this is to start from a state of the system and track the successive states defined by the logical rules and the corresponding updates. The first strategy to construct this STG is to change simultaneously at each time step all the variables that can be changed (Figure 4.2). This method is referred to as a **synchronous updating strategy**. In the second method, referred to as a **asynchronous updating strategy**, variables are changed one at a time (Figure 4.3) and therefore each state has as many successors as there are components whose state must be changed according to logical rules (Figure 4.3A). The latter asynchronous method will be used exclusively in the work presented thereafter

We then define attractors of the model as long-term asymptotic behaviors of the system. Two types of attractors are identified: stable states, when the system has reached a model state whose successor in the transition graph is itself; and cyclic attractors, when trajectories in the transition graph lead to a group of model states that are cycling. For both synchronous and asynchronous updating strategies, the toy model shows the existence of **two types of attractors: a stable steady state and a limit cycle**, depending on the initial value of C. There are two disconnected components of the state transition graph for this example that correspond to the two possible values for the input C. If C is initially equal to 0 (inactive), then there exists only one stable state: $A = B = C = 0$. All the trajectories in the state transition graph lead to only one model state. If C is initially equal to 1, then the attractor is a limit cycle. The path in the state tran-

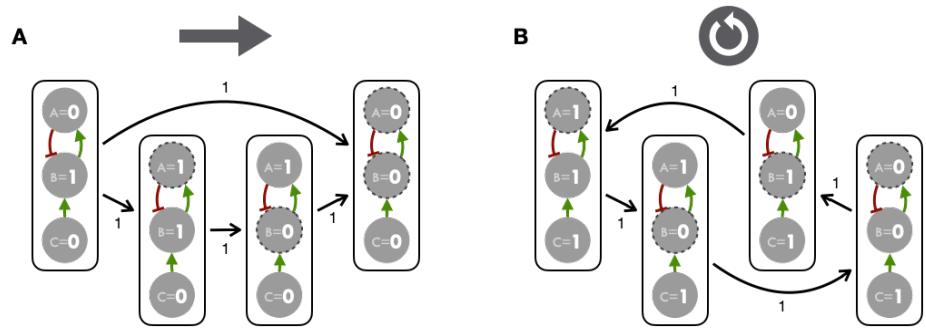

Figure 4.3: **State transition graph and asynchronous updates.** Stable state (A) and limit cycle (B) attractors obtained for the example logical model with asynchronous updates (one update at a time). Figures above/below STG edges correspond to the number of nodes updated in each transition.

sition graph cycles for any initial model state of this connected component. Note that for the asynchronous and synchronous graphs, the precise paths or limit cycles may vary. To conclude, it is important to emphasize and illustrate the characteristics of asynchronous updates in this toy example. In Figure 4.3A, the transition from the initial state ($A = C = 0; B = 1$) suggests two distinct possibilities, so it is necessary to **define additional rules or heuristics to choose between possible transitions**. We will come back to this by specifying the logical modeling implementation chosen in this thesis in section 4.2.

4.1.3 Tools for logical modeling

Numerous tools have been developed to build logical models and study the dynamics of the systems under inverstigation, each with its own specificity. They allow, for example, to represent reglation networks; to edit, modify or infer logical rules; to identify stable states; to reduce models; to visualize graphs of synchronous or asynchronous transitions. Some also allow to integrate temporal data; to discretize expression data; to simulate the model stochastically or to integrate delays; to identify existing models, etc. Among them, we can cite GINsim [Naldi et al., 2018a], BoolNet [Müssel et al., 2010], pyBoolNet [Klarner et al., 2016], BooleanNet [Albert et al., 2008], CellCollective [Helikar et al., 2012], bioLQM [Naldi, 2018], MaBoSS [Stoll et al., 2012, 2017], PINT [Paulevé, 2017], CaspoTS [Ostrowski et al., 2016], or CellNOptR [Terfve et al., 2012]. The interaction between all these tools, their interoperability and complementarity are highlighted in the form of a

notebook jupyter [Naldi et al., 2018b], and some of them are mentioned in section 4.3.

4.2 The MaBoSS framework for logical modeling

In the present study, all simulations have been performed with MaBoSS, a **Ma**rkovian **Bo**olean **S**tochastic **S**imulator whose design is summarized in Figure 4.4 and precisely described by Stoll et al. [2012] and Stoll et al. [2017]. This framework is based on an asynchronous update scheme combined with a continuous time feature obtained with Gillespie algorithm [Gillespie, 1976], allowing simulations to be continuous in time despite the discrete nature of logical modeling.

4.2.1 Gillespie algorithm

Gillespie algorithm provides a stochastic way to choose a specific transition among several possible ones and to infer a corresponding time for this transition. Thus, MaBoSS computation results in one stochastic trajectory as a function of time. To achieve this, transition rates seen as qualitative activation or inactivation rates, must be specified for each node (Figure 4.4A). They can be set either all to the same value by default, in the absence of any indication, or in various levels reflecting different orders of magnitude: post-translational modifications are quicker than transcriptions for instance. They can also be used to vary speeds depending on inputs or even to adapt multi-valued logical mechanisms in a binary framework [Stoll et al., 2012]. These transition rates are translated as transition probabilities in order to determine the actual transition (Figure 4.4B). Indeed, the probability for each possible transition to be chosen for the next update is the ratio of its transition rate to the sum of rates of all possible transitions. Higher rates correspond to transitions that will take place with greater probability, or in other words more quickly.

Thus at each update, the Gillespie algorithm performs the procedure described in Figure 4.4C. Two uniform random variables u and u' are drawn and used respectively to select the transition among the different possibilities (with u) and to infer the corresponding time (with u'). Based on the described formula, time δt follows an exponential law whose average is equal to the inverse of the sum of all possible transition rates (Figure 4.4C). In present work, except otherwise stated, all transition states will be initially assigned to 1.

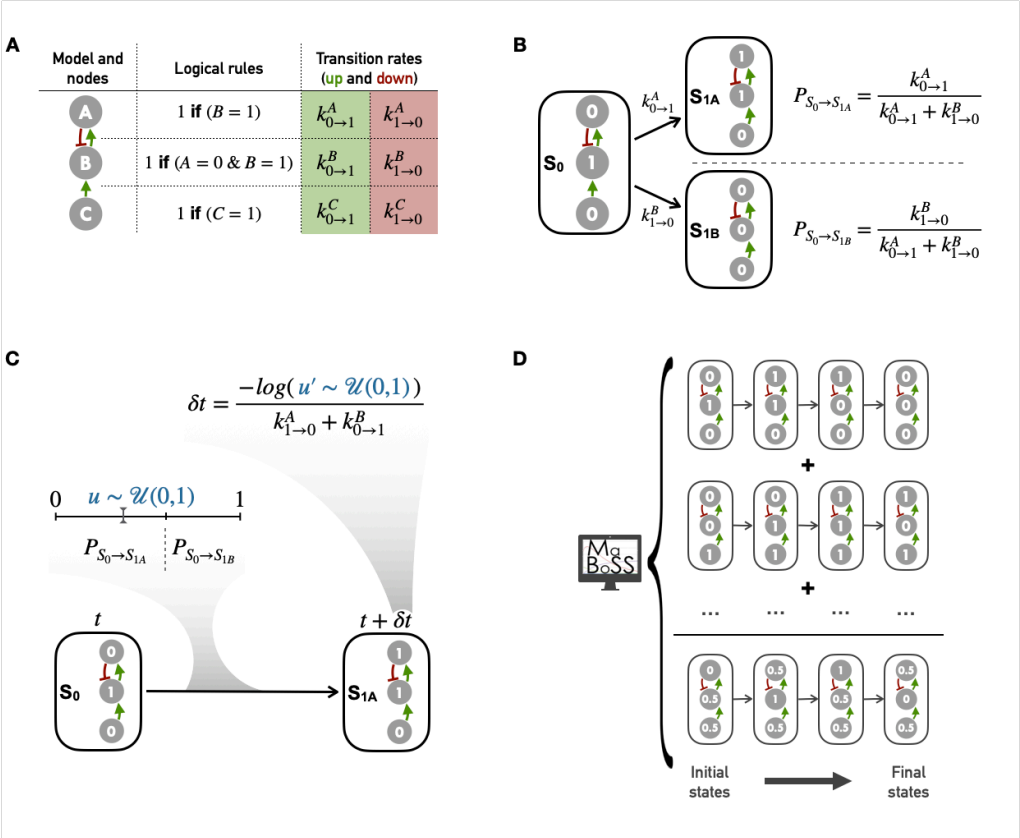

Figure 4.4: **Main principles of MaBoSS simulation framework and Gillespie algorithm.** (A) A logical model with regulatory graph, logical rules and transition rates. (B) A coresponding state transition graph with two possible transitions in asynchronous update for a given initial state; each transition has an associated probability. (C) Random selection of a specific transition and time by the Gillespie algorithm from two uniform random variables. (D) Schematic representation of a logical model simulation with MaBoSS: average trajectory obtained from the mean of many individual stochastic trajectories.

4.2.2 A stochastic exploration of model behaviours

Since MaBoSS computes stochastic trajectories, it is relevant to compute several trajectories in order to get an insight of the average behavior by generating a population of stochastic trajectories over the asynchronous state transition graph (Figure 4.4D). The aggregation of stochastic trajectories can also be interpreted as a description of an heterogeneous population. In fact, in all the examples in next chapters, all simulations have consisted on thousands of computed trajectories. The larger the model, the larger the space of possibilities and the more trajectories are required to explore it. Since several trajectories are simulated, initial values of each node can be defined with a continuous value between 0 and 1 representing the probability for the node to be defined to 1 for each new trajectory. For instance, a node with a 0.6 initial condition will be set to 1 in 60% of simulated trajectories and to 0 in 40% of the cases.

In the present work, we will focus on the "asymptotic"" state of these simulations instead of transient dynamics and we will call **node scores** the asymptotic agregated score obtained by averaging all trajectories at a given final time point. The simulation time should be chosen carefully to ensure that the asymptotic state is achieved, and the term "final state" may be considered as safer. Indeed, asymptotic states are more closely related to logical model attractors than transient dynamics and are therefore less dependent on updating stochasticity and more biologically meaningful [Huang et al., 2009].

All in all this modeling framework is at the intersection of logical modeling and continuous dynamic modeling. If the definition of time remains rather abstract and difficult to interpret experimentally, the stochastic exploration of trajectories makes it possible to refine the purely binary interpretation of the variables.

4.2.3 From theoretical models to data models?

To sum up, logical formalism makes it possible to design fairly quickly and easily models that reflect a priori knowledge of the phenomena being studied. Thus, they allow answering questions for which there is little information on the precise mechanisms involved in a disease or when there is a lack of data related to the expression of genes or the quantity of key proteins, or on the speed of certain processes. Logical models confirm that a network is a good illustration of the underlying biological question. However, in order to propose a patient-specific mechanistic approach, it seems crucial to use

the biological data available. How is this possible in a formalism that is by definition quite abstract?

4.3 Data integration and semi-quantitative logical modeling

The higher level of abstraction of the logical formalism sometimes makes the necessary back and forth between theoretical modeling and experimental or clinical data less easy. However, many theoretical approaches have been developed over the years to enable this dialogue at all stages, from the construction to the validation of a logical model, as summarized in Figure 4.5. This section summarizes some of these approaches to show how the use of biological data enriches logical models and brings them closer to clinical applications in precision medicine and in order to better contextualize the original approach presented in the following chapter. It should be noted that the methods presented below are all applicable to logic models, and illustrated with such examples where possible. However, some methods are not specific to this formalism and can be applied to other models.

4.3.1 Build the regulatory graph

Faced with a biological question (Figure 4.5, first step), it is crucial to identify the main actors in the process in order to define the outline of the model (Figure 4.5, second step). A first approach relies on the existing scientific literature on the topic: which biological species and which interactions have been identified as relevant to my problem? In a more automatic way, it is possible to extract information from different databases in order to establish a first list of biological entities and interactions associated with a biological phenomenon or even a gene of interest [Kanehisa et al., 2012, Perfetto et al. [2016]]. As an example, starting from the study of E2F1 gene as the hub of many regulatory mechanisms, Khan et al. [2017] have reconstructed a dense network of interactions in the vicinity of E2F1, which will be used for the construction of their subsequent model. The main difficulty here is to choose and select the relevant biological information adapted to the context of the model to be created, depending for example on the type of cancer studied or the desired level of precision.

But if the literature can be considered as processed data, it is also possible to use directly experimental data related to the problem under study. Key actors of biological processes identified by statistical analysis, such as

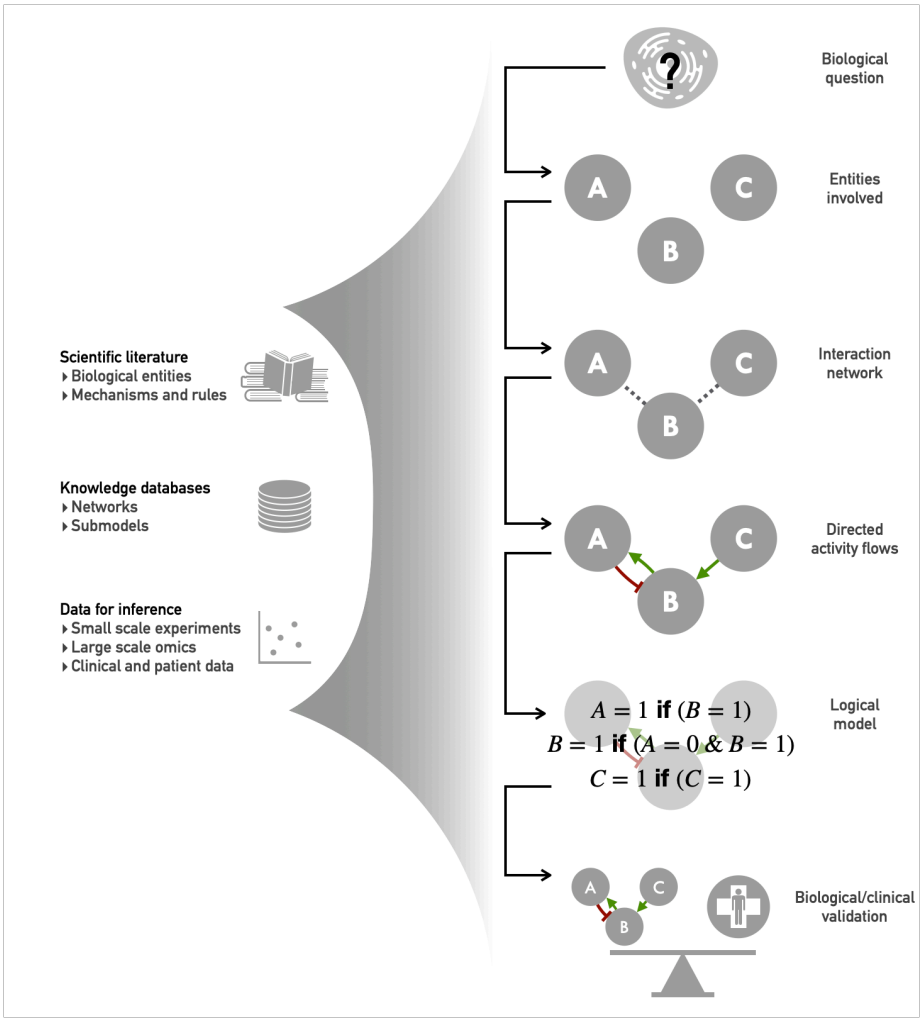

Figure 4.5: **Data integration in logical modeling.** The main types of data used are shown on the left; the essential steps of the logical modeling are shown linearly on the right.

differentially expressed genes or the most frequently mutated genes in a patient cohort, are selected and used as a starting point for the construction of the model [Remy et al., 2015]. More comprehensive approaches can use differential analysis tools on signaling pathways, rather than individual genes, to choose the relevant processes to include by contrasting different groups of patients based on their grades, metstatic status, resistance to treatments etc. [Martignetti et al., 2016, Montagud et al. [2017]]. Similarly, the study of regulatory networks involving transcription factors may justify the use of ChIP-seq data to identify possible new transcriptional regulations not previously listed [Collombet et al., 2017].

Once the main actors have been identified, it is necessary to infer the links between them (Figure 4.5, third and fourth step). However, starting from a list of genes and proteins of interest, how can we ensure that the regulatory relationships are complete and relevant? While a careful reading of the literature can provide locally interesting information, the use of omics data is also a resource that can be declined to different levels of precision. The major interest of these methods, assuming that the data are adequate and sufficiently massive, is to be able to extract information as large as the dataset, potentially on hundreds of entities, and above all specific to the object of study: a cancer subtype or a particular cell line can thus generate their own interaction network [Lefebvre et al., 2010]. Inference methods extract biological knowledge hidden in large databases, summarize it and represent it via networks. Many methods construct coexpression networks, which are non-oriented graphs, with different metrics and methods [Margolin et al., 2006, Vert et al., 2007]. Other approaches seek to infer causal relations between components, allowing the reconstruction of directed graphs where the links between entities are oriented, and sometimes even signed as activating or inhibiting regulations. These methods often make use of time series [Hill et al., 2016] or perturbation data [Meinshausen et al., 2016], but also more recently from observational data [Verny et al., 2017]. The information extracted from the data is then directly readable in the form of activity flows as described in the SBGN standards [Novère et al., 2009], thus providing a representation adapted to the construction of qualitative and a fortiori logical models [Le Novere, 2015]. Closer to the objective of defining logical models, certain methods allow the study and inference of co-regulation expressed with logical operators [Elati et al., 2007], thus facilitating the passage from the definition of an interaction network to the construction of a true logical model.

4.3.2 Define the logical rules

Precision must then be taken further by defining the logical rules that complete the network (Figure 4.5, fifth step). The first source of aggregated data to define logical rules is the scientific literature. The modeller looks for the state of knowledge on a given regulatory mechanism and translates it into a logical rule, according to the desired level of precision. For example, it has been observed that the protein kinase AKT can stabilize the oncogene MDM2 by phosphorylation, which leads to the degradation of p53 by forming a complex with it: this example can be translated by a simple inhibition relationship of AKT on p53 if this level of precision is considered sufficient or else intermediate species such as MDM2 can be used [Cohen et al., 2015]. Then, the effect of inhibition must be defined: can MDM2 alone inhibit p53 or does the presence of other activators outweigh this effect? This kind of considerations allows to define the logical combinations between the different inputs of a network node. In some cases, experimental data can be used to answer such questions: is a single activator sufficient or is the presence of all activators necessary? Which of the activator or inhibitor prevails in the case of simultaneous presence? While this information is often found in the literature, one should generate one's own experimental data to ensure an answer tailored to the study context, using a variety of experimental molecular biology techniques. For example, in order to elucidate the relationship between Foxo1 and Cebpa in a model of differentiation of myeloid and lymphoid cells, Collombet et al. [2017] first established the physical relationship between these species by ChIP-seq before determining the nature of this relationship using an ectopic expression experiment of Foxo1 in macrophage cells.

Other, more global approaches have been developed in recent years, driven by the influx of data from high-throughput sequencing techniques. Based on this rich and complex data, it has become possible to infer entire logical models, with precisely defined rules and interactions [Ostrowski et al., 2016]. The algorithms CellNOpt [Terfve et al., 2012] and caspo [Videla et al., 2017] provide two examples of these approaches, and more recently the SCNS tool described a graphical interface to infer logical models from single cell data [Woodhouse et al., 2018]. This model-inference goes beyond simpler structure-inference by defining the logical rules but it is generally based on a predefined topological structure to which time series or perturbation data are added. These data provide access to the response dynamics of a system. By questioning the way the system reacts, these data are therefore richer than a snapshot and thus facilitate the transition from correlation to causality, and thus the inference of logical rules. In practice,

the use of proteomic or phospho-proteomic data is often recommended because these data account for the activity of the protein and are in fact the closest to the cellular response: [Ostrowski et al., 2016, Terfve et al., 2012, 2015]. In spite of the richness of this type of data, model inference is sometimes still an under-determined problem that can lead to a large number of models with different logical equally compatible with the data. In such situations, it is then a matter of choosing the model on the basis of biological relevance criteria or of accepting to use families of models instead of limiting oneself to a single model [Videla et al., 2017]. In all cases, constructing logical rules directly from data specific to the problem can make it possible to obtain logical rules that are also specific to the context or the system under study [Saez-Rodriguez et al., 2011b]. For example, the inference of logical models specific to one or some cancer cell lines is a powerful tool to study their particularities [Razzaq et al., 2018].

4.3.3 Validate the model

Finally, the data can be used to validate the biological or clinical relevance of the models (Figure 4.5, sixth step). Compared to a system of differential equations, logical modelling has the particularity of being more abstract and therefore less directly reliable to an experimental reality for its validation. A system of differential equations can be compared to the chemical kinetics of the biological system under study. Compared to continuous formalisms, the dynamics of logic model simulation is more delicate to take into account but it is possible to verify it qualitatively, for example by validating the cyclic nature of activation trajectories for a model simulating the cell cycle [Fauré et al., 2006] or cellular decisions as a function of the activation signal [Calzone et al., 2010]. A second, more frequent approach consists in looking at the model's steady states and associating them with physiological conditions [Weinstein et al., 2017, Cohen et al. [2015]]. Finally, a third strategy focuses on the asymptotic state reached during the stochastic simulation of the model(s), a state representing a mixture of the different steady states according to the probability that the model has of reaching them.

In many models, to facilitate the analysis, nodes representing phenotypes have been added as "read-out" of the activity of certain entities. Thus, if a model includes a node named *Proliferation*, it will then be simpler to draw interpretations from the simulations performed with the model that will be linked to experimental observations of tumor growth or cell proliferation [Grieco et al., 2013, Steinway et al., 2015]. To validate these models, the activity of phenotypes, when forcing some node activity to 0 or 1, is

compared with the results of gene mutations reported in experiments carried out on mice or cell lines [Fauré et al., 2006, Cohen et al., 2015]. Another similar method for validating the relevance of a logic model is based on the analysis of the effects of different therapeutic molecules. The mechanistic nature of logic modeling makes it relatively easy to simulate the effect of these molecules. It is possible to simulate the effect of an inhibitory molecule by forcing the activity of its target to 0 and to compare with data [Zañudo et al., 2017, Iorio et al., 2016, Knijnenburg et al., 2016].

Beyond validation, some studies have predicted new therapeutic targets based on logical models, for instance by pointing out weaknesses in the topology of a regulatory system [Sahin et al., 2009]. Taking advantage of the ease of modeling and multiplying combinations of therapeutic molecules, logical modeling has also proved fruitful in predicting the best therapeutic combinations and their synergies, in the context of gastric cancers for example [Flobak et al., 2015]. Experimental confirmation of the predictions resulting from the modeling is then the ultimate stage in the validation of a logic model, completing the fruitful round trip between models and data.

Chapter 5

Personalization of logical models: method and prognostic validation

"All happy families are alike; each unhappy family is unhappy in its own way."

Leo Tolstoy (Anna Karenina, 1877)

Now that logical modeling has been introduced in detail, it is possible to come back to the question that structures this part and to refine it. **Is it possible to use routine omics data to obtain logical models that provide qualitative clinical interpretation?**. We thus propose a sequential approach, separating the model construction process from the integration of biological data. A generic logical model is first built, based on the literature knowledge, and the data are then used to specify the model. Indeed, the model as defined from the literature is often generic in the sense that it summarizes the state of knowledge on a probably heterogeneous pathology or population. Assuming that this general regulatory scheme provides a relevant framework for the system, it may then be relevant to use more precise omics data to impose biologically sourced constraints on the model: inactivation of a gene in a patient, activation of a protein or a signalling pathway by overexpression or phosphorylation, etc. This approach, called **PROFILE** (PeRsonalization OF logIcaL ModEls), allows the integration of both discrete (mutations) and continuous

data (RNA expression levels, proteins) based on the MaBoSS software, and leads to specific models of a cell line or a patient.

Publications

This chapter presents the method developed during the thesis to personlize logical models, i.e. generate patient-specific models from a single generic one. The description of the method and analyses on patient data from TCGA have been comprehensively described in Béal et al. [2019] and briefly summarized in Béal et al. [2020]. Analyses on cell lines are unpublished.

5.1 From one generic model to data-specific models with PROFILE method

The PROFILE method is summarized in Figure 5.1 and the different steps are successively described in the following subsections.

5.1.1 Gathering knowledge and data

The first steps are therefore to build a logical model adapted to the biological question (Figure 5.1, upper left) and to collect omics data that will be used to personalize the model (Figure 5.1, upper right). The construction of the model can be based on literature or data (see previous chapter). In the latter case, the data used to build the model will preferably be distinct from those used to personalize the model.

5.1.1.1 A generic logical model of cancer pathways

In this chapter, which is essentially methodological in nature, we will use a published logical model of cancer pathways to illustrate our PROFILE methodology. It is based on a regulatory network summarizing several key players and pathways involved in cancer mechanisms: RTKs, PI3K/AKT, WNT/β-catenin, TGF-β/Smads, Rb, HIF-1, p53 and ATM/ATR [Fumia and Martins, 2013]. The later analyses will be mainly focused on two read-out nodes, *Proliferation* and *Apoptosis*. Based on the model's logical rules *Proliferation* node is activated by any of the cyclins (CyclinA, CyclinB, CyclinD, and CyclinE) and is, thus, an indicator of cyclin activity as an abstraction and simplification of the cell cycle behavior. *Apoptosis* node is

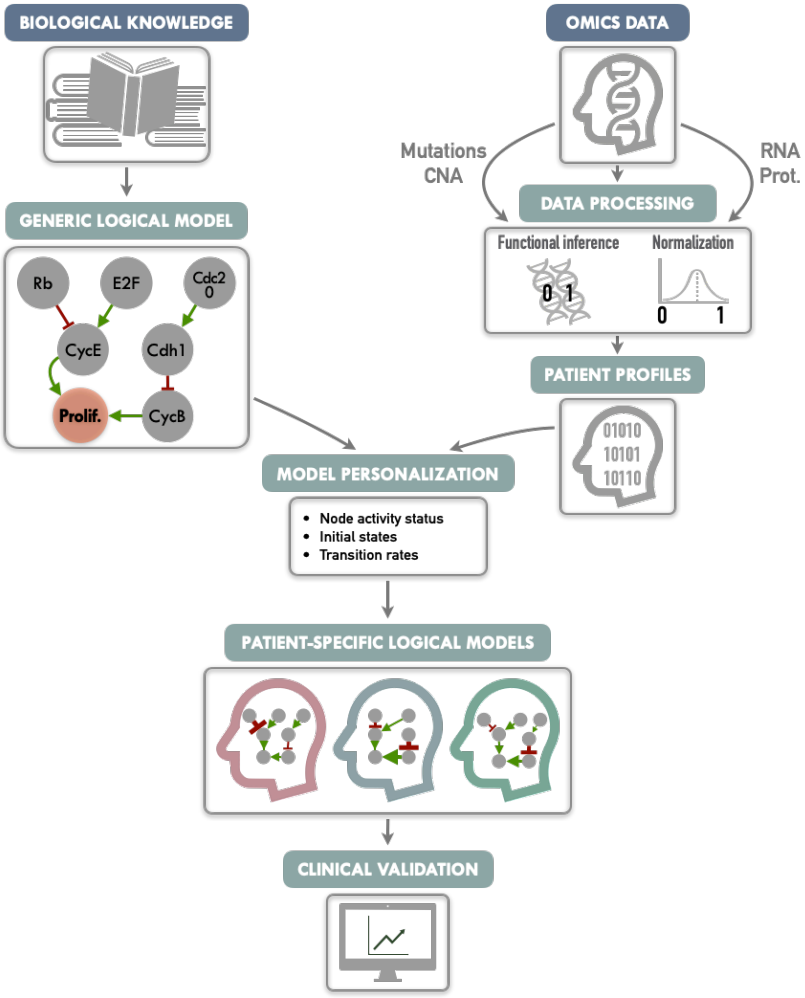

Figure 5.1: **Graphical abstract of PROFILE method to personalize logical models with omics data.** On the one hand (upper left), a generic logical model, in a MaBoSS format is derived from literature knowledge to serve as the starting-point. On the other hand (upper right), omics data are gathered (e.g., genome and transcriptome) as data frames, and processed through functional inference methods (for already discrete genome data) or binarization/normalization (for continuous expression data). The resulting patient profiles are used to perform model personalization, i.e. adapt the generic model with patient data. The merging of the generic model with the patient profiles creates a personalized MaBoSS model per patient. Then, biological or clinical relevance of these patient-specific models can be assessed.

regulated by Caspase 8 and Caspase 9. The generic model of Fumiã and Martins (2013) contains 98 nodes and 254 edges. Further details and visual representation are provided in section B.1 and Figure B.1. Model files are available in MaBoSS format in a dedicated GitHub repository.

5.1.1.2 Cancer data to feed the models

In order to showcase the method, breast-cancer patient data are gathered from METABRIC studies [Curtis et al., 2012, Pereira et al., 2016]. 1904 patients have data for both mutations, copy number alterations, RNA expression and clinical status (e.g relapse, survival). This number rises to 2504 patients if we only look at the mutations. Additional analyses were also performed based on the smaller and clinically less complete TCGA breast cancer data [Network et al., 2012]. These are detailed in Béal et al. [2019] but not included in this thesis. A more comprehensive description of these two databases can be found in section A.3.

In addition to these examples proposed in the original article, an application to cell line data is proposed in section 5.2.1 to link to the next chapters. A.1 In all cases, samples and cell lines will sometimes be referred to as patients for the sake of simplicity.

Details on the different datasets can be found in the appendix, section A.

5.1.2 Adapting patient profiles to a logical model

Before describing precisely the methodologies for using the data to generate patent-specific models, it is important to understand that these data will need to be transformed. This is the transformation of raw omics data into processed profiles that can be used directly in logical modeling.

5.1.2.1 Functional inference of discrete data

Since the logical formalism is itself discrete, the integration of discrete data is more straightforward. The most natural idea, used in many previous works, is to interpret the effect of these alterations and to encode it discreetly in the model. For instance, a deleterious mutation is integrated into the model by setting the corresponding node to 0 and ignoring the logical rule associated to it. For activating mutation, the node is set to 1. The main obstacle is therefore to estimate the functional impact of the alterations in order to translate them as well as possible in the model.

For mutations, based on the variant classification provided by the data, inactivating mutations (nonsense, frame-shift insertions or deletions and mutation in splice or translation start sites) are assumed to correspond to loss of function mutations and therefore the corresponding nodes of the model are forced to 0. Then, missense mutations are matched with OncoKB database [Chakravarty et al., 2017]: for each mutation present in the database, an effect is assessed (gain or loss of function assigned to 1 and 0, respectively) with a corresponding confidence based on expert and literature knowledge. Mutations targeting oncogenes (resp. tumor-suppressor genes), as defined in the 2020+ driver gene prediction method [Tokheim et al., 2016], are assumed to be gain of function mutations (resp. loss of function) and therefore assigned to 1 (resp. 0). To rule potential passenger mutations out, each assignment requires that the effect of the mutation has been identified as significant by predictive software based on protein structure such as SIFT [Kumar et al., 2009] or PolyPhen [Adzhubei et al., 2010].

For integration of copy number alterations, we use the discrete estimation of gain and loss of copies from GISTIC algorithm processing [Mermel et al., 2011]. The loss of both alleles of a gene (labelled -2) can thus be interpreted as a 0. Conversely, a significant gain of copies (labelled +2) denotes a gene that tends to be more highly expressed although the interpretation is more uncertain.

5.1.2.2 Normalization of continuous data

The integration of continuous data, such as RNA expression levels, in logical modeling is more difficult. The stochastic framework of MaBoSS provides however some possibilities. The main continuous mechanistic parameters of MaBoSS are the initial conditions of each node (its initial probability of being activated among the set of simulated stochastic trajectories) and the transition rates associated with the nodes (its probability to have its transition performed in an asynchronous update). In order to facilitate the use of continuous data through one of these two possibilities, we propose to transform them so that the values are continuous between 0 and 1, what we will refer to hereafter as normalized data. It is assumed that these continuous data can be good proxies of activity, 0 corresponding to a very low level of activity of the biological entity and 1 to a very high level. This assumption will have to be explained and justified each time: high level of expression of an RNA or significant phosphorylation of a protein interpreted as continuous markers of an important biological activity for example.

One of the assumptions of our analysis is that the interpretation of con-

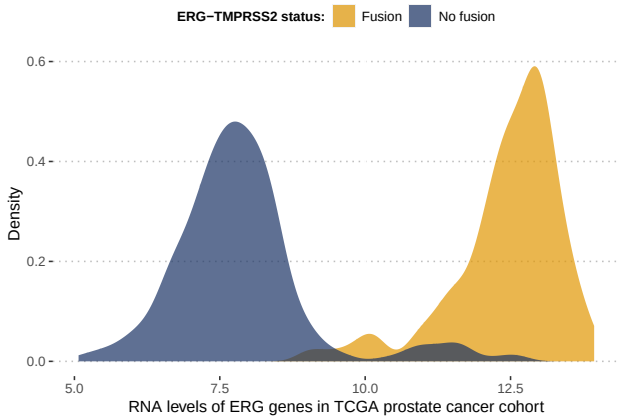

Figure 5.2: **Bimodal distribution of ERG gene in TCGA prostate cancer cohort.** This bimodality is largely explained by the fusion status of ERG gene. Patients for whom the gene has fused with TMPRSS2 have a much higher level of RNA expression for ERG.

tinuous data can only be relative and not absolute. It is indeed difficult to define an absolute threshold of RNA level at which a gene will be considered as activated. This may depend on contexts, technologies or even the way in which the data have been processed. On the other hand, it is possible to estimate that a gene is over-expressed for a patient compared to a cohort of interest. In contrast, the effect of a mutation can be estimated more independently. Thus, the continuous data will be normalized for the whole cohort studied, for each gene individually. In order to retain biological information as much as possible, distribution patterns are identified and normalized in different ways (Figure 5.4). We will illustrate the process by taking the example of the expression data expressed with continuous RNA levels. Beforehand, genes with no variation in expression level or too many missing values are discarded from the analysis. Then, we seek to identify first the genes that have a **bimodal** distribution. Indeed, these naturally fit into a binary formalism and this bimodality often has an underlying biological explanation. As an example, in the TCGA prostate cancer cohort (used in section 6.3), a gene called ERG has a bimodal distribution when looking at RNA levels in all patients. This distribution is almost entirely explained by an underlying genetic alteration that is the fusion of the ERG gene with the TMPRSS2 gene promoter (Figure 5.2), which is very common in this cancer [Tomlins et al., 2005]. In the data we identify bimodal patterns based on three distinct criteria: **Hartigan's dip test of unimodality, Bimodality Index (BI) and kurtosis**. The dip

test measures multi-modality in a sample using the maximum difference between empirical distribution and the best unimodal distribution, i.e., the one that minimizes this maximum difference [Hartigan and Hartigan, 1985]. Values below 0.05 indicate a significant multi-modality. In PROFILE, this dip statistic is computed using the R package *diptest*. The Bimodality Index (BI) evaluates the ability to fit two distinct Gaussian components with equal variance [Wang et al., 2009]. Once the best 2-Gaussian fit is determined, along with the respective means μ_1 and μ_2 and common variance σ, the standardized distance δ between the two populations is given by

$$\delta = \frac{|\mu_1 - \mu_2|}{\sigma}$$

and the BI is defined by

$$BI = [\pi(1-\pi)]^{1/2}\delta$$

where π is the proportion of observations in the first component. In PROFILE, BI is computed using the R package *mclust*. Finally, the kurtosis method corresponds to a descriptor of the shape of the distribution, of its tailedness, or non-Gaussianity. A negative kurtosis distribution, especially, defines platykurtic (flattened) distributions, and potentially bimodal distributions. It has been proposed as a tool to identify small outliers subgroups or major subdivisions (Teschendorff et al., 2006). In our case, we focus on negative kurtosis distributions to rule out non-relevant bimodal distributions composed of a major mode and a very small outliers' group or a single outlier. Although dip test, BI and negative kurtosis criteria emerge as similar tools in the sense that they select genes whose values can be clustered in two distinct groups of comparable size, we choose to combine them in order to correct their respective limits and increase the robustness of our method. For that, we consider that all three conditions (Dip test, Bimodality Index and kurtosis) must be fulfilled in order for a gene to be considered as bimodal. The thresholds of each test are inspired by those advocated in the papers presenting the tools individually. Dip test is a statistical test to which the classical 0.05 threshold has been chosen. In the article describing BI, authors explored a cut-off range between 1.1 and 1.5 and we chose 1.5 for the present work. Regarding kurtosis, the usual cut-off is 0, but since this criterion does not directly target bimodality, this criterion has been relaxed to $K < 1$. Several examples of the relative differences and complementarities between these criteria can be seen in Figure 5.3.

Non-bimodal genes are further classified as unimodal or zero-inflated distributions, looking at the position of the distribution density peak. Then,

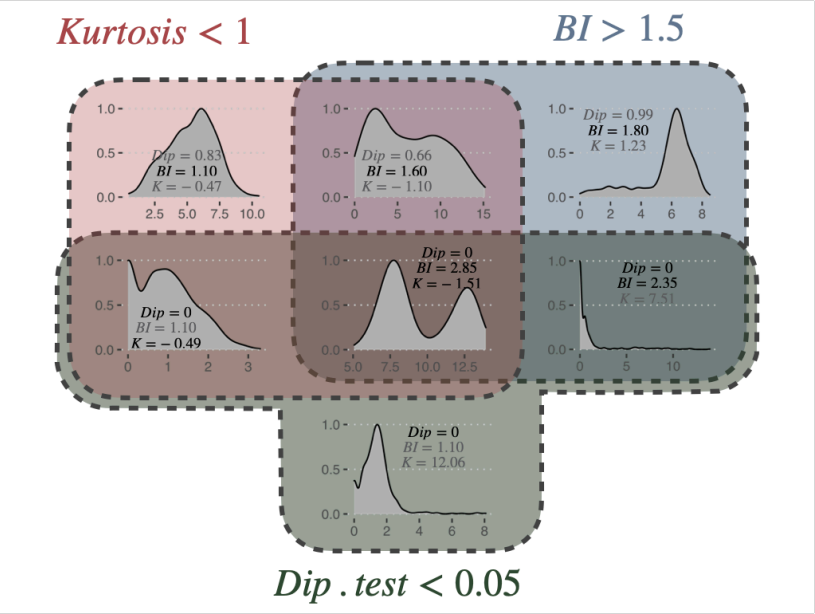

Figure 5.3: **Bimodality criteria and their combinations.** Examples of gene expression distributions for the different combinations of bimodality criteria: Dip test, Bimodality Index (BI) and kurtosis (K). Plots are organized in a Venn diagram.

based on this three category classification of genes, a pattern-preserving normalization can be performed, as summarized in Figure 5.4. For a bimodal gene i, a 2-component Gaussian mixture model is fitted using *mclust* R package resulting in a lower mode $M_{i,0}$ and an upper mode $M_{i,1}$. Denoting $X_{i,j}$ the expression value for gene i and sample j, $X_{i,j}$ has a probability to belong to $M_{i,0}$ or $M_{i,1}$ such as $P[X_{i,j} \in M_{i,0}] + P[X_{i,j} \in M_{i,1}] = 1$. For these bimodal genes, the normalization processing is defined as:

$$X_{i,j}^{norm} = P[X_{i,j}] \in M_{i,1}$$

For unimodal distributions, we transform data through a sigmoid function in order to maintain the most common pattern which is unimodal and nearly-symmetric:

$$X_{i,j}^{norm} = \frac{1}{1 + e^{-\lambda(X_{i,j} - median(X_i))}}$$

Since the slope of the function depends on λ, we adapt it to the dispersion of initial data in order to maintain a significant dispersion in $[0, 1]$ interval: more dispersed unimodal distributions are mapped with a gentle

slope, peaked distributions with a steep one. We map the median absolute deviation $MAD(X_i) = median(|X_i - median(X_i)|)$ on both sides of the median respectively to 0.25 and 0.75 to ensure a minimal dispersion of the mapping. First, the the proposed mapping results in:

$$\lambda = \frac{log(3)}{MAD(X_i)}$$

Last, zero-inflated distributions are transformed by linear normalization of the initial distribution:

$$X_{i,j}^{norm} = \frac{X_{i,j} - min(X_i)}{max(X_i - min(X_i))}$$

The transformation is applied to data between 1st and 99th quantiles to be more robust to outliers. Values outside this range are respectively assigned to 0 and 1. All the categoriation of distributions and the subsequent normalizations are summarized in Figure 5.4. With the help of the categories described here, it is also possible to binarize the continuous data quite simply. This binarization is required for some methods of network inference or logical modeling but will not be used in the examples presented beloww. The reader may refer to Béal et al. [2019] for more details.

5.1.3 Personalizing logical models with patient

It is now possible to redefine more precisely the ways of integrating data into a logic model defined with MaBoSS, as sketched at the beginning of the previous section. Personalization is defined here as the specification of a logical model with data from a given patient: each patient has a personalized model tailored to his/her data, so that all personalized models are different specifications of the same logical model, using data from different patients (Figure 5.1). Based on MaBoSS formalism and the processed patient data, there are several possibilities to personalize a generic logical model with patient data. One possibility to have patient-specific models is to force the value of the variables corresponding to the altered genes in a given patient, i.e.,constraining some model nodes to an inactive (0) or active (1) state (Figure 5.5A). In order to constrain a node to 0 (resp. 1), the initial value of the node is set to 0 (resp. 1) and $k_{0\to1}$ (resp. $k_{1\to0}$) to 0 to force the node to maintain its defined state. For instance, the effect of a p53 inactivating mutation can be modeled by setting the node TP53 in the model and its initial condition to 0 and ignoring the logical rule of TP53 variable. These modifications are referred to as node activity in the logical

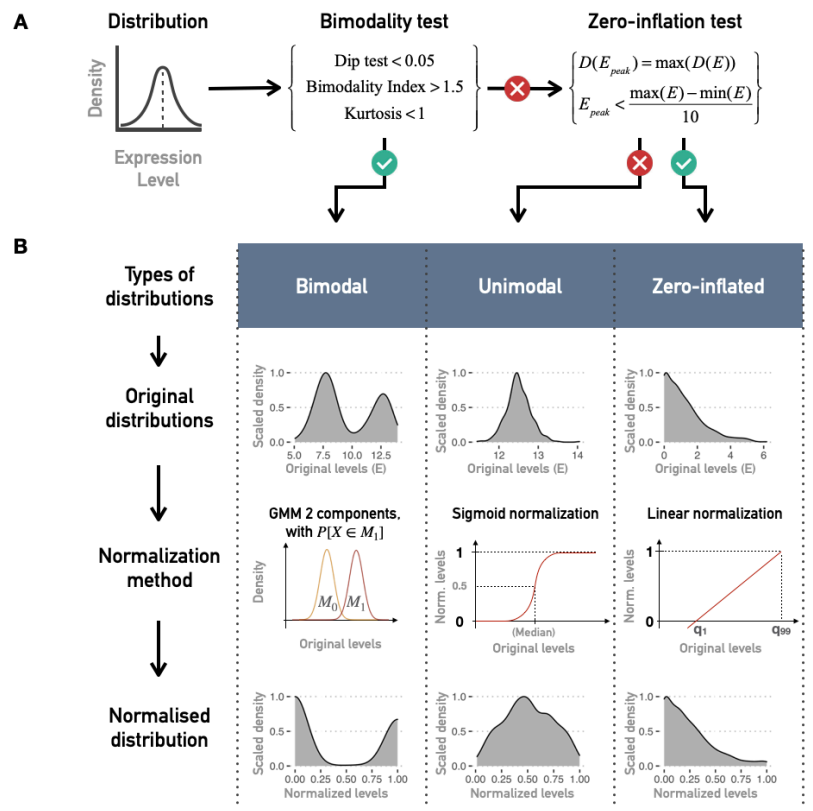

Figure 5.4: **Normalization of continuous data for logical modeling.** (A) Combinations of tests and criteria to classify distributions of continuous data (such as gene expression for one gene and all patients) as bimodal, unimodal or zero-inflated. (B) Normalization methods for each kind of distribution.

model. Because of the type of data used, this personalization method is referred to as **discrete personalization**. It has also been called *strict node variants* in Béal et al. [2019] because this data integration overwrites the logical rules.

Another possible strategy is to modify the initial conditions of the variables of the altered genes according to the results of the normalization. These initial conditions can capture different environmental and genetic conditions. Nevertheless, in the course of the simulation, these variables will be prone to be updated depending on their logical rules. Finally, as MaBoSS uses Gillespie algorithm to explore the STG, data can be mapped to the transition rates of this algorithm. In the simplest case, all transition rates of the model are set to 1, meaning that all possible transitions

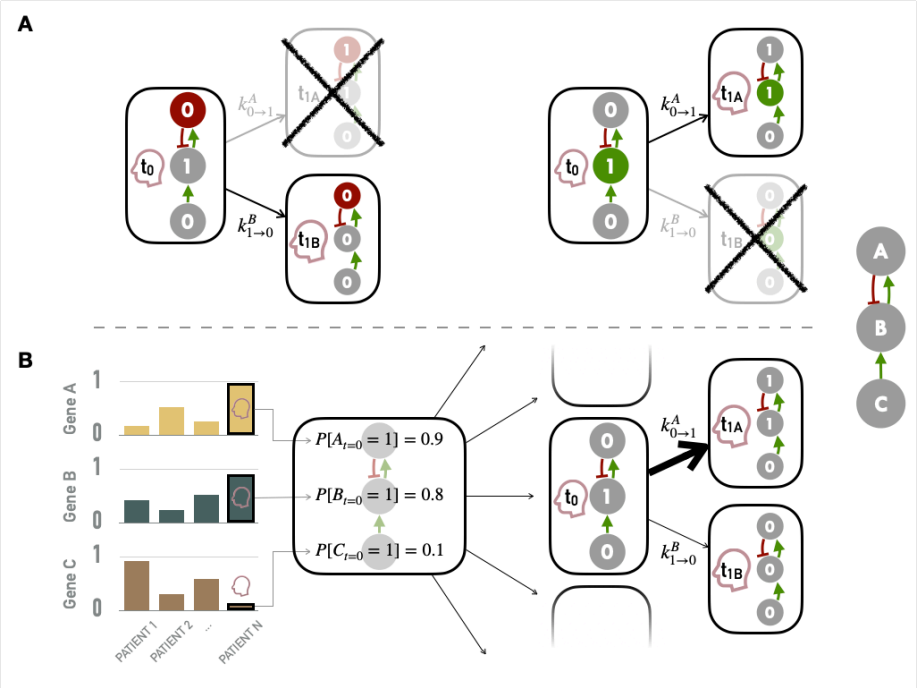

Figure 5.5: **Methods for personalization of logical models.** (A) Personalization with discrete data, such as mutations, with some nodes forced to 0 based on loss of function alteration (left) or 1 based on gain of function/constitutive activation (right). (B) Personalization with continuous data used to define the initial conditions of nodes and to influence the transitions rates and the subsequent probabilities of transition in asynchronous updates.

are equally probable. Alternatively, it is possible to separate the speed of processes by setting the transition rates to different values to account for what is known about the reactions: more probable reactions will have a larger transition rate than less probable reactions [Stoll et al., 2012]. For this, different orders of magnitude for these values can be used. They are set according to the activation status of the node (derived from normalized values) and an amplification factor F, designed to generate a higher relative difference in the transition rates, and are therefore defined for each node i and sample j:

$$k_{i,j}^{0\to1} = F^{2(X_{i,j}^{norm}-0.5)}$$

$$k_{i,j}^{1\to0} = \frac{1}{k_{i,j}^{0\to1}}$$

Thus, if a gene has a value of 1 based on its RNA profile, $k_{0\to1}$ (resp. $k_{1\to0}$) will be 10^2 (resp. 10^{-2}) with an amplification factor of 100. This amplification factor is therefore a hyper-parameter of the method. Very low values will have no impact while higher values will make some transitions almost impossible and the method will then approach the discrete personalization described above. Some quantitative illustrations of the influence of F are provided in Béal et al. [2019]. The integration of continuous data through the initial conditions of the nodes and the transition rates are combined to form a second personalization method called **continuous personalization** and described in Figure 5.5B. This method has also been called *soft node variants* to emphasize its difference with discrete/strict personalization: it may influence the trajectories in the solution state space leading to a change in probabilities of the resulting stable state but it does not overwrite the logical rules. To illustrate a little more explicitly the impact of continuous personalization, if a given node has a normalized value of 0.8 after data processing (based on proteins levels for instance), it will be initialized as 1 in 80% of the stochastic trajectories, its transition rate $k_{0\to1}$ will be increased (favoring its activation) and its transition rate $k_{1\to0}$ will be decreased (hampering its inactivation). These changes increase the probability that this node will remain in an activated state close to the one inferred from the patient's data, while maintaining the validity of its logical rule. Thus, continuous personalization appears as a smoother way to shape logical models' simulations based on patient data. In summary, different types of data can be used, with different integration methods. Note that it is quite natural to use genetic alterations (mutations, CNA) to specify definitive changes in models (such as those of discrete personalization) since this corresponds to biological reality. Conversely, continuous alterations in expression or phosphorylation are subject to modification and regulation, thus justifying their interpretation in a less strong and definitive way (such as continuous personalization). Finally, it follows from these definitions that there are different strategies for personalizing a logical model since discrete and continuous personalizations can each use different types of data; and moreover, these two strategies can be combined. **Except otherwise stated, mutations (resp. RNA or protein) will always be integrated using discrete (resp. continuous) personalization and the joint integration of both types of data will therefore combine both methods.** The relative merits of the different personalization strategies will be discussed below.

5.2 An integration tool for high-dimensional data?

Once the method has been defined, it is imperative to study its validity and possible limitations. This comes down to answering the question: do personalized models capture a biological reality, and in our case do they discriminate between different types of cancer?

5.2.1 Biological relevance in cell lines

These questions can be addressed using cell line data. Using the logical model of cancer pathways from Fumia and Martins [2013], it is possible to study the 663 cell lines from different types of tumors by integrating their processed omics profiles to the generic logical model to obtain as many personalized models. If we focus on the read-out of *Proliferation*, one of the easiest to interpret, there are several ways to study its relevance. For each cell line and each personalization strategy (and corresponding data type) we can define a personalized model and derive the asymptotic value the *Proliferation* node, called *Proliferation* score. This score is therefore a priori different for all cell lines that present a different molecular profile. For the whole population of cell lines, this score can be confronted with other markers of proliferation such as the levels of Ki67 [Miller et al., 2018], here replaced as an example by the RNA levels of the corresponding MKI67 gene. It can then be observed that the simulated *Proliferation* indicator, derived from the personalized models, correlates positively with the biomarker, but only when RNA has been used in the personalization (Figure 5.6A). The correlation makes qualitative sense, but the heterogeneity appears to be very large and most of the variability is not captured by the models. This heterogeneity is also visible by focusing on some types of cancer (Figure 5.6B). Thus this kind of comparison only validates the models' ability to retrieve a RNA biomarker (not used in personalization) when they themselves integrate other RNA data. It is also consistent that scores from models personalized with mutations only have less uniform distributions due to the discrete nature of the data and the many identical profiles: many cell lines are not distinguishable by mutations only.

It is possible to go one step further by comparing these personalized *Proliferation* scores with the doubling time of the cell lines, i.e the time it takes for the cell line population to double. A cell line described as proliferative (high *Proliferation* score) should thus have a low doubling time. This can be observed qualitatively by using a subgroup of cell lines for

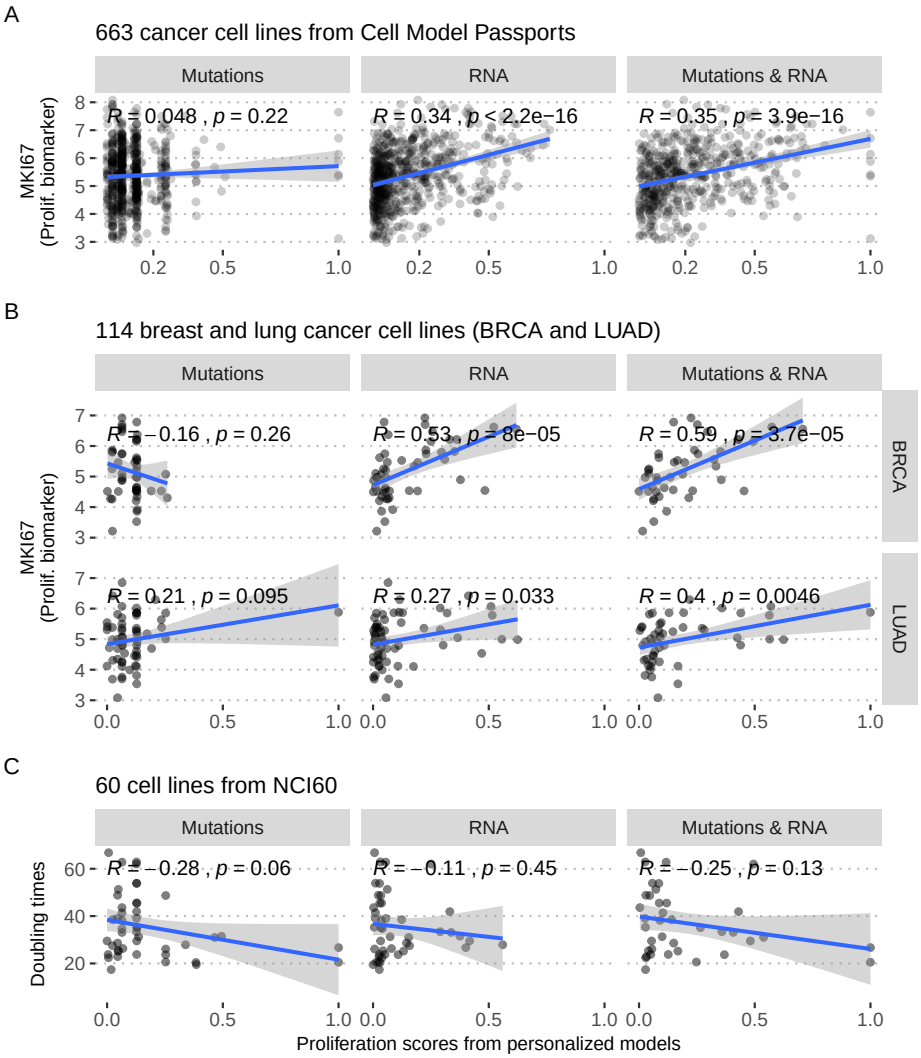

Figure 5.6: **Validation of personalized *Proliferation* scores in cell lines.** (A) Comparison with MKI67 proliferation biomarker for all cancer cell lines. (B) Same with breast (BRCA) and lung (LUAD) cancer only. (C) Comparison with doubling times in a subset of 60 cell lines.

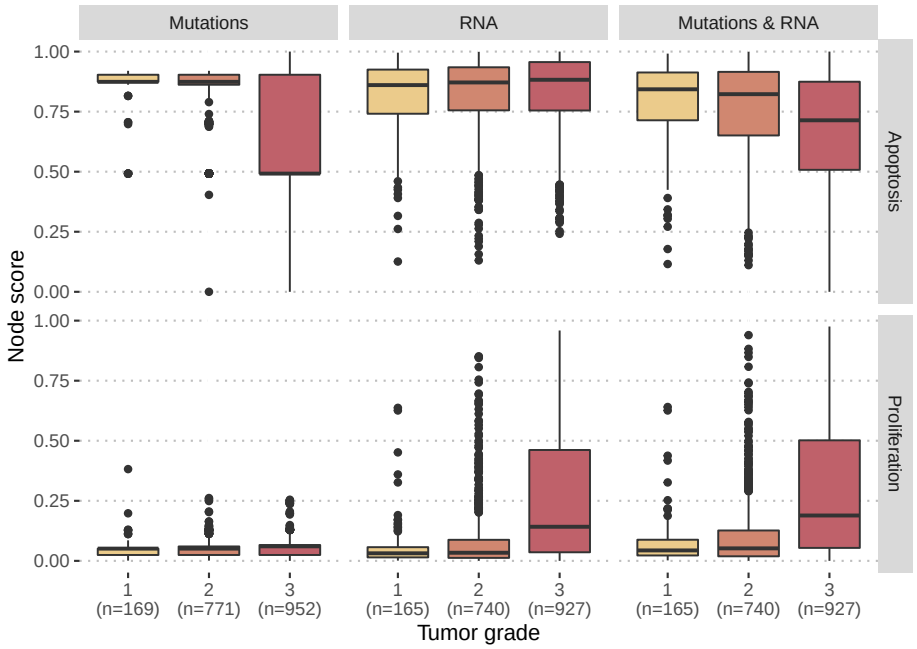

Figure 5.7: **Comparaison of personalized scores with tumour grades for breast cancer patients in METABRIC cohort.** Comparisons are provided for different personalization strategies (with mutations and/or RNA) and two different model nodes (*Proliferation* and *Apoptosis*).

which this information is available (Figure 5.6C). These correlations are not significant and once again summarize a large heterogeneity. Predicting doubling times is, however, a rather difficult task, even with the help of more flexible machine learning methods [Kurilov et al., 2020].

5.2.2 Validation with patient data

By analogy with the validations proposed for other mechanistic models [Fey et al., 2015], it is also possible to evaluate the prognostic value of personalized logical models on patient data. It is also possible to reproduce with patient data analyses of the same type as those previously performed with the MKI67 biomarker, as was done in Béal et al. [2019], but we focus here on the more clinical applications of the personalized mechanistic models. For example, when studying breast cancer patients in the METABRIC cohort, *Proliferation* and *Apoptosis* scores differ according to tumour grade. The more advanced tumours (grade 3) are associated with higher *Proliferation* scores and lower *Apoptosis* scores (Figure 5.7). This is in line with the nat-

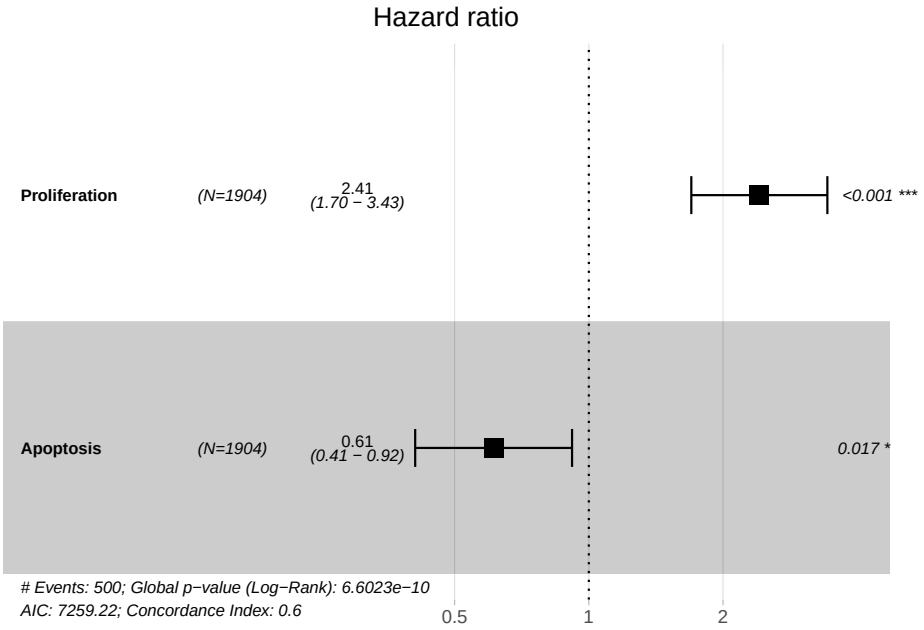

Figure 5.8: **Hazard ratios for *Proliferation* and *Apoptosis* in a survival Cox model in METABRIC cohort.** Higher *Proliferation* (resp. *Apoptosis*) scores correspond to higher (resp. lower) probabilities of death.

ural interpretation that could be given since proliferation is by definition a sign of cancer progression while apoptosis, a programmed death of defective cells, is on the contrary a protective mechanism. While these trends are monotonous and clearly significant for the third strategy using both mutations and RNA ($p < 10^{-12}$ with Jonckheere–Terpstra test for ordered differences among classes, for both nodes), this is not the case when the two types of data are used separately: mutations (resp. RNA) are not sufficient to personalize *Proliferation* (resp. *Apoptosis*) scores in a meaningful way. The personalisation method therefore seems to be able to combine discrete and continuous data in such a way that some of the biological information is preserved.

This comparison to clinical data can be extended to patient survival data in the same cohort. If we focus on the strategy integrating both mutations and RNA, we observe that in a Cox model of survival, *Proliferation* is significantly associated with a higher risk of event while *Apoptosis* is associated with a lower risk, which is again consistent (Figure 5.8). In a more schematic and visual way, it is possible to transform these continuous *Proliferation* and *Apoptosis* scores into binary indicators (using medians) and observe their impact on survival, as has been done in previously men-

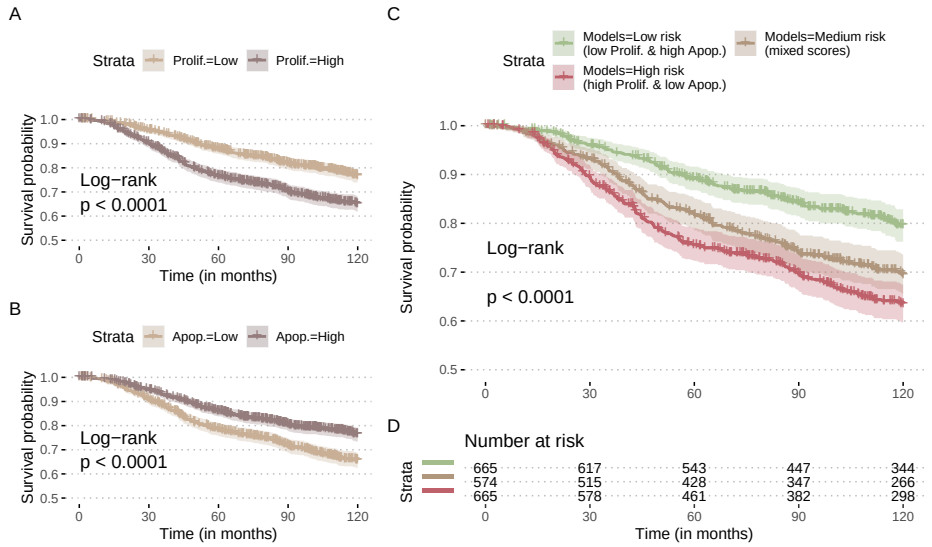

Figure 5.9: **Prognostic value of *Proliferation* scores for breast cancer patients in METABRIC cohort.** (A) Survival curve for overall survival stratified with *Proliferation* scores from personalized models integrating mutations and RNA; scores have binarized based on median and survival censored at 120 months. (B) Same with *Apoptosis* scores. (C) Survival curve stratified with combinations of *Proliferation* and *Apoptosis* scores, based on the same thresholds, and the corresponding number of patients at risk (D).

tioned studies [Fey et al., 2015, Salvucci et al., 2019b]. We then observe the same behaviour for the two personalized scores (Figure 5.9A and B). Interestingly, if we combine the indicators to create groups that are expected to be of very bad prognosis (high *Proliferation*, low *Apoptosis*) or of very good prognosis (low *Proliferation*, high *Apoptosis*), we further discriminate patients and confirm the qualitatively meaningful interpretation of the personalized scores. It should be noted that the clinical validations presented here remain voluntarily simple and quite close to those proposed in similar articles. Discussions and statistical developments will be proposed in Part III.

5.2.3 Perspectives

In summary, this kind of application of personalized models allows the integration of quite heterogeneous and moderately dimensional biological data in a constrained framework that orders the relationships between variables

and guides interpretations. Comparison with external biological or clinical data then makes it possible to verify the absence of major contradictions in the definition of the model. However, the interest of these mechanistic approaches in this type of task appears as quite moderate compared to statistical models. The qualitative aspect is not necessarily compensated here by the integration of knowledge into the structure of the model, especially in examples that use an extremely broad logical model, which has not been specifically designed for the problems to which it is applied. It is then necessary to study the application of these personalized models to more suitable problems, in which the explicitly mechanistic nature of the models can be exploited, for instance the response to perturbations.

Chapter 6

Personalized logical models to study an interpret drug response

"TBD."
?

The notion of modeling

Publications

This chapter extends the method presented in the previous chapter to investigate drug response with personalized logical models. The first application to cell lines of all cancer types was presented orally at ISMB2020 in Basel but is not published. The example concerning BRAF in melanomas and colorectal cancers is under revision and the corresponding pre-print is available as Béal et al. [2020]. Finally, the work on prostate cancer presented in a third section has also been submitted.

6.1 One step further with drugs

6.1.1 Methods

6.1.2 Brute force

6.2 A case study on BRAF in melanoma and colorectal cancers

6.3 Limitations and perspectives illustrated by a prostate cancer study

A model is first of all an ambiguous object and a polysemous word. It therefore seems necessary to start with a semantic study. Among the many

APPENDIX A

About datasets

A.1 Cell lines

Several analyses in previous chapters are based on data derived from cell lines. Among the different databases, the ones used in the thesis are briefly described below. Please refer to corresponding references for additional details.

A.1.1 Omics profiles

The omics profiles of cancer cell lines have been downloaded from Cell Model Passports [van der Meer et al., 2019] containing genotypic and phenotypic information about more than 1000 cell lines. Among the available data used in this thesis are the exome sequencing, copy number variations and RNA-sequencing.

A.1.2 Drug screenings

Information about response to treatments is retrieved from Genomics of Drug Sensitivity in Cancer Database (GDSC, Yang et al. [2012]). In order to allow detailed analyses at the level of cancer types, we will restrict ourselves here to tissues represented by at least 20 cell lines and highlighted in dark grey in Figure A.1A. Most of the 663 cell lines in this subcohort have a complete profile with all omics data (mutations, CNA and expression) and

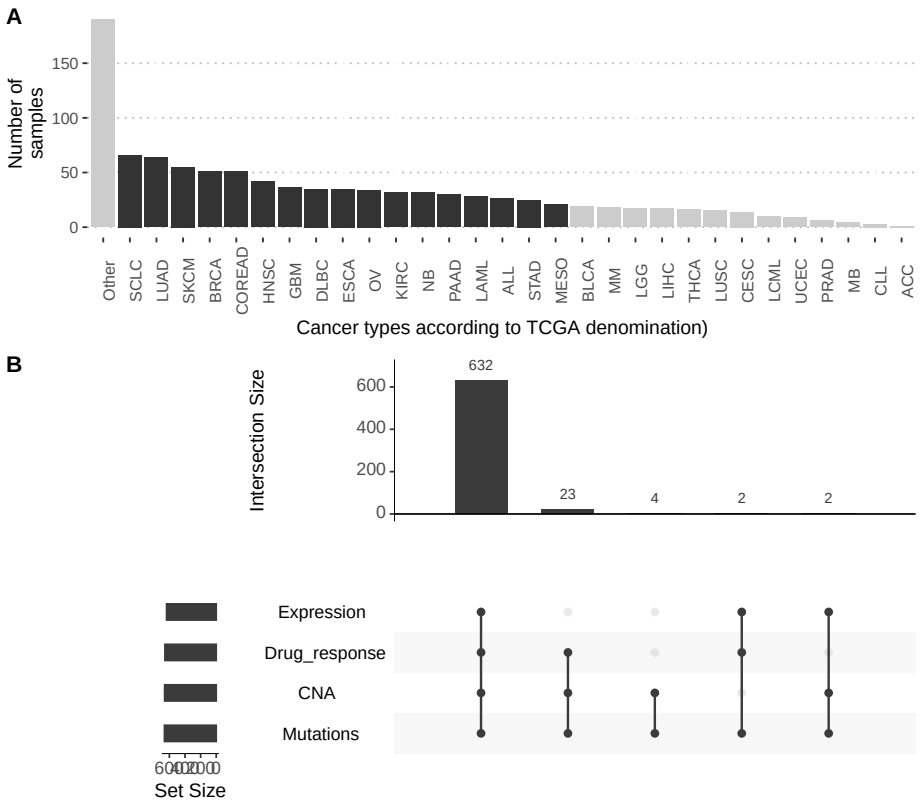

Figure A.1: **Distribution of cancer types and data types in GDSC-associated dataset.** (A) Distribution of cell lines per cancer types, highlighting the ones selected in this thesis with more than 20 cell lines. (B) Availibility of data for the 663 selected cell lines in 17 different cancer types.

drug responses. However, not all cell lines have necessarily been tested for all drugs.

The cell lines are treated with increasing concentration of drugs and the viability of the cell line relative to untreated control is measured. The dose-response relative viability curve is fitted and then used to compute the half maximal inhibitory concentration (IC_{50}) and the area under the dose-response curve (AUC) [Vis et al., 2016], both being represented in Figure. Since the IC_{50} values are often extrapolated outside the concentration range actually tested, we will focus on the AUC metric for all validation with drug screening data. AUC is a value between 0 and 1: values close to 1 mean that the relative viability has not been decreased, and lower values correspond to increased sensitivity to inhibitions. In cases where the ranges of concentrations tested for different drugs vary, comparison of their AUC

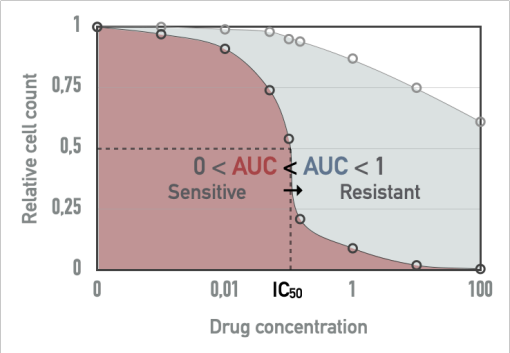

Figure A.2: **Drug screening metrics in cell lines.** Based on a tested drug concentration range, IC_{50} and area under the dose-response curve (AUC) can be computed. For a given drug, red AUC corresponds to a more sensitive cell line than blue AUC.

values does not have a simple and straightforward interpretation.

A.1.3 CRISPR-Cas9 screening

On top the previous drug response characterization, some CRISPR-Cas9 screenings have been performed on cancer cell lines. Very basically, this involves using single-guide RNAs (sgRNAs) to direct the targeted inhibition of certain genes. Conceptually, screening is not very different from drug screening since it allows the sensitivity of cell lines to the inhibition of certain targets to be studied. However, this technology makes it possible to target many more different genes since it is based on RNA guide synthesis and not on the existence of drugs with an affinity for the target of interest. Schematically, sreening is therefore broader (thousands of genes), less biased (any gene can be targeted a priori) and more precise (much lower off-target effect).

Among the various databases available, the ones used in this thesis have been downloaded from Cell Model Passports and come from Sanger Institute [Behan et al., 2019] and Broad Institute [Meyers et al., 2017]. Both databases present CRISPR inhibition results for thousands of genes for a few hundred cell lines among those presented in the previous section. The Sanger dataset for instance includes 324 cell lines, and 238 in common with the subcohort previously described in the previous section and in Figure A.1.

Among the different metrics, the examples presented in this thesis will focus on scaled Bayesian factors to assess the effect of CRISPR targeting

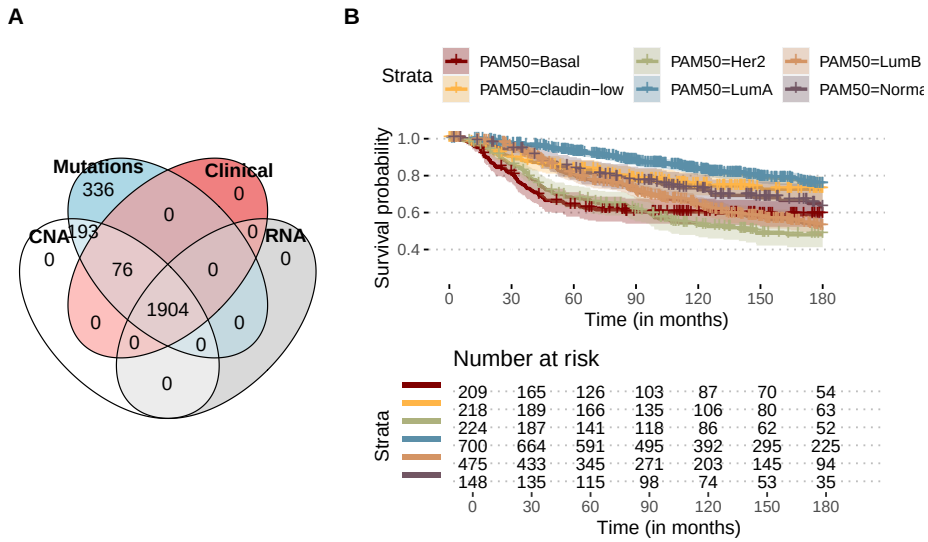

Figure A.3: **Available omics and survival in METABRIC Breast Cancer dataset**. (A) Number of patients for each omics type and their combinations, depicted as a Venn diagram. (B) Overall survival probability for all patients with clinical follow-up, stratified per breast cancer PAM50 subtype; administrative censoring at 180 months.

of genes. These scores are computed based on the fold change distribution of sgRNA [Hart and Moffat, 2016]. The highest values indicate that the targeted gene is essential to the cell fitness.

A.2 Patient-derived xenografts

Distributions and some figures

A.3 Patients

A.3.1 METABRIC

METABRIC dataset is large breast cancer dataset with more than 2000 patients [Pereira et al., 2016]. Mutations, CNA, expression (transcriptomics micro-array) and clinical data are available for a majority of patients (Figure A.3A), with 1904 patients for whom all the data is available. One of the particular features of these data is to propose a very long clinical follow-up, over more than 10 years (Figure A.3B).

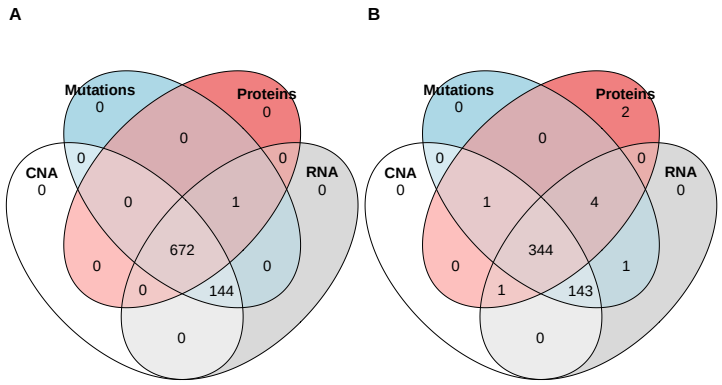

Figure A.4: **Available omics for TCGA Breast and Prostate cancer**. (A) Number of patients for each omics type and their combinations, depicted as a Venn diagram, in TCGA BRCA (Breast Invasive Carcinoma) study. (B) Same for the TCGA PRAD (Prostate Adenocarcinoma) study.

A.3.2 TCGA: Breast cancer

Another reference database for breast cancer is the one from the TCGA consortium [Network et al., 2012]. The cohort is smaller than METABRIC and its clinical follow-up is more limited. In contrast, the omics data are more comprehensive and include RNA sequencing and relative quantification ofsproteins with RPPA technology (Figure A.4A).

A.3.3 TCGA: Prostate cancer

Similarly, for prostate cancer, reference can be made to data from the TCGA study [Abeshouse et al., 2015], which has the same type of data but for a smaller number of patients than the breast cancer (Figure A.4B).

APPENDIX B

About logical models

Several logical models of cancer are used in this thesis and some additional descriptive elements about them are given below.

B.1 Generic logical model of cancer pathways

For this thesis, a published Boolean model from [Fumia and Martins, 2013] has been used to illustrate our PROFILE methodology. This regulatory network summarizes several key players and pathways involved in cancer mechanisms such as RTKs, PI3K/AKT, WNT/β-catenin, TGF-β/Smads, Rb, HIF-1, p53 and ATM/ATR. An input node *Acidosis* has been added, along with an output node *Proliferation* used as a readout for the activity of any of the cyclins (*CyclinA*, *CyclinB*, *CyclinD* and *CyclinE*). This slightly extended model contains 98 nodes and 254 edges and its inputs are *Acidosis*, *Nutrients*, *Growth Factors* (GFs), *Hypoxia*, *TNFalpha*, *ROS*, *PTEN*, *p14ARF*, *GLI*, *FOXO*, *APC* and *MAX*. Its outputs are *Proliferation*, *Apoptosis*, *DNA_repair*, *DNA_damage*, *VEGF*, *Lactic_acid*, *GSH*, *GLUT1* and *COX412*.

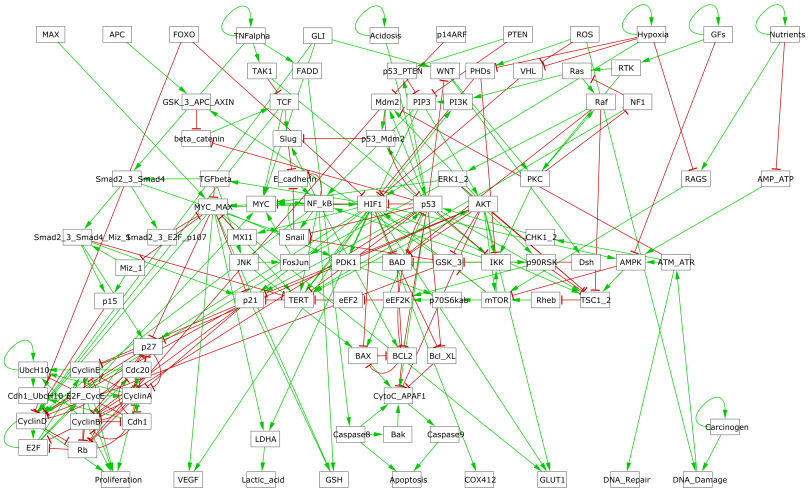

Figure B.1: **GINSIM representation of the logical model described in Fumia and Martins [2013].**

B.2 Logical model of BRAF pathways in melanoma and colorectal cancer

B.3 Logical model of prostate cancer

APPENDIX C

About causality

C.1 Theoretical framework

References

Bibliography

Adam Abeshouse, Jaeil Ahn, Rehan Akbani, Adrian Ally, Samirkumar Amin, Christopher D Andry, Matti Annala, Armen Aprikian, Joshua Armenia, Arshi Arora, et al. The molecular taxonomy of primary prostate cancer. *Cell*, 163(4):1011–1025, 2015.

Wassim Abou-Jaoudé, Pauline Traynard, Pedro T Monteiro, Julio Saez-Rodriguez, Tomáš Helikar, Denis Thieffry, and Claudine Chaouiya. Logical modeling and dynamical analysis of cellular networks. *Frontiers in genetics*, 7:94, 2016.

Ivan A. Adzhubei, Steffen Schmidt, Leonid Peshkin, Vasily E. Ramensky, Anna Gerasimova, Peer Bork, Alexey S. Kondrashov, and Shamil R. Sunyaev. A method and server for predicting damaging missense mutations. *Nature Methods*, 7(4):248–249, April 2010. ISSN 1548-7091. doi: 10.1038/nmeth0410-248. URL https://www.nature.com/nmeth/journal/v7/n4/full/nmeth0410-248.html.

István Albert, Juilee Thakar, Song Li, Ranran Zhang, and Réka Albert. Boolean network simulations for life scientists. *Source code for biology and medicine*, 3(1):16, 2008.

Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. Molecular biology of the cell. garland science. *New York*, 1392, 2007.

Philipp M Altrock, Lin L Liu, and Franziska Michor. The mathematics of cancer: integrating quantitative models. *Nature Reviews Cancer*, 15(12): 730–745, 2015.

Alexander RA Anderson and Vito Quaranta. Integrative mathematical oncology. *Nature Reviews Cancer*, 8(3):227–234, 2008.

Robyn P Araujo and DL Sean McElwain. A history of the study of solid tumour growth: the contribution of mathematical modelling. *Bulletin of mathematical biology*, 66(5):1039–1091, 2004.

Peter Armitage and Richard Doll. The age distribution of cancer and a multi-stage theory of carcinogenesis. *British journal of cancer*, 8(1):1, 1954.

Daniela M Bailer-Jones. Scientists' thoughts on scientific models. *Perspectives on Science*, 10(3):275–301, 2002.

Ruth E Baker, Jose-Maria Pena, Jayaratnam Jayamohan, and Antoine Jérusalem. Mechanistic models versus machine learning, a fight worth fighting for the biological community? *Biology letters*, 14(5):20170660, 2018.

Albert-Laszlo Barabasi and Zoltan N Oltvai. Network biology: understanding the cell's functional organization. *Nature reviews genetics*, 5(2):101–113, 2004.

Dominique Barbolosi, Joseph Ciccolini, Bruno Lacarelle, Fabrice Barlési, and Nicolas André. Computational oncology—mathematical modelling of drug regimens for precision medicine. *Nature reviews Clinical oncology*, 13(4):242, 2016.

Emmanuel Barillot, Laurence Calzone, Philippe Hupe, Jean-Philippe Vert, and Andrei Zinovyev. *Computational systems biology of cancer*. CRC Press, 2012.

Jonas Béal, Arnau Montagud, Pauline Traynard, Emmanuel Barillot, and Laurence Calzone. Personalization of Logical Models With Multi-Omics Data Allows Clinical Stratification of Patients. *Frontiers in Physiology*, 2019. ISSN 1664-042X. doi: 10.3389/fphys.2018.01965. URL https://www.frontiersin.org/article/10.3389/fphys.2018.01965/full.

Jonas Béal, Lorenzo Pantolini, Vincent Noël, Emmanuel Barillot, and Laurence Calzone. Personalized logical models to investigate cancer response to braf treatments in melanomas and colorectal cancers. *bioRxiv*, 2020.

Fiona M Behan, Francesco Iorio, Gabriele Picco, Emanuel Gonçalves, Charlotte M Beaver, Giorgia Migliardi, Rita Santos, Yanhua Rao, Francesco Sassi, Marika Pinnelli, et al. Prioritization of cancer therapeutic targets using crispr–cas9 screens. *Nature*, 568(7753):511, 2019.

Nicola Bellomo, NK Li, and Ph K Maini. On the foundations of cancer modelling: selected topics, speculations, and perspectives. *Mathematical Models and Methods in Applied Sciences*, 18(04):593–646, 2008.

Sébastien Benzekry, Clare Lamont, Afshin Beheshti, Amanda Tracz, John ML Ebos, Lynn Hlatky, and Philip Hahnfeldt. Classical mathematical models for description and prediction of experimental tumor growth. *PLoS Comput Biol*, 10(8):e1003800, 2014.

Upinder S Bhalla and Ravi Iyengar. Emergent properties of networks of biological signaling pathways. *Science*, 283(5400):381–387, 1999.

Raffaella Bianucci, Antonio Perciaccante, Philippe Charlier, Otto Appenzeller, and Donatella Lippi. Earliest evidence of malignant breast cancer in renaissance paintings. *The Lancet Oncology*, 19(2):166–167, 2018.

Erhan Bilal, Janusz Dutkowski, Justin Guinney, In Sock Jang, Benjamin A Logsdon, Gaurav Pandey, Benjamin A Sauerwine, Yishai Shimoni, Hans Kristian Moen Vollan, Brigham H Mecham, et al. Improving breast cancer survival analysis through competition-based multidimensional modeling. *PLoS computational biology*, 9(5), 2013.

Mehdi Bouhaddou, Anne Marie Barrette, Alan D Stern, Rick J Koch, Matthew S DiStefano, Eric A Riesel, Luis C Santos, Annie L Tan, Alex E Mertz, and Marc R Birtwistle. A mechanistic pan-cancer pathway model informed by multi-omics data interprets stochastic cell fate responses to drugs and mitogens. *PLoS computational biology*, 14(3):e1005985, 2018.

Ivana Bozic, Tibor Antal, Hisashi Ohtsuki, Hannah Carter, Dewey Kim, Sining Chen, Rachel Karchin, Kenneth W Kinzler, Bert Vogelstein, and Martin A Nowak. Accumulation of driver and passenger mutations during tumor progression. *Proceedings of the National Academy of Sciences*, 107 (43):18545–18550, 2010.

Peter Allen Braithwaite and Dace Shugg. Rembrandt's bathsheba: the dark shadow of the left breast. *Annals of the Royal College of Surgeons of England*, 65(5):337, 1983.

Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L Siegel, Lindsey A Torre, and Ahmedin Jemal. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 68(6):394–424, 2018.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001a.

Leo Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231, 2001b.

Helen M Byrne. Dissecting cancer through mathematics: from the cell to the animal model. *Nature Reviews Cancer*, 10(3):221–230, 2010.

Jonas Béal, Elizabeth Rémy, and Laurence Calzone. Modélisation logique et données omiques : de la construction des modèles à la médecine personnalisée. In Elisabeth Rémy and Cédric Lhoussaine, editors, *Approche symbolique de la modélisation et de l'analyse des systèmes biologiques*. ISTE, 2020.

Laurence Calzone, Laurent Tournier, Simon Fourquet, Denis Thieffry, Boris Zhivotovsky, Emmanuel Barillot, and Andrei Zinovyev. Mathematical modelling of cell-fate decision in response to death receptor engagement. *PLoS computational biology*, 6(3), 2010.

Laurence Calzone, Emmanuel Barillot, and Andrei Zinovyev. Logical versus kinetic modeling of biological networks: applications in cancer research. *Current Opinion in Chemical Engineering*, 21:22–31, 2018.

Debyani Chakravarty, Jianjiong Gao, Sarah Phillips, Ritika Kundra, Hongxin Zhang, Jiaojiao Wang, Julia E. Rudolph, Rona Yaeger, Tara Soumerai, Moriah H. Nissan, Matthew T. Chang, Sarat Chandarlapaty, Tiffany A. Traina, Paul K. Paik, Alan L. Ho, Feras M. Hantash, Andrew Grupe, Shrujal S. Baxi, Margaret K. Callahan, Alexandra Snyder, Ping Chi, Daniel C. Danila, Mrinal Gounder, James J. Harding, Matthew D. Hellmann, Gopa Iyer, Yelena Y. Janjigian, Thomas Kaley, Douglas A. Levine, Maeve Lowery, Antonio Omuro, Michael A. Postow, Dana Rathkopf, Alexander N. Shoushtari, Neerav Shukla, Martin H. Voss, Ederlinda Paraiso, Ahmet Zehir, Michael F. Berger, Barry S. Taylor, Leonard B. Saltz, Gregory J. Riely, Marc Ladanyi, David M. Hyman, José Baselga, Paul Sabbatini, David B. Solit, and Nikolaus Schultz. OncoKB: A Precision Oncology Knowledge Base. *JCO Precision Oncology*, (1):1–16, May 2017. ISSN 2473-4284. doi: 10.1200/PO.17.00011. URL http://ascopubs.org/doi/full/10.1200/PO.17.00011.

David P. A. Cohen, Loredana Martignetti, Sylvie Robine, Emmanuel Barillot, Andrei Zinovyev, and Laurence Calzone. Mathematical Modelling of Molecular Pathways Enabling Tumour Cell Invasion and Migration. *PLOS Computational Biology*, 11(11):e1004571, November 2015. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1004571. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004571.

Randall J Cohrs, Tyler Martin, Parviz Ghahramani, Luc Bidaut, Paul J Higgins, and Aamir Shahzad. Translational medicine definition by the european society for translational medicine, 2015.

Collins. *The Collins English Dictionary*. HarperCollins, 2020. URL https://www.collinsdictionary.com/dictionary/english/model. Model.

Samuel Collombet, Chris van Oevelen, Jose Luis Sardina Ortega, Wassim Abou-Jaoude, Bruno Di Stefano, Morgane Thomas-Chollier, Thomas Graf, and Denis Thieffry. Logical modeling of lymphoid and myeloid cell specification and transdifferentiation. *Proceedings of the National Academy of Sciences*, 114(23):5792–5799, June 2017. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1610622114. URL https://www.pnas.org/content/114/23/5792.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

Francis Crick. Central dogma of molecular biology. *Nature*, 227(5258): 561–563, 1970.

Christina Curtis, Sohrab P. Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar M. Rueda, Mark J. Dunning, Doug Speed, Andy G. Lynch, Shamith Samarajiwa, Yinyin Yuan, Stefan Gräf, Gavin Ha, Gholamreza Haffari, Ali Bashashati, Roslin Russell, Steven McKinney, Anita Langerød, Andrew Green, Elena Provenzano, Gordon Wishart, Sarah Pinder, Peter Watson, Florian Markowetz, Leigh Murphy, Ian Ellis, Arnie Purushotham, Anne-Lise Børresen-Dale, James D. Brenton, Simon Tavaré, Carlos Caldas, and Samuel Aparicio. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486 (7403):346–352, April 2012. ISSN 0028-0836. doi: 10.1038/nature10983. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3440846/.

Helen Davies, Graham R Bignell, Charles Cox, Philip Stephens, Sarah Edkins, Sheila Clegg, Jon Teague, Hayley Woffendin, Mathew J Garnett, William Bottomley, et al. Mutations of the braf gene in human cancer. *Nature*, 417(6892):949–954, 2002.

Hidde De Jong. Modeling and simulation of genetic regulatory systems: a literature review. *Journal of computational biology*, 9(1):67–103, 2002.

Thomas S Deisboeck, Le Zhang, Jeongah Yoon, and Jose Costa. In silico cancer modeling: is it ready for prime time? *Nature Clinical Practice Oncology*, 6(1):34–42, 2009.

Antonio Del Sol, Rudi Balling, Lee Hood, and David Galas. Diseases as network perturbations. *Current opinion in biotechnology*, 21(4):566–571, 2010.

Li Ding, Matthew H Bailey, Eduard Porta-Pardo, Vesteinn Thorsson, Antonio Colaprico, Denis Bertrand, David L Gibbs, Amila Weerasinghe, Kuanlin Huang, Collin Tokheim, et al. Perspective on oncogenic processes at the end of the beginning of cancer genomics. *Cell*, 173(2):305–320, 2018.

Eytan Domany. Using high-throughput transcriptomic data for prognosis: a critical overview and perspectives. *Cancer research*, 74(17):4612–4621, 2014.

Yotam Drier, Michal Sheffer, and Eytan Domany. Pathway-based personalized analysis of cancer. *Proceedings of the National Academy of Sciences*, 110(16):6388–6393, 2013.

Renato Dulbecco. A turning point in cancer research: sequencing the human genome. *Science*, 231:1055–1057, 1986.

Mohamed Elati, Pierre Neuvial, Monique Bolotin-Fukuhara, Emmanuel Barillot, François Radvanyi, and Céline Rouveirol. LICORN: learning cooperative regulation networks from gene expression data. *Bioinformatics*, 23(18):2407–2414, September 2007. ISSN 1367-4803. doi: 10.1093/bioinformatics/btm352. URL https://academic.oup.com/bioinformatics/article/23/18/2407/236890.

Adrien Fauré, Aurélien Naldi, Claudine Chaouiya, and Denis Thieffry. Dynamical analysis of a generic boolean model for the control of the mammalian cell cycle. *Bioinformatics*, 22(14):e124–e131, 2006.

Dana Ferranti, David Krane, and David Craft. The value of prior knowledge in machine learning of complex network systems. *Bioinformatics*, 33(22):3610–3618, 2017.

Dirk Fey, Melinda Halasz, Daniel Dreidax, Sean P Kennedy, Jordan F Hastings, Nora Rauch, Amaya Garcia Munoz, Ruth Pilkington, Matthias Fischer, Frank Westermann, et al. Signaling pathway models as biomarkers: Patient-specific simulations of jnk activity predict the survival of neuroblastoma patients. *Sci. Signal.*, 8(408):ra130–ra130, 2015.

Gary William Flake. *The computational beauty of nature: Computer explorations of fractals, chaos, complex systems, and adaptation.* MIT press, 1998.

Åsmund Flobak, Anaïs Baudot, Elisabeth Remy, Liv Thommesen, Denis Thieffry, Martin Kuiper, and Astrid Lægreid. Discovery of drug synergies in gastric cancer cells predicted by logical modeling. *PLoS computational biology*, 11(8), 2015.

Andrew Cadle Fowler, Anna C Fowler, and AC Fowler. *Mathematical models in the applied sciences*, volume 17. Cambridge University Press, 1997.

Roman Frigg and Stephan Hartmann. Models in science. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2020 edition, 2020.

Fabian Fröhlich, Thomas Kessler, Daniel Weindl, Alexey Shadrin, Leonard Schmiester, Hendrik Hache, Artur Muradyan, Moritz Schütte, Ji-Hyun Lim, Matthias Heinig, et al. Efficient parameter estimation enables the prediction of drug response using a mechanistic pan-cancer pathway model. *Cell Systems*, 7(6):567–579, 2018.

Herman F Fumia and Marcelo L Martins. Boolean network model for cancer pathways: predicting carcinogenesis and targeted therapy outcomes. *PloS one*, 8(7), 2013.

Hui Gao, Joshua M Korn, Stéphane Ferretti, John E Monahan, Youzhen Wang, Mallika Singh, Chao Zhang, Christian Schnell, Guizhi Yang, Yun Zhang, et al. High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response. *Nature medicine*, 21 (11):1318, 2015.

Daniel T Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, 22(4):403–434, December 1976. ISSN 0021-9991. doi: 10.1016/0021-9991(76)90041-3.

Luca Grieco, Laurence Calzone, Isabelle Bernard-Pierrot, François Radvanyi, Brigitte Kahn-Perles, and Denis Thieffry. Integrative modelling of the influence of mapk network on cancer cell fate decision. *PLoS computational biology*, 9(10):e1003286, 2013.

William C Hahn, Christopher M Counter, Ante S Lundberg, Roderick L Beijersbergen, Mary W Brooks, and Robert A Weinberg. Creation of human tumour cells with defined genetic elements. *Nature*, 400(6743): 464–468, 1999.

Steven I Hajdu. A note from history: landmarks in history of cancer, part 1. *Cancer*, 117(5):1097–1102, 2011a.

Steven I Hajdu. A note from history: landmarks in history of cancer, part 2. *Cancer*, 117(12):2811–2820, 2011b.

Steven I Hajdu. A note from history: landmarks in history of cancer, part 3. *Cancer*, 118(4):1155–1168, 2012a.

Steven I Hajdu. A note from history: landmarks in history of cancer, part 4. *Cancer*, 118(20):4914–4928, 2012b.

Steven I Hajdu and Farbod Darvishian. A note from history: landmarks in history of cancer, part 5. *Cancer*, 119(8):1450–1466, 2013.

Steven I Hajdu and Manjunath Vadmal. A note from history: Landmarks in history of cancer, part 6. *Cancer*, 119(23):4058–4082, 2013.

Douglas Hanahan and Robert A Weinberg. The hallmarks of cancer. *cell*, 100(1):57–70, 2000.

Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *cell*, 144(5):646–674, 2011.

Traver Hart and Jason Moffat. Bagel: a computational framework for identifying essential genes from pooled library screens. *BMC bioinformatics*, 17(1):164, 2016.

J. A. Hartigan and P. M. Hartigan. The Dip Test of Unimodality. *The Annals of Statistics*, 13(1):70–84, March 1985. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1176346577. URL http://projecteuclid.org/euclid.aos/1176346577.

Yehudit Hasin, Marcus Seldin, and Aldons Lusis. Multi-omics approaches to disease. *Genome biology*, 18(1):83, 2017.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

Suzanne Hector, Markus Rehm, Jasmin Schmid, Joan Kehoe, Niamh McCawley, Patrick Dicker, Frank Murray, Deborah McNamara, Elaine W Kay, Caoimhin G Concannon, et al. Clinical application of a systems model of apoptosis execution for the prediction of colorectal cancer therapy responses and personalisation of therapy. *Gut*, 61(5):725–733, 2012.

Michelle Heijblom, Linda M Meijer, Ton G van Leeuwen, Wiendelt Steenbergen, and Srirang Manohar. Monte carlo simulations shed light on bathsheba's suspect breast. *Journal of biophotonics*, 7(5):323–331, 2014.

Tomáš Helikar, John Konvalina, Jack Heidel, and Jim A Rogers. Emergent decision-making in biological signal transduction networks. *Proceedings of the National Academy of Sciences*, 105(6):1913–1918, 2008.

Tomáš Helikar, Bryan Kowal, Sean McClenathan, Mitchell Bruckner, Thaine Rowley, Alex Madrahimov, Ben Wicks, Manish Shrestha, Kahani Limbu, and Jim A Rogers. The cell collective: toward an open and collaborative approach to systems biology. *BMC systems biology*, 6(1): 96, 2012.

N Lynn Henry and Daniel F Hayes. Cancer biomarkers. *Molecular oncology*, 6(2):140–146, 2012.

MA Hernán and JM Robins. Causal inference: What if. *Boca Raton: Chapman & Hill/CRC*, 2020.

C Gordon Hewitt. Conservation of wild life in canada, 1917.

Manuel Hidalgo, Frederic Amant, Andrew V Biankin, Eva Budinská, Annette T Byrne, Carlos Caldas, Robert B Clarke, Steven de Jong, Jos Jonkers, Gunhild Mari Mælandsmo, et al. Patient-derived xenograft models: an emerging platform for translational cancer research. *Cancer discovery*, 4(9):998–1013, 2014.

Steven M Hill, Laura M Heiser, Thomas Cokelaer, Michael Unger, Nicole K Nesser, Daniel E Carlin, Yang Zhang, Artem Sokolov, Evan O Paull, Chris K Wong, et al. Inferring causal molecular networks: empirical assessment through a community-based effort. *Nature methods*, 13(4): 310, 2016.

Jorrit J Hornberg, Frank J Bruggeman, Hans V Westerhoff, and Jan Lankelma. Cancer: a systems biology disease. *Biosystems*, 83(2-3):81–90, 2006.

Sui Huang, Ingemar Ernberg, and Stuart Kauffman. Cancer attractors: a systems view of tumors from a gene network dynamics and developmental perspective. 20(7):869–876, 2009.

Francesco Iorio, Theo A Knijnenburg, Daniel J Vis, Graham R Bignell, Michael P Menden, Michael Schubert, Nanne Aben, Emanuel Gonçalves,

Syd Barthorpe, Howard Lightfoot, et al. A landscape of pharmacogenomic interactions in cancer. *Cell*, 166(3):740–754, 2016.

Hemant Ishwaran. Variable importance in binary regression trees and forests. *Electronic Journal of Statistics*, 1:519–537, 2007.

François Jacob and Jacques Monod. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of molecular biology*, 3(3):318–356, 1961.

Katarzyna Jastrzebski, Bram Thijssen, Roelof JC Kluin, Klaas de Lint, Ian J Majewski, Roderick L Beijersbergen, and Lodewyk FA Wessels. Integrative modeling identifies key determinants of inhibitor sensitivity in breast cancer cell lines. *Cancer research*, 78(15):4396–4410, 2018.

Ahmedin Jemal, Elizabeth M Ward, Christopher J Johnson, Kathleen A Cronin, Jiemin Ma, A Blythe Ryerson, Angela Mariotto, Andrew J Lake, Reda Wilson, Recinda L Sherman, et al. Annual report to the nation on the status of cancer, 1975–2014, featuring survival. *JNCI: Journal of the National Cancer Institute*, 109(9):djx030, 2017.

Siân Jones, Xiaosong Zhang, D Williams Parsons, Jimmy Cheng-Ho Lin, Rebecca J Leary, Philipp Angenendt, Parminder Mankoo, Hannah Carter, Hirohiko Kamiyama, Antonio Jimeno, et al. Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *science*, 321(5897):1801–1806, 2008.

Minoru Kanehisa, Susumu Goto, Yoko Sato, Miho Furumichi, and Mao Tanabe. Kegg for integration and interpretation of large-scale molecular data sets. *Nucleic acids research*, 40(D1):D109–D114, 2012.

Stuart Kauffman. Homeostasis and differentiation in random genetic control networks. *Nature*, 224(5215):177–178, 1969.

Faiz M Khan, Stephan Marquardt, Shailendra K Gupta, Susanne Knoll, Ulf Schmitz, Alf Spitschak, David Engelmann, Julio Vera, Olaf Wolkenhauer, and Brigitte M Pützer. Unraveling a tumor type-specific regulatory core underlying e2f1-mediated epithelial-mesenchymal transition to predict receptor protein signatures. *Nature communications*, 8(1):1–15, 2017.

Hiroaki Kitano. Computational systems biology. *Nature*, 420(6912):206–210, 2002.

Hannes Klarner, Adam Streck, and Heike Siebert. Pyboolnet: a python package for the generation, analysis and visualization of boolean networks. *Bioinformatics*, 33(5):770–772, 2016.

Theo A Knijnenburg, Gunnar W Klau, Francesco Iorio, Mathew J Garnett, Ultan McDermott, Ilya Shmulevich, and Lodewyk FA Wessels. Logic models to predict continuous outputs based on binary inputs with an application to personalized cancer therapy. *Scientific reports*, 6(1):1–14, 2016.

Alfred G Knudson. Mutation and cancer: statistical study of retinoblastoma. *Proceedings of the National Academy of Sciences*, 68(4):820–823, 1971.

Tarja Knuuttila and Andrea Loettgers. Modelling as indirect representation? the lotka–volterra model revisited. *The British Journal for the Philosophy of Science*, 68(4):1007–1036, 2017.

Pamela K Kreeger and Douglas A Lauffenburger. Cancer systems biology: a network modeling perspective. *Carcinogenesis*, 31(1):2–8, 2010.

Prateek Kumar, Steven Henikoff, and Pauline C. Ng. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols*, 4(7):1073, June 2009. ISSN 1750-2799. doi: 10.1038/nprot.2009.86. URL https://www.nature.com/articles/nprot.2009.86.

I Kuperstein, E Bonnet, HA Nguyen, D Cohen, E Viara, L Grieco, S Fourquet, L Calzone, C Russo, M Kondratova, et al. Atlas of cancer signalling network: a systems biology resource for integrative analysis of cancer data with google maps. *Oncogenesis*, 4(7):e160–e160, 2015.

Roman Kurilov, Benjamin Haibe-Kains, and Benedikt Brors. Assessment of modelling strategies for drug response prediction in cell lines and xenografts. *Scientific reports*, 10(1):1–11, 2020.

Eric S Lander. Initial impact of the sequencing of the human genome. *Nature*, 470(7333):187–197, 2011.

Eric S Lander, Lauren M Linton, Bruce Birren, Chad Nusbaum, Michael C Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William FitzHugh, et al. Initial sequencing and analysis of the human genome. 2001.

Nicolas Le Novere. Quantitative and logic modelling of molecular and gene networks. *Nature Reviews Genetics*, 16(3):146–158, 2015.

Celine Lefebvre, Presha Rajbhandari, Mariano J. Alvarez, Pradeep Bandaru, Wei Keat Lim, Mai Sato, Kai Wang, Pavel Sumazin, Manjunath Kustagi, Brygida C. Bisikirska, Katia Basso, Pedro Beltrao, Nevan Krogan, Jean Gautier, Riccardo Dalla-Favera, and Andrea Califano. A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers. *Molecular Systems Biology*, 6:377, June 2010. ISSN 1744-4292. doi: 10.1038/msb.2010.31.

Gaelle Letort, Arnau Montagud, Gautier Stoll, Randy Heiland, Emmanuel Barillot, Paul Macklin, Andrei Zinovyev, and Laurence Calzone. Physiboss: a multi-scale agent-based modelling framework integrating physical dimension and cell signalling. *Bioinformatics*, 35(7):1188–1196, 2019.

Jianfang Liu, Tara Lichtenberg, Katherine A Hoadley, Laila M Poisson, Alexander J Lazar, Andrew D Cherniack, Albert J Kovatich, Christopher C Benz, Douglas A Levine, Adrian V Lee, et al. An integrated tcga pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell*, 173(2):400–416, 2018.

AJ Lotka. Principles of physical biology. *Baltimore: Waverly*, 1925.

Matteo Manica, Ali Oskooei, Jannis Born, Vigneshwari Subramanian, Julio Sáez-Rodríguez, and María Rodríguez Martínez. Toward explainable anticancer compound sensitivity prediction via multimodal attention-based convolutional encoders. *Molecular Pharmaceutics*, 2019.

Adam A. Margolin, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo Dalla Favera, and Andrea Califano. ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. *BMC Bioinformatics*, 7(1):S7, March 2006. ISSN 1471-2105. doi: 10.1186/1471-2105-7-S1-S7. URL https://doi.org/10.1186/1471-2105-7-S1-S7.

Nick I Markevich, Jan B Hoek, and Boris N Kholodenko. Signaling switches and bistability arising from multisite phosphorylation in protein kinase cascades. *The Journal of cell biology*, 164(3):353–359, 2004.

Loredana Martignetti, Laurence Calzone, Eric Bonnet, Emmanuel Barillot, and Andrei Zinovyev. Roma: representation and quantification of module activity from target expression data. *Frontiers in genetics*, 7:18, 2016.

Nicolai Meinshausen, Alain Hauser, Joris M Mooij, Jonas Peters, Philip Versteeg, and Peter Bühlmann. Methods for causal inference from gene

perturbation experiments and validation. *Proceedings of the National Academy of Sciences*, 113(27):7361–7368, 2016.

Craig H. Mermel, Steven E. Schumacher, Barbara Hill, Matthew L. Meyerson, Rameen Beroukhim, and Gad Getz. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biology*, 12:R41, April 2011. ISSN 1474-760X. doi: 10.1186/gb-2011-12-4-r41. URL http://dx.doi.org/10.1186/gb-2011-12-4-r41.

Robin M Meyers, Jordan G Bryan, James M McFarland, Barbara A Weir, Ann E Sizemore, Han Xu, Neekesh V Dharia, Phillip G Montgomery, Glenn S Cowley, Sasha Pantel, et al. Computational correction of copy number effect improves specificity of crispr–cas9 essentiality screens in cancer cells. *Nature genetics*, 49(12):1779–1784, 2017.

Matthew Meyerson, Stacey Gabriel, and Gad Getz. Advances in understanding cancer genomes through second-generation sequencing. *Nature Reviews Genetics*, 11(10):685–696, 2010.

Iain Miller, Mingwei Min, Chen Yang, Chengzhe Tian, Sara Gookin, Dylan Carter, and Sabrina L Spencer. Ki67 is a graded rather than a binary marker of proliferation versus quiescence. *Cell reports*, 24(5):1105–1112, 2018.

Arnau Montagud, Pauline Traynard, Loredana Martignetti, Eric Bonnet, Emmanuel Barillot, Andrei Zinovyev, and Laurence Calzone. Conceptual and computational framework for logical modelling of biological networks deregulated in diseases. *Briefings in Bioinformatics*, 2017. doi: 10.1093/bib/bbx163.

Á C Murphy, Birgit Weyhenmeyer, Jasmin Schmid, Seán M Kilbride, Markus Rehm, Heinrich J Huber, C Senft, J Weissenberger, V Seifert, M Dunst, et al. Activation of executioner caspases is a predictor of progression-free survival in glioblastoma patients: a systems medicine approach. *Cell death & disease*, 4(5):e629–e629, 2013.

Christoph Müssel, Martin Hopfensitz, and Hans A Kestler. Boolnet?an r package for generation, reconstruction and analysis of boolean networks. *Bioinformatics*, 26(10):1378–1380, 2010.

Aurélien Naldi. Biolqm: a java toolkit for the manipulation and conversion of logical qualitative models of biological networks. *Frontiers in physiology*, 9:1605, 2018.

Aurélien Naldi, Céline Hernandez, Wassim Abou-Jaoudé, Pedro T Monteiro, Claudine Chaouiya, and Denis Thieffry. Logical modeling and analysis of cellular regulatory networks with ginsim 3.0. *Frontiers in physiology*, 9, 2018a.

Aurélien Naldi, Céline Hernandez, Nicolas Levy, Gautier Stoll, Pedro T Monteiro, Claudine Chaouiya, Tomáš Helikar, Andrei Zinovyev, Laurence Calzone, Sarah Cohen-Boulakia, et al. The colomoto interactive notebook: accessible and reproducible computational analyses for qualitative biological networks. *Frontiers in physiology*, 9:680, 2018b.

Nicholas E Navin. The first five years of single-cell cancer genomics and beyond. *Genome research*, 25(10):1499–1507, 2015.

Cancer Genome Atlas Network et al. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61, 2012.

Cancer Genome Atlas Research Network et al. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061, 2008.

Chiara Nicolò, Cynthia Périer, Melanie Prague, Carine Bellera, Gaëtan MacGrogan, Olivier Saut, and Sébastien Benzekry. Machine learning and mechanistic modeling for prediction of metastatic relapse in early-stage breast cancer. *JCO Clinical Cancer Informatics*, 4:259–274, 2020.

Nicolas Le Novère, Michael Hucka, Huaiyu Mi, Stuart Moodie, Falk Schreiber, Anatoly Sorokin, Emek Demir, Katja Wegner, Mirit I. Aladjem, Sarala M. Wimalaratne, Frank T. Bergman, Ralph Gauges, Peter Ghazal, Hideya Kawaji, Lu Li, Yukiko Matsuoka, Alice Villeger, Sarah E. Boyd, Laurence Calzone, Melanie Courtot, Ugur Dogrusoz, Tom C. Freeman, Akira Funahashi, Samik Ghosh, Akiya Jouraku, Sohyoung Kim, Fedor Kolpakov, Augustin Luna, Sven Sahle, Esther Schmidt, Steven Watterson, Guanming Wu, Igor Goryanin, Douglas B. Kell, Chris Sander, Herbert Sauro, Jacky L. Snoep, Kurt Kohn, and Hiroaki Kitano. The Systems Biology Graphical Notation. *Nature Biotechnology*, 27(8): 735–741, August 2009. ISSN 1546-1696. doi: 10.1038/nbt.1558. URL https://www.nature.com/articles/nbt.1558.

Peter C Nowell. The clonal evolution of tumor cell populations. *Science*, 194(4260):23–28, 1976.

CNAM Oldenhuis, SF Oosting, JA Gietema, and EGE De Vries. Prognostic versus predictive value of biomarkers in oncology. *European Journal of Cancer*, 44(7):946–953, 2008.

M. Ostrowski, L. Paulevé, T. Schaub, A. Siegel, and C. Guziolowski. Boolean network identification from perturbation time series data combining dynamics abstraction and logic programming. *Biosystems*, 149: 139–153, November 2016. ISSN 0303-2647. doi: 10.1016/j.biosystems.2016.07.009.

Loïc Paulevé. Pint: A Static Analyzer for Transient Dynamics of Qualitative Networks with IPython Interface. In *CMSB 2017 - 15th conference on Computational Methods for Systems Biology*, volume 10545 of *Lecture Notes in Computer Science*, pages 370 – 316. Springer, 2017. doi: 10.1007/978-3-319-67471-1_20.

T Pawson and N Warner. Oncogenic re-wiring of cellular signaling pathways. *Oncogene*, 26(9):1268–1275, 2007.

Bernard Pereira, Suet-Feung Chin, Oscar M. Rueda, Hans-Kristian Moen Vollan, Elena Provenzano, Helen A. Bardwell, Michelle Pugh, Linda Jones, Roslin Russell, Stephen-John Sammut, Dana W. Y. Tsui, Bin Liu, Sarah-Jane Dawson, Jean Abraham, Helen Northen, John F. Peden, Abhik Mukherjee, Gulisa Turashvili, Andrew R. Green, Steve McKinney, Arusha Oloumi, Sohrab Shah, Nitzan Rosenfeld, Leigh Murphy, David R. Bentley, Ian O. Ellis, Arnie Purushotham, Sarah E. Pinder, Anne-Lise Børresen-Dale, Helena M. Earl, Paul D. Pharoah, Mark T. Ross, Samuel Aparicio, and Carlos Caldas. The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. *Nature Communications*, 7, May 2016. ISSN 2041-1723. doi: 10.1038/ncomms11479. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4866047/.

Livia Perfetto, Leonardo Briganti, Alberto Calderone, Andrea Cerquone Perpetuini, Marta Iannuccelli, Francesca Langone, Luana Licata, Milica Marinkovic, Anna Mattioni, Theodora Pavlidou, et al. Signor: a database of causal relationships between biological entities. *Nucleic acids research*, 44(D1):D548–D554, 2016.

Charles M Perou, Stefanie S Jeffrey, Matt Van De Rijn, Christian A Rees, Michael B Eisen, Douglas T Ross, Alexander Pergamenschikov, Cheryl F Williams, Shirley X Zhu, Jeffrey CF Lee, et al. Distinctive gene expression

patterns in human mammary epithelial cells and breast cancers. *Proceedings of the National Academy of Sciences*, 96(16):9212–9217, 1999.

Charles M Perou, Therese Sørlie, Michael B Eisen, Matt Van De Rijn, Stefanie S Jeffrey, Christian A Rees, Jonathan R Pollack, Douglas T Ross, Hilde Johnsen, Lars A Akslen, et al. Molecular portraits of human breast tumours. *nature*, 406(6797):747–752, 2000.

Athanasios Polynikis, SJ Hogan, and Mario di Bernardo. Comparing different ode modelling approaches for gene regulatory networks. *Journal of theoretical biology*, 261(4):511–530, 2009.

Angela Potochnik. *Idealization and the Aims of Science*. University of Chicago Press, 2017.

Gibin G Powathil, Maciej Swat, and Mark AJ Chaplain. Systems oncology: towards patient-specific treatment regimes informed by multiscale mathematical modelling. In *Seminars in cancer biology*, volume 30, pages 13–20. Elsevier, 2015.

Misbah Razzaq, Loïc Paulevé, Anne Siegel, Julio Saez-Rodriguez, Jérémie Bourdon, and Carito Guziolowski. Computational discovery of dynamic cell line specific boolean networks from multiplex time-course data. *PLoS computational biology*, 14(10):e1006538, 2018.

E Premkumar Reddy, Roberta K Reynolds, Eugenio Santos, and Mariano Barbacid. A point mutation is responsible for the acquisition of transforming properties by the t24 human bladder carcinoma oncogene. *Nature*, 300 (5888):149–152, 1982.

Elisabeth Remy, Sandra Rebouissou, Claudine Chaouiya, Andrei Zinovyev, François Radvanyi, and Laurence Calzone. A modeling approach to explain mutually exclusive and co-occurring genetic alterations in bladder tumorigenesis. *Cancer research*, 75(19):4042–4052, 2015.

Jason A Reuter, Damek V Spacek, and Michael P Snyder. High-throughput sequencing technologies. *Molecular cell*, 58(4):586–597, 2015.

Arturo Rosenblueth and Norbert Wiener. The role of models in science. *Philosophy of science*, 12(4):316–321, 1945.

Julio Saez-Rodriguez, Leonidas G Alexopoulos, MingSheng Zhang, Melody K Morris, Douglas A Lauffenburger, and Peter K Sorger. Comparing signaling networks between normal and transformed hepatocytes using discrete logical models. *Cancer research*, 71(16):5400–5411, 2011a.

Julio Saez-Rodriguez, Leonidas G. Alexopoulos, MingSheng Zhang, Melody K. Morris, Douglas A. Lauffenburger, and Peter K. Sorger. Comparing Signaling Networks between Normal and Transformed Hepatocytes Using Discrete Logical Models. *Cancer Research*, 71(16):5400–5411, August 2011b. ISSN 0008-5472, 1538-7445. doi: 10.1158/0008-5472.CAN-10-4453. URL http://cancerres.aacrjournals.org/content/71/16/5400.

Özgür Sahin, Holger Fröhlich, Christian Löbke, Ulrike Korf, Sara Burmester, Meher Majety, Jens Mattern, Ingo Schupp, Claudine Chaouiya, Denis Thieffry, et al. Modeling erbb receptor-regulated g1/s transition to find novel targets for de novo trastuzumab resistance. *BMC systems biology*, 3(1):1, 2009.

Manuela Salvucci, Maximilian L Würstle, Clare Morgan, Sarah Curry, Mattia Cremona, Andreas U Lindner, Orna Bacon, Alexa J Resler, Aine C Murphy, Robert O'Byrne, et al. A stepwise integrated approach to personalized risk predictions in stage iii colorectal cancer. *Clinical Cancer Research*, 23(5):1200–1212, 2017.

Manuela Salvucci, Arman Rahman, Alexa J Resler, Girish M Udupi, Deborah A McNamara, Elaine W Kay, Pierre Laurent-Puig, Daniel B Longley, Patrick G Johnston, Mark Lawler, et al. A machine learning platform to optimize the translation of personalized network models to the clinic. *JCO clinical cancer informatics*, 3:1–17, 2019a.

Manuela Salvucci, Zaitun Zakaria, Steven Carberry, Amanda Tivnan, Volker Seifert, Donat Kögel, Brona M Murphy, and Jochen HM Prehn. System-based approaches as prognostic tools for glioblastoma. *BMC cancer*, 19(1):1092, 2019b.

Yardena Samuels, Zhenghe Wang, Alberto Bardelli, Natalie Silliman, Janine Ptak, Steve Szabo, Hai Yan, Adi Gazdar, Steven M Powell, Gregory J Riggins, et al. High frequency of mutations of the pik3ca gene in human cancers. *Science*, 304(5670):554–554, 2004.

Francisco Sanchez-Vega, Marco Mina, Joshua Armenia, Walid K Chatila, Augustin Luna, Konnor C La, Sofia Dimitriadoy, David L Liu, Havish S Kantheti, Sadegh Saghafinia, et al. Oncogenic signaling pathways in the cancer genome atlas. *Cell*, 173(2):321–337, 2018.

Charles L Sawyers. The cancer biomarker problem. *Nature*, 452(7187):548–552, 2008.

Zsofia K Stadler, Kasmintan A Schrader, Joseph Vijai, Mark E Robson, and Kenneth Offit. Cancer genomics and inherited risk. *Journal of Clinical Oncology*, 32(7):687, 2014.

Steven Nathaniel Steinway, Jorge Gomez Tejeda Zañudo, Paul J Michel, David J Feith, Thomas P Loughran, and Reka Albert. Combinatorial interventions inhibit tgfβ-driven epithelial-to-mesenchymal transition and support hybrid cellular phenotypes. *NPJ systems biology and applications*, 1:15014, 2015.

Gautier Stoll, Eric Viara, Emmanuel Barillot, and Laurence Calzone. Continuous time boolean modeling for biological signaling: application of gillespie algorithm. *BMC systems biology*, 6(1):116, 2012.

Gautier Stoll, Barthélémy Caron, Eric Viara, Aurélien Dugourd, Andrei Zinovyev, Aurélien Naldi, Guido Kroemer, Emmanuel Barillot, and Laurence Calzone. Maboss 2.0: an environment for stochastic boolean modeling. *Bioinformatics*, 33(14):2226–2228, 2017.

Michael R Stratton, Peter J Campbell, and P Andrew Futreal. The cancer genome. *Nature*, 458(7239):719–724, 2009.

Kristin R Swanson, Carly Bridge, JD Murray, and Ellsworth C Alvord Jr. Virtual and real brain tumors: using mathematical modeling to quantify glioma growth and invasion. *Journal of the neurological sciences*, 216(1): 1–10, 2003.

Camille Terfve, Thomas Cokelaer, David Henriques, Aidan MacNamara, Emanuel Goncalves, Melody K. Morris, Martijn van Iersel, Douglas A. Lauffenburger, and Julio Saez-Rodriguez. CellNOptR: a flexible toolkit to train protein signaling networks to data using multiple logic formalisms. *BMC Systems Biology*, 6(1):133, October 2012. ISSN 1752-0509. doi: 10.1186/1752-0509-6-133.

Camille D. A. Terfve, Edmund H. Wilkes, Pedro Casado, Pedro R. Cutillas, and Julio Saez-Rodriguez. Large-scale models of signal propagation in human cells derived from discovery phosphoproteomic data. *Nature Communications*, 6:8033, September 2015. ISSN 2041-1723. doi: 10.1038/ncomms9033. URL https://www.nature.com/articles/ncomms9033.

René Thomas. Boolean formalization of genetic control circuits. *Journal of theoretical biology*, 42(3):563–585, 1973.

René Thomas and Richard d'Ari. *Biological feedback*. CRC press, 1990.

Collin J. Tokheim, Nickolas Papadopoulos, Kenneth W. Kinzler, Bert Vogelstein, and Rachel Karchin. Evaluating the evaluation of cancer driver genes. *Proceedings of the National Academy of Sciences*, 113(50):14330–14335, December 2016. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1616440113. URL http://www.pnas.org/content/113/50/14330.

Cristian Tomasetti and Bert Vogelstein. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science*, 347(6217):78–81, 2015.

Cristian Tomasetti, Luigi Marchionni, Martin A Nowak, Giovanni Parmigiani, and Bert Vogelstein. Only three driver gene mutations are required for the development of lung and colorectal cancers. *Proceedings of the National Academy of Sciences*, 112(1):118–123, 2015.

Scott A Tomlins, Daniel R Rhodes, Sven Perner, Saravana M Dhanasekaran, Rohit Mehra, Xiao-Wei Sun, Sooryanarayana Varambally, Xuhong Cao, Joelle Tchinda, Rainer Kuefer, et al. Recurrent fusion of tmprss2 and ets transcription factor genes in prostate cancer. *science*, 310(5748):644–648, 2005.

Christophe Trefois, Paul MA Antony, Jorge Goncalves, Alexander Skupin, and Rudi Balling. Critical transitions in chronic disease: transferring concepts from ecology to systems medicine. *Current opinion in biotechnology*, 34:48–55, 2015.

Dénes Türei, Tamás Korcsmáros, and Julio Saez-Rodriguez. Omnipath: guidelines and gateway for literature-curated signaling pathway resources. *Nature methods*, 13(12):966, 2016.

John J Tyson, Katherine C Chen, and Bela Novak. Sniffers, buzzers, toggles and blinkers: dynamics of regulatory and signaling pathways in the cell. *Current opinion in cell biology*, 15(2):221–231, 2003.

John J Tyson, William T Baumann, Chun Chen, Anael Verdugo, Iman Tavassoly, Yue Wang, Louis M Weiner, and Robert Clarke. Dynamic modelling of oestrogen signalling and cell fate in breast cancer cells. *Nature Reviews Cancer*, 11(7):523–532, 2011.

Dieudonne van der Meer, Syd Barthorpe, Wanjuan Yang, Howard Lightfoot, Caitlin Hall, James Gilbert, Hayley E Francies, and Mathew J Garnett. Cell model passports—a hub for clinical, genetic and functional datasets of preclinical cancer models. *Nucleic acids research*, 47(D1):D923–D929, 2019.

Laura J Van't Veer, Hongyue Dai, Marc J Van De Vijver, Yudong D He, Augustinus AM Hart, Mao Mao, Hans L Peterse, Karin Van Der Kooy, Matthew J Marton, Anke T Witteveen, et al. Gene expression profiling predicts clinical outcome of breast cancer. *nature*, 415(6871):530–536, 2002.

David Venet, Jacques E Dumont, Vincent Detours, et al. Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS Comput Biol*, 7(10):e1002240, 2011.

J Craig Venter, Mark D Adams, Eugene W Myers, Peter W Li, Richard J Mural, Granger G Sutton, Hamilton O Smith, Mark Yandell, Cheryl A Evans, Robert A Holt, et al. The sequence of the human genome. *science*, 291(5507):1304–1351, 2001.

Louis Verny, Nadir Sella, Séverine Affeldt, Param Priya Singh, and Hervé Isambert. Learning causal networks with latent variables from multivariate information in genomic data. *PLOS Computational Biology*, 13 (10):e1005662, October 2017. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1005662. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005662.

Jean-Philippe Vert, Jian Qiu, and William S. Noble. A new pairwise kernel for biological network inference with support vector machines. *BMC Bioinformatics*, 8(Suppl 10):S8, 2007.

Santiago Videla, Julio Saez-Rodriguez, Carito Guziolowski, and Anne Siegel. caspo: a toolbox for automated reasoning on the response of logical signaling networks families. *Bioinformatics*, 33(6):947–950, March 2017. ISSN 1367-4803. doi: 10.1093/bioinformatics/btw738. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5351548/.

Alejandro F Villaverde and Julio R Banga. Reverse engineering and identification in systems biology: strategies, perspectives and challenges. *Journal of the Royal Society Interface*, 11(91):20130505, 2014.

Daniel J Vis, Lorenzo Bombardelli, Howard Lightfoot, Francesco Iorio, Mathew J Garnett, and Lodewyk FA Wessels. Multilevel models improve precision and speed of ic50 estimates. *Pharmacogenomics*, 17(7): 691–700, 2016.

Vito Volterra. Fluctuations in the abundance of a species considered mathematically, 1926.

Emily A Vucic, Kelsie L Thu, Keith Robison, Leszek A Rybaczyk, Raj Chari, Carlos E Alvarez, and Wan L Lam. Translating cancer 'omics' to improved outcomes. *Genome research*, 22(2):188–195, 2012.

Frederick YM Wan. *Mathematical models and their analysis*, volume 79. SIAM, 2018.

Jing Wang, Sijin Wen, W. Fraser Symmans, Lajos Pusztai, and Kevin R. Coombes. The Bimodality Index: A Criterion for Discovering and Ranking Bimodal Signatures from Cancer Gene Expression Profiling Data. *Cancer Informatics*, 7:199–216, August 2009. ISSN 1176-9351. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2730180/.

James D Watson and Francis HC Crick. Molecular structure of nucleic acids. *Nature*, 171(4356):737–738, 1953.

Robert A Weinberg. *The biology of cancer*. Garland science, 2013.

Nathan Weinstein, Luis Mendoza, Isidoro Gitler, and Jaime Klapp. A network model to explore the effect of the micro-environment on endothelial cell behavior during angiogenesis. *Frontiers in physiology*, 8:960, 2017.

Steven Woodhouse, Nir Piterman, Christoph M Wintersteiger, Berthold Göttgens, and Jasmin Fisher. Scns: a graphical tool for reconstructing executable regulatory networks from single-cell genomic data. *BMC systems biology*, 12(1):59, 2018.

Wanjuan Yang, Jorge Soares, Patricia Greninger, Elena J Edelman, Howard Lightfoot, Simon Forbes, Nidhi Bindal, Dave Beare, James A Smith, I Richard Thompson, et al. Genomics of drug sensitivity in cancer (gdsc): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic acids research*, 41(D1):D955–D961, 2012.

Jorge Gómez Tejeda Zañudo, Maurizio Scaltriti, and Réka Albert. A network modeling approach to elucidate drug resistance mechanisms and predict combinatorial drug treatments in breast cancer. *Cancer convergence*, 1(1):5, 2017.

RÉSUMÉ

Au delà de ses mécanismes génétiques, le cancer peut-être compris comme une maladie de réseaux qui résulte souvent de l'interaction entre différentes perturbations dans un réseau de régulation cellulaire. La dynamique de ces réseaux et des voies de signalisation associées est complexe et requiert des approches intégrées. Une d'entre elles est la conception de modèles dits mécanistiques qui traduisent mathématiquement la connaissance biologique des réseaux afin de pouvoir simuler le fonctionnement moléculaire des cancers informatiquement. Ces modèles ne traduisent cependant que les mécanismes généraux à l'oeuvre dans certains cancers en particulier.
Cette thèse propose en premier lieu de définir des modèles mécanistiques personnalisés de cancer. Un modèle générique est d'abord défini dans un formalisme logique (ou Booléen), avant d'utiliser les données omiques (mutations, ARN, protéines) de patients ou de lignées cellulaires afin de rendre le modèle spécifique à chacun. Ces modèles personnalisés peuvent ensuite être confrontés aux données cliniques de patients pour vérifier leur validité. Le cas de la réponse clinique aux traitements est exploré en particulier dans cette thèse. La représentation explicite des mécanismes moléculaires par ces modèles permet en effet de simuler l'effet de différents traitements suivant leur mode d'action et de vérifier si la sensibilité d'un patient à un traitement est bien prédite par le modèle personnalisé correspondant. Un exemple concernant la réponse aux inhibiteurs de BRAF dans les mélanomes et cancers colorectaux est ainsi proposé.
La confrontation des modèles mécanistiques de cancer, ceux présentés dans cette thèse et d'autres, aux données cliniques incite par ailleurs à évaluer rigoureusement leurs éventuels bénéfices dans la cadre d'une utilisation médicale. La quantification et l'interprétation de la valeur de certains modèles à visée pronostique est brièvement présentée avant de se focaliser sur le cas particulier des modèles capables de sélectionner le meilleur traitement pour chaque patient en fonction des ses caractéristiques moléculaires. Un cadre théorique est proposé pour étendre les méthodes d'inférence causale à l'évaluation de tels algorithmes de médecine de précision. Une illustration est fournie à l'aide de données simulées et de xénogreffes dérivées de patients
L'ensemble des méthodes et applications décrites tracent donc un chemin, de la conception de modèles mécanistiques de cancer à leur évaluation grâce à des modèles statistiques émulant des essais cliniques.

MOTS CLÉS

Modélisation, Cancer, Modèle mécanistique, Biostatistiques, Inférence causale, Médecine de précision.

ABSTRACT

Beyond its genetic mechanisms, cancer can be understood as a network disease that often results from the interaction between different perturbations in a cellular regulatory network. The dynamics of these networks and associated signaling pathways are complex and require integrated approaches. One approach is to design mechanistic models that translate the biological knowledge of networks in mathematical terms to simulate the molecular features of cancers in a computer-readable form. However, these models only reflect the general mechanisms at work in cancers.
This thesis proposes to define personalized mechanistic models of cancer. A generic model is first defined in a logical (or Boolean) formalism, before using omics data (mutations, RNA, proteins) from patients or cell lines in order to make the model specific to each one profile. These personalized models can then be compared with the clinical data of patients in order to validate them. The response to treatment is investigated in particular in this thesis. The explicit representation of the molecular mechanisms by these models allows to simulate the effect of different treatments according to their targets and to verify if the sensitivity of a patient to a drug is well predicted by the corresponding personalized model. An example concerning the response to BRAF inhibitors in melanomas and colorectal cancers is thus presented.
The comparison of mechanistic models of cancer, those presented in this thesis and others, with clinical data also encourages a rigorous evaluation of their possible benefits in the context of medical use. The quantification and interpretation of the value of certain prognostic models is briefly presented before focusing on the particular case of models able to recommend the best treatment for each patient according to his molecular profile. A theoretical framework is defined to extend causal inference methods to the evaluation of such precision medicine algorithms. An illustration is provided using simulated data and patient derived xenografts.
All the methods and applications put forward a possible path from the design of mechanistic models of cancer to their evaluation using statistical models emulating clinical trials.

KEYWORDS

Modeling, Cancer, Mechanistic model, Biostatistics, Causal inference, Precision medicine.